Methods in Molecular Biology 1115

Springer Protocols

Pascale Besse *Editor*

Molecular Plant Taxonomy

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Molecular Plant Taxonomy

Methods and Protocols

Edited by

Pascale Besse

UMR C53 PVBMT Université de la Réunion – Cirad, Université de la Réunion, Ile de la Réunion, France

Humana Press

Editor
Pascale Besse
UMR C53 PVBMT Université de la Réunion – Cirad
Université de la Réunion
Ile de la Réunion, France

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-62703-766-2 ISBN 978-1-62703-767-9 (eBook)
DOI 10.1007/978-1-62703-767-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013957132

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Plant taxonomy is an ancient discipline nowadays facing new challenges with the availability of a vast array of molecular approaches allowing reliable genealogy-based classifications. Although the primary focus of plant taxonomy is on the delimitation of species, molecular approaches also provide a better understanding of evolutionary processes, a particularly important issue for some taxonomic complex groups. This book describes laboratory protocols based on the use of nucleic acids and chromosomes for plant taxonomy, as well as guidelines for phylogenetic analysis of molecular data. It also provides introductive and application review chapters, which are of great importance to put the protocols into a broader perspective.

A thorough historical overview of plant taxonomy is provided in the introductive chapter by *Germinal Rouhan* and *Myriam Gaudeuil* (Chapter 1). Strengths, limitations and the future of molecular techniques with regard to plant taxonomy are also explored, as compared to the use of classical morphological and anatomical data. Guidelines are then given by *Pascale Besse* (Chapter 2) to choose the best appropriate molecular technique depending on the plant taxonomic survey envisaged. Both chapters are prerequisite readings to understand the concepts underlying the "plant taxonomy" discipline and to fully appreciate the strengths and limits of each molecular technique presented in this book.

One of the advantages of molecular techniques for plant taxonomy is that analyses can be performed at early developmental stages, from living plant material as well as from voucher herbarium specimens. This allows an "integrative" approach combining modern molecular data with taxonomic description of reference species. Two chapters describe protocols for DNA extraction. *Kessa Semagn* (Chapter 3) provides various leaf tissue sampling methods particularly handy for field collection and a reliable DNA extraction protocol. *Lenka Záveská Drábková* (Chapter 4) proposes various DNA extraction protocols specifically designed for dried herbarium specimens.

Chapters 5–7 present protocols for classical sequencing of various genomic sequences commonly used in plant taxonomic and phylogenetic studies. Chloroplast DNA is one of the most commonly used DNA in plant taxonomy, *Berthold Heinze, Agnieszka Koziel-Monte* and *Daniela Jahn* (Chapter 5) describe primers and protocols used for the sequencing of chloroplast DNA as well as other useful methods (PCR-RFLP and dHPLC). Although mitochondrial DNA was given less attention as a target for plant taxonomy, *Jérôme Duminil* (Chapter 6) demonstrates how useful such DNA regions can be to resolve taxonomic issues and provide a detailed sequencing protocol. Finally, the internal transcribed spacer (ITS) of the nuclear ribosomal RNA genes is a region of choice to be sequenced in taxonomic studies. The advantages and possible drawbacks of using this region are presented, and a sequencing protocol is provided by *Pascale Besse* (Chapter 7). Recently, the availability of new DNA sequencing (NGS—next generation sequencing) and high-troughput genotyping methods

has opened a new era for plant molecular taxonomy. *David Edwards, Manuel Zander, Jessica Dalton-Morgan* and *Jacqueline Batley* (Chapter 8) provide detailed protocols for Single Nucleotide Polymorphisms (SNPs) discovery and genotyping.

Simple and reliable PCR-based methods other than direct sequencing can be utilized to provide useful molecular markers for resolving plant taxonomic issues. They rely on polymerase chain reaction (PCR) amplification of specific or anonymous regions in the genome and their analysis following electrophoretic size-fractioning to reveal size (length) variations. The isolation of microsatellite loci and their study (by PCR amplification) is a very important component of the molecular tool-kit available for plant taxonomy, particularly at low taxonomic levels (generally at the population level). *Hélène Vignes* and *Ronan Rivallan* (Chapter 9) provide a protocol to isolate such regions through the construction of microsatellite-enriched libraries. Relatively high-throughput multi-locus genotyping methods are also available. Some of these methods target anonymous regions of the genome without the need for any prior knowledge on any of the taxon DNA sequences. *Kantipudi Nirmal Babu, Muliyar Krishna Rajesh, Kukkumgai Samsudeen, Divakaran Minoo, Erinjery Jose Suraby, Kallayan Anupama and Paul Ritto* (Chapter 10) describe protocols for the simplest of these methods: randomly amplified polymorphic DNA (RAPD) as well as more recent derived techniques. *Luis F. Goulao* and *Cristina M. Oliveira* (Chapter 11) provide detailed protocols for other methods such as amplified fragment length polymorphism (AFLP) and derived techniques based on simultaneous microsatellite amplification such as inter simple sequence repeats (ISSR) and selective amplification of microsatellite polymorphic loci (SAMPL). *Ruslan Kalendar* and *Alan Schulman* (Chapter 12) describe the usefulness and protocols for multi-locus tagging of LTR (long terminal repeats)-retrotransposons present in plant genomes such as inter-retrotransposon amplification polymorphism (IRAP), retrotransposon-microsatellite amplification polymorphism (REMAP) and inter-Primer Binding Site Polymorphism (iPBS).

Finally, *Alexandre De Bruyn, Darren P. Martin* and *Pierre Lefeuvre* (Chapter 13) provide detailed and step-by-step guidelines to the analyses of molecular data using phylogenetic reconstruction methods based on DNA sequences, using freely available computer programs.

As hybridization and polyploidization are very important components of plant speciation particularly for some plant groups, specific cytogenetic techniques may need to be developed in addition to molecular studies, as detailed in Chapters 14–16. *Jaume Pellicer* and *Ilia J Leitch* (Chapter 14) provide protocols for plant genome size estimation using flow cytometry, a technique that can be combined with direct chromosomal observations as in fluorochrome banding and FISH (fluorescent in situ hybridization), for which detailed protocols are given by *Sonja Siljak-Yakovlev, Fatima Pustahija, Vedrana Vicic* and *Odile Robin* (Chapter 15). These can help to determine ploidy level and assess genome organisation. Moreover, *Nathalie Piperidis* (Chapter 16) details the GISH (genomic in situ hybridisation) protocol, a powerful technique that can be used to resolve the parental origin of inter-specific and inter-generic hybrid plant species.

The final two chapters are examples of applications on the use of molecular data for the taxonomic revision of some plant groups: Malvaceae, by *Timothee Le Péchon and Luc Gigord* (Chapter 17) and Proteaceae, by *Peter H. Weston* (Chapter 18). These provide a clear illustration of the exciting contributions molecular approaches can make to plant taxonomy.

The reviews and protocols that appear as chapters in this book were selected to provide conceptual as well as technical guidelines to plant taxonomists and geneticists. Despite the present craze for "DNA barcoding", it is now clear that using solely the two chloroplastic genes *mat*K and *rbc*L for resolving plant taxonomy will not be sufficient, particularly in

some plant groups. We rather highly recommend that molecular techniques are used in an "integrative taxonomy" approach, combining nucleic acid and cytogenetic data together with other crucial information (taxonomy, morphology, anatomy, ecology, reproductive biology, biogeography, paleobotany), which will help not only to best circumvent species delimitation but as well to resolve the evolutionary processes in play. This is of great importance as speciation is indeed a dynamic process.

Ile de la Réunion, France *Pascale Besse*

Contents

Preface . . . . . *v*
Contributors . . . . . *xi*

1 Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives 1
Germinal Rouhan and Myriam Gaudeul

2 Guidelines for the Choice of Sequences for Molecular Plant Taxonomy 39
Pascale Besse

3 Leaf Tissue Sampling and DNA Extraction Protocols 53
Kassa Semagn

4 DNA Extraction from Herbarium Specimens 69
Lenka Záveská Drábková

5 Analysis of Variation in Chloroplast DNA Sequences 85
Berthold Heinze, Agnieszka Koziel-Monte, and Daniela Jahn

6 Mitochondrial Genome and Plant Taxonomy 121
Jérôme Duminil

7 Nuclear Ribosomal RNA Genes: ITS Region 141
Pascale Besse

8 New Technologies for Ultrahigh-Throughput Genotyping in Plant Taxonomy 151
David Edwards, Manuel Zander, Jessica Dalton-Morgan, and Jacqueline Batley

9 Development of Microsatellite-Enriched Libraries 177
Hélène Vignes and Ronan Rivallan

10 Randomly Amplified Polymorphic DNA (RAPD) and Derived Techniques . . . 191
Kantipudi Nirmal Babu, Muliyar Krishna Rajesh, Kukkumgai Samsudeen, Divakaran Minoo, Erinjery Jose Suraby, Kallayan Anupama, and Paul Ritto

11 Multilocus Profiling with AFLP, ISSR, and SAMPL 211
Luis F. Goulao and Cristina M. Oliveira

12 Transposon-Based Tagging: IRAP, REMAP, and iPBS 233
Ruslan Kalendar and Alan H. Schulman

13 Phylogenetic Reconstruction Methods: An Overview 257
Alexandre De Bruyn, Darren P. Martin, and Pierre Lefeuvre

14 The Application of Flow Cytometry for Estimating Genome Size and Ploidy Level in Plants 279
Jaume Pellicer and Ilia J. Leitch

15 Molecular Cytogenetics (FISH and Fluorochrome Banding): Resolving Species Relationships and Genome Organization 309
Sonja Siljak-Yakovlev, Fatima Pustahija, Vedrana Vicic, and Odile Robin

16 GISH: Resolving Interspecific and Intergeneric Hybrids 325
Nathalie Piperidis

17 On the Relevance of Molecular Tools for Taxonomic Revision in Malvales, Malvaceae s.l., and Dombeyoideae 337
Timothée Le Péchon and Luc D.B. Gigord

18 What Has Molecular Systematics Contributed to Our Knowledge of the Plant Family Proteaceae? 365
Peter H. Weston

Index *399*

Contributors

Kallayan Anupama • *Indian Institute of Spices Research, Kozhikode, Kerala, India*

Jacqueline Batley • *School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD, Australia; ARC Centre for Integrated Legume Research, University of Queensland, Brisbane, QLD, Australia*

Pascale Besse • *UMR C53 PVBMT Université de la Réunion – Cirad, Université de la Réunion, Ile de la Réunion, France*

Alexandre De Bruyn • *Pôle de Protection des Plantes, CIRAD, UMR PVBMT, Université de la Réunion, Saint-Pierre, France*

Jessica Dalton-Morgan • *School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD, Australia; ARC Centre for Integrated Legume Research, University of Queensland, Brisbane, QLD, Australia*

Lenka Záveská Drábková • *Department of Taxonomy, Institute of Botany, Academy of Sciences of the Czech Republic, Pruhonice, Czech Republic*

Jerôme Duminil • *Service Evolution Biologique et Ecologie, CP160/12, Faculté des Sciences, Université Libre de Bruxelles, Brussels, Belgium; Biodiversity International, Forest Genetic Resources Programme, Sub-Regional Office for Central Africa, Yaoundé, Cameroon*

David Edwards • *School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD, Australia; Australian Centre for Plant Functional Genomics, School of Land Crop and Food Sciences, University of Queensland, Brisbane, QLD, Australia*

Myriam Gaudeul • *Muséum national d'Histoire naturelle, UMR CNRS 7205 "Origine, Structure et Evolution de la Biodiversité", Paris, France*

Luc D.B. Gigord • *Conservatoire Botanique National de Mascarin, Saint Leu, La Réunion, France*

Luis F. Goulao • *Tropical Research Institute (IICT, IP), Lisboa, Portugal*

Berthold Heinze • *Department of Genetics, Federal Research Centre for Forests, Vienna, Austria*

Daniela Jahn • *Department of Genetics, Federal Research Centre for Forests, Vienna, Austria*

Ruslan Kalendar • *MTT/BI Plant Genomics, Institute of Biotechnology, University of Helsinki, Helsinki, Finland*

Agnieszka Koziel-Monte • *Department of Genetics, Federal Research Centre for Forests, Vienna, Austria*

Pierre Lefeuvre • *Pôle de Protection des Plantes, CIRAD, UMR PVBMT, Université de la Réunion, Saint-Pierre, France*

Ilia J. Leitch • *Jodrell Laboratory, Royal Botanic Gardens, Surrey, UK*

Darren P. Martin • *Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa*

Divakaran Minoo • *Providence Women's College, Kozhikode, Kerala, India*

Kantipudi Nirmal Babu • *Indian Institute of Spices Research, Kozhikode, Kerala, India*
Cristina M. Oliveira • *CEER, Centro de Engenharia dos Biossistemas, Instituto Superior de Agronomia Technical University of Lisbon, Lisboa, Portugal*
Timothée Le Péchon • *Chinese Academy of Sciences, Chengdu Institute of Biology, Chengdu, Sichuan, China*
Jaume Pellicer • *Jodrell Laboratory, Royal Botanic Gardens, Surrey, UK*
Nathalie Piperidis • *SRA Limited, Mackay, QLD, Australia*
Fatima Pustahija • *Laboratory Ecology, Systematic and Evolution, UMR 8079, CNRS-UPS-AgroParisTech, University Paris-Sud, Orsay, France; Faculty of Forestry, University of Sarajevo, Sarajevo, Bosnia and Herzegovina*
Muliyar Krishna Rajesh • *Central Plantation Crops Research Institute, Kasaragod, Kerala, India*
Paul Ritto • *Indian Institute of Spices Research, Kozhikode, Kerala, India*
Ronan Rivallan • *Cirad UMR AGAP, Montpellier, France*
Odile Robin • *Laboratory Ecology, Systematic and Evolution, UMR 8079, CNRS-UPS-AgroParisTech, University Paris-Sud, Orsay, France*
Germinal Rouhan • *Muséum national d'Histoire naturelle, UMR CNRS 7205 "Origine, Structure et Evolution de la Biodiversité", Paris, France*
Kukkumgai Samsudeen • *Central Plantation Crops Research Institute, Kasaragod, Kerala, India*
Alan H. Schulman • *MTT/BI Plant Genomics, Institute of Biotechnology, University of Helsinki, Helsinki, Finland; Biotechnology and Food Research, MTT Agrifood Research Finland, Jokioinen, Finland*
Kassa Semagn • *International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya*
Sonja Siljak-Yakovlev • *Laboratory Ecology, Systematic and Evolution, UMR 8079, CNRS-UPS-AgroParisTech, University Paris-Sud, Orsay, France*
Erinjery Jose Suraby • *Indian Institute of Spices Research, Kozhikode, Kerala, India*
Vedrana Vicic • *Department of Molecular Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia*
Hélène Vignes • *Cirad UMR AGAP, Montpellier, France*
Peter H. Weston • *The Royal Botanic Gardens and Domain Trust, Sydney, NSW, Australia*
Manuel Zander • *School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD, Australia; Australian Centre for Plant Functional Genomics, School of Land Crop and Food Sciences, University of Queensland, Brisbane, QLD, Australia; ARC Centre for Integrated Legume Research, University of Queensland, Brisbane, QLD, Australia*

Chapter 1

Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives

Germinal Rouhan and Myriam Gaudeul

Abstract

Taxonomy is the science that explores, describes, names, and classifies all organisms. In this introductory chapter, we highlight the major historical steps in the elaboration of this science that provides baseline data for all fields of biology and plays a vital role for society but is also an independent, complex, and sound hypothesis-driven scientific discipline.

In a first part, we underline that plant taxonomy is one of the earliest scientific disciplines that emerged thousands of years ago, even before the important contributions of Greeks and Romans (e.g., Theophrastus, Pliny the Elder, and Dioscorides). In the fifteenth to sixteenth centuries, plant taxonomy benefited from the Great Navigations, the invention of the printing press, the creation of botanic gardens, and the use of the drying technique to preserve plant specimens. In parallel with the growing body of morpho-anatomical data, subsequent major steps in the history of plant taxonomy include the emergence of the concept of natural classification, the adoption of the binomial naming system (with the major role of Linnaeus) and other universal rules for the naming of plants, the formulation of the principle of subordination of characters, and the advent of the evolutionary thought. More recently, the cladistic theory (initiated by Hennig) and the rapid advances in DNA technologies allowed to infer phylogenies and to propose true natural, genealogy-based classifications.

In a second part, we put the emphasis on the challenges that plant taxonomy faces nowadays. The still very incomplete taxonomic knowledge of the worldwide flora (the so-called taxonomic impediment) is seriously hampering conservation efforts that are especially crucial as biodiversity enters its sixth extinction crisis. It appears mainly due to insufficient funding, lack of taxonomic expertise, and lack of communication and coordination. We then review recent initiatives to overcome these limitations and to anticipate how taxonomy should and could evolve. In particular, the use of molecular data has been era-splitting for taxonomy and may allow an accelerated pace of species discovery. We examine both strengths and limitations of such techniques in comparison to morphology-based investigations, we give broad recommendations on the use of molecular tools for plant taxonomy, and we highlight the need for an integrative taxonomy based on evidence from multiple sources.

Key words Classification, Floras, DNA, History, Molecular taxonomy, Molecular techniques, Morpho-anatomical investigations, Plant taxonomy, Species, Taxonomic impediment

---

Taxonomy can justly be called the pioneering exploration of life on a little known planet Wilson (2004)
The goal of discovering, describing, and classifying the species of our planet assuredly qualifies as big science. Wheeler et al. (2004)

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_1, © Springer Science+Business Media New York 2014

1 Introduction

Adapting the famous aphorism of Theodosius Dobzhansky [1], could we dare to say that nothing in biology makes sense except in the light of taxonomy? Maybe yes, considering that most of biology relies on identified—and so described—species that are end products of taxonomy. Taxonomic information is obviously crucial for studies that analyze the distribution of organisms on Earth, since they need taxonomic names for inventories and surveys. But names are also needed to report empirical results from any other biological study dealing with, e.g., biochemistry, cytology, ecology, genetics, or physiology: even if working an entire life on a single species, e.g., *Arabidopsis thaliana* (L.) Heynh., a molecular biologist will focus all his/her research on numerous plants that all represent this species as delimited by taxonomy. Thus, taxonomy provides names, but it is not only a "biodiversity-naming" service: it is also a scientific discipline requiring theoretical, empirical, and epistemological rigor [2]. Names represent scientific hypotheses on species boundaries, and to put forward such hypotheses involves gathering information from characters of the organisms and adopting a species concept (*see* **Note 1** for an overview of the main species concepts). Morphology, anatomy, and genetics are the main sources of characters used in today's plant taxonomy. Not without noting that these types of characters all bring potentially valuable evidence, the focus of this book is on the use of nucleic acids—and chromosomes—for a reliable and efficient taxonomy.

Before discussing how to choose genomic regions to be studied in order to best deal with particular taxonomic issues (Chapter 2), this chapter aims to summarize the history of taxonomy and to highlight that plant molecular taxonomy emerged from an ancient discipline that has been, and is still, central for other scientific disciplines and plays a vital role for society. We will also give a brief overview of the general background into which plant taxonomy is performed today and propose some general considerations about molecular taxonomy.

2 Taxonomy and Taxon: Terminology and Fluctuating Meanings

It is not before 1813 that the Swiss botanist Augustin Pyramus De Candolle (1778–1841) invented the neologism "taxonomy" from the Greek ταξις (order) and νόμος (law, rule) and published it for the first time in his book "Théorie élémentaire de la Botanique" ("Elementary Theory of Botany", 3). He defined this scientific discipline as the "theory of the classifications applied to the vegetal kingdom," which he considered as one of the three components of

botany along with glossology—"the knowledge of the terms used to name plant organs"—and phytography, "the description of plants in the most useful way for the progress of science."

Much later, the Global Biodiversity Assessment of the United Nations Environment Programme (UNEP; 4) defined taxonomy as "the theory and practice of classifying organisms," including the classification itself but also the delimitation and description of taxa, their naming, and the rules that govern the scientific nomenclature. Today, depending on the authors, taxonomy is viewed either as a synonym for the "systematics" science—also called biosystematics [5, 6]—including the task of classifying species, or only as a component of systematics restricted to the delimitation, description, and identification of species. This latter meaning of taxonomy emerged lately, with the advent of phylogenetics as another component of systematics that allows classifications based on the evolutionary relationships among taxa [7].

Thus, it is ironical that taxonomy and systematics, which deal in particular with classifications and relationships between organisms, often themselves require clarifications on their relative circumscriptions and meanings before being used [8]. This book will consider plant taxonomy in the broadest sense, from, e.g., species delimitation based on different molecular techniques—Chapters 5–13—to focuses on phylogenetic reconstruction methods (Chapter 14).

Incidentally, it is interesting to note that the word "taxon"—plural, taxa—was invented much later (Lam, in ref. [9]) than "taxonomy": a taxon is a theoretical entity intended to replace terms such as "taxonomic group" or "biodiversity unit" [10], and "taxon" refers to a group of any rank in the hierarchical classification, e.g., species, genus, or family.

3 A Historical Perspective to Plant Taxonomy

3.1 One of the Earliest Scientific Disciplines

Delimiting, describing, naming, and classifying organisms are activities whose origins are obviously much older than the word "taxonomy"—which dates back to the nineteenth century; see above. The use of oral classification systems likely even predated the invention of the written language ca. 5,600 years ago. Then, as for all vernacular classifications, the precision of the words used to name plants was notably higher for plants that were used by humans. There was no try to link names and organisms in hierarchical classifications since the known plants were all named following their use: some were for food, others for medicines, poisons, or materials. As early as that time, several hundreds of plant organisms of various kinds were identified, while relatively few animals were known and named—basically those that were hunted or feared [11].

These early classifications that were exclusively utilitarian persisted until the fifteenth to sixteenth centuries although some

major advances were achieved, mainly by Ancient Greeks and Romans. It was perceptible that Greeks early considered plants not just as useful but also as beautiful, taking a look at paintings in Knossos (1900 BC) that indeed show useful plants like barley, fig, and olive but also narcissus, roses, and lilies. The Greek Theophrastus (372–287 BC), famous as the successor of Aristotle at the head of the Lyceum, is especially well known as the first botanist and the author of the first written works on plants. Interested in naming plants and finding an order in the diversity of plants, he could have been inspired by Aristotle who started his "Metaphysics" book by the sentence: "*All men by nature desire to know.*" Theophrastus is indeed the first one to provide us with a philosophical overview of plants, pointing out important fundamental questions for the development of what will be later called taxonomy, such as "what have we got?" or "how do we differentiate between these things?" He was moreover the first one to discuss relationships among plants and to suggest ways to group them not just based on their usefulness or uses. Thus, in his book "Enquiry into Plants," he described ca. 500 plants—probably representing all known plants at that time—that he classified as trees, shrubs, undershrubs, and herbs. He also established a distinction between flowering and nonflowering plants, between deciduous and evergreen trees, and between plants that grew in water and those that did not. Even if 80 % of the plants included in his works were cultivated, he had realized that "most of the wild kinds have no names, and few know about them," highlighting the need to recognize, describe, and name plants growing in the wild [12]. Observing and describing the known plants, he identified many characters that were valuable for later classifications. For instance, based on his observations of plants sharing similar inflorescences—later named "umbels"—he understood that, generally, floral morphology could help to cluster plants into natural groups and, several centuries later, most of these plants showing umbels were indeed grouped in the family Umbelliferae—nowadays Apiaceae.

Theophrastus was way ahead of his time, to such a point that his botanical ideas and concepts became lost during many centuries in Europe. But his works survived in Persia and Arabia, before being translated back into Greek and Latin and rediscovered in Europe in the fifteenth century. During this long Dark Age for botany—like for all other natural sciences—in Europe, the Roman Pliny the Elder (23–79 AD) and the Greek Dioscorides (~40–90 AD), in the first century AD, have however been two important figures. Although they did not improve the existing knowledge and methods about the description, naming, or classifications of plants, they compiled the available knowledge and their written works were renowned and widely used. The *Naturalis Historia* of Pliny (77 AD) was indeed a rich encyclopedia of the natural world, gathering 20,000 facts and observations reported by other authors,

mostly from Greeks like Theophrastus. At the same time in Greece, plants were almost only considered and classified in terms of their medical properties. The major work of Dioscorides *De Materia Medica* (ca. 77 AD) was long the sole source of botanical information (but at that time, botany was only considered in terms of pharmacology) and was repeatedly copied until the fifteenth century in Europe. The Juliana's book—*Juliana Anicia Codex*, sixth century (Figure 1)—is the most famous of these copies, well known because it innovated by adding beautiful and colorful plant illustrations to the written work of Dioscorides. If some paintings could be seen as good visual aids to identification—which should be considered as an advance for taxonomy—others were however fanciful [12]. All those plant books, called "herbals" and used by herbalists—who had some knowledge about remedies extracted from plants—throughout the Middle Ages, did not bring any other substantial progress.

3.2 Toward a Scientific Classification of Plants

With the Renaissance, the fifteenth and sixteenth centuries saw the beginning of the Great Navigations—e.g., C. Columbus discovered the New World from 1492; Vasco da Gama sailed all around Africa to India from 1497; and F. Magellan completed the first circumnavigation of the Earth in 1522—allowing to start intensive and large-scale naturalist explorations around the world: most of the major territories, except Australia and New Zealand, were discovered as soon as the middle of the sixteenth century, greatly increasing the number of plants that were brought back in Europe by either sailors themselves or naturalists on board. At that time, herbalists still played a major role in naming and describing plants, in association with illustrators who were producing realistic illustrations. But naming and classifying so numerous exotic and unknown plants from the entire world would not have been possible without three major inventions. Firstly, the invention of the Gutenberg's printing press with moveable type system (1450–1455) made written works on plants largely available in Europe—the first Latin translation of Theophrastus' books came out in 1483. Secondly, the first botanic gardens were created in Italy in the 1540s, showing the increasing interest of the population for plants and allowing teaching botany. Thirdly, in the botanic garden of Pisa, the Italian Luca Ghini (1490–1556) invented a revolutionary method for preserving—and so studying—plants, consisting in drying and pressing plants to permanently store them in books as "hortus siccus" (dried garden), today known as "herbaria," or "herbarium specimens." These perennial collections of dried plants were—and are still—a keystone element for plant taxonomy and its development: from that time, any observation and experimental result could be linked to specific plant specimens available for further identification, study of morphology, geographic distribution, ecology, or any other features. In short, Ghini provided with herbaria the basis of reproducibility that is an essential part of the scientific method [13].

Fig. 1 Painting of a *Cyclamen* plant, taken from the Juliana's book, showing the flowering stems rising from the upper surface of the rounded corm. According to Dioscorides, those plants were used as purgative, antitoxin, skin cleanser, labor inducer, and aphrodisiac

A student of Ghini, Andrea Cesalpino (1519–1603), was the first one since the Ancient Greeks to take over the work of Theophrastus and to discuss it. He highlighted that plants should be classified in a more natural and rational way than the solely utilitarian thinking. Convinced that all plants have to reproduce, he

provided a new classification system primarily based on seeds and fruits: in *De Plantis Libri XVI* (1583), he described 1,500 plants that he organized into 32 groups such as the Umbelliferae and Compositae—currently Apiaceae and Asteraceae, respectively. Cesalpino also made a contribution to the naming of plant, sometimes adding adjectives to nouns designing a plant, e.g., he distinguished *Edera spinosa*—spiny ivy—from *Edera terrestris*, creeping ivy. This could be seen as a prefiguration of the binomial naming system that was established in the eighteenth century and is still used in taxonomy. But the science of scientific naming was only starting and plants—like other living beings—were usually characterized by several words forming polynomial Latin names: for instance, tomato was designed as *Solanum caule inermi herbaceo, foliis pinnatis incisis*, which means "Solanum with a smooth herbaceous stem and incised pinnate leaves" [14] (Fig. 2).

Cesalpino contributed to the emergence of the concept of natural classification, i.e., a classification reflecting the "order of Nature." This latter expression involved different interpretations and classifications through the history of taxonomy, but a natural classification was always intended to reflect the relationships among plants. Because the evolutionary thought was not developed yet, it basically resulted in clustering plants with similar morphological features. So, it must be noted that the distinction between artificial and natural classifications—respectively named "systems" and "methods" at the end of the eighteenth century—is a modern interpretation of the past classifications. Taking advantage of both technical progresses like microscopy—in the seventeenth century—and scientific methods inspired by Descartes (1596–1650), several attempts were made to reach such a natural classification. For example, Bachmann—also known as Rivin or Rivinus (1652–1723)—based his classification on the corolla shape in *Introductio ad rem herbariam* in 1690. All together, the major interest of these classifications is that they triggered investigations on many morpho-anatomical characters that could be used by later taxonomists to describe and circumscribe plant species. The British John Ray (1627–1705) innovated by not relying anymore on a single characteristic to constitute groups of plants: he suggested natural groupings "from the likeliness and agreement of the principal parts" of the plants, based on many characters—mostly relative to leaves, flowers, and fruits. He documented more than 17,000 worldwide species in *Historia Plantarum* (1686–1704) and distinguished flowering vs. nonflowering plants and plants with one cotyledon—that he named "monocotyledons"—vs. plants with two cotyledons, "dicotyledons." Ray also played a major role in the development of plant taxonomy—and more generally of plant science—by creating the first text-based dichotomous keys that he used as a means to classify plants [15].

Fig. 2 Herbarium specimen from the Tournefort's Herbarium (housed at the Paris National Herbarium, Muséum national d'Histoire naturelle, MNHN) displaying a label with the hand-written polynomial name *Aconitum caeruleum, glabrum, floribus consolid(ae) regalis*

In contrast to Ray and his method intended to be natural, his French contemporary Joseph Pitton de Tournefort (1656–1708) explored, in his "Elements de Botanique" (1694), the possibility of classifying plants based on only few characters related to the corolla of flowers, creating an artificial system. The success of Tournefort's system resulted from the ease to identify the groups of plants based on the number and relative symmetry of the petals of a flower. Within his system, Tournefort precisely defined 698 entities—*Institutiones rei herbariae*, 1700—each being called a genus (plural: genera). The genus concept was new and contributed to a better structuration of the classification.

3.3 Naming Plant Names: Major Advances by Linnaeus

In spite of the numerous new ideas and systems produced from the sixteenth to the middle of the eighteenth century, names of plants still consisted in polynomial Latin names, i.e., a succession of descriptors following the generic name. This led to a rather long, complicated, and inoperative mean to designate plants and became problematic in the context of the Great Explorations, which allowed the discovery of more and more plants from all over the world (major explorations with naturalists on board included, e.g., the circumnavigation of La Boudeuse under Bougainville from 1766 to 1769 and the travels to the Pacific of J. Cook between 1768 and 1779). To overcome this impediment involving the naming of plants, the Swedish Carolus Linnaeus (1707–1778) took a critical step forward for the development of taxonomy.

He suggested dissociating the descriptors of the plant from the name itself, because according to him, the name should only serve to designate the plant. Therefore, he assigned a "trivial name" to each plant (more than 6,000 plants in *Species Plantarum*, 1753) [16], and this name was binomial, only consisting of two words: the "genus" followed by the "species," e.g., *Adiantum capillus-veneris* is a binomen created by Linnaeus that is still known and used as such to designate the Venus-hair fern. Although there had been some attempts of binomials as early as Theophrastus (followed by Cesalpino and a few others), Linnaeus succeeded in popularizing his system as new, universal—applied for all plants and, later on, even for animals in *Systema Naturae* [17]—and long lasting. Truly, the *Species Plantarum* [16] has been a starting point for setting rules in plant taxonomy. Used since Linnaeus until today, the binomial system along with other principles for the naming of plants was developed, standardized, synthesized, and formally accepted by taxonomists into a code of nomenclature—initially called "laws of botanical nomenclature" [18] and nowadays called the International Code of Botanical Nomenclature (ICBN). The current code is slightly evolving every 6 years, after revisions are adopted at an international botanical congress.

Linnaeus also proposed his own artificial classification. With the goal to describe and classify all plants—and other living beings—that

were "put on Earth by the Creator," he grouped them based on the number and arrangement of stamens and pistils within flowers—contrary to Tournefort, who only focused on petals. He called this classification a "sexual system," referring to the fundamental role of flowers in sexual reproduction (Fig. 3). This system included five hierarchical categories: varieties, species, genera, orders—equivalent to current families—and classes.

3.4 The Advent of the Theory of Evolution and Its Decisive Impact on Taxonomy

The end of the eighteenth century was conducive to revolutionary ideas in France, including new principles to reach the natural classification. Studying how to arrange plants in space for creating the new royal garden of the Trianon in the Palace of Versailles, Bernard de Jussieu (1699–1777) applied the key principle of subordination of characters, which will be published in 1789 by his nephew Antoine Laurent de Jussieu (1748–1836) in *Genera Plantarum* [19]. Bernard and A. L. de Jussieu stated that a species, genus, or any other taxon of the hierarchical classification should group plants showing character constancy within the given taxon, as opposed to the character variability observed among taxa. Since not all characters are useful at the same level of the classification, the principle of subordination led to a character hierarchy: characters displaying higher variability should be given less weight than more conserved ones in plant classifications. As a result, B. and A. L. de Jussieu subordinated the characters of flowers—judged more variable and therefore less suitable at higher levels—to the more conserved characters of seeds and embryos. It was the first application of this principle in taxonomy, and it could be interpreted today as a way to limit homoplasy, though the concept of homoplasy had not been elaborated yet [20].

Whereas botanical taxonomy had long been preponderant and faster in its development than its zoological counterpart, the trend was reversed at the beginning of the nineteenth century, especially with the application of the principle of subordination of characters to animals by the French biologists Jean-Baptiste Lamarck (1744–1829) and Georges Cuvier (1769–1832). New questions then arose in the mind of taxonomists, who were not only interested in naming, describing, and classifying organisms anymore, but also in elucidating how the observed diversity had been generated. Early explanatory theories included the theory of the transmutation of species, proposed by Jean-Baptiste de Lamarck in 1809 in his "Philosophie zoologique" [21]. This was the first theory to suggest the evolution of species, although it involved several misleading assumptions such as the notion of spontaneous generations. Charles Darwin (1809–1882) published his famous theory of evolution in "On the Origin of Species" (1859) [22] and introduced the central concept of descent with modification that later received extensive support and is still accepted today. This implied that useful characters in taxonomy, the so-called homologous characters,

Fig. 3 Linnaeus' sexual system as drawn by G. D. Ehret for the Hortus Cliffortianus (1735–1748); this illustration shows the 24 classes of plants that were defined by Linnaeus according to the number and arrangements of stamens

are those inherited from a common ancestor. Darwin indeed predicted that "our classifications will come to be, as far as they can be so made, genealogies" (Darwin 1859, p. 486) [22]. In other words, since the history of life is unique, only one natural classification is possible that is reflecting the phylogeny. This latter word was however not coined by Darwin himself, but by Ernst Haeckel (1834–1919) in 1866 in his "Generelle Morphologie der Organismen" [23], who is commonly known for the first illustration of a phylogeny, although Dayrat [24] evidenced that all Haeckel's illustrations should not be interpreted as real evolutionary-based phylogenetic trees [23] (Fig. 4). However, Darwin did not provide any new techniques or approaches to reconstruct the phylogeny or assist practicing taxonomists in their work [25], and, in spite of his major contributions, plant taxonomists therefore kept applying the method of classification described by B. and A. L. de Jussieu even after the onset of the evolutionary thought.

3.5 New Methods and New Sources of Characters for a Modern Taxonomy

In the 1960s, facing the subjectivity of the existing methods to reconstruct phylogenies, the new concept of numerical taxonomy proposed an entirely new way of examining relationships among taxa. Robert Sokal (1926–2012) and Peter Sneath (1923–2011) started developing this concept in 1963 [26] and elaborated it as an objective method of classification. The method consisted in a quantitative analysis of overall similarities between taxa, based on a characters-by-taxa data matrix—with characters divided into character states—and resulting in pairwise distances among taxa. But this method was not based on any evolutionary theory and the resulting diagrams could therefore not be reasonably interpreted in an evolutionary context, or as an evolutionary classification. Nevertheless, this theory flourished for a while, greatly benefiting from rapid advances in informatics.

A crucial change in the way botanists practice taxonomy occurred with the development of the cladistic theory and reconstruction of phylogenies—using diagrams called cladograms—to infer the evolutionary history of taxa. Willi Hennig (1913–1976) initiated this revolution with his book "Grundzüge einer Theorie der Phylogenetischen Systematik," published in 1950 [27], but his ideas were much more widely diffused in 1966 with the English translation entitled "Phylogenetic Systematics" [28]. The primary principle of cladism, or cladistics, is not to use the overall similarity among taxa to reconstruct the phylogeny, since similarity does not necessarily reflect an actual close evolutionary relationship. Instead, Hennig only based the phylogenetic classification on derived characters, i.e., the characters that are only inherited from the last common ancestor to two taxa—as opposed to the primitive characters. Every taxonomic decision, from a species definition to a system of higher classification, was to be treated as a provisional hypothesis,

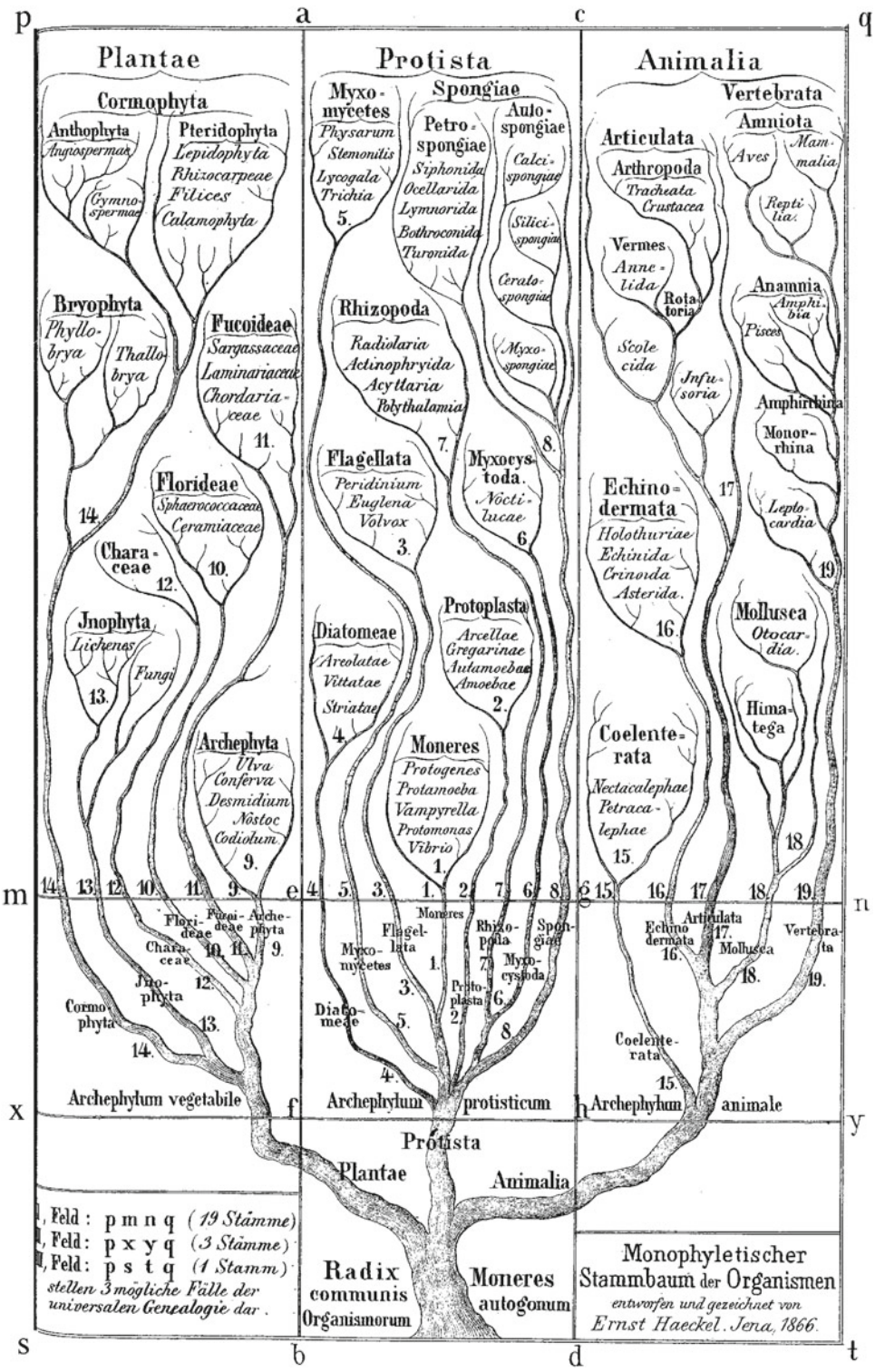


Fig. 4 Illustration from "Monophyletischer Stammbaum der Organismen" (Haeckel 1866): plants form one of the three main branches of the monophyletic genealogical tree of organisms

potentially falsifiable by new data [29]. This new method benefited from an increasing diversity of sources of characters to be considered, thanks to the important technological advances accomplished in the 1940s and 1950s in cytology, ecology, and especially in genetics.

The discovery of the double helical structure of the DNA molecule in 1953, by James Watson and Francis Crick, followed by the possibility to target specific fragments of the genome for selectively amplifying DNA—the polymerase chain reaction (PCR) was invented by Karry Mullis in 1986 [30]—has dramatically changed biology. In particular, the introduction of DNA sequence data has been era-splitting for plant taxonomy, offering access to numerous characters and statistical approaches. Thus, at the turn of the twenty-first century, the use of molecular data and new tree-building algorithms—with probabilistic approaches—led the Angiosperm Phylogeny Group (APG) to better circumscribe all orders and families of flowering plants [31–33] and to improve to a great extent our understanding of the classification based on evolutionary relationships. Many long-standing views of deep-level relationships were drastically modified at the ordinal level and to a lesser extent at the familial level. One of the most striking changes is the abandonment of the long-recognized monocot-dicot split, since Monocots—class Liliopsida—were found to be derived from within a basal grade of families that were traditionally considered as dicots—class Magnoliopsida. Thus, the more or less intuitive classifications proposed since the beginning of the twentieth century [34–38] have progressively been less used, as a consequence of the modifications brought to the classification by molecular results [39].

Taxonomy took advantage of molecular data not only for improving plant classification or species delineation, but also for species-level identification with the development of the DNA barcoding initiative since the early 2000s. DNA barcoding is based on the premise that a short standardized DNA sequence can allow distinguishing individuals of different species because genetic variation between species is expected to exceed that within species. It was first promoted by Paul Hebert for animals [40] and later supported by international alliances of research organizations like the Consortium for the Barcode of Life (CBOL; http://barcoding.si.edu) that includes a plant working group, or the China Plant Barcode of Life Group.

The long history of plant taxonomic research and its numerous contributors, for both theoretical concepts and the practical accumulation of knowledge, allowed the development of an independent, complex, and sound hypothesis-driven scientific discipline that explores, describes, documents the distribution, and classifies taxa. It is clearly not restricted to, e.g., identifying specimens and establishing species lists, but it nevertheless also provides basic knowledge that is required to address a wide range of research questions and serve stakeholders in government agencies and

international biodiversity organizations (for management of agriculture pests, development of new pharmaceutical compounds, control of trade in endangered species, management of natural resources, etc. [29, 41–45]). However, taxonomy is faced with the enormous existing plant diversity, and one still unanswered question resides in the extent of plant diversity: how many species are there on Earth?

4 Plant Taxonomy Today: Current Challenges, Methods, and Perspectives

4.1 How Many Plant Species Are There?

Linnaeus' *Species Plantarum*, published in 1753, was one of the first key attempts to document the diversity of plants on a global scale [16]. In this work, Linnaeus recognized more than 6,000 species but erroneously concluded that "the number of plants in the whole world is much less than commonly believed, I ascertained by fairly safe calculation […] it hardly reaches 10,000" [16]. Later on, in 1824, the Swiss A. P. de Candolle, in his *Prodromus Systematis Naturalis Regni Vegetabilis* [46], aimed to produce a flora of the world: he included 58,000 species in seven volumes. Today, we know that the magnitude of plant diversity is much larger, although we are uncertain of the exact number of plant species.

There are two questions in estimating the total number of plant species: the first one is how many species have already been described; the second one is how many more species are presently unknown to science.

Our uncertainty about the number of described species is mostly due to the fact that taxonomists sometimes gave different names to the same species inadvertently, especially in the past due to poor communication means between distant scientists. This led to the existence of multiple names for a single biological entity, a phenomenon called synonymy. As a consequence, we know that more than 1,600,000 vascular plant names were published, as evidenced by the International Plant Names Index (IPNI) [47], but they would actually represent only 223,000–422,000 accepted species—depending on the method of calculation [43, 48 and references therein 49, 50]. In addition, the disagreement on a single species concept (*see* **Note 1**) among plant taxonomists means that species counts can easily differ by an order of magnitude or more when the same data are examined by different botanists [51]. This leads to a taxonomic inflation, i.e., an increased number of species in a given group that is not due to an actual discovery of new species [52–54]. In practice, this can occur when, e.g., different botanists do not recognize the same number of species in a given taxonomic group—the "splitters" vs. the "lumpers"—or when one botanist describes subspecies while another one elevates them to the rank of species.

The estimation of the total number of plant species on Earth is also obviously hampered by our uncertainty about the extent of the unknown plant diversity: how many more species are there to discover? The exploration of plant diversity allows the discovery of ca. 2000 new plant species every year [43, 44, 55] although part of which may turn out to be synonyms based on future thorough monographic revisions. Based on a model of the rates of plant species description, Joppa et al. [56] estimated that there should be an increase of 10–20 % in the current number of flowering plant species. This means that, based on the estimation of 352,000 currently known flowering plant species [43], they predicted the actual diversity between 390,000 and 420,000 species for this group. Meanwhile, Mora et al. [57] used higher taxonomy data, i.e., they extrapolated the global number of plant species based on the strong negative correlation between the taxonomic rank and the number of higher taxa—which is better known than the total number of species. As a result, focusing on land plants, they suggested an expected increase of 38 % in the number of species, from 215,000 in Catalogue of Life [58] to 298,000 predicted species.

These numbers make clear that our knowledge of plant diversity is still very incomplete and that even estimates of its magnitude remain highly controversial and speculative, highlighting the need for more taxonomic studies.

4.2 Current Threats on Plant Diversity, the Taxonomic Impediment, and Some Initiatives to Overcome It

At the Sixth Conference of the Parties to the Convention on Biological Diversity (CBD) held in 2002, more than 180 countries adopted the Global Strategy for Plant Conservation (GSPC). It included 16 specific targets that were to be achieved by 2010, with the goal to halt the loss of plant diversity [43]. The Strategy was updated in 2010 (at the tenth meeting of the Conference of the Parties), and it is now implemented within the broader framework of the Strategic Plan for Biodiversity 2011–2020. The first and most fundamental target of the Strategy was initially to complete a "widely accessible working list of known plant species, as a step towards a complete world flora" [43, 55]. After the completion of this list in 2010 [49], Target 1 was slightly modified to developing "an online flora of all know plants" (http://www.cbd.int/gspc/targets.shtml). This target aims to provide baseline taxonomic information, i.e., a list of the accepted names for all known plant species, linked to their synonyms but also to biological information such as geographic distribution and basic identification tools. Since species are basic units of analysis in several areas of biogeography, ecology, and macroevolution and also the currency for global biodiversity assessments [59], the lack of such taxonomic information is a critical bottleneck for research, conservation, and sustainable use of plant diversity [43] and was called the "taxonomic impediment" at the Second Conference of the Parties to the CBD (decision II/8). This is especially critical at a time when biodiversity faces its sixth

extinction crisis: most newly described species occur in hotspots of diversity, often in tropical dense forests, where protected areas are scarce, the level of habitat destruction (due to anthropic activities) is high, and the impact of climate change is strong. Newly described species are also likely to be characterized by locally low abundance and small geographic ranges, enhancing their risk of extinction [60]. Therefore, botanists must engage into a race to describe and name species before they go extinct. This is especially true since plants still lag far behind many animal groups in contributing to global conservation planning, despite their essential role in structuring most ecosystems [61]. In addition to the major conservation concern, there are a multitude of possible concrete examples of beneficial application of taxonomic discovery such as the identification of new wild species adaptable for agriculture, timber or fibers, new genes for enhancement of crop productivity, and new classes of pharmaceuticals. Also, basic taxonomic knowledge is a prerequisite to monitor and anticipate the spread of invasive plants and to better understand ecosystem services [42, 44].

Several factors limit the efficiency of botanists in documenting plant diversity. However, recent improvements and future optimistic perspectives must also be underlined, and numerous contributions have been made to imagine and propose what the "taxonomy for the twenty-first century" should and could be (*see* refs. [29, 62] and the whole theme issue of the *Philosophical Transactions of the Royal Society of London, Series B*, that they coordinated; *see* also refs. [63–65]).

First, one limiting factor is the general lack of funding and, in particular, the lack of resources devoted to the basic field activity of collecting new material [51, 62, 63, 66]. Field explorations are also made difficult by practical limitations such as ease of access to remote areas or safety concerns in some parts of the world that may be politically unstable [41, 42, 55]. However, we currently know a renewed age of exploration and discovery, supported by several national or international initiatives. This is particularly true in the United States, where the "Planetary Biodiversity Inventories: Mission to an (Almost) Unknown Planet" program, launched in 2003, aims to complete the world species inventory for some selected taxa, with individual project awards of ca. US$3 million over a 5-year duration (http://nsf.gov/pubs/2006/nsf06500/nsf06500.htm) [51, 67]. In Europe, several major museums and botanical gardens established the Consortium of European Taxonomic Facilities (CETAF) in 1996, which in turn created the European Distributed Institute of Taxonomy (EDIT) in 2006, for a 5-year period, under the European Union Sixth Framework Programme. This worldwide network of excellence brought together 29 leading European, but also North American and Russian institutions with the goal to increase both the scientific basis and capacity for biodiversity conservation. Developing countries

also participated to this international effort by developing either national or multinational similar programs, e.g., in Brazil, Mexico, and Africa [68]. On a global scale, the Global Taxonomic Initiative (GTI) was launched in 1998 by the Conference of the Parties to the CBD and was later related to the GSPC, in order to remove or reduce the "taxonomic impediment." In addition to institutional breakthroughs, modern means of travel have facilitated access to remote places where many species occur. As a result, although today botanical expeditions could probably not be as prolific as those reported during the great naturalists explorations of the eighteenth and nineteenth centuries in terms of new species descriptions (e.g., in 1770, Sir Joseph Banks collected specimens representing as many as 110 new genera and 1,300 new species in Australia; White 1772 in 55), important discoveries occurred in the recent past and provide evidence for the vitality of contemporary botanists: for instance, the Malagasy endemic *Takhtajania perrieri* (Capuron) Baranova and J.-F. Leroy (Winteraceae) was first collected in 1909 and thought to have gone extinct but was rediscovered in 1994 [69], i.e., almost 90 years after its first collection. Other examples suggest that some showy, sometimes abundant, plants still remain to be described, even in geographical areas that are supposed to be well prospected: a new genus and species of conifer, *Wollemia nobilis* W. G. Jones, K. D. Hill, and J. M. Allen, was observed in the 1990s only ca. 150 km from Sydney (Australia) and was shown to belong to a well-known family of charismatic trees (Araucariaceae), including only two other genera [70]. In 2007, Thulin and collaborators reported the discovery of a conspicuous and dominating tree in the Somali National Regional State (Ogaden) in eastern Ethiopia [71, 72]. This tree, *Acacia fumosa* Thulin (Fabaceae), covers an area as large as Crete but was hitherto unknown to science. The location of this species in an African war zone and the inaccessibility of the area probably explain that it had never been collected and remained undescribed so far. To cite a last example of recent striking botanical discovery, we can mention the description of a new palm genus and species from Madagascar, *Tahina spectabilis* J. Dransf. and Rakotoarin [73]. The trees grow to 18 m high and leaves reach 5 m in diameter, making them the most massive palms ever found in Madagascar. However, the small census size (less than a hundred individuals), limited habitat, and rare reproduction events lead to serious conservation concerns for the species.

A second crucial issue for enhancing our knowledge of plant diversity is the lack of taxonomic expertise. This is at least partly due to the lack of credit given to works of descriptive taxonomy (e.g., species lists, floras, or monographs) compared to peer-reviewed publications in high-impact journals [43, 66, 74, 75]. The global number of species described over time has increased over the past 250 years [56, 76], but this remains clearly not

sufficient to counteract the increasing rate of species extinctions, and many species are at risk of disappearing before being described. Although taxonomists have most likely increased the efficiency of their efforts since the mid-1700s, the involvement of more numerous people into the tasks of exploring and describing the biodiversity is needed. In the Unites States, the NSF's Partnerships for Enhancing Expertise in Taxonomy (PEET) program allowed the training of new generations of taxonomists since 1995 [74, 77] and enjoyed much success. In addition, in some regions (e.g., in Costa Rica or Papua New Guinea), local people called "parataxonomists" contribute to specimens collection and species recognition based on rough morphological criteria, in collaboration with taxonomic experts [42]. This is also in line with the growing body of "citizen scientists," who are often amateurs and offer their help to accumulate data, e.g., on the presence/absence of a given species in a given region or the distribution of a morphological character across space. Because they are usually organized as large networks, they represent an immense and increasingly important workforce and make possible some tasks that would otherwise not have been possible because of, e.g., limited time and funding [76]. However, sound knowledge and experience of professional taxonomists remain critical [43, 62, 63, 66, 78] and capacity building in tropical countries—where the greatest diversity of life is concentrated—should therefore be a priority [55].

A third identified impediment to our taxonomic knowledge was—and is still, to a certain extent—the problem of communication and coordination, of tracing the accumulated publication records, of deciphering the complex synonymy, and of chasing the scattered (and sometimes in poor condition) material, especially type specimens that are housed in herbaria around the world [55, 62, 63]. Worldwide natural history collections contain over 350 million plant specimens [79], and their importance has recently been made even more prominent by the finding that they house many new species that remain to be described [80]: researchers analyzing the time lapse between flowering plant sample collection and new species recognition estimated that only 16 % were described within 5 years of being collected for the first time and that nearly 25 % of new species descriptions involved specimens of more than 50 years old. The median time lag between the earliest specimen collection and the publication of the new species description in a monograph was ca. 30 years. Therefore, although one limiting step of species discovery may be the capacity to undertake field work (as suggested above), the examination of existing herbarium specimens by experts is another bottleneck. This is however now partly overcome by increased international collaborations and a better access to information and specimens, thanks to modern data-sharing technologies [43–45, 57, 65]. As an example, a major step is about to be accomplished, thanks to funding from the

Andrew W. Mellon Foundation and subsequent institutional commitments to database and image name-bearing type specimens—on which the species original descriptions are based—and deposit these data in the central repository JSTOR Plant Science [78]. At an even larger scale, several major herbaria—including the Paris Herbarium, which is one of the biggest/richest in the world with *ca*. eight million specimens of vascular plants, pers. obs.—are currently carrying out large-scale digitization of all their specimens, in order to make them freely available as high-quality photographs on the Web. In the United States, the National Science Foundation (NSF), through its Advancing Digitization of Biological Collections (ADBC) program, developed a strategic plan for a 10-year coordinated effort to digitize and mobilize images and data associated with all biological research collections of the country in a freely available online platform. This will ensure increased accessibility of all valuable information and will be made possible by the establishment of a central National Resource for Digitization of Biological Collections (called iDigBio for "Integrated Digitized Biocollections").

For a better diffusion of taxonomic revisions, Godfray [62] claimed the need of a "unitary web-based and modernized taxonomy" (*see* also ref. [81]). Without opting for such a drastic evolution, a recent revision of the International Code of Botanical Nomenclature (ICBN) nevertheless tends to encourage a change dynamics toward electronic publications: at the International Botanical Congress held in Melbourne in July 2011, purely electronic descriptions were judged valid for the publication of new species (Art. 29), as opposed to the previous requirement to publish in traditional, printed publication [82]. Also, whereas the current taxonomic knowledge is mostly made available in paper format as monographs, floras, and field guides, many internet taxonomy initiatives exist and catalogue species names, lists of museums specimens, identification keys, and/or other biological information. These websites include IPNI (www.ipni.org), The Plant List (www.theplantlist.org), GBIF (www.gbif.org), Species 2000/ITIS Catalogue of Life (www.catalogueoflife.org), Tree of Life (www.tolweb.org), and Encyclopedia of Life (www.eol.org) to cite only a few of them (*see* refs. [51, 62])

4.3 Molecular Taxonomy and the Need for an Accelerated Pace of Species Discovery

In addition to increased efforts toward exploration in the field, various initiatives to promote and develop taxonomic expertise, generalization of collaborative work, and improved access to natural history collections and literature, major advances in technology also provide new opportunities to facilitate and accelerate the rate of species discovery at a time of increasing need to monitor and manage biodiversity. The goal of accelerating the pace of species discovery was made especially clear by the promoters of the DNA barcode initiative [83, 84], but more generally, the use of

molecular tools for taxonomic purpose emerged in the 1990s—or even in the 1970s if considering allozyme markers—and has quickly become an area of intense activity.

Today, most recognized species have been delineated and described based on morphological evidence: in general, they have been delimited based on one or more qualitative or quantitative morphological characters that show no—or very little—overlap with other species [85]. The initial enthusiasm for molecular taxonomy most probably came from the additional and complementary information that it provided. Also, molecular taxonomy requires an expertise that is nowadays more broadly distributed than that for thorough morphological investigations, it makes use of tools that are not specific to a particular group of plants, and it may appear more prone to scientific publications in peer-reviewed journals than more traditional, taxonomic studies. We synthesize, here, several other characteristics—both strengths and limitations—of molecular taxonomy that one should keep in mind when initiating taxonomic studies using molecular tools.

4.3.1 Strengths and Limitations of Molecular Taxonomy

First, it must be noted that the resemblance criterion within a species, on which is based the morphological approach to delimit species, suffers exceptions and can lead to erroneous conclusions. Before the various reproductive systems of plants were well understood, male and female individuals from a single—e.g., dioecious—species were sometimes described as two distinct species based on morphological investigations. For example, in the orchid genus *Catasetum* Rich. ex Kunth, plants are functionally dioecious (i.e., with female and male flowers situated on distinct individuals) and can morphologically differ so much from each other that taxonomists of the nineteenth century assigned individuals of the same species to different genera (*Monachanthus* Lindl. and *Myanthus* Lindl.) [86]. Other species descriptions incorporated characters that were in fact due to anther-smut disease caused by the fungus *Microbotryum violaceum* (Pers.) G. Deml and Oberw.: anthers of infected plants are filled with dark-violet fungal spores instead of yellow pollen [87]. As a result, *Silene cardiopetala* Franch., for example, was distinguished from *Silene tatarinowii* Regel by its dark anthers but should likely be treated as the same species. More generally, because the phenotype of a plant is influenced both by its genotype and by its environment—and the interaction between the genotype and the environment, called phenotypic plasticity—the observations of herbarium specimens collected in the field may be somewhat misleading. Molecular taxonomy should avoid this possible bias since it is based on neutral markers that are in principle independent of environmental conditions. However, the influence of the environment is mostly true for vegetative characters and usually less problematic for reproductive characters. In addition, the use of several morphological characters should limit the problem since all traits are unlikely to be affected in the same way [88].

Second, several studies showed that, in comparison to the traditional morphological criterion for delimiting species, molecular tools sometimes allow the detection of additional, so-called cryptic, species that could not be distinguished on morphological grounds only. This may happen when species emerged in the recent past, due to morphological stasis or to morphological convergences [89]. The existence of such cryptic species was reported, e.g., on temperate or tropical plants [90, 91] and references therein; for an animal example, *see* ref. [92].

Third, in addition to the primary goal of species delimitation, the use of genetic tools may allow to better understand the evolutionary process at work within taxonomically complex groups, where taxa are sometimes difficult—or even impossible—to delineate. These groups are often characterized by uniparental reproduction—e.g., self-fertilization or apomixis—and reticulate evolution, due to, e.g., hybridization and introgression, which preclude the delineation of discrete and unambiguous taxonomic entities. In such cases—e.g., in the genera *Sorbus*, *Epipactis*, and *Taraxacum* [93–95], respectively, cited in ref. 96]—principles of conservation biology suggest that the evolutionary processes that generate and maintain diversity should themselves be preserved because they are even more important than the presently observed taxa [96, 97]. In this perspective, molecular tools can yield very useful information, usually based on a population sampling.

Fourth, from a practical point of view, a key strength of molecular taxonomy is that it can be performed on any life stage—even some that bear no or only few morphological characters such as seeds, seedlings, or fern gametophytes [98, 99]—and almost any type of material, e.g., leaves, cambium [100–103], bark [104], dry wood [105], and roots [106, 107]. Therefore, the use of molecular characters for taxonomic purposes appears especially suitable for organisms that require years before flowering and/or fully developing or when access to some other key—e.g., reproductive—characters is difficult.

The ubiquitous character of the DNA molecule in living beings can also become a problem, and care should be taken to only isolate DNA from the target material and exclude DNAs of any other animal, vegetal, or fungal organisms living around or in the plant under study—e.g., parasitic insects, epiphyllous mosses, and endophytic fungi.

Another practical limitation of molecular taxonomy is the cost, as molecular lab facilities and often rather expensive consumables are needed. This cost may be especially limiting in developing countries [108, 109], although it is ever decreasing, thanks to the spread of molecular analyses that are more and more commonly employed and to technological advances that allow cheaper and less time-consuming analyses—see below.

An important parameter that is shared by "traditional" and molecular taxonomy studies is sampling strategy and sampling effort. Taxonomy is based on a comparative approach that requires the investigation of as many specimens/samples as possible in order to catch all the extent of natural variation. Therefore, the quality of taxonomic studies partly relies on a thorough sampling of specimens/samples to be surveyed, and a biased sampling may cause erroneous conclusions. As an example, *Marsilea azorica* Launert and Paiva, which was thought to be a local endemic and critically endangered species of the Azores archipelago, was recently shown to be conspecific to an Australian native species that is widely cultivated and invasive in Florida, *Marsilea hirsuta* R. Br. [110]: because the spread of *M. hirsuta* out of Australia was not documented when the *Marsilea* specimens from the Azores were examined by Launert and Paiva in 1983 [111], the botanists did not include the Australian taxa into their survey and erroneously described the species as new to science.

On more theoretical and conceptual grounds, some claim that, in comparison with "traditional"—typically morphology-based—taxonomy, the use of molecular tools may avoid bias due to the subjectivity of a given taxonomist, who could have a priori ideas on species delimitation. However, the acquisition of a molecular dataset also implies some more or less subjective choices, e.g., on the distinction of orthologs vs. paralogs, on defining character homology when sequences of different lengths must be aligned to form a square matrix, or on the statistical analysis to carry out after the data are produced (*see* refs. [88, 112–114]); the latter choice, on data analysis, is closely related to the adoption of a given species concept (*see* **Note 1**). Also, because our current technological capacities do not allow the routine inclusion of the whole genome in taxonomic analyses, choices must be made on the genomic compartment(s) to survey—nuclear, mitochondrial, or chloroplastic—the molecular technique(s) to use and the precise, individual marker(s) to consider (Chapter 2). The choice of a limited number of markers is required, in practice, although multiple independent loci might often be necessary to solve the possible disagreement between gene trees and species trees and to uncover the common reticulate evolution—due to horizontal DNA transfer, hybridization, and polyploidization events—and incomplete lineage sorting in plants [51, 88, 114–116]. The extent of genome coverage by molecular markers is partly dependent on the molecular technique that is used, and there is often a trade-off between the possibility—due to time and cost limitations—of surveying numerous markers and the information content provided by each marker. For example, it is usually achievable to include a large number of anonymous markers based on length polymorphism—such as RAPD or AFLP markers—but the number of DNA regions that could be sequenced,

representing highly informative data, is much more limited with the traditional Sanger method. However, rapid advances in next-generation sequencing (NGS) technologies have resulted in a huge cost reduction and offer incredible new opportunities for producing billions of base pairs of accurate DNA sequence data in a few hours [117–119]. Studying whole chloroplast genomes or multiple nuclear loci may therefore become routine even in non-model species and would obviously revolutionize plant molecular taxonomy. Then, the main bottleneck would probably be cleaning up and assembling the sequence reads to generate useable data, and major improvements in bioinformatics would be needed to deal with such huge amounts of data [117, 118].

Another limitation of molecular taxonomy is the possible lack of genetic divergence when sister species have very recent origins because they will share alleles due to recent ancestry and, if reproductive isolation is not complete, to ongoing gene flow—i.e., hybridization. This lack of genetic variation can nevertheless be accompanied by some level of morphological differentiation, leading to the exact symmetrical situation to cryptic taxa—where one could observe genetic but no morphological distinction; see above. The absence or extremely weak genetic divergence was observed, e.g., in the young and species-rich neotropical genus *Inga* Mill. (Fabaceae) [120], and striking examples of such morphological diversification but weak genetic variation are also provided by cases of adaptive radiations, where species rapidly adapt to different environments—e.g., in the Hawaiian silverswords (Asteraceae) [121], the Asian genus *Rheum* L. [122], or the widespread columbine genus *Aquilegia* L. [123]; for more examples, *see* ref. [124]. In such cases of recent diversification, the delimitation of species will usually be based on allele frequency changes rather than diagnostic changes [125] and can benefit from recently developed coalescent-based methods (e.g., refs. [126, 127]). The time required for genetic divergence to build up after speciation will depend on the mutation and fixation rates—and the fixation rate depends on the number of reproductively effective individuals. Because of different fixation rates between (diploid) nuclear and (haploid) organelle genomes, studies based on nuclear vs. organelle DNA markers may yield contrasted results on species limits. Such contrasted results are also made likely by the horizontal organelle DNA transfers that occasionally occur, especially among closely related species.

Molecular markers can also suffer from homoplasy, i.e., markers can show similar character states that, however, do not derive from a common ancestor. In this case, they do not inform on the genealogy of taxa and, because they do not reflect a shared evolutionary history, they may be misleading on evolutionary and, as a consequence, on taxonomic relationships. This is especially problematic for highly variable markers—such as microsatellites—and for DNA sequences that are only composed of four types of

monomer (A, C, G, and T): as a result, a substitution at any one position has a high probability of being a reversal or a convergence, i.e., of being homoplasic [88, 114]. It is therefore critical to take this caveat into account when analyzing and interpreting molecular data.

Another drawback of molecular taxonomy is that name-bearing type specimens often do not permit DNA analyses because of non-optimal drying and storage conditions, resulting in DNA deterioration [128]; the same limit also obviously applies to most plant fossils, which do not contain DNA. Consequently, a comparison of the supposedly new species with known species may not be possible on a molecular basis and prevent a rigorous taxonomic, comparative approach. As part of their "plea for DNA taxonomy," Tautz et al. [125] proposed to identify neotypes for all known species in cases of unavailable genetic information from the original types so that these neotypes could constitute new reference records for further studies. However, this proposal received very limited support (*see* refs. [112, 129]). Besides, recent progresses have been made in the extraction of DNA from herbarium specimens [130, 131] and the genetic analysis of such material will very likely benefit from NGS technologies (*see* Chapter 4) [119]. But so far, given the usually low-quantity and often degraded DNA that is extracted from herbarium specimens, the most commonly employed molecular techniques are microsatellite markers—because their short length makes amplification more likely than that of longer DNA stretches (*see* ref. [132])—and organelle DNA sequencing, because their multiple copies per cell represent more abundant template DNA for PCR than nuclear loci. Most of the published studies report the successful exploitation of specimens up to ca. 100–150 years old, with DNA sequences produced usually ca. 500 pb long [133–137]. But the kind of material—e.g., presence of PCR-inhibiting substances [138]—and the speed and method of drying appear more important than the actual age of the sample [133–135, 139], and some botanists managed to obtain DNA sequences from even older specimens and/or longer DNA regions, e.g., Ames and Spooner sequenced ca. 440-bp DNA fragments from potato material from the early eighteenth century [140], and Andreasen et al. [139] sequenced 800-bp DNA fragments from a specimen collected in the late eighteenth century. The successful use of aged seeds has also been reported [130, 141].

4.3.2 The Definitive Need for an Integrative Taxonomy

The use of molecular data in plant taxonomy has been era-splitting and highly successful in many instances, but we also highlighted some limits and cautions to consider when adopting this approach. Most importantly, a species description solely based on molecular evidence would obviously seem critically disconnected from the natural history of the species, e.g. its life-history traits, ecological requirements, co-occurring species, and biotic interactions.

As such, molecular tools may indeed accelerate the rate of species discovery but would actually be a poor contribution to our knowledge and understanding of plant diversity and evolution. Such a use of molecular taxonomy could even end up with the exact opposite of the expected outcome if funders only aim to basically delineate and count species with no other ambition; indeed, gathering further biological information is an essential prerequisite to make a general use of the taxonomic knowledge, efficiently preserve the existing diversity, and allow its continued evolution. Botanists have long realized this and promoted the use of multiple independent sources of data and/or the use of several analytical methods on the same dataset to corroborate the delimitation and provide a thorough and detailed description of species. As early as 1961, Simpson (p. 71) wrote "It is an axiom of modem taxonomy that the variety of data should be pushed as far as possible to the limits of practicability" [6]. In agreement, Alves and Machado [142] wrote that "Taxonomy should be based on all available evidence." This awareness gave rise to the advent of what is now called "integrative taxonomy," where taxonomic hypotheses are cross-validated by several lines of evidence [29, 114, 142–147, and many others]. As sources of relevant characters, many fields of biology might contribute to taxonomic studies: they include morpho-anatomy which takes advantage of new techniques such as scanning electron microscopy (SEM), remotely operable digital microscopy, computer-assisted tomography, confocal laser microscopy, and automatic image processing for morphometry [147, 148] and cytogenetics—*see* Chapters 14–16—but also palynology, physiology, chemistry (production of secondary compounds), breeding relationships, and ecological niche modelling. We are not aware of currently available examples in plant taxonomy, but for animals, *see* refs. [149–151]. Other sources of information will also most probably be more widely used in the future, such as transcriptomics [152, 153], metabolomics [154], proteomics [88], and even phenomics. Munck et al. [155] showed, in barley, that the fingerprint of a near-infrared spectrum from an individual represents a coarse-grained overview of the whole physiochemical composition of its phenome, with the phenomic profile resulting from the combined effects of the entire genome, proteome, and metabolome [156]. The diversity of approaches involved in modern plant taxonomy is consistent with the observations by Joppa et al. [76] that (1) today's biologists who describe species are not only contributing to the field of taxonomy but also active in other fields/disciplines and (2) that most new species are nowadays described by several authors, whereas descriptions by a single author were common around 1,900.

It is also clear that end users of taxonomy such as conservation planners need an operational, character-based, and cheap way to

discriminate species [84, 108, 142]. This could tend to diminish the perceived potential of molecular taxonomy, but in this perspective and in spite of the shortcomings that we have just underlined, molecular taxonomy obviously has a great role to play. DNA can aid to delimit taxa and to group specimens among which to find morphological—or other types of—affinities in further investigations (*see* refs. [157, 158]). Such clusters of individuals, characterized by close genetic relationships, are sometimes referred to as "molecular operational taxonomic units" (MOTU) [159], before their genuine taxonomic status is evaluated by gathering additional data. Markmann and Tautz [160] called this approach, based on an initial molecular assessment, the "reverse taxonomy" (*see* also refs. [143, 144]). The fruitful link between "traditional" and molecular taxonomy should be accompanied by an analogous link between herbarium vouchers, plant samples for DNA extraction, and DNA extracts [161, 162]. The curation of such collections and the maintenance of a dynamic link between them will provide a long-lasting and reliable framework for taxonomic investigations and will permit the critical reevaluation of taxa delimitations at any time, based on both herbarium and DNA material.

Duminil and Di Michele [88] reviewed studies comparing species delimitations based on morphological traits and molecular markers. They found both cases of congruence and incongruence between the two types of data. As suggested above, cases of incongruence were due either to stronger molecular discrimination between species—suggesting the existence of cryptic species—or, on the contrary, to stronger morphological differentiation, due to processes like local adaptation, phenotypic plasticity, or neutral morphological polymorphism (e.g., ref. [163]). Conflicting results often trigger more in-depth studies using as many loci as possible and, if possible, loci that originate from different genomes, with the goal to better understand the patterns and processes of plant evolution and diversification (*see* refs. [164–170]; *see* also Chapters 17–18 for practical case studies).

Taxonomic circumscriptions are scientific hypotheses, which are ideally validated by evidence from multiple sources, and molecular methods offer the opportunity to yield high-potential information. However, there is not a single, best method to be used in all plant groups and the molecular taxonomist will have to face multiple questions: before anything, it is necessary to identify the optimal sampling strategy, the most suited genomic compartment(s) to examine, the right technique(s) to use, and the adequate method(s) of statistical analysis to extract the relevant information about species limits and relationships [113, 114]. In addition to the complementarity of "traditional" and genetic approaches, molecular taxonomy itself will often require to gather and compare patterns based on several types of

data—e.g., nuclear vs. cytoplasmic markers or markers with different rates of evolution. The goal of this book is to present the possible alternatives of molecular taxonomy, their practical implications in the lab, current analytical tools that are available, and theoretical consequences for data interpretation. The empirical and analytical approaches used for a molecular taxonomic study, together with the conclusions drawn from the data, will also obviously depend on the species concept that is adopted and on the choice of operational criteria to delimit species (*see* **Note 1**). After taxa delimitation and description, a further goal of taxonomy is to propose a natural classification using phylogenetic reconstruction methods—*see* Chapter 13.

5 Notes

Species Concepts and Contemporary Criteria for Species Delimitation

1. Species delimitation obviously depends on what a species is, and, although the species is often seen as the fundamental unit of evolution, its definition has long remained highly debated.

   The existence of species itself is somewhat controversial, especially in plants where asexuality, hybridization, and polyploidy may render the definition and delimitation of species complex and fuzzy. Some argue that species are "arbitrary constructs of the human mind," while others claim that they are objective, discrete entities. Reviewing the available data (both in plants and animals), Rieseberg et al. [171] showed that discrete phenotypic clusters exist in most genera (>80 %), although the correspondence of taxonomic species to these clusters is poor (<60 % and not different between plants and animals). In addition, crossability experiments indicate that as much as 70 % of plant taxonomic species and 75 % of plant phenotypic clusters correspond to reproductively independent lineages.

   The proliferation of alternative species concepts really started in the 1970s. It gave rise to several decades of debate and taxonomic instability because many concepts were incompatible in that they lead to the recognition of different species boundaries and different number of species. This was called the "species problem."

   Morphological approaches have dominated species delimitation for centuries, starting with the purely typological (i.e., essentialist) pre-Darwinian view. But most contemporary biologists are familiar with the idea that species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such

groups (the "Biological Species Concept"), whether or not they differ in phenotypic characters that are readily apparent.

However, another unified species concept has now emerged. It originated as early as the beginning of the twentieth century (with, e.g., E. B. Poulton), became well established during the period of the modern evolutionary synthesis (with the great leaders T. Dobzhansky, E. Mayr, G. G. Simpson, and S. Wright), and was recently largely promoted by de Queiroz [172, 173]. This unified concept reconciles previous, at least partially, incompatible species concepts. It considers species as separately evolving metapopulation lineages and is called the "General (metapopulation) Lineage Concept." Other properties of species, which used to be treated as necessary (and sufficient) properties to recognize a species as such (e.g., reproductive isolation, monophyly; *see* Table 1), are now only seen as different lines of evidence or "operational criteria" relevant to assessing lineage separation. The unified species concept is actually not a new concept, but simply the clear separation of the theoretical concept from the operational criteria that are used for the empirical application of this concept.

Operational criteria can be either tree based or non-tree based (e.g., direct tests of crossability, indirect estimates of gene flow, statistical clustering algorithms) [174], and new methods are still being developed (e.g., analyzing multilocus genetic data in a coalescent framework). Criteria differ in their suitability to some particular species (e.g., sexual vs. asexual), their requirements in terms of type of data and sampling, and their strengths and limitations. It must also be noted that most of them will require researchers to make some qualitative judgements at some point.

The commonly observed incompatibility between various criteria stems from the fact that various properties actually arise at different stages in the process of speciation: as lineages diverge, they become distinguishable in terms of quantitative traits, diagnosable in terms of fixed character states, reproductively incompatible, they evolve distinct ecologies, and they pass through polyphyletic, paraphyletic, and monophyletic stages. These changes commonly do not occur at the same time, and they are not even necessarily expected to occur in a specific order. De Queiroz [172] qualifies this transition period, from one ancestral species to two divergent species, a "grey zone," where alternative species definitions can come into conflict. But as lineages diverge, the number of species criteria satisfied will increase and allow a highly corroborated hypothesis of lineage separation and species delimitation.

Table 1
Some alternative contemporary species concepts/criteria

Name of the species concept/criterion	Definition of the species	Major contributor(s)	Refs.
Interbreeding species concept [forms the basis for the General (metapopulation) Lineage Concept]	A group of potentially interbreeding populations	Wright (1940); Mayr (1942); Dobzhansky 1950	175–177
Isolation species concept[a] [often called the biological species concept]	A group of potentially interbreeding populations that is reproductively isolated from other such groups	Poulton (1904); Mayr (1942); Dobzhansky (1970)	177–179
Phenetic species concept	A group that forms a phenetic cluster (quantitative difference)	Sokal and Crovello (1970)	180
Ecological species concept	A group that shares the same niche or adaptive zone	Van Vaalen (1976)	181
Evolutionary species concept[a] [corresponds closely to the General (metapopulation) Lineage Concept]	A lineage (i.e., an ancestral-descendant sequence of populations) evolving separately from others and with its own evolutionary role and tendencies	Simpson (1951); Wiley (1978)	182, 183
Phylogenetic species concept—character diagnosability version	An irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent (fixed qualitative character) The diagnostic character can be from any trait (e.g., morphological or molecular) and of any significance (e.g., a single base pair)	Cracraft (1989)	184
Phylogenetic species concept—reciprocal monophyly version	A group that shows monophyly (consisting of an ancestor and all of its descendants and commonly inferred from the possession of shared derived character states)	Rosen (1979); Donoghue (1985); Mishler (1985)	185–187
Genealogical species concept	A group that shows monophyly for all (or at a consensus of) gene genealogies in the genome	Baum and Shaw (1995)	188
Genotypic species concept	A group recognizable on the basis of multiple, unlinked, inherited genetic markers A pair of such genotypic clusters is recognizable if the frequency distribution of genotypes is bimodal or multimodal, and strong heterozygote deficits and linkage disequilibria are evident between the clusters	Mallet (1995)	189
Cohesion species concept[a]	A group that is characterized by cohesion mechanisms, including reproductive isolation, recognition mechanisms, ecological selection, as well as genealogical distinctiveness	Templeton (1998)	190

[a]Combined species concepts, i.e., concepts using a combination of morphological, ecological, phylogenetic, and reproductive criteria

References

1. Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teach 35:125–129
2. Carvalho d M R, Bockmann FA, Amorim DS, de Vivo M, de Toledo-Piza M, Menezes NA, de Figueiredo JL, Castro RMC, Gill AC, McEachran JD, Compagno LJV, Schelly RC, Britz R, Lundberg JG, Vari RP, Nelson G (2005) Revisiting the taxonomic impediment. Science 307:353
3. Candolle AP (1813) Théorie Élémentaire de la botanique, ou Exposition des principes de la classification naturelle et de l'art de décrire et d'étudier les végétaux.
4. Heywood VH, Watson RT (1995) Global biodiversity assessment. Cambridge University Press, Cambridge
5. Mayr E (1969) Principles of systematic zoology. McGraw-Hill, New York
6. Simpson GG (1961) Principles of animal taxonomy. New York, Columbia Univ. Press
7. Tillier S (2000) Systématique - Ordonner la diversité du Vivant. Rapport sur la science et la technologie de l'Académie des sciences n°11. Éditions Tec & Doc.
8. Small E (1989) Of biological systematics (or taxonomy of taxonomy). Taxon 38:335–356
9. Sprague JL, Lanjouw J, Andreas CH (1948) Chron Bot 12(1/2):12
10. Morton CV (1957) The misuse of the term taxon. Taxon 6(5):155
11. Raven PH (2004) Taxonomy: where are we now? Philos Trans R Soc Lond B Biol Sci 359:729–730
12. Pavord A (2005) The naming of names: the search for order in the world of plants. Bloomsbury, New-York
13. Funk VA, Hoch PC, Prather LA, Wagner WL (2005) The importance of vouchers. Taxon 54:127–129
14. Knapp S (2012) What's in a name? A history of taxonomy. http://www.nhm.ac.uk/nature-online/science-of-natural-history/taxonomy-systematics/history-taxonomy. Accessed Jan 2012
15. Griffing LR (2011) Who invented the dichotomous Key? Richard Waller's watercolors of the herbs of Britain. Am J Bot 98:1911–1923
16. Linnaeus C (1753) Species Plantarum. Holmiae: Impensis Laurentii Salvii.
17. Linnaeus C (1758) Systema naturae, 10th edn. Holmiae: Impensis Laurentii Salvii.
18. Candolle AP (1867) Lois de la nomenclature botanique adoptées par le Congrès international de botanique: tenu à Paris en août 1867. H. Georg, Geneva
19. Jussieu AL (1789) Genera plantarum. Herissant, Paris
20. Philippe H, Lecointre G, VanLe HL, LeGuyader H (1996) A critical study of homoplasy in molecular data with the use of a morphologically based cladogram, and its consequences for character weighting. Mol Biol Evol 13:1174–1186
21. Lamarck, JBPAM (1809) Philosophie zoologique. Paris
22. Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London
23. Haeckel E (1866) Generelle Morphologie der Organismen. Reimer, Berlin
24. Dayrat B (2003) The roots of phylogeny: how did Haeckel build his trees? Syst Biol 52:515–527
25. Davis PH, Heywood PH (1963) Principles of angiosperm taxonomy. Oliver and Boyd, Edinburgh and London
26. Sneath PHA, Sokal RR (1963) Principles of numerical taxonomy, 7th edn. W. H. Freeman, San Francisco
27. Hennig W (1950) Grundzüge einer Theorie der phylogenetischen Systematik
28. Hennig W (1966) Phylogenetic Systematics (tr. D. Davis and R. Zangerl). Univ. of Illinois Press, Urbana.
29. Godfray HCJ, Knapp S (2004) Taxonomy for the twenty-first century – Introduction. Philos Trans R Soc Lond B Biol Sci 359:559–569
30. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA invitro – the polymerase Chain-reaction. Cold Spring Harb Symp Quant Biol 51:263–273
31. Bremer K, Chase MW, Stevens PF, Anderberg AA, Backlund A, Bremer B, Briggs BG, Endress PK, Fay MF, Goldblatt P, Gustafsson MHG, Hoot SB, Judd WS, Kallersjo M, Kellogg EA, Kron KA, Les DH, Morton CM, Nickrent DL, Olmstead RG, Price RA, Quinn CJ, Rodman JE, Rudall PJ, Savolainen V, Soltis DE, Soltis PS, Sytsma KJ, Thulin M, Grp AP (1998) An ordinal classification for the families of flowering plants. Ann Mo Bot Gard 85:531–553
32. Bremer B, Bremer K, Chase MW, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Fay MF, Goldblatt P, Judd WS, Kallersjo M, Karehed J, Kron KA, Lundberg J, Nickrent DL, Olmstead RG, Oxelman B, Pires JC, Rodman JE, Rudall PJ, Savolainen V, Sytsma KJ, van der Bank M, Wurdack K, Xiang JQY, Zmarzty S, Grp AP (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot J Linn Soc 141:399–436

33. Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Moore MJ, Olmstead RG, Rudall PJ, Sytsma KJ, Tank DC, Wurdack K, Xiang JQY, Zmarzty S, Grp AP (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121
34. Bessey CE (1915) The phylogenetic taxonomy of flowering plants. Ann Mo Bot Gard 2:109–164
35. Cronquist A (1981) An integrated system of classification of flowering plants. Columbia University Press, New-York
36. Stebbins GL (1974) Flowering plants: evolution above the species level. Belknap, Cambridge
37. Takhtajan A (1997) Diversity and classification of flowering plants. Columbia University Press, New-York
38. Thorne RF (1976) A phylogenetic classification of the Angiospermae. Evol Biol 9:35–106
39. Soltis DE, Soltis PS, Endress PK, Chase MW (2005) Phylogeny and evolution of angiosperms. Sinauer associates, Sunderland
40. Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond Ser B Biol Sci 270:313–321
41. May RM (2004) Tomorrow's taxonomy: collecting new species in the field will remain the rate-limiting step. Philos Trans R Soc Lond B Biol Sci 359:733–734
42. May RM (2011) Why worry about how many species and their loss? Plos Biol 9
43. Paton AJ, Brummitt N, Govaerts R, Harman K, Hinchcliffe S, Allkin B, Lughadha EN (2008) Towards target 1 of the global strategy for plant conservation: a working List of all known plant species – progress and prospects. Taxon 57:602–611
44. Wilson EO (2003) The encyclopedia of life. Trends Ecol Evol 18:77–80
45. Wilson EO (2004) Taxonomy as a fundamental discipline. Philos Trans R Soc Lond B Biol Sci 359:739
46. Candolle AP (1824–1873) Prodromus systematis naturalis regni vegetabilis. Parisii: Sumptibus Sociorum Treuttel et Würtz.
47. International Plant Names Index (2008) Published on the Internet. http://www.ipni.org. Accessed Apr 2012
48. Scotland RW, Wortley AH (2003) How many species of seed plants are there? Taxon 52:101–104
49. The Plant List (2010) Version 1. Published on the Internet. http://www.theplantlist.org/. Accessed 1 Jan
50. Wortley AH, Scotland RW (2004) Synonymy, sampling and seed plant numbers. Taxon 53: 478–480
51. Mallet J, Willmott K (2003) Taxonomy: renaissance or Tower of Babel? Trends Ecol Evol 18:57–59
52. Isaac NJB, Mallet J, Mace GM (2004) Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol Evol 19:464–469
53. Meiri S, Mace GM (2007) New taxonomy and the origin of species. Plos Biol 5:1385–1386
54. Pillon Y, Chase MW (2007) Taxonomic exaggeration and its effects on orchid conservation. Conserv Biol 21:263–265
55. Crane PR (2004) Documenting plant diversity: unfinished business. Philos Trans R Soc Lond B Biol Sci 359:735–737
56. Joppa LN, Roberts DL, Pimm SL (2011) How many species of flowering plants are there? Proc R Soc B Biol Sci 278:554–559
57. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? Plos Biol 9
58. Bisby FA, Roskov YR., Orrell TM, Nicolson D, Paglinawan LE et al (2010) Species 2000 & ITIS Catalogue of Life: 2010 Annual Checklist. Digital resource at http://www.catalogueoflife.org/annual-checklist/2010. Species 2000: Reading, UK.
59. Caldecott JO, Jenkins MD, Johnson TH, Groombridge B (1996) Priorities for conserving global species richness and endemism. Biodivers Conserv 5:699–727
60. Joppa LN, Roberts DL, Myers N, Pimm SL (2011) Biodiversity hotspots house most undiscovered plant species. Proc Natl Acad Sci USA 108:13171–13176
61. Callmander MW, Schatz GE, Lowry PP (2005) IUCN Red List assessment and the global strategy for plant conservation: taxonomists must act now. Taxon 54:1047–1050
62. Godfray HCJ (2002) Challenges for taxonomy – the discipline will have to reinvent itself if it is to survive and flourish. Nature 417:17–19
63. Funk VA (2006) Floras: a model for biodiversity studies or a thing of the past? Taxon 55:581–588
64. Wheeler QD, Raven PH, Wilson EO (2004) Taxonomy: impediment or expedient? Science 303:285
65. Wheeler QD, Knapp S et al (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. SystBiodivers 10(1):1–20
66. Ebach MC, Valdecasas AG, Wheeler QD (2011) Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. Cladistics 27:550–557

67. Ronquist F, Gardenfors U (2003) Taxonomy and biodiversity inventories: time to deliver. Trends Ecol Evol 18:269–270
68. Joly CA (2006) Taxonomy: programmes developing in the south too. Nature 440:24
69. Schatz GE, Lowry PP, Ramisamihantanirina A (1998) Takhtajania perrieri rediscovered. Nature 391:133–134
70. Jones WG, Hill KD, Allen JM (1995) Wollemia nobilis, a new living Australian genus and species in the Araucariaceae. Telopea 6:173–176
71. Mabberley DJ (2009) Exploring Terra Incognita. Science 324:472
72. Thulin M (2007) Acacia fumosa sp nov (Fabaceae) from eastern Ethiopia. Nord J Bot 25:272–274
73. Dransfield J, Rakotoarinivo M, Baker WJ, Bayton RP, Fisher JB, Horn JW, Leroy B, Metz X (2008) A new coryphoid palm genus from madagascar. Bot J Linn Soc 156:79–91
74. Agnarsson I, Kuntner M (2007) Taxonomy in a changing world: seeking solutions for a science in crisis. Syst Biol 56:531–539
75. Crisci JV (2006) One-dimensional systematist: perils in a time of steady progress. Syst Bot 31:217–221
76. Joppa LN, Roberts DL, Pimm SL (2011) The population ecology and social behaviour of taxonomists. Trends Ecol Evol 26:551–553
77. Rodman JE, Cody JH (2003) The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. Syst Biol 52:428–435
78. Bebber DP et al (2012) Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. Proc R Soc B 279:2269–2274
79. Thiers B (2011) Index herbariorum: a global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/ih/
80. Bebber DP, Carine MA, Wood JRI, Wortley AH, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, Robson NKB, Scotland RW (2010) Herbaria are a major frontier for species discovery. Proc Natl Acad Sci USA 107:22169–22171
81. Godfray HCJ, Clark BR, Kitching IJ, Mayo SJ, Scoble MJ (2007) The Web and the structure of taxonomy. Syst Biol 56: 943–955
82. Knapp S, McNeill J, Turland NJ (2011) Changes to publication requirements made at the XVIII International Botanical Congress in Melbourne – what does e-publication mean for you? BMC Evol Biol 11:250
83. Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Syst Biol 54:852–859
84. Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philos Trans R Soc B Biol Sci 360:1805–1811
85. Wiens JJ (2007) Species delimitation: new approaches for discovering diversity. Syst Biol 56:875–878
86. Pannell JR (2009) Mating-system evolution: succeeding by celibacy. Curr Biol 19: R983–R985
87. Hood ME, Antonovics J (2003) Plant species descriptions show signs of disease. Proc R Soc Lond Ser B Biol Sci 270:S156–S158
88. Duminil J, Di Michele M (2009) Plant species delimitation: a comparison of morphological and molecular markers. Plant Biosyst 143:528–542
89. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I (2007) Cryptic species as a window on diversity and conservation. Trends Ecol Evol 22:148–155
90. Grundt HH, Kjolner S, Borgen L, Rieseberg LH, Brochmann C (2006) High biological species diversity in the arctic flora. Proc Natl Acad Sci USA 103:972–975
91. Pillon Y, Hopkins HCF, Munzinger J, Amir H, Chase MW (2009) Cryptic species, gene recombination and hybridization in the genus Spiraeanthemum (Cunoniaceae) from New Caledonia. Bot J Linn Soc 161:137–152
92. Dulvy NK, Reynolds JD (2009) Biodiversity skates on thin ice. Nature 462:417
93. Robertson A, Newton AC, Ennos RA (2004) Multiple hybrid origins, genetic diversity and population genetic structure of two endemic Sorbus taxa on the Isle of Arran, Scotland. Mol Ecol 13:123–134
94. Squirrell J, Hollingsworth PM, Bateman RM, Tebbitt MC, Hollingsworth ML (2002) Taxonomic complexity and breeding system transitions: conservation genetics of the Epipactis leptochila complex (Orchidaceae). Mol Ecol 11:1957–1964
95. van Dijk PJ (2003) Ecological and evolutionary opportunities of apomixis: insights from Taraxacum and Chondrilla. Philos Trans R Soc Lond B Biol Sci 358:1113–1121
96. Ennos RA, French GC, Hollingsworth PM (2005) Conserving taxonomic complexity. Trends Ecol Evol 20:164–168
97. Ennos RA, Whitlock R, Fay MF, Jones B, Neaves LE, Payne R, Taylor I, De Vere N, Hollingsworth PM (2012) Process-based species action plans: an approach to conserve contemporary evolutionary processes that sustain diversity in taxonomically complex groups. Bot J Linn Soc 168:194–203

98. Li FW, Tan BC, Buchbender V, Moran RC, Rouhan G, Wang CN, Quandt D (2009) Identifying a mysterious aquatic fern gametophyte. Plant Syst Evol 281:77–86
99. Van Deynze A, Stoffel K (2006) High-throughput DNA extraction from seeds. Seed Sci Technol 34:741–745
100. Asif MJ, Cannon CH (2005) DNA extraction from processed wood: a case study for the identification of an endangered timber species (Gonystylus bancanus). Plant Mol Biol Rep 23:185–192
101. Colpaert N, Cavers S, Bandou E, Caron H, Gheysen G, Lowe AJ (2005) Sampling tissue for DNA analysis of trees: trunk cambium as an alternative to canopy leaves. Silvae Genet 54:265–269
102. Rachmayanti Y, Leinemann L, Gailing O, Finkeldey R (2006) Extraction, amplification and characterization of wood DNA from Dipterocarpaceae. Plant Mol Biol Rep 24:45–55
103. Tibbits JFG, McManus LJ, Spokevicius AV, Bossinger G (2006) A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees. Plant Mol Biol Rep 24:81–91
104. Novaes RML, Rodrigues JG, Lovato MB (2009) An efficient protocol for tissue sampling and DNA isolation from the stem bark of Leguminosae trees. Genet Mol Res 8:86–96
105. Deguilloux MF, Pemonge MH, Petit RJ (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. Proc R Soc B Biol Sci 269: 1039–1046
106. Hiiesalu I, Opik M, Metsis M, Lilje L, Davison J, Vasar M, Moora M, Zobel M, Wilson SD, Partel M (2012) Plant species richness belowground: higher richness and new patterns revealed by next-generation sequencing. Mol Ecol 21:2004–2016
107. Kesanakurti PR, Fazekas AJ, Burgess KS, Percy DM, Newmaster SG, Graham SW, Barrett SCH, Hajibabaei M, Husband BC (2011) Spatial patterns of plant diversity below-ground as revealed by DNA barcoding. Mol Ecol 20:1289–1302
108. Dunn CP (2003) Keeping taxonomy based in morphology. Trends Ecol Evol 18:270–271
109. Santos LM, Faria LRR (2011) The taxonomy's new clothes: a little more about the DNA-based taxonomy. Zootaxa 3025:66–68
110. Schaefer H, Carine MA, Rumsey FJ (2011) From European priority species to invasive weed: Marsilea azorica (Marsileaceae) is a misidentified alien. Syst Bot 36:845–853
111. Launert GOE, Paiva JAR (1983) Iconographia selecta florae Azoricae. Coimbra 2:159
112. Lipscomb D, Platnick N, Wheeler Q (2003) The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol Evol 18:65–66
113. Sites JW, Marshall JC (2004) Operational criteria for delimiting species. Annu Rev Ecol Evol Syst 35:199–227
114. Stace CA (2005) Plant taxonomy and biosystematics – does DNA provide all the answers? Taxon 54:999–1007
115. Linder CR, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution UN plants. Am J Bot 91:1700–1708
116. Vriesendorp B, Bakker FT (2005) Reconstructing patterns of reticulate evolution in angiosperms: what can we do? Taxon 54:593–604
117. Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. Am J Bot 99:175–185
118. Harrison N, Kidner CA (2011) Next-generation sequencing and systematics: what can a billion base pairs of DNA sequence data do for you? Taxon 60:1552–1566
119. Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot 99:349–364
120. Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM (2001) Rapid diversification of a species-rich genus of neotropical rain forest trees. Science 293:2242–2245
121. Baldwin BG, Sanderson MJ (1998) Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc Natl Acad Sci USA 95:9402–9406
122. Wang AL, Yang MH, Liu JQ (2005) Molecular phylogeny, recent radiation and evolution of gross morphology of the rhubarb genus Rheum (Polygonaceae) inferred from chloroplast DNA trnL-F sequences. Ann Bot 96:489–498
123. Hodges SA, Arnold ML (1994) Columbines – a geographically widespread species flock. Proc Natl Acad Sci USA 91:5129–5132
124. Linder HP (2008) Plant species radiations: where, when, why? Philos Trans R Soc B Biol Sci 363:3097–3105
125. Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. Trends Ecol Evol 18:70–74
126. Hey J, Pinho C (2012) Population genetics and objectivity in species diagnosis. Evolution 66:1413–1429
127. Knowles L, Carstens B (2007) Delimiting species without monophyletic gene trees. Syst Biol 56:887–895
128. Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, Bakker

FT (2011) DNA damage in plant herbarium tissue. Plos One 6
129. Seberg O, Humphries CJ, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends Ecol Evol 18:63–65
130. Lister DL, Bower MA, Howe CJ, Jones MK (2008) Extraction and amplification of nuclear DNA from herbarium specimens of emmer wheat: a method for assessing DNA preservation by maximum amplicon length recovery. Taxon 57:254–258
131. Wandeler P, Hoeck PEA, Keller LF (2007) Back to the future: museum specimens in population genetics. Trends Ecol Evol 22: 634–642
132. Cozzolino S, Cafasso D, Pellegrino G, Musacchio A, Widmer A (2007) Genetic variation in time and space: the use of herbarium specimens to reconstruct patterns of genetic variation in the endangered orchid Anacamptis palustris. Conserv Genet 8:629–639
133. Erkens RHJ, Cross H, Maas JW, Hoenselaar K, Chatrou LW (2008) Assessment of age and greenness of herbarium specimens as predictors for successful extraction and amplification of DNA. Blumea 53:407–428
134. Drabkova L, Kirschner J, Vlcek C (2002) Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of Juncaceae. Plant Mol Biol Rep 20:161–175
135. Jankowiak K, Buczkowska K, Szweykowska-Kulinska Z (2005) Successful extraction of DNA from 100-year-old herbarium specimens of the liverwort Bazzania trilobata. Taxon 54:335–336
136. Korpelainen H, Pietilainen M (2008) Effort to reconstruct past population history in the fern Blechnum spicant. J Plant Res 121:293–298
137. Savolainen V, Cuenoud P, Spichiger R, Martinez MDP, Crevecoeur M, Manen JF (1995) The use of herbarium specimens in DNA phylogenetics – evaluation and improvement. Plant Syst Evol 197:87–98
138. Ribeiro RA, Lovato MB (2007) Comparative analysis of different DNA extraction protocols in fresh and herbarium specimens of the genus Dalbergia. Genet Mol Res 6:173–187
139. Andreasen K, Manktelow M, Razafimandimbison SG (2009) Successful DNA amplification of a more than 200-year-old herbarium specimen: recovering genetic material from the Linnaean era. Taxon 58:959–962
140. Ames M, Spooner DM (2008) DNA from herbarium specimens settles a controversy about origins of the European potato. Am J Bot 95:252–257
141. Walters C, Reilley AA, Reeves PA, Baszczak J, Richards CM (2006) The utility of aged seeds in DNA banks. Seed Sci Res 16:169–178
142. Alves RJV, Machado MD (2007) Is classical taxonomy obsolete? Taxon 56:287–288
143. DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc B Biol Sci 360:1905–1916
144. DeSalle R (2006) Species discovery versus species identification in DNA barcoding efforts: response to rubinoff. Conserv Biol 20:1545–1547
145. Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH (2010) Integrative taxonomy: a multisource approach to exploring biodiversity. Annu Rev Entomol 55:421–438
146. Dayrat B (2005) Towards integrative taxonomy. Biol J Linn Soc 85:407–415
147. Wheeler QD (2005) Losing the plot: DNA "barcodes" and taxonomy. Cladistics 21: 405–407
148. Corney DPA, Clark JY, Tang HT, Wilkin P (2012) Automatic extraction of leaf characters from herbarium specimens. Taxon 61(1):231–244
149. Raxworthy CJ, Ingram CM, Rabibisoa N, Pearson RG (2007) Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (Phelsuma) from Madagascar. Syst Biol 56:907–923
150. Rissler LJ, Apodaca JJ (2007) Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (Aneides flavipunctatus). Syst Biol 56:924–942
151. Wiens JJ, Graham CH (2005) Niche conservatism: integrating evolution, ecology, and conservation biology. Annu Rev Ecol Evol Syst 36:519–539
152. Brautigam A, Gowik U (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol 12:831–841
153. Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J (2012) Targeted enrichment strategies for next-generation plant biology. Am J Bot 99:291–311
154. Rodriguez-Fernandez JI, De Carvalho CJB, Pasquini C, De Lima KMG, Moura MO, Arizaga GGC (2011) Barcoding without DNA? Species identification using near infrared spectroscopy, Zootaxa, pp 46–54
155. Munck L, Jespersen BM, Rinnan A, Seefeldt HF, Engelsen MM, Norgaard L, Engelsen SB (2010) A physiochemical theory on the applicability of soft mathematical models-experimentally interpreted. J Chemometr 24:481–495

156. Cruickshank RH, Munck L (2011) It's barcoding Jim, but not as we know it. Zootaxa 2933:55–56
157. Andres-Sanchez S, Rico E, Herrero A, Santos-Vicente M, Martinez-Ortega MM (2009) Combining traditional morphometrics and molecular markers in cryptic taxa: towards an updated integrative taxonomic treatment for Veronica subgenus Pentasepalae (Plantaginaceae sensu APG II) in the western Mediterranean. Bot J Linn Soc 159:68–87
158. Schlick-Steiner BC, Seifert B, Stauffer C, Christian E, Crozier RH, Steiner FM (2007) Without morphology, cryptic species stay in taxonomic crypsis following discovery. Trends Ecol Evol 22:391–392
159. Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philos Trans R Soc B Biol Sci 360:1935–1943
160. Markmann M, Tautz D (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Philos Trans R Soc B Biol Sci 360:1917–1924
161. Pleijel F, Jondelius U, Norlinder E, Nygren A, Oxelman B, Schander C, Sundberg P, Thollesson M (2008) Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. Mol Phylogenet Evol 48:369–371
162. Puillandre N, Bouchet P, Boisselier-Dubayle MC, Brisset J, Buge B, Castelin M, Chagnoux S, Christophe T, Corbari L, Lambourdiere J, Lozouet P, Marani G, Rivasseau A, Silva N, Terryn Y, Tillier S, Utge J, Samadi S (2012) New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Mol Ecol Resour 12:396–402
163. Gemeinholzer B, Bachmann K (2005) Examining morphological and molecular diagnostic character states of Cichorium intybus L. (Asteraceae) and C-spinosum L. Plant Syst Evol 253:105–123
164. Bacon CD, McKenna MJ, Simmons MP, Wagner WL (2012) Evaluating multiple criteria for species delimitation: an empirical example using Hawaiian palms (Arecaceae: Pritchardia). BMC Evol Biol 12:23
165. Barrett CF, Freudenstein JV (2011) An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. Mol Ecol 20:2771–2786
166. Koffi KG, Heuertz M, Doumenge C, Onana JM, Gavory F, Hardy OJ (2010) A combined analysis of morphological traits, chloroplast and nuclear DNA sequences within santiria trimera (Burseraceae) suggests several species following the biological species concept. Plant Ecol Evol 143:160–169
167. Ley AC, Hardy OJ (2010) Species delimitation in the Central African herbs Haumania (Marantaceae) using georeferenced nuclear and chloroplastic DNA sequences. Mol Phylogenet Evol 57:859–867
168. Meudt HM, Lockhart PJ, Bryant D (2009) Species delimitation and phylogeny of a New Zealand plant species radiation. Bmc Evol Biol 9
169. Schmidt-Lebuhn AN (2007) Using amplified fragment length polymorphism (AFLP) to unravel species relationships and delimitations in Minthostachys (Labiatae). Bot J Linn Soc 153:9–19
170. Zeng YF, Liao WJ, Petit RJ, Zhang DY (2010) Exploring species limits in two closely related Chinese oaks. Plos One 5
171. Rieseberg LH, Troy TE, Baack EJ (2006) The nature of plant species. Nature 440:524–527
172. de Queiroz K (2005) Ernst Mayr and the modern concept of species. Proc Natl Acad Sci USA 102:6600–6607
173. de Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56(6):879–886
174. Sites JW, Marshall JC (2003) Delimiting species: a renaissance issue in systematic biology. Trends Ecol Evol 18(9):462–470
175. Wright S (1940) The statistical consequences of Mendelian heredity in relation to speciation. In: Huxley J (ed) The new systematics. Oxford University Press, London, pp 161–183
176. Mayr E (1942) Systematics and the origin of species. Columbia University Press, New York
177. Dobzhansky T (1950) Mendelian populations and their evolution. Am Nat 84:401–418
178. Poulton EB (1904) What is a species? Proceedings of the Entomological Society of London 1903: lxxvii–cxvi.
179. Dobzhansky T (1970) Genetics of the evolutionary process. Columbia University Press, New York
180. Sokal RR, Crovello TJ (1970) The biological species concept: a critical evaluation. Am Nat 104:107–123
181. Van Valen L (1976) Ecological species, multispecies, and oaks. Taxon 25:233–239
182. Simpson GG (1951) The species concept. Evolution 5:285–298
183. Wiley EO (1978) The Evolutionary species concept reconsidered. Syst Zool 21:17–26
184. Cracraft J (1989) Speciation and its ontology: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In: Otte D, Endler JA (eds) Speciation and its consequences. Sinauer Associates, Sunderland, pp 28–59
185. Rosen DE (1979) Fishes from the uplands and intermontane basins of Guatemala:

revisionary studies and comparative geography. Bull Am Mus Nat His 162:267–376
186. Donoghue MJ (1985) A critique of the biological species concept and recommendations for a phylogenetic alternative. Bryologist 88: 172–181
187. Mishler BD (1985) The morphological, developmental, and phylogenetic basis of species concepts in bryophytes. Bryologist 88: 207–214
188. Baum DA, Shaw KL (1995) Genealogical perspectives on the species problem. In: Hoch PC, Stephenson AG (eds) Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St. Louis, pp 289–303
189. Mallet J (1995) A species definition for the modern synthesis. Trends Ecol Evol 10: 294–299
190. Templeton AR (1998) Species and speciation: geography, population structure, ecology, and gene trees. In: Howard DJ, Berlocher SH (eds) Endless forms: species and speciation. Oxford University Press, New York, pp 32–43

Chapter 2

Guidelines for the Choice of Sequences for Molecular Plant Taxonomy

Pascale Besse

Abstract

This chapter presents an overview of the major plant DNA sequences and molecular methods available for plant taxonomy. Guidelines are provided for the choice of sequences and methods to be used, based on the DNA compartment (nuclear, chloroplastic, mitochondrial), evolutionary mechanisms, and the level of taxonomic differentiation of the plants under survey.

Key words Nuclear DNA, Chloroplast DNA, Mitochondrial DNA, Repeated DNA, Low-copy DNA, Evolution, Molecular plant taxonomy

1 The Plant Genome and Regions Targeted for Molecular Plant Taxonomy

The nuclear genome in plants is very complex as in many eukaryotes, as illustrated by the "C-value enigma" [1, 2]: although the overall haploid DNA content (C-value) increases with apparent biological complexity, some species have more DNA in their haploid genome than some more complex organisms. Also, for a similar level of biological complexity, some species, such as plants, exhibit a surprisingly wide range of C-values (Fig. 1). This apparent discrepancy can be in part explained by the occurrence of variable amounts of repetitive DNA in the genomes (Fig. 1), most of which is constituted by noncoding sequences [3].

1.1 Repeated Nuclear DNA Sequences

Most nuclear sequences targeted in molecular taxonomy experiments belong to the category of highly repetitive DNA. Nuclear ribosomal RNA genes (nrDNA) are tandemly (side by side) repeated and located at a few loci in plant genomes [4–6] (Fig. 2). These, and particularly the ITS (internal transcribed spacers) [7, 8], have long been widely used for resolving plant taxonomic issues, initially using restriction analysis and then sequencing (Chapter 7). Microsatellite markers, also called STR (simple tandem repeats) or SSR (simple sequence repeats), are tandem repeats of small

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_2,

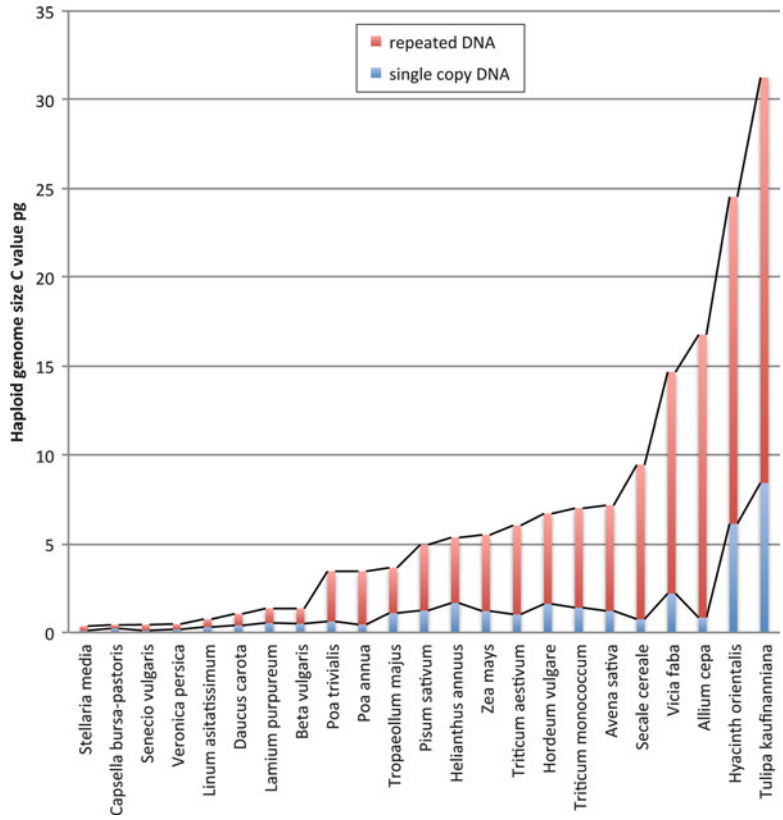


Fig. 1 Haploid genome size and composition for different plant species (graph built from data taken in [50])

stretches of noncoding DNA sequences, discovered in 1989 and named after the discovery of minisatellite and satellite DNA which exhibited a similar tandem structure [9] (Fig. 2). Microsatellites are widely used for diversity studies either as powerful single locus markers easily amplified by PCR (Chapter 9) or in multi-locus profiling methods revealing regions between adjacent SSRs (inter-simple sequence repeats, ISSR) by PCR amplification (Fig. 3) (Chapter 11).

Transposable elements (TEs) represent another class of repeated DNA, but the elements are dispersed across the genome instead of being tandemly repeated and these also can represent an important part of the plant nuclear genome. Two main classes of TEs exist in plants: class I retrotransposons (which transpose through a RNA copy which is then reverse transcribed into DNA and inserted at a new site) and class II transposons (which are excised and transpose directly as DNA) (Fig. 4). Class I retrotransposons are more numerous in genomes than class II as the original copy of the transposon is retained after transposition. In maize, for example, LTR (long terminal repeats)-retrotransposons represent up to 70 % of the nuclear genome [10]. Class I transposons (either LTR-retrotransposons or non-LTR-SINEs (short interspersed nuclear

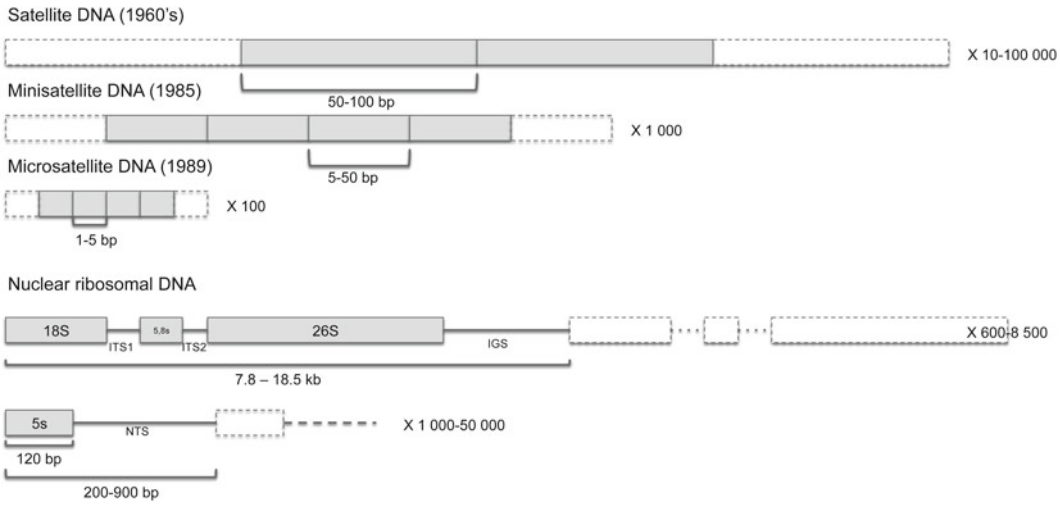


Fig. 2 Tandem repeat sequences used for molecular plant taxonomy: structure and number of tandem repeats

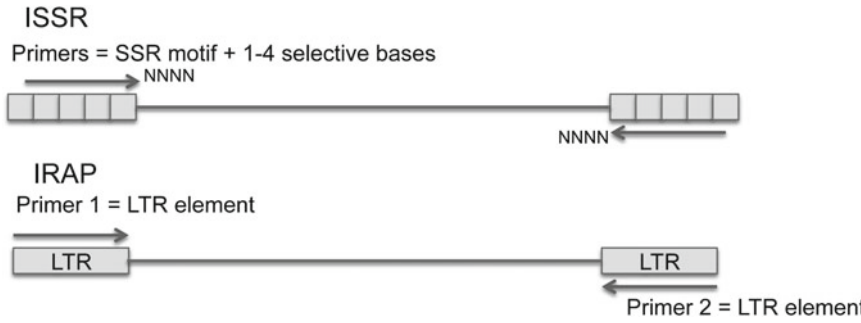


Fig. 3 Multi-locus profiling methods using either SSR or retrotransposons as anchors

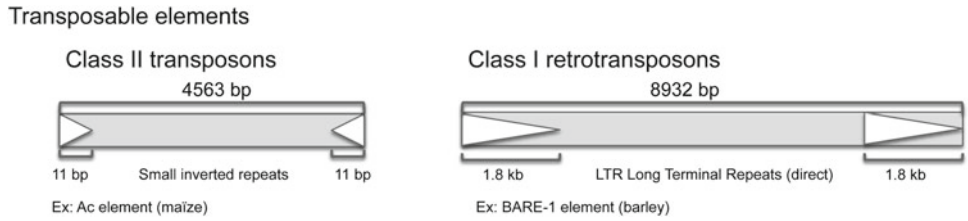


Fig. 4 Transposable elements in plants

elements)) are now commonly used for phylogenetic and taxonomic studies. Many studies use multi-locus PCR-based profiling methods such as inter-retrotransposons amplified polymorphism (IRAP) (Fig. 3) which amplifies regions between adjacent LTR repeats of LTR-retrotransposons (Chapter 12). As described for eukaryotes [11], SINEs are considered as perfect markers and are also being sequenced to build robust plant phylogenies although these studies

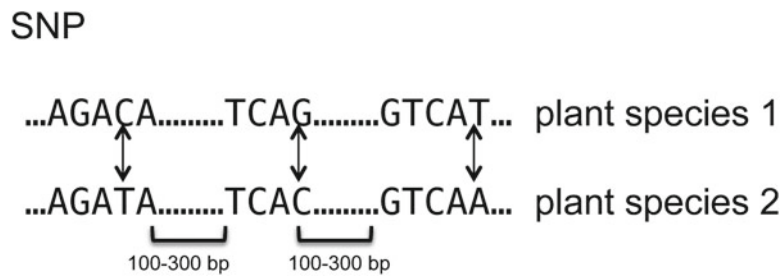


Fig. 5 SNPs in plants

are restricted to a limited number of plant species (mainly cultivated species) for which SINEs have been described and isolated [12, 13].

1.2 Low-Copy Nuclear Genes

Contrary to ribosomal DNA nuclear genes, low-copy nuclear genes (LCNG) do not suffer the possible disadvantages of concerted evolution, paralogy, and homoplasy [7, 8, 14, 15] that can be particularly limiting for taxonomic studies in recent hybrids or polyploids (*see* Chapter 7). However, care must be taken if using low-copy genes belonging to multigenic families for which paralogy and concerted evolution issues might still be problematic [16].

Despite their advantages, single-copy nuclear genes have not so much been used for plant taxonomy as they are much more difficult to isolate and characterize, contrary to cpDNA or ribosomal nuclear DNA which has been extensively used because they are easily amplified using universal primers [16] (Chapters 5 and 7). This situation is however changing rapidly [15, 17]. With the availability and affordability of new sequencing technologies [18], it is now becoming feasible to assess variations at a wide range of single or low-copy genes in nuclear genomes giving access to powerful phylogenomic analyses [19, 20]. Rather than sequencing complete genes for all accessions, single nucleotide polymorphisms (SNPs) can be searched for and analyzed (Fig. 5) (Chapter 8), and various sequence-based SNP assays can then be designed [21].

1.3 Anonymous Sequences

Many molecular technologies also rely on revealing variations at randomly picked anonymous sequences in genomes. In such techniques, the importance is not the nature of the target sequence itself, but rather the high throughput of the technology, which allows revealing numerous markers (loci) covering the genome. The aim is to give an as accurate as possible view of the genome diversity. This is the case for amplified fragment length polymorphism (AFLP) (Chapter 11), randomly amplified polymorphic DNA (RAPD) (Chapter 10), and associated techniques (Fig. 6) which use primers with arbitrary sequence to amplify genomic regions. Some multi-locus profiling techniques use a combination of AFLP associated with the revelation of either SSR loci (selective amplification of microsatellite polymorphic loci, SAMPL) (Chapter 11) or

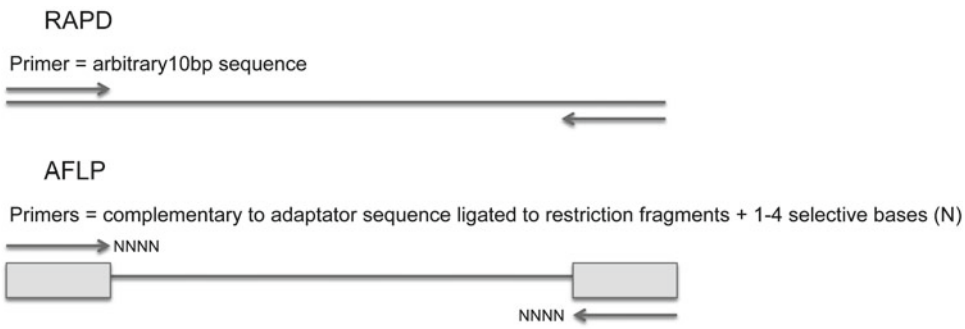


Fig. 6 Markers revealed by RAPD and AFLP

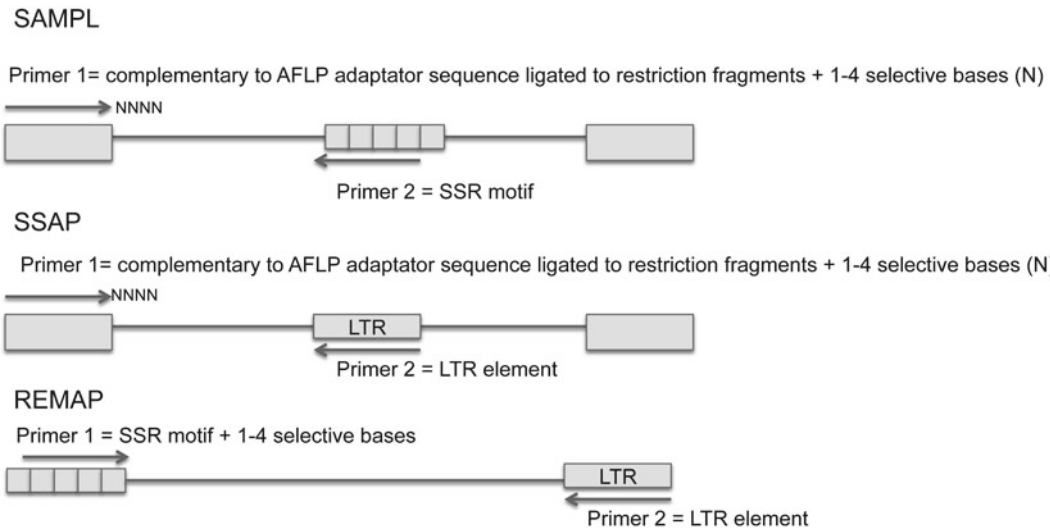


Fig. 7 Multi-locus profiling methods using a combination of anchors based on AFLP, SSR, or LTR

LTR-retrotransposons (sequence-specific amplified polymorphism, SSAP) (Chapter 12); others combine anchor primers in both SSR and LTR-retrotransposon conserved regions (retrotransposon-microsatellite amplification polymorphism, REMAP) (Chapter 12) (Fig. 7). A new technology termed DArT (diversity array technology) [22, 23] was also recently developed. It uses high-throughput DNA-array technology to reveal polymorphisms between individuals without any prior sequence information knowledge and is therefore applicable to non-model species.

1.4 Organellar DNA

In plants, the genetic information is also carried on the mitochondrial as well as chloroplast genomes (organellar DNA). Although mitochondrial genome (mtDNA) has received little attention in plant taxonomic studies (but *see* Chapter 6) because of numerous rearrangements and low levels of sequence variation, chloroplast DNA (cpDNA) has been widely used in molecular plant phylogeny (Chapter 5) through sequencing, restriction, or chromatography.

2 Evolutionary Considerations

The molecular clock hypothesis suggests that nucleotide substitutions occur at a roughly constant rate between and within evolutionary lineages across time [24] and has given rise to different models to estimate this evolutionary rate and its constancy [25]. According to the neutral theory of evolution, the speed of this rate (the amount of molecular variation accumulated over time) depends on the structural and functional constraints of the molecule [26]. This can be illustrated by noncoding DNA molecules (such as introns or intergenic sequences) evolving much faster than coding DNA as they accumulate more variations over time. Also it is now well admitted that third position bases in codons evolve much faster than other positions due to the redundancy of the genetic code [26] (less functional constraint on the third position allows for more variations to accumulate over time). Most markers generated using RAPD or AFLP technology have been shown by genome-mapping experiments to cluster around the centromeres of chromosomes [27–30], a heterochromatin region with mainly noncoding sequences. Consequently, these markers often reveal an important amount of variation.

The evolutionary rate of a molecule is also driven by its evolutionary mechanisms. Microsatellite markers are the most variable molecules known to date. They are mostly noncoding molecules and vary in length (due to the variation in the number of tandem repetitions or VNTR) due to replication slippage (SMM model [31]), which occurs at a high frequency (10^{-6} to 10^{-2}) in plants [32]. Microsatellites with shorter motifs and greater number of repeats are more prone to replication slippage and are thus the most variable [33]. ISSR, SAMPL, and REMAP markers, which use a microsatellite locus as an anchor, also beneficiate to a certain extent from the microsatellite length hyper-variability. Minisatellite sequences that tend to evolve through unequal crossing-over (IAM model [31]), which is a phenomenon with greater frequency than simple base mutations, also vary in length (i.e., number of tandem repeats) with great frequency. Both types of sequences have been for this reason used for generating powerful DNA fingerprints in human [34, 35] and subsequently in numerous species including plants.

Most tandemly repeated sequences in the genome evolve through what is known as "concerted evolution" or molecular drive [36, 37], which involves mechanisms such as unequal crossing-over or biased gene conversion. Over time, the sequences that compose a family of tandem repeats within an individual genome are maintained similar, thanks to this concerted evolution [6, 38–40]. Such sequences also tend to be maintained identical through close lineages within a species and will therefore display a slower evolutionary rate than molecules without concerted evolution.

In the cpDNA, like in the nDNA, intergenic noncoding sequences evolve faster than coding sequences. For example, by testing seven different sequences on a range of land plants, [41] classified these sequences by order of variation as follows: *psbK-psbI* > *trn*H-*psb*A > *atp*F-*atp*H > *mat*K > *rpo*B > *rpo*C1 > *rbc*L, illustrating that cpDNA intergenic regions are more variable than coding regions. Globally, in plants, organellar sequences evolve more slowly than nuclear sequences: mtDNA evolves three times slower than cpDNA, which in turn evolves two times slower than nDNA (average synonymous substitution rates per site per year for mtDNA and cpDNA are $0.2–1.0 \times 10^{-9}$ and $1.0–3.0 \times 10^{-9}$, respectively [42]) (Chapter 6). Even the most variable of intergenic regions in cpDNA is less variable than nuclear ITS: ITS reveals 2.81 % sequence divergence in a range of plant families compared to 1.24 % divergence for *trn*H-*psb*A, one of the most variable intergenic cpDNA regions [43].

Finally, class I TEs are good classification criteria to evaluate species phylogenetic relationships; their mode of transposition ("copy–paste" mode) makes them numerous and implies no ambiguity in the ancestral state definition, which is, for a given locus, the absence of TE [11, 12]. Class II TEs are less appropriate for phylogenetic issues mainly because of their direct mode of transposition ("cut–paste" mode) which, associated with possibilities of horizontal transfer, can lead to erroneous classifications (TE phylogenetic trees not concordant with species phylogenetic history) [44, 45].

3 Choice of Sequences for Molecular Taxonomy

These evolutionary considerations are of primary importance when one wants to use a DNA sequence to infer phylogenetic relationships between a set of accessions. Two questions have to be considered when starting a molecular taxonomy project:

1. What is the degree of time divergence between the accessions under study? Do we want to address variations at the intraspecific level (population level) or are we comparing species from the same genus or different genera from the same family or above?
2. What is the evolutionary rate of the molecule that will be used to infer relationships between accessions?

The rule to keep in mind is that the further we need to go in evolutionary times, the slower the molecule must evolve. Going too far with too much diverging sequences will lead to homoplasy (characters identical by state, not by descent) through convergence or reversion. On the opposite, slow evolving sequences will not be enough in discriminating for groups that have evolved recently (Fig. 8).

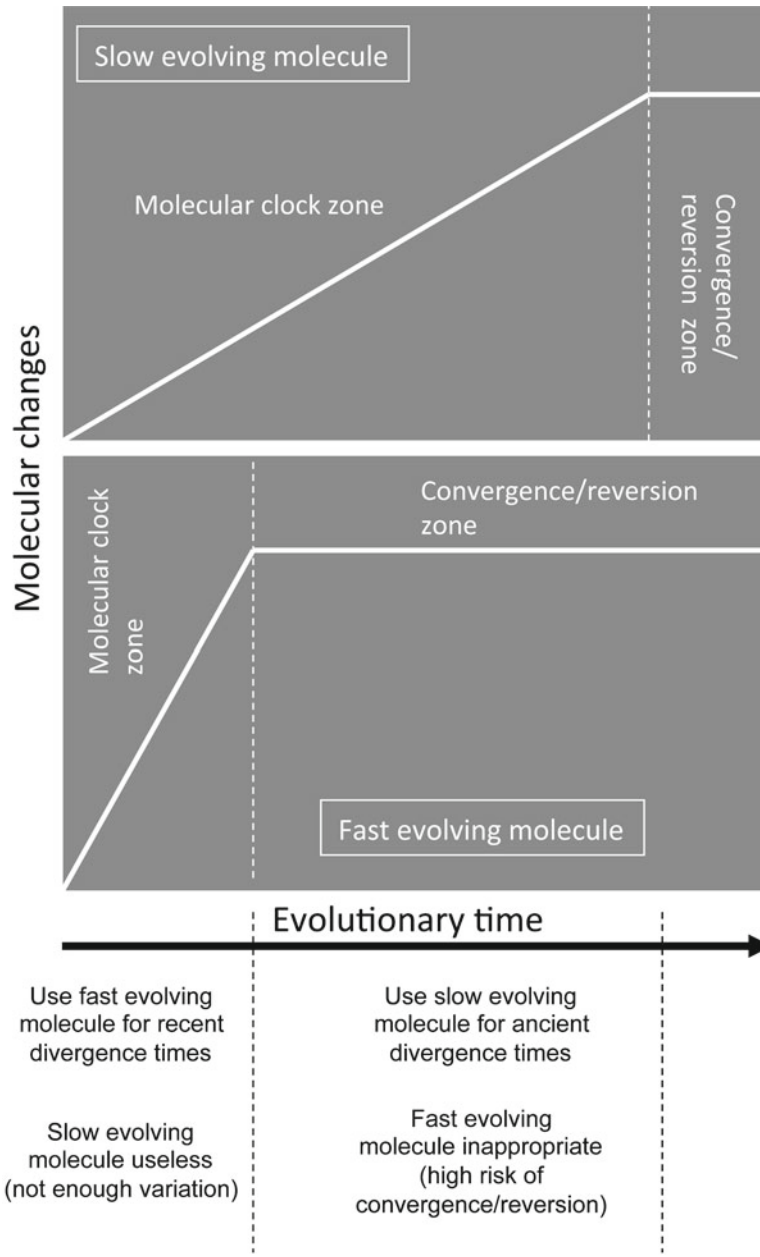


Fig. 8 Illustration of the usefulness of rapidly evolving versus slow evolving sequences in molecular taxonomy assessment of recently or anciently diverged groups. The curvilinear relationship between molecular changes and time is represented theoretically starting with a constant accumulation rate (molecular clock hypothesis) which plateaus as a consequence of the saturation of the sequence over time. The faster the sequence evolves, the faster the plateau is reached

Figure 9 illustrates this rule: if a very slow evolving sequence is used, it might be unable to differentiate the two hypothetical species under study (Fig. 9a). A sequence with an intermediate rate of evolution and concerted evolution would allow the identification

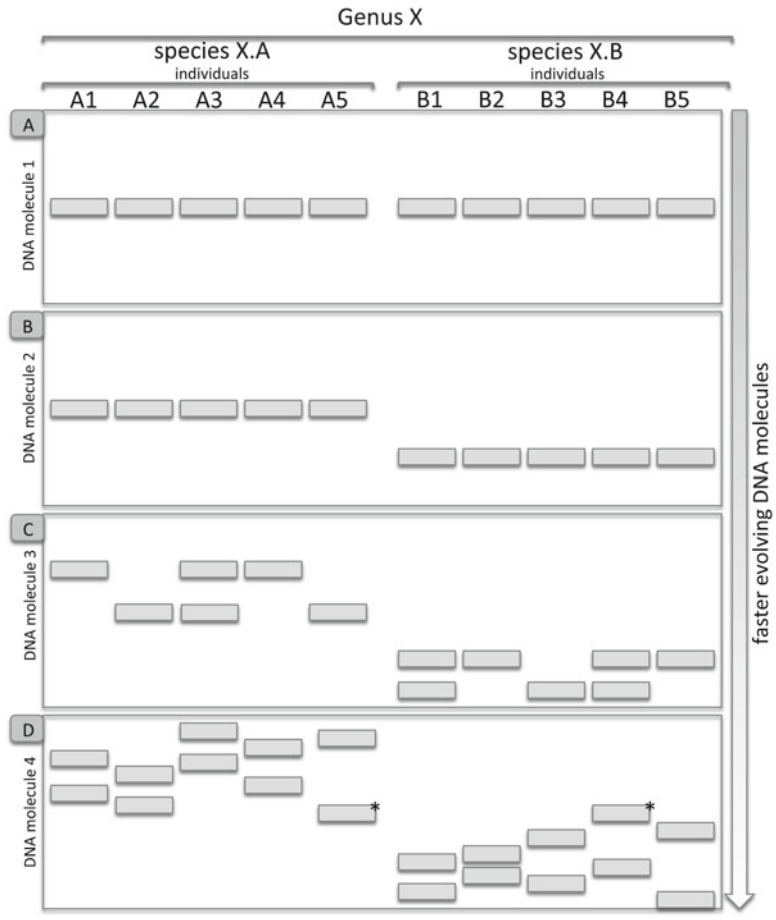


Fig. 9 Illustration of the differentiation power of DNA molecules depending on their evolutionary rates

of each species, but would be unable to reveal any intraspecific variability (Fig. 9b). To reach such level of informativeness, one would need to use a single-copy gene (Fig. 9c) or a microsatellite marker (Fig. 9d), but the latter, due to high evolutionary rate, may generate homoplasy (*) which could lead to erroneous interpretations if comparing species A and B, as individual B4 would appear more related to species A than to individuals from species B. Such rapidly evolving sequences are therefore not appropriate for studying relationships at too high taxonomic levels.

Guidelines for the choice of sequences to be used depending on the level of taxonomic divergence are illustrated (Fig. 10). It must be kept in mind that the level of taxonomic differentiation can vary considerably depending on the species group; therefore one always needs to perform preliminary tests of various sequences on a representative subset of accessions to assess their power in differentiating our own individuals, species, or genera of interest.

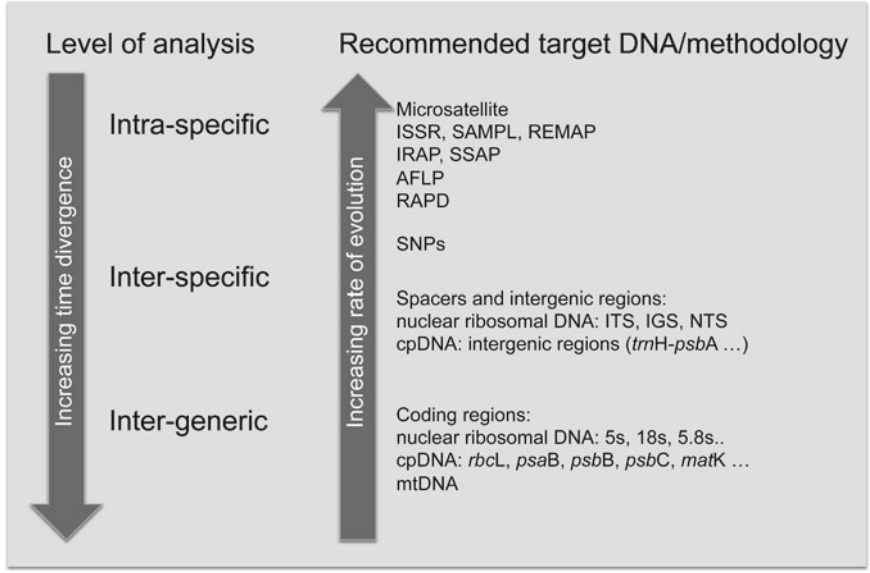


Fig. 10 General guidelines for the choice of markers to be used for plant taxonomy

4 Genetic Considerations

Knowledge of the mode of inheritance of the molecules under study is also of great importance. Nuclear sequences are inherited in a Mendelian fashion, with contribution from both parents. Organellar (chloroplastic and mitochondrial) sequences are almost always uniparentally inherited (generally maternally, but see [46]). This can have important consequences when building a molecular phylogeny, as individuals or species of interspecific origin will appear inconsistently on the trees generated with each type of markers (Fig. 11): a species B of hybrid origin will be grouped with its mother species A using cytoplasmic sequences, although it will appear different from it on the nuclear tree.

AFLP, RAPD, ISSR, and other multi-locus profiling methods generate >90 % dominant markers [47]. The polymorphism revealed is mainly due to mutations in the hybridization region of one of the primers, leading to either amplification of the locus (presence) or null allele (absence of amplification), i.e., a dominant system (Fig. 12). Consequently, such methods provide only biallelic markers.

On the other hand, microsatellites are very powerful monolocus markers as they are multiallelic and codominant (Fig. 12). They are indeed widely used in molecular ecology and population genetic studies as heterozygous loci can be clearly identified and allelic frequencies can be calculated to test for deviations from Hardy-Weinberg equilibrium. One microsatellite multiallelic marker provides as much genetic information as four to ten biallelic AFLP markers [48].

SNP markers are monolocus, codominant, but are biallelic. Indeed, they evolve through the infinite sites (IAM) model: given

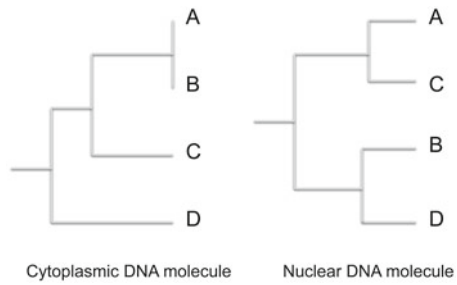


Fig. 11 A hypothetic phylogeny involving a hybrid species B whose maternal parent is species A

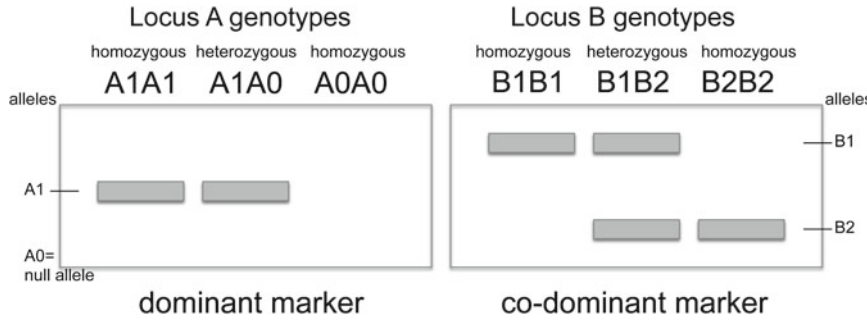


Fig. 12 Different genetic profiles: dominant versus codominant markers

the low rate of substitutions in genomes (the average synonymous substitution rate in plant nuclear genome is about $5.0–30.0 \times 10^{-9}$ per site per year [42]), the probability of more than one mutation at a given site is negligible; therefore each SNP is almost exclusively found only with two different states among the four possible (A, G, C, or T). For population genetic studies, it will be necessary to compensate the low allelic diversity of SNP markers by increasing the number of studied loci (2–6 times more SNP locus are needed as compared to microsatellites [49] to reach the same level of informativeness).

5 Analyzing Results

Fragment length data (different band sizes visualized and coded after electrophoretic separation) will only be analyzed using distance-based methods (e.g., UPGMA or neighbor joining), whereas sequence data will be analyzed either using distance-based methods or more powerfully using character-based methods (e.g., using maximum parsimony or maximum likelihood), allowing true phylogenetic trees to be constructed rather than phenetic trees (Chapter 13). Always remember that the tree built is a sequence tree, not a species trees. For all the reasons discussed above, using different sequences can lead to different trees reflecting the different evolutionary patterns of the sequences under study.

6 Further Exploration: Chromosomal Organization

In plants, genome organization is very complex and polyploidy can be an important speciation mode. It will be almost impossible to differentiate, for example, a diploid species from a related autopolyploid species in a phylogenetic tree. Molecular taxonomy can be greatly enhanced in some taxonomic complex plant groups by assessing not only phylogenetic relationships but also genome organization to determine introgression, hybridization, or polyploidization (by analyzing either chromosomes or simply genome size) (Chapters 14–16).

References

1. Gregory T (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev 76:65–101
2. Thomas C (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237–256
3. Schmidt T, Heslop-Harrison JS (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci 3:195–199
4. Hamby RK, Zimmer EA (1992) Ribosomal RNA as a phylogenetic tool in plant systematics. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Chapman & Hall, New York
5. Schaal BA, Learn GH (1988) Ribosomal DNA variations between and among plant populations. Ann Mo Bot Gard 75:1207–1216
6. Hillis DM, Dixon MT (1991) Ribosomal DNA : molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453
7. Alvarez IA, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogenet Evol 29:417–434
8. Poczai P, Hyvönen J (2010) Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep 37:1897–1912
9. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445
10. SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82:37–44
11. Ray DA (2007) SINEs of progress: mobile element applications to molecular ecology. Mol Ecol 16:19–33
12. Deragon JM, Zhang X (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. Syst Biol 55:949–956
13. Schmidt T (1999) LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol 40:903–910
14. Feliner GN, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol Phylogenet Evol 44:911–919
15. Zimmer EA, Wen J (2013) Using nuclear gene data for plant phylogenetics: progress and prospects. Mol Phylogenet Evol 66:539–550
16. Small RL, Cronn RC, Wendel JF (2004) Use of nuclear genes for phylogeny reconstruction in plants. Aust Syst Bot 17:145–170
17. Schlötterer C (2004) The evolution of molecular markers – just a matter of fashion? Nat Rev Genet 5:63–69
18. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8:3–17
19. Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. Am J Bot 99: 312–319
20. Timme RE, Bachvaroff TR, Delwiche CF (2012) Broad phylogenomic sampling and the sister lineage of land plants. PLoS One 7:1–8
21. Syvänen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942
22. Jaccoud D, Peng K, Feinstein D, Killian A (2001) Diversity arrays: a solid state technology for sequence information independant genotyping. Nucleic Acids Res 29:1–7
23. James KE, Schneider H, Ansell SW, Evers M, Robba L, Uszynski G, Pedersen N, Newton AE, Russell SJ, Vogel JC, Kilian A (2008) Diversity arrays technology (DArT) for pangenomic evolutionary studies of non-model organisms. PLoS One 3:1–11
24. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel H (eds) Evolving genes and proteins. Academic, New York, pp 97–166
25. Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating divergence

times from molecular data on phylogenetic and population genetic timescales. Annu Rev Ecol Syst 33:707–740
26. Kimura M, Ohta T (1974) On some principles governing molecular evolution*(population genetics/mutational pressure/negative selection/random drift). Proc Natl Acad Sci USA 71:2848–2852
27. Saliba-Colombani V, Causse M, Gervais L, Philouze J (2000) Efficiency of RFLP, RAPD, and AFLP markers for the construction of an intraspecific map of the tomato genome. Genome 43:29–40
28. Qi X, Stam P, Lindhout P (1998) Use of locus-specific AFLP markers to construct a high-density molecular map in barley. Theor Appl Genet 96:376–384
29. Saal B, Wricke G (2002) Clustering of amplified fragment length polymorphism markers in a linkage map of rye. Plant Breed 121:117–123
30. Young WP, Schuppet JM, Keim P (1999) DNA methylation and AFLP marker distribution in the soybean genome. Theor Appl Genet 99: 785–792
31. Shriver MO, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993
32. Bhargava A, Fuentes FF (2010) Mutational dynamics of microsatellites. Mol Biotechnol 44:250–266
33. Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28:1040–1050
34. Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific fingerprints of human DNA. Nature 316:76–79
35. Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. Nat Rev Genet 5: 739–752
36. Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299: 111–117
37. Dover G (1994) Concerted evolution, molecular drive and natural selection. Curr Biol 4: 1165–1166
38. Plohl M, Luchetti A, Meštrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409:72–82
39. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152
40. Ganley ARD, Kobayashi T (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res 17:184–191
41. Hollingsworth ML, Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan RO, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour 9:439–457
42. Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. PNAS 84:9054–9058
43. Kress WJ, Wurdack KJ, Zimmer EA, Weig LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. PNAS 102:8369–8374
44. Capy P, Anxolabehere D, Langin T (1994) The strange phylogenies of transposable elements: are horizontal transfers the only explanation. Trends Genet 7–12
45. Syvanen M (1994) Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet 28:237–261
46. McCauley DE, Sunby AK, Bailey MF, Welch ME (2007) Inheritance of chloroplast DNA is not strictly maternal in Silene vulgaris (Caryophyllaceae): evidence from experimental crosses and natural populations. Am J Bot 94: 1333–1337
47. Bensch S, Akesson M (2005) Ten years of AFLP in ecology and evolution: why so few animals? Mol Ecol 14
48. Mariette S, Corre VL, Austerlitz F, Kremer A (2002) Sampling within the genome for measuring within population diversity: trade-offs between markers. Mol Ecol 11:1145–1156
49. Morin PA, Luikart G, Wayne RK, The SNP workshop group (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19: 208–216
50. Flavell R, Bennett M, Smith J, Smith D (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12:257–259

Chapter 3

Leaf Tissue Sampling and DNA Extraction Protocols

Kassa Semagn

Abstract

Taxonomists must be familiar with a number of issues in collecting and transporting samples using freezing methods (liquid nitrogen and dry ice), desiccants (silica gel and blotter paper), and preservatives (CTAB, ethanol, and isopropanol), with each method having its own merits and limitations. For most molecular studies, a reasonably good quality and quantity of DNA is required, which can only be obtained using standard DNA extraction protocols. There are many DNA extraction protocols that vary from simple and quick ones that yield low-quality DNA but good enough for routine analyses to the laborious and time-consuming standard methods that usually produce high quality and quantities of DNA. The protocol to be chosen will depend on the quality and quantity of DNA needed, the nature of samples, and the presence of natural substances that may interfere with the extraction and subsequent analysis. The protocol described in this chapter has been tested for extracting DNA from eight species and provided very good quality and quantity of DNA for different applications, including those genotyping methods that use restriction enzymes.

Key words Blotter paper, CTAB, Desiccant, DNA extraction, Dry ice, Ethanol, Leaf sampling, Silica gel

1 Introduction

The availability of direct polymerase chain reaction (PCR) system helps to amplify DNA in a few easy steps that eliminate the need for standard DNA extraction (isolation) protocols. A number of commercial companies supply kits for such direct PCR. These include, among others, the Thermo Scientific Finnzymes' Phire® Plant Direct PCR Kit, the PlantDirect™ Multiplex PCR System from GenScript, the Sigma-Aldrich Extract-N-Amp™ Plant Kit, the New England Biolabs Phire® Plant Direct PCR Kit, and the Terra PCR Direct Polymerase Mix from Clontech. Most Direct PCR protocols use either a small piece of plant material directly as template in the PCR mix or add a small volume of the dilution buffer containing crushed plant material in the PCR mix. Some of the template DNA sources for direct PCR include (a) alkali-treated intact plant tissue [1, 2], (b) leaf squashes or juices [3, 4],

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_3,

and (c) FTA card (Whatman Inc., Clifton, NJ). FTA card is a paper specially treated with chemicals that inactivates pathogens, protects the DNA from degradation, and allows the cards to be stored at room temperature for extended periods of time [5]. Plant tissue is physically crushed on the card; small discs are punched from the card, washed with aqueous buffers, and dried; and a small disk (1.2 or 2 mm in diameter) is used as templates for PCR (http://www.whatman.com/References/WGI_1397_PlantPoster_V6.pdf). For most applications, however, a reasonably good quality and quantity of DNA is required, which can only be obtained using standard DNA extraction protocols. Therefore, a fast, simple, high-throughput, and reliable extraction protocol for genomic (nuclear, mitochondrial, and chloroplast DNA) or organelle DNA from leaves, seeds [6–8], roots [9], flower petals [10], and pollen grains [11] is a prerequisite for most molecular marker studies. This review and protocol focuses only on leaves as most molecular studies use leaf DNA.

Researchers must be familiar with a number of issues prior to collecting leaf samples including the molecular technique chosen for genotyping, the nature of the tissue to be collected (soft and thin leaves for mesophytes, hard and/or thick leaves for xerophytes, and succulent leaves), the quantity of tissue to be collected, the geographical distance of the sampling sites, anticipated duration of the collection mission, and the mode of transport from the sampling site to the laboratory. All of these factors should be considered in order to prepare sufficient quantity of the required materials prior to the sample collection mission.

DNA can be extracted from fresh, frozen, lyophilized, preserved, or dried leaf samples. Fresh or fresh-frozen leaves in liquid nitrogen (http://en.wikipedia.org/wiki/Liquid_nitrogen) are ideal for DNA extractions. However, it is difficult to immediately place samples in liquid nitrogen in the field because (a) it requires a liquid nitrogen container with a wide mouth and a leakage-proof lid; (b) liquid nitrogen evaporates from such relatively large tank in a shorter period of time; and (c) there are safety concerns surrounding transporting liquid nitrogen in vehicles and in air planes. Dry ice (http://en.wikipedia.org/wiki/Dry_ice) is a solid form of carbon dioxide, and it can be used as an alternative to liquid nitrogen. Dry ice can be easily transported in Styrofoam cooler boxes (http://www.mrboxonline.com) that are well sealed in packing tape. It can last up to a week in these cooler boxes, although the duration depends on the climatic conditions of the sampling sites, the quantity of samples, and the thickness/insulation of the Styrofoam boxes. There are little or no safety concerns for road transportation of dry ice when compared to liquid nitrogen. However, airlines are reluctant to transport dry ice. Gel pack (http://www.coldchaintech.com/products/gel-packs.php) can also be used for sampling and transporting plant materials from the

field. They last for up to a week but their heavy weight can cause samples to break up into pieces. Also, the cooling capacity of gel packs is not as good as both liquid nitrogen and dry ice.

Many taxonomists rely on silica gel (http://en.wikipedia.org/wiki/Silica_gel) to dry their plant samples [12]. This method is generally considered effective for most species as the rate of desiccation is rapid enough to prevent DNA degradation. However, the improper use of silica gel (using too little silica gel relative to the amount of leaf tissue and not breaking up large leaves into smaller pieces) often results in DNA degradation that may not be suitable for applications that requires high molecular weight DNA, such as restriction fragment length polymorphism [13] and genotyping by sequencing [14]. Some researchers reported the use of sodium chloride/cetyltrimethylammonium bromide (NaCl/CTAB) preservation solution [15, 16], while others used a wide range of preservative chemicals, such as ethanol, acetone, 2-propanol (isopropanol), diethyl ether, and ethyl acetate [17, 18]. Another option is to dry the leaves at ambient temperature by pressing them on blotter (absorptive tissue) paper that removes the moisture from the leaves. The blotter paper must be changed a couple of times until the samples are properly dried. We used dry ice, ethanol, isopropanol, NaCl/CTAB solution, silica gel, and blotter papers for collecting leaf samples from fields for eight species and extracted DNA after storage up to a month at room temperature except those samples transported in dry ice that were stored in freezers (−20 or −80 °C) after arrival in to the lab. All the six leaf sampling methods provided reasonably good DNA quality that work well for microsatellite or simple sequence repeat (SSR) markers. However, the DNA isolated from samples collected in dry ice, silica gel, and blotter paper were consistently high quality (Fig. 1) than the preservative methods for applications that require restriction digestion of the isolated DNA. The quality of DNA isolated from NaCl/CTAB preserved samples was very good on agarose gels, but the ratio of absorbance at the 260 and 230 nm in spectrophotometers was consistently lower than 1.7 for all species that we tested. Such lower ratios are indictions for the presence of contaminants (impurities) that may affect the use of DNA for some applications.

Once the plant samples are brought into the lab, the next step is to prepare them for DNA extraction. Depending on the purpose of the research, DNA can be extracted from a single plant or bulks of several plants. The heterogeneous nature of outcrossing species contributes to high levels of within accession or variety genetic variability that requires for the analysis of a large number of representative individuals per accession. As genotyping of a large number of individual samples per accession is costly and time consuming, several studies have used the bulked method [19–24] for quick molecular studies. Plant bulking or DNA pooling can be used for such purposes. In plant bulking, researchers harvest leaf

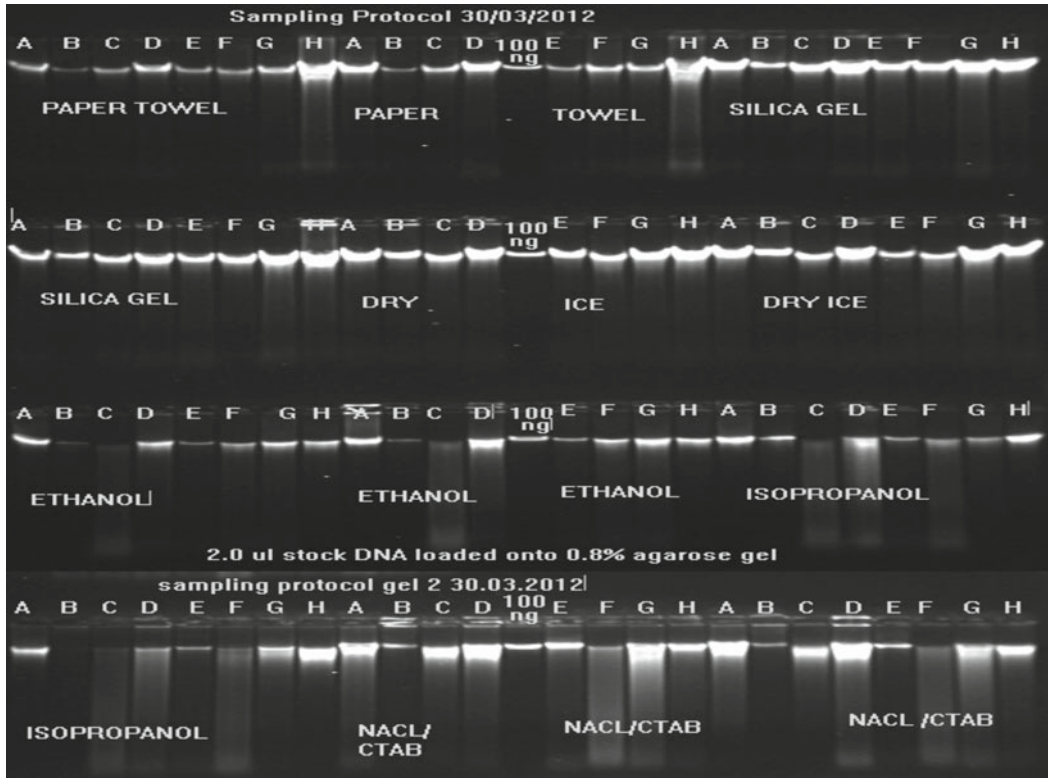


Fig. 1 The quality of DNA extracted from leaf samples collected using different options: dry ice, 100 % ethanol, 100 % isopropanol, silica gel, tissue paper, and NaCl/CTAB. Each sample (A, cassava; B, sweet potato; C, orange; D, bougainvillea; E, banana; F, grasses; G, maize; and H, papaya) was extracted in duplicate

samples from a number of individuals from an accession and bulk about equal tissue from each individual to form a representative sample before DNA isolation. In DNA pooling, leaf samples are harvested separately from a number of individuals; the DNA is isolated independently and then mixed in equimolar amounts to create a representative DNA sample before any marker analysis. The bulking strategy significantly reduces the burden of handling large number of samples per accession and significantly reduces the genotyping cost. The numbers of individual plants used to form a bulked sample varied from 2 to 30 individuals per accession [25, 26]. The bulk sizes are often larger for outcrossing than for self-pollinating species, as outcrossing species are generally more heterogeneous within an accession. The optimal bulk size also depends on the sensitivity of detection of the particular marker technique used. However, there are still inconsistencies in terms of the number of individuals to be used per bulk and the number of bulks per accession or population. Several studies [19–23] used a single bulk of 8–15 individuals per bulk, but some studies [19, 20] suggested using two bulks of 15 individuals each (a total of 30 individuals) per accession or variety. We recently genotyped 218

open-pollinated maize varieties in two sets, with each set containing equal bulks of leaf tissue from 15 individuals, with 51 fluorescently labeled simple sequence repeat (SSR) markers, and found high correlation (0.85–0.97) between the two bulks and the combined dataset (Semagn K, unpublished). However, bulking methods can complicate genotyping, allele calling, and allele frequency estimation compared to individual analysis.

DNA extraction can be carried out using reagents prepared in the lab (homemade extraction buffers) or one of the several commercially available DNA extraction kits. Some of the commercially available DNA extraction kits include the Qiagen DNeasy Plant DNA extraction kit (http://www.qiagen.com/products/genomicdnastabilizationpurification/animalplantsamples.aspx), the Nucleon PhytoPure system from Nudeon Bioscience (http://www.gen-probe.com/pdfs/downloads/PhytoPure-datasheet.pdf), the Promega wizard genomic DNA purification kit (http://www.promega.com/resources/articles/pubhub/enotes/wizard-genomic-dna-purification-kit-and-the-isolation-of-plant-genomic-dna), the Zymo Research ZR plant/seed DNA kits (http://www.zymoresearch.com/product/zr-plantseed-dna-miniprep-d6020), the Norgen genomic DNA isolation kit (http://www.norgen-biotek.com/display-product.php?ID=75), and the Bioneer ExiPrep plant genomic DNA kit (http://us.bioneer.com/products/instrument/16Proordering.aspx).

For homemade extraction buffers, there are many protocols available, and the protocol to be chosen will depend on the quality and quantity of DNA needed, nature of samples, and the presence of natural substances that may interfere with the extraction and subsequent analysis. The protocols vary from simple and quick ones [4, 27–29] that yield low-quality DNA but good enough for routine analyses to the laborious and time-consuming standard methods [30, 31] that usually produce high quality and quantities of DNA. Most DNA extraction protocols involve breaking (through grinding) cell walls and membranes in order to release the cellular constituents. Membrane lipids are removed using detergents, such as sodium dodecyl sulfate (SDS), cetyltrimethylammonium bromide (CTAB), or mixed alkyltrimethylammonium bromide (MATAB). The released DNA should be protected from endogenous nucleases, and ethylenediaminetetraacetic acid (EDTA) is often included in the extraction buffer to chelate magnesium ions that is a necessary cofactor for nucleases. DNA extracts often contain a large amount of RNA, proteins, polysaccharides, tannins, and pigments, which may interfere on the use of the DNA for downstream applications. Most proteins are removed by adding a protein-degrading enzyme (proteinase K), denaturation at 65 °C, and precipitation using chloroform and isoamyl alcohol. RNAs are normally removed using RNA-degrading enzyme called RNase A. Polysaccharide-like contaminants are removed using NaCl and CTAB [31, 32].

As DNA will be released along with other compounds (lipids, proteins, carbohydrates, and/or phenols), it must be separated from others by centrifugation. The DNA in the aqueous phase will then be transferred into new tubes and precipitated in salt solution (e.g., sodium acetate) or alcohol (100 % isopropanol or ethanol), redissolved in sterile water or buffer (either 10 mM Tris-HCl, pH 7.5–8.3 or a combination of 10 mM Tris and 0.1 mM EDTA, pH 8.0). Finally, the concentration of the extracted DNA should be measured using 0.8 % or 1 % agarose gel electrophoresis and/or spectrophotometer. Agarose gel is useful to check whether the DNA is degraded or not, but estimating DNA concentration by visually comparing band intensities of the extracted DNA with a molecular ladder or lambda (λ) DNA of known concentration is too subjective. A spectrophotometer measures either the intensity of absorbance of DNA solution or using fluorescent dyes. The most commonly used technique for measuring nucleic acid concentration is the determination of absorbance at 260 nm. The ratio of absorbance at 260 and 230 nm is used as measures for the presence of contaminants or impurities that may interfere in PCR or restriction digestion. NanoDrop spectrophotometer is one of the most commonly used spectrophotometers for such purpose, but it is highly unreliable for applications that require accurate DNA quantifications. Nevertheless, it works for most routine applications. The major disadvantages of all absorbance-based spectrophotometers include (a) the relative contribution of nucleotides and single-stranded nucleic acids to the signal, (b) the interference caused by contaminants that are common in DNA extraction, (c) the inability to distinguish between DNA and RNA, and (d) the inability of the method to tell the quality of the DNA that is highly required for certain applications. Fluorescent-based DNA quantification using either fluorescent spectrophotometers or plate readers is highly reliable, but it requires purchasing of a kit containing Quant-iT PicoGreen dsDNA reagent and λ-DNA (http://es-mx.invitrogen.com). The fluorescent kit may cost up to one US Dollar per sample. DNA quantification using spectrophotometers can be carried out on a single sample or on multiple samples (ranging from 8 to 1,536 samples at a time).

There are three possible outcomes to any DNA extraction: (a) there is no DNA; (b) the DNA appears as sheared (too fragmented), which is an indication of degradation for different reasons; and (c) the DNA appears as whitish thin threads (good quality DNA) or brownish threads (DNA in the presence of oxidation from contaminants such as phenolic compounds). The researcher, therefore, needs to test different protocols in order to find out the best one that works for the species under investigation. As most laboratories currently use the 96-well plate format both for DNA extraction and for PCR, this protocol describes processing of 192 samples (2 × 96) using 1. 2 mL strip tubes and strip caps. If users wish to extract fewer samples using Eppendorf Safe-Lock

1.5 or 2.0 mL tubes or 14–15 and 50 mL test tubes, they need to find an optimal proportion between the quantity of leaf tissue to be used and the volume of the reagents needed for extraction. This protocol has been tested for extracting DNA from eight species (banana, bougainvillea, cassava, grasses, maize, orange, papaya, and sweet potato) and provided very good quality (Fig. 1) and quantity of DNA for different applications, including genotyping by sequencing.

2 Materials

2.1 Leaf Sampling

1. Option 1: Preservatives—absolute ethanol, isopropanol, or NaCl/CTAB solution at a ratio of 3 and 35 g, respectively.
2. Option 2: Desiccant—silica gel.
3. Option 3: Blotter (tissue) papers.
4. Option 4: Styrofoam boxes containing dry ice.

2.2 DNA Extraction

1. 1.2 mL strip tubes with strip caps or 96 deep well plates with self-closing mats.
2. 4 mm metal grinding balls or 3 mm tungsten carbide beads.
3. Metal grinding ball dispenser (optional).
4. Plate centrifuge.
5. Shaking water bath.
6. Fume hood.
7. Dissecting scissors.
8. Dissecting forceps.
9. Lyophilizer or freeze dryer (optional).
10. Agarose gel electrophoresis facility (including gel casting trays, combs, electrophoresis tank, and power supply).
11. Tissue grinder (e.g., GenoGrinder or Tissue Lyser).
12. DNA extraction stock solutions: 1 M Tris, pH 7.5; 0.5 M EDTA pH 8.0 (*see* **Note 1**); 5 M NaCl; mercaptoethanol. Mercaptoethanol must be used under fume hood.
13. CTAB.
14. CTAB DNA extraction buffer that contains 0.20 M Tris, pH 7.5; 0.05 M EDTA, pH 8.0; 2 M NaCl; 2 % CTAB; and 1 % β-mercaptoethanol.
15. Liquid nitrogen or dry ice.
16. Single and 8-channel pipettes of different volume.
17. Pipette tips of different sizes.
18. Glass wares: beakers, flasks, reagent bottles of different sizes, measuring cylinders of different sizes.

19. Balances.
20. Weighing boats.
21. Heaters and stirrers.
22. RNase A.
23. Proteinase K (optional).
24. Ethanol (70 % and absolute).
25. Isopropanol.
26. Chloroform (must be used in fume hood).
27. Isoamyl alcohol.
28. Gloves.
29. Kleenex tissue paper.
30. Double-distilled water.
31. 0.1× TE (10 mM Tris, pH 7.5–8.3 and 0.1 mM EDTA, pH 8.0).
32. Agarose.
33. Ethidium bromide or GelRed.
34. Electrophoresis buffers: 10× Tris–borate EDTA (TBE) or 50× Tris–acetate EDTA (TAE).
35. Size markers with known concentrations—lambda (λ) DNA, 1 kb DNA ladder, and λ DNA-HindIII digest (optional).
36. Gel documentation system (gel camera, UV tray with white and UV light).
37. Spectrophotometer or plate reader (optional).

3 Methods

3.1 Leaf Sampling

Sample very young leaves from the shoot tips of a single healthy plant or bulks of about equal leaf tissue from several healthy plants per accession using one of the options described below. The three leaves closest to the shoot tip are the best; the younger the leaves, the better the DNA quality and quantities.

1. *Sampling using preservatives*: harvest very young leaves from the shoot tips, fold them, and put them into a 14–15 mL test tube. Write the sample number on a small piece of paper with a pencil and insert this into the tube. Add about 8 mL of 100 % ethanol, isopropanol, or NaCl/CTAB solution; close the caps tightly; and store at room temperature.
2. *Sampling using silica gel*: harvest very young leaves from the shoot tips, tear or cut the leaf material into smaller pieces, and put them into 3×4 inches Ziploc bags (*see* **Note 2**). Write the sample number on a small piece of paper and insert this into the

Ziploc bag. Add sufficient silica gel (deep blue in color) and close the Ziploc bag. The ratio of leaf tissue to silica gel should be at least 1:5, preferably 1:10. Check the Ziploc bags at a regular interval, and replace the silica gel when its color changes from blue to pink. Keep replacing the silica gel until the blue color remains stable. Most leaf samples completely dry within 24 h.

3. *Sampling using blotter (tissue) papers*: harvest small young leaves from the shoot tips, flatten the leaf surface without overlapping the leaves, and press them on to a pile of blotter papers. Write the sample number on a small piece of paper and insert between the blotter papers. To secure the samples, fold all four sides of the blotter papers and staple them or insert them in paper bags.
4. *Sampling using dry ice*: harvest young leaves from the shoot tips and put them in perforated Ziploc bags (*see* **Note 3**). Write the sample number on a small piece of paper with a pencil and insert this into the Ziploc bags. Immediately put the Ziploc bags with the leaf samples into a Styrofoam box containing dry ice. To minimize evaporation of the dry ice, always keep the lid (cover) of the Styrofoam box very tight by securing with a packing tape.

3.2 DNA Extraction

1. Assemble strip tubes (eight tubes per strip) into racks (hereafter referred as collection tubes) and label each strip.
2. Add two stainless steel grinding balls per tube and precool the collection tubes in liquid nitrogen or dry ice.
3. Harvest about 100–150 mg of fresh leaf sample (from a single plant or from bulks of several plants per accession), cut into pieces with a scissor, and transfer (using forceps) into the collection tubes (*see* **Note 4**). At all times, keep the collection tubes and samples in liquid nitrogen or dry ice. If a freeze dryer (lyophilizer) is available, freeze-dry the samples for 3–4 days as described in the user's manual. For leaf samples collected using one of the four options described above (preservatives, silica gel, blotter paper or dry ice), use an equivalent amount of tissue.
4. Prepare a fresh CTAB DNA extraction buffer (*see* **Note 5**) and keep this extraction buffer in 65 °C water bath until the samples are ready for grinding.
5. Sandwich each plate containing collection tubes between adapter plates and balance them using an ordinary weighing balance. Fix the sandwiched plates into the tissue grinder and grind the samples into fine powder as described in the user's manual. We often grind samples by shaking for 2 min at full (1×) speed of strokes/min using a GenoGrinder-2000 or at 25 Hz with a Tissue Lyser. Remove the plates, exchange the

position of the plates (outer plates towards inner and vice versa) (for fresh samples—deep in liquid nitrogen), and grind for an additional of 2 min (*see* **Note 6**). If the samples are not ground into a fine powder, the total grinding time can be increased by up to 6 min, but the shorter the grinding time, the better the DNA quality. Spin down tubes until the centrifuge reaches about 188 relative centrifugal force (RCF) to bring the ground tissue into the bottom of the tube. Longer centrifugation time makes dispersion difficult after adding the extraction buffer.

6. Using 8-channel pipettes, add 600 μL of the preheated (at 65 °C) CTAB extraction buffer per sample and tap each tube to disperse the powder. Alternatively, assemble the collection tubes in the tissue grinder and grind the samples for about 30 s to disperse/homogenize the powder/tissue with the extraction buffer. This time can be extended by up to 2 min if the tissue has not been properly ground but prolonged grinding causes DNA degradation.
7. Incubate the collection tubes at 65 °C water bath for 30–60 min with continuous gentle rocking (*set the RPM between 20 and 30; a higher rotation will result in degraded DNA*). Invert or gently tap tubes once in every 10 min to properly homogenize the tissue with the extraction buffer (*see* **Note 7**).
8. Remove plates from the water bath and allow them to cool for 5–10 min in a fume hood. Gently mix or tap the samples and centrifuge them for 10 min at 2312 RCF or for 5 min at 6795 RCF.
9. Prepare new collection tubes and label them.
10. Remove caps and discard them. Using 8-channel pipette, transfer a fixed volume of about 400–500 μl aqueous phase into the new collection tubes and add 500–600 μL chloroform:isoamyl alcohol (24:1) down the side of the tubes. Mix very gently with continuous rocking for about 5 min (or about 50 times). *DNA is very fragile, so do not use any stronger force, as it can cause degradation.*
11. Centrifuge at 2312 RCF for 10 min or at 6795 RCF for 5 min.
12. Prepare new collection tubes and label them. Transfer the upper aqueous layer and repeat the chloroform:isoamyl alcohol wash. Centrifuge as described above.
13. Transfer a fixed volume of about 300–400 μL of the upper aqueous layer into fresh collection tubes. *Do not transfer the layer containing chloroform (any trace transfer of chloroform will affect PCR). If the aqueous phase of the sample looks dirty, repeat chloroform isoamyl alcohol wash for the third time.*
14. Add 600–700 μL of cold (stored at −20 °C) isopropanol (2-propanol) and mix very gently for about 5 min (or gently

invert for about 50 times) to precipitate the nucleic acid. *Optional step: Keep tubes in at –20 °C freezer for about 1 h, remove the tubes from the freezer and leave the tubes on the bench for about 10 min, and gently invert the tubes until whitish floating material appears in the tubes.*

15. Centrifuge at 2312 RCF for 15 min or at 6795 RCF for 10 min to form a pellet at the bottom of the tube (*see* **Note 8**). Discard the supernatant.
16. Add 600 μL of 70 % ethanol; tap the tubes gently (vortex the tubes for 15–20 s) to let the pellet float for ease in washing.
17. Centrifuge at 2312 RCF for 10 min or at 6795 RCF for 5 min and discard ethanol by decantation.
18. Repeat the ethanol wash and centrifugation. Discard ethanol by decantation again.
19. Allow pellet to air-dry in a fume hood (or using the 37 °C incubator) until the ethanol evaporates completely (until the smell of ethanol disappears). This takes about 30–60 min. Don't overdry the pellet as it will be difficult to dissolve it. Any remaining ethanol smell indicates the pellet is not completely dry.
20. Dissolve the DNA pellet using one of the two options. *Option-1*: Add 100–200 μL 0.1× TE and incubate the tubes for about 45–90 min at 45 °C water bath with gentle rocking every 10 min. Make sure the pellets are completely dissolved with no suspended or floating objects. Leave the samples for 5–10 min on the bench to cool down and add 3–5 μL RNase A at a concentration of 10 μg/μL, mix by gently tapping or inverting tubes, spin down to collect to the bottom of the tubes, and incubate in a 37 °C water bath (or incubator) for about 2 h. Denature the RNase A by incubating the collection tubes at 65 °C water bath. *Option-2*: Add 100–200 μL of 0.1× TE and 3–5 μL RNase A at a concentration of 10 μg/μL, mix with gentle inversion, and leave the tubes overnight at 4 °C. On the second day, incubate the tubes for 2 h in a 37 °C water bath with gentle rocking every 10 min. Denature the RNase A by incubating the collection tubes at 65 °C water bath.

3.3 DNA Quality and Quantity Checks

1. Assemble the gel casting trays and combs as described in the user's manual.
2. Prepare 0.8 % agarose gel as follows: add 0.8 g of agarose in a 250 mL flask containing 100 mL of 1× TBE or TAE buffer and boil in a microwave oven until the agarose completely dissolves. Cool to about 60 °C, add 5 μL of gel stain (ethidium bromide or GelRed) to get a final concentration of 0.5 μg/mL of the gel volume, and mix very gently. Carefully pour the gel mix into the gel casting tray and quickly remove air bubbles, if any, using pipette tips. Leave the gel to polymerize for 30–60 min at room temperature (*see* **Note 9**).

3. In 1.5 or 2.0 mL Eppendorf tube, mix 165 μL of 6x bromophenol blue loading dye and 165 μL double-distilled water (ddH_20) and transfer 3 μL of this mix in to each of the tubes in Costar plate using multichannel pipette. Add 2 μL of the DNA solution into the loading dye mix and pipette up and down to mix. Briefly centrifuge the Costar plate to collect samples to the bottom of the tube and load all content on to the agarose gel.
4. For estimating DNA concentration by comparing fluorescent intensity with a known concentration of lambda (λ) DNA, mix 2.5 μL of 1,000 ng/μL lambda DNA, 50 μl 6× bromophenol blue loading dye, and 122.5 μL ddH_20 (*see* **Note 10**). Load 3.5 and 7.0 μL of this mix per lane to get a band with an equivalent concentration of 50 ng and 100 ng, respectively. Keep the remaining lambda DNA mix at 4 °C for use with other gels.
5. Run the gel at 70–80 V for about 45–60 min.
6. View the gel under UV light, take a picture, and save the gel image.
7. Estimate the concentration of sample DNA by comparing the fluorescent intensity of the bands with the 50 or 100 ng λ DNA.
8. On the agarose gel, check whether the RNA has been completely removed before proceeding to the absorbance-based spectrophotometer (e.g., NanoDrop) quantification. Measure DNA concentration using spectrophotometer. A ratio of 1.8 for sample absorbance at A260/A280 is generally accepted as "pure for DNA." A260/A230 ratio is a secondary measure of nucleic acid purity for the presence/absence of co-purified contaminants. A260/A230 ratio of 1.8–2.2 is acceptable but lower values (<1.7) indicate the presence of contaminants that may interfere with downstream applications. If facilities are available, measure DNA concentration using Quant-It PicoGreen reagent as described in the user's manual.
9. Based on the estimates from agarose gel, NanoDrop, or fluorescent spectrophotometer, normalize the DNA concentration into 100 or 200 ng/μL. If necessary, repeat the agarose concentration gel to check whether the concentration of all samples after normalization is comparable by loading a mix of 2 μL of the normalized DNA and 1.5 μL of the 6× bromophenol blue loading dye and 1.5 μL of ddH_20.
10. Prepare the required working dilutions for PCR or ship the stock DNA into a genotyping service provider.
11. For applications that require high DNA purity, such as GBS, carry out a restriction digest test by randomly selecting at least one sample per strip (about 12 samples per plate) and make sure the DNA samples are completely digested.

4 Notes

1. EDTA will not dissolve easily and it is important to add either a NaOH solution or pellets slowly with stirring until it completely dissolves. Adjust the pH to 8.0.
2. Cutting the leaves into small pieces increases the surface area of the leaf that is exposed to the silica, and speeds up the drying process.
3. For proper air circulation during freezing and thawing, the Ziploc bags should be perforated using paper puncher.
4. Prepare the sample layout that clearly indicates the position of each sample on the 96-well collection tubes. Use short sample identification names instead of long names.
5. The stock solutions can be prepared and stored at room temperature, but the extraction buffer should be freshly prepared just before DNA extraction.
6. Dipping the samples into liquid nitrogen helps to produce a fine powder for samples that are not completely freeze dried.
7. Before inverting the tubes, press the strip caps down so that the buffer won't splash while inverting.
8. Centrifugation while the tubes are still very cold will result in either a very small pellet or no pellet at all.
9. Air bubbles distort DNA migration during electrophoresis and must be removed before the gel polymerizes.
10. Other size markers with known concentrations, such as 1 kb DNA ladder and λ DNA-HindIII digest, can also be used for this purpose.

Acknowledgements

This protocol was optimized for the molecular breeding component of the Drought Tolerant Maize for Africa (DTMA) project, which is funded by funded by the Bill and Melinda Gates Foundation. The author would like to thank Veronica Ogugo for testing the protocol in different species.

References

1. Klimyuk VI, Carroll BJ, Thomas CM, Jones JDG (1993) Alkali treatment for rapid preparation of plant material for reliable PCR analysis. Plant J 3:493–494
2. Clancy JA, Jitkov VA, Han F, Ullrich SE (1996) Barley tissue as direct template for PCR: a practical breeding tool. Mol Breed 2: 181–183
3. Langridge U, Schwall M, Langridge P (1991) Squashes of plant tissue as substrate for PCR. Nucleic Acids Res 19:6954
4. Ikeda N, Bautista NS, Yamada T, Kamijima O, Ishii T (2001) Ultra-simple DNA extraction method for marker-assisted selection using microsatellite markers in rice. Plant Mol Biol Report 19:27–32

5. Lin JJ, Fleming R, Kuo JMBF, Saunders JA (2000) Detection of plant genes using a rapid, nonorganic DNA purification method. Biotechniques 28:346–350
6. Gao S, Martinez C, Skinner DJ, Krivanek AF, Crouch JH, Xu Y (2008) Development of a seed DNA-based genotyping system for marker-assisted selection in maize. Mol Breed 22:477–494
7. Marsal G, Baiges I, Canals JM, Zamora F, Fort F (2011) A fast, efficient method for extracting DNA from leaves, stems, and seeds of Vitis vinifera L. Am J Enol Vitic 62:376–381
8. Praveen M, Nanna RS (2009) A simple and rapid method for DNA extraction from leaves of tomato, tobacco and rape seed. J Phytol 1:388–390
9. Salim K, Qureshi MI, Kamaluddin, Tanweer A, Abdin MZ (2007) Protocol for isolation of genomic DNA from dry and fresh roots of medicinal plants suitable for RAPD and restriction digestion. Afr J Biotechnol 6:175–178
10. Lin J, Ritland K (1995) Flower petals allow simpler and better isolation of DNA for plant RAPD analyses. Plant Mol Biol Report 13: 210–213
11. Reddy GM, Coe E (1990) Isolation of nucleic acids from pollen grains. Maize Genet Coop News Lett 1990, 43
12. Chase MW, Hills HH (1991) Silica gel: an ideal material for field preservation of leaf samples for DNA studies. Taxon 40:215–220
13. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331
14. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379
15. Rogstad SH (1992) Saturated NaCl-CTAB solution as a means of field preservation of leaves for DNA analyses. Taxon 41:701–708
16. Storchova H, Hrdlickova R, Chrtek J Jr, Tetera M, Fitze D, Fehrer J (2000) An improved method of DNA isolation from plants collected in the field and conserved in saturated NaCl/CTAB solution. Taxon 49:79–84
17. Fukatsu T (1999) Acetone preservation: a practical technique for molecular analysis. Mol Ecol 8:1935–1945
18. Flournoy LE, Adams RP, Pandy RN (1996) Interim and archival preservation of plant specimens in alcohols for DNA studies. Biotechniques 20:657–660
19. Dubreuil P, Warburton M, Chastanet M, Hoisington D, Charcosset A (2006) More on the introduction of temperate maize into Europe: large-scale bulk SSR genotyping and new historical elements. Maydica 51:281–291
20. Warburton ML, Setimela P, Franco J, Cordova H, Pixley K, Banziger M et al (2010) Toward a cost-effective fingerprinting methodology to distinguish maize open-pollinated varieties. Crop Sci 50:467–477
21. Etten JV, Lopez MRF, Monterroso LGM, Samayoa KMP (2008) Genetic diversity of maize (Zea mays L. ssp. mays) in communities of the western highlands of Guatemala: geographical patterns and processes. Genet Resour Crop Evol 55:303–317
22. Eschholz TW, Peter R, Stamp P, Hund A (2008) Genetic diversity of Swiss maize (Zea mays L. ssp. mays) assessed with individuals and bulks on agarose gels. Genet Resour Crop Evol 55:971–983
23. Reif JC, Hamrit S, Heckenberger M, Schipprack W, Peter MH, Bohn M et al (2005) Genetic structure and diversity of European flint maize populations determined with SSR analyses of individuals and bulks. Theor Appl Genet 111:906–913
24. Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA 88: 9828–9832
25. Pacek P, Sajantila A, Syvanen AC (1993) Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. PCR Methods Appl 2:313–317
26. Fu YB (2003) Applications of bulking in molecular characterization of plant germplasm: a critical review. Plant Gen Res 1:161–167
27. Wang H, Qi MQ, Cutler AJ (1993) A simple method of preparing plant samples for PCR. Nucleic Acids Res 21:4153–4154
28. Williams CE, Ronald PC (1994) PCR template-DNA isolated quickly from monocot and dicot leaves without tissue homogenization. Nucleic Acids Res 22:1917–1918

29. Post RV, Post LV, Dayteg C, Nilsson M, Forster BP, Tuvesson S (2003) A high-throughput DNA extraction method for barley seed. Euphytica 130:255–260
30. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci 81:8014–8018
31. Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321–4325
32. Paterson AH, Brubaker CL, Wendel JF (1993) A rapid method for extraction of cotton (Gossypium spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Report 11:122–127

Chapter 4

DNA Extraction from Herbarium Specimens

Lenka Záveská Drábková

Abstract

With the expansion of molecular techniques, the historical collections have become widely used. Studying plant DNA using modern molecular techniques such as DNA sequencing plays an important role in understanding evolutionary relationships, identification through DNA barcoding, conservation status, and many other aspects of plant biology. Enormous herbarium collections are an important source of material especially for specimens from areas difficult to access or from taxa that are now extinct. The ability to utilize these specimens greatly enhances the research. However, the process of extracting DNA from herbarium specimens is often fraught with difficulty related to such variables as plant chemistry, drying method of the specimen, and chemical treatment of the specimen. Although many methods have been developed for extraction of DNA from herbarium specimens, the most frequently used are modified CTAB and DNeasy Plant Mini Kit protocols. Nine selected protocols in this chapter have been successfully used for high-quality DNA extraction from different kinds of plant herbarium tissues. These methods differ primarily with respect to their requirements for input material (from algae to vascular plants), type of the plant tissue (leaves with incrustations, sclerenchyma strands, mucilaginous tissues, needles, seeds), and further possible applications (PCR-based methods or microsatellites, AFLP).

Key words DNA extraction, Herbarium specimens, Difficult plant tissues, PCR, Microsatellites, AFLP

1 Introduction

Hundreds of protocols for DNA preparation from various types of tissues have been published over the last few decades. Plant and mainly herbarium plant samples DNA extraction present frequently a challenge in the first stage of each study, because of the extraction of any given taxon may require time consuming optimization of the extraction protocols. The problem of DNA extraction is crucial for further analyses of herbarium samples. The satisfactory quality of DNA is essential for the success of the whole molecular study. Most future molecular taxonomic studies will probably be partly or entirely based on DNA extracts from herbarium specimens because of the easy accessibility and richness of herbarium collections.

However, DNA isolation from dried specimens usually requires some modifications to frequently used protocols [1] because of the

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_4,

small amount of dry herbarium tissue available. The herbarium material is dried and stored on herbarium sheets in packages. If the specimens are air-dried at up to 42 °C [2], they contain a useful amount of high molecular weight DNA. Air-drying is considered to be better than the preservation of tissues in silica gel or anhydrous $CaSO_4$ [2]. In general, old air-dried material that has not been treated with chemical preservatives, high temperatures, or microwaves has the best chance of yielding useful DNA [2]. To preserve DNA well, it is necessary to dry plants as fast as possible. Extraction results depend on how the plant material is prepared, how many times the collection is disinfected, and the type of chemicals or procedures used. For instance, DNA was seriously degraded in leaves that were microwaved [3–5], boiled in water, or immersed in chemical solutions. Another important factor is the regular herbarium treatment used to keep specimens free of pests. Fumigation methods have been changed from time to time [6], making it difficult to be sure about the DNA quality. The extent of DNA degradation in herbarium specimens appears to be related to the condition of the fresh leaf rather than the year in which it was dried [7]. The DNA from herbarium specimens have been satisfactorily obtained from vascular plants about 200 years old and 100 years old from lichens.

Obtaining high-quality DNA depends on the extraction technique used. Traditional plant DNA extraction protocol by Doyle and Doyle [8] was used 6,394 times to date (e.g., cited in WoS). This method is widely used for herbarium material but sometimes with modifications [9, 10]. Several DNA isolation techniques that are useful for dry plant tissue from herbarium specimens have been described [1, 2, 11, 12]. Herbarium specimens have been frequently used during the last decade [9, 13–21]. The most convenient organ to sample is the leaf. However, also the seeds are efficient and inexpensive source for DNA [22].

There are many different protocols more or less satisfactorily used for DNA extraction from herbarium samples of different group of plants with specific types of tissues. It is not possible to present all of them in this book. I do not give detailed literature evidence for many different protocols used but a simple comparison of articles published during the last 10 years clearly shows most frequently used both modified CTAB method [8] and DNeasy Plant Mini Kit (Qiagen). I selected a few of the most used protocols during the last 10 years. This chapter should serve as a tool for projects involving DNA extraction from herbarium specimens of different plants.

1.1 Main Isolation Difficulties

One problem with extraction from herbarium specimens is a very low yield of plant material. Many people work with plant taxa that are rare or grown in inaccessible locations, making it difficult to obtain fresh plant material. The use of dried plants from historical

collections becomes essential for representative taxonomic sampling. Another problem is the quantity of the suitable tissue available. Many plants have a very limited leaf tissue volume, and the sampling for a non problematic extraction (yielding a sufficient amount of DNA) would cause serious damage to the herbarium specimen.

Undoubtedly, especially good homogenization is essential. Another crucial point is a longer and repeated precipitation. Many protocols for DNA extraction use liquid nitrogen for grinding the plant material. The homogenization of plant material is easier and faster, but the simultaneous processing of multiple samples in mortars in one laboratory table leads to the loss of DNA and contamination of the samples. When PCR products are analyzed by sequencing, the contamination is revealed. Other techniques, such as RFLP and RAPD, do not detect this type of mistake.

A good alternative is the use of bead mills. These cylinders disrupt the plant tissues in microcentrifuge tubes in the Mixer Mill (or Tissue Lyser) without risk of contamination. Insufficient disruption of starting material leads to low yield and comprised purity. Pulverizing plant material with a Mixer Mill is easier and produces DNA of more reliable quality than grinding with liquid nitrogen in a mortar [23].

2 Materials

Plants have cell walls mostly comprised of cellulose or some other complex polysaccharide or other chemical compounds or have tissues with mucilaginous substances. All these compounds may influence the quality and yield of extracted DNA even in fresh samples and herbarium samples as well. Furthermore, extraction of DNA from herbarium specimens has always been difficult due to the preservation conditions or liquids in which specimens are preserved.

2.1 Key to the Decision Among the Protocols

2.1.1 According to Plant Group

1	Vascular plants or conifers	*Method 1*
2	Mosses	*Method 1*
3	Lichens	*Method 6*
4	Mushrooms	*Method 8*
5	Algae	*Method 7*

2.1.2 According to Type of Plant Tissue

1	Plant material with leaf incrustation or containing sclerenchyma strands	*Method 1*
2	Plant material containing polysaccharides or phenolic compounds	*Method 4*
3	Plant material containing mucilaginous tissues	*Method 5*
4	Plants with needles	*Method 1*
5	Seeds	*Method 3*

2.1.3 According to Type of Next Procedures

1	DNA extraction for PCR-based methods	*Method 1*
2	DNA extraction for AFLP	*Method 2*
3	DNA extraction for microsatellites	*Method 9*

2.2 General Equipment for all Protocols

1. Manual pipettes.
2. Centrifuge for microcentrifuge tubes.
3. Vortex.
4. Thermal heating block or water bath for incubation and pre-heating of buffers (up to 65 °C).
5. Equipment for sample disruption and homogenization (Tissue Lyser or Mixer Mill) including Tissue Lyser Adapter Set and cylinders (tungsten carbide beads or ceramic cylinders).
6. 1.5–2 mL microcentrifuge tubes.
7. Disposable tips.
8. Ice.
9. Personal protection equipment (lab coat, gloves).

2.3 Method 1: DNeasy Plant Mini Kit (QIAgen) for Plants with Leaves Containing Sclerenchyma Strands

1. Buffers from DNeasy Plant Mini Kit (AP1, AP2, AP3/E, AW, AE).
2. RNase A (100 mg/mL).
3. 100 % ethanol.
4. Tungsten carbide beads.
5. QIAshredder Mini spin column.
6. DNeasy Mini spin column.

2.4 Method 2: DNeasy Plant Mini Kit (QIAgen) Modified for AFLP

The following are in addition to the items needed for Method 1:

1. Proteinase K (19.45 mg/mL).
2. Mortar and pestle.
3. Liquid nitrogen.
4. Quartz sand.

2.5 Method 3: DNA Extraction from Seeds

The following are in addition to the items needed for Method 1:

1. Tween 20.
2. Liquid nitrogen.

2.6 Method 4: The STE/CTAB Method for Microscale DNA Extraction from Polysaccharide-Rich Plants

1. STE (Sucrose–Tris–EDTA): 0.25 M sucrose, 0.03 M Tris, 0.05 M EDTA.
2. 2× CTAB (cetyltrimethylammonium bromide) extraction buffer: 100 mM Tris–HCl (pH = 8.0), 1.4 M NaCl, 20 mM EDTA (pH = 8.0), 2 % (w/v) PVPP (polyvinyl polypyrrolidone), 0.1 % (v/v) β-mercaptoethanol (include to the solution immediately prior to use).
3. Chloroform.
4. Isopropanol.
5. 80 % ethanol.
6. TE buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
7. Liquid nitrogen.

2.7 Method 5: Modified CTAB Adapted Method for Mucilaginous Tissues

1. 2× CTAB (cetyltrimethylammonium bromide) extraction buffer: 100 mM Tris–HCl (pH = 8.0), 1.4 M NaCl, 20 mM EDTA (pH = 8.0), 2 % (w/v) PVPP (polyvinyl polypyrrolidone), 0.1 % (v/v) β-mercaptoethanol (include to the solution immediately prior to use).
2. β-mercaptoethanol.
3. SEVAG: chloroform/isoamyl alcohol 24:1.
4. Ice-cold isopropanol.
5. Isopropanol.
6. TE Buffer : 10 mM Tris–HCl (pH = 8), 1 mM EDTARNase (10 mg/mL).
7. 2.5 M Sodium Acetate (NaOAc).
8. 95 % ethanol.
9. 70 % ethanol.
10. Vacuum desiccator.

2.8 Method 6: Modified CTAB Method for Fungi and Lichen-Forming Fungi

1. Extraction Buffer : 1 % (w/v) CTAB, 1 M NaC1, 100 mM Tris, 20 mM EDTA, 1 % (w/v) PVPP.
2. Precipitation Buffer: 1 % (w/v) CTAB, 50 mM Tris–HC1, 10 mM EDTA, 40 mM NaCl.
3. 1.2 M NaC1.

4. SEVAG: chloroform/isoamyl alcohol 24:1.
5. RNase A (10 mg/mL).
6. Isopropanol.
7. 70 % ethanol.
8. TE buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
9. Liquid nitrogen.

2.9 Method 7: CTAB/HNO_3 Method for Algae

1. 2 % CTAB extraction buffer: 100 mM Tris–HCl (pH = 8.0), 1.4 M NaCl, 20 mM EDTA, 0.1 % (w/v) PVPP, 0.2 % (v/v) β-mercaptoethanol (added freshly).
2. Binding Buffer: 6 M NaI, 0.1 M Na_2SO_3.
3. HNO_3 (1 mL, 5 M).
4. Washing Buffer: 20 mM Tris–HCl (pH 8), 1 mM EDTA, 0.1 mM NaCl solution, 18 mL 100 % ethanol.
5. TE buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
6. 0.45-μm membrane filter (Whatman).
7. Silica gel.

2.10 Method 8: DNA Extraction from Dried Mushrooms Using Enzymatic Digestion and Glass-Fiber Filtration (EDGF)

1. Proteinase K (20 mg/mL).
2. Lysis buffer (LB): 100 mM NaCl, 50 mM Tris–HCl (pH 8.0), 10 mM EDTA (pH 8.0), 0.5 % (w/v) SDS.
3. Binding buffer (BB): 6 M GuSCN, 20 mM EDTA (pH 8.0), 10 mM Tris–HCl (pH 6.4), 4 % (v/v) Triton X-100.
4. Binding mix (BM): 50 mL of ethanol (96 %) thoroughly mixed with 50 mL of BB.
5. Protein wash buffer (PWB): 70 mL of ethanol (96 %), 26 mL of BB.
6. Wash buffer (WB): ethanol (60 %), 50 mM NaCl, 10 mM Tris–HCl (pH 7.4), 0.5 mM EDTA (pH 8.0).
7. TE buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
8. PCR plate (e.g., Sorenson 96-well UltraAmp).
9. PALL collar.
10. Glass-fiber filtration (GF) membrane (Whatman).
11. Aluminum cover.

2.11 Method 9: NucleoSpin Plant II kit (Macherey-Nagel) Used for Microsatellites

1. Buffers from NucleoSpin Plant II kit (PL1, PL2, PC, PW1, PW2, PE).
2. RNase A (100 mg/mL).
3. 96–100 % ethanol.
4. NucleoSpin Plant II Column.

3 Methods

3.1 DNeasy Plant Mini Kit (QIAgen)

Several commercially available DNA extraction kits are very popular for high-quality extracted DNA and easy to use. For extraction of herbarium specimens, it is usually necessary to modify the manufacturer's protocol. The most valuable part of the kits is silica-gel-membrane spin columns for convenient extraction of high-quality DNA especially for PCR.

All procedures should be carried out at room temperature unless different conditions are specified (e.g., sample incubation on ice).

3.1.1 Method 1: DNeasy Plant Mini Kit (QIAgen) for Plants with Leaves Containing Sclerenchyma Strands

This extraction protocol was modified for monocots [13] and is useful for all PCR-based applications. PCR requires only minute amounts of DNA; it suggests that herbarium collections will become more valuable as sources of material for molecular studies and analyses based on PCR technique [13]. However, herbarium samples do require special extraction and reaction conditions (the most crucial points are emphasized in Subheading 4).

Mechanical disruption of plant material proved to be a limiting step when handling multiple samples in parallel [23]. Therefore, the tissue should be ground in the Mixer Mill (Tissue Lyser) with tungsten carbide beads or ceramic cylinders. This procedure is optimal for sufficient homogenization of hard leaf structure of e.g., *Juncaceae*, *Cyperaceae*, and *Pinaceae*. This extraction is presented according the QIAgen protocol with a modification for dried samples. These modifications were introduced mainly in the laboratory of Institute of Botany Copenhagen University (G. Petersen, personal communication).

1. Place 0.5–1 g of dried leaf tissue together with 3 mm tungsten carbide beads (2 or 3 pieces) into a 1.5 mL microcentrifuge tube. Place the tubes into the Tissue Lyser Adapter Set and fix into the clamps of the Tissue Lyser. Grind the samples for 1–3 min at 30 Hz (*see* **Note 1**).
2. Add 450 μL Buffer AP1 and 4 μL RNase A to a maximum of 20 mg dried disrupted plant tissue and vortex powerfully (*see* **Note 2**).
3. Incubate the mixture for 30 min at 65 °C for cell lysis. Mix 2 or 3 times during incubation by inverting tube.
4. Add 130 μL Buffer AP2 to the lysate, mix, and incubate for 5 min on ice.
5. Pipet the lysate into the QIAshredder Mini spin column placed in a 1.5 μL collection tube and centrifuge for 2 min at 20,000 × *g* (14,000 rpm) (*see* **Note 3**).
6. Transfer the flow-through fraction from the previous step into a new tube without disturbing the cell-debris pellet. Usually 450 μL of lysate is recovered (*see* **Note 4**).

7. To 450 µL lysate add 675 µL Buffer AP3/E. Reduce the amount of Buffer AP3/E to 1.5 volumes if different volume of lysate than 450 µL is obtained. Mix it by pipetting. It is important to pipet Buffer AP3/E directly onto the cleared lysate and to mix immediately.
8. Pipet 650 µL of the mixture from the previous step, including any precipitate that may have formed, into the DNeasy Mini spin column placed in a 2 mL collection tube. Centrifuge for 1 min at 6,000 × *g*, and discard the flow-through. Reuse the collection tube in the next step.
9. Repeat previous step with remaining sample. Discard flow-through and collection tube. Place the DNeasy Mini spin column into a new 2 mL collection tube, add 500 µL Buffer AW, and centrifuge for 1 min at 6,000 × *g*. Discard the flow-through and reuse the collection tube in the next step.
10. Add 500 µL Buffer AW to the DNeasy Mini spin column and centrifuge for 2 min at 20,000 × *g* to dry the membrane (*see* **Note 5**).
11. Transfer the DNeasy Mini spin column to a 1.5 mL or 2 mL microcentrifuge tube and pipet 50 µL Buffer AE directly onto the DNeasy membrane. Incubate for 10 min at room temperature (15–25 °C) and then centrifuge for 1 min at 6,000 × *g* to elute.
12. Repeat the elution step.
13. Store at –20 or –80 °C (*see* **Notes 6** and **7**).

3.1.2 Method 2: DNeasy Plant Mini Kit (QIAgen) Modified for AFLP

This extraction protocol was modified for dried vascular plants [24] and successfully used for AFLP.

1. Grind the plant tissue in a mortar with quartz sand and about 3 mL liquid nitrogen into a very fine powder.
2. Preheat a total of 500 µL AP1 buffer (60 °C), add to the sample, and grind until a mixture is completely homogeneous.
3. After grinding add 4 µL RNase and 4 µL Proteinase K.
4. Transfer the mixture to an Eppendorf tube and incubate at 60 °C for 1 h.
5. Add 150 µL AP2 buffer and follow the Qiagen extraction protocol to the final step, in which elute the DNA with 50 µL preheated (60 °C) AE buffer.
6. Store at –20 or –80 °C (*see* **Note 8**).

3.1.3 Method 3: DNA Extraction from Seeds

This extraction protocol was modified [22] for seeds of vascular plants.

1. Remove seed coats from seeds after a preliminary 10-min soak in 10 % bleach solution containing a drop of Tween 20.

2. Ground whole embryos or separated embryonic axes and cotyledons into a powder in the presence of liquid nitrogen.
3. Extract and purify DNA using the DNeasy Mini Kit according to Subheading 3.1.1 or the manufacturer's instructions.

3.2 CTAB Modified Methods

The CTAB extraction methods are based on the well-established CTAB extraction procedure [8]. However, there are some modifications for different types of plant tissues. Two protocols modified for mucilaginous tissues and fungi and lichen-forming fungi follow.

3.2.1 Method 4: The STE/CTAB Method for Microscale DNA Extraction from Polysaccharide-Rich Plants

This extraction protocol was modified [25] for polysaccharide-rich plant tissues.

1. Place 0.5–1 g of dried leaf tissue in a microcentrifuge tube with a sterile grinder. Snap freeze by suspending the tube in liquid nitrogen and grind to a fine powder (*see* **Note 9**).
2. Add 1 mL of freshly made STE to the ground plant tissue. Vortex, then centrifuge at 2,000 × *g* for 10 min. Discard supernatant and repeat STE wash.
3. Add 600 μL of CTAB solution and incubate at 60 °C for 40 min with occasional shaking.
4. Add 600 μL chloroform and shake vigorously to homogenize. Pulse centrifuge to 7,000 × *g*.
5. Remove upper aqueous layer with a wide-bore pipette tip into a new microcentrifuge tube. Add 600 μL of room-temperature isopropanol and invert gently.
6. Leave at room temperature for 1–5 min and transfer DNA pellet using a wide-bore pipette tip into a microcentrifuge tube containing 800 μL of 80 % ethanol. Wash pellet by gently inverting several times. Remove DNA pellet to a new microcentrifuge tube and repeat ethanol wash.
7. Dry the pellet and suspend in 30–60 μL of TE.

3.2.2 Method 5: Modified CTAB Adapted Method for Mucilaginous Tissues

This extraction protocol was modified [26] for plant mucilaginous tissues. The DNA obtained by this extraction can be used not only for PCR-based techniques, but also for AFLP.

1. Add 750 μL of 2× CTAB buffer and 3.0 μL of β-mercaptoethanol to Eppendorf tubes.
2. Grind 0.5–1.0 g of tissue with liquid nitrogen and sterilized sand until finely powdered.
3. Add a spatula-tip of powdered tissue to each tube and mix well.
4. Incubate in a water bath at 55–60 °C for 1–5 h, mixing every 15 min.
5. Add 700 μL of SEVAG to each tube and mix thoroughly. Centrifuge at 9,240 × *g* for 10–15 min. Transfer the aqueous phase to a new Eppendorf tube.

6. Add 0.33 volume of ice-cold isopropanol and store at −30 °C for at least 1 h.
7. Spin at 9,240–13,305 × *g* for 10 min at room temperature. Discard supernatant without disturbing the pellet. Vacuum dry. Repeat **steps 6** and 7 two to four times if the aqueous phase is viscous.
8. Resuspend pellet in 100–200 μL of TE. Add 1–2 μL of RNase. Mix well and incubate for 30 min at 37 °C.
9. Add 20 μL (0.1 vol) of NaOAc and 500 μL (2–2.5 vol) ice-cold 95 % ethanol and store at −20 °C for ≥30 min. Spin at 9,240–13,305 × *g* for 5 min. Discard supernatant.
8. Wash pellet with 1 mL of 70 % ethanol. Do not disturb the pellet. Spin at 9,204 × *g* for 4 min and pour off ethanol. Vacuum-dry pellet. Do not overdry (*see* **Note 10**).
10. Resuspend pellet in 100–200 μL of TE. Store at −20 °C.

3.2.3 Method 6: Modified CTAB Method for Fungi and Lichen-Forming Fungi

This extraction protocol was modified [7] adapted for fungi and lichen-forming fungi [27].

The best results were obtained from liquid nitrogen frozen samples [7]. Samples can be disrupted without liquid nitrogen by grinding the material in powdered glass. DNA extracted in this way gave good amplifications, although the total DNA yield was reduced compared to liquid nitrogen preparations. A mortar and pestle can also be used without additional abrasives, although this did not prove practical for either large numbers or small amounts of material.

1. Put 3–100 mg of material into 1.5 mL tubes and place in a container with liquid nitrogen for 5–10 min. Then remove from the container and place in an insulated rack. Add liquid nitrogen to the tube and grind the material with a sterile pre-cooled sharp glass bar. Sterilize glass bars in flame immediately prior to use.
2. Add 0.5 mL of pre-warmed extraction buffer to the ground material. Add PVPP to the buffer immediately prior to use. Mix the tubes by inverting several times and then heated in a water bath for 30 min at 70 °C before adding one volume of SEVAG. Mix by inverting the tube and centrifuged for 5 min at 10,000 × *g* at room temperature.
3. Collect the upper aqueous phase in a new tube and discard the slurry and lower layers (*see* **Note 11**).
4. Add two volume of precipitation buffer to the supernatant and mix well by inversion for 2 min.

5. Centrifuge the mixture for 15 min at 13,000 × *g* at room temperature and collect the pellet.
6. Resuspend the pellet in 350 µL of 1.2 M NaCl and add one volume of SEVAG. Mix vigorously and centrifuge for 5 min at 10,000 × *g* at room temperature.
7. For RNA-free DNA add 2 µL of RNase A to the sample and incubate at 37 °C for 30 min.
8. Remove the upper phase to a new tube and add 0.6 vol of isopropanol. Mix by inversion and place the tube at −20 °C for 15 min.
9. Collect the final pellet by centrifugation for 20 min at 13,000 × *g* at 4 °C. Wash the final pellet with 1 mL of 70 % ethanol and recollect by centrifugation for 3 min at 13,000 × *g* at 4 °C. Then drain the pellet and dry at 50 °C prior to resuspension in either PCR grade water or TE Buffer.

3.2.4 Method 7: CTAB/HNO_3 Method for Algae

This extraction protocol was modified [28] for brown macroalgae.

1. Grind the plant tissue and add 2 % CTAB buffer.
2. Clarify the binding buffer by filtration through a 0.45-µm membrane filter.
3. Prepare silica fines by placing 20–30 g of silica gel into c. 500 mL of milliQ-filtered deionized water and stirring for c. 1 h. After stirring, allow the silica to settle for c. 15 min.
4. Transfer the supernatant to 50 mL plastic tubes and centrifuge for 5 min at 1,250 × *g*.
5. Remove most of the supernatant from each 50 mL tube; leave only a small amount for resuspension of the pelleted particles and subsequent consolidation into one tube.
6. Transfer aliquots (c. 1 mL) of the consolidated particles to 2 mL plastic tubes.
7. Add HNO_3 to each 2 mL tube prior to heating at 95–100 °C for 30 min in a vented hood.
8. After cooling, centrifuge the tubes at 13,000 × *g* for 1 min and discard the supernatant.
9. Wash the silica pellet by resuspending in c. 2 mL milliQ-filtered deionized water and centrifuge for 1 min at 13,000 × *g*. Discard the supernatant.
10. Repeat the washing step five times prior to a final resuspension with an equal volume of milliQ-filtered deionized water.
11. Add 6.8 mL of the washing buffer.
12. Elute in the TE buffer.

3.3 Method 8: DNA Extraction from Dried Mushrooms Using Enzymatic Digestion and Glass-Fiber Filtration (EDGF)

This extraction protocol was described [29] for animal tissues and modified [30] for dried mushrooms.

1. Add a small amount of sample (1–2 mm^3) to each well of a 96-well PCR plate. Instruments should be flame sterilized between samples to avoid cross contamination. Last well can be left blank and used as a negative control.
2. Mix 5 mL of LB and 0.5 mL of Proteinase K (20 mg/mL) in a sterile container and dispense 100 μL to each well. Cover each row with caps and incubate at 56 °C overnight (8–16 h) to allow digestion.
3. Centrifuge at 1,000 × *g* for 1 min.
4. Add 100 μL of BM to each sample. Mix by pipetting up and down few times.
5. Remove cap strips/cover and transfer the lysate (about 150 μL) from the wells of microplate into the wells of the PALL glass-fiber filtration (GF) plate placed on top of a square-well block. Seal the plate with adhesive cover.
6. Centrifuge at 1,500 × *g* for 10 min to bind DNA to the GF membrane.
7. Add 250 μL of PWB to each well of the GF plate. Seal with a new adhesive cover and centrifuge at 1,500 × *g* for 5 min. Discard the flow-through.
8. Add 300 μL of WB to each well of the GF plate. Seal with a new cover and centrifuge at 1,500 × *g* for 10 min.
9. To avoid incomplete WB removal, open the cover to relieve the vacuum that may have formed in the wells, seal the plate again, and centrifuge the plates again at 1,500 × *g* for 5 min. Discard the flow-through.
10. Repeat **steps 8** and **9**.
11. Remove the cover. Place the GF plate on a clean square-well block and incubate at 56 °C for 30 min to evaporate residual ethanol.
12. Position a PALL collar on a collection plate and place plate and collar on top of a clean square-well block. Place GF PALL plate with DNA bound to the membrane on top of a PCR plate. Dispense 50 μL of 0.1× TE buffer or water, pre-warmed at 56 °C, directly onto the membrane of each well of GF plate and incubate at room temperature for few minutes and then seal plate.
13. Centrifuge at 1,500 × *g* for 10 min to collect the eluted DNA. Remove the GF plate and discard it.
14. Cover DNA plate with aluminum cover. Keep at 4 °C for temporary storage or at −20 °C for long-term storage.

3.4 Method 9: NucleoSpin Plant II Kit (Macherey-Nagel) Used for Microsatellites

This extraction protocol was used [31] prior to microsatellite data analysis. The standard protocol uses Lysis Buffer PL1, which is based on the established CTAB extraction procedure [8]. Alternatively, the SDS-based Lysis Buffer PL2 is provided which requires subsequent protein precipitation by potassium acetate (Precipitation Buffer PL3).

1. Homogenize up to 20 mg dry weight plant (*see* **Note 12**).
2. Transfer the resulting powder to a new tube and add 400 μL Buffer PL1. Vortex the mixture thoroughly (*see* **Note 13**). Alternatively, transfer the resulting powder to a new tube and add 300 μL Buffer PL2. Vortex the mixture thoroughly. If the sample cannot be resuspended easily, additional Buffer PL2 can be added. Note that the volumes of RNase A and Buffer PC have to be increased proportionally (*see* **Note 14**).
3. Add 10 μL RNase A solution and mix sample thoroughly. Note that the volumes of RNase A have to be increased proportionally. Incubate the suspension for 10 min at 65 °C. Alternatively, add 75 μL Buffer PL3, mix thoroughly, and incubate for 5 min on ice to precipitate SDS completely (*see* **Note 15).**
4. Place a NucleoSpin Filter into a new collection tube (2 mL) and load the lysate onto the column. Centrifuge for 2–5 min at 11,000 × *g*, collect the clear flow-through, and discard the NucleoSpin Filter. If not all liquid has passed the filter, repeat the centrifugation step. If a pellet is visible in the flow-through, transfer the clear supernatant to a new 1.5 mL microcentrifuge tube.
5. Add 450 μL Buffer PC and mix thoroughly by pipetting up and down (5 times) or by vortexing.
6. Place a NucleoSpin Plant II Column into a new collection tube (2 mL) and load a maximum of 700 μL of the sample (*see* **Note 16**). Centrifuge for 1 min at 11,000 × *g* and discard the flow-through.
7. Preheat Buffer PE to 65 °C.
8. Add 400 μL Buffer PW1 to the NucleoSpin Plant II Column. Centrifuge for 1 min at 11,000 × *g* and discard flow-through.
9. Add 700 μL Buffer PW2 to the NucleoSpin Plant II Column. Centrifuge for 1 min at 11,000 × *g* and discard flow-through.
10. Add another 200 μL Buffer PW2 to the NucleoSpin Plant II Column. Centrifuge for 2 min at 11,000 × *g* in order to remove wash buffer and dry the silica membrane completely.
11. Place the NucleoSpin Plant II Column into a new 1.5 mL microcentrifuge tube. Pipette 50 μL Buffer PE (65 °C) onto the membrane. Incubate the NucleoSpin Plant II Column for 5 min at 65 °C. Centrifuge for 1 min at 11,000 × *g* to elute the DNA. Repeat this step with another 50 μL Buffer PE (65 °C) and elute into the same tube.

4 Notes

1. Proper grinding of plant samples with a Tissue Lyser or Mixer Mill is the crucial step. The plant tissue should be ground to a fine powder after the disruption. However, for some plants one disruption step may not be sufficient. In that case repeat the disruption for 1 min at 30 Hz till the sample is not thoroughly and equally homogenized.
2. It is necessary to remove tissue clumps, because tissue clumps will not lyse properly and therefore decrease yield of DNA. If the small amount of sample is expected, use longer precipitation or repeat it.
3. It may be necessary to cut the end of the pipet tip to apply the lysate to the QIAshredder Mini spin column. The QIAshredder Mini spin column removes most precipitates and cell debris, but a small amount will pass through and form a pellet in the collection tube.
4. It is crucial not to disturb the pellet. In case you do that, repeat **step 5**. For herbarium specimens usually less lysate is recovered. In this case, determine the volume for the next step.
5. It is important to dry the membrane of the DNeasy Mini spin column since residual ethanol may interfere with subsequent reactions. Discard flow-through and collection tube.
6. Preferably short-term storage in TE (or AE buffer) at –25 °C, for long-term storage use –80 °C.
7. The exclusion of samples based on visualization of total DNA on agarose gel alone is gratuitous. This statement is also valid for other techniques as AFLP (*see* below). For PCR the best results require short length of products (optimum of 300–350 to 500 bp). A higher number of PCR cycles are recommended.
8. Use short AFLP fragments, up to 300 bp (depending on the quality/quantity of DNA and chromatograms). To compensate for using only part of the chromatogram, it may be necessary to increase the number of primer combinations in order to obtain a sufficient number of polymorphic fragments. Even samples for which DNA appearance on the agarose gel showed small amount and/or low quality may in some cases work well for AFLP.
9. To obtain a fine powder is the most crucial step.
10. If the pellet is disturbed, centrifuge again.
11. Do not disturb the lower layers and the pellet. If do so, centrifuge again.
12. Proceed with cell lysis using Buffer PL1 or alternatively with Buffer PL2 or Buffer PL3 (*see* below). Choose the buffer

according to plant tissue used: Buffer PL1 is based on the established CTAB procedure. Additionally, the SDS-based Lysis Buffer PL2 is provided by manufacturer which requires subsequent protein precipitation by potassium acetate (Precipitation Buffer PL3).

13. If the sample cannot be resuspended easily additional Buffer PL1 can be added.
14. The standard protocol uses Lysis Buffer PL1, which is based on the established CTAB procedure. For some plant species Lysis Buffers PL1 and PL2 can be used with similar results.
15. The SDS-based Lysis Buffer PL2 is provided which requires subsequent protein precipitation by potassium acetate (Precipitation Buffer PL3).
16. The maximum loading capacity of the NucleoSpin Plant II Column is 700 μL. For higher sample volumes, repeat the loading step.

Acknowledgement

The study was supported by GAČR 206/07/P147, GAČR P506/11/0774, and Institutional Research Plan AV0Z60050516.

References

1. Rogers SO (1994) Phylogenetic and taxonomic information from herbarium and mumified DNA. In: Adams RP et al (eds) Conservation of plant genes II.: utilization of ancient and modern DNA. Miss Bot Gard, Monogr, Missouri Botanical Garden Press, St. Louis, vol 48
2. Tailor JW, Swann EC (1994) Dried samples: soft tissues, DNA from herbarium specimens. In: Herrmann B, Hummel S (eds) Ancient DNA. Springer, Verlag
3. Hall DW (1981) Microwave: a method to control herbarium insects. Taxon 30:818–819
4. Hill SR (1983) Microwave and the herbarium specimen: potential dangers. Taxon 32:614–615
5. Bacci M, Checcuccii A, Checcuccii G, Palandek MR (1983) Microwave drying of herbarium specimens. Taxon 34:649–653
6. Metsger DA, Byers SC (1999) Managing the modern herbarium, an interdisciplinary approach. Society for the preservation of natural history collections, Washington DC, p 384
7. Rogers SO, Bendich AJ (1994) Extraction of total cellular DNA from plants, algae, and fungi. In: Gelvin SB, Schilperoort RA (eds) Plant Molecular Biology Manual, 2nd ed., Kluwer Academic Publishers, Dordrecht. The Netherlands D1:1–8
8. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15
9. Ribeiro RA, Lovato MB (2007) Comparative analysis of different DNA extraction protocols in fresh and herbarium specimens of the genus *Dalbergia*. Genet Mol Res 6:173–187
10. Agostini G, Lüdtke R, Echeverrigaray S, de Souz-Chies TT (2011) Genomic DNA extraction from herbarium samples of *Cunila* D. Royen ex L. (Lamiaceae) and *Polygala* L. (Polygalaceae). Conserv Genet Resour 3:37–39
11. Wittzell H (1999) Chloroplast DNA variation and reticulate evolution in sexual and apomictic sections of dandelions. Mol Ecol 8:2023–2035
12. Ristaino JB, Groves CT, Parra GR (2001) PCR amplification of the Irish potato famine pathogen from historic specimens. Nature 411(6838):695–697
13. Drábková L, Kirschner J, Vlček Č (2002) Historical herbarium specimens in molecular taxonomy of the Juncaceae: a comparison of DNA extraction and amplification protocols. Plant Mol Biol Rep 20(2):161–175
14. De Castro O, Menale B (2004) PCR amplification of Michele Tenore's historical specimens and

facility to utilize an alternative approach to resolve taxonomic problems. Taxon 53:147–151

15. Jankowiak K, Buczkowska K, Szweykowska-Kulinska Z (2005) Successful extraction of DNA from 100-year-old herbarium specimens of the liverwort *Bazzania trilobata*. Taxon 54: 335–336
16. Asif MJ, Cannon CH (2005) DNA extraction from processed wood: a case study for identification of an endangered timber species (*Gonystylus bancanus*). Plant Mol Biol Rep 23(2):185–192
17. Erkens RHJ, Cross H, Maas JW, Hoenselaar K, Chatrou LW (2008) Assessment of age and greenness of herbarium specimens as predictors for successful extraction and amplification of DNA. Blumea 53:407–428
18. Lister DL, Bower MA, Howe CJ, Jones MK (2008) Extraction and amplification of nuclear DNA from herbarium specimens of emmer wheat: a method for assessing DNA preservation by maximum amplicon length recovery. Taxon 57:254–258
19. Andreasen K, Manktelow M, Razafimandimbison SG (2009) Successful DNA amplification of a more than 200-year-old herbarium specimen: recovering genetic material from the Linnaean era. Taxon 58:959–962
20. Poczai P, Taller J, Szabo I (2009) Molecular genetic study of a historical Solanum (Solanaceae) herbarium specimen collected by Paulus Kitaibel in the 18th century. Acta Bot Hung 51:337–346
21. Sohrabi M, Myllis L, Soili S (2010) Successful DNA sequencing of a 75 year-old herbarium specimen of *Aspicilia aschabadensis* (J. Steiner) Mereschk. Lichenologist 42:626–628
22. Walters C, Reilley AA, Reeves PA, Baszczak J, Richards CM (2006) The utility of aged seeds in DNA banks. Seed Sci Res 16:169–178
23. Csaikl UM, Bastion H, Brettschneider R, Gauch S, Metr A, Schauerte M, Schulz F, Sperisen C, Vornam B, Ziegenhagen B (1998) Comparative analysis of different DNA extraction protocols: a fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. Plant Mol Biol Rep 16:69–86
24. Lambertini C, Frydenberg J, Gustafsson MHG, Brix H (2008) Herbarium specimens as a source of DNA for AFLP fingerprinting of *Phragmites* (Poaceae): possibilities and limitations. Pl Syst Evol 272:224–231
25. Shepherd LD, McLay TGB (2011) Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. J Plant Res 124:311–314
26. Cota-Sánchez JH, Remarchuk K, Ubayasena K (2006) Ready-to-use DNA extracted with a CTAB method adapted for herbarium specimens and mucilaginous plant tissue. Plant Mol Biol Rep 24:161–167
27. Cubero OF, Crespo A, Fatehi J, Bridge PD (1999) DNA extraction and PCR amplification method suitable for fresh, herbarium-stored, lichenized, and other fungi. Pl Syst Evol 216: 243–249
28. Ivanova NV, Dewaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Mol Ecol Notes 6:998–1002
29. Dentinger BTM, Margaritescu S, Moncalvo J-M (2010) Rapid and reliable high-throughput methods of DNA extraction for use in barcoding and molecular systematics of mushrooms. Mol Ecol 10:628–633
30. Hoarau G, Coyer JA, Stam TW, Olsen JL (2007) A fast and inexpensive DNA extraction/purification protocol for brown macroalgae. Mol Ecol Notes 7:191–193
31. Malenica N, Šimon S, Besendorfer V, Malecić E, Kontić JK, Pejić I (2011) Whole genome amplification and microsatellite genotyping of herbarium DNA revealed the identity of an ancient grapevine cultivar. Naturwissenschaften 98:763–772

Chapter 5

Analysis of Variation in Chloroplast DNA Sequences

Berthold Heinze, Agnieszka Koziel-Monte, and Daniela Jahn

Abstract

This chapter introduces and reviews methods for analyzing variation in chloroplast DNA, mainly by polymerase chain reaction (PCR) and subsequent revelation of polymorphisms. Sources for chloroplast primers are discussed, as well as methods such as Sanger sequencing, PCR followed by restriction fragment length polymorphism (RFLP), gel electrophoresis, fragment analysis on automated DNA sequencers, denaturing high-performance liquid chromatography (dHPLC), and next-generation sequencing (NGS). A special section deals with peculiarities of chloroplast DNA variation, such as tandem repeats and mini- and microsatellites.

Key words Chloroplast DNA, PCR, PCR-RFLP, Sanger sequencing, dHPLC, Next-generation sequencing, Gel electrophoresis, Tandem repeats

1 Introduction

1.1 Chloroplast DNA Analysis

Chloroplast DNA sequences have been used in molecular plant taxonomy from an early stage onwards [1]. The topic has developed along two different lines: one "camp" of researchers has been interested in the use of chloroplast gene sequences for developing "deep" phylogenies that often involved tens of taxa, while the other "camp" has explored chloroplast DNA sequences for distinguishing lines within a species or between closely related species, subspecies, varieties, and the like. From an early point onwards, it became clear that different sets of chloroplast sequences must be used for these diverging uses.

Variation in these sequences was first analyzed with restriction enzymes (e.g., [2]). Purified chloroplast DNA preparations were digested and run on electrophoresis gels. The availability of a cloned set of probes representing the whole chloroplast genome of *Petunia* [3] made possible the use of Southern blotting. With this technique, the chloroplast DNA fragments on the gels could be identified and ordered (i.e., mapped; e.g., [4]). Around this time, the first complete DNA sequence of a chloroplast genome, the one

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_5,

of tobacco, *Nicotiana tabacum*, was reported [5, 6], a major breakthrough in chloroplast DNA research. This species still serves as a model until today [7–11]. In addition to the reports cited, Bock [12] and Ravi et al. [13] provide current reviews of chloroplast DNA research.

With this tobacco and a few other sequences available, it became possible to design polymerase chain reaction (PCR) primers to amplify regions of the chloroplast genome that would allow sequencing of chloroplast genes [1]. Taberlet et al. [14] introduced the concept of using PCR to amplify highly variable intron and spacer regions. The "universal chloroplast primers" (ucp) which they suggested are still in wide use today, and Google Scholar lists 3209 citations of their article (November 2013). Both "camps" mentioned above have subsequently contributed numerous primer sequences to the scientific literature.

Methods of analysis for the chloroplast gene sequences in arrays of samples have thus gone from restriction enzymes and Southern blots with labeled probes to the use of polymerase chain reaction (PCR) based methods. These in turn can be categorized into direct sequencing methods and indirect methods like PCR followed by restriction fragment length polymorphism (RFLP) analysis or other methods that detect sequence polymorphisms. In our lab at BFW, we have been working on denaturing high-performance liquid chromatography (dHPLC) as one such method, for which a protocol is presented here, and we are currently experimenting with TILLING (Targeting Induced Local Lesions IN Genomes, [15]). TILLING is, more generally speaking, a method for heteroduplex DNA polymorphism detection just like dHPLC and like others based, e.g., on DNA mobility in gels (e.g., [16]). Specifically, in TILLING, heteroduplexes are generated by either amplification from two polymorphic chromosomes or from a mixture of DNAs from different individuals. The latter can be adopted for chloroplast DNA typing, and we are currently experimenting with the sensitivity of this system (Jahn, Till and Heinze, unpublished).

Until a few years ago, whole books were dedicated to the subject of identifying variation among DNA molecules and their carrier organisms [17]. Nowadays, fewer such methods remain in routine use, and DNA sequencing of (mostly) PCR products, either in-house or by dedicated genome centers or commercial service providers, has become much more widespread. We describe below our routine protocols for analyzing PCR products of chloroplasts. Although methods for sequencing are widely published (e.g., [18, 19]), the availability of newer machines and dedicated commercial kits justifies writing such a method paper.

1.2 PCR Primers from the Literature

The advantage in chloroplast PCR is the conserved nature of many sequence elements and their position adjacent to variable sequences. This means that primers designed from one species will often be

successful in related plant species, sometimes even in distantly related ones. Another approach is to compare sequences from several species and identifying conserved sequence stretches for primer design. PCR products may reveal variable sequences (polymorphisms) for studying relationships between or even within taxa.

These approaches were first followed for particular genes: first *rbcL*, then *atpB/E*, *psbA,* and others [1]. Primers published in these early years include those for the *rpoC1* and *rpoC2* genes [20], for *rpoB* [21], and for *matK* [22]. A collection is presented by Graham and Olmstead [23], who discuss the merits of 17 individual genes (or highly conserved introns) for resolving phylogenies at a higher phylogenetic level.

Taberlet et al. [14] and their idea of amplifying highly variable introns and spacers opened up this field for studying closely related species or even individuals within species. While their original primers are still used by many (as the regions they are amplifying are exceptionally variable), other authors have since suggested many more primer sequences designed with the same rationale of either amplifying spacers and introns or genes (or gene fragments). Some of the larger collections include [24–30] and finally [31], where an online database is presented. Collectively, it is now possible with these primers to amplify consecutive (overlapping) stretches of chloroplast DNA including gene, intron, and spacer sequences. Maintaining the online database [31] is a challenge in the face of constantly appearing new articles that propose new primer sequences. Some of the more recent examples introducing primer collections include, among others, [32–40]. We are therefore in the process of improving the Internet database ([31]; http://bfw.ac.at/rz/bfwcms2.web?dok=4977) into a "web 2.0" version, where users can submit suggestions for new primers, which are then checked automatically and manually before addition to the database (Heinze, Edlinger and Jahn, in preparation).

1.3 PCR Primers for Species Identification

Nakamura et al. [41] apparently were among the first to propose a specific region (*rpl16–rpl14*) to be sequenced in order to establish species identity for plants from a chloroplast DNA sequence. Brunner et al. [42] have made use of the Taberlet et al. [14] universal chloroplast primers for exactly this purpose, and thus identified below-ground roots at the species level. Borsch et al. [43] have further evaluated the primers for "deep phylogeny" uses. The Consortium for the Barcode of Life [44] has since approximately 2005 ([45]; *see* Fig. 1 in [46]) revived this general idea (apparently without taking notice of Nakamura et al. [41]). In a combination of database and laboratory research, the consortium and others have considered a number of candidates (e.g., [47–53], and, with a somewhat deviating rationale, Borsch et al. [33]). Taberlet et al. [54] have reviewed the use of their originally proposed primers for species identification (in the sense of DNA barcoding). The consortium finally adopted the combination of *matK* and *rbcL*

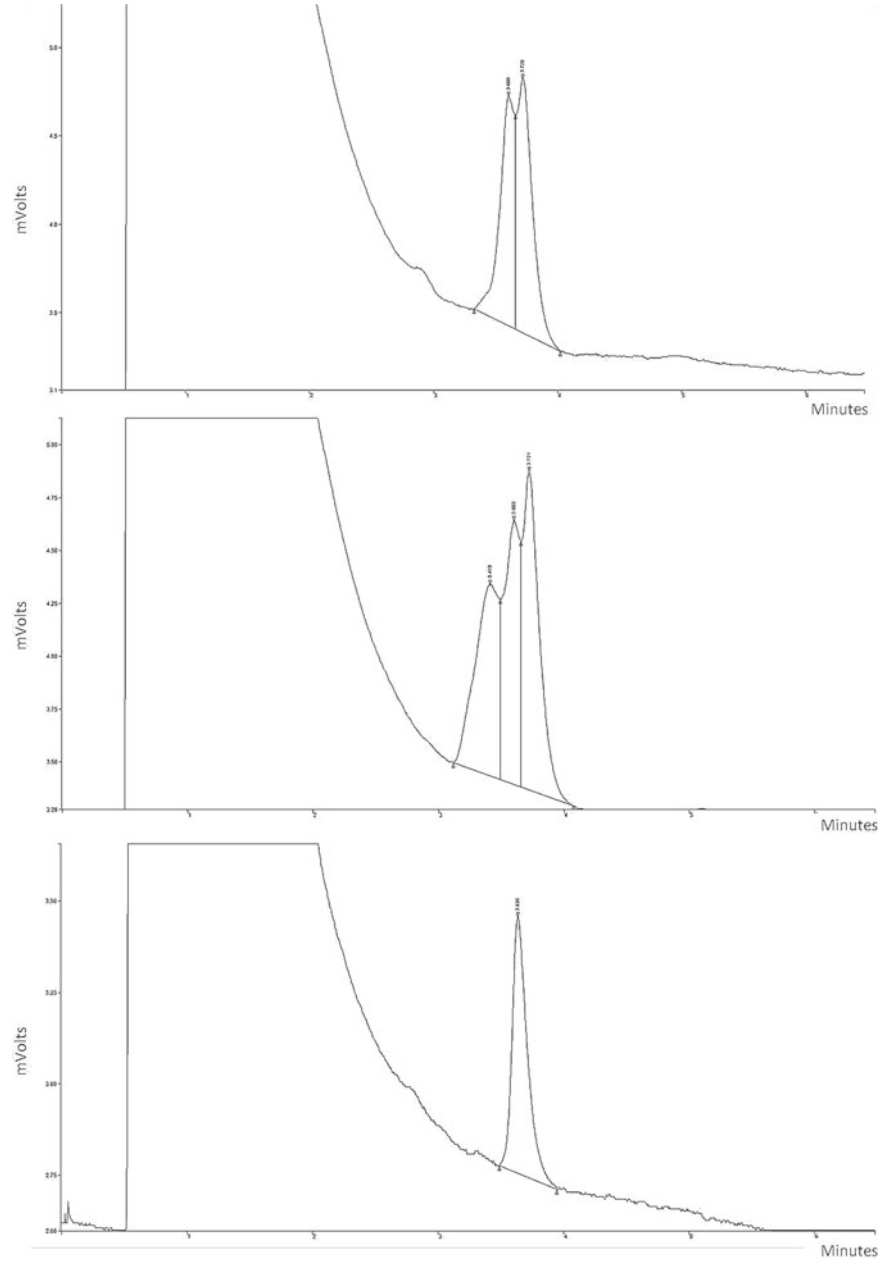


Fig. 1 dHPLC chromatogrammes. The *upper panel* shows a double peak, the *central panel* a triple peak, and the *lower panel* a single peak with a shoulder

as the universal plant barcode [44], a step that is necessary in order to concentrate database development and sequencing efforts. However, skeptical voices have echoed throughout the literature, questioning whether this choice, or indeed the idea of a universal plant barcode, is feasible [55–58].

On the other hand, certain chloroplast DNA sequences, even in otherwise conserved regions, often diverge so much between monocots and dicots or between angiosperms and gymnosperms (or conifers), so that different primer sets have to be used for either group [59, 60].

1.4 Analyzing Variation in Chloroplast DNA PCR Products

Sequencing products of PCR with chloroplast primers for different individuals one by one is still the most common method of obtaining sequence information. Dideoxynucleotide sequencing (the Sanger method) with fluorescent dye terminator nucleotides is now standard in many molecular biology laboratories.

Size differences in PCR amplification products of chloroplast DNA may be visible directly on electrophoresis gels (especially after polyacrylamide electrophoresis), but usually, the size differences are small and not readily detectable. Testing various restriction enzymes for their ability to cut a given fragment is a simple and straightforward method to assess the variability of the amplified DNA (called PCR-RFLP—polymerase chain reaction—restriction fragment length polymorphism), although it does not give ultimate resolution. For the ease and speed of obtaining results, however, the method is still in wide use.

Because of the time-consuming procedure for most polyacrylamide gel systems, we start many chloroplast DNA surveys with "simple" agarose gels. However, our gels are optimized for the purpose, i.e., medium throughput and high resolution. Consequently, our system is a low-cost option that may be attractive for student labs, smaller institutions, or laboratories in remote locations.

As an alternative, we have used denaturing high-performance liquid chromatography (dHPLC) of chloroplast DNA fragments for detecting polymorphisms. Like with many other HPLC systems, the principle of dHPLC is ion-paired reverse-phased chromatography with a stationary phase of a hydrophobic support media (divinylbenzene, DVB) and a mobile phase composed of an aqueous buffer with triethylamine acetate (TEAA), to which an organic solvent is added during the run, thus forming a gradient. Due to the fact that TEAA has counter ions of the opposite charge, it pairs with the negatively charged phosphate group of the DNA (resulting in the "ion pair"), so that the DNA is coated in a hydrophobic layer and is absorbed to the column matrix. Given the buffer gradient and a set column temperature, the separating principle in this application is whether DNA duplex molecules are perfect or not: polymorphisms like single nucleotide polymorphisms or

insertions/deletions will cause heteroduplex (mixed) DNA double strands to form, if both types are present during heating and cooling phases. The heteroduplex molecules will form small bulges that cause these DNA molecules to dissociate earlier from the column matrix than perfect double strands. The DNA amount in the effluent is monitored by an ultraviolet-visible (UVis) detector, and the resulting chromatogram can be examined for characteristic "double peaks" or more generally aberrant peak shapes and patterns that form in the presence of heteroduplexes [61]. Due to the fact that polymorphism analysis with the dHPLC system is only possible relative to a baseline (standard) DNA sequence, a (homozygous) sample must be defined as standard for analysis, and its DNA must be mixed individually with the DNAs of each of the samples before the dHPLC run, in order to form heteroduplexes. Chloroplast DNA of a single plant is usually "homozygous" in a sense that only one type of chloroplast DNA is present.

1.5 Special Sequence Features of Chloroplast DNA

It was noted early [62] that in the first and very "successful" regions proposed for amplifying intergenic spacers, long stretches of mononucleotide repeats (As or Ts) would make sequencing highly challenging (but *see* [63] who may have overlooked this note). Such stretches were later identified in many locations in the genome, and it was found that some of them are conserved. Powell et al. [64] therefore proposed to utilize these regions as polymorphic markers (called "chloroplast microsatellites" or "simple sequence repeats", SSRs). Weising and Gardner [65] presented a set of conserved primers that could be used in many angiosperms. Similar sets were later proposed specifically for conifers [66] and for monocotyledons [67, 68]; and [34] enlarged the sets considerably. As it is necessary for these regions to reliably distinguish length variations of single base pairs despite considerable "stuttering" (replication slippage of the PCR enzyme), capillary sequencers or long polyacrylamide gels have to be used for fragment separation.

"Minisatellites" in the chloroplast are a similar phenomenon; their repeat unit consists of more nucleotides (e.g., around 15 bp; [69–71]). This repeat type is already detectable on an agarose gel, especially if the PCR fragments are digested or rather small [72]. It makes a difference to phylogenetic or phylogeographic networks whether these polymorphisms are considered in an analysis, or not [73].

1.6 Next-Generation Sequencing

More recent developments make use of new "mass sequencing" or "next-generation sequencing" approaches. While there are examples coming up in the literature that have successfully used these methods for establishing whole chloroplast genome sequences in a number of species (e.g., [74, 75]), and some of them have even moved into taxonomic applications [73], this field is advancing at a very high speed, and protocols developed today may be outdated tomorrow. Therefore, we only add a few considerations concerning next-generation sequencing (*see* Subheadings 2.9 and 3.9).

2 Materials

2.1 Chloroplast PCR Primers from Literature and from the Web

The Internet provides ample opportunities for searching species-specific primers from the primary literature.

1. One useful starting point is GenBank (www.ncbi.nlm.nih.gov/genbank/), where sequences of a taxon or group of taxa of interest often include a host of chloroplast sequences, and the GenBank entries of these often include information of any journal publications (where the specific primer sequences should be mentioned).
2. Based on primer collections published in the literature, a database was collated a few years ago [31] and is present on the Internet at: http://bfw.ac.at/rz/bfwcms2.web?dok=4977. The sets from the literature were supplemented with newly designed primers, in order to achieve coverage of most parts of "standard" chloroplast genomes with moderate density. The database is searchable, and results can be formatted as a table. Mark, copy, and paste allows one to extract parts or complete tables from the database.

2.2 PCR Amplification of Chloroplast DNA Fragments

1. Plant DNA preparations (*see* **Note 1** on DNA isolation) of good quality (*see* **Note 2**) in solution at a concentration of 10–50 ng/μL (*see* **Note 3**).
2. DNA polymerase: Most commercial thermostable DNA polymerases can be used in chloroplast DNA PCR (*see* **Notes 4** and **5**). They come with their specific buffers (either including $MgCl_2$ or with an extra vial with $MgCl_2$).
3. Deoxynucleotide triphosphates (dNTPs): Standard molecular biology grade deoxynucleotide triphosphate (dNTP) solutions are also available commercially.
4. PCR machine: The variety of "PCR machines" (thermocyclers) is ever increasing—any model for standard PCR can be used (*see* **Note 6**).

2.3 Agarose Gel Electrophoresis

1. Agarose: high resolution agarose and standard (high gel strength) agarose (molecular biology grade each).
2. 10× Tris (tris(hydroxymethyl)aminomethane)–borate (TBE) buffer: 10,8 g Tris base, 5,5 g boric acid, 0,7 g ethylenediaminetetraacetic acid (EDTA)-Na_2 in100 mL H_2O. This is diluted to 0.5× for use (final concentrations: 89 mM Tris base, 89 mM boric acid, 2 mM EDTA–Na_2).
3. Gel apparatus including tank with lid, gel trays, and combs is a standard part of laboratories.
4. Power supply for delivering a constant flow of electric current to be set in the range of 80–200 V; it is also usually a standard part of laboratory equipment (for safety, *see* **Note 7**).
5. Ethidium bromide concentrate (10 mg/μL) and staining solution (0.5 mg/mL)—for safety, *see* **Note 8**, and also *see* **Note 9**.

2.4 PCR-RFLP

1. Use material described in Subheading 2.2 for PCR.
2. Restriction enzymes: Preparations of many manufacturers can be used, as most commercial restriction enzymes work directly with PCR products. Enzymes with four or six nucleotides at their recognition sites are preferred, because the amplified fragments are comparatively small (*see* **Note 10**).
3. The restriction enzymes are delivered with a vial of concentrated buffer, usually 5× or 10× concentrated.
4. Materials for agarose (Subheading 2.3) or polyacrylamide (Subheading 2.6) electrophoresis.

2.5 Sanger Sequencing

Dideoxynucleotide sequencing (the Sanger method) with fluorescent dye terminator nucleotides is now standard in many molecular biology laboratories.

1. PCR products (as obtained in Subheading 3.2).
2. PCR or sequencing primers (as described in Subheading 2.1).
3. Kits can be obtained from many manufacturers. The kits contain all of the components necessary for the sequencing reaction; only primers and template DNA have to be provided by the user. If sequencing reactions are loaded onto a certain capillary electrophoresis machine that requires specific fluorescent dyes, the choice of sequencing kit is determined by this specificity of the "DNA sequencer" machine. In our example, we will use the kit from Beckman Coulter (DTCS kit) as we will go on to analyze the reaction products on a Beckman Coulter CEQ 8000 analyzer, but kits and machines of other manufacturers are equally useful. Polyacrylamide sequencing gels may still be used for sequencing purposes (*see* **Note 11**).
4. Capillary electrophoresis automatic sequencer, e.g., Beckman Coulter CEQ 8000, or alternatively polyacrylamide sequencing gel apparatus (*see* Subheadings 2.6 and 3.6).

2.6 Polyacrylamide Gel Electrophoresis

2.6.1 Polyacrylamide Gels

1. Vertical polyacrylamide gel apparatus (e.g., C.B.S. Scientific Company, Del Mar, CA) with spacers and combs.
2. Strong detergent (e.g., Deconex 11, Borer Chemie, Zuchwil, Switzerland).
3. Bind silane and repel silane (Pharmacia Biotech).
4. Stock solutions of acrylamide/bisacrylamide are available commercially; these are easier in handling, compared to making up the solutions from acrylamide powder (safety: *see* **Note 12**).
5. Urea.
6. 10× Tris–borate-EDTA (TBE) buffer (*see* Subheading 2.3).
7. Tetramethylethylenediamine (TEMED; approximately 40 µL per 20 mL gel).

8. Ammonium persulfate (APS; 1 % solution w/v in H_2O), freshly made.
9. Formamide loading buffer: 80 % formamide, 10 mM EDTA, 0.02 % bromophenol blue.
10. DNA ladders as markers (e.g., 10 bp ladder and 100 bp ladder, Invitrogen).
11. Power supply for constant power at approx. 40 W, e.g., BioRad PowerPac 1000 or similar (*see* **Note 7**).
12. White plastic trays for staining.
13. Shaker for staining, e.g., IKA Vibrax VXR.

2.6.2 Silver Staining of Polyacrylamide Gels

1. Fixing solution: 10 % acetic acid.
2. Distilled water for washing.
3. Impregnation solution: 1 L distilled water, 2 mL 37 % formaldehyde, 1 g silver nitrate.
4. Developing solution: 1 L distilled water, 30 g sodium carbonate, 2 mL of 37 % formaldehyde, 200 µL of 10 % sodium thiosulfate.
5. Scanner for digitizing gel image (any standard computer scanner can be used).

2.6.3 Fluorescent Dye Staining

1. Ethidium bromide staining solution: *see* Subheading 2.3 and for safety again, **Note 8**.
2. SYBR Gold working solution: prepared from commercially available concentrate from Molecular Probes, Inc., by adding 20 µL of concentrate to 200 mL of 1× TBE buffer, stored at 4 °C (*see* **Note 13**).
3. Standard transilluminator (302 nm with 6×15 W tubes, e.g., Hoefer Scientific Instruments).
4. Imaging system (e.g., Gel Print 2000i, Bio Photonics Corporation, Ann Arbor, Michigan) with 600 nm band pass filter.
5. Software for analyzing gel images, e.g., Gene Profiler, Scanalytics, VA, USA.

2.7 Denaturing High-Performance Liquid Chromatography (dHPLC)

2.7.1 Specific PCR for dHPLC

1. For dNTPs and plant DNA template, as well as a thermocycler, *see* Subheading 2.2. Best results are achieved, in our hands, if plant DNA extracts are diluted fivefold with water before PCR.
2. Polymerase (e.g., Phire II, Finnzymes).
3. 5× buffer detergent free (e.g., Finnzymes), 5× buffer (Finnzymes), *see* **Note 14**.
4. Forward and reverse PCR primers (4 µM each)—*see* Subheading 2.1.
5. $MgCl_2$ (50 mM).

6. Nanodrop DNA spectrometer (Thermo Scientific).
7. 96-well microtiter plate.

2.7.2 Formation of Heteroduplex DNA for dHPLC

1. PCR products of **step 1** of Subheading 3.7 for samples to be tested and for a standard DNA (from one specific plant) to which every sample in the test is compared (*see* Subheading 1.4).
2. Thermocycler (as in Subheadings 2.2 and 2.7.1).

2.7.3 dHPLC Runs

1. Bio-inert gradient HPLC system with UV detector, e.g., Varian Inc. (now part of Agilent) HELIX system or Transgenomic's WAVE system (a suitable HPLC system can also be composed of single components meeting the same specifications).
2. Columns: Varian Helix DVB or Transgenomic's Wave columns contain nonporous microbeads of 2 μm divinylbenzene (Agilent, after taking over Varian Inc., has recently stopped delivery of the Varian Helix DVB column; replacement products can be obtained from standard HPLC suppliers).
3. "Buffer A": 100 mM triethylammonium acetate (TEAA), 0.1 mM EDTA.
4. "Buffer B": 25 % acetonitrile (v/v) added to buffer A.
5. Plasmid pUC18 cut with HaeIII used as a control standard (available from Sigma).

2.8 Chloroplast Micro- and Minisatellites

Materials are mentioned in previous Subheadings: primers, 2.1; PCR amplification, 2.2; agarose gels, 2.3; PCR-RFLP, 2.4; capillary electrophoresis, 2.5; and polyacrylamide electrophoresis, 2.6.

2.9 Next-Generation Sequencing (NGS)

This is usually done as a service by specialized central laboratories or companies.

1. Microgram amounts of pure DNA from each sample to be subjected to NGS are necessary for most current methods.
2. Data delivered by service providers are often one very big (gigabyte size) file, so good computer network connections are necessary, as well as a computer with fairly good RAM and ROM sizes.

3 Methods

3.1 Use of Databases and Alignments for Finding Suitable PCR Primers

PCR primers specific for chloroplast DNA are presented in an online database (*see* Subheading 2.1). Alternatively, additional primers can be designed by aligning available chloroplast DNA sequences of interest (as obtained, for instance, from GenBank; *see* also Subheading 2.1).

1. Description of the online database (*see* Subheading 2.1): The database is a big table which can be searched, filtered, and ordered. There are data columns for primer name; forward/reverse direction (relative to the *Nicotiana tabacum* chloroplast genome); the genes in which the primer resides (if any); primer length; primer sequence; position of the 5′ nucleotide of the primer in the completely sequenced chloroplast genomes of *Nicotiana tabacum*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Oryza sativa*, *Eucalyptus globulus*, *Acorus calamus*, *Atropa belladonna*, *Marchantia polymorpha*, *Spinacia oleracea*, *Pinus thunbergii*, *Oenothera sp.*, *Zea* mays, and *Medicago truncatula*; and reference citations. The primers in the current database table can be ordered by any of these columns. A filter function makes it possible to search for primer or gene names, forward or reverse primers (relative to the tobacco sequence), or free text (e.g., within citations).
2. Ordering function: The most important use of the ordering function is to order primers along the chloroplast DNA sequences of the currently included reference species. By marking, copying, and pasting these numbers into a spreadsheet calculation program, lengths of PCR fragments by many pairs of primers can be calculated. Adequate primer orientation, however, has to be confirmed by comparing the gene order between *Nicotiana* and the species of interest for the particular region—this can be done by ordering the table for positions of the primers in one of the two species and check the column of the other species for continuity in this region. If working with a species that is not present among those in the database, it is a good idea to check any recently published molecular phylogeny for the one among those species present in the database that is closest to the species of particular interest (e.g., if working with a grass species, try *Zea mays* first).
3. Another use of the ordering function is to identify primers with similar names or around particular regions. The "primer show" button at the upper right-hand side invokes a function that will show the selected primers in any DNA sequence that is pasted into the upper window in FASTA format. For example, newly appearing complete chloroplast DNA sequences in GenBank can be retrieved in this format and utilized (the function can be very slow though with large data sets).
4. Designing primers from alignments: In the original database article [31], the method of Antoniw [76] was employed for deducing suitable primer sequences from multiple alignments. This suite of programs was written in DOS, which is not in use much anymore (though it may be possible to run them inside an MS-DOS window in the Windows operating system). We have often designed primers for specific lineages "manually"

(e.g., [77–79]). Since then, bioinformatics has made great leaps forward, and programs like Amplicon [80], Greene SCPrimer [81], or iCODEHOP [82] automate these processes on today's computer systems. Using the iCODEHOP principles, Contreras-Moreira et al. [83] have designed a web-based software that allows the user to first align DNA sequences and build a phylogenetic tree and then to select portions of the tree for finding primers specific for the chosen clades (*see* **Note 15**). The same approach is also useful for choosing or designing sequencing primers.

3.2 PCR Amplification

"Generic" PCR conditions that favor successful amplification of chloroplast DNA with combinations of primers from the database [31] are standard PCR conditions, as shown in the following:

1. Use 5–25 ng template plant DNA, 2 mM Mg^{2+}, 0.25 U DNA polymerase per 10 μL reaction. We prepare such reactions in volumes of 10–25 μL, depending on further processing steps (*see* **Note 16**).
2. Run PCR program in two parts: with the first 10 cycles at 70 °C annealing, followed by 32 cycles at 55 or 50 °C annealing ([31, 78]; *see* **Note 6**).

3.3 Agarose gel Electrophoresis Method

Agarose gel electrophoresis procedures are standard, but it may be worth saying that we are using very thin combs (less than 1 mm thick), and include ethidium bromide in the gels, but not in the buffer. Here is a short description:

1. Mix agarose for the special purpose: Routinely, we mix high resolution and high gel strength agaroses, e.g., 3:1 w/w. This gives higher gel strength while still providing adequate resolution. Higher resolution is need in PCR-RFLP (*see* Subheading 3.6) and for chloroplast minisatellites (*see* Subheading 3.8). Add the desired quantity, depending on gel concentration, to the desired volume of 0.5× TBE buffer. For instance, our gel apparatus and trays hold 100 μL of gel; a 1.5 % gel (suitable for resolving PCR-RFLP fragments of 150–2,000 bp size) would require 1.5 g of agarose in total, made up of 1.125 g high resolution agarose and 0.375 g of high gel strength agarose. Carefully boil the suspension in a microwave oven (retardation of boiling may occur), cool down for 2 or 3 min, and add approx. 20 μL ethidium bromide concentrate (*see* **Note 8** for safety issue; the exact amount will depend on the specific setting of gel apparatus, transilluminator, and gel imaging system). We add ethidium bromide into the gel as this allows taking pictures during the course of the run (as in **step 3**). Put thin combs in place and let the gel set.
2. When the gel has set, put it into the electrophoresis tank, carefully pull out the comb, and load the samples. Connect to power supply and run at constant voltage (e.g., 120 V in our setting).

3. After the front has migrated approx. 2.5 cm, take the gel with its tray out of the apparatus and take a first picture over the transilluminator using the gel imaging system. This is especially recommended when performing PCR-RFLP (*see* Subheading 3.4), as small restriction fragments might only be visible early on (*see* also Subheading 3.4, **step 4**).
4. Put the gel back in, run it until the front has reached the bottom of the gel, and take a picture as in **step 3**.
5. If further resolution is desired and seems possible, put gel back in and run further (*see* Subheading 3.4).

3.4 PCR-RFLP

The following is a description of our standard methods (e.g., [84, 85]):

1. Check PCR fragment (obtained with Subheading 3.2) quality and quantity on an agarose gel (*see* Subheading 3.3). Watch for absence of shadow or double bands and other artifacts. Good resolution on the gels is necessary to reveal some of those artifacts. If too many, or any strong artifacts, are present, either try to optimize PCR conditions until a single band is obtained or consider cutting the right band from the agarose gel and purifying it (commercial kits are available for this).
2. Use, e.g., 5 μL of a typical PCR for the restriction reaction. It contains the components of the PCR buffer, i.a. Mg^{2+}. In our PCR protocols (*see* Subheading 3.2), a standard concentration of 2 mM of this ion is used. The mix recommended below takes this into consideration, trying to propose a compromise between exactly adjusting Mg^{2+} levels (to 10 mM in most restriction enzyme buffers) and minimal pipetting steps. Add 0.7 μL of the 10× restriction enzyme buffer, milliQ water to 8 μL (including the PCR volume which is, in this example, 5 μL), and 0.05 U of restriction enzyme (*see* **Note 17**) for each sample (by preparing a mix with these three components and adding aliquots of 3 μL to each sample; *see* **Note 18**). Mix gently and incubate at 37 °C (for most enzymes) for several hours (we usually digest PCR fragments overnight).
3. Perform gel electrophoresis on either agarose (*see* Subheading 3.2) or polyacrylamide gels (*see* Subheading 3.6).
4. Interpret banding patterns: the sum of the sizes of all bands in a restriction enzyme pattern from chloroplast DNA should equal the size of the original fragment. Check for restriction site polymorphisms (where some bands in certain samples are not cut, while they "disappear" and "decay" into separate bands in other individuals), length polymorphisms of some restriction fragments (which can be rather small), and any abnormalities. For instance, it is difficult to detect small restriction fragments (below 100 bp) on agarose gels. DNA bands bind ethidium bromide proportionally to their length/size, so

shorter bands will bind less dye and remain darker as well as more fuzzy (due to enhanced movement of smaller molecules in the gel matrix). We therefore routinely take the gel out of the apparatus during the run (after e.g., 20–30 min) and take a picture over the UV box in order to capture the small fragments (as our gels contain ethidium bromide, *see* Subheading 3.2). Primer-dimer artifacts from PCR can irritate here, suggesting the presence of small bands when in reality these are not results of restriction of the PCR product; primer-dimers can be particularly strong though, which would be untypical for restriction fragments of small size. Another hint is that bands should appear more prominent in a steady relationship to their sizes, i.e., longer bands in a pattern (a gel lane) are best stained, and each shorter band should gradually appear less intensively stained. If a band breaks this pattern, it can be a (brighter) double band (two bands of very similar sizes on top of each other) or an (fainter) artifact (e.g., often an intermediate product of incomplete fragment digestion; *see* **Note 19**). Note down the approximate sizes of all restriction sizes accepted to be "real" digestion products of the original PCR fragment and compare to its original size. Also note down any variants present only in some samples and describe the nature of the variation (e.g., restriction site polymorphism or approximate size of the length polymorphism of a specific restriction fragment; *see* **Notes 20** and **21**).

3.5 Sanger Sequencing of Chloroplast DNA PCR Fragments

The following steps are necessary: checking the quality and quantity of the PCR products (by standard agarose gel electrophoresis; *see* Subheading 3.3), purification of the PCR product by phosphatase and primer digestion and/or silica column kits and/or alcohol precipitation, the actual sequencing reaction, another alcohol precipitation, and the final application to a capillary or polyacrylamide slab gel sequencer.

1. We start by using 250 ng of PCR product (obtained following Subheading 3.2). The volume containing this amount (estimated from agarose gel electrophoresis, *see* Subheading 3.3, or spectroscopic measurement, *see* Subheading 3.7, **step 4**) is digested by adding exonuclease and shrimp alkaline phosphatase in order to destroy remaining primers and triphosphate nucleotides (dNTPs), using the enzyme manufacturers' recommendations. Both reactions can proceed at the same time in the same tube. Alternatively, PCR clean-up kits, offered by many different manufacturers, can be used. These kits work by binding DNA to silica (columns). The treated PCR fragments are precipitated preferentially with two volumes ethanol followed by centrifugation (*see* **Note 22** for precipitation in 96 well plates). The pellet is dissolved in 2.5 μL H_2O (milliQ quality).

2. Prepare mix for sequencing from DTCS kit (Beckman Coulter): The kit can be used at 1:4 dilution with good quality templates. For each sample to be sequenced, mix 2 μL of DTCS Quick Start MasterMix from kit, 0.6 μL Tris–HCl (pH 8.0, 1 M), 5.1 μL H_2O, and 0.3 μL $MgCl_2$ (50 mM; often comes with PCR polymerase). For sequencing a fragment from both ends (using the original primers from PCR), distribute this diluted mix into two equal parts and add primers. For each sample and primer reaction, use 8 μL of the diluted mix, 0.8 of a 4 μM primer stock, and water to 19 μL. Add 1 μL of the purified and dissolved PCR product per reaction. Put into PCR machine and cycle between following temperatures: 96 °C for 20 s, 50 °C for 20 s, and 60 °C for 4 min (*see* **Note 23**). After 30 cycles, cool to 4 °C and recover the samples. For stopping the sequencing reaction, add the following freshly made mix to each sample: 1 μL sodium acetate (3 M stock), 0.4 μL EDTA (0.5 M stock), 1 μL glycogen (20 mg/mL, contained with DTCS kit), and water to 5 μL. Mix gently. Proceed with another ethanol precipitation (*see* **step 1**). Dry the pellets by leaving the tubes or plates open on the bench for approx. 15 min (no visible traces of liquid must be visible after drying).
3. The dried pellet is dissolved in sample loading solution (SLS, de-ionized formamide) for the Beckman Coulter CEQ 8000 sequencer. Add 30 μL of SLS to each sample, move gently to dissolve DNA (5 min). Transfer dissolved sample into CEQ sample plate and run sequencing electrophoresis. We routinely use a set of conditions (LFRb: capillary temperature 57 °C, denature sample at 90 °C for 120 s, inject sample at 2.0 kV for 15 s, separate at 6.0 kV for 60 min) that takes about 1 h for each sample and resolves up to 600 bp.
4. Commercial service providers and local resource centers offer the sequencing of custom samples as a paid-for service. Usually, either unpurified PCRs (as they come from the machine) or self-purified fragments (using methods as outlined above in **step 1**) can be sent. The DNA amount and concentration to be sent are specified by the service provider. Likewise, the customer has to supply the oligonucleotide(s) for initiating the sequencing reaction. Prices and offers are extremely variable, and it pays to "shop around" for current offers.
5. Sequencing result files use a file format specific to the sequencing system for displaying the electropherograms. The most common file formats (e.g., by the ABI system) can be recognized by many computer programs. A standard format is *.scf (note, however, that the MS Windows computer systems uses this extension for system files, which can cause problems). A useful freeware is BIOEDIT [86], which was recently updated

to version 7.2.3 (http://www.mbio.ncsu.edu/bioedit/page2.html). This software can not only show the electropherograms, but also assist in all the other standard operations with sequences, e.g., alignments (*see* **Note 24**).

6. Phylogenetic analysis routines are described elsewhere in this volume. A few considerations, though, have to be added on the choice of marker or sequence system (*see* **Notes 25**–**27**).
7. For the special case of DNA barcodes (*see* Subheading 1.3), raw data processing involves cleaning the sequences as described above. Subsequently, they are usually compared to a database of barcodes, in order to identify those species with the highest similarity to the sample. Such database can be found online, and it is possible to simply "cut and paste" the query sequences directly into a box in an Internet browser window. For example, the Barcode of Life Data Systems Webserver (http://www.boldsystems.org) offers an "identification machine" that works on this principle. Of course, such similarity searches can also be done in the standard DNA databases, e.g., GenBank, by using their BLAST facilities. As attractive as this sounds, our practical experience is such that depending on the individual quality of the entries, species identification based on the barcodes can be a nontrivial exercise (*see* **Note 28**).

3.6 Polyacrylamide Gel Electrophoresis Methods

Polyacrylamide gel electrophoresis is also a standard technique and hence, described elsewhere [19], but methods for staining are more varied nowadays than previously [17]. We will describe the silver, ethidium bromide, or SYBR Gold staining methods to visualize DNA.

1. Carefully wash the glass plates of the electrophoresis apparatus with a strong detergent (e.g., Deconex 11, Borer Chemie). Cover one surface of one plate with bind silane and a surface of the second plate with repel silane (Pharmacia Biotech), following the recommendations of the manufacturer and [19].
2. For preparation of a gel sandwich, use 0.2 mm × 42 cm spacers (C.B.S. Scientific Company) for 20 cm wide gels. Pour gel by mixing, just before pouring, the appropriate volume of acrylamide stock solution (acrylamide/bisacrylamide 37.5:1; for safety issue, *see* **Note 12**), urea to give a 7 M end concentration, 10× TBE buffer, TEMED (approximately 40 μL per 20 mL gel), and freshly made solution of APS 1 % (approximately 500 μL per 20 mL gel) to give an 8 % polyacrylamide gel. Pour gel and leave to polymerize for approximately 30 min. UV light and higher temperatures enhance polymerization.
3. Mix 1 to 4 μL of PCR product with 4 μL of formamide loading buffer, denature at 94 °C for 5 min, snap cool on ice, and apply to the gel slots. Use DNA ladders as markers in some lanes,

preferentially not the outermost lanes, as they might become subject to a "smiling" effect. Electrophorese the samples through the denaturing polyacrylamide gels in 1× TBE buffer for 2–2.5 h at 40 W constant power and 50 °C using a temperature probe attached to the glass plate.

4. Dismantle gel apparatus (the gel should stick only to the surface treated with bind silane). Prepare for staining (*see* **Note 29**).
5. For silver staining [87], all steps are performed in white plastic trays on a shaker (e.g., Vibrax VXR, IKA, Staufen, Germany) at room temperature (around 20 °C in Central Europe). After 25 min incubation in fixing solution, the plate with the gel sticking to it is washed in distilled water for 10 min. For impregnation, the gel is gently shaken for 30 min in the appropriate solution (*see* Subheading 2.6). Then the gel is rinsed shortly with developing solution to get rid of the excess silver nitrate. Developing is performed by shaking the gel in new developing solution until the DNA fragments become visible on the white background of the tray (this is the step most dependent on the level of ambient temperature). Placing the gel again in fixing solution for a few minutes will stop developing at this point. To obtain documentation in a digital form, the gel can be scanned using a digital scanner (most often, the interesting part of the gel does not exceed standard DIN A4/US letter size; otherwise, scan parts of the gel subsequently).
6. Methods for ethidium bromide and SYBR Gold staining are similar. Place the glass plate with the gel sticking to it, facing upwards, in a plastic tray. Distribute 50 mL of staining solution (either ethidium bromide or SYBR Gold; *see* **Notes 8** and **13** for safety, respectively) over the range of the gel where products are expected to appear and incubate for 15–20 min in the dark. After this time, pour off the stain, discard it, and rinse the gel shortly with distilled water. The glass plate with the gel facing down is put onto a standard transilluminator by placing thin spacers under the edges of the glass. The camera lens is focused on the glass. A photo of the interesting part of the gel is taken like for standard agarose gels stained with ethidium bromide, through a 600 nm band pass filter with an imaging system. The photograph can be stored and used in digital form for further analysis (*see* **Notes 30–33**).

3.7 dHPLC

This section describes the method for PCR specifically for subsequent dHPLC analysis, heteroduplex formation, and dHPLC runs and interpretation.

1. The PCR master mix for a 96-well plate (30 μL each sample) consists of 570.24 μL 5× detergent-free buffer, 63.36 μL 5× buffer (*see* **Note 34**), 31.68 μL $MgCl_2$, 63.36 μL dNTPs

(10 mM each), 158.4 µL forward and reverse primer each (at 4 µM concentration), 2120.98 µL H_2O, and 1.58 µL polymerase. With the same master mix amount, prepare PCR product of a standard plant sample. This sample is necessary for heteroduplex formation (*see* **step 5**).

2. Add 30 µL master mix to 1.0 µL DNA in each well.
3. Put plate into thermocycler and start the following program: initial denaturation at 98 °C (50 s), then cycle between denaturation at 98 °C, annealing at 55 °C (or a better suitable temperature, depending on the primer pair) for 50 s, and elongation at 72 °C (2 min) for 34 cycles. End with a final extension phase at 72 °C for 5 min, then hold at 4 °C.
4. Measure concentration of the PCR products on the Nanodrop or a similar photometer and load 5 µL on a 1.5 % agarose gel for checking quality (*see* Subheading 3.3). PCR product concentration should be in the 100–200 ng/µL range, and a single, clear band should be present.
5. For forming heteroduplex DNA, first calculate necessary sample volumes. For dHPLC analysis, 5 µL injection volume are required. To avoid total evaporation during analysis, mix twice the injection volume plus 20 %. Mix 6 µLPCR product of each individual sample with 6 µL PCR product from standard sample (*see* **Note 35**) in a 96-well microtiter plate. Mix at least once the standard DNA sample with itself and at least three times standard DNA with distilled H_2O. Additionally, use an H_2O blank. As standard DNA and blanks should be distributed over the plate, this will result in approximately 80–90 wells available for samples.
6. The amplicon mixes undergo a post PCR denaturation and then a re-annealing step with a cooling gradient from 95 to 65 °C. Place tubes or multi-well plates containing the sample-standard mixes in PCR machine and run the following temperature profile: 3 min at 95 °C, slow cooling over 30 min to 65 °C, and hold at 65 °C for 3 min. Snap cool on ice.
7. Starting up the dHPLC system: This is a summary of the standard operating procedure (SOP) of the manufacturer (in our case, Varian Inc. Helix system). Boot up the dHPLC system and turn the lamp of the UV detector on. Fill the buffer reservoirs and check buffer storage tanks. Check the system and peek tubings for any leakage or bubbles. Start wash procedure and check syringe for any bubbles (tap gently to loosen them). Equilibrate system in order to obtain stable oven temperature and a stable baseline at the detector. Adjust the needle height for sampling. Open the Univ50 method (Varian Helix system) and equilibrate again. Inject 5 µL of the pUC18 HaeIII standard to test column performance (*see* **Note 36**). Repeat the

standard injection two more times. Compare the retention times with those given by the manufacturer in the SOP. If resolution is not satisfactory, the column has to be cleaned by injecting 0.5 M EDTA or even flushed with 75 % acetonitrile. Once stable retention times and satisfactory patterns are achieved, the system can be adjusted to the parameters necessary for particular sample injection. Create such a method as suggested by http://insertion.stanford.edu/melt. Set the oven to the recommended temperature.

8. dHPLC method building: If the sample sequence is totally unknown, it is necessary to get some kind of a reference sequence to get a clue where to start method building. For instance, try to retrieve the sequence of the nearest relative present, e.g., in GenBank. For a dHPLC run, the criteria to consider are the working temperature for the column (at which the specific PCR amplicon denatures partly) and the flow rate of both buffers. The changing concentrations of the buffers during the gradient are very important in order to optimize the retention time of the molecules. To estimate the correct melting temperature of the double strand, a helpful tool is the melt program from Stanford (http://insertion.stanford.edu/melt). There, one can upload the known sequence or a reference sequence, and the output will be a recommended melting temperature plus the concentration of the buffers at which the heterozygous samples should be analyzed.
9. With this hint, test parameters with a test set of samples and start optimization. To "move" a peak, it is recommended to adjust the timings of change of the buffer concentrations. Important to note is that each fragment needs different parameter sets for good peak resolution. If there is no reference sequence, the buffer concentrations and the melt temperature have to be determined empirically, which makes it a bit time-consuming to set up the parameters. One hint is to start with a recommended dHPLC universal method and approach the melt temperature by increasing the temperature by 2 °C in each consecutive injection, until the peak in a mixed sample starts to split up. In order to get an idea how a universal method adjustment looks like, consult Table 1.
10. dHPLC operation—running samples: Load the method created (on the basis of the Stanford Melt program recommendations). Set oven to the correct temperature and adjust buffer concentrations and flow rates. Equilibrate the system by running the baseline buffer concentration until a stable detector baseline is achieved. Create your sample list on the autosampler module of the system: sample IDs, blanks, flushing intervals, and standard DNA injections (*see* **Note 37**). Fill a fresh 96-well microtiter plate with twice the injection volume of

Table 1
Typical dHPLC method

Time	% Buffer A	% Buffer B	Flow rate (mL/min)
0:00	55	45	0.45
0:30	50	50	0.45
6:00	32	68	0.45
7:00	32	68	0.45
7:01	55	45	0.45
8:50	55	45	0.45

Time is given in minutes:seconds
Buffer percentages are on a volume basis
Flow is total flow (buffers A + B combined). For further explanations, *see* text

5 μL (plus a 20 % surplus, i.e., 12 μL in total) of the mixed samples after having formed heteroduplexes and put the plate into the autosampler. Close the plate with an aluminum seal to prevent evaporation and for dark storage of the samples. Check the needle height settings once again.

11. The dHPLC system can be run automatically and unattended overnight if attention is given to possible sample evaporation and the environmental temperature is controlled. At the time when the analysis can start, the defined standard (the PCR product of the standard DNA sample, which has been hybridized with each individual sample) must be run alone as well in order to compare the patterns. When choosing a standard method for the dHPLC runs, each run will take about 9.5 min (8.5 min. analysis, 30 s equilibration, 30 s re-equilibration).
12. Interpretation of the results: To make interpretation of the peaks as easy as possible, the optimal parameter settings should be carefully selected. Within the analyzing software (provided by the Varian Inc. Helix system), up to 6 results can be stacked in order to compare patterns. The recommendation is to always check the sample runs against the standard alone. In the first round of interpretation, coarsely differentiated patterns should be identified (*see* Fig. 1); samples that do not appear to clearly fall into one of the categories should be reexamined in a second round of interpretation, once a better overview regarding the types of patterns has been achieved (*see* **Note 38**).
13. Each identified category (bin) then consists of samples that have identical or very similar DNA sequences, as the dHPLC system is sensitive to single base mismatches or insertions/deletions in a 700 bp fragment under optimal conditions. Sequencing a small number of samples (using methods in Subheading 3.5) from each bin will usually confirm this assumption.

3.8 Methods for Analyzing Tandem Repeats in Chloroplast DNA: "Microsatellites" or "Simple Sequence Repeats" (SSRs) and "Minisatellites"

Long stretches of mononucleotide repeats (most often, As or Ts are present in longer stretches) make amplification and sequencing problematic. "Universal" primers and methods for such "chloroplast microsatellites" are available from consulting [34, 64–68]. Capillary sequencers or long polyacrylamide gels have to be used for fragment separation because of the "stuttering" problem. Length differences are usually small. To circumvent this problem, amplified fragments can be treated with a restriction enzyme that shortens the DNA carrying both the mononucleotide repeat and the fluorescent label (i.e., cuts off the non-labeled primer and a part of the fragment adjacent to it). The remaining fragment is shorter, so the relative size change due to a single nucleotide is bigger, making detection easier. Flores-Renteria and Whipple [88] have recently suggested redesigning one primer to contain a part of the mononucleotide repeat. In their hands, stuttering and minute size changes are less of a problem with such redesigned primers.

1. Capillary electrophoresis or analysis on long denaturing polyacrylamide gels (*see* Subheading 3.6) is usually necessary to reveal exact length differences for these "microsatellites" (*see* **Notes 39** and **40**).
2. "Minisatellites" in the chloroplast are a related type of sequence polymorphism. They consist of "motifs" of 5–30 bases ([69–71]; *see* **Note 41**). Initially identified on agarose gels of digested amplification products (as detailed in Subheading 3.4), we now routinely analyze such a locus [72] on a capillary electrophoresis system after restriction digestion, with one primer labeled with fluorescent dyes. This requires that a restriction enzyme is chosen which cuts between the tandem repeats and the unlabeled primer, but not on the other side of the repeat (*see* **Note 42**).

3.9 Next-Generation Sequencing

1. Several research groups have aligned complete chloroplast DNA sequences with fragment sequences retrieved from GenBank in order to study polymorphisms in the alignments (e.g., [75, 89]). The experience from both cited studies is that a surprising amount of hitherto unexplored polymorphisms can be identified from this endeavor. The drawbacks are that (1) these sequences again represent commonly sequenced regions, so the advantage of having all regions of the chloroplast genome in the complete genomes is not fully exploited, and (2) sequences submitted to GenBank, but not formally published in peer-reviewed literature, in our view sometimes represent questionable levels of sequence quality (i.e., may contain sequencing errors) and may therefore be less useful [75].
2. On the other hand, Besnard et al. [90] have completely sequenced eight different olive (*Olea europaea*) chloroplast genomes. Markers were developed for two single nucleotide polymorphisms (SNPs) and 62 length polymorphisms (including mononucleotide repeats or "chloroplast microsatellites").

Surprisingly, only 40 polymorphic loci (2/3 of the tested markers) resulted from this work (*see* **Note 43**). In the future, it is generally expected that the cost for mass sequencing will further come down. This will open new perspectives for the analysis of variation in the chloroplast genome. A recent study by Morris et al. [91] is an early example (*see* **Note 44**).

4 Notes

1. Plant DNA useful for analysis of chloroplast DNA can be obtained by almost any standard DNA extraction protocol, or DNA isolation kit, in most cases (e.g., Sigma or Qiagen plant DNA isolation kits). However, using additives in the initial extraction buffer, or indeed, during the cell disruption process, can improve DNA quality by inhibiting browning reactions early on [92]. The kits are often available in a high-throughput format, e.g., for extracting DNA from (2×) 96 samples in parallel. Most DNA extraction protocols yield DNA in solution, which can be directly put forward to enzymatic reactions like PCR. An alternative method of obtaining DNA for PCR consists of squashing leaves onto protective membranes, drying this membrane, and punching a small disc from it for use in PCR (Whatman FTA cards). DNA from most (green) plant tissues will contain sufficient amounts of organelle DNA for this purpose, as PCR can make use of even tiny amounts of template DNA. Usual estimations for chloroplast DNA content in total genomic DNA preparations from green plant material are in the range of 5–10 %. Furthermore, the complexity of chloroplast DNA is much lower than that of nuclear or mitochondrial DNA (e.g., approximately 1:1,000 for chloroplast to nuclear genome complexity for *Arabidopsis*). This means that each chloroplast DNA sequence will be present in many copies, even if the total amount of DNA is low. This is because green plant tissue contains numerous chloroplasts, and each chloroplast contains numerous copies of the chloroplast DNA "chromosome" [12]. Lutz et al. [93], while experimenting with DNA extraction methods yielding purer nuclear DNA for next-generation sequencing experiments, have tried to determine the levels of chloroplast and mitochondrial DNA "contamination" in total plant DNA preparations. Their results indicate that between 0.4 and 15 % chloroplast DNA may be present in DNA isolated with the commonly used CTAB method, while this proportion could be suppressed to a level of 0.2–9.1 % using their recommended nuclei isolation protocols. Nock et al. [94] have counted chloroplast reads in next-generation sequencing experiments with total DNA and report

similar percentages (2.77–11.63 %) for chloroplast DNA in total DNA preparations. The copy number ratio for chloroplast to (single copy) nuclear genes (determined by quantitative PCR) from the experiments ranged from 56 to 245 in total DNA preparations (data recalculated from [93]). If these results hold under more general conditions, the rationale is that it is between roughly 50 and 250 times "easier" to amplify chloroplast DNA fragments than such of nuclear genomic origin from total DNA preparations, but also that next-generation sequencing of such DNA preparations may end in a frequency differential between chloroplast and nuclear DNA in this order of magnitude (*see* [8]). It is interesting to note that Tuskan et al. [95], when assembling the nuclear and chloroplast genomes for *Populus trichocarpa* using a whole genome shotgun (Sanger) approach, arrived at a ratio within this range (410×/7.5× = 54.7) when their coverages of the chloroplast and nuclear genomes are compared (*see* Supplementary Information pp. 4 and 12, respectively, [95]).

2. The main criterion for DNA quality is absence of enzyme inhibitors, as practically all of the downstream processes depend on enzymatic steps. Most DNA extraction methods based on binding of DNA to silica membranes in the presence of chaotropic salts result in DNA of sufficient quality. In general, DNA precipitation by alcohols (ethanol or 2-propanol) tends to retain some enzyme inhibitors to a higher degree. However, our main experience is that the source and storage conditions of plant tissues have the highest influence on DNA quality. Tissues that tend to wilt, brown, or otherwise deteriorate quickly after harvesting are usually problematic. To circumvent these problems, use only healthy looking plant material and either freeze plant material immediately after harvest (preferentially at −80 °C) or put it on silica gel for immediate dehydration (a method more amenable to field work). In some cases, it may be necessary to experiment with different types of tissues, as different classes of problematic substances (e.g., polysaccharides, polyphenols, fats or oils) are present in various tissue types. Another consideration is feasibility of collecting material. For instance, for tall trees that have leaves (or buds) only high up in the canopy, which may be difficult to access, isolating DNA from small cambium scrapings of the trunk is a recommended alternative [96–98].
3. Useful DNA concentrations for subsequent PCR amplifications are in the range of 10–50 ng/μL, but even less DNA can be amplified successfully in many cases. Often, the extraction kits yield higher initial DNA concentrations. Diluting this DNA also dilutes any remnant enzyme inhibitors, thus making subsequent analyses easier.

4. Most commercial thermostable polymerase preparations come with buffers, and often further components like Mg^{2+} solutions, additives to improve the amplification reaction, or even with deoxynucleotide triphosphates (dNTPs). Certain downstream processes, e.g., dHPLC, demand for absence of detergents in the PCR solution that is applied to the HPC columns. Many manufacturers supply polymerase buffers without such detergents. If PCR buffer components are not revealed by the manufacturer, information about the presence or absence of detergents in their buffer formulations can almost certainly be obtained by directly contacting them.
5. DNA polymerase properties and primer annealing temperatures must match. Some recently introduced thermostable DNA polymerases with a remarkable activity (e.g., Finnzymes PHIRE I polymerase) best work at primer annealing temperatures of 60 °C or above. Many primers obtained from the literature, however, have annealing temperatures in a more standard range of 50–58 °C. Using our standard PCR program outlined above, we have recently obtained good success by using Finnzymes PHIRE II polymerase. This is also a highly processive enzyme that is designed for fast PCRs with short holding temperatures during PCR cycling. This enzyme works at great dilutions with more standard cycling conditions, thus great savings are possible. As the enzyme is designed to work at annealing temperatures below 60 °C as well, less care must be taken when primers with lower annealing requirements are employed. Trial and error will always guide such efforts (*see* following **Note 6**).
6. If experimenting with different primer annealing temperatures, a PCR thermocycler capable of realizing different temperatures across its heating block (temperature gradients) may be useful.
7. Safety note concerning power supplies: Always make sure electric current cannot harm you. Use the following order of manipulations: put lid onto electrophoresis tank, connect wires on the lid if necessary, plug other end of wires into power supply, switch on power supply, adjust voltage or power setting, and press "run" (or similar). For interrupting or dismantling a gel, use reverse order (press "stop" in some power supplies).
8. Safety note concerning ethidium bromide: This chemical is a mutagen, in that it interacts with DNA. It is also toxic, but typical laboratory levels (below 1 μg/mL) are below the toxicity threshold. Use only in a (small) dedicated laboratory room, use gloves for all manipulations, and always wear a dedicated lab coat (which should be closed). Disposal should be in accordance with local applicable regulations (these tend to vary). Our method of disposal is by filtering all solutions through an activated charcoal filter and by collecting all gels in a separate

container for toxic waste (to be incinerated by special waste companies). All surfaces that have come in touch with ethidium bromide are rinsed with isopropanol for decontamination.

9. Agarose gels can be recycled by saving used gels in the refrigerator and boiling them in a microwave oven immediately before the next usage cycle. We often filter the hot, liquid gel at this point to remove any particles that accumulate. This whole procedure will decrease electrophoresis cost considerably. In a similar way, electrophoresis buffer can be used over a period if stored in the fridge between uses, in our case for about 2 weeks, before being replaced with freshly made buffer.
10. On average, a 4-base cutter will cleave random DNA at every 256 positions. However, chloroplast DNA is not "random." Chloroplast DNA is rich in A and T nucleotides, so enzymes with a higher number of those nucleotides in their recognition sequence will, on average, cut more often.
11. Polyacrylamide sequencing gels are still in use for sequencing purposes; laboratories considering to make use of them are referred to [18] or [19], where the use of radioactively labeled nucleotides is suggested. This is no longer necessary, given the advancements in silver staining (*see* Subheading 2.6), but their recipes and methods can be used without radioactive labels and sequencing gels prepared following the rationale of the polyacrylamide gel electrophoresis and silver staining sections below (*see* Subheading 3.6).
12. Safety note for acrylamide/bisacrylamide: These chemicals are potent neurotoxins in liquid and powdered form. Use personal safety equipment (gloves, spectacles, and for handling powder, a protective mask). Dispose of any unpolymerized rests with toxic waste. The polymerized product, however, is an "everyday" chemical.
13. Safety of SYBR Gold: This dye is photosensitive and binds to glass surfaces, so it should be stored in a polypropylene bottle, wrapped in aluminum foil. It can be irritating on skin and in contact with eyes, so wear protective clothing. Possible mutagenic potential is still not adequately known. Dispose of like ethidium bromide (*see* **Note 8**).
14. In order to analyze the samples on the dHPLC system, it is necessary to consider possible detergents in the polymerase buffer (these should be absent as much as possible). Detergent-free buffers are on the market now; these increase the performance of the dHPLC system to a much higher resolution. If detergent-free buffers are not an option, PCR clean-up can be done as well, but will add to the analyzing costs.
15. Although the database [31] already contains a wealth of primers, and some genes or exons are approaching "saturation" for primer sequences (this applies primarily to the *trn* genes),

we sometimes find surprisingly similar sequences in alignment portions even outsides genes/exons.

16. PCR volume is determined by downstream processing of the products. Usually, 5 μL are employed in single restriction enzyme digests or in single dHPLC runs. 5–15 μL are required for sequencing reactions in our hands. Less volume is usually applied to capillary sequencers (where primers carry a fluorescent dye) or to polyacrylamide slab gels.
17. Not all restriction enzymes work with all residual PCR buffers. Moreover, residual inhibitors introduced together with the DNA from the plant extract may prevent DNA cleavage. If no reaction products (other than the original size PCR fragments) are observed, this might be the reason.
18. The total volume of the restriction reaction should be kept low, so that the DNA is not over duly diluted.
19. Some experience is necessary to interpret gel banding patterns, on agarose as well as on polyacrylamide gels. More clarity on the sizes and order of the restriction fragments can be obtained either by using PCR primers that are binding at slightly shifted positions (these will cause the "end" restriction fragments to shift) or by classical restriction mapping with two or more enzymes used alone and in combination. Partial digests can also yield this information, but in our hands they are more challenging. Partial digests can also result after overnight incubation, either when high DNA concentrations are present after PCR or when residual restriction enzyme inhibitors are present (another reason for this may also be activity loss in the enzyme concentrate after prolonged storage). These are difficult to interpret, but easy to spot, as very often, a (faint) band at the size of the original, entire PCR fragment is present in the pattern, as well as a number of intermediate sized bands.
20. Dhingra and Folta [28] and Lin et al. [35] are particularly noteworthy in that their proposed primer sets allow amplification of large parts of the chloroplast genome with a minimal number of PCRs (long-range PCR). The drawback is that for such long PCR products, template DNA quality must be high, and PCR conditions need to be optimized considerably. The more complicated the PCR setup, the higher the danger of unwanted products, e.g., such from nuclear copies of chloroplast DNA (*see* also **Note 21**).
21. It became clear early on [99, 100] that plant nuclear genomes contain numerous fragments of apparent recent evolutionary origin from the chloroplast. Indeed, most plant genome sequences examined today contain nuclear copies of all chloroplast DNA dispersed through their chromosomes (e.g., [95, 101]). These nuclear copies of chloroplast DNA are a potential threat for PCR

analysis, if co-amplified with genuine chloroplast DNA, because they may contain mutations relative to "genuine" chloroplast DNA of that individual plant [95]. However, as shown above (*see* **Note 1**), the high frequency differential (1:50–1:250) makes it unlikely that these nuclear copies are amplified from total DNA preparations. Furthermore, the nuclear insertions of chloroplast DNA are often short and mutated, and it is therefore less likely that both primer binding sites of a PCR fragment are present.

22. PCR fragments can be precipitated in 96-well plates, if a suitable centrifuge is available. Care must be taken when mixing and turning the plates. We turn the plates over onto filter paper for removing supernatants (after ethanol precipitation and washing) and centrifuge the plate upside down, with a layer of filter paper underneath, at low speed, to get rid of remaining liquid. The pellets in the plate are then dried by leaving the plate (in its original position) in the flow hood for about 10 min.
23. It may be necessary to use somewhat higher primer annealing temperatures for certain fragment/primer combinations, in order to avoid sequencing artifacts.
24. Before aligning sequences, they should be checked for ambiguities that results from imperfect sequencing reactions. Primer sequences at both ends of a fragment should be trimmed (they can be retained temporarily, if that supports the alignment process, but should be removed before making any phylogenetic inferences). It is good practice to sequence DNA from both ends. By comparing these two sequences (e.g., in BIOEDIT, [86]), the quality of the sequences can be assessed best. If these complementary sequences are not available, it is strongly recommended to visually check the peaks in the electropherograms for any abnormalities, along the whole sequence.
25. The nature of variation in noncoding regions of the chloroplast is variable, but some types can be differentiated [102]. Sequence variation of chloroplast spacers and introns often consists of very small changes [103]. Insertions of single or only very few nucleotides are abundant. Substitutions (true single nucleotide polymorphism, SNPs) are present, but these alone would not give a robust picture of sequence variation. The very short insertion/deletions (indels) often manifest themselves in the mononucleotide repeats. Another form of variation typical for chloroplasts is short inversions [102, 103]. Chloroplast genes often terminate in stem-loop structures in their 3′ untranslated regions, or such stem-loops are present in introns. Apparently, the loops in these structures occasionally flip position [102], similar to the inversion in mitochondrial DNA identified by Dumolin-Lapegue et al. [104]. These peculiarities make it difficult to distinguish sequencing errors or

ambiguities from true variation. We suspect that such ambiguities have found their way into databases and the literature. In a recent comparison of two completely sequenced date palm chloroplast genomes with GenBank entries for short fragments, very few polymorphisms between the two complete chloroplasts contrasted with a high incidence of polymorphisms in (often unpublished) GenBank entries of short chloroplast fragments [75].

26. Borsch and Quandt [103] have recently reviewed the use of noncoding regions of the chloroplast genome to a remarkable depth. They list several widely used regions that make comparisons among studies feasible: the introns in *petD*, *rpl16*, *rps16* and *trnK*, and the spacers *trnS-trnG*, *trnT-trnF*, *psbA-trnH*, and *atpB-rbcL*. Apart from microsatellites and stem-loop structures which pose the issue of homoplasy (back-mutation), mutational hotspots exist also elsewhere and need to be treated with great caution in phylogenetic studies, or excluded from the data matrix altogether [103]. Along similar lines, Ravi et al. [105] found that their complete sequence of the *Morus* chloroplast moved phylogenetically closer to *Cucumis* and *Lotus* if only coding regions were considered but closer to *Eucalyptus* if noncoding regions were included in the analysis. Watts et al. [32] have suggested conserved primer pairs for chloroplast group II intron loops that are highly variable but less subject to the problems listed above (for the *rpl16*, *petD*, *atpF*, *petB*, and *ndhA* introns). Ochoterena [106] has reviewed methods of sequence alignment for noncoding DNA, which must take into consideration that different evolutionary constraints apply.
27. Comparisons of gene sequences should also take into account RNA editing, which seems to be frequent at least in some plant species (e.g., *Hevea brasiliensis*, 51 RNA editing sites, [107]). In *Hevea*, 48 of these sites are in gene coding sequences, but three additional ones are in introns. Schmitz-Linneweber et al. [108] report that *Atropa belladonna* and *Nicotiana tabacum*, two members of the Solanaceae, show a high similarity of their chloroplast genomes' coding regions, but sites subject to RNA editing differ remarkably. RNA editing is guided by factors imported from the nucleus. Therefore, nuclear genomic divergence during speciation may drive chloroplast divergence at those sites.
28. Concerning barcoding, it may be necessary to take a pragmatic view and assess the limits to what the proposed general barcode can be used but on the other hand develop or suggest other or additional regions for specific questions, i.e., for specific groups of taxa. For example, Schroeder et al. [109] have compared generally recommended barcoding primers

(23 published primer pairs) and newly designed primers specific for the genus *Populus* for their discriminating power between species in the genus. Their general conclusion was that while not all the new primers yielded amplification products for all species, these more specific primers designed on the basis of the complete sequence of the *Populus trichocarpa* chloroplast genome [95] were much more reliable, and some of the intergenic regions had high levels of polymorphism (e.g. *trnG-psbK*, or *psbK-psbI*).

29. There are a variety of methods available for detection of PCR products on polyacrylamide gels. Radioactive detection involves the incorporation of radioactively labeled nucleotides or primers during PCR. Nonradioactive methods are preferred by many because of health and safety issues. Silver staining and fluorescence dyes are the choices in this respect. Although many different protocols have been suggested, silver staining is known to be time-consuming because it involves many steps, and it requires some level of skill. Ethidium bromide (excitation 310 and 526 nm, emission 605 nm) is a cheap fluorescent dye routinely used for agarose gels, but the sensitivity of this method leaves a lot to be desired [110, 111]. More sensitive fluorescence dyes have become available [19, 111]: SYBR Green I and II (excitation 497 nm, emission 520 nm) and SYBR Gold (excitation 310 and 495 nm, emission 537 nm)—the latter one is recommended also for single-stranded DNA (ssDNA) analysis. They are reported to be less mutagenic (Amersham Biosciences, Fluorescence Imaging fact sheet).
30. To investigate the effects of different filters on image quality, photos can also be taken through an SYBR Green/Gold gel stain photographic filter (Molecular Probes, Oregon, USA) and a UV-blocking filter (UV 390, O–Hase, Hama, Germany).
31. Both silver staining and SYBR Gold treatment usually produce sharp and clearly visible products. Even differences in allele sizes about 1 or 2 bp can be distinguished with both staining methods. However, on silver stained gels, the intensity of bands seems to be higher especially for bands of higher molecular weight, which appear stronger. 1 μL of standard PCR product is readily detectable on the gel after silver staining in our laboratory setup, whereas the optimum for SYBR Gold is between 2 and 4 μL. Silver staining is more sensitive for overloading and higher concentrations of loaded DNA, resulting in smears. A readable banding pattern of good quality is usually also obtained after subsequent treatment of an SYBR Gold stained gel with the silver staining procedure, which may combine the advantages of both methods (but at least double the workload).

32. Taking photos through three different filters and their combinations, we experienced that the use of appropriate filters for documentation for the fluorescent staining is important. Fragments are usually clearly visible on a gel when photographed with 600 nm band pass filter or the specially designed SYBR Green/Gold gel stain photographic filter. The narrow window of transmittance for the 300 nm region in the SYBR Green/Gold filter causes that the tubes of the transilluminator become visible on the picture, making typing more inconvenient. Using the SYBR Green/Gold filter in combination with the standard photographic UV-blocking filter (390 nm) does not eliminate this effect. As can be expected, using the UV-blocking filter alone results in a poor image with very low band intensities.
33. Comparison of required reagents and equipment, relative costs, handling, and waste disposal for both presented methods reveal that, at least in our experience, the SYBR Gold process offers much advantage over silver or let alone ethidium bromide staining. For detection of DNA stained with ethidium bromide and SYBR Gold fluorescent dyes, the same visualization equipment can be used. Plastic films for attaching gels instead of covering glass plates each time with bind and repeal silane would ease handling for silver staining. Such gels can also be permanently stored not only in the digital but also in the physical form. This is not possible with SYBR Gold staining because current films produce additional background fluorescence when illuminated with UV light boxes. Without such a plastic support, fresh covering of glasses especially with bind silane is necessary every four times with SYBR Gold, but more often with silver staining. SYBR waste disposal is quick and efficient, while we find handling silver waste more laborious and costly. We find that if chemicals can be bought in higher quantities, and with higher gel throughput, costs tend to decrease. From a single pack of SYBR Gold, we can stain 100 or more gels. The main cost component for silver staining is silver nitrate. As a minimum amount of staining solution (1 L in our case) has to be prepared for immersing the gel, staining a single gel is much more expensive. These costs come down when the solution is used for several gels. However, silver staining solutions have a shorter life time (1 week in our laboratory).
34. The enzyme cannot operate in completely detergent-free buffer.
35. At this step, both PCR products (sample and standard) have to be at fairly equal concentrations. Adjust if necessary.
36. This consists of known fragments, and the resulting dHPLC pattern at a non-denaturing temperature can be checked for

resolution of the fragments, which will give hints on column performance.

37. To achieve optimal resolution during extended runs (e.g., a 96 or 384 sample plate), flushing intervals become important. Flushing with buffer B alone, or with acetonitrile at 75 %, should be programmed every 20 samples to avoid peak "movement" or shoulders within the peaks.
38. For instance, only identify clear single peaks and doublets, triplets, and so on in the first round and concentrate on retention time delays or unclear peak splits in the second round.
39. Precisely determining these small length differences is not easy, because they are often below 1 % of the total length of the PCR fragment. While within an experiment, this is not so much an issue, it becomes more problematic when repeating analyses on different automatic DNA sequencer models or in different years in the same laboratory.
40. Single nucleotide sequencing as suggested by [112] may be particularly useful for this type of variation: In this method, just a single dideoxynucleotide triphosphate is introduced during the sequencing reaction. This makes the interpretation of the resulting pattern much easier, if only the microsatellite tract is of interest.
41. In an extension of our recent study [72], a repeat of approximately 15 bp was discovered in the *rpl16* intron of several *Populus* species. Polymorphisms among, but also within species, were apparent.
42. Vachon and Freeland [73] have compared large phylogeographic networks of haplotypes by either including or excluding tandem repeat (micro- and minisatellites) polymorphisms. Their conclusion is that such repeats should be excluded from large-scale studies, while they may provide valuable insight in more local (fine-scale) studies. Figure 2 shows one such example of insertions/deletions of variable sizes in our recent work in *Gentianella*—some of these insertions seem to be tandem repeats (Jahn and Heinze, unpublished). Indeed, studying chloroplast microsatellites is widespread for intraspecific phylogeographic work.
43. Complete chloroplast genome sequences often reveal a surprising picture on variation, as exemplified by the recent examples of Korean pine (*Pinus koraiensis*, [113]) and pepper (*Capsicum annuum*, [114]). Several complete chloroplasts from the Solanaceae family were known before, but pepper was found to contain large indels, and tandem repeats were particularly frequent in pepper.

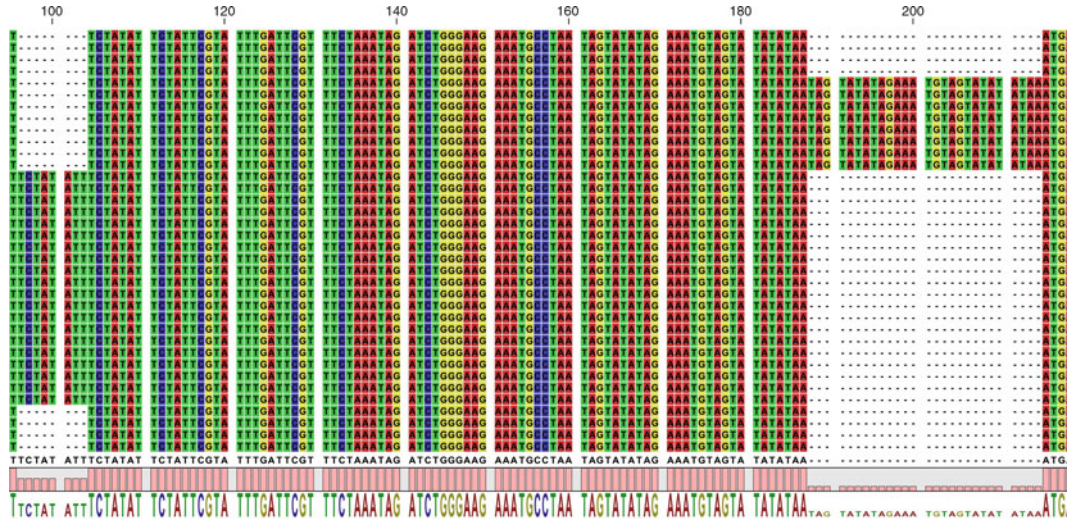


Fig. 2 Schematic alignment of *Gentianella* chloroplast DNA sequence showing presence of small insertions/deletions that are often tandemly repeated sequence tracts

44. Morris et al. [91] have produced random whole-genome shotgun sequences from 24 individuals of switchgrass (*Panicum virgatum*) at low coverage. As chloroplast DNA was abundant in their DNA preparations, whole chloroplast DNA sequences could be reconstructed, and polymorphisms identified. The individuals in their study were labeled by using "barcode" sequences in the next-generation sequencing run. We have recently pooled "leftover" chloroplast PCR products from a wide range of studies and subjected the pool to Illumina sequencing (Heinze and Jahn, unpublished). Simple bioinformatics analysis allowed us to identify the molecular causes of many of the PCR-RFLP and dHPLC pattern differences that we have observed previously from the same samples. Once these polymorphisms and others that could not be detected with the previous approaches are known, it is straightforward to design the best assay that allows the analysis of these polymorphisms in large sets of samples with more traditional methods.

Acknowledgements

We thank Irena Nanista, Bianca Widmar, Ingrid Gerstl, and Renate Slunsky for technical assistance in the laboratory over the years. Funding for projects was by the European Commission—research project RAP—Realising Ash' Potential, QLK5-2001-00631; Jubiläumsfond der Stadt Wien (Migrationsforschung); and the Austrian Science Funds (P 22716-B16 *Gentianella*).

References

1. Clegg MT, Zurawski G (1992) Chloroplast DNA and the study of plant phylogeny: present status and future prospects. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Chapman & Hall, New York, pp 1–13
2. Bovenberg WA, Kool AJ, Nijkamp HJJ (1981) Isolation, characterization and restriction endonuclease mapping of the *Petunia hybrida* chloroplast DNA. Nucleic Acids Res 9:503–517
3. Palmer JD et al (1983) Chloroplast DNA evolution and the origin of amphidiploid *Brassica* species. Theor Appl Genet 65:181–189
4. Clark EM, Izhar S, Hanson MR (1985) Independent segregation of the plastid genome and cytoplasmic male sterility in *Petunia* somatic hybrids. Mol Gen Genet 199:440–445
5. Shinozaki K et al (1986) The complete nucleotide sequence of the tobacco chloroplast genome. Plant Mol Biol Rep 4:111–147
6. Sugiura M (1987) Structure and function of the tobacco chloroplast genome. J Plant Res 100:407–436
7. Sugiura M (1992) The chloroplast genome. Plant Mol Biol 19:149–168
8. Olmstead RG, Sweere JA, Wolfe KH (1993) Ninety extra nucleotide in *ndhF* gene of tobacco chloroplast DNA: a summary of revisions to the 1986 genome sequence. Plant Mol Biol 22:1191–1193
9. Wakasugi T et al (1998) Updated gene map of tobacco chloroplast DNA. Plant Mol Biol Rep 16:231–241
10. Sugiura M (2003) History of chloroplast genomics. Photosynth Res 76:371–377
11. Yukawa M, Tsudzuki T, Sugiura M (2005) The 2005 version of the chloroplast DNA sequence from tobacco (*Nicotiana tabacum*). Plant Mol Biol Rep 23:359–365
12. Bock R (2007) Structure, function, and inheritance of plastid genomes. In: Bock R (ed) Cell and molecular biology of plastids. Springer, Berlin, pp 29–63
13. Ravi V et al (2008) An update on chloroplast genomes. Plant Syst Evol 271:101–122
14. Taberlet P et al (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol Biol 17: 1105–1109
15. Till BJ (2006) A protocol for TILLING and Ecotilling in plants and animals. Nat Protoc 1:2465–2477
16. Hauser MT et al (1998) Generation of codominant PCR-based markers by duplex analysis on high resolution gels. Plant J 16:117–125
17. Hoelzel AR (1998) Molecular genetic analysis of populations. IRL Press at Oxford University Press, Oxford
18. Hoelzel AR, Green A (1998) PCR protocols and population analysis by direct DNA sequencing and PCR-based DNA fingerprinting. In: Hoelzel AR (ed) Molecular genetic analysis of populations. IRL Press at Oxford University Press, Oxford, pp 201–235
19. Sambrook J, Russell DW (2001) Molecular cloning, 3rd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
20. Liston A (1992) Variation in the chloroplast genes rpoC1 and rpoC2 of the genus *Astralagus* (Fabaceae): evidence from restriction site mapping of a PCR-amplified fragment. Am J Bot 79:953–961
21. Zeltz P et al (1993) Editing of the chloroplast rpoB transcript is independent of chloroplast translation and shows different patterns in barley and maize. EMBO J 12: 4291–4296
22. Steele KP, Vilgalys R (1994) Phylogenetic analysis of Polemoniaceae using nucleotide sequences of the plastid gene matK. Syst Bot 19:126–142
23. Graham SW, Olmstead RG (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am J Bot 87:1712–1730
24. Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4:129–131
25. Dumolin-Lapegue S, Pemonge M-H, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol 6:393–398
26. Small RL et al (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot 85:1301–1315
27. Grivet D et al (2001) Genome walking with consensus primers: application to the large single copy region of chloroplast DNA. Mol Ecol Notes 1:345–349
28. Dhingra A, Folta KM (2005) ASAP: amplification, sequencing & annotation of plastomes. BMC Genomics 6:176
29. Shaw J et al (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am J Bot 92:142–166
30. Shaw J et al (2007) Comparison of whole chloroplast genome sequences to choose

noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am J Bot 94:275–288
31. Heinze B (2007) A database of PCR primers for the chloroplast genomes of higher plants. Plant Methods 3:4
32. Watts CD et al (2008) The D4 set: primers that target highly variable intron loops in plant chloroplast genomes. Mol Ecol Resour 8:1344–1347
33. Borsch T et al (2009) The petD group II intron as a species level marker: utility for tree inference and species identification in the diverse genus *Campanula* (Campanulaceae). Willdenowia 39:7–33
34. Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9:673–690
35. Lin C-P et al (2010) Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biol Evol 2:504–517
36. Scarcelli N et al (2011) A set of 100 chloroplast DNA primer pairs to study population genetics and phylogeny in monocotyledons. PLoS ONE 6:e19954
37. Prince LM (2011) Plastid primers for phylogenetics. Rancho Santa Ana Botanic Garden, Claremont, California, USA. https://sites.google.com/site/plastidprimersforphylogenetics/. Accessed 2 Apr 2012
38. Haider N, Wilkinson M (2011) A set of plastid DNA-specific universal primers for flowering plants. Russ J Genet 47:1066–1077
39. Haider N (2011) Chloroplast-specific universal primers and their uses in plant studies. Biol Plant 55:225–236
40. Ebert D, Peakall R (2009) A new set of universal de novo sequencing primers for extensive coverage of noncoding chloroplast DNA: new opportunities for phylogenetic studies and cpSSR discovery. Mol Ecol Resour 9:777–783
41. Nakamura I et al (1997) A proposal for identifying the short ID sequence which addresses the plastid subtype of higher plants. Breed Sci 47:385–388/394
42. Brunner I et al (2001) Molecular identification of fine roots of trees from the Alps: reliable and fast DNA extraction and PCR-RFLP analyses of plastid DNA. Mol Ecol 10: 2079–2087
43. Borsch T et al (2003) Non-coding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol 16:558–576
44. CBOL Plant Working Group (2009) A DNA barcode for land plants. Proc Natl Acad Sci USA 106:12794–12797
45. Chase MW et al (2005) Land plants and DNA barcodes: short-term and long-term goals. Phil Trans R Soc B-Biol Sci 360:1889–1895
46. Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6:e19254
47. Presting GG (2006) Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. Can J Bot 84:1434–1443
48. Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. Can J Bot 84:335–341
49. Chase MW et al (2007) A proposal for a standardised protocol to barcode all land plants. Taxon 56:295–299
50. Fazekas AJ et al (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. Public Libr Sci 3:e2802
51. Lahaye R et al (2008) DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci USA 105:2923–2928
52. Ford CS et al (2009) Selection of candidate coding DNA barcoding regions for use on land plants. Bot J Linn Soc 159:1–11
53. Hollingsworth ML et al (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour 9:439–457
54. Taberlet P et al (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35:e14
55. Spooner DM (2009) DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. Am J Bot 96:1177–1189
56. Fazekas AJ et al (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Mol Ecol Resour 9:130–139
57. Ran J-H et al (2010) A test of seven candidate barcode regions from the plastome in *Picea* (Pinaceae). J Integr Plant Biol 52:1109–1126
58. Arca M et al (2012) Deciduous trees and the application of universal DNA barcodes: a case study on the circumpolar *Fraxinus*. PLoS ONE 7:e34089
59. Tsumura Y et al (1995) Molecular phylogeny of conifers using RFLP analysis of PCR-amplified specific chloroplast genes. Theor Appl Genet 91:1222–1236
60. Parks M, Liston A, Cronn R (2011) Newly developed primers for complete ycf1 amplification in *Pinus* (Pinaceae) chloroplasts with possible family-wide utility. Am J Bot 98: e185–e188
61. Oefner PJ, Underhill PA (1998) Detection of nucleic acid heteroduplex molecules by denaturing high-performance liquid

chromatography and methods for comparative sequencing., US Patent 5,795,976 (18 August 1998)
62. Gielly L, Taberlet P (1994) The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. Mol Biol Evol 11:769–777
63. Devey DS, Chase MW, Clarkson JJ (2009) A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. Taxon 58:7–15
64. Powell W et al (1995) Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. Curr Biol 5:1023–1029
65. Weising K, Gardner R (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42:9–19
66. Vendramin GG et al (1996) A set of primers for the amplification of 20 chloroplast microsatellites in Pinaceae. Mol Ecol 5:595–598
67. Provan J et al (2004) Universal primers for the amplification of chloroplast microsatellites in grasses (Poaceae). Mol Ecol Notes 4:262–264
68. Chung S-M, Staub JE (2003) The development and evaluation of consensus chloroplast primer pairs that possess highly variable sequence regions in a diverse array of plant taxa. Theor Appl Genet 107:757–767
69. Blasko K et al (1988) Variation in copy number of a 24-base pair tandem repeat in the chloroplast DNA of *Oenothera hookeri* strain Johansen. Curr Genet 14:287–292
70. Hipkins VD et al (1995) A mutation hotspot in the chloroplast genome of a conifer (Douglas-fir: *Pseudotsuga*) is caused by variability in the number of direct repeats derived from a partially duplicated tRNA gene. Curr Genet 27:572–579
71. Cafasso D et al (2001) Characterization of a minisatellite repeat locus in the chloroplast genome of *Orchis palustris* (Orchidaceae). Curr Genet 39:394–398
72. Fussi B, Lexer C, Heinze B (2010) Phylogeography of *Populus alba* (L.) and *Populus tremula* (L.) in Central Europe: secondary contact and hybridisation during recolonisation from disconnected refugia. Tree Genet Genomes 6:439–450
73. Vachon N, Freeland JR (2011) Phylogeographic inferences from chloroplast DNA: quantifying the effects of mutations in repetitive and non-repetitive sequences. Mol Ecol Resour 11:279–285
74. Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol 7:84
75. Khan A et al (2012) The chloroplast genome sequence of date palm (*Phoenix dactylifera* L. cv. 'Aseel'). Plant Mol Biol Rep 30:666–678
76. Antoniw J (1995) A new method for designing PCR primers specific for groups of sequences and its application to plant viruses. Mol Biotechnol 4:111–119
77. Noh EW, Lee JS (1997) Molecular genetic analysis of *Populus* chloroplast DNA. In: Klopfenstein NB, Chun YW, Kim MS, Ahuja MR (eds) Micropropagation, genetic engineering, and molecular biology of *Populus*. US Dept. of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, pp 143–149
78. Heinze B (1998) PCR-based chloroplast DNA assays for the identification of native *Populus nigra* and introduced poplar hybrids in Europe. Forest Genet 5:31–38
79. Hamza-Babiker N et al (2009) Chloroplast DNA identification of eight closely related European *Salix* species. Austrian J Forest Sci 126:175–193
80. Jarman SN (2004) Amplicon: software for designing PCR primers on aligned DNA sequences. Bioinformatics 20:1644–1645
81. Jabado OJ et al (2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. Nucleic Acids Res 34: 6605–6611
82. Boyce R, Chilana P, Rose TM (2009) iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucleic Acids Res 37:W222–W228
83. Contreras-Moreira B et al (2009) primers4clades: a web server that uses phylogenetic trees to design lineage-specific PCR primers for metagenomic and diversity studies. Nucleic Acids Res 37:W95–W100
84. Lexer C et al (2005) Barrier to gene flow between two ecologically divergent *Populus* species, *P. alba* (white poplar) and *P. tremula* (European aspen): the role of ecology and life history in gene introgression. Mol Ecol 14: 1045–1057
85. Turkec A, Sayar M, Heinze B (2006) Identification of sweet cherry cultivars (*Prunus avium* L.) and analysis of their genetic relationships by chloroplast sequence-characterised amplified regions (cpSCAR). Genet Resour Crop Evol 53:1635–1641
86. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98
87. Bassam BJ, Caetano-Anollés G (1993) Silver staining of DNA in polyacrylamide gels. Appl Biochem Biotechnol 42:181–188

88. Flores-Renteria L, Whipple AV (2011) A new approach to improve the scoring of mononucleotide microsatellite loci. Am J Bot 98:e51–e53
89. Mariotti R et al (2010) Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. BMC Plant Biol 10:1–13
90. Besnard G et al (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. BMC Plant Biol 11:80
91. Morris GP, Grabowski PP, Borevitz JO (2011) Genomic diversity in switchgrass (*Panicum virgatum*): from the continental scale to a dune landscape. Mol Ecol 20:4938–4952
92. Vroh Bi I (1996) Improved RAPD amplification of recalcitrant plant DNA by the use of activated charcoal during DNA extraction. Plant Breed 115:205–206
93. Lutz K et al (2011) Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. BMC Biotechnol 11:1–9
94. Nock CJ et al (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9:328–333
95. Tuskan GA et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604
96. Lin H, Walker MA (1997) Extracting DNA from cambium tissue for analysis of grape rootstocks. Hortscience 32:1264–1266
97. Colpaert N et al (2005) Sampling tissue for DNA analysis of trees: trunk cambium as an alternative to canopy leaves. Silvae Genetica 54:265–269
98. Tibbits J et al (2006) A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees. Plant Mol Biol Rep 24:81–91
99. Ayliffe MA, Timmis JN (1992) Tobacco nuclear DNA contains long tracts of homology to chloroplast DNA. Theor Appl Genet 85:229–238
100. Ayliffe MA, Scott NS, Timmis JN (1998) Analysis of plastid DNA-like sequences within the nuclear genomes of higher plants. Mol Biol Evol 15:738–745
101. Huang CY, Ayliffe MA, Timmis JN (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422:72–76
102. Kim K-J, Lee H-L (2005) Widespread occurrence of small inversions in the chloroplast genomes of land plants. Mol Cells 19:104–113
103. Borsch T, Quandt D (2009) Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. Plant Syst Evol 282: 169–199
104. Dumolin-Lapegue S, Pemonge MH, Petit RJ (1998) Association between chloroplast and mitochondrial lineages in oaks. Mol Biol Evol 15:1321–1331
105. Ravi V et al (2006) The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis. Tree Genet Genomes 3:49–59
106. Ochoterena H (2009) Homology in coding and non-coding DNA sequences: a parsimony perspective. Plant Syst Evol 282:151–168
107. Tangphatsornruang S et al (2011) Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. Gene 475:104–112
108. Schmitz-Linneweber C et al (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. Mol Biol Evol 19:1602–1612
109. Schroeder H, Hoeltken AM, Fladung M (2011) Differentiation of *Populus* species using chloroplast single nucleotide polymorphism (SNP) markers - essential for comprehensible and reliable poplar breeding. Plant Biol 14:374–381
110. Sanguinetti CJ, Dias Neto E, Simpson AJG (1994) Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. Biotechniques 17:915–919
111. Rahman MH, Jaquish B, Khasa PD (2000) Optimization of PCR protocol in microsatellite analysis with silver and SYBR® stains. Plant Mol Biol Rep 18:339–348
112. Guicking D et al (2008) Single nucleotide sequence analysis: a cost- and time-effective protocol for the analysis of microsatellite- and indel-rich chloroplast DNA regions. Mol Ecol Resour 8:62–65
113. Noh EW et al (2011) Plastid genes psaM and ndhB are differentially degraded between hard and soft pines. In: 4th IUFRO conference on the breeding and genetic resources of five-needle pines, Book of Abstracts. Institute of Monitoring of Climatic and Ecological Systems SB RAS (IMCES), Tomsk, Russia, p 33
114. Jo YD et al (2011) Complete sequencing and comparative analyses of the pepper (*Capsicum annuum* L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome. Plant Cell Rep 30:217–229

Chapter 6

Mitochondrial Genome and Plant Taxonomy

Jérôme Duminil

Abstract

The lability in size, structure, and sequence content of mitochondrial genome (mtDNA) across plant species has sharply limited its use in taxonomic studies. However, due to the new opportunities offered by the availability of complete mtDNA sequence in plant species and the subsequent development of universal primers, the number of mtDNA-based molecular studies has recently increased. Historically, universal primers have enabled to characterize mtDNA polymorphism mainly by the RFLP technique. This methodology has been progressively replaced by Sanger DNA sequencing, which actually provides the full phylogenetic information content of a DNA fragment (single nucleotide, insertion/deletion, and single sequence repeat length polymorphism). This chapter presents a sequencing working protocol to be routinely used in mtDNA-based phylogenetic studies.

Key words Cytoplasmic DNA, Organelle DNA, Botany, Phylogeny, Plants, Sequences, Genetic diversity, Polymorphism

1 Introduction

The mitochondrial genome (mtDNA) originated from a eubacterial ancestor. More specifically, it is now widely accepted that the mitochondria originated from a single endosymbiotic event which involved a α-proteobacteria-like organism and a common cellular ancestor of eukaryotes [1]. This symbiotic relationship between a primitive eukaryote nucleus and an aerobic bacteria—the future mitochondria—has enabled the eukaryote to evolve an aerobic lifestyle. In relation with this new endosymbiotic habit, the "resident" mitochondrial genome has undergone a reductive evolution, characterized among other things by a loss of coding capacity [2]. The gene content reduction of mitochondrial genomes has been primarily attributable to either gene loss or mitochondria-to-nucleus gene transfers [3]. This process has been interpreted as a consequence of deleterious accumulation in organelle genomes [2] and as a necessity for multicellular organisms to keep the function originally coded by organelle genomes. Gene transfer from the mitochondria to the nucleus has been demonstrated to be an ongoing

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_6,

process in plants [4], which explains that the mitochondrial gene content is heterogeneous across species [5, 6] and the difficulty to develop universal primers across species.

In sharp contrast to the relative small and homogeneous size of animal mtDNA (usually between 16 and 20 kb; [7]) and fungal mtDNA (between 19 and 100 kb; [8]), plant mtDNA is large and variable in size (between 104 kb, in the moss *Anomodon rugelii*, and 11.3 Mb, in the angiosperm *Silene conica*; [9]). Importantly, the "genome size" of the angiosperm mitochondrion refers to the size of the "master cycle DNA"—"which is a presumptive circular molecule consisting of all the DNA sequences present at substantial stoichiometry in the mitochondrion" [10]—the plasmid-like molecules, and the substoichiometric DNA molecules being excluded. Angiosperm mitochondrial genome size is not related to gene content [9]. MtDNA size variation among species is mainly related to differences in the size of intergenic regions (introns, intergenes, repeated sequences and alien sequences of chloroplast, and nuclear origin; [11]). Plastid-derived and nuclear-derived nucleotide sequences represent, respectively, from 1.6 to 8.8 % and from 0.1 to 13.4 % of the mtDNA [12]. However, a recent comparative study has demonstrated that genome size variation within the *Silene* genus (complete mtDNA sequences are now available for four species of the genus, *see* Table 1) might be related to changes in recombinational processes [9].

Due to the presence of numerous repeated regions and to the coexistence of more than one type of mitochondrial genome in a cell (heteroplasmy), recombinations are frequent within the mtDNA and gene arrangement in higher plants varies enormously [13]. Besides the large size, recombination activity is the most distinctive feature of these genomes [14]. Gene arrangement of mtDNA in higher plants varies enormously due to the presence of repeated regions, source of recombination within and between mtDNA genomes [13]. Fortunately, mtDNA coding sequences are highly conserved, facilitating the identification of conserved regions within which universal primers can be defined [15, 16].

In opposition to animals, plant mtDNA evolves very slowly. Comparing coding sequence silent (synonymous) substitution rates among plant genomes (nrDNA, cpDNA, mtDNA), Wolfe et al. [17] have demonstrated that the mtDNA evolves three times slower ($0.2–1.1 \times 10^{-9}$ substitutions per synonymous site per year) than the cpDNA ($1.1–2.9 \times 10^{-9}$ substitutions per synonymous site per year), which in turn evolves two times slower than the nrDNA (up to 31.5×10^{-9} substitutions per synonymous site per year). These results were further confirmed by Gaut et al. [18] based on the comparison of genes from all three genomes between maize and rice. Interestingly, as outlined by Muse [19], the similarity obtained between Wolfe and Gaut studies, albeit different levels of evolutionary divergence were addressed, might indicate that plant nucleotide substitution features have been constant along higher

Table 1
List of available Viridiplantae species mtDNA complete sequence

Species	Viridiplantae clade	Accession number[a]	Year of publication	Genome size (pb)	Protein	tRNA
Scenedesmus obliquus	Chlorophyta	NC_002254	1990	42781	20	33
Chlamydomonas reinhardtii	Chlorophyta	NC_001638	1994	15758	8	17
Prototheca wickerhamii	Chlorophyta	NC_001613	1994	55328	36	29
Chlamydomonas eugametos	Chlorophyta	NC_001872	1998	22897	14	13
Pedinomonas minor	Chlorophyta	NC_000892	1999	25137	11	12
Pseudendoclonium akinetum	Chlorophyta	NC_005926	2004	95880	72	27
Ostreococcus tauri	Chlorophyta	NC_008290	2005	44237	43	35
Nephroselmis olivacea	Chlorophyta	NC_008239	2006	45223	40	30
Oltmannsiellopsis viridis	Chlorophyta	NC_008256	2006	56761	36	27
Polytomella capuana	Chlorophyta	NC_010357	2008	12998	7	13
Dunaliella salina	Chlorophyta	NC_012930	2009	28331	9	12
Micromonas sp.	Chlorophyta	NC_012643	2009	47425	39	40
Polytomella sp.	Chlorophyta	NC_013472, NC_016918	2009	16083	7	13
Pycnococcus provasolii	Chlorophyta	NC_013935	2010	24321	18	18
Coccomyxa sp.	Chlorophyta	NC_015316	2011	65497	31	29
Polytomella parva	Chlorophyta	NC_016916, NC_016917	2012	16153	7	13
Anomodon rugelii	Streptophyta	NC_016121	2011	104239	46	27
Arabidopsis thaliana	Streptophyta	NC_001284	1997	366924	117	24
Beta vulgaris	Streptophyta	NC_015099	2010	364950	150	29
Beta vulgaris	Streptophyta	NC_002511	2000	368801	140	31
Beta macrocarpa	Streptophyta	NC_015994	2011	385220	156	31
Boea hygrometrica	Streptophyta	NC_016741	2011	510519	32	31
Brassica rapa	Streptophyta	NC_016125	2011	219747	78	21
Brassica juncea	Streptophyta	NC_016123	2011	219766	78	21
Brassica napus	Streptophyta	NC_008285	2003	221853	79	20
Brassica carinata	Streptophyta	NC_016120	2011	232241	69	20

(continued)

Table 1
(continued)

Species	Viridiplantae clade	Accession number[a]	Year of publication	Genome size (pb)	Protein	tRNA
Brassica oleracea	Streptophyta	NC_016118	2011	360271	77	21
Carica papaya	Streptophyta	NC_012116	2010	476890	39	22
Chaetosphaeridium globosum	Streptophyta	NC_004118	2002	56574	46	31
Chara vulgaris	Streptophyta	NC_005255	2003	67737	46	30
Chlorokybus atmophyticus	Streptophyta	NC_009630	2007	201763	58	31
Citrullus lanatus	Streptophyta	NC_014043	2010	379236	39	21
Cucumis sativus	Streptophyta	NC_016005, NC_016004, NC_016006	2011	1684592	37	26
Cucurbita pepo	Streptophyta	NC_014050	2010	982833	38	16
Cycas taitungensis	Streptophyta	NC_010303	2008	414903	39	29
Lotus japonicus	Streptophyta	NC_016743	2011	380861	34	23
Marchantia polymorpha	Streptophyta	NC_001660	1992	186609	76	32
Megaceros aenigmaticus	Streptophyta	NC_012651	2009	184908	48	21
Mesostigma viride	Streptophyta	NC_008240	2002	42424	41	29
Millettia pinnata	Streptophyta	NC_016742	2011	425718	37	27
Nicotiana tabacum	Streptophyta	NC_006581	2005	430597	156	27
Oryza sativa	Streptophyta	NC_011033	2002	490520	53	25
Oryza sativa	Streptophyta	NC_007886	2006	491515	54	39
Oryza rufipogon	Streptophyta	NC_013816	2010	559045	41	18
Phaeoceros laevis	Streptophyta	NC_013765	2010	209482	38	23
Phoenix dactylifera	Streptophyta	NC_016740	2012	715001	43	21
Physcomitrella patens	Streptophyta	NC_007945	2007	105340	42	27
Pleurozia purpurea	Streptophyta	NC_013444	2009	168526	69	31
Ricinus communis	Streptophyta	NC_015141	2011	502773	37	23
Silene latifolia	Streptophyta	NC_014487	2010	253413	30	13
Silene vulgaris	Streptophyta	NC_016406, NC_016170, NC_016402, NC_016415	2012	427138	11	3
Silene noctiflora	Streptophyta	NC_016393[b]	2012	6727869	29	20

(continued)

Table 1
(continued)

Species	Viridiplantae clade	Accession number[a]	Year of publication	Genome size (pb)	Protein	tRNA
Silene conica	Streptophyta	NC_016249[b]	2012	11318806	32	18
Sorghum bicolor	Streptophyta	NC_008360	2006	468628	32	21
Treubia lacunosa	Streptophyta	NC_016122	2011	151983	69	30
Tripsacum dactyloides	Streptophyta	NC_008362	2006	704100	33	21
Triticum aestivum	Streptophyta	NC_007579	2005	452528	39	34
Vigna radiata	Streptophyta	NC_015121	2011	401262	32	19
Vitis vinifera	Streptophyta	NC_012119	2009	773279	74	34
Zea luxurians	Streptophyta	NC_008333	2006	539368	32	20
Zea mays	Streptophyta	NC_007982	2004	569630	163	33
Zea perennis	Streptophyta	NC_008331	2006	570354	32	20
Zea mays	Streptophyta	NC_008332	2006	680603	42	19

[a]Using this accession number, you can access the complete mtDNA sequence at http://www.ncbi.nlm.nih.gov/nuccore
[b]This accession number correspond to the chromosome 1; other accession numbers corresponding to the other chromosomes are available at http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=33090&opt=organelle

plant evolution. Though this low molecular evolutionary rate appeared to concern most of plant species, some exceptions were demonstrated (e.g., within *Pelargonium*, *Plantago*, *Silene*; [9, 20, 21]). Recently the generality of slow synonymous sequence evolution in mitochondrial genomes has been investigated across a large and taxonomically widely distributed set of seed plants [22]. According to this study, earlier findings were confirmed for roughly 80–90 % of the studied species, indicating that a surprising high number of taxa depart from this common pattern (presenting either an accelerated or a slower synonymous substitution rates). Moreover, Mower et al. [22] demonstrated that both patterns of faster and slower evolutionary rates can be found at different genes within the same species supporting the idea that all genes evolve independently of the others. Albeit this observation might be related to different artifacts [22], independent evidences for mutation rate variation among genes were acquired [23]. Therefore, the general idea remains that mtDNA evolves at a slow rate and that mtDNA polymorphism is very low within one single species and even between closely related species. This partly explains the limited use of mtDNA in phylogeography and phylogeny, though the demonstration of molecular rate heterogeneity within some plant

lineages [9, 22, 23] might support the idea that it is worth investigating for mtDNA polymorphism within a given species.

Mitochondrial genomes are generally maternally inherited in angiosperms, though some species have been shown to present either a paternal inheritance or a biparental inheritance [24]. In conifers, paternal inheritance has been demonstrated. The uniparental inheritance of organellar genomes (oDNA; either mitochondrial DNA or chloroplast DNA), together with slow molecular evolutionary rate, explains their success as molecular markers in phylogeography studies (reviewed in [25]). The mode of inheritance has been shown theoretically and experimentally to have a major effect on the estimation of the among-population genetic differentiation. Maternally inherited genomes generally experience more subdivisions than paternally or biparentally inherited ones [26]. Thus, in conifers, genetic structure is almost always larger at mtDNA markers than at cpDNA markers, while in angiosperms genetic structure is nearly similar at both markers as they are generally maternally inherited [26].

The first land plant complete mtDNA sequence was obtained for the liverwort *Marchantia polymorpha* [27]. The number of Viridiplantae species for whom the complete mtDNA sequence is now available is 63 (16 Chlorophyta, 47 Viridiplantae—of whom 33 Streptophyta—Table 1). Figure 1 represents a cladogram of the 33 Streptophyta species for whom the complete mtDNA sequence is available. In comparison, 241 species of Viridiplantae were completely sequenced for their cpDNA (April 2012, according to http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=2759&hopt=html). The number of plant species mtDNA and cpDNA complete sequences has recently dramatically increased (Fig. 2). Owing to the burst of high-throughput sequencing technologies, this trend might shortly accelerate at an amazing rate.

With the exception of the conifers, mtDNA markers were rarely used in plant phylogeographic studies. This can be explained by the difficulty to find mtDNA genomic variations. The exact level of intraspecific mtDNA polymorphism across plant species is difficult to estimate. Indeed, few studies mentioned that they have screened for mtDNA polymorphism and did not find anything [28]. Whereas the polymorphism content of different cpDNA markers has been accurately estimated across plant species providing some rules on the ideal candidate cpDNA loci to be used in phylogeography or phylogeny studies (e.g., [29]), such information is clearly lacking for mtDNA. Thus, the current trend is to test as many mtDNA loci as possible using available universal primers [15, 16, 30–33]. Universal primers are based on the conserved nature of the exonic sequences of mtDNA across species, enabling the identification of consensus regions within coding sequences. The comparison of full complete mtDNA sequences can unravel the presence of a conserved microsynteny in some mtDNA genomic regions, which can be used to define primer pairs as universal as

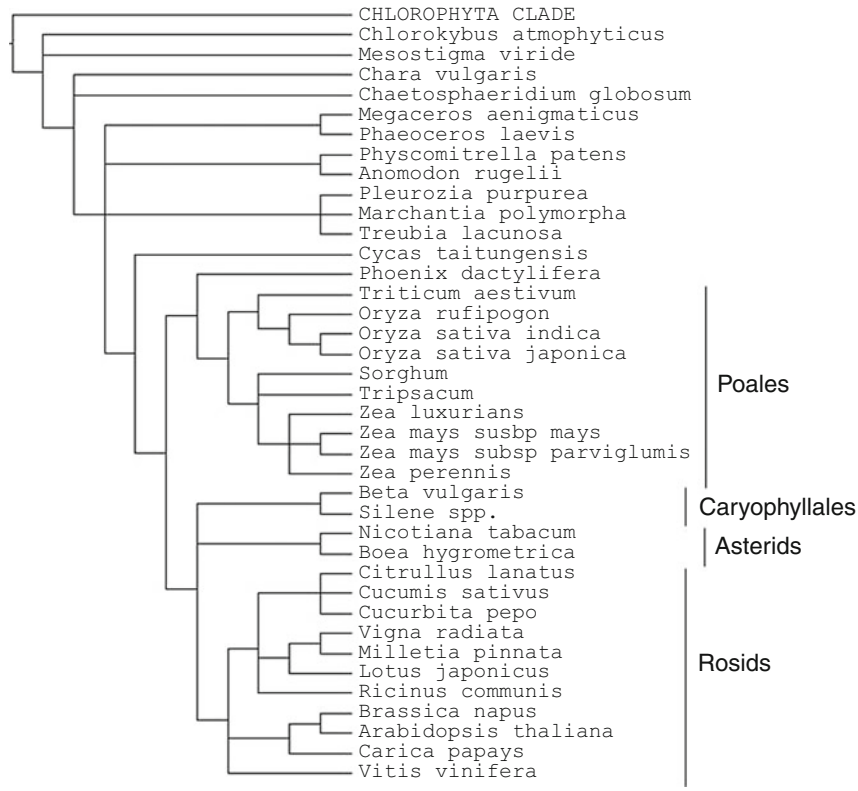


Fig. 1 Cladogram of the Streptophyta species for whom the complete mtDNA sequence is available (rooted on the Chlorophyta clade)

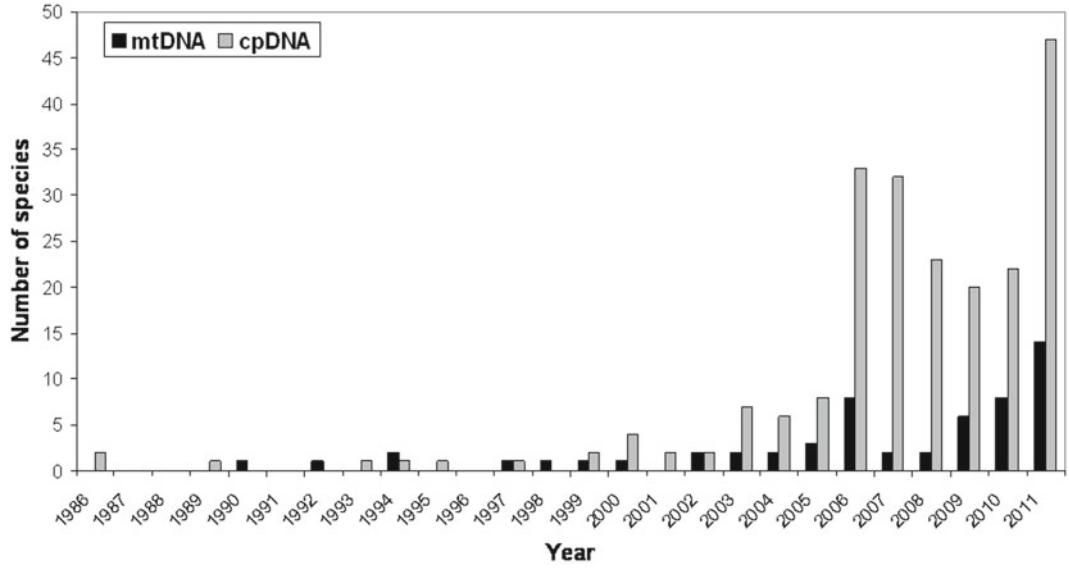


Fig. 2 Evolution of the number of plant species (Chlorophyta + Streptophyta) for which mtDNA and cpDNA complete sequences were obtained over time (1986–2011)

possible across plant species. The choice of the candidate loci depends on the taxonomic level that will be addressed by the phylogenetic study. At the lowest taxonomic level (intraspecies or inter-closely related species), the focus will be on mtDNA intergenic or intronic sequences. Instead, at higher taxonomic level (polymorphisms among a set of species in a phylogenetic framework), the focus will be on mtDNA coding sequences.

Historically, the screening of mtDNA polymorphism was performed using the PCR-RFLP [32, 34–39] or the RFLP-SSR technique [32, 35, 40]. MtDNA fragments were amplified using universal primers, and only the polymorphism corresponding to mutations at restriction enzymes cleaving DNA site was studied. Alternatively many studies made use of the polymorphism observed at minisatellite loci (VNTR, variable number of tandem repeat) [41–47]. Few authors screened the polymorphism of mtDNA-SSR [48]. To date, mtDNA polymorphism is predominantly investigated via Sanger sequencing approaches [33, 35, 49–55]. Awaiting the development of high-throughput sequencing in the field, Sanger sequencing still remains the best approach in molecular phylogenetic studies. The present chapter provides a sequencing working protocol to be used in such studies.

Besides the advantage of using non-recombining molecular markers that originate from haplotype genome (oDNA), a second advantage is the absence of heterozygosity in contrast to nrDNA. This has a practical consequence in terms of DNA sequence acquisition, as using mtDNA or cpDNA will not necessitate DNA phase reconstruction (reconstruction of the two haplotypes that correspond to the two DNA copies of a given nrDNA marker) and/or DNA cloning. In other words, the exact sequence of a given oDNA marker can be recovered directly from a PCR product.

2 Materials

All work has to be performed in sterile conditions. All work surfaces have to be preliminary washed using bleach and alcohol.

2.1 Equipment

1. Micropipette and micropipette tips for dispensing from very small volumes (<1 μL) to larger ones (up to 1,000 μL).
2. Microtubes (1.5 and 0.5 mL).
3. PCR plates.
4. Sequencing plates.
5. Thermocycler.
6. Centrifuge.
7. Plastic sealing tape.
8. Aluminum sealing foil.

9. Gel electrophoresis system.
10. Microvolume spectrophotometer.
11. Sequencer.
12. Oven (facultative).
13. PCR product cleanup kit (optional, *see* Subheading 3.2).
14. Speed vacuum (optional, *see* Subheading 3.2.2).

2.2 PCR Reagents

1. Pure and high-quality DNA from plant tissues.
2. Taq polymerase (store at −20 °C and do not leave at room T°C for long time).
3. Taq polymerase buffers (provided with the Taq polymerase).
4. dNTPs (10 mM).
5. $MgCl_2$ (25 mM) (optional, as $MgCl_2$ might be contained in the Taq polymerase buffer).
6. Ultrapure water.
7. PCR primers (10 μM).

2.3 Electrophoresis Reagents

1. Agarose.
2. Ultrapure water.
3. 10× TBE electrophoresis buffer: 108 g Tris base, 55 g boric acid, 5.8 g EDTA disodium; adjust the solution with ultrapure water to a final volume of 1 L.
4. Loading dye (e.g., bromophenol blue: 3 mL glycerol (30 %), 25 mg bromophenol blue (0.25 %), dH_2O to 10 mL).

2.4 Sequencing Reagents

1. PCR product (template DNA) of known concentration (ng/μL).
2. PCR primers (10 μM).
3. Cycle sequencing polymerase.
4. Cycle sequencing buffer.
5. Ultrapure, RNAse-free water.

2.5 Purification Reagents

For the PCR product purification (optional, see Subheading 3.2):

1. Exonuclease I (20 U/μL).
2. Shrimp Alkaline Phosphatase (1 U/μL).

For the sequence product purification:

3. EDTA.
4. Sodium acetate.
5. 100 and 70 % ethanol.

2.6 Sequencer Electrophoresis Reagents

1. Formamide.

3 Methods

The sequencing of a DNA locus is done in five steps: (a) PCR amplification of the locus using specific DNA primers, (b) purification of the PCR product, (c) sequencing reaction on the PCR product, (d) purification of the sequencing product, and (e) electrophoresis of the sequencing product and acquisition of the data.

3.1 Amplification of PCR Products Using Universal Primers

If previously designed universal primers are used, readers are advised referring to the PCR protocols presented within the original publications for primer melting temperatures and elongation times (e.g., in [15, 16, 30–33]). A standard PCR protocol can be used:

1. Work on ice.
2. Add 1–2 μL of purified DNA (10–100 ng) in the well of the PCR plate.
3. Centrifuge the plate to ensure that all DNA templates reach the bottom of the wells.
4. Prepare in a microtube the PCR mix containing ultrapure water, Taq polymerase buffer, dNTPs, $MgCl_2$, Forward and Reverse primers (quantities depend on the final volume of the reaction and on the Taq polymerase used, see user manual provided with the Taq polymerase).
5. Mix gently.
6. Spin down briefly.
7. Add the Taq polymerase (taken at the last moment from the freezer and put back to the freezer immediately after use) (*see* **Note 1**).
8. Mix by pipetting in order to homogenize.
9. Distribute the mix in each well avoiding cross contaminations among wells.
10. Centrifuge the plate to ensure all products reach the bottom of the wells.
11. Transfer the plate to the thermocycler, and run a program previously settled up with the settings corresponding to the primer pairs (primer melting temperature, time of elongation, see original publications) and to the Taq polymerase that has been used (see the user manual provided with the Taq polymerase).
12. When the thermocycler run is finished, store the PCR products at 4 °C (short-term storage, one or two days) or –20 °C (long-term storage, more than 2 days) until performing further analysis.

Given that the quality of the DNA amplification step is crucial for the sequencing of the PCR products, potential troubles,

as amplification of nonspecific PCR products, are treated as a note (*see* **Note 2**). The following steps of the protocol rely on a "high-quality" PCR product that contains only one specific DNA fragment.

3.2 PCR Product Purification Before Sequencing

Two different protocols are proposed here. The first one relies on the use of two restriction enzymes (ExoSap protocol), whereas the second one is based on a column purification (column protocol) (*see* **Note 3**).

3.2.1 ExoSap

This protocol of PCR product cleanup is based on the enzymatic activity of the Exonuclease I (ExoI) together with the Shrimp Alkaline Phosphatase (SAP). The ExoI degrades the single-stranded DNA fragments in a 5′ → 3′ direction releasing deoxyribonucleoside 5′-monophosphates in a stepwise manner and leaving 5′-terminal dinucleotides intact. This allows the removal from the PCR mixture of the leftover primers and single-stranded DNA containing a 3′-hydroxyl terminus. The SAP catalyzes the release of 5′- and 3′-phosphate groups from DNA (removal of the dNTPs from the PCR mixture).

1. Before the purification, prepare a master mix containing the two enzymes in a 1.5 mL microtube with 100 μL of SAP (1 U/μL) and 5 μL of ExoI (20 U/μL).
2. Add 1.1 μL of the ExoSap enzyme mix to each 25 μL PCR product (*see* **Note 4**) (directly in the well of the PCR plate).
3. Incubate at 37 °C (working temperature for enzyme activities) for 1 h and at 85 °C for 15 min (inactivation of the enzymes). A thermocycler with the corresponding program can be used at this aim.
4. Store the PCR products at 4 °C or –20 °C until performing further analysis.

3.2.2 Column Protocol

Readers should follow the protocol provided with the PCR product cleanup kit of their choice. Importantly, either a speed vacuum or a centrifuge is necessary for this protocol. Basically, these ready-to-use kits are based on ultrafiltration. PCR products are applied and retained on a filtration membrane, while primers, dNTPs, and salts are eliminated using specific buffers. Once washed, the desired PCR products are recovered from the membrane after the addition of water or low salt buffer.

Store the PCR products at 4 °C or –20 °C until performing further analysis.

3.3 DNA Quantification

DNA quantification can be done either directly on the agarose gel after the electrophoresis of the PCR products or by absorbance measurements.

3.3.1 Quantification on Agarose Gel

PCR products can be quantified respectively to a standard DNA ladder (Fig. 1). To this end, PCR products as well as the DNA ladder have to be analyzed on a 1 % agarose gel electrophoresis system.

1. Prepare the agarose gel and wait until it has completely hardened.
2. During this time, mix 3 μL of the PCR product with 1 μL of a loading dye using a micropipette.
3. Load 3 μL of DNA ladder in the first and last well of the agarose gel with a micropipette. Most ladders have a standard band that corresponds to a standard amount of DNA per μL.
4. Load carefully the mixture obtained in **step 3** into a well of the agarose gel.
5. Load the following PCR products into each well, changing the tips of the micropipette each time to avoid contaminations.
6. Roughly quantify visually the PCR product concentration according to the intensity of the standard.

3.3.2 Quantification by Absorbance Measurements

Using a microvolume spectrophotometer, the PCR product concentration can be accurately quantified. Spectrophotometer-based quantifications are more accurate than gel-based quantification (*see* **Note 5**).

3.4 DNA Sequencing

The present protocol corresponds to a fluorescence-based DNA cycle sequencing reaction. DNA cycle sequencing refers to DNA sequencing based on the Sanger method (dideoxy chain termination method). The sequencing relies on the incorporation of ddNTPs (dideoxy nucleoside triphosphates) and dNTPs by a Taq polymerase. ddNTPs lack an –OH on the 2′-C and 3′-C of the pentose sugar. As a consequence, ddNTPs cannot form a bond with next incoming dNTP. Once integrated by the Taq polymerase during DNA replication, ddNTPs stop chain growth. Due to this property, they are commonly named "terminators" in cycle sequencing kit. This protocol might slightly change from one cycle sequencing kit to another, and readers are advised to read carefully the user manual that corresponds to the kit of their choice. General guidelines are here provided in order to obtain a working sequencing reaction.

Importantly, the sequencing reaction requires either a Forward or a Reverse primer (*see* **Note 6**) and a cycle sequencing polymerase (*see* **Note 7**). Figure 3 provides indications on the main steps of the sequencing protocol.

The respective quantities of these products for one cycle sequencing reaction depend on the cycle sequencing kit. Importantly the quantity of the PCR template depends on its concentration but also on its size (Fig. 3). Greater quantities of longer PCR templates are necessary in order to obtain good-quality sequences.

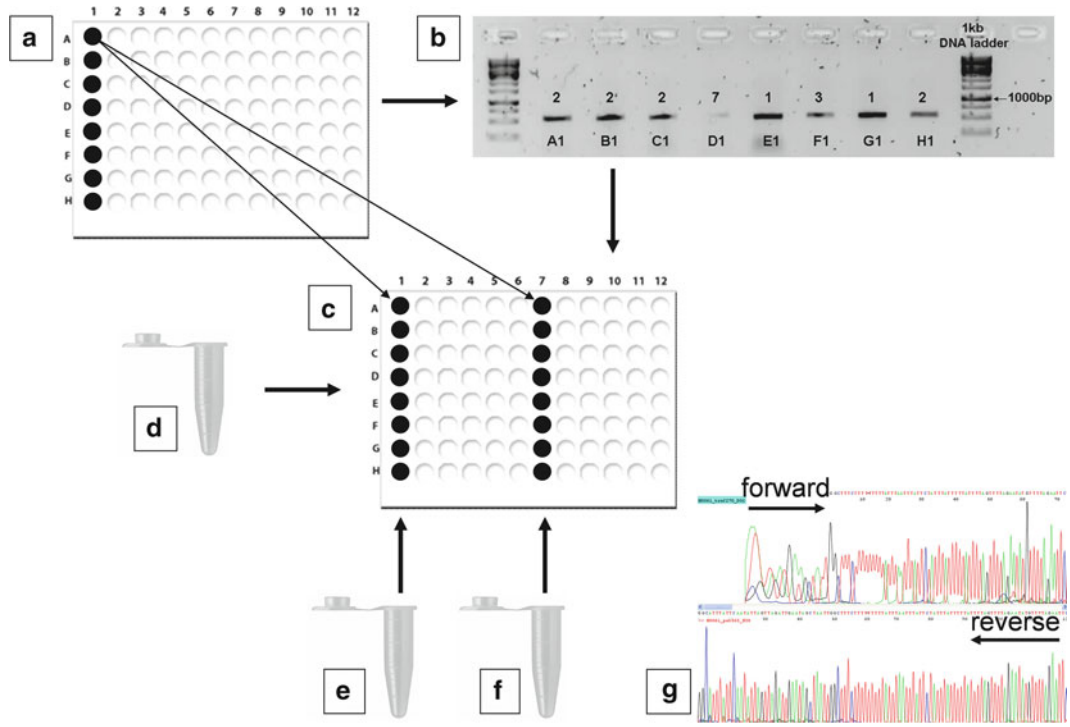


Fig. 3 Example of sequencing protocol. (**a**) Eight individuals were amplified using the two primers, Forward and Reverse (wells A1 to H1 in *black* on the PCR plate). The final PCR product size is about 500 bp length. (**b**) All individuals are loaded on an agarose gel together with a DNA ladder. PCR products can be quantified according to the 1,000 bp band of the DNA ladder. All individuals were successfully amplified (note however that the individual corresponding to the well D1 was poorly amplified). All these PCR products have to be purified before next step, see Subheading 3.2. On the picture of the agarose gel, below the bands the wells of origin of the PCR product from the PCR plate (A1 to H1) are indicated. Above the bands the quantity of PCR product that will be used for the sequencing reaction are reported. For example, 2 μL of the PCR product from the well A1 from the PCR plate (*see* (**a**)) will be loaded in the well A1 of the sequencing plate (*see* (**c**)) and 2 μL of the same PCR product will be loaded in the well A7 of the sequencing plate (*see* (**c**)). (**d**) Once the respective quantities of all samples are loaded on the sequencing plate, RNAse-free water is added in order to complete the final volume PCR product. If this final volume is 7 μL, 5 μL of RNAse-free water will be added in wells A1-B1-C1-H1-A7-B7-C7-H7, 6 μL in wells E1 and E7, and 4 μL in wells F1 and F7. Note that no water is added to wells D1 and D7 given that the final volume has already been reached. (**e**) and (**f**) Prepare a mix containing the cycle sequencing polymerase, the cycle sequencing buffer, and either the Forward primer (in tube (**e**)) or the Reverse primer (in tube (**f**)). Mix gently by pipetting, and distribute the required volume of the mix from tube (**e**) in column 1 and from tube (**f**) in column 7. Put the sequencing plate in a thermocycler using the sequencing program indicated by the provider. (**g**) DNA electrogram of one PCR product sequenced using the Forward primer (*upper part* of the figure) and the Reverse primer (*lower part* of the figure). An alignment of the two sequences allows obtaining a high-quality consensus sequence

The concentrations of the PCR templates are generally heterogeneous. As a consequence, the quantity of each PCR template to be used for the sequencing reaction will also be heterogeneous. It is advisable to apply some rules in order to minimize the quantity to be used in the preparation of the sequencing reactions. Thus, for

PCR templates of 500–1,000 bp, respectively, 1, 3, 5, and 7 μL will be used for highly (>200 ng/μL), moderately (60–100 ng/μL), low (30 ng/μL), and very low (10 ng/μL) concentrated products. Note that PCR templates at very low concentration do not generally provide good-quality sequences and trials have to be done only in "no-alternative" cases.

1. Work on ice.
2. Use a PCR plate that can be used directly in your electrophoresis system for acquisition of the data.
3. First load the DNA template in the PCR plate (*see* **Note 8**).
4. Once all DNA templates are loaded, centrifuge your plate to ensure that all PCR products are at the bottom of the wells.
5. Add the ultrapure water in each well according to the quantity of PCR templates that was loaded (*see* **Note 9**).
6. Centrifuge the plate.
7. Prepare in different 1.5 mL microtubes (each for different DNA primers) a mix that contains the PCR primers, the cycle sequencing polymerase, and the cycle sequencing buffer.
8. Homogenize the mix by pipetting.
9. Distribute the mix in the wells (*see* **Note 10**).
10. Run the plate on a thermocycler using the protocol indicated in the user manual.

3.5 Sequence Product Purification

Different protocols are available. The ethanol/EDTA/sodium acetate purification works well:

1. Spin briefly the PCR plate in order to ensure that PCR products reach the bottom of the wells.
2. Add 1 μL of 125 mM EDTA to each well.
3. Spin briefly in order to ensure that the EDTA reaches the bottom of the wells.
4. Add 1 μL of 3 M sodium acetate to each well.
5. Spin briefly in order to ensure that the sodium acetate reaches the bottom of the wells.
6. Add 50 μL of 100 % ethanol to each well.
7. Seal the plate with aluminum foil and mix by inverting three times.
8. Incubate at room temperature for 15 min.
9. Spin the plate at 1,650 × *g* for 45 min.
10. Proceed to the next step immediately to avoid the pellet resuspension.
11. Invert the plate, place it on an absorbent paper, and spin up to 185 × *g*.

12. Remove from the centrifuge.
13. Add 100 μL of 70 % ethanol to each well.
14. Spin at 1,650 × *g* for 15 min.
15. Invert the plate, place it on an absorbent paper, and spin up to 185 × *g* for 1 min.
16. Remove from the centrifuge.
17. Put the plate in an oven at 60 °C for 30 min in order to ensure that all the ethanol has been removed from the wells of the plates.
18. Resuspend the samples in formamide.

4 Notes

1. High-fidelity Taq polymerase might be used for mtDNA fragments that contain microsatellite-like motif (frequent repetitions of the same mono- or dinucleotides).
2. *Duplicated copies.* It is worth noting that even dealing with oDNA, obtaining a single-band (a unique "orthologous" locus) PCR product using universal primers is far from being a generality. However, obtaining orthologous fragments is critical as researchers might be interested in their comparison in parentage reconstruction methods as done within a species (haplotype network reconstruction) or among species (phylogenetic tree). The amplification of a two-band PCR product results from the amplification of two duplicated copies of the same fragment that are conserved in the region of DNA primer hybridization. Duplicated copies can correspond to the original mtDNA copy and a nuclear copy that has been transferred more or less recently to the nuclear genome [4–6]. In other words, after the duplication, one of the two copies became inactive and evolved as a neutral DNA marker. This inactive fragment that tends to evolve faster than the second still active copy is commonly named a pseudogene (a gene that has lost its function). Pseudogenes have been demonstrated to be more affected by deletions than insertions, which have been interpreted as a mark of selection for genome compactness [56], resulting in a smaller number of copies than the corresponding gene. As a consequence, the copy of interest that is the original focus of the PCR amplification is generally the one with the bigger size (nota bene: checking that this size corresponds to the expected amplification size obtained in other organisms is a valuable proof to identify the target copy). The presence of duplicated copies can be directly demonstrated after electrophoresis as two bands will appear on the gel.

The situation is more complex when dealing with noncoding fragments. These are generally considered neutral, albeit some models of indirect selection for genome size have been hypothesized [57, 58]. Therefore, it is expected that these fragments evolve as pseudogenes, without any direct selective constraint. The duplication of noncoding sequences thus creates two copies that will evolve independently.

In any case, if the duplication event is recent, irrespectively of whether the copy is coding or noncoding, the duplication cannot be visible on the gel, as differential mutation load would not create a detectable size difference. If the two copies present a small difference in size, they can be hopefully separated using a longer and more concentrated agarose gel. If the two copies have exactly the same size, there will not be any alternative but cloning the products and trying to define new PCR primers specific to one of the two copies, which will be almost impossible for recently duplicated copies. A practical advice would be to concentrate on the other loci that can be of interest and do not present such challenging methodological molecular issues.

Heterogeneity in the amplification success among individuals. Some fragments might demonstrate problems of amplification for some of the individuals under study. If these fragments are shown to be interesting (e.g., they reveal informative polymorphisms), one possibility for obtaining a complete data set would be to define internal primers. Relying on the DNA alignment based on the sequences obtained for some of the data, it is actually possible to redefine primers trying to encompass at best the polymorphism. If primers are designed in regions conserved over all taxa, they will allow to increase the amplification success rate.

3. The column protocol is normally more reliable than the ExoSap one but is more expensive. The crucial point concerns the quality of the PCR product. For good-quality product, the ExoSap protocol works fine.
4. PCR products correspond to one single band on the agarose gel. If the concentrations are heterogeneous among the different PCR products, this will be taken into account during the sequencing reaction steps (*see* Subheading 3.4).
5. Always perform the electrophoresis of the PCR products in order to verify if the amplification worked well (positive amplification and one single-band fragment). Based on the picture of the agarose gel, it is quite easy and fast to decide the amount of PCR product to be taken for the sequencing reaction (Fig. 3). Try first this procedure, and if it does not work, use the more time-consuming procedure that relies on absorbance measurements via a spectrophotometer.

6. The sequencing reaction of a PCR product can be done using either the Forward or the Reverse primer (but only one of the two primers in this case, not both as for a PCR amplification). If the Reverse primer is localized on the 5′-part of the PCR product and the Forward primer on the 3′-part, using the Reverse primer within the sequencing reaction, the sequence obtained will correspond to the 5′–3′ sequence (Fig. 3). Using the Reverse primer the 3′–5′ sequence will be obtained. According to the sequencer, the expected length sequence might generally fluctuate between 400 and 600 bp. That means that both Reverse and Forward sequences have to be obtained for locus larger than 800–1,000 bp in order to get the full locus sequence. For smaller PCR products, in principle, only the Reverse or the Forward sequence can be obtained. However, in order to get great confidence in the quality of the sequence, it is advisable to sequence in both sides. Reverse and Forward sequences will have to be aligned in order to infer a consensus sequence that will further be used for analyses.
7. Generally, the quantity of the cycle sequencing polymerase advised in user manuals are in large excess and can be reduced in order to lower the cost. It is worthwhile to perform some preliminary tests.
8. If the sequence has to be obtained in 5′–3′ direction (using the Forward primer) and in 3′–5′ direction (using the Reverse primer), organize your plate accordingly. If you have 48 PCR products to be sequenced in both sides (5′–3′ and 3′–5′), load the first PCR product on well A1 and on well A7, the second PCR product on wells B1 and B7, the third on wells C1 and C7, …, the 9th on wells A2 and A8, etc., until the 48th on wells H6 and H12 (*see* Fig. 3 for a simple example). The first half of the plate (columns 1–6) will then be used with the Forward primer, and the second half of the plate (columns 7–12), with the Reverse primer.
9. For a final volume of 10 μL of sequencing reaction, complete to 7 μl by adding *x*μL of water to the volume of PCR template (Fig. 3).
10. Relying on the previous example (final reaction volume: 10 μL, PCR template + water volume: 7 μL), add 3 μL of the mix that contains the Reverse or the Forward primer, the cycle sequencing polymerase, and buffer.

Acknowledgements

Research funded by the FNRS (grants FRFC 2.4.576.07.F and MIS F.4.519.10.F). Many thanks to Michela Di Michele and Esra Kaymac for their comments on a previous version of the article.

References

1. Gray MW, Burger G, Franz Lang B (2001) The origin and early evolution of mitochondria. Genome Biol 2:1018.1011–1018.1015
2. Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6:263–268
3. Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y, Song K (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci USA 50:27–29
4. Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 408:354–357
5. Adams KL, Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol 29:380–395
6. Adams KL, Qiu YL, Stoutemyer M, Palmer JD (2002) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci USA 99:9905–9912
7. Boore JL (1999) Survey and summary. Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780
8. Bullerwell CE, Gray MW (2004) Evolution of the mitochondrial genome: protist connections to animals, fungi and plants. Curr Opin Microbiol 7:528–534
9. Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10(1):e1001241
10. Kitazaki K, Kubo T (2010) Cost of having the largest mitochondrial genome: evolutionary mechanism of plant mitochondrial genome. J Bot. doi:10.1155/2010/620137
11. Marienfeld J, Unseld M, Brennicke A (1999) The mitochondrial genome of Arabidopsis is composed of both native and immigrant information. Trends Plant Sci 4:495–502
12. Kubo T, Newton KJ (2008) Angiosperm mitochondrial genomes and mutations. Mitochondrion 8:5–14
13. Schuster W, Brennicke A (1994) The plant mitochondrial genome: physical structure, information content, RNA editing, and gene migration to the nucleus. Annu Rev Plant Physiol Plant Mol Biol 45:61–78
14. Woloszynska M (2010) Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. J Exp Bot 61:657–671
15. Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4: 129–131
16. Duminil J, Pemonge MH, Petit RJ (2002) A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial DNA. Mol Ecol Notes 2:428–430
17. Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84: 9054–9058
18. Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci USA 93:10274–10279
19. Muse SV (2000) Examining rates and patterns of nucleotide substitution in plants. Plant Mol Biol 42:25–43
20. Parkinson CL, Mower JP, Qiu YL, Shirk AJ, Song K, Young ND, DePamphilis CW, Palmer JD (2005) Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol Biol 5:1–12
21. Cho Y, Mower JP, Qiu YL, Palmer JD (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc Natl Acad Sci USA 101: 17741–17746
22. Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7:135
23. Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR (2007) Variation in mutation rate and polymorphism among mitochondrial genes of *Silene vulgaris*. Mol Biol Evol 24: 1783–1791
24. Birky CW Jr (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet 35: 125–148
25. Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction. In: Weiss S, Ferrand N (eds) Phylogeography of Southern European refugia: evolutionary perspectives on the origins and conservation of European biodiversity. Springer, Dordrecht, pp 23–101

26. Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. Mol Ecol 14:689–701
27. Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T et al (1992) Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. A primitive form of plant mitochondrial genome. J Mol Biol 223:1–7
28. Marchelli P, Baier C, Mengel C, Ziegenhagen B, Gallo LA (2010) Biogeographic history of the threatened species *Araucaria araucana* (Molina) K. Koch and implications for conservation: a case study with organelle DNA markers. Conserv Genet 11:951–963
29. Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the Tortoise and the hare III. Am J Bot 94:275–288
30. Jeandroz S, Bastien D, Chandelier A, Du Jardin P, Favre JM (2002) A set of primers for amplification of mitochondrial DNA in *Picea abies* and other conifer species. Mol Ecol Notes 2: 389–392
31. Dumolin-Lapegue S, Pemonge MH, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol 6:393–397
32. Jaramillo-Correa JP, Bousquet J, Beaulieu J, Isabel N, Perron M, Bouillé M (2003) Cross-species amplification of mitochondrial DNA sequence-tagged-site markers in conifers: The nature of polymorphism and variation within and among species in *Picea*. Theor Appl Genet 106:1353–1367
33. Froelicher Y, Mouhaya W, Bassene JB, Costantino G, Kamiri M, Luro F, Morillon R, Ollitrault P (2011) New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. Tree Genet Genomes 7:49–61
34. Boonruangrod R, Desai D, Fluch S, Berenyi M, Burg K (2008) Identification of cytoplasmic ancestor gene-pools of *Musa acuminata* Colla and *Musa balbisiana* Colla and their hybrids by chloroplast and mitochondrial haplotyping. Theor Appl Genet 118:43–55
35. Godbout J, Jaramillo-Correa JP, Beaulieu J, Bousquet J (2005) A mitochondrial DNA minisatellite reveals the postglacial history of jack pine (*Pinus banksiana*), a broad-range North American conifer. Mol Ecol 14: 3497–3512
36. Jaramillo-Correa JP, Beaulieu J, Bousquet J (2004) Variation in mitochondrial DNA reveals multiple distant glacial refugia in black spruce (*Picea mariana*), a transcontinental North American conifer. Mol Ecol 13:2735–2747
37. Jose-Maldia LS, Uchida K, Tomaru N (2009) Mitochondrial DNA variation in natural populations of Japanese larch (*Larix kaempferi*). Silvae Genetica 58:234–241
38. Naydenov K, Senneville S, Beaulieu J, Tremblay F, Bousquet J (2007) Glacial vicariance in Eurasia: Mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. BMC Evol Biol 7:233
39. Moriguchi Y, Kang KS, Lee KY, Lee SW, Kim YY (2009) Genetic variation of *Picea jezoensis* populations in South Korea revealed by chloroplast, mitochondrial and nuclear DNA markers. J Plant Res 122:153–160
40. Burban C, Petit RJ (2003) Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. Mol Ecol 12: 1487–1495
41. Bastien D, Favre JM, Collignon AM, Sperisen C, Jeandroz S (2003) Characterization of a mosaic minisatellite locus in the mitochondrial DNA of Norway spruce [*Picea abies* (L.) Karst.]. Theor Appl Genet 107:574–580
42. Yoshida Y, Matsunaga M, Cheng D, Xu D, Honma Y, Mikami T, Kubo T (2012) Mitochondrial minisatellite polymorphisms in fodder and sugar beets reveal genetic bottlenecks associated with domestication. Biologia Plantarum 56(2):369–372
43. Honma Y, Yoshida Y, Terachi T, Toriyama K, Mikami T, Kubo T (2011) Polymorphic minisatellites in the mitochondrial DNAs of *Oryza* and *Brassica*. Curr Genet 57:261–270
44. Fievet V, Touzet P, Arnaud JF, Cuguen J (2007) Spatial analysis of nuclear and cytoplasmic DNA diversity in wild sea beet (*Beta vulgaris* ssp. *maritima*) populations: do marine currents shape the genetic structure? Mol Ecol 16:1847–1864
45. Sperisen C, Büchler U, Gugerli F, Mátyás G, Geburek T, Vendramin GG (2001) Tandem repeats in plant mitochondrial genomes: application to the analysis of population differentiation in the conifer Norway spruce. Mol Ecol 10:257–263
46. Lunt DH, Whipple LE, Hyman BC (1998) Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. Mol Ecol 7:1441–1455
47. Nishizawa S, Kubo T, Mikami T (2000) Variable number of tandem repeat loci in the

mitochondrial genomes of beets. Curr Genet 37:34–38

48. Hosaka K, Sanetomo R (2009) Comparative differentiation in mitochondrial and chloroplast DNA among cultivated potatoes and closely related wild species. Genes Genet Syst 84:371–378
49. Avtzis DN, Aravanopoulos FA (2011) Host tree and insect genetic diversity on the borderline of natural distribution: a case study of *Picea abies* and *Pityogenes chalcographus* (Coleoptera, Scolytinae) in Greece. Silva Fennica 45:157–164
50. Gugger PF, Gonzalez-Rodriguez A, Rodriguez-Correa H, Sugita S, Cavender-Bares J (2011) Southward Pleistocene migration of Douglas-fir into Mexico: phylogeography, ecological niche modeling, and conservation of 'rear edge' populations. New Phytol 189:1185–1199
51. Jaramillo-Correa JP, Beaulieu J, Ledig FT, Bousquet J (2006) Decoupled mitochondrial and chloroplast DNA population structure reveals Holocene collapse and population isolation in a threatened Mexican-endemic conifer. Mol Ecol 15:2787–2800
52. Goodall-Copestake WP, Pérez-Espona S, Harris DJ, Hollingsworth PM (2010) The early evolution of the mega-diverse genus *Begonia* (Begoniaceae) inferred from organelle DNA phylogenies. Biol J Linn Soc 101: 243–250
53. Edwards EJ, Nyffeler R, Donoghue MJ (2005) Basal cactus phylogeny: implications of *Pereskia* (Cactaceae) paraphyly for the transition to the cactus life form. Am J Bot 92:1177–1188
54. Eckert AJ, Tearse BR, Hall BD (2008) A phylogeographical analysis of the range disjunction for foxtail pine (*Pinus balfouriana*, Pinaceae): the role of Pleistocene glaciation. Mol Ecol 17:1983–1997
55. Cun YZ, Wang XQ (2010) Plant recolonization in the Himalaya from the southeastern Qinghai-Tibetan Plateau: geographical isolation contributed to high population differentiation. Mol Phylogenet Evol 56:972–982
56. Petrov DA, Hartl DL (2000) Pseudogene evolution and natural selection for a compact genome. J Hered 91(3):221–227
57. Duminil J, Grivet D, Ollier S, Jeandroz S, Petit RJ (2008) Multilevel control of organelle DNA sequence length in plants. J Mol Evol 66: 405–415
58. Selosse MA, Albert B, Godelle B (2001) Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol Evol 16:135–141

Chapter 7

Nuclear Ribosomal RNA Genes: ITS Region

Pascale Besse

Abstract

Despite possible drawbacks (intraspecific polymorphisms and possible fungal contamination), sequencing of the ribosomal RNA gene ITS region remains one of the most popular nuclear sequences used for plant taxonomy and phylogeny. A protocol for PCR amplification and sequencing of this region using universal plant primers is provided.

Key words Ribosomal DNA, ITS, Sequencing, PCR

1 Introduction

Since early reviews [1] and the general agreement around the necessity to use biparentally inherited nuclear markers together with monoparentally inherited ones such as chloroplast or mitochondrial DNA, nuclear ribosomal RNA genes (nrDNA) have received increasing attention in plant taxonomy and phylogeny. One of the reasons is that such genes provide significant information in phylogenetic research because they are composed of different regions (both coding and noncoding) that are conserved differently and thus provide information at different taxonomic levels [2] (*see* Chapter 2). In particular, spacer regions of nrDNA are useful for plant systematics from species to generic levels [3]. Another related reason for such popularity is that easy PCR amplification is provided by designing PCR primers in conserved coding regions surrounding a more variable spacer region. Ribosomal genes are arranged in tandem repeats and are subjected to concerted evolution, which results in the homogenization of the sequences at the tandem array, individual, population, and species levels through genomic mechanisms like gene conversion and unequal crossing-over [4, 5]. Homogeneous nrDNA sequences are therefore generally found within one genome [2]. This implies reduced levels of intraspecific variation (as compared to interspecific) therefore allowing a reduced intraspecific sampling effort.

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_7,

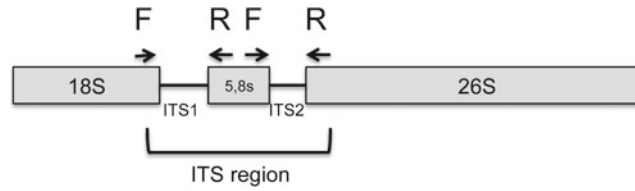


Fig. 1 Structure of the ITS region of the nuclear ribosomal RNA genes and schematic location of primers from Table 1

It also provides ease of analysis (because nrDNA is abundant and uniform) [2].

In plants, nrDNA are generally arranged in two distinct sets of tandem repeats. The first one is composed of nrDNA 5s and the second of nrDNA 18s-5,8s-26s (Chapter 2). The latter is the most frequently used for plant phylogeny and taxonomy. It is present at one or more loci (with hundred to thousands of tandem copies) [6], and when transcriptionally active, these regions are referred to as NORs (nucleolar organizer regions). It comprises different spacer regions. The intergenic spacer IGS, which separates adjacent nrDNA 18s-5,8s-26s units, contains many reiterated subrepeats within its sequence and is very variable both in sequence and in length [3]. This leads to difficulties for correctly aligning IGS sequences. It therefore has not received as much attention as the internal transcribed spacers (ITS), which are flanking the 5,8S RNA gene region (between the 18s and the 26s RNA genes) (Fig. 1). This entire ITS region (ITS1 + 5,8s + ITS2) can be easily amplified using universal primers in the conserved coding regions [3] (Fig. 1), as the total size is up to 700 bp in angiosperms [1], although in some other seed plants such as gymnosperms, it can be much longer, up to 1,500–3,700 bp [7]. The ITS region has become highly popular in plant phylogeny [7], as witnessed by the constant increase of *Embryophyta* ITS sequences available in the NCBI database since 2003 (189,026 ITS sequences as per 17th of June 2013) (Fig. 2). As a comparison, much fewer hits (4,275) are obtained for 5s nrDNA. This region provides different levels of informativeness: the central 5,8s RNA gene is highly conserved due to evolutionary constrains, whereas the surrounding ITS spacers are highly variable and more informative. This can be illustrated by an alignment of these regions made from Poaceae sequences (Fig. 3).

Despite early warnings [8], attention has been focussed only recently on the possible drawbacks in using nrDNA (and therefore ITS) for phylogenetic studies [9]. Concerted evolution does not always act immediately after organismal processes (such as hybridization or polyploidization) or after genomic changes (duplication, recombination) [3, 9]. Concerted evolution efficiency may also vary across loci and taxa [9]. This may induce intraindividual nrDNA polymorphism [8]. If hybrids and allopolyploids are recent, they

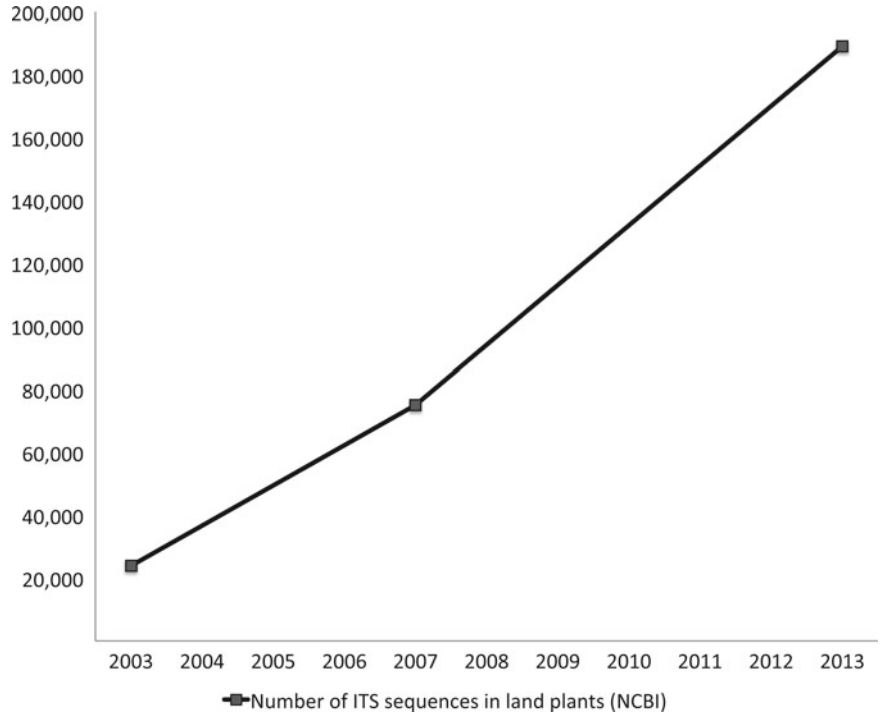


Fig. 2 Number of nucleotide ITS sequences for land plants in NCBI (2003 and 2007 values are from ref. [9])

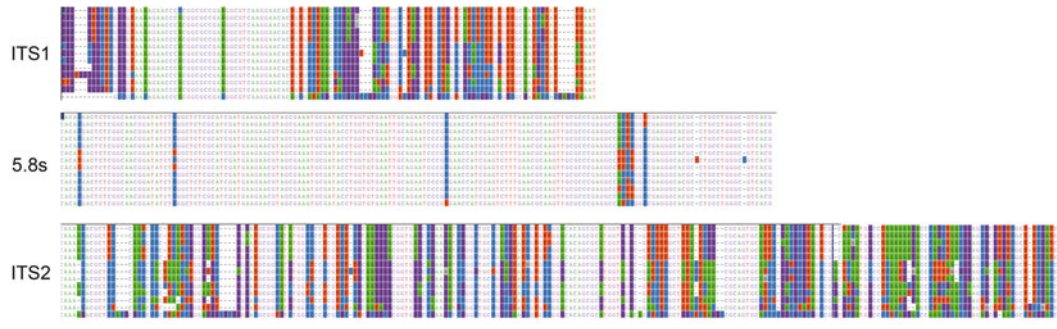


Fig. 3 Variable sites (highlighted) in the ITS1, 5.8s rRNA gene, ITS2 regions of various Poaceae species following alignment with Mega 5 [20]: *Bromus carinatus* (AY367948), *Bromus gunckelii* (AY367947), *Bromus berteroanus* (AY367946), *Bromus striatus* (AY367945), *Bromus cebadilla* (AY367944), *Festuca matthewsii* (AY524836), *Festuca madida* (AY524833), *Festuca novae-zelandiae* (AY524832), *Agropyron cristatum* (L36480), *Thinopyrum bessarabicum* (L36506), *Lolium perenne* (L36517), *Poa alpina* (AY327793), and *Oryza sativa* (DQ996015)

might retain paralogous copies of their nrDNA genes. On the other hand, in some hybrid species or polyploids, one of the parental nrDNAs can be more or less rapidly selectively eliminated [7, 10, 11]. Intermediate cases of partial additivity are also found [7]. The PCR product obtained may therefore represent a mixture of sequences (in various concentrations) sharing the same priming sites, but located at one or more locus on one or more chromosomes, and representing either paralogous or orthologous sequences [7, 8]. It is

important to be aware of the possibility of such intraspecific nrDNA heterogeneity, which may thus result not only from the presence of homeologous loci due to recent hybridization (with or without polyploidization), but also from [7, 11, 12]:

- Low concerted evolution rates: different sequences will coexist within a single locus.
- Duplication.
- Allelic variants (heterozygosity).
- Amplification of nonfunctional copies (pseudogenes) with different evolutionary constraints [13].
- Possibilities of contamination by fungal DNA (as the same primers are used for plants) [7].

Paralogous copies resulting from hybridization or allopolyploidization processes can efficiently be utilized to study these processes. Many examples are reviewed in [3, 9]. The problem is more when paralogous sequences are mistaken for orthologous sequences, which will lead to wrong inference in species relationships [8, 14]. The occurrence of nrDNA intraspecific heterogeneity (not due to hybrids or allopolyploids) has been documented in a range of taxa as reviewed by [14]. Only a detailed study involving cloning and cytogenetics (FISH or GISH to reveal array number and chromosomal distribution) (Chapters 15 and 16) may help resolve the origin of this heterogeneity. Very thorough flowchart diagrams were designed [12] to help unravel part of these problems. The least that should be done are:

(1) To check, after PCR amplification (under stringent conditions), that only one clear band is obtained. Otherwise subsequent cloning and analysis of each PCR product will be required. (2) After sequencing, always do a BLAST search to check for possible contamination (particularly from fungus). (3) It is also very important to verify the congruence of the obtained ITS tree with other marker trees such as chloroplastic gene trees: if differences are detected, they could be due to the hybrid status of some species, but isolated polyphyly could indicate paralogous sequence. A more detailed procedure depicted in very thorough flowchart diagrams is available [12] if necessary.

A strategy to distinguish between paralogous (pseudogene) and orthologous copies by using nucleotide diversification patterns to determine if sequences are functionally constrained using tree-based approaches was also proposed [14]. A comparison of 5,8s and ITS trees is performed. Functional copies should have a slower rate of evolution of 5,8s compared to ITS, whereas pseudogenes should show equal evolutionary rates in 5,8s and ITS.

The ITS region was early proposed [15] as a powerful tool for plant DNA barcoding, following testing on a large plant sampling (99 species covering 80 genera from 53 different families) which

showed more divergence (2,81 %) than the most variable intergenic chloroplastic region *trn*H–*psb*A (1,24 %). Nevertheless, two chloroplastic genes (*mat*K and *rbc*L) have been selected by CBOL in 2009 as the universal plant barcode system [16] mainly because of the previously mentioned possible drawbacks in ITS analysis. Recently, the high success rate of the ITS2 region to identify species in dicotyledons (76.1 %), monocotyledons (74.2 %), gymnosperms (67.1 %), ferns (88.1 %), and mosses (77.4 %) was further demonstrated [17]. In addition, the Chinese Barcoding of Life group [18] conducted a very thorough (6,286 samples from 1,757 angiosperm and gymnosperm species) comparative (with chloroplastic genes *rbc*L and *mat*K and intergene *trn*H–*psb*A) research on ITS efficiency/universality. Themselves and others [19] advocated for the incorporation of the ITS region as a supplementary barcode for land plants. Adding ITS to the plant barcode system (*rbc*L + *mat*K) indeed brings discrimination success from 49.7 to 77.4 %. Furthermore, [18] showed that contrary to what was feared [7], very low problems with fungal contamination (only in 2 % of the samples studied) and a very low occurrence of intraindividual multiple copies of nrDNA (only 7.4 % of the individuals) were detected.

ITS therefore remains a highly suitable and powerful region for resolving plant taxonomic and phylogenetic issues in most plant lineages, as long as one is aware of its possible (but hopefully rare) limits.

2 Materials

All solutions must be made up using sterile deionized water (MilliQ water), and all chemicals must be analytical reagent grade. As in all molecular biology procedures, work surfaces should be cleaned and gloves should be worn for all procedures.

2.1 PCR

1. PCR machine (thermocycler).
2. PCR plates or PCR tubes.
3. Taq polymerase: GoTaq® DNA polymerase (Promega) is well suited, 5 U/μL.
4. Appropriate Taq polymerase buffer (e.g., green flexi buffer for GoTaq®DNA polymerase) (5×).
5. $MgCl_2$ 25 mM (if not present in buffer).
6. dNTPmix (10 mM each) (add 10 μL of each dNTP solution at 100 mM to 60 μL MilliQ water).
7. Universal plant ITS primers (Table 1) (5 μM).
8. Plant DNA (10 ng/μL).
9. Sterile MilliQ water.

Table 1
Universal plant primers for the ITS region (always use a combination of a forward and a reverse primer, *see* Fig. 1)

Primer name	Sequence (5′->3′)	Reference
	18S Forward primers	
ITS1	TCCGTAGGTGAACCTGCGG	[21]
ITS5	GGAAGTAAAAGTCGTAACAAGG	[21]
17SE	ACGAATTCATGGTCCGGTGAAGTGTTC	[22]
	26S Reverse primers	
ITS4	TCCTCCGCTTATTGATATGC	[21]
26SE	TAGAATTCCCCGGTTCGCTCGCCGTTAC	[22]
	5.8S Reverse primers	
ITS2	GCTGCGTTCTTCATCGATGC	[21]
	5.8S Forward primers	
ITS3	GCATCGATGAAGAACGCAGC	[21]

2.2 Electrophoresis

1. Electrophoresis apparatus (gel tray, combs, power supply).
2. Standard transilluminator (302 nm with 6 × 15 W tubes).
3. High resolution agarose and standard agarose (molecular biology grade) (*see* **Note 1**).
4. 10× TRIS (tris(hydroxymethyl)aminomethane)-borate (TBE) buffer: 10.8 g TRIS base, 5.5 g boric acid, 0.7 g ethylenediaminetetraacetic acid (EDTA)-Na_2 in100 mL H_2O. This TBE buffer is diluted to 1× in MilliQ water for use.
5. Fluorescent nucleic acid gel stain: GelRed™ 1,000× in water, (*see* **Note 2**) (ethidium bromide can also be used if preferred).

3 Methods

3.1 PCR

For each PCR (25 µL) (*see* **Note 3**):

1. Deposit 2.5 µL template DNA in tube or well of the plate.
2. Add 1.5 µL $MgCl_2$ 25 mM.
3. Add 0.5 µL dNTPmix 10 mM.
4. Add 1.5 µL of each primer 5 µM (forward and reverse).
5. Add 5 µL of PCR buffer 5×.
6. Add 12.3 µL MilliQ sterile water.
7. Add 0.2 µL (1 U) DNA polymerase.

3.2 PCR Program

1. Pre-denaturation at 95 °C for 3 min.
2. 35 cycles with: 95 °C for 45 s, 60 °C for 45 s, 72 °C for 1 min 30.
3. Final elongation step at 72 °C for 7 min.
4. Maintain at 4 °C.

3.3 PCR Quality Verification on Agarose Gel

1. Prepare a 2 % mixture of agarose in TBE 1× (2 g agarose for 100 mL).
2. Bring to boil in a microwave oven.
3. Add 5 μL RedGel™ for 50 μL 2 % agarose/TBE.
4. Cool down, pour gel. Let gel cool down and prepare for migration.
5. Add 10 μL of PCR solution and loading dye.
6. Deposit in the well.
7. Run migration for appropriate time and observe gel over trans-illuminator (*see* **Note 4**).

3.4 Sequencing

A large number of private companies perform sequencing reaction directly from PCR products that can be sent by express mail (either sealed or vacuum dried in a SpeedVac). Generally the primers used have to be provided (*see* **Note 5**).

4 Notes

1. We use high resolution agarose gels rather than standard agarose gels to check for the purity of the amplified fragment and insure that only one band is amplified. Further routine checks can be made on standard agarose gels, which can be re-thawed and reused 6 times.
2. We prefer using GelRed™ than ethidium bromide (EB) as with standard Ames test, as measured in two bacterial strains, GelRed™ has been confirmed to be substantially safer than EB. GelRed™ is not mutagenic at all dosages in the absence of the S9 fraction. With S9 metabolic activation, GelRed™ showed weak mutagenicity only at the highest dosage (50 μg/plate or 18.5 μg/mL), well above the normal concentration used for gel staining.

   We however use EB safety rules when handling GelRed™: solution pipetting is made under a fume hood, and wear gloves and a lab coat. Whether GelRed™ waste solution can be directly poured into the drain may depend on local regulations despite its nonmutagenicity and noncytotoxicity. Alternatively, GelRed™ solution may be disposed by adding 25–50 mL bleach (regular household bleach) to each gallon (~4 L) of the waste staining solution and letting the mixture react for at least 8 h before pouring the solution to a sink. (Practically, you may

simply accumulate your GelRed™ waste solution in a jar containing appropriate amount of bleach.) For precast gels, you can simply let the gels dry out first and then let the dried waste go in regular trash bag (together with gloves and other wastes which are autoclaved prior to disposal).

3. Generally a PCR mix (**steps 2–7**) is prepared for the desired number of reactions (allow for 10 % variation) and then aliquoted in the wells of the PCR plates or in the PCR tubes containing the DNA samples. Work on ice.
4. At this stage the amplification should give a unique clear band. If more than one band is obtained, try to use more stringent PCR conditions (increase annealing temperature, lower $MgCl_2$ concentration, use species-specific primers rather than universal ones). If the problem still appears, the different bands will have to be cloned and sequenced.
5. Always ask for a double sequencing reaction, e.g., forward and reverse, which helps to check the quality of the sequence (consistently poor reactions despite specific primers and stringent PCR conditions might be due to heterozygous state or heterogeneity in the sequences—paralogous copies).

Acknowledgements

Denis Da Silva and Michel Grisoni (UMR PVBMT Cirad-Université de la Réunion) are acknowledged for their help in the ITS amplification protocol design.

References

1. Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann Mo Bot Gard 82:247–277
2. Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453
3. Poczai P, Hyvönen J (2010) Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep 37:1897–1912
4. Arnheim N (1983) Concerted evolution in multigene families. In: Nei M, Koehn R (eds) Evolution of genes and proteins. Sinauer, Sunderland, MA, pp 38–61
5. Dover G (1994) Concerted evolution, molecular drive and natural selection. Curr Biol 4: 1165–1166
6. Rogers SO, Bendich AJ (1987) Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Mol Biol 9:509–520
7. Alvarez IA, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogenet Evol 29:417–434
8. Buckler ES, Ippolito A, Holtsford TP (1997) The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. Genetics 145:821–832
9. Calonje M, Martın-Bravo S, Dobes C, Gong W, Jordon-Thaden I, Kiefer C, Kiefer M, Paule J, Schmickl R, Koch MA (2009) Non-coding nuclear DNA markers in phylogenetic reconstruction. Plant Syst Evol 282:257–280
10. Bachmann K (1994) Molecular markers in plant ecology Tansley review no. 63. New Phytol 126:403–418

11. Zimmer EA, Wen J (2013) Using nuclear gene data for plant phylogenetics: progress and prospects. Mol Phylogenet Evol 66:539–550
12. Feliner GN, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol Phylogenet Evol 44:911–919
13. Besnard G, Rubio de Casas R, Vargas P (2007) Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (Olea europaea). J Biogeogr 34:736–752
14. Bailey CD, Carr TG, Harris SA, Hughesa CE (2003) Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. Mol Phylogenet Evol 29:435–455
15. Kress WJ, Wurdack KJ, Zimmer EA, Weig LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. PNAS 102:8369–8374
16. CBOL Plant Working Group (2009) A DNA barcode for land plants. PNAS 106: 12794–12797
17. Yao H, Song J, Liu C, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS One 5:e13102
18. Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Chen Z-D, Zhou S-L, Chen S-L, Yang J-B, Fu C-X, Zeng C-X, Yan H-F, Zhu Y-J, Sun Y-S, Chen S-Y, Zhao L, Wang K, Yang T, Duan G-W, China Plant BOL Group (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. PNAS 108:19641–19646
19. Hollingsworth PM (2011) Refining the DNA barcode for land plants. PNAS 108: 19451–19452
20. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
21. White T, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis M, Gelfand D, Sninsky J, White T (eds) PCR protocols: a guide to methods and applications. Academic, New York, USA, pp 315–322
22. Sun Y, Skinner D, Liang G, Hulbert S (1994) Phylogenetic analysis of *Sorghum* and related taxa using internal transcribed spacers of nuclear ribosomal DNA. Theor Appl Genet 89:26–32

Chapter 8

New Technologies for Ultrahigh-Throughput Genotyping in Plant Taxonomy

David Edwards, Manuel Zander, Jessica Dalton-Morgan, and Jacqueline Batley

Abstract

Molecular genetic markers represent one of the most powerful tools for the analysis of variation between plant genomes. Molecular marker technology has developed rapidly over the last decade, with the introduction of new DNA sequencing methods and the development of high-throughput genotyping methods. Single nucleotide polymorphisms (SNPs) now dominate applications in modern plant genetic analysis. The reducing cost of DNA sequencing and increasing availability of large sequence data sets permit the mining of this data for large numbers of SNPs. These may then be used in applications such as genetic linkage analysis and trait mapping, diversity analysis, association studies, and marker-assisted selection. Here we describe automated methods for the discovery of SNP molecular markers and new technologies for high-throughput, low-cost molecular marker genotyping. Examples include SNP discovery using autoSNPdb and wheatgenome.info as well as SNP genotyping using Illumina's GoldenGate™ and Infinium™ methods.

Key words autoSNP, autoSNPdb, Wheatgenome.info, Infinium™, Single nucleotide polymorphism, SNP, SNPlex™

1 Introduction

The application of molecular markers in plant science is now well established. The bulk of variation at the nucleotide level is often not visible at the phenotypic level. This variation can be exploited in molecular genetic marker systems. DNA-based markers have many advantages over phenotypic markers in that they are highly heritable, are relatively easy to assay, and are not affected by the environment. Markers that are transferable between species enable studies of synteny and genome rearrangement across taxa. Modern agricultural breeding is dependent on molecular markers for the rapid and precise analysis of germplasm, trait mapping, and marker-assisted selection. Molecular markers are complementary tools to phenotypic analysis. They can increase our understanding of phenotypic characteristics, their genetic association, and distribution

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_8,

across populations. Furthermore, molecular markers are invaluable as a tool for genome mapping in all systems, offering the potential for generating very high-density genetic maps that can be used to develop haplotypes for genes or regions of interest [1]. Insight into the organization of the plant genome can be obtained by calculating a genetic linkage map using molecular markers. Genetic mapping places molecular genetic markers on linkage groups based on their co-segregation in a population. Single nucleotide polymorphisms (SNPs) are now the principal markers utilized in plant genetic analysis.

Molecular markers have many applications in plant taxonomy. However, to increase throughput and decrease costs, it is necessary to eliminate bottlenecks throughout the SNP discovery and genotyping process, as well as minimize sources of variability and human error to ensure data quality and reproducibility. These new technologies may provide the way forward for the discovery and application of molecular markers in plant taxonomy and will enable the application of markers for a greater diversity of species than currently possible.

1.1 What Are SNPs?

DNA sequence differences are the basic requirement for the study of molecular genetics. SNPs are the ultimate form of molecular genetic marker, as a nucleotide base is the smallest unit of inheritance, and an SNP represents a single nucleotide difference between two individuals at a defined location. There are two different forms of SNPs: transitions (C/T or G/A) and transversions (C/G, A/T, C/A, or T/G) [2]. SNPs are direct markers as the sequence information provides the exact nature of the allelic variants. Furthermore, this sequence variation can have a major impact on how the organism develops and responds to the environment. SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers near a locus of interest [3].

SNPs can differentiate between related sequences, both within an individual and between individuals within a population. The frequency and nature of SNPs in plants is beginning to receive considerable attention. Studies of sequence diversity have been performed for a range of plant species, and these have indicated that SNPs appear to be abundant in plant systems, with 1 SNP every 100–300 bp [4]. SNPs at any particular site could in principle involve four different nucleotide variants, but in practice they are generally biallelic. This disadvantage, when compared with multiallelic markers such as simple sequence repeats (SSRs), is compensated by the relative abundance of SNPs. SNPs are also evolutionarily stable, not changing significantly from generation to generation. The low mutation rate of SNPs makes them excellent markers for studying complex genetic traits and as a tool for understanding genome diversity and evolution [5].

The high density of SNPs makes them valuable for genome mapping, and in particular they allow the generation of ultrahigh-density genetic maps and haplotyping systems for genes or regions of interest and map-based positional cloning. SNPs are used routinely in crop breeding programs [6], for genetic diversity analysis, cultivar identification, phylogenetic analysis, characterization of genetic resources, and association with agronomic traits [1]. The applications of SNPs have been extensively reviewed by Rafalski [1], Gupta et al. [6], Edwards and Batley [7], and Duran et al. [8]. These reviews highlight that for several years SNPs will coexist with other marker systems. However, with the development of new technologies to increase throughput and reduce the cost of SNP assays, along with further plant genome sequencing, the use of SNPs will become more widespread.

1.2 Why Novel Marker Technologies Are Required

During the past two decades, several molecular marker technologies have been developed and applied for plant genome analysis, predominantly assessing the differences between individual plants within a species. These marker technologies have been applied to plant breeding to allow breeders to use the genetic composition or genotype of plants as a criterion for selection in the breeding progress. The recent application of association mapping via linkage disequilibrium (LD) in plants demonstrates the requirement to be able to identify and screen large numbers of markers rapidly and at low cost [9]. The development of technologies that increase marker throughput with reducing cost will broaden the uptake of marker-assisted breeding to include more diverse crops and a greater variety of traits as well as the characterization of diverse wild germplasm.

1.3 New Marker Discovery

Large quantities of sequence data are generated through transcriptome or genome sequencing projects internationally, and these provide a valuable resource for the mining of molecular markers [10, 11]. This will be further accelerated with the application of new sequencing technology from Roche (454) and Illumina (Hi Seq) [12–14].

1.3.1 In Silico SNP Discovery

The challenge of in silico SNP discovery is not the identification of polymorphic bases, but the differentiation of true SNPs from the often more abundant sequence errors. High-throughput sequencing remains prone to inaccuracies as frequent as one error every 100 base pairs. This incorrect base calling impedes the electronic filtering of sequence data to identify potentially biologically relevant polymorphisms. There are several different sources of error which need to be taken into account when differentiating between sequence errors and true polymorphisms.

The frequency of occurrence of a polymorphism at a particular locus provides a measure of confidence in the SNP representing

a true polymorphism and is referred to as the SNP redundancy score. By examining SNPs that have a redundancy score equal than or greater than two (two or more of the aligned sequences represent the polymorphism), the vast majority of sequencing errors are removed. Although some true genetic variation is also ignored due to its presence only once within an alignment, the high degree of redundancy within the data permits the rapid identification of large numbers of SNPs without the requirement for sequence quality scores. However, while redundancy-based methods for SNP discovery are highly efficient, the nonrandom nature of sequence error may lead to certain sequence errors being repeated between runs around locations of complex DNA structure. Therefore, errors at these loci would have a relatively high SNP redundancy score and appear as confident SNPs. In order to eliminate this source of error, an additional independent SNP confidence measure is required. This can be determined by the co-segregation of SNPs to define a haplotype. True SNPs that represent divergence between homologous genes co-segregate to define a conserved haplotype, whereas sequence errors do not co-segregate with a haplotype. Thus, a co-segregation score, based on whether an SNP position contributes to defining a haplotype, is a further independent measure of SNP confidence. By using the SNP score and co-segregation score together, true SNPs may be identified with reasonable confidence.

Three tools currently apply the methods of redundancy and haplotype co-segregation: autoSNP [15, 16], SNPServer [17], and autoSNPdb [18, 19]. The recently developed autoSNPdb combines the SNP discovery pipeline of autoSNP with a relational database, hosting information on the polymorphisms, cultivars, and gene annotations, to enable efficient mining and interrogation of the data [20]. Users may search for SNPs within genes with specific annotation or for SNPs between defined cultivars. autoSNPdb can integrate both Sanger and Roche 454 pyrosequencing data enabling efficient SNP discovery from next-generation sequencing technologies.

The short reads and large data volumes produced by Illumina DNA sequencing require a different approach to mine this data for SNP discovery [21, 22]. By mapping these high-throughput reads to a reference, it is possible to identify very large numbers of confident SNPs [23]. This approach is adopted in the SGSautoSNP pipeline with the output produced in GFF3 format suitable for display in genome viewers such as GBrowse [24] and Biomatters Geneious [25]. To date, this method has been used to identify more than 1.5 million SNPs across the B. napus canola genome and more than 800,000 SNPs across the group 7 chromosomes in wheat.

1.3.2 Identification of SNPs Using autoSNPdb

autoSNPdb implements the autoSNP pipeline within a relational database to enable mining for SNP and indel polymorphisms [18]. A web interface enables searching and visualization of the data, including the display of sequence alignments and SNPs. All sequences are annotated by comparison with GenBank and UniRef90, as well as through comparison with reference genome sequences. The system allows researchers to query the results of SNP analysis to identify SNPs between specific groups of individuals or within genes of predicted function.

The discovery of very large numbers of SNPs across diverse germplasm requires the establishment of custom databases and tools to maintain, integrate, and interrogate this information. In the above examples, we present the application of tools to identify SNPs from next-generation gene expression and whole-genome shotgun data. With the continued growth of next-generation DNA sequencing, it is expected that the requirement for these tools will continue to increase.

1.4 New Genotyping Technologies

1.4.1 New Genotyping Technologies for SNPs

Many new marker technologies involve improving the genotyping of SNPs, reflecting the increasing popularity of these markers. SNPs can be identified within a gene of interest or within close proximity to a candidate gene. Although the SNP may not be directly responsible for the observed phenotype, it can be used for the positional cloning of the gene responsible and as a diagnostic marker. Furthermore, SNPs are useful to define haplotypes in regions of interest. The success of the human HapMap project [26], where a very large number of SNPs were assayed over a range of individuals from different groups, demonstrates the value that can be gained from SNP studies. Reducing costs could enable similar studies to be undertaken to gain a greater understanding of plants.

1.4.2 Illumina GoldenGate™ and Infinium™ Assays

The Illumina GoldenGate™ technology is a novel array technology that is capable of genotyping up to 1,536 polymorphic sites in 384 individuals using custom SNP panels or oligo pool assays (OPAs). A level of redundancy is provided for each locus to increase confidence in genotype calls.

The assay performs allelic discrimination directly on genomic DNA and then generates a synthetic allele-specific PCR template before performing PCR on this artificial template. This is a reversal of conventional SNP genotyping assays which usually use PCR to amplify a SNP of interest and carry out allelic discrimination on the PCR product.

The assay technology works on two allele-specific oligonucleotides (ASOs) and one locus-specific oligonucleotide (LSO) for each SNP. Each ASO consists of a 3′ portion that hybridizes to the DNA at the SNP locus, with the 3′ base complementary to one of the two SNP alleles, and a 5′ portion that incorporates a universal

PCR primer sequence (P1 or P2, each associated with a different allele). The LSOs consist of three parts: at the 5′ end is a SNP locus-specific sequence, the middle contains an address sequence complementary to one of the capture sequences on the array, and there is a universal PCR priming site (P3′) at the 3′ end. After PCR, the amplified products are captured on beads carrying complementary target sequences for the SNP-specific tag of the ligation probe. Each SNP is assigned a different address sequence, which is contained within the LSO. Each of these addresses is complementary to a unique capture sequence represented by one of the bead types in the array. Therefore, the products of the assays hybridize to different bead types in the array, allowing all genotypes to be read simultaneously. The ratio of the two primer-specific fluorescent signals identifies the genotype as either of the two homozygotes or heterozygote. This universal address system, consisting of artificial sequences that are not SNP specific, allows any set of SNPs to be read on a common, standard array, providing flexibility and reducing array manufacturing costs. Custom assays are made on demand by building the address sequences into the SNP-specific assay oligonucleotides.

The introduction of the Infinium assay allowed for larger-scale SNP panel analysis than the GoldenGate assay. Genome-wide genotyping for any species of interest where sequence is available, using custom SNP panels sized between 3,000 and one million, can be performed using the novel Infinium® HD assays. In this assay a whole-genome amplification step, rather than PCR, is used to increase the amount of DNA up to 1,000-fold. The DNA is fragmented and captured on a bead array by hybridization to immobilize SNP-specific primers, followed by extension with hapten-labelled nucleotides. The primers hybridize adjacent to the SNPs and are extended with a single nucleotide corresponding to the SNP allele. The incorporated hapten-modified nucleotides are detected by adding fluorescently labelled antibodies in several steps to amplify the signals. The size of the SNP panel determines sample capacity on the BeadChips with 4, 12, and 24 sample size chips genotyping a maximum of one million, 250,000, and 90,000 SNPs, respectively, with a minimum of labor time. Use of these high-throughput multiplexed assays has been enabled by the advent of second-generation genome sequencing and in silico SNP discovery pipelines from the resultant data. Use of the Infinium assay is expected to increase as greater numbers of reference genomes become available. A custom 50,000 SNP panel has been designed against *Brassica napus* and will be made available to the general research community, with a per-sample cost expected to be less than US$100 for this chip.

The biological applications of SNP technology for both evolutionary and molecular geneticists as well as plant breeders and industry are far-reaching and will be invaluable to our understanding and advancement of the Brassica crop species.

2 Materials

2.1 In Silico SNP Discovery

The autoSNPdb database currently hosts SNPs for wheat, Brassica, barley, and rice and is available at http://autosnpdb.appliedbioinformatics.com.au/. The latest version of the wheat GBrowse database is available at www.wheatgenome.info.

2.2 Illumina GoldenGate Genotyping Assay

2.2.1 User-Supplied

1. DNA samples and controls.
2. 10 mM Tris–HCl pH 8.0, 1 mM EDTA (TE).
3. 2-Propanol.
4. Titanium *Taq* DNA polymerase.
5. Uracil DNA Glycosylase (UDG, optional).
6. 0.1 N NaOH.
7. 100 % EtOH.

2.2.2 Illumina-Supplied

1. MS1 reagent.
2. PS1 reagent.
3. RS1 reagent.
4. OB1 reagent.
5. OPA reagent.
6. AM1 reagent.
7. UB1 reagent.
8. MEL reagent.
9. MMP reagent.
10. IP1 reagent.
11. MPB reagent.
12. UB2 reagent.
13. MH1 reagent.
14. CHB reagent.
15. XC4 reagent.
16. PB1 reagent.
17. BeadChips.

2.2.3 Equipment

1. Qubit® Fluorometer (Invitrogen).
2. Qubit® dsDNA BR assay kit (Invitrogen).
3. GoldenGate Satellite Kit contents: shaker for microplates, hs, 230v (1×), fastener loop, adh, high temp (108×), fastener hook nylon (18×), Illumina hybridization oven (220v) (1×), Hybex 220v, w, microtube block (2×).
4. 96-well 0.2 mL skirted microtiter plate ×1 + 1 for balancing.

5. 0.45 μM clear styrene filter plate with lid.
6. 96-well V-bottom plate.
7. Filter plate adapter.
8. Multichannel pipettes.
9. 96-well cap mat 2× per plate.
10. Large centrifuge capable of accommodating plates.
11. Clear adhesive 96-well plate sealers.
12. Heat sealer.
13. Heat sealer, combi heat-sealing unit.
14. Adapter plate, combi heat-sealing unit (96-Well PCR Plate Carrier).
15. Heat-sealing foil sheets, Thermo-Seal.
16. BeadChip Wash Rack and Glass Tray.
17. Infinium Hybridization Chamber and Gasket 1×.
18. Wash dishes 2×.
19. Wash rack.
20. Multi-Sample BeadChip Alignment Fixture.
21. Vacuum desiccator.
22. Self-locking tweezers.
23. Staining rack.
24. HiScan machine.
25. Raised-bar magnetic plate.
26. Thermocycler/PCR machine.

2.3 Illumina Infinium HD Assay

2.3.1 Illumina-Supplied Reagents

Supplied in correct amounts for ordered assay:

1. ATM—Anti-Stain Two-Color Master Mix.
2. FMS—Fragmentation solution.
3. MA1—Multi-Sample Amplification 1 Mix.
4. MA2—Multi-Sample Amplification 2 Mix.
5. MSM—Multi-Sample Amplification Master Mix.
6. PB1—Reagent used to prepare BeadChips for hybridization.
7. PB2—Humidifying buffer used during hybridization.
8. PM1—Precipitation solution.
9. RA1—Resuspension, hybridization, and wash solution.
10. STM—Superior Two-Color Master Mix.
11. TEM—Two-Color Extension Master Mix.
12. XC1—Xstain BeadChip solution 1.

13. XC2—Xstain BeadChip solution 2.
14. XC3—Xstain BeadChip solution 3.
15. XC4—Xstain BeadChip solution 4.

2.3.2 User-Supplied Reagents

1. 0.1 N NaOH: Dissolve 4 g NaOH in 1 L water.
2. 100 % 2-propanol.
3. 100 % ethanol.
4. 95 % formamide/1 mM EDTA: Store at –20 °C.

2.3.3 Equipment

1. Qubit® Fluorometer (Invitrogen).
2. Qubit® dsDNA BR assay kit (Invitrogen).
3. GoldenGate Satellite Kit contents: microplate shaker, hs, 230v (1×), fastener loop, adh, high temp (108×), fastener hook, nylon (18×), Illumina hybridization oven (220v) (1×), Hybex 220v, w, microtube block 2×.
4. 96-well 0.8 mL microtiter plate ×1 + 1 for balancing.
5. Multichannel pipettes.
6. Cap mat 2×.
7. Large centrifuge capable of accommodating plates.
8. Foil seal.
9. Heat sealer.
10. Heat sealer, combi heat-sealing unit.
11. Adapter plate, combi heat-sealing unit (96-Well PCR Plate Carrier).
12. Heat-sealing foil sheets, Thermo-Seal.
13. BeadChip Wash Rack and Glass Tray.
14. Infinium Hybridization Chamber and Gasket 1×.
15. Te-Flow Flow-Through Chambers—4 per plate.
16. Wash dish 2×.
17. Wash rack.
18. Multi-Sample BeadChip Alignment Fixture.
19. Water Circulator.
20. Flow-Through Chamber with Illumina temperature probe.
21. Vacuum desiccator.
22. Self-locking tweezers.
23. Staining rack.
24. Wash dishes 2×.
25. BeadChips 4×.
26. HiScan machine.

3 Methods

3.1 Searching for SNPs in Wheat Disease-Associated Genes

1. Navigate to http://autosnpdb.appliedbioinformatics.com.au/ and select "wheat" from the drop-down menu and click option 2 (a) to search the database using keywords (*see* **Note 1**) (Fig. 1).
2. Type the keyword "disease" into box 3 and click the GO icon (*see* **Note 2**) (Fig. 2).
3. A total of 497 records were retrieved; to view an example record, select number 13. wheat_rep_c3572_a1_c1.
4. Figure 3 displays an overview of the positions of the SNPs in the top graphic with the SNP redundancy score represented by line height on the Y axis and SNP co-segregation score represented by the color of the line. Clicking on the SNP centers the alignment in the lower graphic on the SNP position. The lower graphic displays the sequence alignment and nucleotide sequence at the SNP position. Clicking "view sequence" in the top frame displays the consensus sequence for the assembly, while the lower frames may be expanded to view the annotation (Fig. 4).

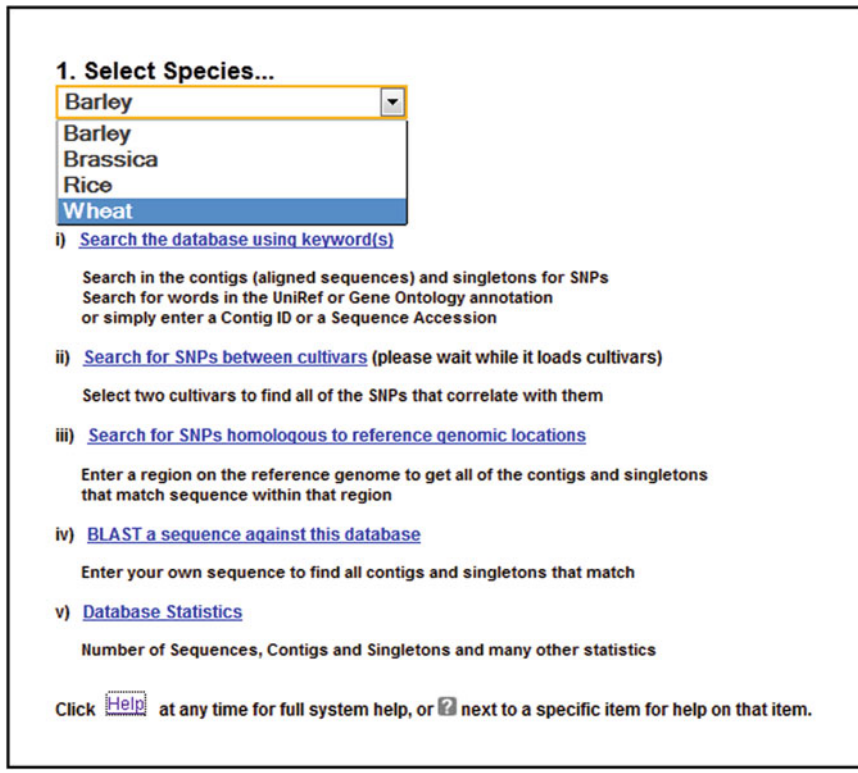


Fig. 1 Selection of species for autoSNPdb search

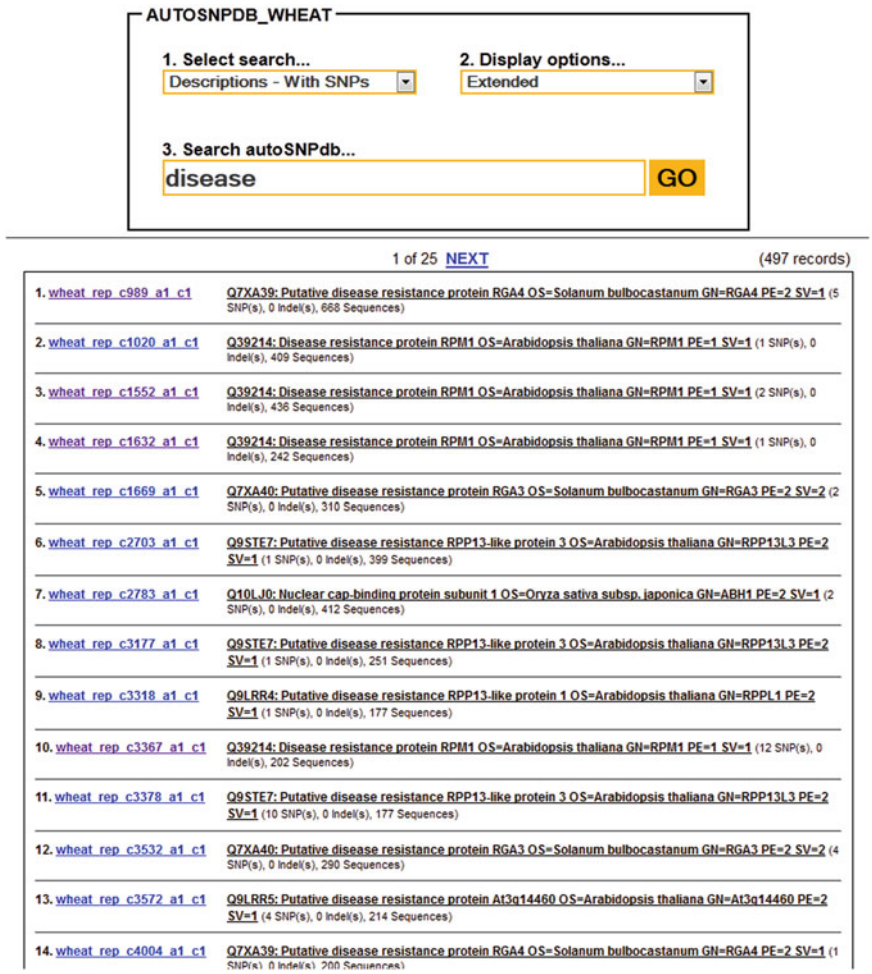


Fig. 2 Presentation of assemblies annotated with the term "disease" and which have predicted SNPs

3.2 Identification of SNPs Using Wheatgenome.info

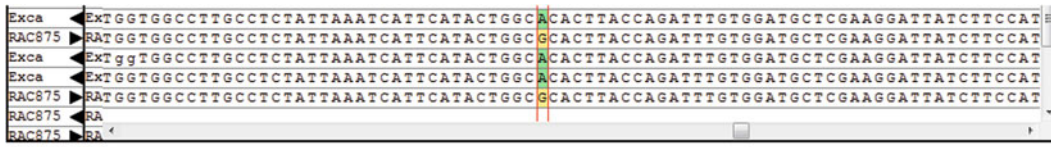
Assemblies and syntenic builds for each of the bread group 7 chromosome arms have been produced [27, 28] and are hosted in a GBrowse2 database at wheatgenome.info for public access prior to publication [29]. Each wheat chromosome arm has been annotated with predicted genes, UniRef90 gene similarities, as well as intervarietal SNPs discovered through the re-sequencing of Australian bread wheat varieties [30]. As well as annotation keyword searches, a BLAST portal enables sequence similarity searches of assembled wheat chromosome arm data. DNA or protein query sequence can be uploaded or pasted in the web-based form in FASTA format. The results are displayed in three sliding windows: the Overview window, Region window, and Details window. The reference view can be dragged and zoomed. Several tracks of annotation are available, including UniRef90, Genes, Contigs, SNPs, and Exons. All of these features can be expanded by clicking the

Fig. 3 Overview of record number 13. wheat_rep_c3572_a1_c1

associated plus button, and each feature provides a link to show the feature details (Fig. 1). In the absence of a finished wheat reference genome upon which to base crop improvement efforts, this database represents the first opportunity for wheat researchers to interact with chromosome-scale gene-based sequence scaffolds in an intuitive and user-friendly manner.

3.2.1 Searching for SNPs in Wheat Leucine-Rich Repeat Genes on Chromosome 7A Using Wheatgenome.info

1. Using a standard web browser, navigate to www.wheatgenome.info and select the link to the Wheat 7A GBrowse database.
2. Type lrr (abbreviation for leucine-rich repeat) and click the "search" icon.
3. A total of 78 regions are identified and presented graphically (Fig. 5).
4. As an example, select UniRef90_Q6R2Q1 which is the second link down on the syntenic build. This takes you to the annotation viewing window (Fig. 6). To expand the selection, select "show 20 kbp" from the drop-down window (*see* **Note 3**).



0 UniRef90 Annotation(s)

30 GenBank Annotation(s)

4 Gene Ontology Annotation(s)

GO Accession	Term Type	GO Name	GO Annotation Count
GO:0000166	molecular_function	nucleotide binding	4
GO:0005524	molecular_function	ATP binding	4
GO:0006915	biological_process	apoptosis	4
GO:0006952	biological_process	defense response	4

1 Reference Genome Location(s)

Genome	Chromosome	Region
Brachypodium	2	51395753-51398555

17 Swiss-Prot Annotation(s)

Name	Description	E-Value	Score (bits)
Q9LRR5	Putative disease resistance protein At3g14460 OS=Arabidopsis thaliana GN=At3g14460 PE=2 SV=1	4e-54	545
Q9LRR4	Putative disease resistance RPP13-like protein 1 OS=Arabidopsis thaliana GN=RPPL1 PE=2 SV=1	3e-44	459
Q7XA40	Putative disease resistance protein RGA3 OS=Solanum bulbocastanum GN=RGA3 PE=2 SV=2	3e-36	391
Q7XA39	Putative disease resistance protein RGA4 OS=Solanum bulbocastanum GN=RGA4 PE=2 SV=1	3e-35	382
Q7XBQ9	Disease resistance protein RGA2 OS=Solanum bulbocastanum GN=RGA2 PE=1 SV=1	8e-35	378
Q7XA42	Putative disease resistance protein RGA1 OS=Solanum bulbocastanum GN=RGA1 PE=2 SV=2	2e-33	367
Q39214	Disease resistance protein RPM1 OS=Arabidopsis thaliana GN=RPM1 PE=1 SV=1	4e-15	208
A9QGV6	Inactive disease susceptibility protein LOV1 OS=Arabidopsis thaliana GN=LOV1 PE=3 SV=1	8e-14	197
A7XGN8	Disease susceptibility protein LOV1 OS=Arabidopsis thaliana GN=LOV1 PE=1 SV=1	3e-13	192
Q9C646	Probable disease resistance protein RXW24L OS=Arabidopsis thaliana GN=RXW24L PE=2 SV=1	4e-13	191
O04093	Putative inactive disease susceptibility protein LOV1 OS=Arabidopsis thaliana GN=LOV1 PE=5 SV=3	5e-13	190
O22727	Probable disease resistance protein At1g61190 OS=Arabidopsis thaliana GN=At1g61190 PE=2 SV=1	2e-12	185
P0C8S1	Probable disease resistance RPP8-like protein 2 OS=Arabidopsis thaliana GN=RPP8L2 PE=1 SV=1	6e-12	181
Q8W474	Probable disease resistance protein At1g58390 OS=Arabidopsis thaliana GN=At1g58390 PE=2 SV=3	2e-11	176
Q9LVT4	Probable disease resistance protein At5g47250 OS=Arabidopsis thaliana GN=At5g47250 PE=2 SV=1	6e-11	172
O64973	Disease resistance protein RPS5 OS=Arabidopsis thaliana GN=RPS5 PE=1 SV=2	3e-10	166
Q9SX38	Putative disease resistance protein At1g50180 OS=Arabidopsis thaliana GN=At1g50180 PE=2 SV=1	7e-10	163

Fig. 4 autoSNPdb showing the overview of the SNPs in this assembly and the aligned sequences with the SNPs highlighted

5. The annotated UniRef gene is highlighted in yellow, and several SNPs can be observed in the viewer as red triangles. Additional SNPs can be seen by zooming out or scrolling to the left or right.
6. Click on a selection of red triangles to view information about each SNP (Fig. 7). Click on the contig at the base of Fig. 6 to download the related sequence information (Fig. 8) for the design of sequence-based SNP assays.

3.3 Illumina GoldenGate Genotyping Assay

This method describes the GoldenGate Genotyping Assay for 96 samples on three Illumina BeadChips.

The method describes the GoldenGate protocol for Single-Use DNA (SUD).

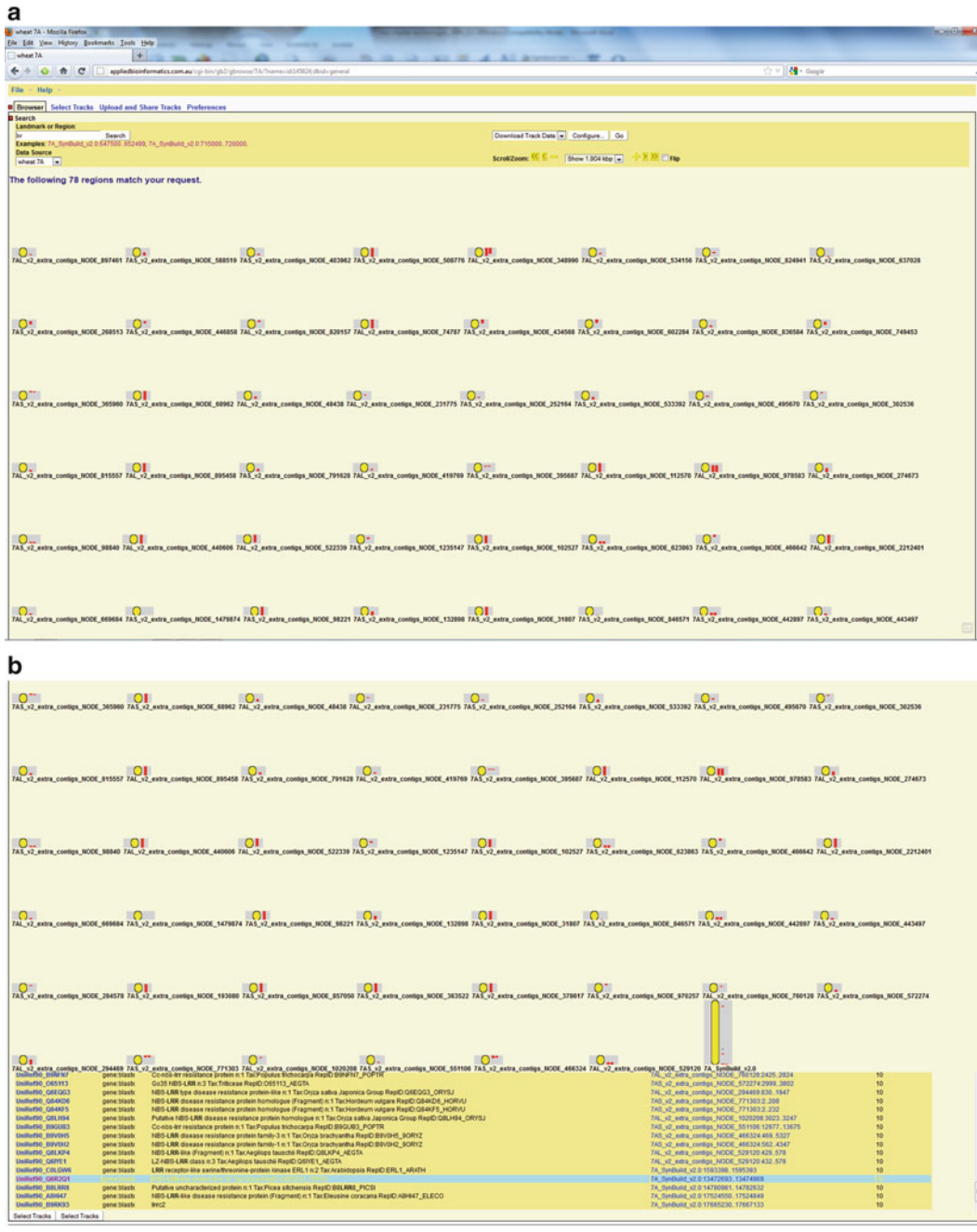


Fig. 5 The GBrowse search identified 78 regions which are displayed graphically as well as a list. Both graphic and list features link to further detailed information about the region

Unless stated, all centrifugation and vortexing steps are for 1 min.

1. Quantitate samples using the Qubit dsDNA BR assay (Invitrogen). Normalize all samples in a 96-well format to 50 ng/μL by adding Tris–HCl 10 mM, pH 8.5 (*see* **Note 5**).

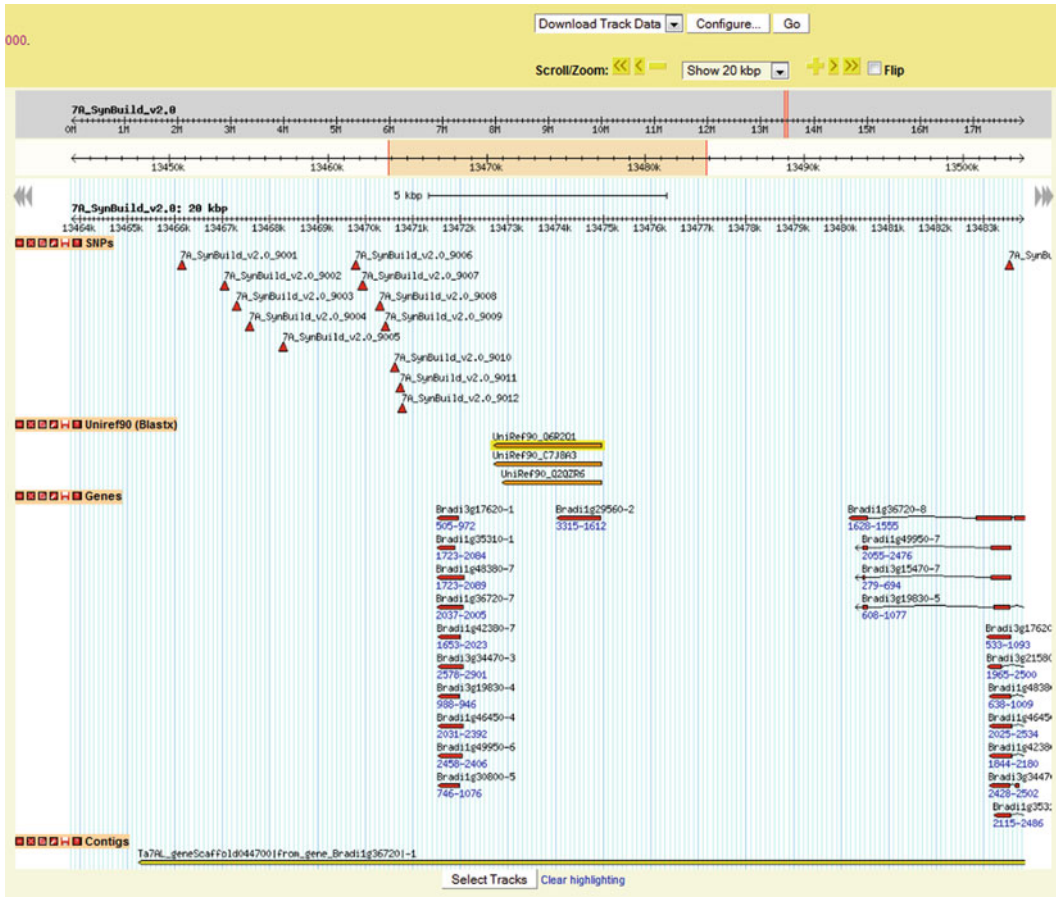
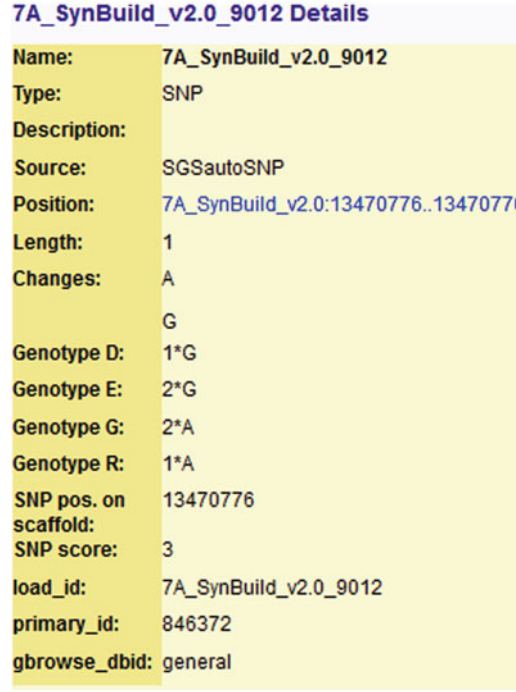


Fig. 6 The GBrowse viewer graphically represents features in tracks across the screen in relation to their respective genomic position

2. Dispense 5 µL MS1 reagent followed by 5 µL of normalized DNA sample into each well of the 0.2 mL plate.
3. Apply microplate foil heat seal to the plate using the heat sealer (3 s) ensuring all wells are completely sealed.
4. Pulse centrifuge plate to 250 × *g* and vortex at 2,300 rpm for 20 s, then pulse centrifuge a second time to 250 × *g* (*see* **Note 6**).
5. Incubate the plate at 95 °C for exactly 30 min.
6. Pulse centrifuge plate to 250 × *g* and carefully remove heat seal.
7. Dispense 5 µL PS1 reagent into each well and seal with a clear adhesive film before pulse centrifuging the plate to 250 × *g*.
8. Vortex at 2,300 rpm for 20 s or until solution is uniformly blue in color and then centrifuge the plate to 3,000 × *g* for 20 min (*see* **Note 7**).
9. Remove the plate from the centrifuge and decant the liquid by inverting the plate and smacking it down onto an absorbent



7A_SynBuild_v2.0_9012 Details

Name:	**7A_SynBuild_v2.0_9012**
Type:	SNP
Description:	
Source:	SGSautoSNP
Position:	7A_SynBuild_v2.0:13470776..13470776
Length:	1
Changes:	A
	G
Genotype D:	1*G
Genotype E:	2*G
Genotype G:	2*A
Genotype R:	1*A
SNP pos. on scaffold:	13470776
SNP score:	3
load_id:	7A_SynBuild_v2.0_9012
primary_id:	846372
gbrowse_dbid:	general

Fig. 7 Clicking on the red triangle displays information on each SNP in a new browser window, including the position in the reference and the genotype for each of the four varieties (*see* **Note 4**)

Ta7AL_geneScaffold044700|from_gene_Bradi1g36720|-1 Details

Name:	**Ta7AL_geneScaffold044700|from_gene_Bradi1g36720|-1**
Type:	contig
Description:	
Source:	SyntenicBuild
Position:	7A_SynBuild_v2.0:13465246..13490667 (- strand)
Length:	25422
load_id:	Ta7AL_geneScaffold044700|from_gene_Bradi1g36720|-1
primary_id:	1295
gbrowse_dbid:	general

```
>Ta7AL_geneScaffold044700|from_gene_Bradi1g36720|-1 class=Sequence position=7A_SynBuild_v2.0:13465246..13490667 (- strand)
AGTGACGCCA GGAAATCAAA CACTGGCTAA ATCGCTTGCT TGAGAGGATT TACACTGTGA AACTCCTCCT TTGTCAGGTA
CCAGAGGAAT ACATCTGATT GGAAAGTTTG GGGTTGTATG TTGTTTCTGA ATTCAGACCA TTAGTGTTCG TACTAGCTAC
CCGATGACTG GAGATATATT TGAGAAGTCT CCACTGTCCA GTCCAGTCTG ATGATAATAT TGCTCTGGAA GCCAAATGGA
TATGTCTTCA CTTTGAAAAT CATGTGCTGC AACGCATGAA GCCACAATCT GTTATTTTCA TGTACAGTTA TTGTCAGATG
TGAGTTTCTT TTGAAAAATG GCGAGGACTG CCTGTCTATC TGCAGCAATA AAATCCATAT ACTGCCAAGT TTACCATTTC
TCTTCGATAG TAATAACTGG GATCATGCCG ACAACCTGTG TAAGAAGCAG CATACCAATT ACTCATGGAT AATAATTGAG
ATTTTATTTC AGTTATCAGC TGCATGGTGT GGTGGGGTGC AGCCACTACT TTCCCCGAAT CATATCCCTG GAGCTTCGTT
TCAGGAGTAC ACATCAGCCA GCTTAAAGCC TGAAACATTG ATGGACCTGC TTGAAAGACA ATGACTTCAA TAATAGTTGA
ATTGTAGACC GAATAGAACA AGTTGCAATT CCAAGAAAAA GACAAGAAGT TATTGAGAGA AGGATACGTG AAACAATGAA
ATCCCTGCTG CAAAGTTGAC CATTTATTTC CCGAGAAACA ATTGATCCTG GGGACCTTTG TACGATATAT AGCAAACTAA
TTACTCCTAT ATATCATTGT ATTGTCACCA CACAACTGAA ATAAGTAACT TTTTTCACTT ATCTCCTGCA TGAAGGATGA
TGAAATTTGA GCAAGAGTGG CCCAAATTTG ACGATGCCGT GGGGTCTACA TCCTTATTTC AGGAGTATAC ATCAGCCAGC
TAAAATAAAA GATGATGCCG TGGGGTCTGC ATCCTTACTG TCATCTACAA CGCTCAAACT TAATGACAAT GACTTCTCAA
CTAACTATTA ATTGCATCAT TGACGAACTA GAACAAGCTA TCATTCTCAT GTAAGGGACA AGCGCCAAAT CGCCAATCAG
TTTCCTATTT AGTTATAGTA GAAGTATATC GCTTCTACCT ACCACTAGTT GATTCCTATA ATAGCGTCCA TTTGTGATCA
TTCCATAGAT CAACTCGAGT GCCTCAACAA ATATGGATGC ACGCTCTGCC TCTACTCCTA CTAAATTTGG TTCATCTGTT
TTTGCGCATC TCAGCTCGTG AATTCCTCTT GCCGGGCTCC TCTCTCTCCA TAGAGGATAG CTCCCATATG CTCCGTTCAC
CAGATGGCAC TTTCACCTGT GGCTTCAAAA ACATTTCCCG AAATGCCTCC GTCTTCTCTA TTTGGTTCTC CAATACAGTT
GACAAGACTG TCGTCTGGAG CGCAAATCAT CTCCATCCTG TGTACTCCTG GGGATCCCGA GTCATGCTAT ATGCGGATGG
CGGAATGGTT CTAGAGGATT ACAATGGACA GTCCGTGTGG GAAAACAGTG TCAACTCTTC ATCAACAGCT GTAAAAGCTC
AGTTATTGGA CACCGGGAAC CTCATCGTGA AGGGCCAAGG TGATATTATT CTATGGCAAA GCTTTCATTC ACCTACTGAT
ACATTGCTGC CCTATCAGAA CATTACCAGT GCTACCAAGT TGGTATCTAC TAGTAGGTTA CTTATTCCTG GGCGCTACAG
CTTTCATTTT GATGATCAAC ATCTACTCAC ATTGTTTGAT GATGAGAAGG ATATCTCTTT TATCTACTGG CCAAATCCTA
```

Fig. 8 Clicking on the contig opens a new window displaying the sequence of the contig enabling the design of sequence-based SNP assays

pad and tap the inverted plate onto the pad to blot excess supernatant (*see* **Note 8**).

10. Dry the plate by either centrifuging at 8×*g* for 1 min (with plate inverted) or allowing the plate to dry at room temperature for about 15 min.
11. To resuspend DNA, dispense 10 μL of RS1 reagent into each well of the plate and seal the plate with a clear adhesive microplate film.
12. Pulse centrifuge to 250×*g* and then vortex at 2,300 rpm for 1 min or until the blue pellet is completely resuspended.
13. Pulse centrifuge the plate to 250×*g*.
14. Dispense 10 μL of OPA reagent and 30 μL of OB1 reagent to each well of a new plate (called ASE plate).
15. Carefully remove the heat seal from the first plate (SUD plate), and transfer 10 μL of sample from each well into the corresponding well of the ASE plate.
16. Heat-seal the ASE plate using a microplate heat seal, ensuring all wells are completely sealed.
17. Pulse centrifuge the ASE plate to 250×*g* and then vortex at 1,600 rpm for 1 min or until all beads are completely resuspended.
18. Place the sealed plate on the 70 °C heat block and close the lid.
19. Immediately reset the temperature to 30 °C and allow the plate to cool to 30 °C for about 2 h. The plate may remain on the heating block for up to 16 h (*see* **Note 9**).
20. Remove the plate from the heating block and centrifuge to 250×*g*.
21. Place the plate on the raised-bar magnetic plate for around 2 min or until the beads are captured by the magnets to the side of the tube. Carefully remove the heat seal from the plate.
22. Using an 8-channel pipette with new tips, remove and discard the liquid from all the wells, leaving the beads. Inspect the pipette tips after removing the solution to ensure that no beads have been removed. If this has happened, return the solution to the same wells and wait for the magnet to re-collect the beads, before once more removing the liquid.
23. Dispense 50 μL of AM1 reagent into each well of the plate (still on raised-bar magnetic plate) before sealing the plate with microplate clear adhesive film.
24. Vortex the plate at 1,600 rpm for 20 s or until the beads are resuspended and place the plate on the magnetic plate for 2 min or until the beads are completely captured by the magnets.

25. Remove the clear film, taking care to avoid splashing from the wells. Remove the AM1 reagent from each well using an 8-channel pipette.
26. Repeat **steps 23–25** once.
27. Remove the plate from the raised-bar magnetic plate and dispense 50 μL of UB1 reagent to each well using an 8-channel pipette before placing the plate back on the magnetic plate until the beads are captured.
28. Remove the UB1 reagent from each well, ensuring no beads are removed.
29. Repeat **steps 27–28** once.
30. Remove the plate from the magnetic plate and dispense 37 μL of MEL reagent to each well and then seal with clear adhesive film and vortex the plate 1,600–1,700 rpm for 1 min or until beads are resuspended.
31. Incubate the plate on the preheated 45 °C heating block for exactly 15 min and leave at room temperature if continuing immediately or store at 4 °C for up to 1 h.
32. While plate is preheating (**step 31**), prepare the PCR plate (**steps 33–35**).
33. Add 64 μL of DNA polymerase into the MMP reagent tube as well as the optional 50 μL Uracil DNA Glycosylase and mix contents by inverting the tube.
34. Using an 8-channel pipette, add 30 μL of this mixture into each well of the PCR plate and seal with clear adhesive microplate film.
35. Centrifuge to 250 × *g* and place the PCR plate in a light-protected location (such as drawer/cupboard).
36. Remove plate from heating block (**step 31**) and place on the raised-bar magnetic plate until beads are captured and remove the clear adhesive film from the plate.
37. Remove the supernatant from all wells of the plate using an 8-channel pipette, leaving the beads in the wells.
38. Leaving the plate on the magnetic plate, add 50 μL of UB1 reagent to each well. Wait for all beads to be recaptured before removing the supernatant from all the wells of the plate.
39. Add 35 μL of IP1 to each well of the plate before sealing with clear adhesive film and vortexing at 1,800 rpm for 1 min or until beads are resuspended.
40. Place the plate on the 95 °C heating block for 1 min, then place on the magnetic plate until beads are captured.
41. Transfer 30 μL of supernatant from each well to the corresponding well of the PCR plate (made at **steps 32–35**) and discard plate with beads.

42. Seal the PCR plate with a PCR plate-sealing film appropriate for your thermocycler, transfer the PCR plate to the thermocycler, and run the following program (~2 h 45 min):
    (a) 37 °C for 10 min.
    (b) 95 °C for 3 min.
    (c) 3 cycles of:
        95 °C for 35 s.
        56 °C for 35 s.
        72 °C for 2 min.
    (d) 72 °C for 10 min.
    (e) Hold at 4 °C for 5 min.
43. Pulse centrifuge PCR plate to 250 × *g* before adding 20 μL resuspended MPB reagent into each well of the plate. Pipette solution in the PCR plate up and down several times to mix beads with PCR product.
44. Transfer mixed solution from each well into the corresponding well of a filter plate (70 μL solution in each well) and discard PCR plate.
45. Cover filter plate with its lid and store at room temperature, in a light-protected place, for 1 h.
46. Place the filter plate adapter onto an new 96-well V-bottom plate (waste plate) and place the filter plate containing the PCR plate onto the filter plate adapter.
47. Centrifuge to 1,000 × *g* for 5 min at 25 °C, remove filter plate lid, and add 50 μL UB2 reagent to each well, replace lid, and centrifuge to 1,000 × *g* for 5 min at 25 °C.
48. Using a multichannel pipette, add 30 μL MH1 reagent to each well of a new plate (INT plate) and replace the waste plate with the INT plate. Orient the INT plate so that well A1 of the filter plate matches well A1 of the INT plate. Discard waste plate.
49. Add 30 μL 0.1 N NaOH to each well of the filter plate, and centrifuge to 1,000 × *g* for 5 min at 25 °C. At the end, no beads should be visible in the wells of the INT plate (bottom plate in filter assembly).
50. Gently mix the contents of the INT plate by moving side to side without splashing and then seal the INT plate with 96-well cap mat and store in the dark until ready to dispense onto a BeadChip.
51. Dispense 15 μL of each sample (from INT plate) onto the inlet ports along each side of the BeadChips, ensuring that each sample flows to cover the entire bead stripe.
52. Position the BeadChips in the Hyb Chamber (containing CHB reagent as a humidifying buffer) and replace the lid. Correctly orientate the chamber in the oven and incubate for exactly 30 min at 60 °C with the rocker at speed setting 5.

53. After 30 min, reset the temperature to 45 °C and incubate for 16–18 h.
54. Remove the BeadChips from the oven and cleanly remove the IntelliHyb seals from the BeadChips, one at a time, before sliding into the prepared wash rack submerged in the dish containing PB1 reagent.
55. Wash the chips by gentle agitation of the wash rack for 1 min before transferring to a second wash dish containing PB1. Repeat the 1 min agitation wash step.
56. Transfer to a third dish containing XC4 reagent and slowly move the wash rack up and down ten times. Let it soak for 5 min.
57. Remove the wash rack to a tube rack in one smooth, rapid motion and use self-locking tweezers to slide each BeadChip from the wash rack to the tube rack.
58. Place the entire tube rack in a vacuum desiccator and start the vacuum, using at least 508 mmHg. Dry under vacuum for 50–55 min.
59. Image BeadChips on HiScan system.
60. Import data to GenomeStudio software.
61. Analyze results.

3.4 Illumina Infinium HD Assay

While this protocol is written from the perspective of assaying 96 samples using Infinium chips which hold 24 samples each, other combinations are possible and can be easily accommodated into the protocol.

Unless stated, all centrifugation and vortexing steps are for 1 min.

1. Quantitate samples using the Qubit dsDNA BR assay (Invitrogen). Normalize all samples in a 96-well PCR plate to 50 ng/μL by adding Tris–HCl 10 mM, pH 8.5 (*see* **Note 5**).
2. Dispense 20 μL MA1 followed by 4 μL DNA sample into each well of the 0.8 mL plate and seal with a cap mat.
3. Vortex at 1,600 rpm, centrifuge at 280×*g*, then incubate at room temperature for 10 min.
4. Dispense 34 μL MA2 and 38 μL MSM into each well before resealing for vortexing and centrifuging as in **step 3**.
5. Incubate resealed plate in a 37 °C oven for 20–24 h.
6. Before opening the plate, centrifuge at 50×*g*.
7. Add 25 μL FMS to each well, reseal, and vortex as before. Centrifuge again at 50×*g*.
8. Incubate on a heating block at 37 °C for 1 h.
9. Add 50 μL PM1 to each well, seal, and vortex as before.
10. Incubate for a further 5 min on the 37 °C heating block. Centrifuge at 50×*g*.

11. Add 155 μL 2-propanol to each well then seal plate with the second, fresh cap mat.
12. Mix by inverting the plate at least ten times then incubate at 4 °C for 30 min.
13. Prepare a balance plate before centrifuging at 2,000 × *g* and 4 °C for 20 min. This should produce pale blue pellets in the bottom of the wells (*see* **Note 10**).
14. Immediately decant supernatant by smoothly and rapidly inverting the plate onto an absorbent pad prepared on the bench. Remove all liquid by tapping the plate firmly for 1 min on the pad. Ensure the pellets are completely dry by leaving the plate inverted at room temperature for 1 h.
15. Resuspend pellets in 23 μL RA1 then seal with a foil seal using a heat sealer.
16. Incubate in a 48 °C oven for 1 h.
17. Vortex plate at 1,800 rpm then centrifuge at 280 × *g*.
18. Prepare and assemble Hyb Chamber as recommended by Infinium user manual, including adding 400 μL PB2 to each of the 8 reservoirs.
19. Denature samples by incubating on a 95 °C heating block for 20 min. Incubation for a further 30 min at room temperature is then followed by centrifuging at 280 × *g*.
20. Prepare 4 BeadChips by unpackaging and placing in Hyb Chamber inserts.
21. Dispense 12 μL of each sample onto the inlet ports along each side of the BeadChips, ensuring that each sample flows to cover the entire bead stripe.
22. Position the BeadChips in the Hyb Chamber and replace the lid. Correctly orientate the chamber in the oven and incubate for 16–24 h at 48 °C with the rocker at setting 5.
23. Prepare XC4 reagent for the following day by adding 330 mL 100 % ethanol and shaking vigorously to mix. Leave at room temperature until needed.
24. Before opening the chamber, allow to cool on the bench for 25 min.
25. Cleanly remove the IntelliHyb seals from the BeadChips, one at a time, before sliding into the prepared wash rack submerged in the dish containing PB1. It is important that the chips should not be allowed to dry out before the Flow-Through Chamber is assembled.
26. Wash the chips by gentle agitation of the wash rack for 1 min before transferring to a second wash dish containing PB1. Repeat the 1 min agitation wash step.

27. Prepare Multi-Sample BeadChip Alignment Fixture containing PB1. Transfer the BeadChips to the Alignment Fixture and assemble Flow-Through Chamber comprising the clear spacers, glass back plates, and metal clamps, using the Alignment Bar to correctly align components. Trim excess spacer.
28. Prepare Chamber Rack connected to the Water Circulator so that the temperature is 44 °C, calibrated with an Illumina® Te-Flow Thermometer Assembly. Place Flow-Through Chamber assemblies into Chamber Rack once the desired temperature is reached.
29. Perform the Single-Base Extension section of the protocol without interruption by dispensing the following reagents into the reservoir of each Chamber assembly (**steps 30–38**).
30. Add150 μL RA1. Incubate for 30 s. Repeat 5 times.
31. Add 450 μL XC1. Incubate for 10 min.
32. Add 450 μL XC2. Incubate for 10 min.
33. Add 200 μL TEM. Incubate for 15 min.
34. Add 450 μL 95 % formamide/1 mM EDTA. Incubate for 1 min. Repeat once.
35. Incubate 5 min.
36. Begin ramping the Chamber Rack temperature to the temperature indicated on the STM tube or to 37 °C if none is shown.
37. Add 450 μL XC3. Incubate for 1 min. Repeat once.
38. Wait for the Chamber Rack to reach the desired temperature before continuing.
39. Once the second temperature has been reached, continue with the staining section of the protocol by dispensing the following reagents into the reservoir of each Chamber assembly (**steps 40–50**).
40. Add 250 μL STM and incubate for 10 min.
41. Add 450 μL XC3 and incubate for 1 min. Repeat once, and then wait 5 min.
42. Add 250 μL ATM and incubate for 10 min.
43. Add 450 μL XC3 and incubate for 1 min. Repeat once, and then wait 5 min.
44. Add 250 μL STM and incubate for 10 min.
45. Add 450 μL XC3 and incubate for 1 min. Repeat once, and then wait 5 min.
46. Add 250 μL ATM and incubate for 10 min.
47. Add 450 μL XC3 and incubate for 1 min. Repeat once, and then wait 5 min.

48. Add 250 μL STM and incubate for 10 min.
49. Add 450 μL XC3 and incubate for 1 min. Repeat once, and then wait 5 min.
50. Move the Chamber assemblies to the lab bench and place horizontally.
51. Carefully disassemble the Chamber assemblies, one at a time, and place the BeadChips in the prepared staining rack submerged in the wash dish containing PB1. Perform staining by moving the rack up and down ten times then leave to incubate for a further 5 min.
52. Transfer to a second wash dish containing freshly poured XC4. Repeat the staining process.
53. Remove the staining rack to a tube rack in one smooth, rapid motion and use self-locking tweezers to slide each BeadChip from the staining rack to the tube rack.
54. Place the entire tube rack in a vacuum desiccator and start the vacuum, using at least 508 mmHg. Dry under vacuum for 50–55 min.
55. Image BeadChips on HiScan system.
56. Import data to GenomeStudio software.
57. Analyze results.

4 Notes

1. In addition to searching the database using keywords, it is possible to search for SNPs which differentiate between two cultivars, search for genes with similarity to regions of a reference genome, or search using a nucleotide sequence using BLAST.
2. By selecting the appropriate option in drop-down menu 1, it is also possible to search annotations for all sequences, not only those which contain SNPs, as well as limit the search to gene ontologies, sequence accession, or contig ID.
3. If all the tracks are not visible, click on "Select Tracks" at the top of the page and tick the boxes to view the relevant tracks.
4. At the time of writing, the database holds SNP information for four wheat varieties. Data for a further 12 varieties is being processed and is expected to go online by the end of 2012.
5. While 50 ng/μL is the ideal concentration, concentrations in the range of 10–100 ng/μL can be accommodated. The more important property is the quality of the DNA. It is recommended to run the samples on a 1 % agarose gel before use in this assay to ascertain any levels of degradation or contamination.

6. When using the vortex, ensure that the plate is firmly strapped to the vortex platform to prevent the plate from moving. You may place a hand on the plate during vortexing to ensure this.
7. If there is any delay before continuing on to **step 9**, repeat this step.
8. This step must be done firmly enough to decant all the supernatant from the wells. It may not work if done lightly and cross-contamination may occur.
9. This allows for the user to leave the plate overnight and continue the protocol the following day.
10. If there is any delay before continuing on to **step 14**, repeat this step.

References

1. Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100
2. Edwards D, Forster JW, Chagné D, Batley J (2007) What are SNPs? In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association mapping in plants. Springer, New York, pp 41–52
3. Batley J, Edwards D (2007) SNP applications in plants. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) Association mapping in plants. Springer, New York, pp 95–102
4. Edwards D, Forster JW, Cogan NOI, Batley J, Chagné D (2007) Single nucleotide polymorphism discovery. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) Association mapping in plants. Springer, New York, pp 53–76
5. Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942
6. Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535
7. Edwards D (2007) Bioinformatics and plant genomics for staple crops improvement. In: Kang MS, Priyadarshan PM (eds) Breeding major food staples. Blackwell, Oxford, UK, pp 93–106
8. Duran C, Appleby N, Edwards D, Batley J (2009) Molecular genetic markers: discovery, applications, data storage and visualisation. Curr Bioinformatics 4:16–27
9. Duran C, Eales D, Marshall D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Appleby N, Batley J, Basford K, Edwards D (2010) Future tools for association mapping in crop plants. Genome 53:1017–1023
10. Batley J, Edwards D (2009) Mining for single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) molecular genetic markers. In: Posada D (ed) Bioinformatics for DNA sequence analysis. Humana, Totowa, NJ, USA, pp 303–322
11. Batley J, Edwards D (2009) Genome sequence data: management, storage, and visualization. Biotechniques 46:333–336
12. Imelfort M, Duran C, Batley J, Edwards D (2009) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnol J 7:312–317
13. Berkman PJ, Lai K, Lorenc MT, Edwards D (2012) Next generation sequencing applications for wheat crop improvement. Am J Bot 99:365–371
14. Lee H, Lai K, Lorenc MT, Imelfort M, Duran C, Edwards D (2012) Bioinformatics tools and databases for analysis of next generation sequence data. Brief Funct Genomics 2:12–24
15. Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422
16. Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91
17. Savage D, Batley J, Erwin T, Logan E, Love CG, Lim GAC, Mongin E, Barker G, Spangenberg GC, Edwards D (2005) SNPServer: a real-time SNP discovery tool. Nucleic Acids Res 33:W493–W495
18. Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D (2009) AutoSNPdb: an annotated single nucleotide

polymorphism database for crop plants. Nucleic Acids Res 37:D951–D953

19. Duran C, Appleby N, Vardy M, Imelfort M, Edwards D, Batley J (2009) Single nucleotide polymorphism discovery in barley using autoSNPdb. Plant Biotechnol J 7:326–333
20. Lai K, Lorenc MT, Edwards D (2012) Genomic databases for crop improvement. Agronomy 2: 62–73
21. Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J 10:743–749
22. Hayward, A.; Vighnes, h. G.; Delay, G.; Samian, M. R.; Manoli, S. M.; Stiller, J.; McKenzie, M.; Edwards, D.; Batley, J. (2012) Second generation sequencing for gene discovery in the brassicaceae. *Plant Biotechnol J*. 10(6):750–759
23. Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotechnol J 7:1–8
24. Arnaoudova EG, Bowens PJ, Chui RG, Dinkins RD, Hesse U, Jaromczyk JW, Martin M, Maynard P, Moore N, Schardl CL (2009) Visualizing and sharing results in bioinformatics projects: GBrowse and GenBank exports. BMC Bioinformatics 10(Suppl 7):A4
25. Meintjes P, Duran C, Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Thierer T, Ashton B, Heled J (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649
26. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, Chang LY, Huang W, Liu B, Shen Y, Tam PKH, Tsui LC, Waye MMY, Wong JTF, Zeng CQ, Zhang QR, Chee MS, Galver LM, Kruglyak S, Murray SS, Oliphant AR, Montpetit A, Hudson TJ, Chagnon F, Ferretti V, Leboeuf M, Phillips MS, Verner A, Kwok PY, Duan SH, Lind DL, Miller RD, Rice JP, Saccone NL, Taillon-Miller P, Xiao M, Nakamura Y, Sekine A, Sorimachi K, Tanaka T, Tanaka Y, Tsunoda T, Yoshino E, Bentley DR, Deloukas P, Hunt S, Powell D, Altshuler D, Gabriel SB, Qiu RZ, Ken A, Dunston GM, Kato K, Niikawa N, Knoppers BM, Foster MW, Clayton EW, Wang VO, Watkin J, Gibbs RA, Belmont JW, Sodergren E, Weinstock GM, Wilson RK, Fulton LL, Rogers J, Birren BW, Han H, Wang HG, Godbout M, Wallenburg JC, Larcheveque P, Bellemare G, Todani K, Fujita T, Tanaka S, Holden AL, Lai EH, Collins FS, Brooks LD, McEwen JE, Guyer MS, Jordan E, Peterson JL, Spiegel J, Sung LM, Zacharia LF, Kennedy K, Dunn MG, Seabrook R, Shillito M, Skene B, Stewart JG, Valle DL, Clayton EW, Jorde LB, Belmont JW, Chakravarti A, Cho MK, Duster T, Foster MW, Jasperse M, Knoppers BM, Kwok PY, Licinio J, Long JC, Marshall PA, Ossorio PN, Wang VO, Rotimi CN, Royal CDM, Spallone P, Terry SF, Lander ES, Lai EH, Nickerson DA, Abecasis GR, Altshuler D, Bentley DR, Boehnke M, Cardon LR, Daly MJ, Deloukas P, Douglas JA, Gabriel SB, Hudson RR, Hudson TJ, Kruglyak L, Kwok PY, Nakamura Y, Nussbaum RL, Royal CDM, Schaffner SF, Sherry ST, Stein LD, Tanaka T, Int HapMap C (2003) The international HapMap project. Nature 426:789–796
27. Berkman PJ, Skarshewski A, Lorenc MT, Lai K, Duran C, Ling EYS, Stiller J, Smits L, Imelfort M, Manoli S, McKenzie M, Kubalakova M, Simkova H, Batley J, Fleury D, Dolezel J, Edwards D (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. Plant Biotechnol J 9:768–775
28. Berkman PJ, Skarshewski A, Manoli S, Lorenc MT, Stiller J, Smits L, Lai K, Campbell E, Kubalakova M, Simkova H, Batley J, Dolezel J, Hernandez P, Edwards D (2012) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. Theor Appli Genet 124:423–432
29. Lai K, Berkman PJ, Lorenc MT, Duran C, Smits L, Manoli S, Stiller J, Edwards D (2012) WheatGenome.info: an integrated database and portal for wheat genome information. Plant Cell Physiol 53:1–7
30. Edwards D, Wilcox S, Barrero RA, Fleury D, Cavanagh CR, Forrest KL, Hayden MJ, Moolhuijzen P, Keeble-Gagnère G, Bellgard MI, Lorenc MT, Shang CA, Baumann U, Taylor JM, Morell MK, Langridge P, Appels R, Fitzgerald A (2012) Bread matters: a national initiative to profile the genetic diversity of Australian wheat. Plant Biotechnol J 10: 703–708

Chapter 9

Development of Microsatellite-Enriched Libraries

Hélène Vignes and Ronan Rivallan

Abstract

Among the molecular markers used for plant genetic studies, microsatellite markers are easy to implement and can provide suitable codominant markers for molecular taxonomy. Here we describe a method to obtain enriched libraries in microsatellite loci from genomic DNA, using capture method with synthetic oligonucleotide probes.

Key words Microsatellite, Genomic-enriched libraries

1 Introduction

Microsatellites, also called simple sequence repeats (SSRs) [1], are small repeats of one, two, three, or four tandemly arranged nucleotides that are ubiquitous components of eukaryotic genomes. They have a high level of polymorphism due to mutation affecting the number of repeat units. Their variable length polymorphism can be revealed by polymerase chain reaction (PCR) [2] with unique flanking primers [3] that generate codominant markers. Microsatellites have a Mendelian heritability [4] and have a potential advantage of reliability, reproducibility, discrimination, standardization, and cost-effectiveness [5]. All these characteristics make them a suitable tool for genetic analysis, diversity analysis, population structure studies, genetic mapping, and quantitative traits analysis.

Since the early 1990s, microsatellite loci have been obtained by screening plant genomic libraries. Genomic libraries contain only 0.5–8 % of microsatellite inserts, as reported by previous authors [6, 7], this low rate means that it is necessary to check an important number of clones for their discovery. The use of enriched library has been proposed, and different methods are available to construct a microsatellite-enriched library, all based on the hybridization of a synthetic oligonucleotide microsatellite sequence onto genomic DNA. These libraries increase the level of positive clones from at least 20–90 % [8–10]. Here we expose a common protocol

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_9,

to produce genomic libraries enriched in microsatellites loci; this protocol was first used and published for tropical crops [11] and has since been used to obtain microsatellites markers on more than 200 plant genomes. We choose to use the two more simplified dinucleotide microsatellite motifs GA and GT with no self-hybridization capability. The main idea is to restrict the DNA with a four base pair recognition site endonuclease, hybridize with two GA and GT synthetic biotinylated oligoprobes for microsatellite capture and to clone the enriched DNA fragments.

2 Materials

All steps need to use ultrapure water prepared to obtain a sensitivity of 18 MΩ cm at 25 °C. Solutions resulting of the different protocol steps must be stored at −20 °C.

2.1 DNA Restriction

1. *Rsa*I restriction endonuclease and reaction buffer: 10 mM Bis-Tris-Propane-HCl, 10 mM $MgCl_2$, 1 mM Dithiothreitol pH 7.0.
2. TAE buffer 1×: 0.04 M Tris-acetate and 0.001 M EDTA pH 8.0.
3. Agarose gel 1.2 %, weigh 0.6 g biomolecular grade agarose, transfer to a 200 mL erlen, add 50 mL 1× TAE, and dissolve in a microwave oven. Carefully poor the gel when the temperature is below 60 °C.
4. 40 mM spermidine.
5. Ethidium bromide 0.5 μg/L bath (*see* **Note 1**). The ethidium bromide bath must be prepared by adding 500 μL of a commercial 10 mg/L solution to 1 L of ultrapure water.

2.2 Adaptor Ligation and PCR Preamplification

1. Synthetic oligonucleotide primers proposed by Edwards et al. [10] (*see* **Note 2**) used for the ligation of a blunt self-complementary adaptor.
2. Rsa21 primer 10 μM 5′-CTCTTGCTTACGCGTGGACTA-3′.
3. Rsa25 primer 10 μM 5′-pTAGTCCACGCGTAAGCAAGA GCACA-3′.
4. T4 DNA ligase and 5× reaction buffer: 250 mM Tris–HCl, pH 7.6, 50 mM $MgCl_2$, 5 mM ATP, 25 % polyethylene glycol-8000.
5. *Taq* DNA polymerase 5 unit/μL (Sigma).
6. 10× dNTP 2 mM.
7. 10× reaction buffer: 100 mM Tris–HCl pH 8.3, 500 mM KCl, 15 mM $MgCl_2$.
8. Use elements 2, 3, and 4 from Subheading 2.1.

2.3 Purification of the Preamplification Product Using QIAquick PCR Purification Kit

1. QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).
2. 1.5 or 2 mL microcentrifuge tubes.
3. 3 M sodium acetate, pH 5.0.
4. Before use, add ethanol (96–100 %) to the provided PE buffer.

2.4 Selection of DNA Fragments Containing Microsatellites

1. Streptavidin MagneSphere Paramagnetic Particles (Promega, Madison, USA) (*see* **Note 3**).
2. Separation Stand (12-position) 1.5 mL (Promega).
3. Microsatellite 5′biotin-labelled oligo probes: biotin $I_5(GA)_8$ and biotin $I_5(GT)_8$ (*see* **Note 4**).
4. Heating block.
5. Magnet Separation Stand (12-position) 1.5 mL (Promega).
6. 20× SSC (150 mM NaCl, 15 mM sodium citrate). To prepare 1 L, dissolve 175.3 g of NaCl and 88.2 g of sodium citrate in 800 mL water. Adjust the pH to 7.0 with a few drops of 14 N solution of HCl. Adjust the volume to 1 L with ultrapure water. Dispense into aliquots. Sterilize by autoclaving.
7. 0.5× SCC and 0.1× SSC.

2.5 PCR Amplification of the Selected Fragments

1. Rsa21 primer 10 μM.
2. Elements 3, 4, and 5 from Subheading 2.2.
3. Elements 2, 3, and 4 from Subheading 2.1 are also needed.

2.6 Cloning in pGEM-T

Use pGEM®-T Vector System I Promega containing:

1. pGEM®-T Vector (50 ng/μL).
2. 12 μL Control Insert DNA (4 ng/μL).
3. 100 units T4 DNA ligase.
4. 200 μL 2× rapid ligation buffer: 60 mM Tris–HCl pH 7.8, 20 mM $MgCl_2$, 20 mM DTT, 2 mM ATP, 10 % polyethylene glycol-8000.

2.7 Transformation in Bacteria

1. LB agar plates: mix 10.0 g tryptone, 5 g yeast extract, 10.0 g NaCl, and 20 g of agar in approximately 700 mL of water. Bring the final volume to 1 L. Adjust pH to 7.0 with 5 N NaOH. Pour this into a 2 L flask and autoclave 20 min.
   Let LB agar cool to ~55 °C and add ampicillin to a final concentration of 100 mg/L, 5-bromo-4-chloro-3-inodlyl-β-D-galactopyranoside (X-gal) to a final concentration of 80 μg/mL (prepared in dimethylformamide (DMF)), and isopropyl-1-thio-β-D-galactopyranoside (IPTG) to a final concentration of 20 mM (prepared in sterile dH_2O).
2. SOB medium: mix 20.0 g of tryptone, 5.0 g of yeast extract, and 0.5 g of NaCl in deionized H_2O to a final volume of 1 L. Pour this into a 2 L flask and autoclave 20 min. Add 10 mL of filter-sterilized

1 M $MgCl_2$ and 10 mL of filter-sterilized 1 M $MgSO_4$ prior to the preparation of the SOC medium (per 100 mL).

3. SOC medium: prepare fresh immediately before use. Add 1 mL of filter-sterilized 2 M glucose in previously prepared SOB medium.
4. BD Falcon polypropylene round-bottom tubes.
5. XL1-Blue competent cells (Stratagene) containing bacteria, pUC18 control plasmid (0.1 ng/μL in TE buffer), and β-Mercaptoethanol solution (1.42 M).

2.8 Positive Clone Screening for Microsatellite Motif and Preparation for Sequencing

1. Synthetic oligonucleotide primers: four primers are needed for positive clones screening, two from the flanking regions of the cloning site and two for the microsatellite motif hybridization.
   AGE1 5′-AAACAGCTATGACCATGATTAC-3′ 100 μM.
   AGE2 5′-TTGTAAAACGACGGCCAGTG-3′ 100 μM.
   $(GA)_{12}$ 100 μM.
   $(GT)_{12}$ 100 μM.
2. Elements 3, 4, and 5 from Subheading 2.2.
3. Use elements 2, 3 (but prepare a 1 % agarose gel), and 4 from section.
4. Laminar flux cabinet.
5. PCR 96 well plate.
6. Autoclaved toothpick.
7. 96 Tips replicator.

3 Methods

The enrichment technique is illustrated in Fig. 1.

3.1 DNA Restriction

1. Aliquot 40 μL (250 ng/μL) of total genomic DNA solution into an Eppendorf tube. Mix 100 units of RsaI endonuclease (10 unit/μg of DNA), 10 μL of reaction buffer, 10 μL of 40 mM spermidine, and ultrapure water to 100 μL. Incubate overnight at 37 °C.
2. Control the restriction by agarose gel electrophoresis. Add 1 μL of loading dye to 10 μL of the restriction and load on a 50 mL 1.2 % agarose gel, running at 60 V, stop the migration after 20 min, reveal by coloration using ethidium bromide (*see* **Note 1**), and take a picture. A regular smear must be obtained without band Fig. 2.

3.2 Adaptor Ligation and Preamplification

1. Mix 10 μL of restricted DNA, 2 μL of oligo RSA21 (10 μM), 2 μL of oligo RSA25 (10 μM), 20 μL of reaction buffer, 4 units of T4 DNA ligase, and ultrapure water to 100 μL. Incubate 2 h at 20 °C.

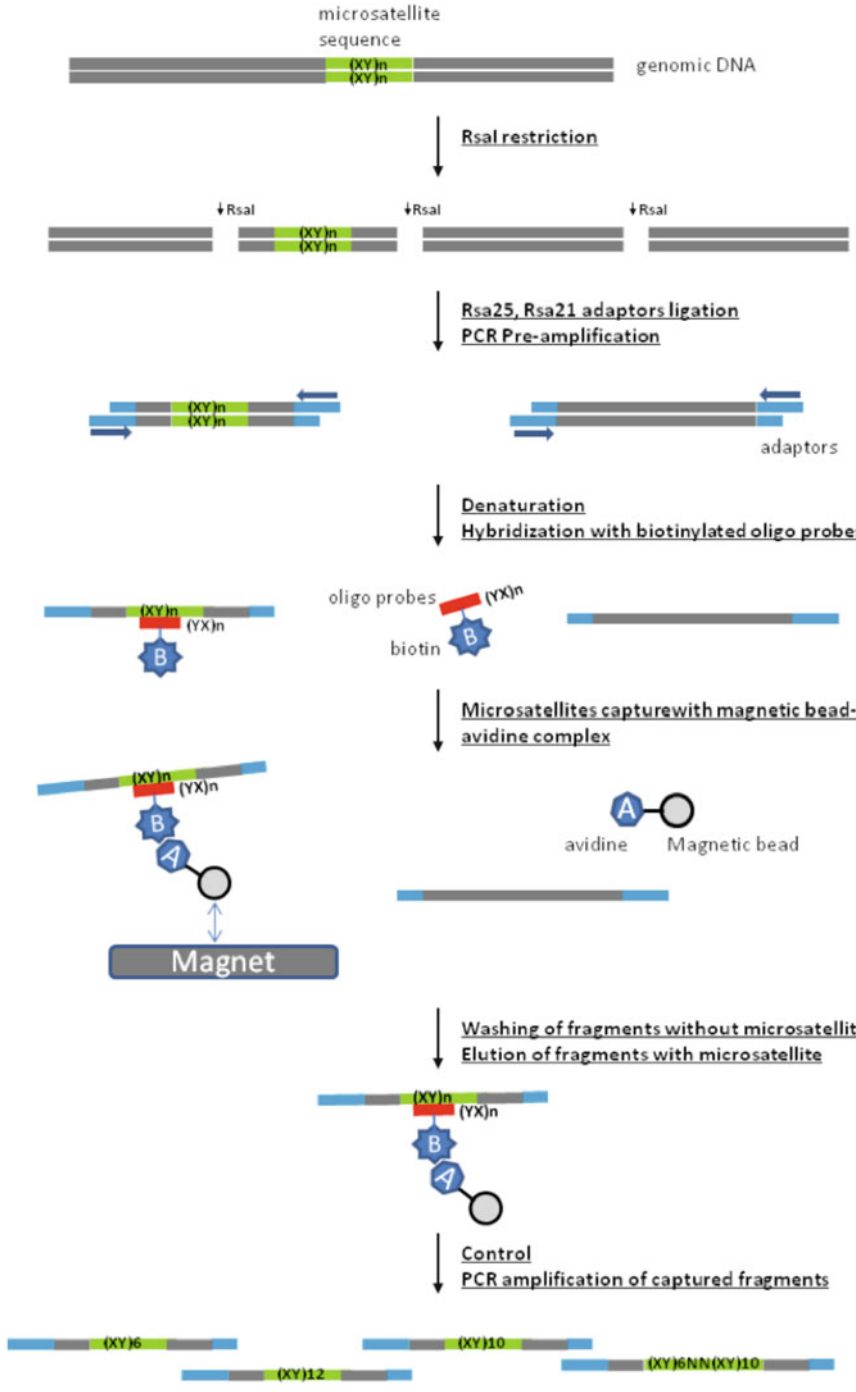


Fig. 1 Diagram of the genomic library enrichment technique in microsatellite loci

2. Mix 2 μL of the ligated DNA solution, 2 μL of dNTP, 2.5 μL of reaction buffer, 1 unit of *Taq* DNA polymerase, and ultrapure water to 25 μL.

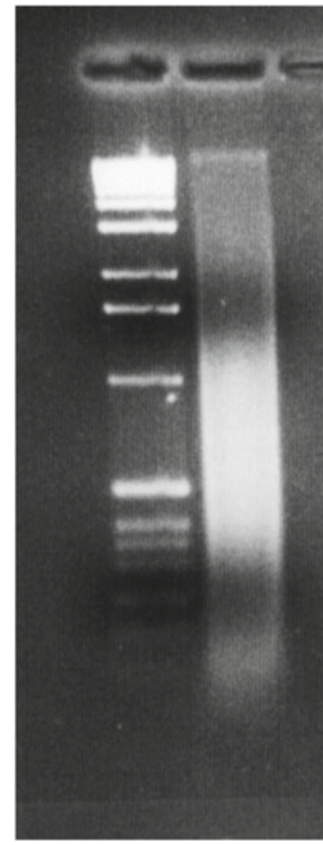


Fig. 2 Migration on a 1.2 % agarose gel of the restricted genomic DNA

3. Run the PCR program: 1 cycle at 95 °C 4 min followed by 20 cycles with 94 °C 30 s, 60 °C 1 min, 72 °C 2 min, and a final step at 72 °C for 8 min.
4. Mix 10 µL of the amplification product with 1 µL of loading dye and load in a 50 mL 1.2 % agarose gel, run at 60 V for 45 min.

   The gel must reveal a smear between 100 and 1,500 bp like on Fig. 3.

3.3 Purification of the Preamplification Product Using QIAquick PCR Purification Kit

All the centrifugation must be performed at 10,000 × *g*.

1. Add 5 volumes of Buffer PB to 1 volume of the PCR sample and mix. For example, add 500 µL of Buffer PB to 100 µL PCR product.
2. Check that the color of the mixture is yellow. If the color of the mixture is orange or violet, add 10 µL of 3 M sodium acetate, pH 5.0, and mix. The color of the mixture will turn to yellow.
3. Place a QIAquick spin column in a provided 2 mL collection tube.

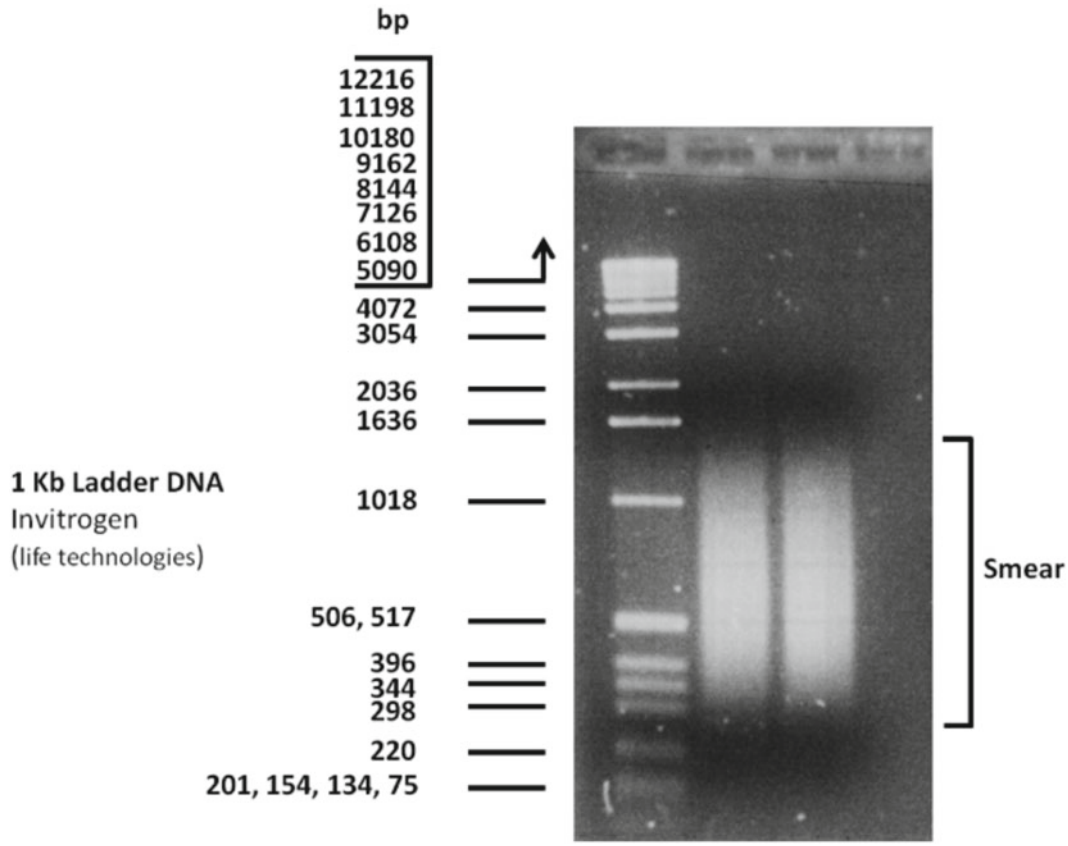


Fig. 3 Migration on a 1.2 % agarose gel of the preamplification of ligated DNA fragments

4. To bind DNA, apply the sample to the QIAquick column and centrifuge for 30–60 s.
5. Discard flow-through. Place the QIAquick column back into the same tube. Collection tubes are reused to reduce plastic waste.
6. To wash, add 0.75 mL Buffer PE to the QIAquick column and centrifuge for 30–60 s.
7. Discard flow-through and place the QIAquick column back in the same tube. Centrifuge the column for an additional 1 min.
8. Place QIAquick column in a clean 1.5 mL microcentrifuge tube.
9. To elute DNA, add 50 μL of water to the center of the QIAquick membrane (*see* **Note 5**) and centrifuge the column for 1 min; repeat the elution with 50 μL of water.

3.4 Microsatellites Fragments Selection

1. Mix 600 μL of MagneSphere (one tube of Promega kit, Streptavidin MagneSphere Paramagnetic Particles), place the tube on the magnetic support, wait 30 s, and discard carefully

the supernatant. Resuspend in 300 µL of 0.5× SSC. Repeat these two steps three times. Resuspend in 100 µL of 0.5× SSC.

2. Add 400 µL of ultrapure water to 100 µL purified DNA from step 3.3.9 and incubate at 95 °C for 15 min.
3. Add 3 µL of each $I_5(GA)_8$ and $I_5(GT)_8$ biotinylated microsatellite oligo (50 µM) and 13 µL of 20× SSC.
4. Wait at room temperature for 20 min and mix slowly by inverting the tube every 2 min.
5. Mix the 100 µL prewashed MagneSphere beads from **step 1** with the 516 µL hybridization mix.
6. Incubate 10 min at room temperature and shake slowly continuously. Place the tube on the magnetic support, wait for 30 s, and discard carefully the supernatant. Resuspend in 300 µL of 0.1× SSC. Repeat three times.
7. Resuspend the magnetic bead in 100 µL of water, place the tube on the magnetic support, wait for 30 s, and take the supernatant. Resuspend the magnetic bead another time in 150 µL of water, place the tube on the magnetic support, wait for 30 s, and take the supernatant of water. Combine the two supernatants and freeze at –20 °C this selected fragments solution.

3.5 Selected Fragments Amplification

1. Mix 10 µL of selected fragments solution, 4 µL oligo Rsa21 (10 µM), 10 µL 10× PCR buffer, 8 µL dNTP (2.5 mM), 2.5 units of *Taq* DNA polymerase, and ultrapure water to 100 µL.
   Run the PCR program: cycle: 95 °C 1 min, followed by 20 times 94 °C 40 s, 60 °C 1 min, 72 °C 2 min, and final step 72 °C 5 min.
2. Mix 10 µL of the amplification product with 1 µL of loading dye and load in a 1.2 % agarose gel; run at 60 V for 45 min (*see* **Note 6**).

3.6 Cloning in pGEM-T

1. Mix a 2 µL aliquot of the amplification product (*see* **Note 7**), 1 µL of pGEM-T plasmid, 10 µL of 2× buffer, 1 µL of ligase, and 6 µL of ultrapure water.
2. Mix the reaction by pipetting. Incubate 1 h at room temperature.
3. Centrifuge the tubes containing the ligation reaction to collect contents at the bottom of the tube.

3.7 Transformation in XL1-Blue Competent Cells

1. Prepare 10 LB + ampicillin + X-gal + IPTG plates (130 mm diameter) for platting the transformation product and add two other small plates (90 mm diameter) for determining transformation efficiency (*see* **Note 8**).

2. Equilibrate the plates at room temperature prior to plating.
3. 1 h before the manipulation, transfer one tube of XL1-Blue bacteria suspension from –80 °C freezer to a –20 °C freezer.
4. Prepare a water bath at 42 °C.
5. 5 min before the manipulation, transfer XL1-Blue bacteria suspension to ice.
6. In two 15 mL BD Falcon tubes (ref. 2059) (*see* **Note 9**), place 40 µL of XL1-Blue suspension and add 0.7 µL of the provided mercaptoethanol solution to each tube.
7. Wait for 10 min in ice while swirling gently every 2 min.
8. In the first tube, add 2 µL of ligation product (*see* **Note 10**), and in the second tube add 1 µL of the provided solution uncut pUC18 control plasmid for determination of the transformation efficiency of the competent cells; mix slowly the two tubes.
9. Incubate the tubes in ice for 30 min, without mixing.
10. Heat pulse, by incubating both tubes in a water bath at 42 °C for 45 s (very precise and do not shake) (*see* **Note 11**).
11. Incubate the tubes in ice for 2 min.
12. Add 450 µL of preheated at 37 °C SOC medium.
13. Incubate the tubes at 37 °C for 1 h with shaking at 200/225 rpm.
14. Plate 50 µL of the transformation mix on each 130 mm LB ampicillin + X-gal + IPTG agar plate. For the transformation control, a 1:10 dilution with SOC medium is recommended for plating.
15. Incubate the plates overnight at 37 °C. If performing blue-white color screening, incubate the plates at 37 °C for at least 17 h to allow color development (*see* **Notes 12** and **13**).

3.8 Positive Clone Screening Using PCR Between Plasmid Sequence and Microsatellite Motif

1. Under laminar flux cabinet, prepare PCR 96 well plate, with 20 µL filtered (0.25 µ) ultrapure water in each well.
2. With a toothpick take a white clone and put in water. Proceed line by line of the well plate and discard each toothpick. When the PCR plate is full, transfer a half volume to a second (replicate) PCR plate.
3. Using 96 tips replicator, make a subculturing on a new LB + ampicillin + X-gal + IPTG plate. Incubate the plates overnight at 37 °C.
4. Perform two PCR with the two different primers AGE1 and AGE2 (*see* **Note 14**)

PCR mix for 4 plates: 10× buffer	1,025 µL
dNTP 2.5 mM	820 µL
Oligo AGE1 or AGE2 (100 µM)	25 µL
Oligo $(GA)_{12}$ (100 µM)	25 µL
Oligo $(GT)_{12}$ (100 µM)	25 µL
Ultrapure water	4,050 µL
Taq polymerase 5 unit/µL	80 µL

Add a 15 µL aliquot in each well, add one drop of mineral oil, and cover plate with an adhesive plastic.

PCR cycle: 95 °C 4 min, following by 30 times 94 °C 30 s, 60 °C 45 s, 72 °C 1 min 30 s, and 72 °C 8 min.

5. Loading 15 µL of the amplification product in a 300 mL 1.0 % agarose gel, run at 80 V for 1 h 30 min (*see* **Note 13**). Reveal by coloration using ethidium bromide (*see* **Note 1**) and take a picture. When a microsatellite motif is present in the probe, the fragment is amplified between microsatellite motif GA or GT and the plasmid oligo (AGE 1 or AGE2) and the gel must reveal fragments.

3.9 Positive Clone Size Determination and Fragment Selection for Sequencing

1. To determine the size of cloned fragment fragments, perform a PCR using both AGE1 and AGE2 primers on positive clones obtained in Subheading 3.8, **step 5**

PCR mix for 1 plate: 10× buffer	660 µL
dNTP 2.5 mM	300 µL
Oligo AGE1 (100 µM)	25 µL
Oligo AGE2 (100 µM)	25 µL
Ultrapure water	3,778 µL
Taq polymerase 5 unit/µL	50 µL

Add a 45 µL aliquot in each well containing 5 µL of bacterial culture, add one drop of mineral oil, and cover plate with an adhesive plastic.

PCR cycle: 95 °C 4 min, following by 30 times 94 °C 30s, 60 °C 1 min, 72 °C 2 min, and 72 °C 8 min.

2. Loading 10 µL of the amplification product in a 300 mL 1.0 % agarose gel, run at 80 V for 1 h 30 min (*see* **Note 15**). Reveal by coloration using ethidium bromide (*see* **Note 1**) and take a picture.
3. A first choice of amplified fragments is based on size between 500 and 1,200 bp. Before sequencing, the selected fragments must be purified using a QIAquick PCR Purification Kit like in step 3.3.

3.10 Sequencing and Primer Design

1. Send selected clones for sequencing to a commercial provider using the Sanger method [12].
2. Plasmidic and adaptor flanking sequences are eliminated, and the presence of SSRs is detected using SAT, a microsatellite analysis tool [13]. It is used to collect sequence information and facilitate the design of PCR primers and their flanking sequences. The following criteria must be considered for sequence selection: uniqueness, adequate flanking sequence size, and non-repetitive flanking regions. Finally, primer pairs are designed using Primer 3 software [14].

4 Notes

1. Ethidium bromide is an intercalating agent of double-stranded DNA; it is a potent mutagen product. Handle only with gloves and precaution.
2. All the PCR primers used in the present protocol are prepared with lyophilized oligonucleotides and dissolved in ultrapure water to a final concentration of 100 μM and stored at −20 °C. Rsa21 and Rsa25 are diluted to 10 μM with ultrapure water before use.
3. Store the Streptavidin MagneSphere® Paramagnetic Particles at 4 °C. Do not freeze the Streptavidin MagneSphere® Paramagnetic Particles, as this will reduce their performance.
4. Biotinylated oligonucleotide must be aliquoted by 10 μL fractions and stored at −20 °C; they must be thawed 15 min before use.
5. Ensure that the elution buffer is dispensed directly onto the QIAquick membrane for complete elution of bound DNA.
6. The gel must reveal a similar or lesser size than the preamplification smear.
7. The pGEM®-T Vector Systems have been optimized using a 1:1 molar ratio of the Control Insert DNA to the Vectors. If initial experiments with your PCR product are suboptimal, ratio optimization may be necessary. Ratios from 3:1 to 1:3 provide good initial parameters. The concentration of PCR product should be estimated by comparison to DNA mass standards on a gel.
8. Blue-white color screening for recombinant plasmids is available when the host strain contains the *lacI*q*Z*D*M15* gene on the F′ episome with a plasmid that provides α-complementation. When lacZ expression is induced by IPTG in the presence of the chromogenic substrate X-gal, colonies containing plasmids with inserts will be white, while colonies containing plasmids without inserts will be blue.

9. Use of 14-mL BD Falcon polypropylene round-bottom tubes: it is important that 14-mL BD Falcon polypropylene round-bottom tubes (BD Biosciences Catalog #352059) are used for the transformation protocol, since other tubes may be degraded by β-mercaptoethanol. In addition, the duration of the heat pulse has been optimized using these tubes.
10. Centrifuge the tubes containing the ligation reactions to collect contents at the bottom of the tube.
11. Optimal transformation efficiency is observed when cells are heat-pulsed at 42 °C for 45–50 s. Efficiency decreases sharply when cells are heat-pulsed for <45 s or for >60 s. Do not exceed 42 °C.
12. For the pUC18 control, expect 50 colonies ($\geq 1 \times 108$ cfu/μg pUC18 DNA). For the experimental DNA, the number of colonies will vary according to the size and form of the transforming DNA, with larger and non-supercoiled DNA producing fewer colonies.
13. To facilitate blue-white screening, color can be enhanced by subsequent incubation of the plates for 3 h at 4 °C.
14. To test the microsatellite motif capture 2 PCRs were performed using one of the motif used for the capture and the two primers located on either side of the cloning site on the plasmid.
15. The loading design of the deposit on the gel should be as readable as possible, it is best to place two or three lines of PCR plate by row of the gel wells (depends of the number of well and the gel size).

References

1. Tautz D, Rentz P (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genome. Nucleic Acids Res 12: 4127–4138
2. Mullis K, Faloona S, Scharf R, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor Symp Quant Biol 51:263–273
3. Beckman JS, Soller M (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Biotechnology 8:930–932
4. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millaseau P, Vaysseix G, Lathrop M (1992) A second generation map of the human genome. Nature 359:794–801
5. Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.), comparisons with data from RFLPs and pedigree. Theor Appl Genet 95:163–173
6. Akagi H, Yokozeki Y, Inagaki A, Fujimura T (1996) Microsatellite DNA markers for rice chromosome. Theor Appl Genet 93: 1071–1077
7. Panaud O, Chen X, McCouch SR (1997) Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (*Oryza sativa* L.). Mol Gen Genet 252:597–607
8. Karagyozov L, Kalcheva ID, Chapman VM (1993) Construction of random small insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Res 21(16): 3911–3912

9. Kijas JMH, Fowler JCS, Garbett CA, Thomas MR (1994) Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particle. Biotechniques 16(4):656–660
10. Edwards KJ, Barker JHA, Daly A, Jones C, Karp A (1996) Microsatellite libraries enriched for several microsatellite sequences in plants. BioTechniques 20:758–760
11. Billotte N, Lagoda PJL, Risterucci AM, Baurens FC (1999) Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. Fruits 54:277–288
12. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12): 5463–5467
13. Dereeper A, Argout X, Billot C, Rami J-F, Ruiz M (2007) SAT, a flexible and optimized Web application for SSR marker development. BMC Bioinformatics 8:465
14. Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, NJ, pp 365–386

Chapter 10

Randomly Amplified Polymorphic DNA (RAPD) and Derived Techniques

Kantipudi Nirmal Babu, Muliyar Krishna Rajesh, Kukkumgai Samsudeen, Divakaran Minoo, Erinjery Jose Suraby, Kallayan Anupama, and Paul Ritto

Abstract

Understanding biology and genetics at molecular level has become very important for dissection and manipulation of genome architecture for addressing evolutionary and taxonomic questions. Knowledge of genetic variation and genetic relationship among genotypes is an important consideration for classification, utilization of germplasm resources, and breeding. Molecular markers have contributed significantly in this respect and have been widely used in plant science in a number of ways, including genetic fingerprinting, diagnostics, identification of duplicates and selecting core collections, determination of genetic distances, genome analysis, developing molecular maps, and identification of markers associated with desirable breeding traits. The application of molecular markers largely depends on the type of markers employed, distribution of markers in the genome, type of loci they amplify, level of polymorphism, and reproducibility of products. Among many DNA markers available, random amplified polymorphic DNA (RAPD) is the simplest and cost-effective and can be performed in a moderate laboratory for most of its applications. In addition RAPDs can touch much of the genome and has the advantage that no prior knowledge of the genome under research is necessary. The recent improvements in the RAPD technique like AP-PCR, SCAR, DAF, SRAP, CAPS, RAMPO, and RAHM can complement the shortcomings of RAPDs and have enhanced the utility of this simple technique for specific applications. Simple protocols for these techniques are presented.

Key words RAPD, AP-PCR, SCAR, DAF, SRAP, CAPS, RAMPO, RAHM, DNA fingerprinting, Genetic diversity, Population and evolutionary genetics

1 Introduction

The advent of polymerase chain reaction (PCR) and subsequent emergence of DNA-based markers have provided plant taxonomists easy and reliable techniques to study the extent and distribution of variation in species gene pools and to answer typical evolutionary and taxonomic questions which were not previously possible with only phenotypic methods. Properties desirable for ideal DNA markers include highly polymorphic nature, codominant inheritance,

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_10,

and frequent occurrence in the genome, easy access, easy and fast assay, and high reproducibility. DNA marker systems based on PCR include random amplified polymorphic DNAs (RAPDs) [1], amplified fragment length polymorphism (AFLPs) [2] (*see* Chapter 11), microsatellites/simple sequence repeats (SSRs) [3] (*see* Chapter 9), and single-nucleotide polymorphisms (SNPs) [4] (*see* Chapter 9). Although the sequencing-based molecular techniques provide better resolution at intra-genus and above level [5], it is expensive and laborious. Frequency data from markers such as random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and microsatellites provide the means to classify individuals into nominal genotypic categories and are mostly suitable for intraspecies genotypic variation study. Compared to other PCR-based techniques, which vary in detecting genetic differences and applicability to particular taxonomic levels, RAPD is a cost-effective tool for taxonomic studies.

RAPDs is an adaptation of the PCR which relies on the rationale that at low stringency, a given synthetic oligonucleotide primer is likely to find a number of sequences in the template DNA to which it can anneal when these sites are close to each other and lie in opposite orientations, and the DNA sequence between the sites will be amplified to produce a DNA fragments characteristic of that genome. Multiple bands of different sizes produced from the same genomic DNA constitute a "fingerprint" of that genome [1]. Patterns from different individuals and species will vary as a function of how similar the genomic DNA sequences are between samples. RAPD polymorphisms result from either chromosomal changes in the amplified regions or base changes that alter primer binding. This assay has the advantage of being readily employed, requiring very small amounts of genomic DNA, and eliminating the need for blotting and radioactive detection. As RAPD require initial genome information, it provides markers in regions of the genome previously inaccessible to analysis. RAPD-derived estimates of genetic relationships are in good agreement with pedigree, RFLP, and isozyme data [6, 7].

DNA fingerprinting for cultivar or varietal identification has become an important tool for estimating genetic diversity for plant breeding, germplasm management, utilization [8], monitoring genetic erosion, and removing duplicates from germplasm collections [9]. As RAPD markers could gain information about genetic similarities or differences that are not expressed in phenotypic information, RAPD analysis becomes an inexpensive tool to characterize germplasm collections [10], to understand the pattern of evolution from wild progenitors, and to classify them into appropriate groups.

RAPDs have been successfully applied in estimation of varietal distinctiveness and relatedness of commercially important crops, registration activities like cultivar identification [11], or control of seed purity of hybrid varieties [12]. The potential of RAPD for

varietal identification has been used varietal to know about the variety being exported or sold under various trade names, to settling a lawsuit involving unauthorized commercialization of a patented varieties [13], to identify the cases of adulteration, and even the level of adulteration [14].

As RAPDs make use of arbitrary primers, some of them amplify DNA at highly conserved region, leading to generate polymorphisms at high levels of classification, whereas some will amplify at highly variable region, useful for classification and analyses at and below the species level. This property of RAPD is taxonomically useful at subgeneric level [15], species level [16], and for the analysis of geographic variation. Another application of RAPD is for evaluation of the genetic integrity of somatic embryo derived plants [17].

RAPDs have significant use in ecology in studying mating systems and assigning paternity. In plants, insect pollination might be studied by fingerprinting all the potential pollen sources by RAPDs and comparing the dominant RAPD bands seen in the resulting seeds [18]. RAPDs are useful in hybridization studies to document intergeneric hybridization [19] to identify species-specific bands as well as interspecific hybridization and detection of introgression in both natural and cultivated plant populations [20]. RAPDs may provide insights into organismal evolutions that are overlooked by single-gene comparisons [21].

The RAPD technique has received a great deal of attention from population geneticists [22] because of its simplicity and rapidity in revealing DNA-level genetic variation. The assumption of homology between bands of apparently the same molecular weight from the same primer is potentially another problem for RAPD surveys. Homology between co-migrating bands in different individuals is a good assumption when individuals are from the same population. This may not be true when individuals belong to different species or widely divergent populations [23]. Because the chance of co-migrating bands being homologous becomes less as populations diverge, it was suggested [1, 23] that RAPD analysis gives more accurate estimates between closely related populations and less accurate estimates for distantly related populations.

A disadvantage of RAPD markers is the fact that they are dominant markers and provides no information on heterozygosity. RAPD markers can be converted into codominant markers called SCAR markers (sequence characterized amplified regions) [24]. RAPDs also have shortcomings of reproducibility of data.

The reproducibility of different molecular markers, RAPD, AFLP, and SSR, tested in plants by a network of European laboratories [25] in which an optimal system (genetic screening package) was present was distributed to each of the laboratories. Different experiences were gained in this exchange experiment with the different techniques. RAPDs were found to be easy to perform by all groups, but reproducibility was not achieved to a satisfactory level.

For AFLPs, a single-band difference was observed in one track, while SSR alleles were amplified by all laboratories, but small differences in their sizing were obtained. Hence, RAPD marker identity might be established by fingerprinting a set of standard genotypes by RAPD to facilitate communication and the reproducibility among laboratories, which may be influenced by the independence of RAPD polymorphisms relative to each other and the distribution of polymorphism across genotypes [26].

The RAPD protocol is refined to techniques like SCAR, AP-PCR, DAF, SRAP, CAPS, RAMPO, and RAHM so that some of the current problems such as lack of reproducibility and codominant nature of inheritance will be overcome. Using several strategies, various modifications have been developed in conjunction with RAPD to enhance the ability to detect polymorphism either by using more than one arbitrary primer [27] or by using a degenerate primer in the amplification reaction [28].

Sequence characterized amplified region (SCAR) markers are generated by sequencing RAPD marker termini and designing longer primers (22–24 nucleotide bases long) for specific amplification of particular locus [29, 30]. SCARs are usually dominant markers; however, some of them can be converted into codominant markers by digesting them with tetra-cutting restriction enzymes, and polymorphism can be deduced by either denaturing gel electrophoresis or SSCP [31]. Besides higher specificity it is based on the presence/absence of a single specific amplicon, considerably simplifying the interpretation of the results, especially when a large number of samples are checked. SCARs also allow comparative mapping or homology studies among related species, thus making it an extremely adaptable concept in the near future.

Arbitrary primed polymerase chain reaction (AP-PCR) is a special case of RAPD, wherein discrete amplification patterns are generated by employing single primers of 10–50 bases in length in PCR of genomic DNA. Unlike RAPDs, the oligonucleotide length and primer concentrations are tenfold higher [32], and two cycles of low-stringency annealing conditions to allow mismatches followed by PCR at high stringency and the newly synthesized fragments are radiolabeled using dCTP. AP-PCR-generated fragments are analyzed as plus/minus DNA amplification-based polymorphism [33] due to either sequence divergence at one of the priming sites or insertion/deletion within the amplification region.

DNA amplification fingerprinting (DAF) uses single arbitrary primers as short as 5 bases to amplify DNA using polymerase chain reaction with high multiplex ratio [34]. This marker shares those features common to AP-PCR and RAPDs; namely, it results in plus/minus heritable amplification polymorphism, a preponderance of dominant marker loci, and unknown allelism between fragments of equivalent molecular weight. DAF bands contain many more bands than AP-PCR and RAPD patterns, and the likelihood is increased for observing polymorphism between samples.

DNA amplification fingerprinting (DAF) has found to be promising in many plants for cultivar identification and sex determination [35] and for determination of genetic origin and diversity analysis [36].

The sequence-related amplified polymorphism technique (SRAP), a variation of RAPD, also uses arbitrary primers of 17–21 nucleotides to generate a specific banding pattern aimed to amplify coding sequences (ORFs) in the genome [37] and results in a moderate number of codominant markers. SRAP polymorphism results from two events: fragment size changes due to insertions and deletions, which could lead to codominant markers, and nucleotide changes leading to dominant markers. It has several advantages over other systems: simplicity, reasonable throughput rate, allows easy isolation of bands for sequencing, discloses numerous codominant markers, and allows screening thousands of loci shortly to pinpoint the genetic position underlying the trait of interest. The primers and primer concentration vary for each of the RAPD-derived techniques which increases its utility in various applications (*see* **Note 1**).

To derive greater information from RAPD patterns, the strategy of hybridizing SSR repeat primers to RAPD amplification patterns has been described. The method has been called either random amplified hybridization microsatellites (RAHM) [38] or random amplified microsatellite polymorphism (RAMPO) [39]. In RAHM, RAPD amplification and oligonucleotide screening are combined for detection of microsatellites to provide more information from RAPD gels and also help to reveal microsatellite genomic clones without the time-consuming screening of genomic libraries [38] (*see* Chapter 9). RAMPO combines arbitrarily or semi-specifically primed PCR with microsatellite hybridization to produce several independent and polymorphic genetic fingerprints per electrophoretic gel. In this approach, the amplified products resolve length polymorphism that may be present either at the SSR target site itself or at the associated sequence between the binding sites of the primers [39]. The RAPD binding site actually serves as an arbitrary end point for the SSR-based amplification product, and therefore the products obtained are not as restricted by the relative genomic positions of a specific SSR.

Another strategy is referred as cleaved amplified polymorphic sequences (CAPs), in which sequence information from cloned RAPD bands can be used for analyzing nucleotide polymorphisms. CAPS markers rely on differences in restriction enzyme digestion patterns of PCR fragments caused by nucleotide polymorphism between ecotypes. Sequence information available in data bank of genomic DNA or cDNA sequences or cloned RAPD bands can be used for designing PCR primers for this process. Cleaved amplified polymorphic sequences (CAPS) are PCR- RFLP markers performed by digesting locus specific PCR amplicons with one or more restriction enzymes followed by separation of the digested DNA on agarose or polyacrylamide gels [40, 41]. The sizes of the cleaved and uncleaved amplification products can be adjusted

arbitrarily by the appropriate placement of the PCR primers. Critical steps in the CAPS marker approach include DNA extraction, PCR conditions, and the number or distribution of polymorphic sites.

2 Materials

2.1 Genomic DNA Isolation and Quantification

1. 2× extraction buffer: (2 % cetyltrimethylammonium bromide (CTAB), 100 mM Tris–HCl, pH 8, 20 mM ethylenediaminetetraacetic acid (EDTA), pH 8, 1.4 M NaCl, 1 % polyvinylpolypyrrolidone (PVPP)).
2. Chloroform: isoamyl alcohol (24:1).
3. 100 % ethanol or isopropanol.
4. 70 % alcohol.
5. TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8).
6. RNAse A (10 mg/mL).
7. 50× Tris-Acetate-EDTA (TAE) buffer (pH 8).
8. Agarose.
9. Ethidium bromide (10 mg/mL).
10. 6× loading dye (30 % glycerol, 5 mM EDTA, 0.15 % bromophenol blue, 0.15 % xylene cyanol).
11. MassRuler 1,000 bp DNA ladder.

2.2 Reagents Used for RAPD PCR

1. Taq DNA polymerase with 10× buffer.
2. 10 mM dNTPs: 10 mM each of dATP, dCTP, dGTP, and dTTP.
3. 25 mM $MgCl_2$.
4. 10 μM Primers (operon primers are the most commonly used RAPD primers) (*see* **Notes 2** and **3**).
5. Milli-Q water.

2.3 Sequence Characterized Amplified Region (SCAR)

1. QIAquick gel extraction kit, Qiagen, Germany.

2.3.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.3.2 Reagents for PCR (See Subheading 2.2)

2.3.3 Gel Extraction

2.3.4 Cloning of PCR Amplified Gene

1. PCR amplified and purified product.
2. PCR cloning vector.
3. T4 DNA ligase.
4. 5× ligation buffer.
5. Sterile deionized water.
6. Overnight culture of E coli DH5∝.
7. $CaCl_2$ (100 mM).
8. $MgCl_2$ (25 mM).
9. LB medium.
10. Sterile micro centrifuge tubes and tips.
11. Sterile glycerol (80 %).
12. LB agar with ampicillin (100 µg/mL), X gal (20 µg/mL), and IPTG (40 µg/mL).

2.4 Arbitrary Primed Polymerase Chain Reaction (AP-PCR)

2.4.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.4.2 Reagents for PCR

1. Taq polymerase.
2. 10× PCR buffer.
3. 25 mM $MgCl_2$.
4. 10 mM each of dNTPs.
5. 50 µCi α-[^{32}P] dCTP.
6. 10 µM of each primer.

2.4.3 Electrophoresis

1. 40 % Acrylamide bis-acrylamide.
2. 7.5 M Urea.
3. 10× Tris-Borate-EDTA(TBE) buffer, pH 8.

2.5 DNA Amplification Fingerprinting (DAF)

2.5.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.5.2 Reagents for PCR (See Subheading 2.2)

2.5.3 PAGE Reagents

1. 40 % Acrylamide bis-acrylamide.
2. 7.5 M Urea.
3. 10× Tris-Borate-EDTA(TBE) buffer, pH 8.
   Cover the bottle with aluminum foil and store at 4 °C and use before 1 month.
4. 10 bp MassRuler.
5. 100 bp MassRuler.

2.5.4 Silver Staining Reagents

1. Acetic acid, glacial.
2. Silver nitrate crystal, AR (ACS) ($AgNO_3$).
3. Formaldehyde solution, AR (ACS) (HCHO).
4. Sodium thiosulfate (Na_2S_2O).
5. Sodium carbonate powder, ACS reagent (Na_2CO_3).

6. Ethanol.
7. Silver staining solution (250 mg silver nitrate and 375 μL formaldehyde and 50 μL sodium thiosulfate).
8. Ice-cold developer solution (10 °C) (7.5 g sodium carbonate, 375 μL formaldehyde, and 50 μL sodium thiosulfate (10 mg in 1 mL water) in 250 mL water).
9. Formamide loading dye (80 % formamide, 10 mM EDTA pH 8.0, 1 mg/mL Xylene cyanol 1 mg/mL, bromophenol blue—50 mg).

2.6 The Sequence-Related Amplified Polymorphism Technique (SRAP)

2.6.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.6.2 Reagents for PCR (See Subheading 2.2)

2.6.3 PAGE Electrophoresis (See Subheadings 2.5.3 and 2.5.4)

Primers. The arbitrary primers consists of the following elements: core sequences, which are 13–14 bases long, where the first 10 or 11 bases start at the 5′end, are sequences of no specific constitution ("filler" sequences), followed by the sequence CCGG in the forward primer and AATT in the reverse primer. The purpose for using the "CCGG" sequence in the core of the first set of SRAP primers was to target exons to open reading frame (ORF) regions (*see* **Note 4**).

2.7 Randomly Amplified Microsatellite Polymorphism (RAMPO)

2.7.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.7.2 Reagents Used for RAPD and Microsatellite-Primed PCR (MP-PCR) (See Subheading 2.2)

2.7.3 Hybridization with Microsatellite-Complementary Probes

1. Nylon membrane (Hybond, Amersham).
2. ^{32}P-labeled microsatellite-complementary oligonucleotide probes.

2.7.4 Autoradiography

2.8 Random Amplified Hybridization Microsatellites (RAHM)

2.8.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.8.2 Reagents Used for RAPD PCR (See Subheading 2.2)

2.8.3 Hybridization with Microsatellite-Complementary Probes (See Subheading 2.7.3)

2.8.4 Autoradiography

1. 10- mer primers (Operon Technologies, Alameda, CA, USA)
2. Hybond-N+ filters (Amersham Inc.)
3. Oligonucleotide probes carrying simple sequence repeats (SSR) labeled with Digoxigenin-ddUTP (DIG Oligonucleotide 3'-End Labeling Kit, Boehringer Mannheim)
4. Gel purification - 'Double Geneclean' (BIO 101 Inc., USA)

2.9 Cleaved Amplified Polymorphic Sequences (CAPS)

2.9.1 Genomic DNA Isolation and Quantification (See Subheading 2.1)

2.9.2 Reagents for PCR Conditions (See Subheading 2.2)

2.9.3 Restriction Enzyme Digestion

2.9.4 PAGE Reagents (See Subheading 2.5.3)

2.9.5 Silver Staining Reagents (See Subheading 2.5.4)

1. Restriction enzymes: Mse I, Alu I, Mbo I, and Hae III.
2. Buffer 2 (NEB)—supplied at 10× concentration.
3. 50 mM NaCl.
4. 10 mM Tris–HCl.
5. 10 mM MgCl2.
6. 1 mM DTT pH 7.9 at 25 °C.
7. 100× BSA (10 mg/mL)—use at 1×.

3 Methods

3.1 Isolation of Genomic DNA (Modified Doyle and Doyle [42])

1. Grind 2 g of clean young leaf tissue to fine powder with a pestle and mortar after freezing in liquid nitrogen, transfer it to 10 mL CTAB extraction buffer, and incubate at 60 °C for 1 h.
2. Extract with chloroform: isoamyl (24:1) and centrifuge at 12,378 ×*g* for 10 min at room temperature.
3. Precipitate the DNA with 100 % ethanol or isopropanol and centrifuge at 19,341 ×*g* for 10 min at 4 °C.

4. Wash the DNA with 70 % ethanol and centrifuge at 19,341 ×*g* for 5 min at 4 °C.
5. Dry the pellet and dissolve the DNA in 1× TE buffer.
6. Treat the DNA in solution with RNAse (10 μg/mL) at 37 °C for 30 min.
7. Wash with chloroform: isoamyl alcohol (24:1) and centrifuge at 12,378 ×*g* for 10 min at room temperature.
8. Precipitate with 100 % ethanol and dissolve in 1× TE buffer. Store frozen at −20 °C.

3.2 DNA Quantification

It is an essential step in many procedures where it is necessary to know the amount of DNA that is present when performing techniques such as PCR and RAPDs (*see* **Note 5**).

3.2.1 By Gel Electrophoresis

The comparison of an aliquot of the extracted sample with standard DNAs of known concentration (lambda *Hin* III) can be done using gel electrophoresis.

1. 5 μL of the DNA is mixed with 1 μL of 6× loading dye and loaded onto a 0.8–1 % agarose gel along with 500 ng of lambda *Hin* III digest marker and electrophoresed at 90 V for 30 min.
2. The quantity of extracted DNA is estimated based on the intensity of lambda *Hin* III digest marker bands as the top bands accounts half amount (250 ng) of total loaded amount.
3. The quality of genomic DNA is confirmed for its integrity.

3.2.2 Using UV Spectrophotometer

1. Take 1 mL of TE buffer in a cuvette and calibrate the spectrophotometer at 260 and 280 nm wavelength.
2. Add 2–5 μL of DNA mix properly and record the optical density at both 260 and 280 nm.
3. Estimate the DNA concentration employing the following formula:
   Amount of DNA (μg/μL) = (OD) 260 *50* dilution factor/1,000.
4. Judge the quality of DNA from the ratio of OD values recorded at 260 and 280 nm. Pure DNA has values close to 1.8.
5. Dilute the DNA sample to get 20 ng/μL.

3.3 RAPD

3.3.1 PCR Amplification of Genomic DNA with Primers (See **Notes 2, 3, 6,** and **7**)

1. Amplify 20–50 ng of genomic DNA in a reaction mix containing 1.0 U *Taq* DNA polymerase, 1 μM primer, 1.5–2.0 mM $MgCl_2$, and 0.125 mM each of dNTPs and 1× *Taq* DNA polymerase buffer (*see* **Note 6**).
2. The amplification profile consists of an initial denaturation of 3 min at 94 °C followed by 35–40 cycles of denaturation for 1 min at 94 °C, annealing for 37 °C for 1 min, and extension at 72 °C for 2 min and final extension for 6 min at 72 °C (*see* **Note 7**).

3.3.2 Gel Electrophoresis

1. Amplified RAPD products are separated by horizontal electrophoresis in 1.5 % (w/v) agarose gel, with 1× TAE buffer, stained with ethidium bromide (0.5 μg/mL), and analyzed under ultraviolet (UV) light. The length of the DNA fragments is estimated by comparison with DNA ladder.

3.3.3 Scoring and Interpretation of RAPD Banding Patterns

Variability is then scored as the presence or absence of a specific amplification product.

Polymorphism usually results from mutations or rearrangements either at or between the primer binding sites due to appearance of a new primer site, mismatches at the primer site, and difference in the length of the amplified region between the primer sites due to deletions or insertions in the DNA.

Each gel is analyzed by scoring the present (1) or absent (0) polymorphic bands in individual lanes. The scoring procedure is based on the banding profiles which are clear, transparent, and repeatable (*see* **Notes 8–10**) (Fig. 1).

1. The RAPD profiles are compared between the genotypes to estimate the similarity index. Studies are initiated to assess the similarity/differences between the genotypes using RAPD polymorphism as estimated by paired affinity indices (PAI).
   PAI was calculated by the formula PAI = No. of similar bands/Total no. of bands.
   The PAIs expressed as percentage indicated the similarity (%) between any two genotypes.
2. The binary matrix is transformed into similarity matrix using Dice similarity (NTSYS-PC 2.01; Numerical Taxonomy System of Multivariate Programs) [43] as the Dice coefficient/Jaccard coefficient assigns weights to matches rather than to mismatches and does take shared absences of bands into account.

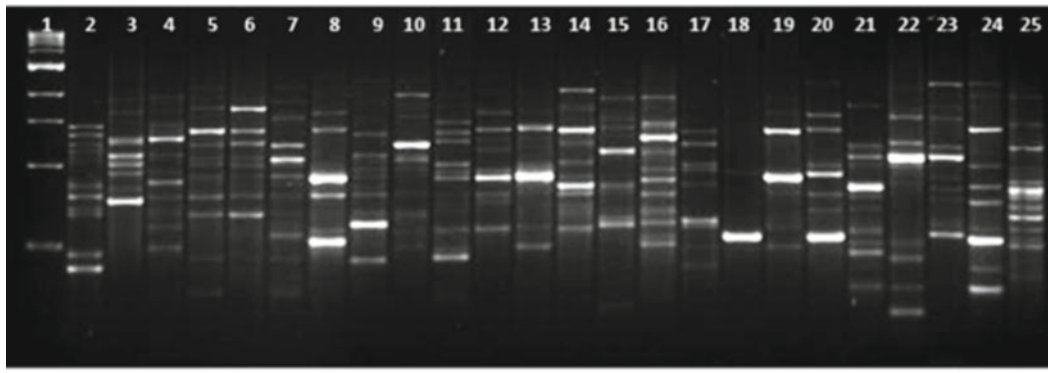


Fig. 1 RAPD polymorphism expressed by "operon primers" OPC-09 in wild and related species of black pepper (*Piper nigrum*) (1) 1 kb ladder, (2) *P. longum*, (3) *P. hapnium*, (4) *P. mullesua*, (5) *P. attenuatum*, (6) *P. argyrophyllum*, (7) *P. hymenophyllum*, (8) *P. bababudani*, (9) *P. trichostachyon*, (10) *P. galeatum*, (11) *P. sugandhi*, (12) *P. psuedonigrum*, (13) *P. nigrum*, (14) *P. schimdti*, (15) *P. wightii*, (16) *P. silentvalyensis*, (17) *P. barberi*, (18) *P. betel*, (19) Cultivated black pepper cv. Karimunda, (20) *P. chaba-1*, (21) *P. chaba-2*, (22) *P. colubrinum-1*, (23) *P. colubrinum-2*, (24) *P. arboreum*, and (25) *P. ornatum*

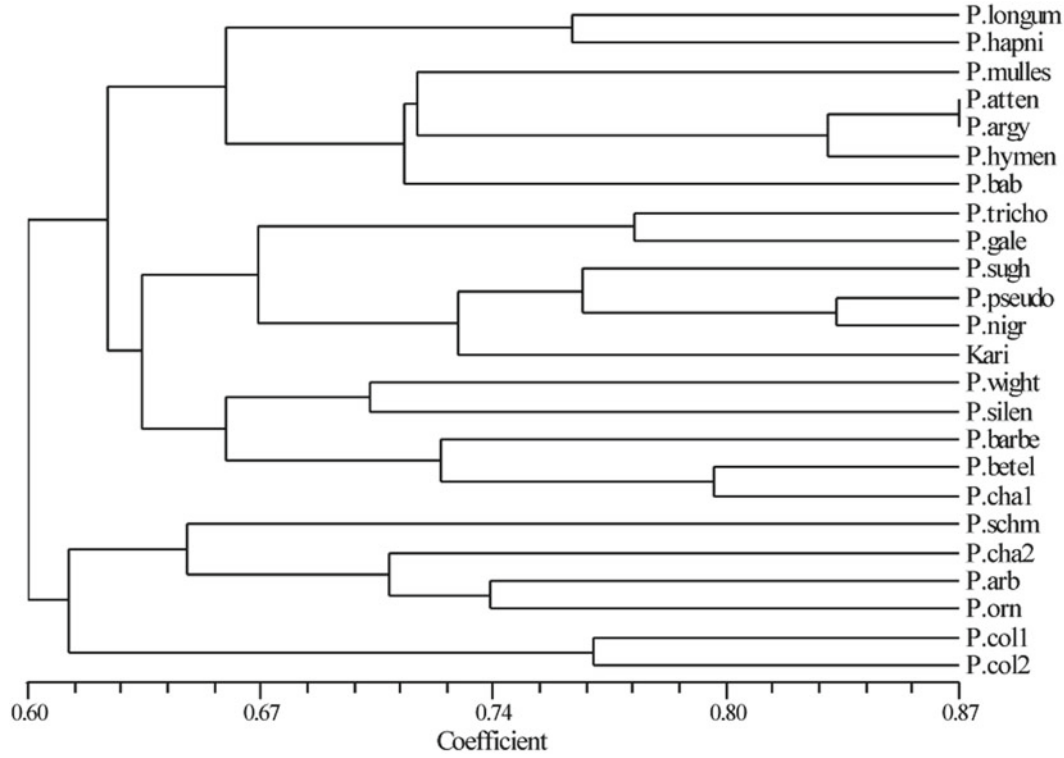


Fig. 2 Dendrogram of interrelationships among wild and related species of black pepper (*Piper nigrum*) (1) *P. longum*, (2) *P. hapnium*, (3) *P. mullesua*, (4) *P. attenuatum* (5) *P. argyrophyllum*, (6) *P. hymenophyllum*, (7) *P. baba-budani*, (8) *P. trichostachyon*, (9) *P. galeatum*, (10) *P. sugandhi*, (11) *P. psuedonigrum*, (12) *P. nigrum*, (13) *P. schimdti*, (14) *P. wightii*, (15) *P. silentvalyensis*, (16) *P. barberi*, (17) *P. betel*, (18) Cultivated black pepper cv. Karimunda, (19) *P. chaba-1*, (20) *P. chaba-2*, (21) *P. colubrinum-1*, (22) *P. colubrinum-2*, (23) *P. arboreum*, and (24) *P. ornatum*

3. The similarity matrix is subjected to a clustering analysis using the unweighted pair group method with arithmetic means (UPGMA; NTSYS-PC 2.0) [43] (*see* **Notes 11** and **12**).
4. The RAPDs matrix can be analyzed using the neighbor-joining (N-J) method and evaluated statistical support for the clusters recovered both in the UPGMA and N-J trees by generating 1,000 bootstrap pseudoreplicates.
5. Dendrograms are then constructed according to the UPGMA, using NTSYS-PC 2.01 [43] (Fig. 2).

3.4 Sequence Characterized Amplified Region (SCAR)

3.4.1 Amplification

1. Genomic DNA is isolated, quantified, and diluted (*see* Subheading 3.1).
2. 20–50 ng of genomic DNA is amplified using random primers (*see* Subheading 3.2).
3. Aliquots (5.0 μL) of RAPD products are separated by horizontal electrophoresis in 1.5 % (w:v) agarose gel, with 1× TAE buffer, stained with ethidium bromide (0.5 μg/mL), and analyzed under ultraviolet (UV) light. The length of the DNA fragments is estimated by comparison with DNA ladder.

3.4.2 RAPD Fragments Selection and Cloning

1. From obtained RAPD fingerprints, the polymorphic RAPD marker bands are selected.
2. These bands are cut, eluted, and purified using QIAquick gel extraction kit, cloned, and sequenced.
3. PCR amplification: For the verification of primers ability to amplify predicted fragment length, primers are tested with isolated DNA.
4. Primer design: New longer and specific primers of 15–30 bp are designed for the DNA sequence, which is called the SCAR (*see* **Note 13**).

3.5 Arbitrary Primed Polymerase Chain Reaction (AP-PCR)

3.5.1 Amplification

1. Amplify 20 ng genomic DNA in a PCR mix containing 0.025 U Taq polymerase and 1× buffer (Stratagene) adjusted to 4 mM with $MgCl_2$, 0.2 mM of each dNTP, and 10 μM primer.
2. Amplification profile consists of an initial denaturation of 94 °C for 5 min followed by 40 °C for 5 min for low-stringency annealing of primer and 72 °C for 5 min for extension for two cycles. This temperature profile is followed by ten high stringency cycles: 94 °C for 1 min, 60 °C for 1 min, and 72 °C for 2 min for 10 cycles.
3. At the end of this reaction, add 90 μL of a solution containing 2.25 U Taq polymerase in 1× buffer, 0.2 mM dNTPs, and 50 μCi α-[^{32}P] dCTP, and the high stringency cycles are continued for an additional 20 or 30 rounds.

3.5.2 Electrophoresis

1. Prepare the 40 % stock 19:1 acrylamide bis-acrylamide solution store it in dark bottles at 4 °C.
2. Prepare 5 % working solution containing 7.5 M urea, 40 % acrylamide bis-acrylamide, TBE buffer, and 10× TBE buffer. Assemble electrophoresis unit by adding 0.5× TBE buffer to upper tank and lower tank.
3. Add 4 μL of the loading buffer to 8 μL of the final amplified reaction mix.
4. Load this sample into the gel and conduct electrophoresis at 18 W for 55 min.
5. The AP-PCR generated fragments are size separated on polyacrylamide and visualized via radiography.

3.6 DNA Amplification Fingerprinting (DAF)

3.6.1 Amplification

1. Amplify 20 ng of genomic DNA in a 10 μL PCR mix containing 0.5 U of Taq polymerase, 200 μM each dNTPs, 0.5 μM primer, and 1× PCR buffer with 2 mM $MgCl_2$ overlaid with a drop of mineral oil.
2. The amplification profile consists of an initial denaturation at 5 min of 94 °C followed by 40 cycles of denaturation for 5 s at 94 °C, annealing at either 35 °C or 45 °C and 30 s at 72 °C.

3. The amplification products are separated in a vertical electrophoresis system using 5 % non-denaturing polyacrylamide gel of 0.5 mm thickness to separate DNA fragments according to their molecular weight.
4. Gel preparation (*see* Subheading 3.5.2).

3.6.2 Silver Staining for DNA Visualization

1. Gently place the gel in 10 % (v/v) glacial acetic acid for 30 min at room temperature.
2. Rinse the gel in deionized water twice for about 2 min each.
3. Immerse the gel in silver staining solution for 20 min.
4. Pour out the silver stain solution and wash the gel quickly with deionized water within 10 s.
5. Immerse the gel in an ice-cold developer solution (10 °C) until optimal image intensity is obtained. Stop the developing process by immersing the gel in 7.5 % ice-cold glacial acetic acid.
6. Transfer gel onto the Whatman paper.
7. Air-dry the gel or dry using gel drier at 70 °C for 30 min.

3.6.3 Gel Interpretation

Scoring can be done by presence or absence of band. Bands are sized and matched directly on gels, autoradiographic or photographic films, or photocopies on transparency overlays.

3.7 Sequence-Related Amplified Polymorphism (SRAP)

3.7.1 Amplification

1. Amplify 20 ng of genomic DNA in a PCR mix containing 1 U of Taq polymerase, 200 μM each dNTPs, 0.1 mM each forward and reverse primer, and 1× PCR buffer with 1.5 mM $MgCl_2$.
2. The amplification profile consists of an initial denaturation at 2 min of 94 °C followed by 5 cycles of denaturation for 1 min at 94 °C, annealing at 35 °C for 1 min and 72 °C for 1 min; followed by 35 cycles of 94 °C for 1 min, 50 °C for 1 min, and 72 °C for 1 min; and followed by 7 min at 72 °C.
3. Polyacrylamide gel electrophoresis (*see* Subheading 3.5.2).
4. Marker analysis: Each polymorphic band can be scored as a single dominant marker.

3.7.2 Sequencing of SRAP Marker Bands

1. After electrophoresis, the gel is exposed overnight to a high-sensitivity film, (Kodak BioMax).
2. Using the exposed film as a blueprint, the gel pieces containing the polymorphic bands are cut and introduced into a dialysis tube.
3. The dialysis tube is placed into the buffer tank of a sequencing gel apparatus, and the DNA was electroeluted in 1× TBE buffer. The application of 2,000 V, which is the same voltage used for running sequencing gels, resulted in the complete electroelution of DNA into buffer from the gel fragment.
4. After ethanol precipitation and TE buffer suspension, the DNA can be used for direct sequencing.

3.8 Randomly Amplified Microsatellite Polymorphisms (RAMPO)

3.8.1 Genomic DNA Is Isolated (See Subheadings 3.1 and 3.2)

3.8.2 Amplification of Genomic DNA with RAPD Primers/Microsatellite Primers

1. The DNA is first amplified with a single arbitrary (*see* Subheading 3.3.1) or microsatellite-complementary PCR primer (MP-PCR) (*see* **Note 14**).
2. The products are separated by on 1.4 % agarose gels, stained with ethidium bromide, and photographed.
3. Before hybridization to a new probe, membranes are stripped by washing in 5 mM EDTA at 60 °C (2× 30 min).

3.8.3 Hybridization with Microsatellite-Complementary Probes

1. The gel is either dried or blotted onto a nylon membrane.
2. Hybridize to a [^{32}P]-labeled, microsatellite-complementary oligonucleotide probe.
3. Hybridization is done overnight at 42 °C containing 20–40 ng/mL of the probe.
4. Filters are washed twice for 5 min at room temperature in 2× SSC; 0.1I SDS followed by two final washing steps (2 × 15 min) at different stringency.
5. The stringency can be varied through temperature (50–65 °C) and salt concentration (1× SSC; 0.1 % SDS to 0.1× SSC; 0.1 % SDS).
6. Positive signals are detected by either chemiluminescence system and documented by exposure to X-ray film for 1–2 h.

3.9 Random Amplified Hybridization Microsatellites (RAHM)

1. The DNA is amplified using RAPD primers (*see* Subheading 3.3.1).
2. The amplified products are separated by gel electrophoresis (*see* Subheading 3.3.2).
3. The polymorphisms on the agarose gel are identified and scored (*see* Subheading 3.3.3).
4. The amplified DNA is then transferred onto Hybond-N+ filters using Southern blot procedures.
5. The filters are then hybridized with radiolabeled oligonucleotide probes carrying simple sequence repeats (SSR).
6. The luminescent signals produced are detected by autoradiography. Hybridizing bands are named random amplified hybridization microsatellites (RAHM).

3.10 Cleaved Amplified Polymorphic Sequences (CAPs)

1. Genomic DNA is isolated (*see* Subheadings 3.1 and 3.2).
2. Amplifying the different CAPS marker locus by PCR.
3. Analyzing the PCR by gel electrophoresis to confirm amplification of DNA and the yield.

4. Mix 5 μL PCR and 10 μL digest mix, incubate at 37 °C for 5 h, and then heat to 65 °C for 5 min.
5. Mix equal parts of digest mix and formamide loading dye. Denature sample by heating at 94 °C for 5 min and then placing tube on ice.
6. Resolve restriction fragments using 1× TBE, 8.25 % polyacrylamide gel.
7. Load 2.5 μL of the denatured sample per lane.
8. Denature by heating at 94 °C for 5 min and then placing tube on ice.
9. Load 3.5 μL of the denatured ladder per lane, equivalent to 117 ng DNA.
10. Run gel at 80 W for approximately 80 min or until the bromophenol blue dye front has reached the bottom of the gel.
11. Follow usual silver staining protocol to stain gel (*see* Subheading 3.6.2).

4 Notes

1. Randomly amplified polymorphic DNA (RAPD) and arbitrarily primed PCR (AP-PCR) use relatively low concentrations (e.g., 0.2 μmol/L) of single short oligonucleotide primers in the PCR with annealing temperatures ranging from 37 to 40 °C, and up to 20 markers can be simultaneously amplified and detected. DNA amplification fingerprinting (DAF) also implements a single short oligonucleotide primer but at a higher concentration (5 μmol/L), and higher annealing temperatures (53–57 °C) using DNA polymerase Stoffel Fragment in PCR.
2. Although the sequences of RAPD primers are arbitrarily chosen, two basic criteria must be met: a minimum of 40 % GC content (50–80 % GC content is generally used) and the absence of palindromic sequence (a base sequence that reads exactly the same from right to left as from left to right). Because G–C bond consists of three hydrogen bridges and the A–T bond of only two, a primer-DNA hybrid with less than 50 % GC will probably not withstand the 72 °C temperature at which DNA elongation takes place by DNA polymerase [1].
3. Data from at least 10 primers with a total of 100 RAPD bands are needed to produce a stable classification [44].
4. The rationale behind primer designing in SRAP is based on the fact that exons are normally in GC-rich regions. The core is followed by three selective nucleotides at the 3′end. The filler sequences of the forward and reverse primers must be different from each other and can be 10 or 11 bases long.

5. The most important factor for reproducibility of the RAPD profile has been found to be the result of inadequately prepared template DNA which could be overcome through choice of an appropriate DNA extraction protocol to remove any contaminants [45]. Differences between the template DNA concentrations of two DNA samples will result in the loss or gain of some bands.
6. RAPD reaction is far more sensitive than conventional PCR because of the length of a single and arbitrary primer used to amplify anonymous regions of a given genome. Optimization of reaction conditions should precede the actual RAPD analysis to get consistent and reproducible results. Following optimizations are essential: template DNA concentration and quality, *Taq* DNA polymerase concentration, Mg^{2+} ion concentration, primer concentration and annealing temperature, and primers suitable for detection of polymorphic loci in the taxa to be analyzed [46].
7. Too many RAPD cycles can increase the amount and complexity of nonspecific background products, while too few cycles give low product yield. The optimum number of cycles will depend mainly upon the starting concentration of target DNA when other parameters are optimized.
8. The probability of a scored RAPD band being scored in replicate data is strongly dependent on the uniformity of amplification conditions between experiments, as well as relative amplification strength of the RAPD band [26].
9. Deleting inconsistent or faint bands or using only those bands that are reproducible introduces false negatives and simply ignoring RAPD artifacts, and using all bands introduces false positive into RAPD data [47].
10. The criteria for selecting scoring bands include reproducibility and consistency—the experiments need to be repeated to achieve reproducible results, thickness, and size of the bands.
11. If estimates of the percent of false-positive and false-negative bands in the RAPD data are available (such as when replicate runs have been made), equations described earlier [48] can be used to determine the actual bias by subtracting the true value from the estimated value. Once the bias is known, it can be used to determine whether the RAPD protocol has been optimized sufficiently to provide accurate enough estimates of the similarities.
12. Other softwares like PAUP, PHYLIP, CLINCH, MaClade, PopGene, and Arlequin can also be used to accomplish the cluster algorithms and for phylogenetic analysis.

13. In SCAR, the longer primer sequence increases the specificity of the PCR and produces results less sensitive to changes in reaction conditions and thus more reproducible than RAPD [49].
14. If RAPD gels are used for RAMPO analysis, banding patterns are generally less complex, less variable, and easier to interpret than those derived from MP-PCR gels [50].

References

1. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18: 6531–6535
2. Vos P, Hogers R, Bleeker M, Reijans M, Van De Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414
3. Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215–222
4. Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535
5. Arif IA, Bakir MA, Khan HA, Al Farhan AH, Al Homaidan AA, Bahkali AH, Al Sadoon M, Shobrak M (2010) A brief review of molecular techniques to assess plant diversity. Int J Mol Sci 11:2079–2096
6. Vierling RA, Nguyen HT (1992) Use of RAPD markers to determine the genetic diversity of diploid, wheat genotypes. Theor Appl Genet 84:835–838
7. dos Santos JB, Nienhuis J, Skroch P, Tivang J, Slocum MK (1994) Comparison of RAPD and RFLP genetic markers in determining genetic similarity among *Brassica oleracea* L. genotypes. Theor Appl Genet 87:909–915
8. Maria D, Angela P, Chialexei L (2008) Characteristics of RAPD markers inbreeding of *Cucumis sativus* L. Roum Biotech Lett 13:3843–3850
9. Khadari B, Breton C, Moutier N, Roger JP, Besnard G, Bervillé A, Dosba F (2003) The use of molecular markers for germplasm management in a French olive collection. Theor Appl Genet 106:521–529
10. Tinker NA, Fortin MG, Mather DE (1993) Random amplified polymorphic DNA and pedigree relationships in spring barley. Theor Appl Genet 85:976
11. Mailer RJ, Scarth R, Fristensk B (1994) Discrimination among cultivars of rapeseed (*Brassica napus* L.) using DNA polymorphism amplified from arbitrary primers. Theor Appl Genet 87:697–704
12. Marshall P, Marchand MC, Lisieczko Z, Landry BS (1994) A simple method to estimate the percentage of hybridity in canola (*Brassica napus*) F1 hybrids. Theor Appl Genet 89:853–858
13. Congiu L, Chicca M, Cella R, Rossi R, Bernacchia G (2000) The use of randomly amplified polymorphic DNA (RAPD) markers to identify strawberry varieties: a forensic application. Mol Ecol 9:229–232
14. Bligh HFJ (2000) Detection of adulteration of Basmati rice with non premium long grain rice. Int J Food Sci Tech 35:257–265
15. Adams RP, Demeke T (1993) Systematic relationships in Juniperus based on random amplified polymorphic DNA. Taxon 42:553–571
16. Wilkie SE, Issac PG, Slater RJ (1993) Random amplified polymorphic DNA (RAPD) markers for genetic analysis in Allium. Theor Appl Genet 86:497–504
17. Isabel N, Tremblay l, Michaud M, Tremblay FM, Bousquet J (1993) RAPDs as an aid to evaluate the genetic integrity of somatic embryogenesis-derived populations of *Picea mariana* (Mill.) B.S.P. Theor Appl Genet 86:81–87
18. Lewis PO, Snow AA (1992) Deterministic paternity exclusion using RAPD markers. Mol Ecol 1:155–160
19. Crawford DJ, Brauner S, Cosner MB, Steussy TF (1993) Use of RAPD markers to document the origin of intergeneric hybrid *Margyracaena skottsbergii* (rosaceae) on the Juan Fernandez Islands. Am J Bot 80:89–92
20. Waugh R, Baird E, Powell W (1992) The use of RAPD markers for the detection of gene introgression in potato. Plant Cell Rep 11:466–469
21. Halima HS, Bahy AA, Tian-Hua H, Da-Nian Q, Xiao-Mei W, Qing-Dong X (2007) Use of random amplified polymorphic DNA analysis for economically important food crops. J Integr Plant Biol 49(12):1670–1680
22. Hedrick P (1992) Shooting the RAPDs. Nature 355:679–680
23. Smith JSC, Williams JGK (1994) Arbitrary primer mediated fingerprinting in plants: case

studies in plant breeding, taxonomy and phylogeny. In: Schierwater B, Streit B, Wagner GP, DeSalle R (eds) Molecular ecology and evolution: approaches and applications. Birkhauser Verlag Basel, Switzerland, pp 5–15
24. Paran I, Michelmore RW (1993) Development of reliable PCR-based markers linked to Downy Mildew resistance genes in lettuce. Theor Appl Genet 85:985–993
25. Jones CJ et al (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol Breed 3:381–390
26. Skroch P, Nienhuis J (1995) Qualitative and quantitative characterization of RAPD variation among snap bean (*Phaseolus vulgaris*) genotypes. Theor Appl Genet 91:1078–1085
27. Challahan LM, Weaver KR, Caetano-Anolles G, Bassam BJ, Gresshoff PM (1993) DNA fingerprinting of turfgrasses. Int Turfgrass Soc Res J 7:761–767
28. Caetano-Anollés G, Gresshoff PM (1994) DNA amplification fingerprinting using arbitrary mini-hairpin oligonucleotide primers. Biotechnology 12:619–623
29. Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA 88:9828–9832
30. Martin GB, Williams JGK, Tanksley SD (1991) Rapid identification of markers linked to a Pseudomonas resistance gene in tomato by using random primers and near-isogenic lines. Proc Natl Acad Sci USA 88:2336–2340
31. Rafalski JA, Tingey SV (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends Genet 9:275–280
32. Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218
33. Welsh J, Honeycutt RS, McClelland M, Sobral BWS (1991) Parentage determination in maize hybrids using the arbitrarily primed polymerase chain reaction (AP-PCR). Theor Appl Genet 82:473–476
34. Caetano-Anollés G, Bassam BJ, Gresshoff PM (1991) DNA amplification finger printing using short arbitrary oligonucleotide primers. Biotechnology 9:553–557
35. Somsri S, Bussabakornkul S (2008) Identification of certain papaya cultivars and sex identification in papaya by DNA amplification fingerprinting (DAF). Acta Hort (ISHS) 787:197–206
36. Luro S (1995) DNA amplified fingerprinting, a useful tool for determination of genetic origin and diversity analysis in citrus. Hort 30(5): 1063–1067
37. Li G, Quiros CF (2001) Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in Brassica. Theor Appl Genet 103:455–546
38. Cifarelli RA, Gallitelli M, Cellini F (1995) Random amplified hybridization microsatellites (RAHM): isolation of a new class of microsatellite containing DNA clones. Nucleic Acid Res 23:3802–3803
39. Richardson T, Cato S, Ramser J, Kahl G, Weising K (1995) Hybridization of microsatellites to RAPD: a new source of polymorphic markers. Nucleic Acids Res 23:3798–3799
40. Konieczny A, Ausubel FM (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-nomically important pathogen based markers. Plant J 4:403–410
41. Jarvis P, Lister C, Szabo V, Dean C (1994) Integration of CAPS markers into the RFLP map generated using recombinant inbred lines of Arabidopsis thaliana. Plant Mol Biol 24:685–687
42. Doyle JJ, Doyle LJ (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15
43. Rohlf FJ (1998) NTSYS-pc numerical taxonomy and multivariate analysis system. Version 2.02. Exeter Publications Setauket, New York
44. Demeke T, Adams RP (1994) The use of RAPD-PCR analysis in plant taxonomy and evolution. In: Griffin HG, Griffin AM (eds) PCR technology: current innovations. CRC Press, Boca Raton, FL
45. Micheli MR, Bova R, Pascale E, D'Ambrosio E (1994) Reproducible DNA fingerprinting with the random amplified polymorphic DNA (RAPD) method. Nucleic Acids Res 22: 1921–1922
46. Wolff K, Schoen ED, Peters-Van RJ (1993) Optimizing the generation of random amplified polymorphic DNA in chrysanthemum. Theor Appl Genet 86:1033–1037
47. Lamboy WF (1994) Computing genetic similarity coefficients from RAPD data: the effects of PCR artifacts. PCR Meth Appl 4:31–37
48. Lamboy WF (1994) Computing genetic similarity coefficients from RAPD data: correcting for the effects of PCR artifacts caused by variation in experimental conditions. PCR Meth Appl 4:38–43
49. Hernandez P, Martin A, Dorado G (1999) Development of SCARs by direct sequencing of RAPD products: a practical tool for introgression and marker-assisted selection of wheat. Mol Breed 5:245–253
50. Davis MJJ, Bailey CS, Smith CK (1997) Increased informativeness of RAPD analysis by detection of microsatellite motifs. Biotechniques 23:285–290

Chapter 11

Multilocus Profiling with AFLP, ISSR, and SAMPL

Luis F. Goulao and Cristina M. Oliveira

Abstract

Molecular markers which sample multiple loci simultaneously, like amplified fragment length polymorphism (AFLP), inter-simple sequence repeats (ISSR), and selective amplification of microsatellite polymorphic loci (SAMPL), produce highly informative fingerprints due to their high effective multiplex ratio and expected heterozygosity. Moreover, these markers can be generated for DNA of any organism without initial investment in primer/probe development or in sequence analyses. The fragments produced can be visualized either by agarose or polyacrylamide gel electrophoresis followed by autoradiography or silver staining or via separation and detection on automatic DNA sequencers. Here, we describe detailed protocols based on the original methods aimed to obtain these markers optimized to be resolved on polyacrylamide gel electrophoresis and detected by silver staining which provides a fast, sensitive, and cost-effective method.

Key words Amplified fragment length polymorphism, DNA markers, Inter-simple sequence repeats, Polyacrylamide gel electrophoresis, Polymerase chain reaction, Selective amplification of microsatellite polymorphic loci

1 Introduction

Multilocus profiling is advantageous to taxonomic and genetic diversity studies due to the increased number of polymorphic loci that can be assayed without compromising the level of polymorphisms observed at the loci. A high multiplex ratio is an advantage in applications aimed at typing individuals or measuring genetic diversity, whereas high information content is required for heritability tests and mapping. A number of such multilocus markers are available to detect polymorphisms in nuclear DNA, of which AFLP, ISSR, and SAMPL are the most popular, with large prominence on the former two. Their high resolving power, coupled with the possibility of producing perfectly usable markers from genomes for which no previous sequence knowledge exists and their technical simplicity, makes these markers a very attractive choice both in basic (e.g., phylogenetic analysis, diversity studies, identification of taxa, population structure, and search for useful genes in individuals or populations) and applied research (e.g., genetic mapping,

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_11,

marker-assisted selection for crop improvement and breeding programs, paternity testing, and food traceability), which remain important issues in the modern scenario.

Amplified fragment length polymorphism (AFLP) combines sequential DNA restriction digestion with polymerase chain reaction (PCR) amplification [1]. A genomic DNA sample is first simultaneously digested with one rare- and one frequent-cut restriction enzyme, and the resulting fragments are amplified using primers partially complementary to adapters ligated to the over-hanged ends. Arbitrary selective nucleotides included at the 3′ end of the primers to reduce the number of amplified fragments generated, making them resolvable in standard sequencing gels. In a single reaction, about 50–100 loci can be analyzed depending on the genome size, the frequency of cut obtained using a given enzyme combination, the number of selective nucleotides, and the resolution of the electrophoresis system. A virtually unlimited number of polymorphisms can be obtained using combinations of different restriction enzymes and selective nucleotides in the primers employed.

Inter-simple sequence repeats (ISSR) is a microsatellite-based method that relies on the ubiquity and hypervariable nature of these regions in eukaryotic genomes. The ISSR method was first reported in 1994 [2, 3] with a rationale nearly identical to the RAPD technique (see previous chapter), except that a single primer is used, composed of a microsatellite sequence anchored at the 3′ or 5′ end of the repeat adjacent regions by one to four selective, often degenerate nucleotides. DNA is amplified between two opposed microsatellites of the same type and allelic polymorphisms occur whenever one genome is missing the sequence repeated or has a deletion or insertion that modifies the distance between repeats. For 5′-anchored primers, polymorphisms also occur due to differences in the length of the microsatellite. The sequences of repeats and anchored nucleotides are arbitrarily selected. Since longer primers can be devised, PCR can be set using annealing temperatures higher than those used to obtain RAPD markers, minimizing the reproducibility problems associated with RAPDs. Hence, ISSR markers are DNA fragments of about 100–3,000 bp located between adjacent, oppositely oriented microsatellite regions. About 10–60 scorable fragments, which are usually below 1,500 bp, from multiple loci are generated simultaneously. Other techniques closely related to ISSR analysis were developed, namely, single primer amplification reaction (SPAR) [3] and directed amplification of minisatellite-region DNA (DAMD) [4], that use a single primer containing only the core motif of a micro- or a minisatellite, respectively. It should be noted that while di- and trinucleotide simple repeats are considered to have a random distribution [5], four-nucleotide repeats seems to be distributed in preferential locations on the chromosomes [6] and thus are less suitable for diversity studies.

Selective amplification of microsatellite polymorphic loci (SAMPL) was developed by [7, 8] to provide a high multiplex ratio marker system that combines the advantages of microsatellites and AFLP markers. The same pre-selective amplification products produced for AFLP analysis (see method below) are used as templates for amplification in the SAMPL protocol. SAMPL differs from AFLP in the primers used for subsequent selective amplification. While in ISSR technique, oligonucleotides primers correspond to known, repeated sequences in various combinations in PCR reactions, to obtain SAMPL markers, the selective amplification is attained using one standard AFLP primer and a primer complementary to microsatellite sequences. The design of SAMPL primers used in the original procedure is based only on compound SSR sequences consisting of two different adjacent dinucleotide repeats [9], but in our lab, we also obtained successful results with the following primer combinations: *Mse*I or *Eco*RI adapter-based primers coupled with ISSR primers and *Mse*I or *Eco*RI adapter-based primers coupled with primers flanking SSR [10]. These primer strategies allow the amplification of any type of repeat structure (not only compound microsatellites) and can be extended to different types of tri-, tetra-, and penta-nucleotide repeats.

The main drawback of multilocus markers comes from the fact that these methods reveal polymorphisms resulting both from length variation between conserved repeat sites (codominant loci) and from individual repeat sites whose presence or absence differs (dominant loci) between genomes. Hence, the dominant inheritance of AFLP, ISSR, and SAMPL markers, preventing a heterozygous individual to be distinguished from a dominant homozygous, must be taken into account in the selection of backcrosses and statistical analysis routines in mapping programs and should be considered while planning some applications, like mapping. Moreover, same size products on the gels do not necessarily have the same sequence, which limits its use in estimation of diversity at the interspecific level. Nevertheless, despite these disadvantages, the high effective multiplex ratio provided by these markers makes them very powerful tools for most of the current genome analyses. Although some works report that the use of different marker systems results in differences in phenetic similarities between a set of genotypes, in our hands, the results obtained with AFLP, ISSR, RAPD, and SSR markers were significantly comparable [11, 12].

2 Materials

All solutions must be made up using sterile double-distilled or deionized water (MilliQ water for enzymatic reactions), and all chemicals must be analytical reagent grade. As in all molecular biology procedures, work surfaces should be cleaned and gloves should be worn for all procedures. Sterile, disposable plasticware should be used wherever possible.

2.1 Amplified Fragment Length Polymorphism (See Note 1)

1. Template genomic DNA to be assayed (*see* **Note 2**): For genomes smaller or larger than 500 Mb, use 50 or 100 ng DNA for template preparations, respectively. Determine DNA concentrations by measuring OD_{260} in a spectrophotometer or NanoDrop® and confirm the measurement and the integrity of DNA by electrophoresing the sample together with a series of phage λ DNA dilutions ranging from 50 to 500 ng in standard 1 % agarose in 1× TAE gels (*see* **Note 3**).
2. Distilled and MilliQ water.
3. Two restriction enzymes: One rare (6 bp) and one frequent (4 bp) cutter and appropriate reaction buffer (available from several suppliers) (*see* **Note 4**).
4. Adaptor/ligation solution: 0.4 mM ATP, 10 mM Tris–HCl, pH 7.5, 10 mM Mg-acetate, 50 mM K-acetate. Prepare 10 mL by weighting 2.2 mg ATP (adenosine 5′-triphosphate disodium salt hydrate), 47.07 mg potassium acetate and dissolve them in 90 mL water. Mix by stirring and, when dissolved, add 100 μL of 1 M Tris–HCl and 100 μL of 1 M magnesium acetate solution. Complete to 10 mL with water.
5. Matching adaptors: *Eco*RI/*Mse*I adaptors at 5 and 50 pmol/μL concentration, respectively. To make a 500 μM solution of adaptor pairs (for *Eco*RI and *Mse*I; *see* **Note 4**), combine 40 μL 1 mM F adaptor (5′-CTCGTAGACTGCGTACC-3′for *Eco*RI and 5′-GACGATGAGTCCTGAG-3′ for *Mse*I) with 40 μL 1 mM R adaptor (5′-AATTGGTACGCAGTCTAC-3′ for *Eco*RI and 5′-TACTCAGGACTCAT-3′ for *Mse*I), incubate at 94 °C for 5 min, and let slowly cool to room temperature (*see* **Note 5**). Spin briefly (10 s) to collect the reaction at the bottom of the tube. Dilute the *Eco*RI adaptor to 5 μM by mixing 2 μL of the 500 μM solution with 198 μL of adaptor/ligation solution and the *Mse*I adaptor to 50 μM by mixing 20 μL of the 500 μM solution with 180 μL of adaptor/ligation solution (*see* **Note 6**).
6. dNTP mix (10 mM each): Add 10 μL of each 100 mM dNTP solution to 60 μL water.
7. T4 DNA ligase (1 U/μL) (available from several commercial suppliers with applicable reaction buffer).
8. AFLP selective pre-amplification primer mixture (*Eco*RI + 1 primer, 5′-GACTGCGTACC*AATTC***A**-3′; *Mse*I + 1 primer, 5′-GATGAGTCCTGAG*TAA***C**-3′; underlined, core region; italic, enzyme-specific region; bold, selective nucleotide): Dilute each primer in water or TE buffer to a 50 ng/μL concentration.
9. AFLP selective amplification primer mixture (*Eco*RI + 3 primer, 5′-GACTGCGTACC*AATTC***ANN**-3′; *Mse*I + 3 primer,

5′-GATGAGTCCTGAG*TAA***CNN**-3′; underlined, core region; italic, enzyme-specific region; bold, selective nucleotide): Dilute each primer in water or TE buffer to a 30 ng/μL concentration.

10. *Taq* DNA polymerase (5 U/μL) and appropriate reaction buffer (available from many commercial suppliers) (*see* **Note 7**).
11. 1 M Tris–HCl (hydroxymethyl) aminomethane: Dissolve 60.57 g Tris–HCl powder in 450 mL water using a magnetic stirring bar. Adjust pH to 7.5 using HCl and complete the volume to 500 mL with water.
12. 500 mM EDTA (pH 8.0) diaminoethane-tetraacetic acid: Dissolve 18.6 g in 80 mL water, stir vigorously using a magnetic stirrer, and adjust the pH to 8.0 using 1 N NaOH (about 2 g of NaOH pellets). Complete volume to 100 mL with water (*see* **Note 8**).
13. TAE buffer: Prepare a 50× stock solution of TAE by weighing 242 g Tris base and dissolving by stirring in approximately 750 mL deionized water. Carefully add 57.1 mL glacial acetic acid and 100 mL of 0.5 M EDTA pH 8.0 or 18.6 g EDTA powder. Adjust the solution to a final volume of 1 L. The pH should be ca. 8.5 and doesn't need adjustment. The working solution of 1× TAE buffer is prepared by diluting the stock solution by 50× in deionized water.
14. TE buffer: 10 mM Tris–HCl, pH 7.5, 1 mM EDTA. To make 100 mL of TE buffer, add 1 mL 1 M Tris–HCl pH 7.5 and 200 μL 500 mM EDTA pH 8.0 to 98.8 mL water, mix, and autoclave before use.
15. $TE_{0.1}$ buffer: 1 mM Tris–HCl, pH 7.5, 100 μM EDTA. Mix 100 μL of TE buffer with 900 μL of water.
16. Bench microcentrifuge.
17. Programmable dry incubators or water baths.
18. Programmable thermal cycler.
19. Microwave oven.
20. Standard horizontal agarose gel electrophoresis apparatus.
21. Power supply.

2.2 Inter-simple Sequence Repeats

1. Genomic DNA.
2. MilliQ water.
3. dNTP mix (10 mM each): Add 10 μL of each 100 mM dNTP solution to 60 μL water.
4. *Taq* DNA polymerase (5 U/μL) and appropriate reaction buffer (available from many commercial suppliers).
5. ISSR primer: Dilute in water or TE buffer to a 20 μM concentration (*see* **Note 9**).

6. Bench microcentrifuge.
7. Programmable thermal cycler.

2.3 Selective Amplification of Microsatellite Polymorphic Loci

1. Template (AFLP pre-selective amplification products, after 1:10 dilution): Obtained as described for the AFLP protocol (Subheading 3.1.5, **step 2**).
2. MilliQ water.
3. dNTP mix (10 mM each): Add 10 μL of each 100 mM dNTP solution to 60 μL water.
4. *Taq* DNA polymerase (5 U/μL) and appropriate reaction buffer (available from many commercial suppliers).
5. Primer 1—AFLP primer at 30 ng/μL (*see* Subheading 2.1, **item 9**).
6. Primer 2—ISSR primer at 10 μM (*see* Subheading 2.2, **item 5**, *diluted 1:1*).
7. Bench microcentrifuge.
8. Programmable thermal cycler.

2.4 Preparation of the Glass Plates for Polyacrylamide Gels

1. Absolute ethanol.
2. 70 % ethanol: Add 350 mL absolute ethanol to 150 mL water (*see* **Note 10**).
3. Binding solution: Always make fresh binding solution in a fume hood. Prepare two 1.5 mL Eppendorf tubes with binding solution by adding, in each tube, 5 μL Bind-Silane® (γ-methacryloxypropyl trimethoxysilane) to 50 μL glacial acetic acid in 945 μL absolute ethanol.
4. Repel solution (dimethyldichlorosilane): Available commercially as Repel Silane®. Use directly from the bottle.
5. 2 M NaOH: In a glass beaker, weigh 80 g NaOH and stir, a little at a time (to keep the heat down), into a large volume of water. Then, dilute the solution to make a final volume of 1 L.
6. Short (390 × 330 × 6 mm) and long (420 × 330 × 6 mm) glass plates.

2.5 Preparation and Casting of Polyacrylamide Gels

1. Acrylamide Long Ranger® Gel 50 % solution: Available from many scientific supply companies.
2. Running buffer: 10× TBE (Tris–Borate–EDTA): dissolve 108 g Tris base [tris(hydroxymethyl) aminomethane], 55 g of boric acid, 7.5 g of EDTA, disodium salt in 800 mL of deionized water. Dilute the buffer to 1 L. Use at 1× strength for running buffer and gel matrix (*see* **Note 11**).
3. Urea 9.6 M: Dissolve 290 g urea in 350 mL of dH_2O. Complete volume to 500 mL with water (*see* **Note 12**).

4. Ammonium persulfate (APS): In an Eppendorf tube, prepare 10 % (w/v) ammonium persulfate by dissolving 100 mg in1 mL of dH_2O (*see* **Note 13**).
5. *N*,*N*,*N*′,*N*′-Tetramethyl-ethylenediamine (TEMED): $(CH_3)_2NCH_2CH_2N(CH_3)_2$.
6. 2× denaturant loading dye: 98 % formamide, 10 mM EDTA, 0.25 % xylene cyanol. To prepare 10 mL, add 200 μL of 0.5 M EDTA pH 8.0 and ca. 2 mg of xylene cyanol (*see* **Note 14**) to 9.8 mL of deionized formamide 99 %. Mix by gentle inverting the tube 5–10 times and store at 4 °C.
7. 2× loading dye for non-denaturing polyacrylamide gels (0.01–0.05 % xylene cyanol; 6 M urea; 0.01 % sucrose): For 10 mL, add ca. 2 mg of xylene cyanol (*see* **Note 14**) and 1 mg sucrose to 6.25 mL of urea 9.6 M solution. Fill up with water to complete a volume of 10 mL. Mix by gentle inverting the tube 5–10 times and store at 4 °C.
8. 10 bp ladder marker (1 μg/μL; available from several suppliers).
9. 100 bp ladder marker (1 μg/μL; available from several suppliers).
10. A pair of 0.4 or 0.35 mm vinyl or Mylar spacers with foam block.
11. A pair of 0.4 or 0.35 mm vinyl or Mylar sharkstooth combs.
12. Sequencing gel rig (e.g., Model S2 Sequencing Gel Electrophoresis Apparatus, Biometra).
13. Three pairs of clamps or a casting holder (optional).
14. High voltage power supply.

2.6 Silver Staining of Gels

Use analytical grade chemicals (*see* **Note 15**). Silver nitrate is very hazardous in case of skin and eye contact, ingestion, and inhalation. Formaldehyde and formamide are suspected carcinogen and teratogen. They are both toxic and irritating through contact, inhalation, and ingestion, so all steps should be performed and should be used in a ventilated fume hood. These substances should be stored in the dark, at room temperature to prevent oxidation. Dispose these solutions in accordance with federal, state, and local environmental control regulations.

1. Fixing/stopper solution: 10 % glacial acetic acid. Add 1.7 L of water to a 2 L graduated cylinder; add 200 mL 100 % glacial acetic acid and make up to a volume of 2 L with water (*see* **Note 16**).
2. 10 mg/mL sodium thiosulfate solution: Weigh 100 mg of $Na_2SO_3.5H_2O$ and add water to a volume of 10 mL (*see* **Note 17**).
3. Impregnation silver solution: 10 % silver nitrate, 0.4 % formaldehyde: Weigh 2 g $AgNO_3$ and dissolve in 1.7 mL of water, in

a glass beaker, with stirring. When dissolved, add 3 mL 37 % formaldehyde (HCOH) and keep stirring for additional 10–15 min (*see* **Note 18**). Add water to a volume of 2 L.

4. 2× developer solution: 3 % Na_2CO_3, 0.4 % formaldehyde, 0.0005 % Na_2SO_3 (*see* **Note 19**). Weigh 60 g Na_2CO_3 and dissolve in 850 mL of water with stirring (*see* **Note 20**). Add 1.5 mL 37 % HCOH and 500 μL of the 10 mg/L sodium thiosulfate solution. Complete to a volume of 1 L with water and mix by stirring or swirling the flask. Add 1 L cold water to make 1× developer solution.
5. Staining tray (stainless steel, glass, or plastic containers—with a dimension to accommodate 33.2×41.6 cm glass plates). The tray must be perfectly clean (*see* **Note 21**).
6. Vertical waving shaker.
7. Ventilated fume hood.

3 Methods

It is imperative that all solutions are thawed completely and all buffers are mixed before use. Keep everything on ice as much as possible, to keep the enzymes from working before you want them and to minimize evaporation. Enzymes should be removed from the freezer immediately before use, placed on ice, and immediately returned to the freezer after use. Prepare master mixes (on ice) whenever possible, incorporating a correction factor to overcome mix losses (outside the tip and small pipetting errors). The easiest way is to multiply the volume of each reagent to 1.03 to obtain 103 % of the needed mixture volume. It is important to amplify each set of samples and a particular primer twice, giving two amplification replicates. A few (maximum 3 %) bands not reproducible are expected in the AFLP analysis.

3.1 Amplified Fragment Length Polymorphism

Adapted from Vos et al. [1] with modifications to cope with non-radioactive labeling and detection:

1. Isolate and purify genomic DNA (*see* **Note 2**).

3.1.1 Restriction Endonuclease Digestion of Genomic DNA

1. For each sample, pipette 250 ng of DNA (in a volume up to18 μL of MilliQ water) into a 1.5 mL Eppendorf tube.
2. Add appropriate restriction buffer to the working concentration (5 or 2.5 μL of 5× or 10× buffer, respectively).
3. Add 2.5 U of each *Mse*I and *Eco*RI (*see* **Note 4**).
4. Complete with water to a volume of 25 μL.
5. Mix by gently tapping the tube and centrifuge briefly (10 s) to collect the content in the bottom of the tube.

6. Incubate for 4 h at 37 °C (*see* **Note 22**).
7. Inactivate the restriction endonucleases incubating the mixture for 15 min at 70 °C and immediately placing the tube on ice.
8. Collect contents by a brief centrifuge spin (10 s).

3.1.2 Ligation of Adaptors

1. To each tube containing the digested DNA, add 24 μL of the ligation/adaptor solution (*see* **Note 23**).
2. Pipette 1 U of T4 DNA ligase to the solution.
3. Complete volume to 50 μL with water if necessary (*see* **Note 24**).
4. Mix gently and centrifuge for 10 s to collect the mixture in the bottom of the tube.
5. Incubate for 4 h—overnight at room temperature (20–24 °C).
6. Perform a 1:10 dilution of the ligation mixture by diluting 10 μL of the reaction mixture with 90 μL of TE buffer (*see* **Note 25**).

3.1.3 Selective Pre-amplification

1. Transfer 5 μL of diluted template DNA from the previous step to a 0.2 or 0.5 mL thin-walled PCR microtube.
2. Add appropriate PCR buffer to the corresponding working concentration (typically 5 μL of 10× PCR buffer in a 50 μL total volume).
3. Add 2.5 μL of the each pre-selective amplification primer (*Eco*RI + 1 + *Mse*I + 1 primer).
4. Add 1 μL of 10 mM dNTPs mix.
5. Add 5 μL of 1.5 mM $MgCl_2$ (*see* **Note 26**).
6. Add 0.5 μL *Taq* DNA polymerase (1 U).
7. Mix gently by tapping the tube and centrifuge briefly (10 s) to collect the reaction at the bottom of the tube.
8. Add MilliQ water to complete a 50 μL total volume (*see* **Note 27**).
9. Put the tubes in a thermal cycler and perform 30 cycles as follows (*see* **Note 28**): 94 °C denaturation for 30 s, 56 °C annealing for 60 s, and 72 °C extension for 60 s (*see* **Note 29**). Soak temperature is 12 °C (*see* **Note 30**). Keep the samples in the fridge until electrophoresed.

3.1.4 Verifying Successful Amplification of Target Sequences

1. Prepare 50 mL 1 % agarose gel in 1× TAE buffer, containing 1.25 μL of GreenSafe dye (*see* **Note 3**).
2. Prepare sample in small tube or on top of a Parafilm® cover: mix 8 μL pre-selective amplification sample with 2 μL loading dye (*see* **Note 3**).
3. Load 500 ng of molecular weight standard (100 bp ladder) in one well and 10 μL of each sample in the other wells of the agarose gel.

4. Run for 10–20 min at 5 V/cm (until the bromophenol blue reaches the bottom of the gel).
5. Visualize under ultraviolet light, in a transilluminator (*see* **Note 31**).

3.1.5 Selective PCR Amplification

1. To prepare the template for selective AFLP amplification, combine, for each reaction product, 3 μL of the pre-selective amplification reaction product with 147 μL $TE_{0.1}$ buffer.
2. Mix by gently tapping the tube and spin for 10 s.
3. Store at 2–6 °C until use (*see* **Note 32**).
4. For each reaction, add 5 μL of the diluted pre-selective PCR products template to a 0.5 or 0.2 mL PCR microtube.
5. Add 2 μL of 10× PCR buffer.
6. Add 1 μL of each selective primer (*Mse*I + 3 + *Eco*RI + 3 primer).
7. Add 0.4 μL10 mM dNTPs mix.
8. Add 2 μL 1.5 mM $MgCl_2$ (*see* **Note 26**).
9. Add 0.2 μL (1 U) *Taq* DNA polymerase.
10. Mix gently and centrifuge briefly to collect reactions at the bottom of the tube.
11. Complete to 20 μL with water (*see* **Note 27**).
12. Amplify by means of a touch-down PCR as follows: one cycle of 94 °C denaturation for 30 s, 65 °C annealing for 30 s, and 72 °C extension for 60 s, followed by 12 cycles with the annealing temperature lowered by 0.7 °C per cycle. Complete with 23 further cycles of 94 °C for 30 s, 58 °C for 30 s, and 72 °C for 60 s (*see* **Notes 29** and **33**). Set soak temperature to 12 °C (*see* **Note 30**). Keep the samples at the fridge until electrophoresed.

3.2 Inter-simple Sequence Repeats

Adapted from Zietkiewicz et al. [2]:

1. For a standard reaction (20 μL volume), add 2 μL of 10× PCR buffer to a 0.5 or 0.2 mL PCR microtube.
2. Add 1 μL of each primer (*see* **Note 34**).
3. Add 0.4 μL 10 mM dNTPs mix.
4. Add 2 μL 1.5 mM $MgCl_2$ (*see* **Notes 26** and **35**).
5. Add 0.2 μL (1 U) *Taq* DNA polymerase (*see* **Note 35**).
6. Complete to 18 μL with water.
7. Mix by gently tapping the tube and spin briefly to collect the mix in the bottom of the tube.
8. Aliquot 18 μL to each tube and add 2 μL of respective DNA sample (diluted at 10–20 ng/μL) (*see* **Note 27**).

9. Transfer the tubes to a thermal cycler programmed as follows: an initial denaturation at 94 °C for 90 s, 30 cycles of denaturing at 94 °C for 30 s, annealing at 48 °C for 45 s, extension at 72 °C for 90 s, and a final extension of 7 min at 72 °C (*see* **Notes 29** and **36**). Store at 4–12 °C (short term) or at–20 °C (long term).
10. Keep the samples at the fridge or proceed to electrophoresis.

3.3 Selective Amplification of Microsatellite Polymorphic Loci

Adapted from Morgante and Vogel [7]:

1. For each reaction add 5 μL of the diluted pre-amplification template PCR products (from Subheading 3.1.5, **step 2**) to a 0.5 or 0.2 mL PCR microtube.
2. Add 2 μL of 10× PCR buffer.
3. Add 1 μL of each primer.
4. Add 0.4 μL 10 mM dNTPs mix.
5. Add 2 μL 1.5 mM $MgCl_2$ (*see* **Notes 26** and **35**).
6. Add 0.2 μL (1 U) *Taq* DNA polymerase (*see* **Note 35**).
7. Mix gently and centrifuge briefly to collect reactions at the bottom of the tube.
8. Complete to 20 μL with water (*see* **Note 27**).
9. Amplify by means of a touch-down PCR as follows: one cycle of 94 °C denaturation for 30 s, 65 °C annealing for 30 s, and 72 °C extension for 60 s, followed by 12 cycles with the annealing temperature lowered by 0.7 °C per cycle. Complete with 23 further cycles of 94 °C for 30 s, 58 °C for 30 s, and 72 °C for 60 s (*see* **Notes 29**, **33**, and **36**). Set soak temperature to 12 °C (*see* **Note 30**). Keep the samples at the fridge until electrophoresed.

3.4 Preparation of the Glass Plates for Polyacrylamide Gels

Failure to perform this step will result in the gel sticking to both plates, and it will be torn and destroyed during the subsequent separation of the glass plates that needs to be done prior to gel staining. Always wear gloves and use perfectly cleaned dry paper towels or Kimwipe paper in all steps. Change gloves between working with Repel Silane and Bind-Silane. Work at room temperature.

1. Thoroughly clean both glass plates twice with dH_2O and 70 % ethanol. Clean gently with Kimwipes and wait for 2 min. Repeat the operation and execute a final polish with new Kimwipe papers (*see* **Notes 37** and **38**).

3.4.1 Preparation of the Large Glass Plate

1. Choose the largest glass to which the gel will be affixed.
2. Pour the binding solution into the center of the glass and, with a dry paper towel, spread gently using a circular motion over the entire surface.

3. Wait 3 min for the binding solution to dry.
4. Repeat **step 2** using the second tube with the binding solution.
5. Let dry for an additional period of 5 min.
6. Gently clean using Kimwipe paper moistened with 95 % ethanol to remove the excess of binding solution.
7. Let dry for 5 min.
8. Wipe the gel plate 2–3 times with 95 % ethanol, using Kimwipes tissues, to remove the excess binding solution.
9. Let dry, while preparing the small glass plate.

*3.4.2 Preparation of the Small Glass Plate (See **Note 39**)*

1. Pour 1.5 mL of undiluted dimethyldichlorosilane (Repel Silane®) into the center of the glass, and, with a dry paper towel, gently spread using a circular motion over the entire surface.
2. Let dry for 5 min.
3. Gently remove the excess Repel Silane® with a dry Kimwipe tissue.
4. Let dry for additional 10 min.
5. Assemble the glass plates by placing the spacers (*see* **Note 40**) on the sides of the larger glass plates. Carefully (*see* **Note 41**) put the smaller glass plate on top. Use a set of 4–6 clamps or an appropriate casting holder (*see* **Note 42**) to clench the glasses in place. Make sure the bottom of both plates and spacers are perfectly aligned with one another to prevent leaking of the unpolymerized polyacrylamide gel.

3.5 Preparation and Casting of Polyacrylamide Gels

Acrylamide has been found to be cancerigenous and neurotoxic, being necessary to avoid skin contact. Protective gloves and eyewear should be worn while handling these products. Carry out all procedures at room temperature.

1. Prepare the acrylamide gel solution: Denaturant polyacrylamide gel (for AFLP and SAMPL markers). Prepare 60 mL of gel (6 % acrylamide, 7.5 M urea, 1× TBE, 0.7 % APS, 0.05 % TEMED) combining 45.8 mL of 9.6 M urea (*see* **Note 43**) with 6 mL 10× TBE and 7.2 mL acrylamide solution 50 % (Long Ranger®) in a 100 mL beaker, mixing by gentle stirring (*see* **Note 44**). Add 450 μL of 10 % ammonium persulfate (APS) to the mixture followed by 32 μL of TEMED. Mix quickly but gently (*see* **Notes 45** and **46**).
2. Prepare the acrylamide gel solution: Non-denaturant polyacrylamide gel (for ISSR markers): For a 60 mL gel (6 % acrylamide, 3 M urea, 1× TBE; 0.7 % APS,0.05 % TEMED), combine 18.72 mL urea 9.6 M (*see* **Note 43**), 6 mL 10× TBE

and 7.2 mL acrylamide 50 % solution (Long Ranger®) and mix in a beaker with gentle stirring. Make up the 60 mL volume with water (*see* **Note 44**). Add 450 μL of 10 % ammonium persulfate (APS) and 32 μL of TEMED to the solution and mix quickly but gently (*see* **Notes 45** and **46**).
Immediately, cast the prepared polyacrylamide gel using a sequencing apparatus (*see* **Note 47**).

3. Using a glass or a sterile plastic pipette, carefully pour the gel solution between the assembled glass plates (*see* **Note 48**). Keep the assembled plates with ca 40 °C angle in relation to both vertical and horizontal planes and always pour the solution at one side with a constant flow to prevent bubble formation (*see* **Note 49**). If any bubbles are noticeable, gently tap the glass plate or move the assembled cast to remove them.
4. Once the cast is filled up with the gel solution, insert the comb(s) into the gel with the teeth facing up (*see* **Note 50**). Start inserting the comb(s) by the edge of the plate. Clamp with three clips and keep it at a 5° angle relative to the surface while the gel polymerizes (*see* **Notes 51** and **52**).
5. After the acrylamide has polymerized, remove the clamps holding the comb(s) and casting stand.
6. Pull out the comb(s) straight by wriggling it gently and smoothly.
7. Remove the glass holding clips or casting clamp.
8. Place the stand with the smaller glass facing back, into the apparatus tank, touching the base of the lower buffer reservoir.
9. Fill the upper and lower reservoirs with 1× TBE buffer (*see* **Note 53**).
10. Mark the level of the buffer in the upper chamber with a pen marker for subsequent check for possible leakage (*see* **Note 54**).
11. Gently flush the wells thoroughly with running buffer using a discardable Pasteur pipette (*see* **Note 55**).
12. Gently insert the shark toothcomb between the glass plates with teeth facing downwards.
13. Fix the safety cover on top on the upper buffer chamber to prevent evaporation of buffer and pre-run the gel at constant power (e.g., ca. 55 W for a Model S2 Apparatus) for ca. 20 min.
14. For electrophoresis under denaturing conditions (to resolve AFLP and SAMPL products), denature the PCR products (5 μL each sample); prepare two tubes with molecular weight marker (1 μL of 10 bp ladder) mixed with equal volume of 2× denaturant loading dye for 3 min at 95 °C. Immediately, transfer the denatured samples to ice to prevent annealing.

15. For electrophoresis under non-denaturing conditions (to resolve ISSR products), mix equal volumes of PCR products and non-denaturing loading dye (3 μL each). Also prepare two tubes with molecular weight marker (1 μL of 100 bp ladder) mixed with equal volume of loading dye.
16. Load the samples into each well. Load 3–5 μL of AFLP/ISSR/SAMPL products. Load the two extreme wells with the appropriate DNA ladder.
17. After loading all the samples and markers, close the lid of the upper buffer chamber.
18. Allow the gel to run at constant power until the xylene cyanol dye reaches ca. 2/3 of the gel (*see* **Notes 56–58**).
19. Remove the plates carefully from apparatus and remove excess buffer by blotting the bottom of the cast on a stack of paper towels.
20. Remove the spacers and separate the plates carefully so that the gel should retain on the smaller glass plate (*see* **Note 59**).
21. Take the larger glass plate with the gel attached and proceed to the gel silver staining protocol to visualize the bands.

3.6 Silver Staining of Gels

Adapted from Bassam et al. [13]. Work in a ventilated fume hood. Use ultrapure deionized water and analytical grade reagents. Since most chemicals are unstable, the solutions must be prepared the day of use. Use a vertical waving shaker. The same staining tray can be used in all steps. As silver staining is a temperature-dependent process, lab temperature should be controlled (to 18–24 °C). Keep all silver staining solutions protected from light. Carry out all procedures, including gel agitation, in a fume hood and wear gloves at all stages. The solutions recipes are given per 2 L to be directly suited for staining gels of the dimensions used.

1. Before beginning, put the fixing/stopper, the impregnation and the developer solutions at 4 °C (*see* **Note 60**). Keep an additional 1 L of water refrigerated at 4 °C.
2. Fix the gel with 2 L of fixing solution during at least 30 min (*see* **Notes 61** and **62**).
3. Rinse with 2 L deionized water 3 times × 2 min.
4. Incubate gel for 20 min in 4 °C cold silver impregnation solution for 30 min in the dark (*see* **Note 63**). Keep the gel in the same tray during all subsequent steps.
5. Rinse once with dH_2O for no more than 15 s (*see* **Note 64**).
6. Develop by soaking the gel with developer solution until the bands are revealed (typically 2–5 min; *see* **Note 65**).
7. Stop the reaction by dispersing 1 L of 10 % acetic acid in the developer solution (stopper solution, reserved from **step 2**) for at least 5 min.

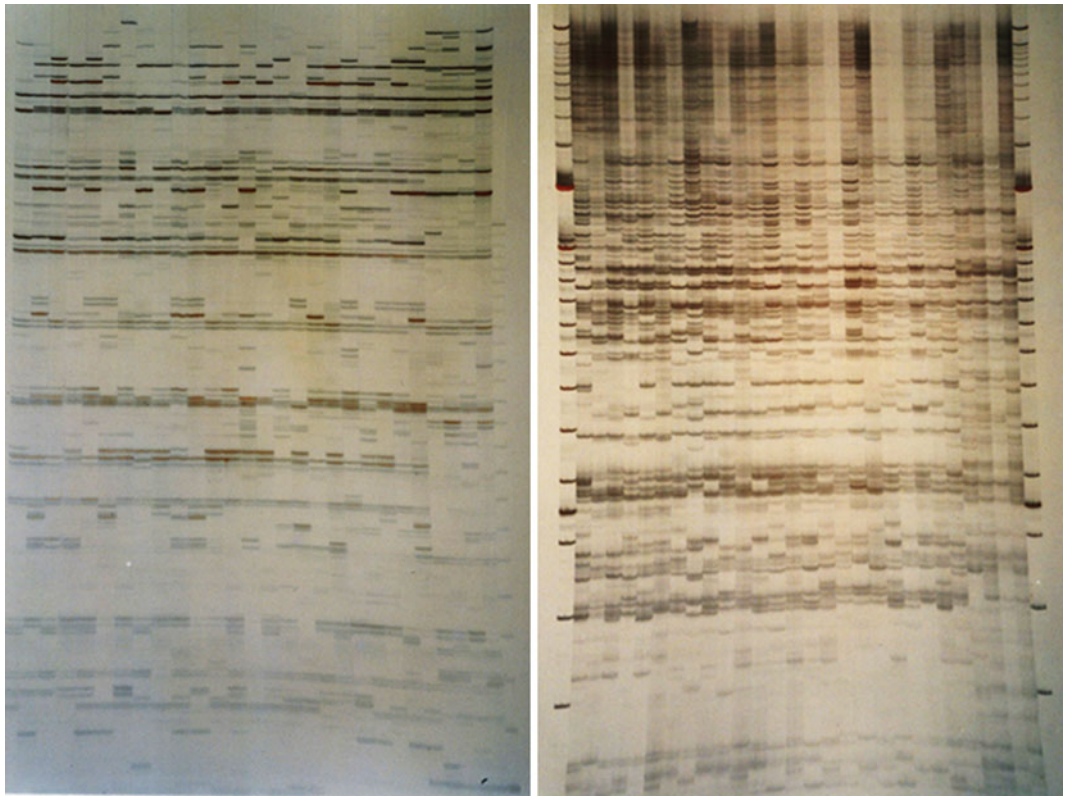


Fig. 1 Example of AFLP (*left*) and ISSR (*right*) amplification profiles generated from 28 plum cultivars using primer combination M-CTT/E-ACT for AFLP and HVH(TG)7 for ISSR, respectively (from Goulão et al. [12]). Both gels were obtained following the experimental protocols described

8. Wash with deionized water for 10 min and let dry on the bench in a near vertical position.
9. Digitally scan the gel for permanent record (Fig. 1).

4 Notes

1. All reagents required for AFLP analysis can be alternatively obtained from Invitrogen Life Technologies (Paisley PA4 9RF, UK) as kits, except for *Taq* DNA polymerase (available from many commercial manufactures).
2. The success of the AFLP technique is dependent upon complete restriction digestion of genomic DNA; therefore, much care should be taken to isolate high quality, intact DNA, without contaminating nucleases or inhibitors. DNA should be dissolved in ultrapure MilliQ water or TE pH 8.0 buffer.
3. To prepare the gel, weigh 0.5 g agarose into an Erlenmeyer flask, add 50 mL 1× TAE buffer, and heat to dissolve in a microwave oven. Let it cool to ca. 50 °C, add 1.25 μL of

GreenSafe, pour into the gel cast, and then place the well combs. Let it polymerize (ca. 20–40 min). Typically, load a total of 10 μL sample into the gel (mix 2 μL of loading dye ((0.0025 % bromophenol blue; 30 % (v/v) glycerol) with 8 μL sample)). Let the gel run at 5–8 V/cm, until the dye reaches the end of the gel, and examine the gel on a UV transilluminator. Ethidium bromide can be used as a cheaper substitute of the GreenSafe dye. Note that ethidium bromide is a powerful mutagen and should be handled carefully, always with gloves.

4. Selecting a 4 bp cutter and 6 bp cutter enzymes produces small DNA fragments in the optimal size range (50 bp–1 kb) to be amplified and separated on denaturing polyacrylamide gels. *Eco*RI and *Mse*I enzymes are used in typical AFLP procedures. Due to primer design and amplification strategy, *Eco*RI–*Mse*I fragments are preferentially amplified, rather than *Eco*RI–*Eco*RI or *Mse*I–*Mse*I fragments. Other enzymes can be selected but, preferably, they should not be methylation sensitive. Make sure that the chosen enzymes are active in the same reaction buffer.
5. The adapters come as single strands, so the two strands of each adapter must be annealed to each other before they can be used. The easiest way is to use a thermal cycler programmed as follows: 94 °C for 90 s, 65 °C for 10 min, 37 °C for 10 min, 25 °C for 10 min, and 4 °C for 10 min.
6. Different concentration of adapters is used since rare-cut enzyme produces less restricted ends than the frequent-cut one.
7. In our conditions, although most of the current *Taq* DNA polymerases in the market can be used, some brands fail to produce satisfactory results.
8. The disodium salt of EDTA will not go into solution until the pH of the solution is adjusted to ca. 8.0 by the addition of NaOH. For tetrasodium EDTA, use 226.1 g of EDTA and use HCl to adjust pH.
9. Di-, tri-, or tetra-repeats can be used to design primers. Anchored primers are recommended to avoid floating of the primer in the satellite sequence. Primers anchored in the 3′ end will produce polymorphisms considering both the length of the satellite and the distance between satellites. Examples of ISSR primers are (5′–3′) as follows: $(GA)_8YG$, $(AG)_8YC$, $(AG)_8YT$, $(CA)_8R$, $(AGC)_4YT$, $(AGC)_4YR$, $VHV(GT)_7$, $VHV(TG)_7$, $HVH(CA)_7$, $HVH(TG)_7$, $DBD(AC)_7$, and $(AGC)_4YR$ (Y = pyrimidine, B = every base except A, D = every base except C, H = every base except G, V = every base except T).
10. Measure the two solutions separately to avoid errors due to volume contraction.
11. Undissolved white clumps may be formed and can be dissolved by placing the bottle in a warm (ca. 65 °C) water bath. pH adjustment is not necessary. Sterilization is not needed for

general use, but for longtime storage, autoclave and filter through a 0.45 μM membrane are recommended. Store the bottle of 10× buffer solution at room temperature.

12. This solution should be prepared with gentle warming (37 °C) and vigorous stirring until the urea is totally dissolved. Add the urea to the water in small doses to help dissolving. Store at room temperature protected from light (wrap the bottle with aluminum foil or use a dark bottle).
13. Always prepare fresh (immediately before use) and store at 4 °C until use.
14. The most convenient way is to dip a spatula into the xylene cyanol flask and then dip it again in the solution. Traditional loading dyes also have bromophenol blue, but we do not consider being needed, since electrophoresis is finished after this dye runs out of the gel.
15. We have found poor staining usually results from low-quality or old reagents.
16. Due to the exothermic character of the reaction, always add acid to the water and never add water to the acid.
17. Make a fresh stock solution every month.
18. Formaldehyde should be added to the developer at most 1 h before use, and the stock solution should be stored at room temperature since cold storage will inactivate it by polymerization.
19. The developer solution should be used at about 8 °C. This is conveniently done by preparing 1 L of the solution with the double concentration needed (2×) and storage at room temperature. Then, immediately before use, complete with 1 L of cold water (keep always a bottle stored at 4 °C). Alternatively, the solution can be prepared at 1× concentration and incubated on ice for about 15 min prior to use.
20. Make sure the water is swirling when the sodium carbonate is added or it will clump and take much longer to dissolve.
21. Silver reacts with nucleic acids and proteins leading to high background staining and spots in the gels.
22. Longer incubations (e.g., overnight) are acceptable, but care must be taken to prevent star activity, particularly due to high (>5 % v/v) glycerol concentrations which may occur due to evaporation during overnight incubations.
23. Make sure to completely thaw and mix the ligation buffer since it contains ATP, which is rather unstable. Furthermore, the stock tube of this reagent should be aliquoted it into small tubes to reduce freeze/thaw cycles.
24. Not needed if using adaptor solution as suggested and DNA ligase at 1 U/μL concentration.

25. Store the unused portion of the reaction mixture at –20 °C to be used in future works.
26. Commonly it can be omitted if the reaction buffer already contains magnesium.
27. Add 1–2 drops of mineral oil if using a thermal cycler without "heated lid" option.
28. Do not denature samples prior to PCR because it reduces the annealing efficiency of the primers.
29. Minor optimizations may be needed according to different polymerase brands and thermal cycler machines. The conditions reported were optimized using a Biometra UNO II (Biometra, Göttingen, Germany) thermal cycler.
30. The typical soaking temperature is 4 °C, but keeping the thermal cycler at 12 °C saves energy and is not detrimental to the samples.
31. Expected result is a smear in the 100–1,000 bp range. Sometimes bands are visible through the smear. UV radiation exposure can damage the cornea and burn to the skin. Always use protective goggles.
32. Store the unused portion as aliquots, at –20 °C.
33. PCR is started at a very high annealing temperature to obtain optimal primer selectivity. In the following steps the annealing temperature is lowered gradually to a temperature at which efficient primer binding occurs. This temperature is then maintained for the rest of the PCR cycles.
34. The conditions may require optimization. It should be noted that ISSR primers contain repetitive regions and can be more difficult to amplify, so increased concentrations of primer (up to 20 μM) may be necessary. In our hands, 30–50 ng of template DNA results in optimal amplification.
35. It can be needed to determine optimal concentrations of $MgCl_2$ and *Taq* DNA polymerase form different brands of enzymes, or when analyzing DNA from different species.
36. It should be noted that ISSR primers contain repetitive regions and can be more difficult to amplify, requiring optimization.
37. Cleaning with Kimwipes or paper towels must be gentle enough not to remove the bind and repel compounds from the glass plates but strong enough to completely dry the plates (no drops should be visible). Keep polishing until it "squeaks."
38. Before reusing the glass plates, incubate them overnight submerged in a 2 M NaOH solution and then rise abundantly in tap water.
39. The small glass plate is to be separated from the cast without gel adhering to it.

40. Using silver staining detection, best results are obtained with 0.3–0.4 mm spacers and shark-tooth combs.
41. Verify if the glass plates are completely dry (no drops should be visible) and do not allow the treated surfaces to come into contact with one another at any time.
42. Several gel apparatus offer convenient casting locks (e.g., the S2/S2001) that provide an easy way to cast and seal sequencing gels without the use of tape or clamps, with the additional advantage of setting even pressure over the entire surface. Always check that the inner surface of the casting clamp is clean and free of debris and grease.
43. Urea easily precipitates at low temperatures. Before use, check for total dissolution and, if needed, warm at 37 °C for about 30 min (or until dissolved).
44. Mix gently to prevent gas bubble formation. No degas is needed in our hands. If proved necessary, perform this step via vacuum application for 10 min, using a Kitasato flask.
45. Add the TEMED immediately prior to pouring the gel and work quickly after its addition to complete pouring the gel before the acrylamide polymerizes.
46. Most gel formulations allow only approximately 5 min before starting polymerization. Work quickly to be able to load all the gel solution inside the gel cast.
47. When pouring the gel from the top, make sure the bottom fill port of the casting clamp is sealed to prevent the gel solution to leak from the bottom.
48. Alternatively, use a syringe that will hold the volume of gel solution to be poured and with a nozzle that will fit in between the glass plate sandwich. Pour the acrylamide gel solution into a barrel of the syringe and invert the syringe to expel any trapped air that has entered the barrel. Introduce the nozzle of the syringe into the notched region the plates. Gently but quickly expel the mixed solution from the syringe, filling the space almost to the top. Hold the assembled plate sandwich at a 35–40° angle on one bottom corner so that the gel solution flows evenly down along the lower side spacer. Maintain a constant, even flow to reduce the chance of forming bubbles in the solution.
49. Gently lower a bit of the angle of the glass plates while pouring the gel during casting.
50. This step aims at creating a perfectly flat surface in the top of the gel, permitting the sample to be uniformly loaded. Disturbances in the surface of the gel will result in "waving" of the bands.
51. Takes 2 h to overnight. Better results are obtained with an overnight polymerization. In such conditions, the top of the

gel to be formed should be protected with paper towels wet with 1× TBE (running buffer).

52. A simple way to monitor polymerization is to check for a small amount that was left not casted. Typically, the acrylamide polymerizes completely for 45–120 min at room temperature.
53. The volume of running buffer should be enough to cover the combs inserted to generate the flat surface.
54. Make sure that the glass plates are firmly seated against the inner surface of the casting clamp. Verify that the upper buffer chamber drain valve is in the closed position (down), and fill the upper buffer chamber with running buffer, covering the gel. Make sure that no leakage exists from the upper buffer chamber. Do not start electrophoresis if leakage is observed. If leakage is observed during the gel run, keep adding running buffer to the top chamber of the apparatus to prevent overheating and cracking of the glasses or, if it is not possible, abort electrophoresis immediately.
55. Use a 200 μL micropipette, a discardable Pasteur pipette or a syringe filled with 1× TBE buffer with an attached needle to flush out all the wells. Pipette vigorously up and down several times. Acrylamide gel fragments will prevent homogenous run, leading to distortion of the bands.
56. At 45 W and under our conditions, a complete run takes about 2–2.30 h for AFLP and SAMPL and 2.30–3 h for ISSR.
57. During the electrophoresis, keep monitoring possible buffer leakage and gel temperature. An appropriate indicator placed on to the outer plate near the center of the gel can be used as an option. The temperature should be maintained between 40 and 50 °C.
58. Xylene cyanol co-migrates with ca. 125 bp linear single-stranded DNA of in 6 % denaturing gels and co-migrates with ca. 230 bp linear double-stranded DNA in 6 % non-denaturing gels.
59. The best way to separate the glass plates is to use a pizza wheel. Carefully insert the wheel between the two glasses (use the spacers to create enough room for the wheel at the top contact between the two glass plates) and gently and smoothly wriggle it until the plates separate.
60. Staining is enhanced with cold $AgNO_3$.
61. Fix until no dye is visible on the gel. Overnight fixing is acceptable, if convenient.
62. After fixation, reserve 1 L of the solution to be used in **step 7**.
63. Protect from light by covering the tray with aluminum foil. Impregnation solution can be reused once. When reusing, increase the incubation time to 45 min.

64. Residual $AgNO_3$ on the gel surface and staining tray will increase background staining.
65. The developing time varies. Larger bands (top of the gel) develop first. A certain (but controlled) degree of overstaining in this area of the gel is acceptable in order to visualize the smaller bands.

Acknowledgments

Financially supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal. The authors acknowledge previous coworker Luisa Monte-Corvo for her valuable contributions during different phases of the protocol development.

References

1. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414
2. Zietkiewicz E, Rafalski A, Labuda D (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20:176–183
3. Gupta M, Chyi YS, Romero-Severson J, Owen JL (1994) Amplification of DNA markers from evolutionarily diverse genomes using single primers of SSRs. Theor Appl Genet 89: 998–1006
4. Heath DD, Iwana GK, Delvin RH (1993) PCR primed with VNTR core sequences yield species specific patterns and hypervariable probes. Nucleic Acids Res 21:5782–5785
5. Bell CJ, Ecker JR (1994) Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics 19:137–144
6. Arens P, Odinot P, van Heusden AW, Lindhout P, Vosman B (1995) GATA and GACA repeats are not evenly distributed throughout the tomato genome. Genome 38:84–90
7. Morgante M, Vogel J (1994) Compound microsatellite primers for the detection of genetic polymorphisms. US Patent Appl. no. 08/326456
8. Witsenboer H, Vogel J, Michelmore RW (1997) Identification, genetic localization, and allelic diversity of selectively amplified microsatellite polymorphic loci in lettuce and wild relatives (Lactuca spp.). Genome 40:923–936
9. Paglia G, Morgante M (1998) PCR-based multiplex DNA fingerprinting techniques for the analysis of conifer genomes. Mol Breed 4:173–177
10. Monte-Corvo L, Goulao L, Oliveira CM (2001) ISSR analysis of cultivars of pear and the suitability of molecular markers for clone discrimination. J Am Soc Hortic Sci 126: 517–522
11. Goulao L, Oliveira CM (2001) Molecular characterisation of cultivars of apple (Malus domestica Borkh.) using microsatellite (ISSR and SSR) markers. Euphytica 122:81–89
12. Goulao L, Monte-Corvo L, Oliveira CM (2001) Phenetic characterization of plum cultivars by high multiplex ratio markers: amplified fragment length polymorphisms and inter-simple sequence repeats. J Am Soc Hortic Sci 126:72–77
13. Bassam BJ, Anollés GC, Gresshoff PM (1991) Fast and sensitive silver staining of DNA in polyacrylamide gels. Anal Biochem 196:80–83

Chapter 12

Transposon-Based Tagging: IRAP, REMAP, and iPBS

Ruslan Kalendar and Alan H. Schulman

Abstract

Retrotransposons are a major component of virtually all eukaryotic genomes, which makes them useful as molecular markers. Various molecular marker systems have been developed that exploit the ubiquitous nature of these genetic elements and their property of stable integration into dispersed chromosomal loci that are polymorphic within species. To detect polymorphisms for retrotransposon insertions, marker systems generally rely on PCR amplification between the retrotransposon termini and some component of flanking genomic DNA. The main methods of IRAP, REMAP, RBIP, and SSAP all detect the polymorphic sites at which the retrotransposon DNA is integrated into the genome. Marker systems exploiting these methods can be easily developed and are inexpensively deployed in the absence of extensive genome sequence data. Here, we describe protocols for the IRAP, REMAP, and iPBS techniques, including methods for PCR amplification with a single primer or with two primers, and agarose gel electrophoresis of the product using optimal electrophoresis buffers; we also describe iPBS techniques for the rapid isolation of retrotransposon termini and full-length elements.

Key words Retrotransposon, Molecular marker, IRAP, REMAP, iPBS

1 Introduction

Interspersed repetitive sequences comprise a large fraction of the genome of most eukaryotic organisms, and they are predominantly composed of transposable elements (TEs) [1]. In most species that have been studied, interspersed repeats are distributed unevenly across the nuclear genome, with some repeats having a tendency to cluster around the centromeres or telomeres [2–4]. Following the induction of recombinational processes during meiotic prophase, variation in the copy number of repeat elements and internal rearrangements on both homologous chromosomes can ensue.

Nucleotide sequences matching repetitive sequences showing polymorphism in RFLP analyses have been used as polymerase chain reaction (PCR) primers for the inter-repeat amplification polymorphism marker method [5, 6]. Such repetitive sequences include microsatellites, such as $(CA/GT)_n$ or $(CAC/GTG)_n$, which are distributed throughout the genome. A related approach was

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_12,

developed to generate PCR markers based on amplification of microsatellites near the 3′ end of the Alu (SINE) transposable elements (TEs), called Alu-PCR or SINE-PCR [7]. Successful applications of microsatellite-specific oligonucleotides as PCR primers were first described in the early 1990s [5, 6, 8].

1.1 LTR Retrotransposons

Long terminal repeat (LTR) retrotransposons, or type I transposable elements, replicate by a process of reverse transcription, as do the lentiviruses such as HIV [9]. The retrotransposons themselves encode the proteins needed for their replication and integration back into the genome [10]. Their "copy-and-paste" life cycle means that they are not excised in order to insert a copy elsewhere in the genome. Hence, genomes diversify by the insertion of new copies, but old copies persist. Their abundance in the genome is generally highly correlated with genome size. Large plant genomes contain hundreds of thousands of these elements, together forming the vast majority of the total DNA [11].

Human and other mammalian genomes also contain an abundance of retrotransposons. The majority of these, however, are not LTR retrotransposons but LINEs and SINEs, which replicate by a somewhat different copy-and-paste mechanism [12, 13]. The L1 family of LINE elements and the *Alu* family of SINE elements comprise together roughly 30 % of human genomic DNA and nearly two million copies [14]. Nevertheless, integrated retroviruses, which are remnants of ancient infections, are abundant in mammalian genomes [15]. These elements, called "endogenous retroviruses" (ERVs, HERVs in humans), are functionally equivalent to LTR retrotransposons. The features of integration activity, persistence, dispersion, conserved structure, and sequence motifs and high copy number together suggest that retrotransposons are well-suited genomic features on which to build molecular marker systems [16, 17].

1.2 Retrotransposons as DNA Markers

Retrotransposon-based systems (Fig. 1) detect the insertion of elements hundreds to thousands of nucleotides long, although generally only the insertion joint itself is monitored due to the impracticality of amplifying and resolving long fragments and discriminating their insertion sites. The LTRs that bound a complete retrotransposon contain ends that are highly conserved in a given family of elements. Newly inserted retrotransposons, therefore, form a joint between the conserved LTR ends and flanking, anonymous genomic DNA. Most retrotransposon-based marker systems use PCR to amplify a segment of genomic DNA at this joint. Generally, one primer is designed to match a segment of the LTR conserved with a given family of elements but different in other families. The primer is oriented towards the LTR end. The second primer is designed to match some other feature of the genome. The first retrotransposon method described was SSAP or S-SAP

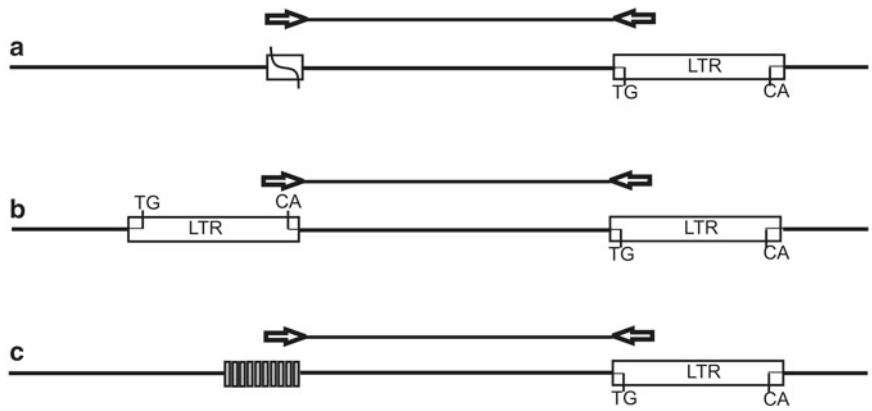


Fig. 1 Retrotransposon-based molecular marker methods. Multiplex products of various lengths from different loci are indicated by the *bars* above or beneath the diagrams of each reaction. Primers are indicated as *arrows.* (**a**) The SSAP method. Primers used for amplification match the adapter (restriction site shown as *empty box*) and retrotransposon (*LTR box*). (**b**) The IRAP method. Amplification takes place between retrotransposons (*left* and *right LTR boxes*) near each other in the genome (*open bar*), using retrotransposon primers. The elements are shown oriented head-to-head using a single primer. (**c**) The REMAP method. Amplification takes place between a microsatellite domain (*vertical bars*) and a retrotransposon, using a primer anchored to the proximal side of the microsatellite and a retrotransposon primer

(sequence-specific amplified polymorphism, *see* Fig. 1a), where one primer matched the end of the *BARE1* retrotransposon of barley and the other matched an AFLP-like restriction site adapter [18].

1.3 Sequence-Specific Amplified Polymorphism

Sequence-specific amplified polymorphism (SSAP) was described by Waugh and coworkers in 1997 [18], but has several origins and forms [19–22]. The SSAP method can be considered to be a modification of AFLP [23] or as a variant of anchored PCR [24]. The method described by Waugh and colleagues [18] has many similarities to AFLP, especially in that two different enzymes are used to generate the template for the specific primer PCR and that selective bases are used in the adapter primer.

In the SSAP procedure, it is important to maximize the sequence complexity of the template for the specific primer amplification, so a single enzyme digestion is used [25]. As with the method described for *BARE1* [18], the adapter primer is selective. This is a matter of convenience, and nonselective primers could be substituted where the enzyme used for digestion has a larger recognition sequence, or if the copy number were lower. In general, LTR ends are convenient for the design of SSAP primers [26]. However, for the *PDR1* retrotransposon in *Pisum*, the LTR is exceptionally short at 156 bp, so a GC-rich primer could be designed corresponding to the polypurine tract (PPT) which is found internal to the 3′ LTR in retrotransposons. For *BARE1* in barley and other high-copy-number families, the number of selective bases may be increased compared to the first version of the

protocol [18, 27]. Furthermore, *BARE1* and most other retrotransposons have long LTRs, necessitating an anchor primer in the LTR near to the external terminus. The main feature of the SSAP procedure that may be modified for various situations is the location of the sequence-specific primer [28]. The choice of this primer is critical and can be modified according to need. For example, internal primer sites have been exploited to describe structural variation within retrotransposons [29], and the primers can be applied to defined sequences other than the LTR or PPT.

1.4 Inter-retrotransposon Amplification Polymorphism (IRAP) and Retrotransposon-Microsatellite Amplification Polymorphism (REMAP)

The IRAP (Fig. 1b) and REMAP (Fig. 1c) methods represent a departure from SSAP, because no restriction enzyme digestion or ligation step is needed and because the products can be resolved by conventional agarose gel electrophoresis without resort to a sequencing apparatus. The IRAP method detects retrotransposon insertional polymorphisms by amplifying the portion of DNA between two retroelements. It uses one or two primers pointing outwards from an LTR and therefore amplifies the tract of DNA between two nearby retrotransposons. IRAP can be carried out with a single primer matching either the 5′ or 3′ end of the LTR but oriented away from the LTR itself or with two or more primers. The two primers may be from the same retrotransposon element family or may be from different families. The PCR products, and therefore the fingerprint patterns, result from amplification of hundreds to thousands of target sites in the genome.

The complexity of the pattern obtained will be influenced by the retrotransposon copy number, which mirrors genome size, as well as by their insertion pattern and by the size of the retrotransposon families chosen for analysis. Furthermore, thousands of products can neither be simultaneously amplified to detectable levels nor resolved on a gel system. Hence, the pattern obtained represents the result of competition between the targets and products in the reaction. As a result, the products obtained with two primers do not represent the simple sum of the products obtained with the primers individually.

If retrotransposons were fully dispersed within the genome, IRAP will either produce products too large to give good resolution on gels or target amplification sites too far apart to produce products with the available thermostable polymerases. This is because retrotransposons generally tend to cluster together in "repeat seas" surrounding "gene islands" and may even nest within each other. For example, the *BARE1* retrotransposon of barley, an abundant superfamily *Copia* element, is present as about 13,000 full-length copies of about 8.9 kb and 90,000 solo LTRs of 1.8 kb in the cultivar Bomi. Given a genome of roughly 5×10^9 bp, these elements comprise 5.6 % of the genome but would occur only about once every 46 kb if they were fully interspersed. Nevertheless, IRAP with *BARE1* primers displays a range of products from 100 bp upwards of 10 kb (Fig. 2).

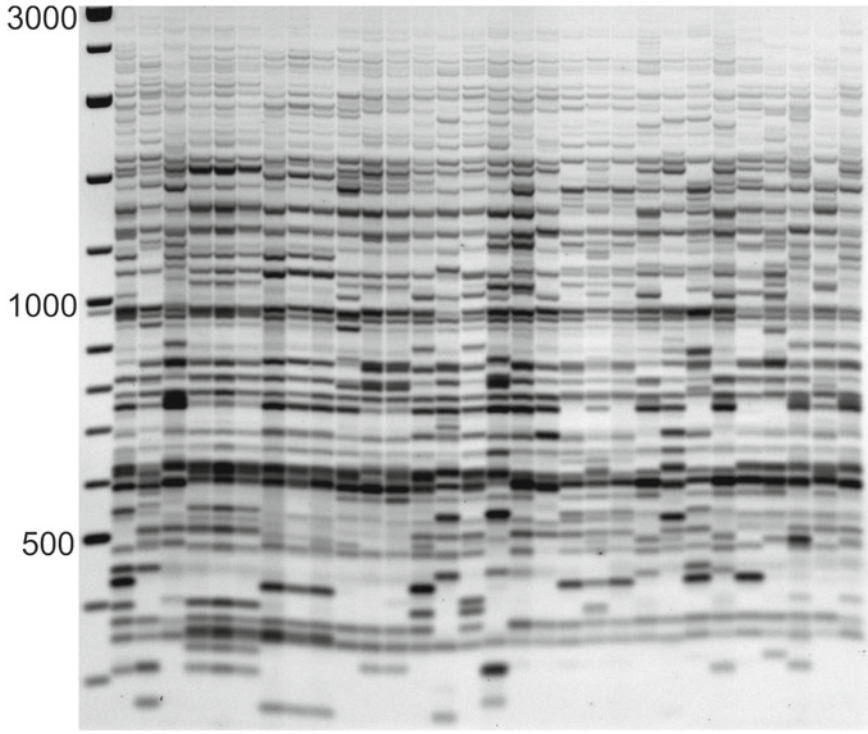


Fig. 2 Utility of IRAP for a diversity analysis of plant species. The phenogram of 30 genotypes of populations of *H. spontaneum* based on IRAP analysis is shown as negative images of ethidium bromide-stained agarose gels following electrophoresis. Results for *BARE1* LTR primer 1369 are shown. A 100 bp DNA ladder is present on the *left*

The REMAP method is similar to IRAP, except that one of the two primers matches an SSR motif with one or more non-SSR anchor nucleotides present at the 3′ end of the primer. Microsatellites of the form $(NN)_n$, $(NNN)_n$, or $(NNNN)_n$ are found throughout plant and animal genomes. In cereals, they furthermore appear to be associated with retrotransposons [30]. Differences in the number of SSR units in a microsatellite are generally detected using primers designed to unique sequences flanking microsatellites. Alternatively, the stretches of the genome present between two microsatellites may be amplified by ISSR [6, 8], in a way akin to IRAP. In REMAP, anchor nucleotides are used at the 3′ end of the SSR primer both to avoid slippage of the primer within the SSR, which would produce a "stutter" pattern in the fingerprint, and to avoid detection of variation in repeat numbers within the SSR. REMAP uses primer types that are shared by IRAP and ISSR. Although it would appear that the SSR primers in REMAP should also yield ISSR products and the LTR primers also IRAP products, in practice this is rarely the case. This is probably due to a combination of factors including both genome structure and competition within the PCRs.

1.5 Inter-primer Binding Site Polymorphism (iPBS)

A major disadvantage of all retrotransposon-based molecular marker techniques is the need for sequence information to design element-specific primers. The primary requirement is the sequence of an LTR end, harvested either from a database or produced by cloning and sequencing the genomic DNA that flanks conserved segments of retrotransposons. Although rapid retrotransposon isolation methods based on PCR with conserved primers for TE have been designed, it maybe still necessary to clone and sequence hundreds of clones to obtain a few good primer sequences. The LTRs contain no conserved motifs, which would allow their direct amplification by PCR.

Several restriction and adaptor-based methods for LTR cloning have been developed, which are based on the conservation of reverse transcriptase domain, especially for the superfamily *Copia* retrotransposons [28]. In general, however, all reverse-transcribing elements, including LTR retrotransposons of superfamily *Gypsy* as well as LINE retrotransposons, can be obtained by PCR with degenerate primers. For example, for *Copia*, two degenerate primers were designed for the RT motifs encoding TAFLHG and, for the reverse primer, the downstream YVDDML, as well as for QMDVKT and YVDDML respectively, in order to amplify the family-specific domain in between [31–33]. For *Gypsy* elements, degenerate primers were designed for the RT motifs encoding RMCVDYR, LSGYHQI, or YPLPRID and for the reverse primers for the domains YAKLSKC or LSGYHQI. The RT-based isolation method is limited, of course, to the families of retrotransposons that contain RT and the chosen domains. Thus, for example, non-autonomous groups such as TRIMs, LARDs, and SINEs cannot be found using this approach [34, 35].

The LTR retrotransposons and all retroviruses contain a conserved binding site for tRNA. Generally $tRNAi^{Met}$ is the most common, but also $tRNA^{Lys}$, $tRNA^{Pro}$, $tRNA^{Trp}$, $tRNA^{Asn}$, $tRNA^{Ser}$, $tRNA^{Arg}$, $tRNA^{Phe}$, $tRNA^{Leu}$, and $tRNA^{Gln}$ can be found. Elongation from the 3′-terminal nucleotides of the respective tRNA results in the conversion of the retroviral or retrotransposon RNA genome to double-stranded DNA prior to its integration into the host DNA. While the process of reverse transcription is conserved among virtually all retroelements, the specific tRNA capture varies for different retroviruses and retroelements. The primer binding sequences (PBS) are almost universally present in all LTR-retrotransposon sequences. Hence, an isolation method for retrotransposon LTRs, which is based on the PBS sequence, has the potential for cloning all possible LTR retrotransposons.

The inter-PBS amplification (iPBS) technique has led to the development of a virtually universal and exceedingly efficient method, which utilizes the conserved parts of PBS sequences, both for direct visualization of polymorphism between individuals, polymorphism in transcription profiles, and fast cloning of LTR segments from genomic DNA, as well as for database searches of LTR retrotransposons (Fig. 3). Many retrotransposons are nested, recombined, inverted, or truncated, yet can be easily amplified

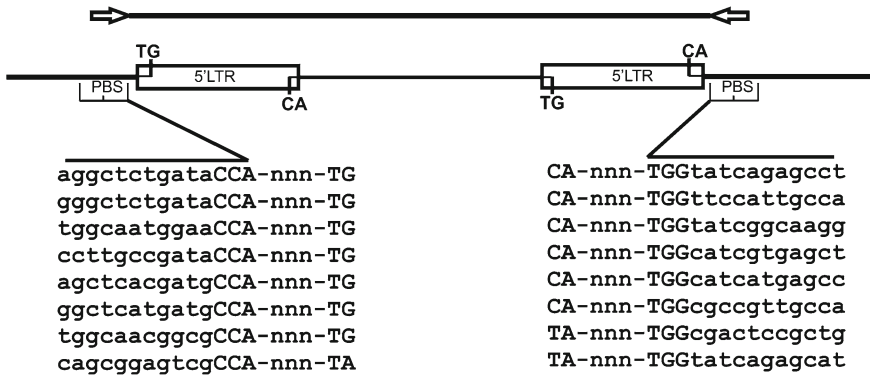


Fig. 3 The inter-PBS amplification (iPBS) scheme and LTR retrotransposon structure. Two nested LTR retrotransposons in inverted orientation amplified from single primer or two different primers from primer binding sites. PCR product contains both LTRs and PBS sequences as PCR primers in the termini. In the figure, general structure for PBS and LTR sequences and several nucleotides long spacer between 5′LTR (5′-CA) and PBS (5′-TGG3′) are schematically shown

using conserved PBS primers in all plant species tested. Fragments of retrotransposons containing a 5′ LTR and part of the internal domain are often located near other entire or similarly truncated retrotransposons. Therefore, PBS sequences are very often located sufficiently near to each other to allow amplification. This situation allows the use of PBS sequences for cloning LTRs. Where the retrotransposon density is high within a genome, PBS sequences can be exploited for detection of their chance association with other retrotransposons. When retrotransposon activity or recombination has led to new genome integration sites, the iPBS method can be used to distinguish reproductively isolated plant lines. In this case, amplified bands derived from a new insertion event or from recombination will be polymorphic, appearing only in plant lines in which the insertions or recombination have taken place.

The PBS primer(s) can amplify nested inverted retrotransposons or related elements' sequences dispersed throughout genomic DNA. The PCR amplification occurs in this case between two nested elements' PBS domains and produces fragments containing the insertion junction between the two nested LTRs. After retrieving LTR sequences of a selected family of retrotransposons, an alignment is made of them to find the most conserved region [36]. The related plant species have conserved regions in LTR for members of the same retrotransposon family. Thus, alignments of several LTR sequences from several species will identify these conserved regions. Subsequently, these conserved domains of LTRs can be used for inverted primers designed for long distance PCR, for cloning of whole retrotransposons, and also for the IRAP, REMAP, or SSAP marker techniques. The iPBS amplification technique shows about the same level of polymorphism in comparison with IRAP and REMAP, and it is an efficient method for the detection of cDNA polymorphism and clonal differences resulting from retrotransposon activities or recombination.

2 Materials

2.1 Reagents

Prepare all solutions using Milli-Q or equivalent ultrapure water and analytical grade reagents.

1. TE buffer (10×): 100 mM Tris–HCl (pH 8.0), 10 mM EDTA. DNA and primers should be stored in a 1× TE solution.
2. Electrophoresis buffer (10× THE): 200 mM Tris-HEPES (pH 8.06), 5 mM EDTA. Weigh 24.2 g Tris-base and 47.7 g HEPES (free acid), add 10 mL 0.5 M EDTA (pH 8.0), dissolve in water; bring final volume to 1 L. Store at +4 °C. While we get best results with 1× THE, standard 1× TBE (50 mM Tris-H_3BO_3, pH 8.8), 1× TAE (40 mM Tris-CH_3COOH, pH 8.0), or 1× TPE (40 mM Tris-H_3PO_4, pH 8.0) buffers may also be used.
3. Gel loading buffer (10×): 20 % (w/w) Polysucrose 400, 100 mM Tris–HCl (pH 8.0), 10 mM EDTA, ~0.01 % (w/w) Orange G, and ~0.01 % Xylene Cyanol FF. Dissolve 20 g Polysucrose 400 (Ficoll 400) in 80 mL 10× TE buffer. Add Orange G and Xylene Cyanol FF according on the desired color intensity. Store at +4 °C.
4. Thermostable polymerase: many types and sources of recombinant thermostable polymerases are effective. Most preferable for PCR use are the recombinant polymerases with 3′–5′ exonuclease proofreading activity that permit a "hot start," such as Phire® Hot Start II DNA Polymerase (Thermo Scientific) from *Pyrococcus furiosus*. Another excellent choice is *Thermus thermophilus* Biotools DNA polymerases (Biotools S.A., Madrid, Spain). Any *Thermus aquaticus* (*Taq*) DNA polymerase is applicable, however. We have tested several *Taq* DNA polymerases, including those of DreamTaq™ (Thermo Scientific), FIREPol® (Solis BioDyne), MasterAmp™ (Epicentre), and GoTaq® (Promega). Other thermostable polymerases, such as that from *Thermus brockianus* (DyNAzyme™ II, Thermo Scientific), was also tested to determine if the choice of polymerase enzyme had an effect on the products amplified. A polymerases mix consisting of 100 U of *Taq* DNA polymerase and 0.5 U *Pfu* DNA polymerase improves amplification of long bands and the accuracy of the PCR. Long distance PCR is performed with DyNAzyme™ EXT (Thermo Scientific) or Phusion® High-Fidelity (Thermo Scientific) or LongAmp™ *Taq* DNA polymerases (New England Biolabs).
5. PCR buffers (1×): several PCR buffers for *Taq* polymerase are suitable for PCR: Buffer 1: 20 mM Tris–HCl (pH 8.8), 2 mM $MgSO_4$, 10 mM KCl, 10 mM $(NH_4)_2SO_4$; Buffer 2: 10 mM Tris–HCl (pH 8.8), 2 mM $MgCl_2$, 50 mM KCl, 0.1 % Triton X-100; Buffer 3: 50 mM Tris–HCl (pH 9.0), 2 mM $MgCl_2$,

15 mM $(NH_4)_2SO_4$, 0.1 % Triton X-100. The PCR and its efficiency depend on which buffer and enzyme combination is used (*see* **Note 1**).

6. Ethidium bromide solution in water, 0.5 mg/mL. Store at room temperature.
7. SYBR Green I gel staining solution (50×) in 50 % dimethyl sulfoxide (DMSO) in Milli-Q water. The 10,000× concentrate is diluted to 50× with 50 % DMSO in Milli-Q water. Store at –20 °C.
8. DNA ladder for electrophoresis GeneRuler™ DNA Ladder Mix (Thermo Scientific), 100–10,000 base range, or similar. DNA ladder diluted with 1× gel loading buffer to final concentration 25 ng/μL.
9. Agaroses: RESolute Wide Range (BIOzym), D1 low EEO (Conda), Premium (Serva), MP (AppliChem). Agaroses of the LE type are not effective for fine resolution of fingerprinting bands. Cambrex NuSieve 3:1 and MetaPhor agaroses have low gel strength and are uncomfortable for gel manipulation; 1 % agaroses with gel strength >1,700 g/cm^2 can be used. Electrophoresis gels of enhanced selectivity can be produced by adding a performed polymer (additive) to a polymerization solution.
10. 10× FastDigest® buffer for restriction DNA with FastDigest® restriction enzymes: *MseI*, *PstI*, *TaqI*, and *TaiI*.
11. MinElute PCR Purification (Qiagen) or similar kit, for PCR products purified.
12. QIAEX II Gel Extraction Kit (Qiagen) or similar kit, for silica-membrane-based purification of DNA fragments from 40 bp to 50 kb from gel or enzymatic reactions.
13. PCR product TA cloning kits, TOPO® TA Cloning® Kit (pCR®2.1 plasmid vector) with TOP10 *E. coli* (Invitrogen) or alternative kits. PCR products should be amplified with Taq, Biotools, or DyNAzyme II DNA polymerases and have single 3′ adenine overhangs. Primers should use a hydroxyl group at 5′ termini. The alternative to TA cloning, BigEasy® Long PCR Cloning Kits (Lucigen) can be effectively used for GC cloning of difficult and long PCR fragments.
14. Competent *E. coli* cells (10^9 cfu/μg), One Shot® TOP10 (Invitrogen), JM109 (Promega).

2.2 Equipment

1. Thermal cycler for 0.2 mL tubes or plates (96 well), with a rapid heating and cooling capacity between 4 and 99 °C, so that the temperature can be changed by 3–5 °C/s, for example, the Mastercycler Gradient (Eppendorf AG) or the PTC-100 Programmable Thermal Controller (Bio-Rad Laboratories).
2. Power supply (minimum 300 V, 400 mA) for electrophoresis.

3. Horizontal electrophoresis apparatus without special cooling. Most commercially available medium- or large-scale horizontal DNA gel electrophoresis systems are suitable, for example, from such suppliers as GE Healthcare, Hoefer, or Bio-Rad. These include the GNA-200, Hoefer HE 99X Max Submarine, BioExpress (Wide Maxi Horizontal Gel System (E-4123-1)), and Sigma (Maxi-Plus). Small electrophoresis boxes and short gel trays are not suitable due to the large number of PCR products that need to be resolved. We routinely employ an apparatus with a run length of 20 cm.
4. Gel comb, a 36 or more well comb, 1 mm thickness, forming 3–4 mm wide wells, with a 1 mm well spacing. This comb is ideal for analysis of any PCR amplification product or DNA restriction enzyme digest. The small space between the slots is important for analysis of banding patterns and for comparing lanes across the gel. Also this thickness of comb improves band resolution.
5. UV transilluminator, for visualization of ethidium bromide-stained or SYBR green-stained nucleic acids, with a viewing area of 20×20 cm.
6. Dark Reader (Clare Chemical Research), for visualization and isolation of SYBR green-stained nucleic acids.
7. Imaging system. A digital gel electrophoresis scanner for detection of ethidium bromide-stained nucleic acids by fluorescence (532 nm green laser) or SYBR green-stained nucleic acids (473 nm blue laser), with a resolution of 50–100 μm. Examples include the FLA-5100 imaging system (Fuji Photo Film GmbH., Germany). Software such as the Aida Image Analyzer and Adobe Photoshop is required for image analysis and manipulation.

3 Methods

3.1 Primer Design

PCR primers are designed to match an LTR sequence near to either its 5′ or 3′ end, with the primer oriented so that the amplification direction is towards the nearest end of the LTR. Generally it is best to base the design on a sequence alignment for representative LTRs from a particular family of elements and to place the primer within the most conserved region for that family. For long LTRs, it is often useful to test primers at several locations within the LTR or internal part of the retrotransposon and in both orientations, particularly if there is evidence for nested insertions in the genome. Primers can be placed directly at the end of the LTR facing outwards, provided that they do not form dimers or loops. For primers placed at the edge of the LTR, one or more additional selective bases can be added at 3′ end in order to reduce the number of amplification targets (Fig. 4). This can be tried in a second round of primer design, if the initial primer yields amplification

TGGGAAGTAGTAAGTAGATGA**TGGGTTGCTCGAGTGACAGAAGC**TTAAACCCCAGTTTATGTGTTG**CTTCGTGAGGGGCTGATTTGGATC**CA
TGGGA-GTAGTAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGG**TTAAACCCCGGTCTATGCGTTG**CTTCGTAAGGGGCTGATTTGGATC**CA
TGGGAAGTAGTAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTAAGGGGCTGATTTGGACC**CA
TGGGAAGTAGTAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTAAGAGGCTGATTTGGATA**CA
TGGGAAGTACTAAGTAGATGA**TGGGTTGCTAGAGTAACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTAAGGGGCTGATTTGGATC**CA
TGGGAAGTCATAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTAAGGGGCTGATTTGGATC**CA
TGGGAAGTAGTAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCATAAGGGGCTGAATTGGATC**CA
TGGGAAGTAATAAGTAGATGA**TGGGTTGCTAGAGTGACAGGAGC**TTAAACCCTAGTTTATGCGTTT**CTTCGTAAGGGGCTGATTTGGATC**CA
TGGGAAGTAGTAAGTAGATGA**TTGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTAAGGGGCTAATTTGGATC**CA
TAGGAAGTAATAAGTAGATTA**TGGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTAAGGGGCTGATTTGGATC**CA
TGGGAAGTAATAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGC**ATAAACCCTAGTTTATGTGTTG**CTTCGTAAGGGGCTGATTTGGATC**CA
TGGGAAGTAATAAGTAGATGA**TGGGTTGCTAGAGTGACATAAGC**TTAAACCCTAGTTTATGCGTTC**CTTTGTTAGGGGCTGATTTGGATC**CA
TGGGAAGTAATAAGTAGATGA**TGGGTTGCTAGAGTGACACAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCATAAGGGGCTGATTTGGATC**CA
TGGGAAGTAATAAGTAGATGA**TGGGTTGCTAGAGTGACACAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCATAAGGGGCTGATTTGGATC**CA
TGGGAATTAATAAGTAGATGA**TGGGTTGCTAGAGTGACAGAAGC**TTAAACCCTAGTTTATGCGTTG**CTTCGTTAGGGGCTGATTTGGATC**CA
GGGGAAGTAATAAGTAGATGA**TGGGTTTCTAGAGTGACAAAAGC**TTAAACCCTAGTTTATGTGTTG**CTTCATTAGGGGCTGATATGGATC**CA

5′-**TGGGTTGCTAGAGTGACAGAAGC** 5'-**CTTCGTAAGGGGCTGATTTGGATC**

Fig. 4 Example of LTR primer design: conserved 3′ end of *Cereba* LTRs and matching primers

Table 1
Microsatellite primers

Microsatellites sequence	Primers sequences (5′–3′)	T_m (°C)[a]
Two-base-repeat microsatellites		
$(CT)_n$	(CT)10G; (CT)10T; (CT)10A	52.1; 51.5; 50.3
$(CA)_n$	(CA)10G; (CA)10T; (CA)10A	56.6; 56.2; 55.9
$(TC)_n$	(TC)10G; (TC)10C; (TC)10A	53.6; 53.5; 52.4
$(TG)_n$	(TG)10G; (TG)10C; (TG)10A	58.3; 58.8; 57.0
$(AG)_n$	(AG)10G; (AG)10C; (AG)10T	53.4; 54.0; 52.9
$(AC)_n$	(AC)10G; (AC)10C; (AC)10T	58.4; 58.3; 57.6
$(GA)_n$	(GA)10T; (GA)10C; (GA)10A	51.3; 52.0; 51.1
$(GT)_n$	(GT)10T; (GT)10C; (GT)10A	56.4; 56.6; 55.1
Three-base-repeat microsatellites		
$(CTC)_n$	(CTC)6G; (CTC)6T; (CTC)6A	58.0; 57.0; 56.7
$(GTG)_n$	(GTG)6C; (GTG)6T; (GTG)6A	61.7; 60.6; 59.7
$(CAC)_n$	(CAC)6G; (CAC)6T; (CAC)6A	61.2; 60.4; 60.0
$(ACC)_n$	(ACC)6G; (ACC)6T; (ACC)6C	63.0; 62.2; 63.0
$(TCG)_n$	(TCG)6G; (TCG)6C; (TCG)6A	61.3; 61.7; 59.9

[a]Oligonucleotide concentration is 200 nM

products containing too many weak individual bands for analysis by gel electrophoresis. If the LTRs are short (<300 bp), the primers may also be designed to match internal regions, but this will concomitantly increase the size of the amplified products.

Database searches can sometimes be used to find unannotated, native LTR sequences matching characterized retrotransposons from other species. However, care should be taken in defining the ends of the LTRs. Generally, mapping of the RT primer binding sites PBS and PPT is needed in order to define the LTR ends with confidence. Microsatellite primers for REMAP or ISSR should be

Table 2
Retrotransposon LTR primers

Name	Sequence	TE, source	T_m (°C)[a]	Optimal annealing T_a (°C)
560	TTGCCTCTAGGGCATATTTCCAACA	*Wis2*, LTR	58.2	58–65
554	CCAACTAGAGGCTTGCTAGGGAC		58.8	60–68
2105	ACTCCATAGATGGATCTTGGTGA		54.9	55–61
2106	TAATTTCTGCAACGTTCCCCAACA		57.3	58–65
2107	AGCATGATGCAAAATGGACGTATCA	*Wilma*, LTR	57.2	58–65
833	TGATCCCCTACACTTGTGGGTCA		59.5	60–68
2108	AGAGCCTTCTGCTCCTCGTTGGGT		64.2	64–72
516	TCCTCGTTGGGATCGACACTCC		60.5	60–68
2109	TACCCCTACTTTAGTACACCGACA	*Daniela*, LTR	56.3	57–62
2110	TCGCTGCGACTGCCCGTGCACA		67.8	68–72
2111	CAGGAGTAGGGTTTTACGCATCC		57.2	58–65
2112	TGCTGCGACTGCCCGTGCACA		66.5	66–72
2113	TACGCATCCGTGCGGCCCGAAC		66.5	66–72
2114	GGACACCCCCTAATCCAGGACTCC	*Fatima*, LTR	62.4	62–68
2115	CAAGCTTGCCTTCCACGCCAAG		61.6	62–67
2116	CGAACCTGGGTAAAACTTCGTGTC		57.9	58–64
2117	AGATCCGCCGGTTTTGACACCGACA		64.1	64–72
728	TGTCACGTCCAAGATGCGACTCTATC	*Sabrina*, LTR	59.8	60–66
2118	GTAGATAATATAGCATGGAGCAATC		50.8	55–61
2119	AGCCACTAGTGAAACCTATGG		54.4	55–63
2120	GTGACCTCGAAGGGATTGACAACC		59.5	60–65
2121	ACTGGATTGATACCTTGGTTCTCAA		55.8	55–62
2122	AGGGAAATACTTACGCTACTCTGC		56.3	57–64
432	GATAGGGTCGCATCTTGGGCGTGAC	*Sukkula*, LTR	63.0	63–68
480	GGAACGTCGGCATCGGGCTG		63.3	63–68
1319	TGTGACAGCCCGATGCCGACGTTCC		66.8	66–72
2123	GGAAAAGTAGATACGACGGAGACGT	*Wham*, LTR	57.8	58–63
483	TCTGCTGAAAACAACGTCAGTCC		57.8	58–63
1623	TGCGATCCCCTATACTTGTGGGT		60.1	60–65
552	CGATGTGTTACAGGCTGGATTCC	*Bagy1*, LTR	57.7	58–64
1369	TGCCTCTAGGGCATATTTCCAACAC	*BARE1*, LTR	59.0	60–65

[a]Oligonucleotide concentration is 200 nM

designed according to two principles: first, the primer length should be between 19 and 22 bases; second, the last base at 3′-end of the primer should be designed as a selective base, which is absent in repeat unit itself. We have provided examples of LTR conservation and consequent primer design for LTRs and microsatellites (Fig. 4 and Tables 1, 2, and 3).

We have designed primers using the FastPCR software [37] or Java Web tools [38]. Occasionally, not all primers (those derived from retrotransposons or SSR primers) will work in the PCR. The genome may contain too few retrotransposon or microsatellite target sites, or they may be too dispersed for the generation of PCR

Table 3
PBS 18-mer primers

Name	Sequence	T_m (°C)[a]	Optimal annealing T_a (°C)
2217	ACTTGGATGTCGATACCA	52.5	51.4
2218	CTCCAGCTCCGATTACCA	56.1	51.0
2219	GAACTTATGCCGATACCA	51.5	53.0
2220	ACCTGGCTCATGATGCCA	59.0	57.0
2221	ACCTAGCTCACGATGCCA	58.0	56.9
2222	ACTTGGATGCCGATACCA	55.7	53.0
2224	ATCCTGGCAATGGAACCA	56.6	55.4
2225	AGCATAGCTTTGATACCA	50.5	55.0
2226	CGGTGACCTTTGATACCA	54.2	53.1
2228	CATTGGCTCTTGATACCA	51.9	54.0
2229	CGACCTGTTCTGATACCA	53.5	52.5
2230	TCTAGGCGTCTGATACCA	54.0	52.9
2231	ACTTGGATGCTGATACCA	52.9	52.0
2232	AGAGAGGCTCGGATACCA	56.6	55.4
2237	CCCCTACCTGGCGTGCCA	65.0	55.0
2238	ACCTAGCTCATGATGCCA	55.5	56.0
2239	ACCTAGGCTCGGATGCCA	60.4	55.0
2240	AACCTGGCTCAGATGCCA	58.9	55.0
2241	ACCTAGCTCATCATGCCA	55.5	55.0
2242	GCCCCATGGTGGGCGCCA	69.2	57.0
2243	AGTCAGGCTCTGTTACCA	54.9	53.8
2244	GGAAGGCTCTGATTACCA	53.7	49.0
2245	GAGGTGGCTCTTATACCA	53.1	50.0
2246	ACTAGGCTCTGTATACCA	50.9	49.0
2249	AACCGACCTCTGATACCA	54.7	51.0
2251	GAACAGGCGATGATACCA	54.3	53.2
2252	TCATGGCTCATGATACCA	52.7	51.6
2253	TCGAGGCTCTAGATACCA	53.4	51.0
2255	GCGTGTGCTCTCATACCA	57.1	50.0
2256	GACCTAGCTCTAATACCA	49.6	51.0
2257	CTCTCAATGAAAGCACCA	52.4	50.0

(continued)

Table 3
(continued)

Name	Sequence	T_m (°C)[a]	Optimal annealing T_a (°C)
2295	AGAACGGCTCTGATACCA	55.0	60.0
2298	AGAAGAGCTCTGATACCA	51.6	60.0
2373	GAACTTGCTCCGATGCCA	57.9	51.0
2395	TCCCCAGCGGAGTCGCCA	66.0	52.8
2398	GAACCCTTGCCGATACCA	57.1	51.0
2399	AAACTGGCAACGGCGCCA	63.4	52.0
2400	CCCCTCCTTCTAGCGCCA	61.6	51.0
2401	AGTTAAGCTTTGATACCA	47.8	53.0
2402	TCTAAGCTCTTGATACCA	49.0	50.0
2415	CATCGTAGGTGGGCGCCA	62.5	61.0

[a]Oligonucleotide concentration is 1,000 nM

products. Alternatively, sequence divergence in ancient retrotransposon insertions or polymorphisms between heterologous primers and native elements may lead to poor amplification. Some primers generate smears under all PCR conditions. Many sources can contribute to this problem, ranging from primer structure to variability in the target site and competition from other target sites. Generally, it is more efficient to design another primer than to try to identify the source of the problem. Furthermore, primers which produce a single, very strong band are not suitable for fingerprinting.

3.2 Template DNA

The quality of template DNA plays an important role in the quality of the resulting fingerprint. Standard DNA extraction methods are sufficient to yield DNA of high quality from most samples. DNA should be free of polysaccharides, pigments, and secondary metabolites. Some tissue materials contain much polysaccharides, pigments, oils, or polyphenols, which can reduce the efficiency of PCR. Furthermore, contaminated DNAs will decline in PCR performance during prolonged (a month or more) periods of storage, due to chemical modification (*see* **Note 2**). Such DNAs (e.g., from *Brassica* spp.) should be extracted, for example, with methods involving guanidine thiocyanate at weakly acidic pH (below pH 6), followed by hot chloroform DNA extraction. DNA templates should be diluted with 1× TE solution for the appropriate working concentration (5 ng/μL) and stored at 4 °C. High-quality DNA can be stored at 4 °C for many years without showing any PCR inhibition or decrease in amplification efficiency for the longer bands.

3.3 Polymerase Chain Reaction

3.3.1 Protocol for IRAP, REMAP, and iPBS

The method described below is for reactions with standard *Taq* polymerase or with proofreading Phire® Hot Start II DNA Polymerase. PCR products can be separated by agarose gel electrophoresis or, if fluorescent-labeled primers are used after *Taq*I digestion of PCR fragments, by sequencing gel systems instead. For separation on sequencing systems, fluorescent, Cy5- or Cy3-labeled primers may be used; no special reaction conditions are needed.

1. The PCR can be set up at room temperature. Prepare a master mix for the appropriate number of samples. The DNA polymerase is the last component added to the PCR mixture. Mix well the master mix and centrifuge the tube. The reaction volume may vary from 10 to 25 μL; 10 μL is enough for running two gels. The final primer concentration(s) in the reaction can vary from 200 to 400 nM for primers in combination. For a single PCR primer, use 400 nM for IRAP and 1,000 nM for iPBS amplification. Although higher primer concentrations increase PCR efficiency and the rapidity of DNA amplification, they also produce over-amplified products.
2. Perform PCR with Phire® Hot Start II DNA Polymerase in a 25 μL reaction mixture containing 25 ng DNA (*see* **Note 3**), 1× Phire Reaction Buffer (containing 1.5 mM $MgCl_2$), 0.2–1 μM primer(s), 200 μM dNTP, 0.2 μL Phire® Hot Start II DNA Polymerase.
3. Alternative protocol for PCR with Taq polymerase: use a 25 μL reaction mixture containing 25 ng DNA, 1× Taq PCR Buffer (including 1.5 mM $MgCl_2$), 0.2–1 μM primer(s), 200 μM dNTP, 0.2 μL (1 U) Taq DNA polymerase (5 U/μL).
4. Centrifuge all tubes or the plate before starting amplification.
5. The PCR with Phire® Hot Start II DNA Polymerase (40 min) should consist of a 2 min initial denaturation step at 98 °C and 30–32 cycles of 5–10 s at 98 °C, 30 s at 55–72 °C (*see* **Note 4**), and 30 s at 72 °C; complete with a 2 min final extension at 72 °C. The denaturation step in PCR needs to be as short as possible. Usually 5 s at 98 °C is enough for most templates. For some templates requiring longer denaturation time, up to 10 s can be used, with the initial denaturation time extended up to 3 min.
6. The standard PCR with Taq polymerase (70 min total) should consist of: a 3 min initial denaturation step at 95 °C; 30–32 cycles of: 15 s at 95 °C, 30 s at 55–72 °C, and 60 s at 72 °C; a 5 min a final extension at 72 °C. PCR thermal conditions can be varied without large effects on the resulting band pattern.
7. The time of the annealing step can vary from 30 to 60 s, and the annealing temperature depends on the melting temperature of the primer; it should be between 55 and 68 °C (60 °C is optimal for almost all primers and their combinations in IRAP and REMAP; *see* **Note 5**).
8. PCRs can be stored at 4 °C overnight.

3.3.2 Long Distance Inverted PCR to Isolate Complete LTR Retrotransposons

Complete LTR retrotransposons can be identified and extracted using long distance PCR with inverted LTR primers and running low numbers of PCR cycles [10–15] to select for abundant elements. The iPBS amplification technique helps with cloning of LTR segments from genomic DNA or with database searches. After retrieving LTR sequences of a selected family of retrotransposons, align them to identify the most conserved regions. The conserved segments of the LTR are used for design of inverted primers for long distance PCR, such as for cloning of whole elements (*see* **Note 6**).

Several primer pairs, oriented away from each other as for inverse PCR, are designed for each identified element. Inverted primers of 25–30 nt with high T_m (>60 °C) need to be designed from the LTR. This allows annealing and polymerase extension in one step at 68–72 °C, thereby increasing the efficiency of the amplification of long fragments. The PCR product will be consist of complementary fragments of both LTRs and the central part of retroelement. To avoid formation of nonspecific PCR products and, assuming a high copy number for the retroelement of interest, the reaction is carried out with a low number of PCR cycles [10–15].

The PCRs can be set up at room temperature.

1. The 100 μL reaction volume contains 1× Phusion™ HF buffer, 100 ng DNA, 300 nM of each primer, 200 μM dNTP, 2 U Phusion™ High-Fidelity DNA Polymerase.
2. The reaction cycle consists of a 1 min initial denaturation step at 98 °C; 10–15 cycles of 10 s at 98 °C, 4 min at 72 °C; a final extension of 5 min at 72 °C.
3. PCRs can be stored at 4 °C overnight.

3.3.3 Digestion of PCR Products with Restriction Endonucleases Without Prior Purification

For separation on sequencing systems, fluorescent, Cy5- or Cy3-labeled primers need to be used. The PCR products are digested using restriction enzymes recognizing four nucleotides when they exceed the system resolution range (generally over 500 bp; Fig. 5). Such enzymes include *Alu*I, *Csp6*I, *Msp*I, *Taï*I, and *Taq*I.

1. Reactions are set up at room temperature. Prepare a 2× master mix for the appropriate number of samples to be amplified.
2. Perform PCR product digestion with *Taï*I restriction enzyme in a 20 μL reaction mixture.
3. To 10 μL PCR products (from IRAP, REMAP, or iPBS amplification), add 10 μL mix of 2× FastDigest® buffer with 1 μL of FastDigest® *Taï*I.
4. Centrifuge all tubes or the plate before starting reaction. Incubate at 65 °C for 30 min.

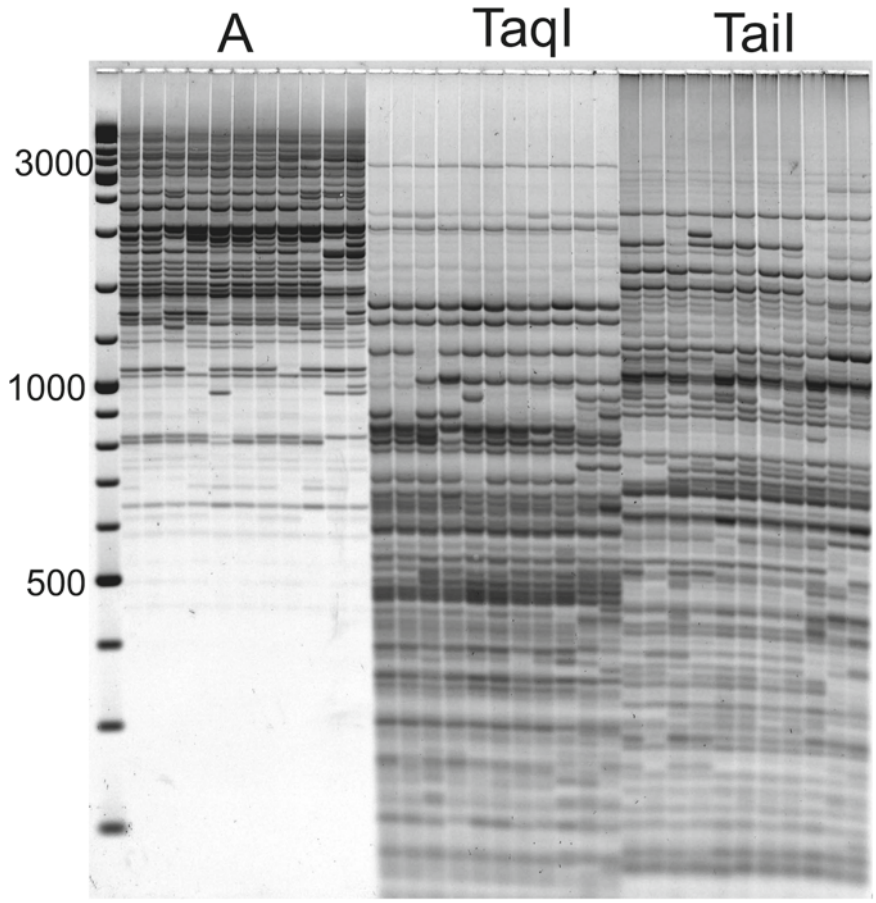


Fig. 5 Digestion of IRAP PCR products with restriction endonucleases. IRAP amplification with *Sukkula* LTR primer (432) is shown (*A*) prior to digestion; (*TaqI*) digestion with *TaqI* and (*TaiI*) with *TaiI* enzymes. A 100 bp DNA ladder is present on the *left*

3.4 Sample Preparation and Loading

Add an equal volume of 2× loading buffer to the completed PCRs in tubes or plates and mix well. Collect the mixture by a short centrifugation (by turning a benchtop microcentrifuge on and immediately off again). Load the gels with a sample volume of 8–10 μL. The DNA concentration plays an important role in gel resolution. Overloaded lanes will result in poor resolution.

3.5 Casting the Agarose Gel

1. Prepare 200 mL of 1.6 % (w/v) agarose containing 1× THE buffer in a 500 mL bottle. This volume is required for one gel with the dimensions 0.4 cm × 20 cm × 20 cm. Dissolve and melt the agarose in a microwave oven. The bottle should be closed, but the plastic cap must not be tightened! The agarose gel must be completely melted in the microwave and then allowed to slowly cool until its temperature drops to about 50–60 °C.

At that point, if desired, add the ethidium bromide solution at a rate of 50 μL per 100 mL, to bring the final concentration to 0.5 μg per mL (alternatively the gel can be stained at the end of the run). Take care not to boil over the agarose. Add ethidium bromide only after removing the agarose from the microwave oven to minimize risks from boilover. The agarose gel must melt and dissolve properly. Small undissolved inclusions will severely hamper the quality of the results. Do not allow the gel to cool unevenly before casting, for example, by leaving it stand on the benchtop or in cool water. The best way to cool the agarose is by shaking it at 37 °C for 15 min. Careful casting of gels is critical to success. Small, undissolved agarose inclusions in the gels will result in bands with spiked smears.

2. Pour the agarose into the gel tray (20×20 cm). Allow the agarose to solidify at room temperature for 1 h minimum. For optimal resolution, cast horizontal gels 3–4 mm thick. The volume of gel solution needed can be estimated by measuring the surface area of the casting chamber and then multiplying by gel thickness.
3. Fill the chamber with 1× THE running buffer until the buffer reaches about 3–5 mm over the surface of the gel.

3.6 Gel Electrophoresis

Select running conditions appropriate to the configuration of your electrophoresis box. For a standard 20×20 cm gel, carry out electrophoresis at a constant 80–100 V for 5–9 h (a total of 700–900 Vh). Electrophoresis may cause the gels to deteriorate after several hours; their temperature should not be allowed to exceed 30 °C, above which electrophoretic resolution will be impaired. Still better results are obtained with a slower run. We routinely use 90 V for 7 h or overnight at 50 V for 14 h (700 Vh). As the end of the run approaches, it is helpful to check the run with a UV transilluminator. For samples with many or large (>500 bp) bands, perform the gel electrophoresis at a constant voltage of 50 V overnight (17 h).

3.7 DNA Visualization

A high-quality gel scanner with good sensitivity and resolution is also very important. Older video systems, which may be suitable for checking the success of restriction digests, cloning reactions, or simple PCRs, are not suitable for analysis of complex banding patterns. DNA can be visualized directly by casting ethidium bromide into gel as described above or by incubating in an ethidium bromide solution of equivalent strength following electrophoresis. The gels are scanned on an FLA-5100 imaging system (Fuji Photo Film GmbH., Germany) or equivalent scanner with a resolution of 50–100 μm, or on a digital gel electrophoresis scanner for detection of ethidium bromide-stained nucleic acids by fluorescence using a second-harmonic-generation (SHG) green laser, 532 nm, or by SYBR Green (Molecular Probes), GelGreen (Biotium), GelStar (Lonza), using a SHG blue laser, 473 nm.

3.8 Cloning PCR Fragment

This protocol is for T/A-end cloning of blunt-ended DNA fragments that are amplified with a proofreading DNA polymerase (*Phusion*™, *Phire*® II, and *Pfu* DNA polymerases). When the PCR mixture contains more than one band of amplified DNA, purify the target fragment by electrophoresis in an agarose gel with SYBR Green I stained DNA. If not purified by gel electrophoresis, PCR-amplified DNA should be prepared for ligation by purification with a QIAEX II Gel Extraction Kit.

1. Non-templated 3′ adenosine is added to the blunt-end PCR fragments by adding 5 U of Taq polymerase per 100 μl reaction volume and 1 μM dATP (a final concentration) directly to PCR mix that is amplified with proofreading DNA polymerase and incubating for 30 min at 72 °C.
2. Mix the samples of DNA with 10× loading buffer and 50× SYBR Green I solution to 1× final concentrations; load them into the slots of the gel. Electrophoresis is performed with 1 % agarose gel in 1× THE buffer and at a constant voltage of 70 V, about 3 h or longer.
3. Following agarose gel electrophoresis, purify the band. Under a Dark Reader, use a sharp scalpel or razor blade to cut out a slice of agarose containing the band of interest and transfer it to a clean, disposable plastic tube.
4. After gel extraction with QIAEX II Gel Extraction Kit, PCR fragments are with TOPO® TA Cloning® Kit.

4 Notes

1. Most enzymes are supplied with their own recommended buffer; these buffers are often suitable for other thermostable polymerases as well. The concentration of $MgCl_2$ ($MgSO_4$) can be varied from 1.5 to 3 mM without influencing the fingerprinting results. A higher $MgCl_2$ concentration can increase the PCR efficiency and allow reduction in the number of PCR cycles from 30 to 28 and also help in PCRs containing somewhat impure DNA. Additional components such as (listed at their final concentration in the reaction buffer) 5 % acetamide, 0.5 M betaine (*N*,*N*,*N*-trimethylglycine), 3–5 % DMSO, 5–10 % glycerol, 5 % PEG 8000, and 5–20 mM TMA (Tetramethylammonium chloride) can increase the PCR efficiency for multiple templates and PCR products [39].
2. The DNA quality is very important, as it is for most PCR-based methods. DNA purification with a spin-column containing a silica-gel membrane is not a guarantee of high DNA quality for all plant samples or tissues. One sign of DNA contamination is that, after some period of time (a month or more) in storage, only short bands can be amplified.

3. The amount of DNA template should be about 1 ng DNA per 1 μL of PCR volume. Much higher DNA concentrations will produce smears between the bands, which is a sign of over-amplification.
4. The result from primer T_m calculation can vary significantly depending on the method used. We have provided a system for T_m calculation and corresponding instructions on the website http://primerdigital.com/tools/ to determinate the T_m values of primers and optional annealing temperature. If using a two-step PCR protocol, where both annealing and extension occur in a single step at 68–72 °C, the primers should be designed accordingly. If necessary, use a temperature gradient to find the optimal temperature for each template-primer pair combination.
5. The optimal annealing temperature (T_a) is the range of temperatures where efficiency of PCR amplification is maximal without nonspecific products. Primers with high T_ms (>60 °C) can be used in PCRs with a wide T_a range compared to primers with low T_ms (<50 °C). The optimal annealing temperature for PCR is calculated directly as the value for the primer with the lowest T_m (T_m^{min}) plus the natural logarithm of fragment size $T_a = T_m^{min} + \ln(L)$, where L is length of PCR fragment [38].
6. Development of a new marker system for an organism in which retrotransposons have not been previously described generally takes about 1 month. The availability of heterologous and conserved primers as well as experience in primer design, sequence analysis, and testing speeds up the development cycle. Routine analysis of samples with optimized primers and reactions may be carried out thereafter. Retrotransposons have several advantages as molecular markers. Their abundance and dispersion can yield many marker bands, the pattern possessing a high degree of polymorphism due to transpositional activity. The LTR termini are highly conserved even between families, yet longer primers can be tailored to specific families. Unlike DNA transposons, the new copies are inserted but not removed. Even intra-element recombination resulting in the conversion of a full-length element to a solo LTR does not affect its performance in IRAP or REMAP. Retrotransposon families may vary in their insertional activity, allowing the matching of the family used for marker generation to the phylogenetic depth required. The primers for different retrotransposons and SSRs can be combined in many ways to increase the number of polymorphic bands to be scored. Furthermore, the length and conservation of primers to the LTRs facilitate cloning of interesting marker bands and the development of new retrotransposons for markers. The IRAP and REMAP fingerprinting patterns can be used in a variety of applications, including measurement

of genetic diversity and population structure [17, 40–43], determination of essential derivation, marker-assisted selection, and recombinational mapping [3, 44]. In addition, the method can be used to fingerprint large genomic clones (e.g., BACs) for the purpose of assembly. The method can be extended, as well, to other prevalent repetitive genomic elements such as MITEs. Although SSAP is somewhat more general than IRAP or REMAP, requiring only a restriction site near the outer flank of a retroelement, its requirement for two additional enzymatic steps introduces the possibility of artifacts from DNA impurities, methylation, and incomplete digestion or ligation. Furthermore, SSAP generally requires selective nucleotides on the 3′ ends of the retrotransposon primers in order to reduce the number of amplification products and increase their yield and resolvability. As for IRAP and REMAP, the resulting subsets of amplifiable products are not additive [18]. The strength of all these methods is that the degree of phylogenetic resolution obtained depends on the history of activity of the particular retrotransposon family being used. Hence, it is possible to analyze both ancient evolutionary events such as speciation and the relationships and similarities of recently derived breeding lines. The IRAP and REMAP can be generalized, furthermore, to other transposable element systems, such as to MITEs, and to other organisms. For example, the SINE element *Alu* of humans has been used in a method called *Alu*-PCR in a way similar to IRAP and REMAP [7, 45, 46].

Acknowledgments

This work was supported by Academy of Finland grant 134079. Anne-Mari Narvanto is thanked for excellent technical assistance.

References

1. Wicker T, Keller B (2007) Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res 17:1072–1081
2. Cheng ZJ, Murata M (2003) A centromeric tandem repeat family originating from a part of Ty3/gypsy-retroelement in wheat and its relatives. Genetics 164:665–672
3. Boyko E et al (2002) Combined mapping of *Aegilops tauschii* by retrotransposon, microsatellite, and gene markers. Plant Mol Biol 48:767–790
4. Lamb JC et al (2007) Plant chromosomes from end to end: telomeres, heterochromatin and centromeres. Curr Opin Plant Biol 10:116–122
5. Meyer W et al (1993) Hybridization probes for conventional DNA fingerprinting used as single primers in the polymerase chain reaction to distinguish strains of *Cryptococcus neoformans*. J Clin Microbiol 31:2274–2280
6. Sivolap IM, Kalendar RN, Chebotar SV (1994) The genetic polymorphism of cereals demonstrated by PCR with random primers. Tsitol Genet 28:54–61
7. Charlieu JP et al (1992) 3′ Alu PCR: a simple and rapid method to isolate human polymorphic markers. Nucleic Acids Res 20: 1333–1337
8. Zietkiewicz E, Rafalski A, Labuda D (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20:176–183

9. Feschotte C, Jiang N, Wessler S (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341
10. Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97: 381–388
11. Schulman AH, Kalendar R (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogenet Genome Res 110:598–605
12. Hedges DJ, Batzer MA (2005) From the margins of the genome: mobile elements shape primate evolution. Bioessays 27:785–794
13. Ostertag EM, Kazazian HH (2005) Genetics: LINEs in mind. Nature 435:890–891
14. Jurka J (2004) Evolutionary impact of human Alu repetitive elements. Curr Opin Genet Dev 14:603–608
15. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 101:14572–14579
16. Leigh F et al (2003) Comparison of the utility of barley retrotransposon families for genetic analysis by molecular marker techniques. Mol Genet Genomics 269:464–474
17. Kalendar R et al (2010) Analysis of plant diversity with retrotransposon-based molecular markers. Heredity 106:520–530
18. Waugh R et al (1997) Genetic distribution of *BARE*-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet 253:687–694
19. Vogel JM, Morgante M (1992) A microsatellite-based, multiplexed genome assay, In *Plant Genome III Conference*. San Diego, CA USA
20. Korswagen HC et al (1996) Transposon Tc1-derived, sequence-tagged sites in *Caenorhabditis elegans* as markers for gene mapping. Proc Natl Acad Sci U S A 93: 14680–14685
21. Van den Broeck D et al (1998) *Transposon Display* identifies individual transposable elements in high copy number lines. Plant J 13:121–129
22. Ellis THN et al (1998) Polymorphism of insertion sites of Ty1-copia class retrotransposons and its use for linkage and diversity analysis in pea. Mol Gen Genet 260:9–19
23. Vos P et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 21:4407–4414
24. Yu G-X, Wise RP (2000) An anchored AFLP- and retrotransposon-based map of diploid *Avena*. Genome 43:736–749
25. Porceddu A et al (2002) Development of S-SAP markers based on an LTR-like sequence from *Medicago sativa* L. Mol Genet Genomics 267:107–114
26. Lee D et al (1990) A *copia*-like element in *Pisum* demonstrates the uses of dispersed repeated sequences in genetic analysis. Plant Mol Biol 15:707–722
27. Syed NH, Flavell AJ (2006) Sequence-specific amplification polymorphisms (SSAPs): a multilocus approach for analyzing transposon insertions. Nat Protoc 1:2746–2752
28. Pearce SR et al (1999) Rapid isolation of plant Ty1-copia group retrotransposon LTR sequences for molecular marker studies. Plant J 19:711–717
29. Vershinin AV, Ellis TH (1999) Heterogeneity of the internal structure of PDR1, a family of Ty1/copia-like retrotransposons in pea. Mol Gen Genet 262:703–713
30. Ramsay L et al (1999) Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J 17:415–425
31. Hirochika H, Hirochika R (1993) Ty1-copia group retrotransposons as ubiquitous components of plant genomes. Jpn J Genet 68: 35–46
32. Flavell AJ et al (1992) *Ty1-copia* group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res 20:3639–3644
33. Ellis THN et al (1998) *Ty1-copia* class retrotransposon insertion site polymorphism for linkage and diversity analysis in pea. Mol Gen Genet 260:9–19
34. Kalendar R et al (2008) *Cassandra* retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci U S A 105:5833–5838
35. Witte CP et al (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A 98:13778–13783
36. Corpet F (1988) Multiple sequence alignment with hierarchical-clustering. Nucleic Acids Res 16:10881–10890
37. Kalendar R, Lee D, Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. Genes, Genomes and Genomics 3:1–14
38. Kalendar R, Lee D, Schulman AH (2011) Java web tools for PCR, *in silico* PCR, and oligonucleotide assembly and analysis. Genomics 98:137–144
39. Kovarova M, Draber P (2000) New specificity and yield enhancer of polymerase chain reactions. Nucleic Acids Res 28:E70
40. Baumel A et al (2002) Inter-retrotransposon amplified polymorphism (IRAP), and retrotransposon-microsatellite amplified polymorphism (REMAP) in populations of the young

allopolyploid species *Spartina angelica* Hubbard (Poaceae). Mol Biol Evol 19:1218–1227
41. Boronnikova SV, Kalendar RN (2010) Using IRAP markers for analysis of genetic variability in populations of resource and rare species of plants. Russ J Genet 46:36–42
42. Belyayev A et al (2010) Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. Mob DNA 1:6
43. Smýkal P et al (2011) Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. Theor Appl Genet 122:1385–1397
44. Manninen OM et al (2006) Mapping of major spot-type and net-type net-blotch resistance genes in the Ethiopian barley line CI 9819. Genome 49:1564–1571
45. Tang JQ et al (1995) Alu-PCR combined with non-Alu primers reveals multiple polymorphic loci. Mamm Genome 6:345–349
46. Shedlock AM, Okada N (2000) SINE insertions: powerful tools for molecular systematics. Bioessays 22:148–160

Chapter 13

Phylogenetic Reconstruction Methods: An Overview

Alexandre De Bruyn, Darren P. Martin, and Pierre Lefeuvre

Abstract

Initially designed to infer evolutionary relationships based on morphological and physiological characters, phylogenetic reconstruction methods have greatly benefited from recent developments in molecular biology and sequencing technologies with a number of powerful methods having been developed specifically to infer phylogenies from macromolecular data. This chapter, while presenting an overview of basic concepts and methods used in phylogenetic reconstruction, is primarily intended as a simplified step-by-step guide to the construction of phylogenetic trees from nucleotide sequences using fairly up-to-date maximum likelihood methods implemented in freely available computer programs. While the analysis of chloroplast sequences from various *Vanilla* species is used as an illustrative example, the techniques covered here are relevant to the comparative analysis of homologous sequences datasets sampled from any group of organisms.

Key words Phylogeny, DNA sequence, Alignment, Phylogenetic tree, Maximum likelihood

1 Introduction

Invented by Haeckel in 1866, and derived from the concept of genealogy, the term "phylogeny" describes the relationships between entities (such as species, genes, or genomes) [1] in such a way as to reflect their evolutionary histories. Given the assumption that measurable similarities between organisms are suggestive of their common evolutionary history, by comparing analogous features of contemporary or fossil organisms (such as beak sizes or the amino acid sequences of some specific gene), phylogeneticists try to infer the pathways of evolutionary change that yielded these features.

The patterns of relationship and evolutionary pathways that are revealed by phylogenetic analyses are most commonly depicted in the form of phylogenetic trees, the branching patterns and branch lengths of which graphically describe the relative relatedness of the different species, individuals, genes, or other entities of interest (often called operational taxonomic units or OTUs) that were used to construct the trees (Fig. 1).

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_13,

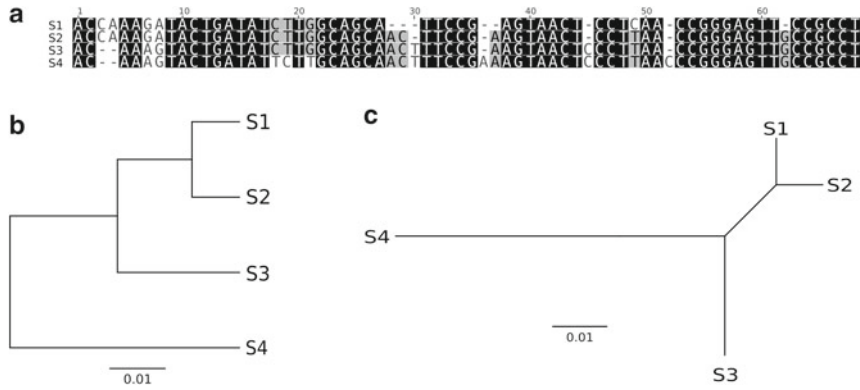


Fig. 1 (**a**) Alignment of four sequences S1, S2, S3 and S4, different shades of *grey* representing polymorphic sites. (**b**) and (**c**), respectively, rooted and unrooted tree of the four sequences presented in the alignment, scale bar corresponding to number of substitutions per site

1.1 Use of Macromolecular Data in Phylogenetic Reconstruction

For many years, phylogenetic trees were constructed based on comparisons among individuals of morphological and physiological characters such as skull morphologies in primates [2] or self-incompatibility in plants [3]. Recent advances in molecular biology and sequencing technologies have created an exponentially increasing volume of DNA and protein sequence data. Since such macromolecular data can be interpreted as a linear string of multistate characters (with four possible states for DNA sequences and 20 possible states for amino acids sequences), their use in phylogeny reconstruction is extremely straightforward. Almost since the moment such data first became available, it was clearly evident that there existed a "molecular clock" such that differences in homologous nucleotide or protein sequences should be good indicators of their relatedness [4, 5]. Furthermore, it was soon realized that the almost universality of the genetic code enabled the use of nucleotide and amino acid sequences to infer phylogenetic relationships even among organisms that were so distantly related that they shared no discernible common morphological or physiological characters. Finally, the mutational processes underlying nucleotide and amino acid sequence evolution are far more amenable to detailed fully probabilistic mathematical modelling than are the gross changes that occur during the evolution of morphological or physiological characters [6].

1.2 Main Methods Used in Phylogenetic Reconstruction

The starting material for constructing any phylogenetic tree from nucleotide or amino acid sequence data is the gathering of sets of evolutionarily related, or homologous, sequences. However, before using these sequences to construct a phylogenetic tree, it is important to ensure that each nucleotide or amino acid in each sequence is compared only with the corresponding homologous nucleotides or amino acids in the other sequences. This preliminary task is performed by aligning the sequences to one another, to obtain a

discrete character matrix called an alignment (see an example in Fig. 1a), within which each row represents one of the sequences and each column a set of homologous nucleotides or amino acids. This step will be described in more detail later, but it should be borne in mind that it is one of the trickiest parts of the whole phylogenetic reconstruction process.

Numerous methods have been developed to infer phylogenetic trees from multiple sequence alignments. Whereas some, called character-based methods, directly use the individual columns of aligned nucleotides or amino acids, others, called distance-based methods, use measures of the overall differences between all pairs of sequences in the alignment (represented as a matrix of pairwise genetic distances). Also, whereas some methods are limited to accommodate only the simplest models of evolution, others offer the capacity to explicitly model various different aspects of the evolutionary process.

Crucially, each method has its own good and bad points, and the choice of one method over another for any particular analysis tends to depend primarily on a compromise between the desired complexity or accuracy of the analysis and the amount of time it will take to run. Nevertheless, with the numerous computational optimizations that have been made to modern phylogenetic reconstruction algorithms and the speed of today's computers, even the various "slow-but-accurate" methods are applicable to most datasets. Applying one such tree construction method, called maximum likelihood (ML), will be the primary focus of the practical component of this chapter.

1.2.1 Distance-Based Methods

Distance-based methods rely on the calculation of genetic distances between each pair of sequences in a dataset with phylogenetic trees being constructed from the resulting distance matrix using a clustering algorithm [7]. The simplest distance measure, the "p-distance" (also called the Hamming distance), corresponds to the proportion of different sites between each pair of taxa. This value is an underestimation of the true evolutionary distance, since the possibility that multiple unseen mutations may have occurred at individual sites is not taken into consideration. Since phylogenetic analyses are concerned primarily with the inference of evolutionary relationships, it is desirable to, in most cases, calculate evolutionary distances rather than simply p-distances. This can be achieved using an explicit model of evolution that corrects the p-distance.

Given a matrix of evolutionary distances, a tree can then be obtained using the unweighted pair grouping with arithmetic mean (UPGMA; [8]), least squares (LS, also called minimum evolution [9]), neighbor-joining (NJ [10]), or BIONJ [11] methods. The primary advantage of such distance-based methods is their computational speed. These methods are therefore ideally suited to the initial exploration of evolutionary relationships between sequences in a dataset. Many of these methods also directly help speed up

slower but more accurate character-based tree construction methods by directing these methods to preferably search for highly likely phylogenetic trees that resemble distance-based trees.

Despite their obvious utility, distance-based methods tend to disregard much of the potentially evolutionarily informative information within an alignment by compressing it into sets of pairwise evolutionary distances [1, 12]. Indeed, all information provided by the distribution of the character states, or relationships between particular characters and the tree, are lost in the process of pairwise-distance calculations [13].

1.2.2 Character-Based Methods

While distance-based methods compress phylogenetic information within a set of sequences into a pairwise-distance matrix, character-based methods take advantage of all the information available in sequences at each homologous site. The most widely used methods are the maximum parsimony (MP; [14, 15]) and maximum likelihood (ML; [16, 17]) methods. Whereas the maximum parsimony methods generally implicitly assume a very simple model of evolution (usually one where all possible nucleotide substitutions are equally probable), the primary appeal of maximum likelihood methods is that they are probabilistic and enable the application of a wide variety of explicit evolutionary models.

The Maximum Parsimony Method

The MP method was, until recently, one of the most widely used for phylogenetic inference. Initially developed for the analysis of morphological traits, its main underlying idea can be best summed up by the principle of Ockham's razor, which states that when several hypotheses with different degrees of complexity are proposed to explain the same phenomenon, one should choose the simplest hypothesis. Framed in a phylogenetic context, this principle proposes that the most believable or parsimonious phylogenetic tree will be the tree that invokes the smallest number of evolutionary changes during the divergence of the sequences it represents [18–20].

A major issue encountered when inferring the most parsimonious tree for a given set of nucleotide sequences is that there will frequently be multiple equally parsimonious trees. In such cases, it is often desirable to produce a strict consensus tree which includes only the topological features that are found in every tree. In a strict consensus tree, it is assumed that unresolved features of the tree topology represent relationships that cannot be resolved due to the studied sequences containing too few phylogenetically informative sites.

Another option when multiple equally parsimonious trees exist is to produce a majority-rule consensus tree which includes only the topological features that are present in at least half of the trees. In many cases, however, such majority-rule consensus trees can have topologies that contradict some of the equally parsimonious trees which should, in fact, be considered just as plausible as the consensus tree [21].

The Maximum Likelihood Method

Maximum likelihood (ML) is a statistical concept popularized by Fischer in the early 1900s [22] that has been widely used in many areas of biology such as population genetics and ecological modelling, and which was first applied to the field of phylogenetics by Joseph Felsenstein in 1973 [23]. The basic concept of likelihood is relatively simple to comprehend: given some data D (in our case, nucleotide or amino acid sequences), under a model of evolution, M (which is explicitly defined and describes the mutation process from one base to another), the likelihood of a set of parameters, θ (tree topology, tree branch lengths, substitution model parameters), corresponds to the probability of obtaining D under the model M with parameters θ. The maximum likelihood estimates of the parameter values included in θ correspond to the set of values that maximize this probability. If the principle is easily understandable, the calculation of the likelihood function can be mathematically complex, and this method can be very computationally expensive.

Although the actual calculation of the likelihood of a particular tree topology and a defined set of parameter values can be quite rapidly computed, finding the set of parameters, including the tree branching orders and branch lengths, for which the likelihood is maximized, is much more computationally daunting due to the incredibly large numbers of trees that must be evaluated even for datasets containing only modest numbers of sequences (e.g., 15 or more).

For this reason, various computational tricks have been devised to cut down on the numbers of trees that need to be evaluated to find the one with the maximum likelihood [24]. Unlike exhaustive search methods, which evaluate all possible phylogenetic trees, so-called exact tree searching methods such as the branch-and-bound method [25] reduce the number of trees that must be evaluated while still guaranteeing that the best tree will be found. However, even exact methods like branch-and-bound are too slow to evaluate datasets with more than a modest number of sequences.

For large datasets, it is usually necessary to use so-called approximate tree searching methods which, while guaranteeing to find trees with high likelihoods, will not always find the tree with the maximum likelihood [26]. The most commonly used approximate tree searching methods involve tree rearrangement approaches such as nearest-neighbor interchange (NNI) and subtree pruning and regrafting (SPR) [27]. These tree rearrangement approaches involve modifications of local branching patterns within small subsections of the tree while leaving the rest of the tree untouched and then testing the new tree to determine whether the rearrangements yield a tree with an improved likelihood. If the new tree has a better likelihood than the original, it is selected for a new round of rearrangements. The process continues until further rearrangements fail to improve the likelihood. The main failing of approximate tree searching methods like NNI and SPR is that just because simple branch rearrangements fail to yield trees

with improved likelihoods, it does not mean that more complex tree rearrangements will not. In many cases, the only pathway to the maximum likelihood tree will involve multiple simultaneous tree rearrangements. Nevertheless, despite the tendency of approximate tree searching methods to become entrapped on local likelihood peaks [26], they will generally very rapidly yield trees with likelihoods close to the maximum.

1.3 Step-by-Step Construction of Phylogenetic Trees

Sequence-based phylogenetic reconstructions can be divided into five main steps: (1) choosing a genome region to study, (2) identifying and retrieving sets of homologous sequences from the same genome region of related individuals, (3) aligning homologous nucleotide/amino acid sites within these sequences, (4) constructing a phylogenetic tree, and (5) visualizing the tree. Although one might assume that the fourth step is the most complex, it is very important to realize that to produce a meaningful tree, none of these steps should be neglected.

Here we describe a robust up-to-date protocol for constructing phylogenetic trees using what is today the most popular available maximum likelihood tree construction software. This chapter is aimed at phylogenetics newbies, and readers interested in a more detailed dissection of the phylogenetic tree construction process are encouraged to consult some of the many excellent reviews and books on advanced tree construction methodologies [13, 24, 28, 29]. We will illustrate the step-by-step construction of a phylogenetic tree using chloroplast *rbc*L sequences sampled from various species of vanilla plants (refer to Chapter 5 on sampling and sequencing protocols).

2 Materials

1. Windows, Mac OS X 10 3+, or Linux Operating System.
2. MEGA5 (http://www.megasoftware.net/) [30].
3. jModelTest2 (http://code.google.com/p/jmodeltest2/) [31].
4. PhyML3 (http://www.atgc-montpellier.fr/phyml/binaries.php) [32].
5. FigTree (http://tree.bio.ed.ac.uk/software/figtree/) [33].
6. Internet connection and an internet browser.
7. Example files located at http://phylorec.blogspot.com/.

3 Methods

3.1 Obtaining a Sequence Dataset

The aim of phylogenetic tree construction is to determine as accurately as possible the evolutionary relationships between groups of organisms. As different sequence datasets can be built to answer

different kinds of question (e.g., do humans and chimpanzees have a more recent common ancestor than either have with gorillas?) or challenge different hypotheses (e.g., chimpanzees are the nearest extant primate relative of humans), before beginning a phylogeny reconstruction, it is important to properly think about the specific biological question (if any) that is being addressed. At this point, it is also crucial to realize that assembling an appropriate sequence dataset and producing an alignment are the most difficult and important part of almost all phylogenetic analyses. If an inappropriate set of sequences is selected (e.g., a set of human and primate sequences where chimpanzee and gorilla sequences are left out) or the alignment produced is of low quality (e.g., if a misalignment error makes it seem that humans are a divergent outlier among the primates), further analysis with the resulting phylogenetic tree could be extremely misleading [34]. Assuming a new DNA sequence was obtained from a newly sampled individual belonging to a particular species, it could be useful to gather sequences belonging to the same or related species from a database. The National Center for Biotechnology Information (NCBI) database and GenBank [35], along with the EMBL Nucleotide Sequence Database, EMBL-Bank [36], and the DNA Database of Japan (DDBJ) [37], host billions of DNA and protein sequence records. Each sequence record is associated with information such as the organism from which the sequence was derived, the author of the sequence, the scientific publications analyzing the sequence, its sampling date and place, and many other pieces of information. Importantly, each sequence is also associated with a set of unique identification numbers, the most important of which is called its accession number. With an accession number in hand, it is very straightforward to retrieve any previously determined and deposited nucleotide or amino acid sequence.

For the purpose of reconstructing the vanilla phylogeny, we generated a fake partial *rbc*L sequence that we will imagine has been obtained using two pairs of PCR primers that amplify two overlapping fragments (*see* ref. 38 for a detailed protocol), and we will use the NCBI database to retrieve homologues of this sequence. Two complementary approaches can be used to query this database and gather the sequence dataset. The first relies on the use of the NCBI basic local alignment search tool (BLAST) (http://blast.ncbi.nlm.nih.gov/Blast.cgi; [39]). BLAST allows one to find a set of sequences with some degree of similarity with a query sequence. This approach is really valuable when no knowledge of the studied sequence is available. BLAST matches are sorted by a "similarity score" (*S* value) and approximate probabilities (*E* values) of the identified similar sequences not being related by evolutionary descent (i.e., that they are similar by chance alone). The selection of homologous sequences is often based upon an *S* value threshold. This common approach consists of taking two sequences to be homologous only if their level of similarity is high enough

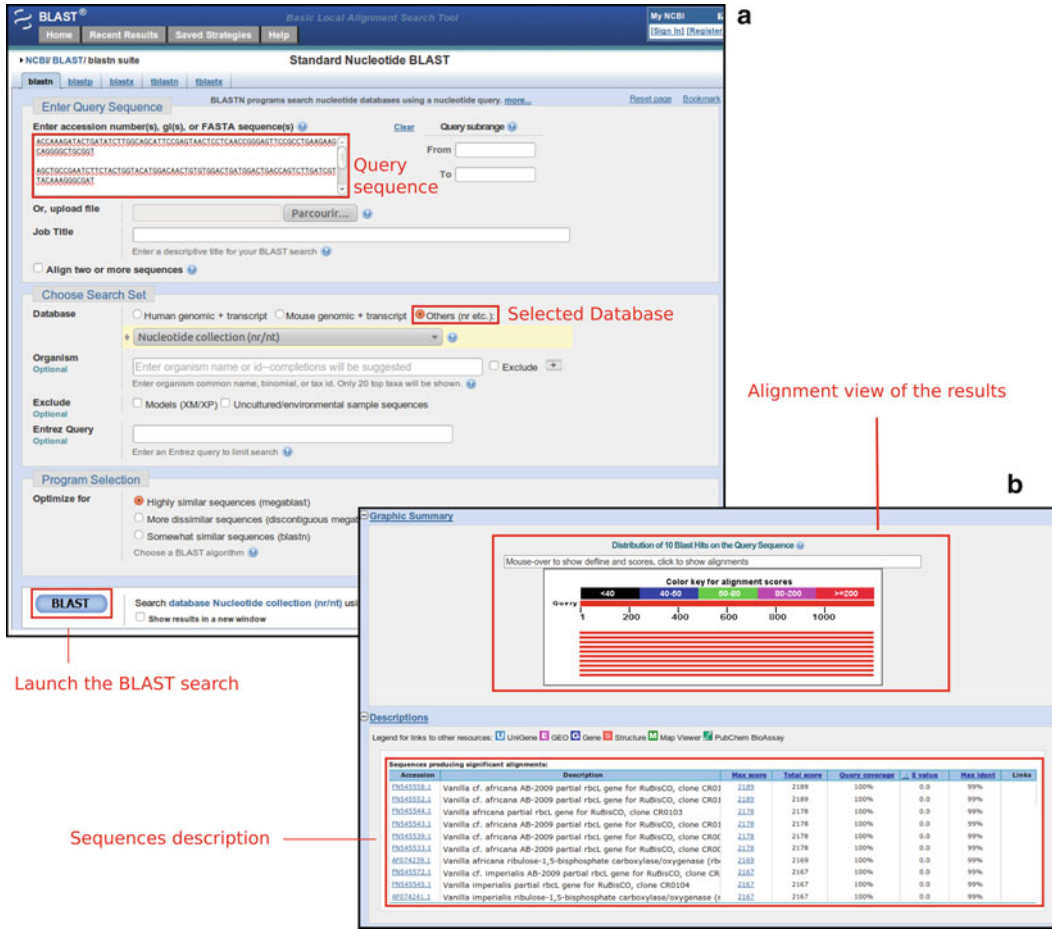


Fig. 2 Screenshots of BLAST with (**a**) search settings and (**b**) results pages

over a large enough proportion of the total query sequence length. Fixing a high level of similarity will also limit our search to sequences from more closely related organisms.

To perform the BLAST with our query sequence, open the NCBI BLAST page on a web browser and select the appropriate BLAST search, here "nucleotide BLAST" in the "basic BLAST" section. Then, copy or upload the query sequence in the "Enter Query Sequence" frame, select "others" in the database "choose search set" section, and then click on the BLAST button (Fig. 2a). The BLAST results will automatically appear (Fig. 2b), presenting the best sequence hits. An inspection of the hits informs us that, as expected, our query is highly related with *Vanilla rbc*L sequences with more than 99 % identity over 100 % of the query length, the first hit corresponding to a partial *rbc*L sequence of *Vanilla africana*. Once the BLAST search is performed, it is possible to download the hit sequences by ticking the appropriate box and clicking the "get selected sequence" button.

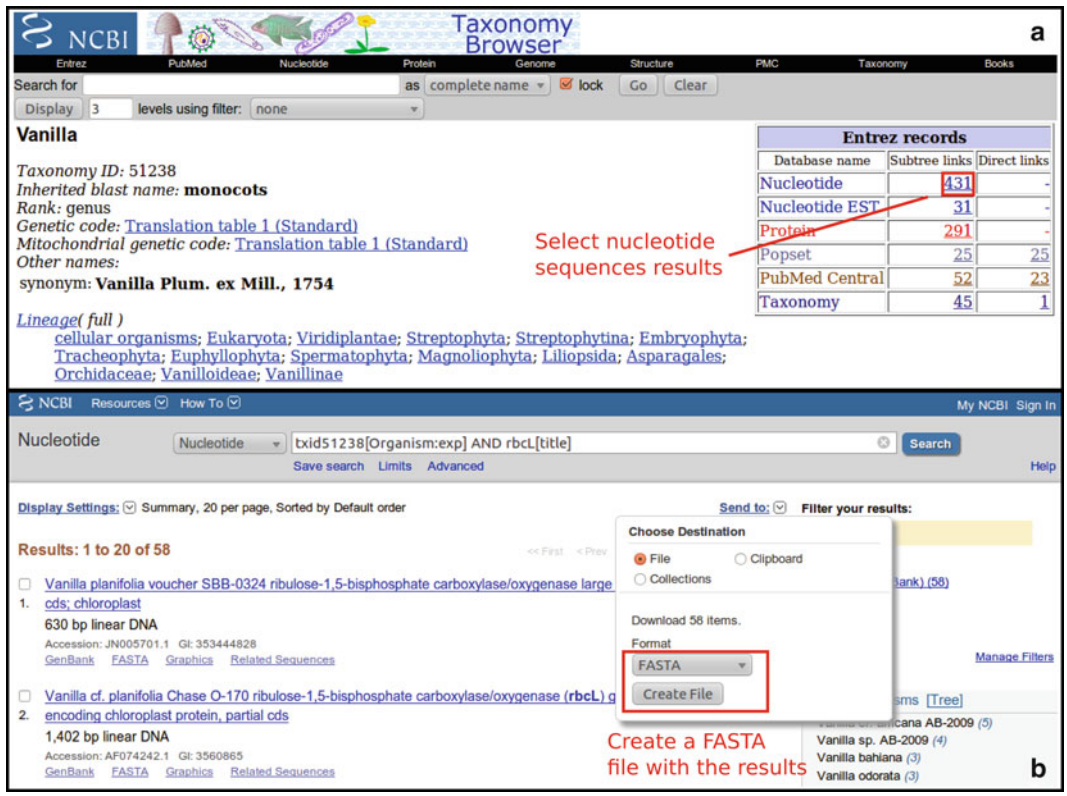


Fig. 3 (**a**) The *Vanilla* genus web page from the Taxonomy Browser Database (NCBI) and (**b**) corresponding nucleotide sequence search results

An alternative approach to obtaining sequences that could eventually be used to construct a *Vanilla* rbcL phylogenetic tree is to use the Taxonomy Browser database (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html). Typing "vanilla" in the search box will present a classification scheme of the results. A further click on the appropriate link (here the "Vanilla" link at the root of the tree diagram) will load a new page presenting all the records for this taxon (Fig. 3a). On the right side of the results page, we select "nucleotide." Note that doing so just restricts the search to the database with the additional query term "txid51238[Organism:exp]," the taxonomy identification "txid51238" being a specific identification code for the genus *Vanilla*. As we are only interested in *rbc*L sequences, we can simply add to the current query the term "AND rbcL[title]" to restrict our search to this specific gene. Beware that without providing the "[title]" designator in conjunction with the "rbcL" term, the search will return any sequence entry with the consecutive letters r-b-c and -l present in any of the sequence record fields (many of which will not be *rbc*L genes records). Note that [Organism] and [title] are two of the many "Entrez query" terms that can be used

a FASTA format

```
>S1
ACCAAAGATACTGATATCTTGGCAGCA---TTCCG--AGTAACT-CCTCAA-CCGGGAGTT-CCGCCT-
>S2
ACCAAAGATACTGATATCTTGGCAGCAAC-TTCCG-AAGTAACT-CCTTAA-CCGGGAGTTGCCGCCT-
>S3
AC--AAAGTACTGATATCTTGGCAGCAACTTTCCG-AAGTAACTCCCTTAA-CCGGGAGTTGCCGCCTT
>S4
AC--AAAGTACTGATATTCTTGCAGCAACTTTCCGAAAGTAACTCCCTTAACCCGGGAGTTGCCGCCTT
```

b Phylip format

```
4 69
S1        ACCAAAGATACTGATATCTTGGCAGCA---TTCCG--AGTAACT-CCTCAA-CCGGGAGTT-CCGCCT-
S2        ACCAAAGATACTGATATCTTGGCAGCAAC-TTCCG-AAGTAACT-CCTTAA-CCGGGAGTTGCCGCCT-
S3        AC--AAAGTACTGATATCTTGGCAGCAACTTTCCG-AAGTAACTCCCTTAA-CCGGGAGTTGCCGCCTT
S4        AC--AAAGTACTGATATTCTTGCAGCAACTTTCCGAAAGTAACTCCCTTAACCCGGGAGTTGCCGCCTT
```

Fig. 4 Examples of (**a**) FASTA and (**b**) PHYLIP formats of multiple sequence alignments containing four sequences

to filter results according to sequence record file fields (such as deposition date and country of origin) or sequence characteristics (such as length). A total of 58 appropriate *rbc*L sequences are available for download simply by clicking the "Send to" button and choosing "file" (Fig. 3b). It is highly recommended that the sequences be downloaded in FASTA format. Among the large variety of alignment formats, the FASTA format is among the simplest and most widely usable by sequence analysis software. Adding or removing a sequence entry from a FASTA file can be performed in a simple text editor by pasting in, or deleting, the name and lines of sequence for that entry (see an example of the FASTA file format with example files and in Fig. 4a).

A further consideration when assembling a dataset is whether a rooted phylogenetic tree is desired. Specifically, a phylogenetic tree can be either rooted or unrooted. Whereas in a rooted tree it is clear which direction sequences are evolving in (usually represented in a left-to-right orientation with the left-most node representing the most recent common ancestor of all the sequences in the tree; Fig. 1b), in an unrooted tree, the direction of evolution is unspecified (Fig. 1c) [21]. It is usually desirable to root phylogenetic trees, and for this reason, different rooting methods have been devised. The most common and generally reliable of these is the outlier rooting method. With this method, an "outlier sequence" is selected that (1) must be homologous to the sequences in the dataset of interest and (2) must be from an organism that is less closely related to the sequences in the dataset than any of these sequences are to one another, and (3) of all the sequences not included in the dataset, the outlier should be as closely related to the sequences in the dataset as possible [40]. For the purposes of our example analysis, we have chosen the *Pseudovanilla ponapensis* partial *rbc*L sequence (accession number AY381131) as an outlier because it meets all of these criteria.

All the sequences (the *Vanilla rbc*L sequences, the *Pseudovanilla rbc*L outlier, and the sequence we are attempting to place phylogenetically) have to be pasted into the same FASTA file. Note that the sequences in the FASTA file downloaded from NCBI each have a very long name that includes a description of the sequence. For convenience, we have used a standard text editor to crop the names to include only the accession number (the "gb field" in the sequence name; final FASTA file available with example files).

3.2 Aligning the Sequences

Multiple sequence alignment involves lining up the homologous sites of homologous nucleotide or amino acid sequences. Specifically, a sequence alignment is essentially a table, or matrix, of data within which each sequence is assigned a separate row in the matrix, with homologous nucleotide or amino acid positions in different sequences lined up into columns. When the alignment is used to construct a phylogenetic tree, the residues in each particular column are assumed to be different states of a homologous trait derived by mutation from common residue in an ancestral sequence (Fig. 1a).

Whereas modern multiple sequence alignment algorithms are fast enough to align a fairly large set of sequences (relative to the kind of analysis we describe in this chapter), it can be an extremely difficult and a time-consuming task to refine the alignments that these algorithms yield. Whereas alignment is trivial when the degree of diversity between the sequences being analyzed is low (as is the case with our *Vanilla rbc*L dataset), it gets considerably more complex as sequences get more divergent. It is very important to realize that none of the frequently used alignment programs is capable of consistently producing perfect alignments even for moderately divergent sequences. Therefore, for both easy and more complex alignments, it is always important to check by eye for obviously misaligned nucleotides and shift these back into alignment by adding and removing alignment gaps.

The MEGA5 [30] computer program implements one of the best free multiple sequence alignment editors, and it is the one that we will use here. It allows the use of two distinct automatic alignment methods, ClustalW [41] and MUSCLE [42]. Sequences can be aligned as a whole or a subset of sites can be selected. This last functionality is very useful when different parts of the sequences have different degrees of diversity (as is commonly the case when the aligned sequences include both coding and noncoding sequences).

It is worth noting that MEGA allows one to alter various alignment parameters that can in some cases improve the quality of the alignments produced by ClustalW and MUSCLE. The most noteworthy of these parameters are the gap open penalty (GOP) and gap extension penalty (GEP) and, in the case of the MUSCLE method, the "maximum iterations" setting. Alignment methods such as ClustalW and MUSCLE focus on maximizing an alignment

score where matching characters in columns are given a positive score and mismatches are given a negative score. To obtain the alignment, gaps (the "-" character) corresponding with hypothetical ancestral insertion/deletion events are introduced into the sequences at sites that maximize the alignment score. Wherever a gap is added in isolation, the alignment score is penalized by the amount specified by the GOP parameter, and when subsequent gaps are added adjacent to an existing gap, the alignment score is penalized by the amount specified by the GEP parameter. Whereas this procedure is performed in a single round with the ClustalW algorithm, multiple iterations of the same process can be performed with the MUSCLE algorithm (controlled by the maximum iterations parameter). Increasing the iteration number will theoretically improve the alignment but will have a cost in terms of speed.

However, even when multiple iterations are performed, the MUSCLE method is considerably faster than the ClustalW method. As a first alignment step, one might align a set of sequences with MUSCLE with default parameters with the maximum iterations setting between 2 and 6. Dependent on the alignment quality obtained in this first step, it is advisable that a second alignment step should be to realign particularly badly aligned regions of the alignment either with the ClustalW method or with a mixture of ClustalW and MUSCLE with GOP and GEP settings that are lower than the default values (i.e., so as to penalize alignment gaps less severely). The third step should then be to edit the alignment by eye focusing particularly on the exact placement of gap characters so as to minimize the number of mismatches in individual alignment columns.

Following this alignment procedure using MEGA, it is recommended that the final alignment be saved in FASTA format by pressing the "Data" menu option, then the "Export Alignment" menu option, and then the "FASTA format" menu option.

Open MEGA5 and click on "Align" and then "Edit/Build alignment." Choose "Retrieve sequences from a file" and select the FASTA file that you obtained in the previous exercise. You will be presented with a matrix of characters with unaligned sequences running in rows. In our example, due to a low level of variability between *rbc*L sequences, one iteration of the MUSCLE algorithm is sufficient to obtain a good alignment. Select all sequences (CTRL+A) and align them with MUSCLE by clicking on the MUSCLE icon (you can also do this via the "Alignment" menu and select the "Align by Muscle" option). Keep gap penalties with their default values, modify "Max Iterations" to one, and then click on the "Compute" button. As emphasized before, any automatically produced alignment must be checked by eye for (1) misaligned sections of sequence (i.e., incorrectly placed gap characters), (2) truncated/partial sequences, and (3) sequences that are either nonhomologous or in the incorrect orientation. Once the overall alignment is completed,

the ends of the alignment should in most cases be trimmed to the size of the sequences of interest (e.g., from the beginning of the *rbc*L start codon to the end of the *rbc*L stop codon).

Once the alignment is performed and has been properly checked, simply export the file in FASTA format. Whereas jModelTest2 [31], the program we next use, will work with many different alignment formats, PhyML3 [32], the program we use after jModelTest2, requires an alignment in PHYLIP format (see example file and Fig. 4b). To obtain a PHYLIP file, the easiest is to use the same converter used with jModelTest2. Simply run the "alter.jar" application available in the jModelTest2/lib repertory, open your last saved FASTA file by clicking on "Load…" under the input box, press the convert button, and choose the "GENERAL" option and then the "PHYLIP" option to save the alignment as a PHYLIP file.

3.3 Determining the Best-Fit Nucleotide Substitution Model

The major advantage of using ML in the context of phylogenetic inference is that it allows the construction of phylogenetic trees within a very well-defined model-based statistical framework. This framework allows one to rationally determine whether the evolution of the observed sequences is more consistent with certain evolutionary models (detailing nucleotide substitution and tree topology parameters) than it is with others. The likelihood score of a particular tree topology and nucleotide substitution model with particular nucleotide substitution parameters is calculated by multiplying the estimated probabilities of each alignment column given the model under consideration. This likelihood score is usually very small even when alignments are very short such that for convenience the likelihood is usually presented in the form of the negative of its natural logarithm (usually denoted –lnL). The smaller the –lnL value is, the more likely the model is [24].

As in every statistical analysis, the model describing the nucleotide substitution process can be either simple (all possible types of substitution are equally probable) or complex (all types of substitution occur at different rates), and the degree of complexity of the model will depend on how informative the data is. Fortunately, programs like jModelTest2 are available to compare different possible models of evolution and identify which models are best supported by the observed sequences. For each of up to 88 separate models jModelTest2, maximum likelihood scores are computed and compared. Four different model comparison procedures are implemented including likelihood ratio tests (hLRT [43]), the Akaike information criterion tests (AIC [44]), the Bayesian information criterion tests (BIC [45]), and the decision theory method (DT [46]). Each of those model selection procedures compares the likelihood of the models, taking into account the fact that in many cases different models have different numbers of free parameters. It has been determined that, in the

phylogenetic context at least, the BIC and DT tests tend to support simpler models that perform as well as more complex models selected using hLRT and AIC tests and that the BIC and DT tests should therefore be preferentially used when deciding on the best-fit model [47, 48].

The models that jModelTest2 evaluates range from the simple Jukes-Cantor 1969 model, [49] where all mutations occur at the same basal rate, to more complex general time reversible (GTR) models where every nucleotide substitution is free to occur at a distinct rate [50]. Three other parameters can increase the complexity of models including accounting for unequal base frequencies (+F), the varying proportions of invariant sites (+I), and variations in the substitution rate across sites (+G). For this last parameter, a given number of discrete rate categories are used to model variations in substitution rates among sites. The default number of substitution rate categories in PhyML3 is four, but this number can be freely changed (although it must be realized that the analysis time will increase linearly with this number).

To begin the model selection procedure, simply start the jModelTest.jar application. Load the data (the alignment in either FASTA or PHYLIP format) with file >open>select your data. The model selection procedure could be time-consuming, and depending of the amount of data, a restricted number of models can be tested. A number of substitution schemes (NSS) of three is straightforward to set up in PhyML3 for later tree reconstruction and will be chosen for the analysis we perform here. Select "+F," "+I," and "+G" before running the analysis by pressing the "Compute Likelihoods" button (Fig. 5a).

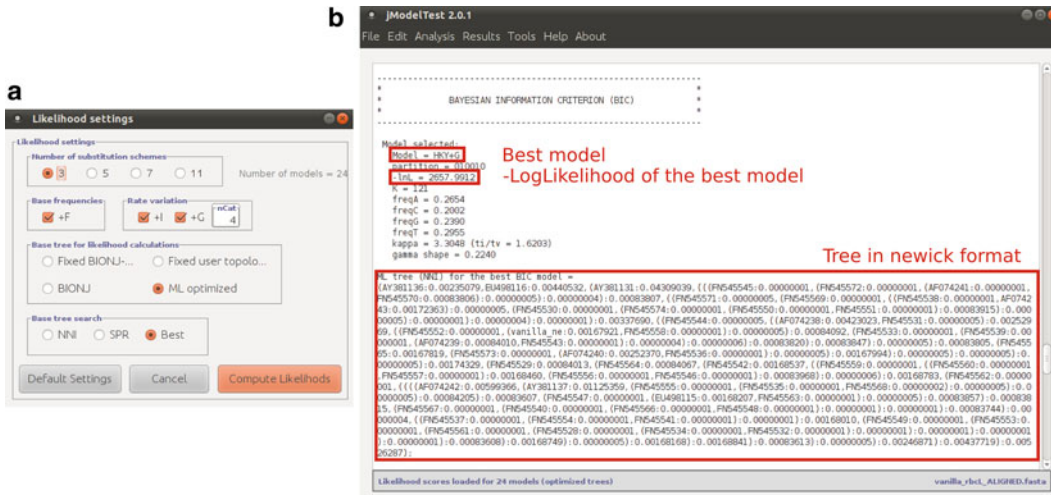


Fig. 5 (**a**) jModelTest2 settings window and (**b**) BIC result for the best-fit model with its associated tree in parenthetic format

Once the analysis is over, select either BIC or DT in the Analysis panel, leaving the confidence interval as is (this function permits one to choose a hierarchical selection of models in order to do model-averaging parameters estimation and will not be further mentioned in this chapter). After running the test, in the output the best supported model will be selected and appears first in the list. In our example, the best supported model according to the BIC test is the HKY + G model (Fig. 5b), a model for which transitions and transversions are free to occur at different rates and where basal substitution rates vary from site to site along the sequences and all bases are not assumed to occur at equal frequency [51].

3.4 Reconstructing a Phylogeny

The jModelTest2 software provides in its log file the phylogenetic tree obtained with the best-fit model (Fig. 5b). The tree is available in a parenthetic format (the Newick format) and can be directly copied from the log to any text editor before being loaded in an appropriate tree viewing program. While useful, this tree lacks any indication of how well supported individual branches within the tree are. To achieve this, it is usually desirable to perform bootstrap tests to identify the branches that are most and least supported by the available data.

To assess the robustness of the ML tree produced by jModelTest2, we will use PhyML3 to analyze the *rbc*L sequences with the HKY + G model indicated to be the best-fit model by jModelTest2. Two tests of branch support are available in PhyML3. The first is the well-established and widely used bootstrap test [52]. In this test, alignment sites (the columns) are sampled with replacement to obtain new "virtual alignments" derived from our real data. Different phylogenetic trees are then inferred for each of these resampled alignments (usually between 100 and 1,000 times alignment + tree combinations), and the percentage of times that particular groups of sequences that cluster in the tree constructed from the real sequences also cluster in the bootstrapped trees is used as a measure of robustness of the branches separating these sequences from the remainder of the tree. The statistical meaning of such bootstrap values is obscure, and it must be remembered that they are in no way "p-values." In fact it is generally accepted that branches that are supported in as few as 70 % of the bootstrap replicates should be considered to be robustly supported.

Considerably less statistically obscure is the alternative to the bootstrap test that is provided by PhyML3: the approximate likelihood ratio test (aLRT, a derivative of the LRT; [53]). This fast nonparametric test provides branch statistics that are essentially approximate *p*-values that indicate whether trees containing the branch have a significantly better likelihood than the best alternative tree topologies where the branch is absent. Besides being much faster to compute than branch supports determined by bootstrapping, branch supports determined by aLRT are also much easier to interpret.

To start PhyML3, simply double click on its icon or launch it from the command line. A screen appears asking for the alignment name. This alignment must be in PHYLIP format (obtained previously) and should be placed in the same directory as the PhyML3 program (otherwise, the full path of the alignment should be provided with the name). A menu will then appear with default parameters for the input data. Four pages of submenu, the input data, the model of substitution, the tree searching, and the branch support menu, must be checked and set to the appropriate values before running the analysis. To navigate between the sub-menus, the "+" and "-" commands are used. For each line of the menu, the value of a given entry is written on the right of the screen. To toggle the value, simply type the letter given between brackets on the left of the line until the correct parameter is set (see the squared "m" letter on the line of the model of nucleotide substitution, Fig. 6). For example, to toggle from DNA to protein, simply type "d." Another "d" will set the value back to DNA. All the default parameters are fine in this sub-menu. However, the "Run ID" option, permitting one to name the output files, can be useful to keep track of the analyses performed with different analysis settings.

In the next sub-menu, we have to set up the substitution model parameters. The best-fit evolutionary model selected with jModelTest2 was HKY + G. To set this model, select the "m" option. Next select the "r" option so as to allow rate variation across sites, and then select the "c" option and make sure that four rate categories are defined.

Once the model is set up, press "+" to reach the tree searching submenu. Here, the tree topology search operations should be set to "Best of NNI and SPR" using the "s" option. This setting is

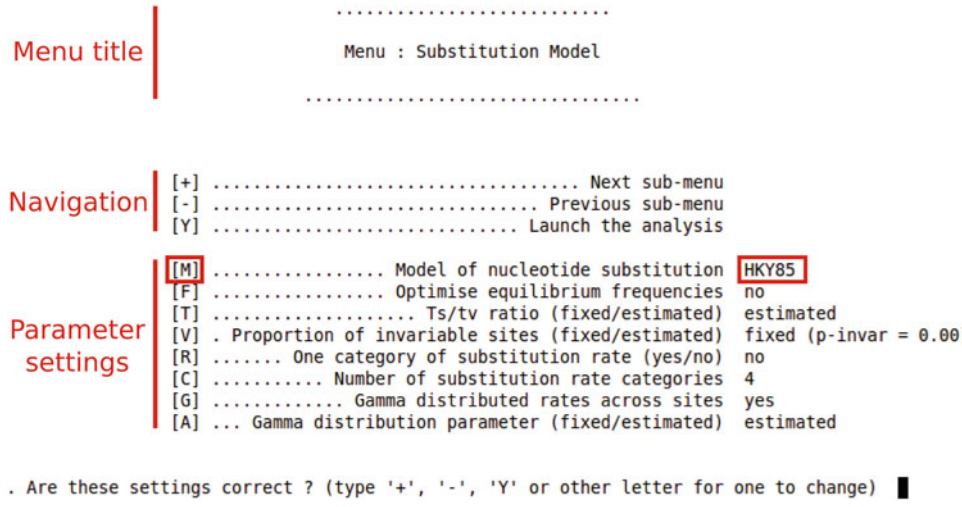


Fig. 6 PhyML substitution model submenu

more computationally intensive than the default setting ("NNI") but often ensures better results. Finally in the last submenu, the bootstrap analysis and aLRT can be turned on and off. Note that turning on the bootstrap analysis will turn off the aLRT and vice versa. Several statistics are available to infer branch confidence with the aLRT method. Documentation about these different statistics can be found in [53] and [54]. Here we will use the SH-like statistic, which have been proved to be more robust to violations of model assumptions [54]. Once every parameter is set appropriately, simply type "y" to launch the analysis.

Once the tree search begins, the program will continuously keep you updated on its progress and will eventually shut down automatically when it perceives it has found what is probably the maximum likelihood tree. This tree will be written to the file <align_name>_phyml_tree.txt<_RunID>.txt, in a parenthetic format, and the run summary statistics will be written to the file <align_name>_phyml_stats.txt<_RunID>.txt.

3.5 Visualization and Editing of Trees

The last step of the phylogeny reconstruction process is the visualization and editing of the trees that PhyML3 produces. Several programs have the ability to take a tree file in parenthetic format and plot a graphic of the tree(s) depicted in the file. The program we present here is called FigTree [33]. After starting the figtree.jar application, simply open the tree file (here, the <align_name>_phyml_tree.txt<_RunID>.txt file) with file>open. A pop-up question window will appear offering the option for you to rename the branch labels (i.e., type either "bootstrap" or "aLRT" depending on the test that was performed). Once the tree is opened, among the many options available are those allowing one to edit, color, and modify the appearance of the tree. To label branches with their bootstrap/aLRT support values, tick the "Node labels" box and select label/bootstrap/aLRT. The scale can be adjusted for convenience.

The *Vanilla* genus has been shown to segregate into three groups: α, β, and γ [38]. Our ML phylogenetic tree indicates that our new sequence sits within the γ group (colored in yellow on the tree) corresponding to Old World and Caribbean species and more precisely with *Vanilla africana* accessions (FN545552 and FN545558). This placement within the tree is consistent with the results obtained in our earlier BLAST search.

The tree presented in FigTree can be printed as is or exported in any one of several different image formats (Fig. 7). Note that the enhanced metafile (.emf) or windows metafile (.wmf) formats are good choices for a further editing of the image file in any graphics editing programs. The Scalable Vector Graphics (.svg) format is also a good choice if one would like to edit the graphics in open-source graphics editors such as Inkscape (www.inkscape.org).

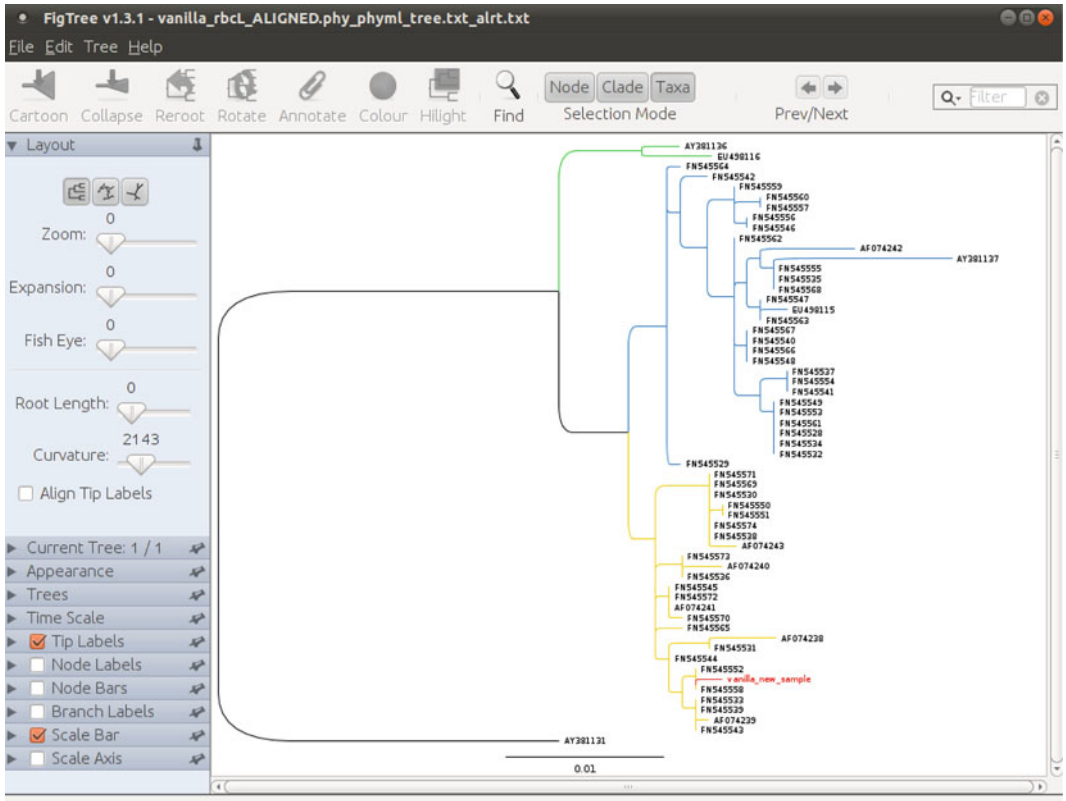


Fig. 7 FigTree representation of the phylogenetic reconstruction of *Vanilla* samples, with colors representing the three distinct clades of the genus (*green*, α; *blue*, β; *yellow*, γ [38]), and the query sequence used in our example shown in *red*. For the purpose of clarity, aLRT values are not displayed (Color figure online)

4 Notes

1. *PhyML special model*
   In the example given above, we select only one of several nucleotide substitution models available in PhyML3. Several other models exist and can be set up in PhyML3 using "custom" settings in the "Substitution model" menu. The model is then set using a binary representation of the substitution matrix. Information on how these models are implemented can be found in both the jModelTest and PhyML manuals.
2. *Other software with (nearly) the full procedure implemented*
   The phylogenetic reconstruction procedure described here is implemented in part or whole in various other computer programs that the user may want to use. Usually these programs still employ the programs we present here or use a similar analysis pipeline. It is important to keep in mind that the software landscape is moving rapidly and faster than the methods.

3. *Amino acid substitution model*
   While we present our phylogenetic reconstruction using nucleotide sequences, it is equally applicable to amino acid sequences. Obviously in the case of amino acid sequences, amino acid and not nucleotide substitution models have to be chosen. Amino acid substitution model selection can be performed using the ProtTest3 program (http://darwin.uvigo.es/software/prottest3/prottest3.html; [55]), designed by the same group that produces jModelTest2.
4. *Recombination*
   As different portions of an organism's genome can have different origins, recombination is a major issue for phylogenetic reconstruction [56]. The evolutionary history of a recombinant sequence cannot usually be depicted by a single phylogenetic tree (the regions of the recombinant genome derived from its two parents should be described by different phylogenetic trees), and attempting to explain the evolutionary history of such sequences with a single tree can be very misleading [56, 57]. This should be borne in mind before any phylogenetic analysis and especially when one chooses the type of data or genetic locus to analyze.
5. *Bayesian phylogenetic inference*
   Another excellent approach for phylogenetic reconstruction relies on Bayesian inference and is implemented in programs such as MrBayes3 (http://mrbayes.sourceforge.net/; [58]) and BEAST (http://beast.bio.ed.ac.uk/Main_Page; [59]). ML and Bayesian inference are known to perform similarly, with respect to determining the true phylogeny underlying the evolution of a set of sequences [60, 61]. Rather than focusing on the inference of a single "best" tree, Bayesian inference focuses on the identification of groups of similarly plausible phylogenetic trees. In this regard, Bayesian inference of phylogenies is favored in applications where one would like to account for phylogenetic uncertainty while inferring, for example, ancestral sequences or sites evolving under natural selection.

Acknowledgments

ADB is supported by the Conseil Général de La Réunion and CIRAD. DPM is supported by the Wellcome Trust. PL is supported by CIRAD and Conseil Régional de La Réunion and European Union (FEDER). The authors wish to thank Dr. Jean-Michel Lett for his helpful comments.

References

1. Darlu P, Tassy P (1993) La reconstruction phylogénétique. Concepts et Méthodes. Masson
2. Groves C (1986) Systematics of the great apes. In: Swindler DR, Erwin J (eds) Comparative primate biology: systematics, evolution and anatomy, vol 1. Liss AR, New York, pp 187–217
3. Hemsley AR, Poole I (2004) The evolution of plant physiology. From whole plants to ecosystems. Elsevier Academic Press, Amsterdam
4. Caputo P (1997) DNA and phylogeny in plants: history and new perspectives. Lagascalia 19:331–344
5. Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366
6. Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York
7. Van de Peer Y (2009) Phylogeny inference based on distance methods. In: Salemmi M, Vandamme AM (eds) The phylogenetic handbook, a practical approach to DNA and protein phylogeny. Cambridge University Press, New York, pp 101–135
8. Michener CD, Sokal RR (1956) A quantitative approach to a problem in classification. Evolution 11:130–162
9. Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284
10. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
11. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685–695
12. Steel MA, Hendy MD, Penny D (1988) Loss of information in genetic distances. Nature 336:118
13. Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland
14. Sober E (1988) Reconstructing the past: parsimony, evolution, and inference. MIT Press, Cambridge
15. Edwards AWF, Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees. In: Heywood VH, McNeill J (eds) Phenetic and phylogenetic classification: a symposium. Systematics Association, London, pp 67–76
16. Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 32:550–570
17. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
18. Farris JS (1970) Methods for computing Wagner trees. Syst Zool 19:83–92
19. Fitch WM (1971) Towards defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416
20. Kluge AG, Farris JS (1969) Quantitative phyletics and the evolution of anurans. Syst Zool 18:1–32
21. Harrison CJ, Langdale JA (2006) A step by step guide to phylogeny reconstruction. Plant J 45:561–572
22. Aldrich J (1997) R. A. Fisher and the making of maximum likelihood 1912–1922. Statist Sci 12:162–176
23. Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25:471–492
24. Schmidt HA, von Haeseler A (2009) Phylogenetic inference using maximum likelihood methods. In: Salemmi M, Vandamme AM (eds) The phylogenetic handbook, a practical approach to DNA and protein phylogeny. Cambridge University Press, New York, pp 181–209
25. Hendy MD, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59:277–290
26. Swofford DL, Sullivan J (2003) Phylogeny inference based on parsimony and other methods using Paup*. In: Salemmi M, Vandamme AM (eds) The phylogenetic handbook, a practical approach to DNA and protein phylogeny. Cambridge University Press, New York, pp 267–312
27. Swofford DL, Olsen GJ (1990) Phylogeny reconstruction. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer Associates, Sunderland, pp 411–501
28. Swofford DL et al (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer Associates, Sunderland, pp 407–514
29. Ronquist F, van der Mark P, Huelsenbeck JP (2009) Bayesian phylogenetic analysis using MrBayes. In: Salemmi M, Vandamme AM (eds) The phylogenetic handbook, a practical approach to DNA and protein phylogeny. Cambridge University Press, New York, pp 210–266
30. Tamura K et al (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
31. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25: 1253–1256
32. Guindon S et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

33. Morariu V et al (2008) Automatic online tuning for fast Gaussian summation. Advances in Neural Information Processing Systems (NIPS) 1–8
34. Hall BG (2007) Phylogenetic trees made easy: a how-to manual, 3rd edn. Sinauer Associates, Sunderland
35. Benson DA et al (1994) GenBank. Nucleic Acids Res 22:3441–3444
36. Cochrane G et al (2009) Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res 37:D19–D25
37. Tateno Y et al (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 30:27–30
38. Bouetard A et al (2010) Evidence of transoceanic dispersion of the genus *Vanilla* based on plastid DNA phylogenetic analysis. Mol Phyl Evol 55:621–630
39. Altschul SF et al (1990) Basic local alignment tool. J Mol Biol 215:403–410
40. Maddison WP, Donoghue MJ, Maddison DR (1984) Outgroup analysis and parsimony. Syst Zool 33:83–103
41. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680
42. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797
43. Posada D, Crandall KA (1998) Model test: testing the model of substitution. Bioinformatics 14:817–818
44. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723
45. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
46. Minin V et al (2003) Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52:674–683
47. Luo A et al (2010) Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. BMC Evol Biol 10:242
48. Ripplinger J, Sullivan J (2008) Does choice in model selection affect maximum likelihood analysis? Syst Biol 57:76–85
49. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic, New York, pp 21–132
50. Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci (Am Math Soc) 17:57–86
51. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174
52. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791
53. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol 55: 539–552
54. Anisimova M et al (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol 60:6 85–699
55. Darriba D et al (2011) ProtTest3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165
56. Ruths D, Nakhleh L (2005) Recombination and phylogeny: effects and detection. Int J Bioinform Res Appl 1:202–212
57. Posada D, Crandall KA (2002) The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 54:396–402
58. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574
59. Drummond AJ et al (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973. doi:10.1093/molbev/mss075
60. Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304–311
61. Mau B, Newton M, Larget B (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12

Chapter 14

The Application of Flow Cytometry for Estimating Genome Size and Ploidy Level in Plants

Jaume Pellicer and Ilia J. Leitch

Abstract

Over the years, the amount of DNA in a nucleus (genome size) has been estimated using a variety of methods, but increasingly, flow cytometry (FCM) has become the method of choice. The popularity of this technique lies in the ease of sample preparation and in the large number of particles (i.e., nuclei) that can be analyzed in a very short period of time. This chapter presents a step-by-step guide to estimating the nuclear DNA content of plant nuclei using FCM. Attempting to serve as a tool for daily laboratory practice, we list, in detail, the equipment required, specific reagents, and buffers needed, as well as the most frequently used protocols to carry out nuclei isolation. In addition, solutions to the most common problems that users may encounter when working with plant material and troubleshooting advice are provided. Finally, information about the correct terminology to use and the importance of obtaining chromosome counts to avoid cytological misinterpretations of the FCM data are discussed.

Key words Chromosome number, DAPI, DNA ploidy level, Genome size, Flow cytometry, Flow histogram, *C*-value, PI, Plant nuclei isolation, Relative fluorescence

1 Introduction

The total amount of DNA in the nucleus of an organism is generally referred to as the genome size, and it is measured either in picograms (pg; i.e., 1×10^{-9} g) or megabase pairs (Mbp, with 1 pg = 978 Mbp, [1]). People started to investigate genome size in plants even before the structure of DNA was worked out, with the first plant to have its genome size estimated being *Lilium longiflorum* in 1951 [2]. Since then the genome sizes of over 8,500 species have been estimated [3] with the data being used not only for practical applications (e.g., how much will it cost to sequence a genome? how many clones are needed for making BAC libraries?) but also for providing valuable insights into many biological fields, including evolution, systematics, ecology, population genetics, and plant breeding (*reviewed in* refs. [4–7]).

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_14,

Over the years, several methods have been used to estimate genome sizes in plants (e.g., Feulgen densitometry, reassociation kinetics). Nevertheless, in recent years, due to a variety of reasons, flow cytometry (FCM) has become the method of choice [8]. Briefly the method involves three steps: (1) a sample of plant tissue is chopped in a suitable buffer to release the nuclei; (2) the nuclei are stained with a fluorochrome that binds quantitatively to the DNA, so the bigger the genome, the more stain that is bound to the DNA; and (3) the nuclei are passed through a flow cytometer which measures the amount of stain bound to each nucleus (Fig. 1a). By preparing a combined sample which includes a plant species with a known DNA amount (reference standard), the

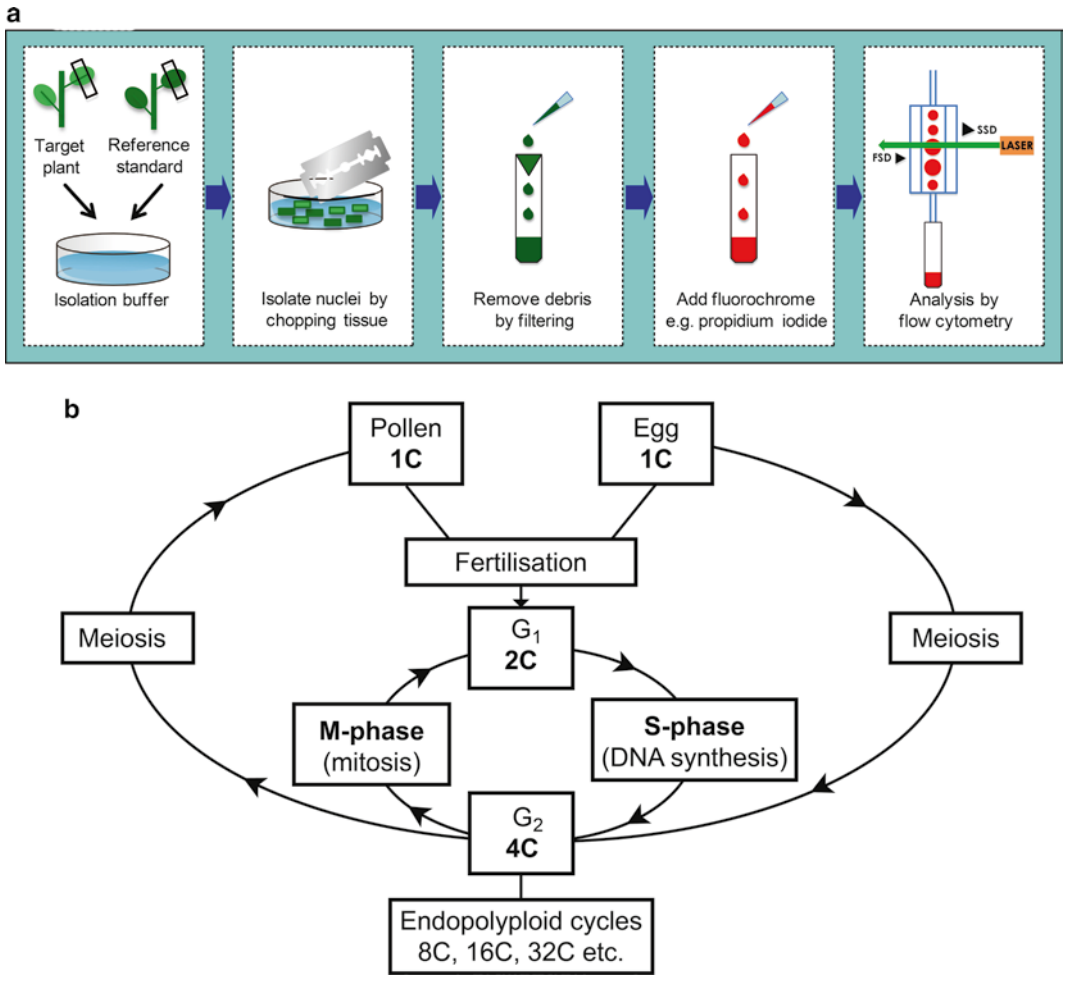


Fig. 1 (**a**) The basic steps involved in the estimation of genome size and ploidy level by flow cytometry. (**b**) Changes in the holoploid *C*-value at different stages of the cell cycle and following meiosis and endopolyploidy (N.B. cells which undergo endopolyploidy (i.e., DNA synthesis not accompanied by mitosis) will have *C*-values greater than 4*C* (i.e., 8*C*, 16*C*, 32*C*, etc., depending on the number of rounds of DNA replication))

relative amount of fluorescence from the target plant can be converted into an absolute genome size. It is important to realize that FCM only gives information about the relative or absolute DNA amount of the isolated nuclei; it does not provide cytological information. Yet without such information, interpretations of the chromosome number and/or ploidy level of the species can be flawed. Subheading 1.2 below highlights the importance of obtaining such cytological data and the pitfalls and errors that can arise without it.

FCM can also be used to estimate the ploidy level of a plant based on comparing the genome size of the target species either with the genome size of a specimen of known ploidy (i.e., determined karyologically) or with an internal standard (in that case the reference standard must be kept constant throughout the experiment and the ploidy level of at least one target sample should be karyologically determined). However, in such cases, the ploidy level is referred to as the "DNA ploidy" to distinguish it from studies where ploidy level has been determined karyologically [9]. Such approaches are now being increasingly used to survey the diversity of cytotypes across plant populations and have uncovered a surprising diversity of hitherto unsuspected ploidy variation in some species [4–6].

This chapter outlines the general method used to estimate genome size and/or determine the DNA ploidy level in plants using FCM (Subheadings 2, 3, and 4). However, given the immense diversity of plants in terms of their morphology (e.g., woody, succulent, herbaceous) and biochemistry (e.g., presence of pigments, tannins, phenolics), problems may well be encountered. The majority of these arise mainly from the interaction between chemicals present in the cell cytoplasm and the binding of the fluorochrome to the DNA [10–16] leading to erroneous results. Thus, this chapter also outlines some of the more commonly encountered problems and ways in which the poor results might be improved to overcome these issues.

In addition to the information given here, it is also suggested that the FLOWer database [17, 18] is consulted as this provides an extensive list of papers which have used FCM to estimate genome size in plants. It is worth checking whether the particular genus of interest has previously been studied by FCM and, if so, whether any particular modifications were made to the buffers and protocols used, to overcome specific problems associated with the particular genus being analyzed. In addition, genome size databases may also be useful to check in order to get some idea about the range of genome sizes one might expect for a given taxa. Examples include the Plant DNA C-values Database [3] which contains data for all the major groups of land plants and three algal lineages, while the more focused database containing genome size data for Asteraceae (GSAD—19) is ideal for specific studies focused on this family of angiosperms. Such prior information can save a lot of time and frustration!

1.1 Terminology Used for Genome Size Studies

Given that the amount of DNA varies throughout the cell cycle (i.e., G_2 nuclei have twice the DNA amount as G_1 nuclei) (Fig. 1b), and following meiosis and endopolyploidy (somatic polyploidy), considerable confusion can arise when discussing genome sizes. To overcome such issues, Greilhuber et al. [20] proposed the following terminology which has now been widely adopted.

1. *Holoploid 1C-value* (abbreviated to 1*C*-value) refers to the amount of DNA in the unreplicated gametic nucleus (e.g., pollen or egg cell of angiosperms) regardless of the ploidy level of the cell. The 2*C*-value represents the amount of DNA in a somatic cell at the G_1 stage of the cell cycle, while the 4*C*-value is the amount in a somatic cell at the G_2 stage, following DNA synthesis (*S*-phase) (*see* Fig. 1b).
2. *Monoploid 1C-value* (abbreviated to 1*Cx*-value) refers to the amount of DNA in the unreplicated monoploid (x) chromosome set. For a diploid organism where $2n = 2x$, the 1*C* and 1*Cx* values are the same; however, for a polyploid organism, the 1*Cx* is always smaller than the 1*C*-value (e.g., for a tetraploid where $2n = 4x$, then $1Cx = 1/2\ 1C$, whereas for a hexaploid where $2n = 6x$, then $1Cx = 1/3\ 1C$).

1.2 The Importance of Cytological Data for Genome Size Studies

As noted above, FCM only measures the total amount of DNA in the nucleus and gives no specific information about the chromosome number or ploidy level of the plant analyzed (although this can be deduced in certain cases as outlined in Subheading 3.2.3). Despite this, many studies report a ploidy level or chromosome number for the analyzed plant which has either been taken from the literature or based on comparisons of DNA amounts found in related species. This is fine in a stable cytological system where there is little variation in chromosome number and size between species. However, in plants, such situations are probably the exception rather than the rule, even between closely related taxa, as many genera show considerable cytological diversity—for example, (1) polyploidy, both within and between species, is frequent, (2) large divergences in genome size among closely related species with the same ploidy have been reported, and (3) increases in ploidy level or chromosome number are not necessarily accompanied by proportional changes in DNA amount.

Examples of problems and misinterpretations of genome size that can arise through assuming the ploidy level and/or chromosome number of a species have been discussed by Suda et al. [9]. Below are a few examples to illustrate the pitfalls that can arise when karyological information is not obtained in parallel with genome size data.

1. In species with a constant chromosome number but a big range in size, an absence of chromosome data could lead to the erroneous suggestion that polyploids may be present to explain the large range of genome sizes encountered. This is illustrated by the genus *Cypripedium* (Orchidaceae) where

most species have a chromosome count of $2n = 20$ but genome size has been shown to vary over tenfold between species ($1C = 4.1$–43.1 pg) [21] (*see* Fig. 2a, b). A similar situation has also been reported in the genus *Artemisia* (Asteraceae), where the diploid species *A. annua* with $2n = 18$ chromosomes has a $1C$-value of 1.75 pg [22], while *A. leucodes*, with the same chromosome number, has a $1C = 7.70$ pg [23]. Without doing a chromosome count, one could easily assume that *A. leucodes* was a polyploid given such differences in genome size.

2. Erroneous assumptions of ploidy level in a studied species can arise when increases in chromosome number via polyploidy have not been accompanied by proportional increases in genome size. This is likely to be a common problem since genome downsizing following polyploidy is frequently encountered in angiosperms [24]. In extreme cases, species with higher ploidy levels may have the same or lower genome sizes than related species of lower ploidy. An example of this is provided by the genus *Physaria* (Brassicaceae) where the high

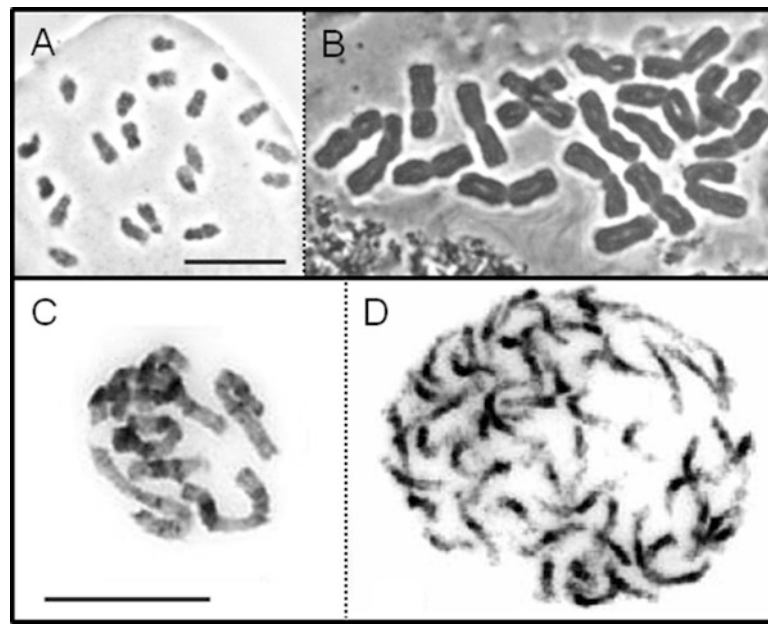


Fig. 2 Examples where chromosome size and number in related species do not correlate with genome size. Chromosomes of (**a**) *Cypripedium molle* ($1C = 4.1$ pg) and (**b**) *C. calceolus* ($1C = 32.4$ pg) taken at the same magnification showing an eightfold range in genome size but a constant chromosome number of $2n = 20$. Image reproduced with permission from ref. 21 (scale bar = 10 μm). Chromosomes of (**c**) diploid *Physaria bellii* ($2n = 2x = 8$; $1C = 2.34$ pg) compared with those from (**d**) the high polyploid *P. didymocarpa* ($2n = 14x = 56$; $1C$ 2.23 pg). Both species have similar genome sizes but very different chromosome numbers and sizes. Image reproduced with permission from ref. 25 (scale bar = 5 μm)

polyploid *P. didymocarpa* with $2n = 14x = 56$ actually has a smaller genome ($1C = 2.23$ pg) than a related diploid *P. bellii* ($2n = 2x = 8$) with $1C = 2.34$ pg [25] (Fig. 2c, d).

3. Nonproportional changes in DNA content have also been reported in different cytotypes of the same species. Once again, this has the potential to lead to erroneous deductions of ploidy level based on genome size data alone. Both increases and decreases in the size of monoploid genomes have been reported with increasing ploidy levels. For example, in *Larrea tridentata* (Zygophyllaceae), the hexaploid cytotype was reported to have just 1.25 times more DNA than the tetraploid [26]. Without chromosome data to support this, it is possible that the hexaploid could have been misidentified as a pentaploid, based on DNA amount alone.

In addition to these examples, it is also important to note that many chromosomal changes and variations (e.g., aneuploidy, chromosome duplications and deletions, sex and supernumerary chromosomes, supernumerary segments) can arise which are detectable as changes in DNA amount. Without identifying these through cytological analysis, further misinterpretations of the data may arise.

Overall, these examples serve to illustrate how serious mistakes can be made in the absence of karyological information. Thus, it is strongly recommended that chromosome counts are made of the plant used for genome size estimation. If this is not possible, then the ploidy level should always be referred to as the "DNA ploidy level" as discussed by Suda et al. [9].

2 Materials

Detailed information about plant tissues, reagents, composition of the isolation buffers, as well as the technical equipment needed to carry out genome size and ploidy estimations using FCM are described below.

2.1 Plant Tissue and Reference Standards

Of the potential plant tissues suitable for genome size estimation, leaf tissue is preferred by researchers because it generally gives the best results. Nevertheless, other plant tissues such as petals, stems (including petioles), roots, pollen grains (including pollinia), and seeds (dried or fresh) [27–29] can be considered as viable alternatives for genome size estimations. When fresh plant tissues are selected, they should be as fresh as possible and collected from young and actively growing parts of the plant as such material is likely to give the best results. Old and senescent tissues will probably result in higher levels of background signal and may contain high proportions of nuclei at the G_2 phase of mitosis.

In addition, silica-dried leaves and herbarium vouchers may be used to estimate DNA ploidy levels [30]. However, given that

DNA deterioration is likely to occur in such samples, the material is not considered suitable for high-quality estimations of genome size in absolute units. Nevertheless, recent studies have suggested that maybe even dried material can be used for genome size estimations if it has been appropriately desiccated [31].

As an alternative to desiccation, a recent publication has demonstrated the suitability of glycerol-preserved nuclei for estimating genome size in absolute units for material up to at least a few weeks old [32]. This method has been designed for field research, and although it still has a few limitations (i.e., only high-quality results are obtained when samples are kept in ice-cold buffer), it demonstrates the efforts that researchers in this discipline may go to in order to overcome problems associated with the current limited time scale available to analyze large batches of fresh material without compromising quality of the results.

Concerning reference standards, we recommend that several species, covering a broad range of genome sizes, are kept growing in the laboratory to enable the most appropriate standard to be selected for each particular analysis. Many species have been used [17] but we summarize some of the most popular ones in Table 1 which work well with FCM.

Table 1
Several reference standard species recommended for genome size estimation

Plant species	1*C* DNA content (pg)	Reference
Oryza sativa L. "IR-36"	0.50	[49]
Raphanus sativus L. "Saxa"	0.55	[50]
Solanum lycopersicum L. "Stupiké polní rané"	0.98	[50]
Vigna radiata (L.) R.Wilczek "Berken"	1.20	[49]
Glycine max Merr. "Polanka"	1.25	[51]
Petunia hybrida Vilm. "PxPc6"	1.42	[52]
Petroselinum crispum (Mill.) Nyman ex A.W.Hill "Champion Moss Curled"	2.22	[53]
Zea mays L. "CE-777"	2.71	[54]
Pisum sativum L. "Express Long"	4.18	[52]
Pisum sativum L. "Ctirad"	4.54	[55]
Pisum sativum L. "Minerva Maple"	4.86	[49]
Secale cereale L. "Daňkovské"	8.09	[55]
Vicia faba L. "Inovec"	13.45	[50]
Allium cepa L. "Ailsa Craig"	16.77	[49]
Allium cepa L. "Alice"	17.42	[55]

2.2 Equipment Needed

1. Set of pipettes with disposable tips (100 μL, 1 mL).
2. Razor blades (double-edged) or scalpel with replaceable blades. A razor blade holder or alternative protective device (e.g., cork or silicon bung) is also recommended.
3. Plastic petri dishes (c. 5–6 cm diameter).
4. Disposable nylon mesh filters (30–42 μm pore size; e.g., Partec, cat. no. 04-0042-2316). Alternatively, regular nylon mesh cut into squares and fitted on disposable tips can be used.
5. Sample tubes suitable for the particular flow cytometer being used (check manufacturer's specifications in each case).
6. 1.5 mL tubes.
7. Sample tube racks.
8. Plastic and/or expanded polystyrene containers to fill with ice.
9. Latex, nitrile, or vinyl gloves. Safety goggles and lab coat.
10. Centrifuge fitted with a rotor suitable for 1.5 mL tubes.
11. Fridge and freezer.
12. Flow cytometer fitted with the light source suitable for excitation of the DNA fluorochrome used in the study (check fluorochrome's excitation and emission spectra to select the suitable excitation sources following the manufacturer's recommendations).
13. Analytical software for evaluation of flow cytometric data (usually provided by the manufacturer of the flow cytometer).
14. Fume cupboard to carry out nuclei isolation using buffers supplemented with either β-mercaptoethanol or DTT (*see* **Note 1**).
15. Cleaning and decontamination solutions for flow systems. Domestic sodium hypochlorite (bleach) diluted 1:5 in distilled water.
16. Calibration particles: fluorescent beads [e.g., Partec, cat. no. 05-4006 (green), 05-4020 (UV)].

2.3 Reagents

2.3.1 Fluorochromes

1. PI (propidium iodide—*see* **Note 2**): Prepare a stock solution of 1 mg/mL and filter through a 0.22 μm filter. Store in 1 mL aliquots at −20 °C (*see* **Note 1**). The working concentration of PI is usually 50 μg/mL.
2. DAPI (4′,6-diamidino-2-phenylindole—*see* **Note 2**): Prepare stock solution of 0.1 mg/mL and filter through a 0.22 μm filter. Store in 1 mL aliquots at −20 °C (*see* **Note 1**). The working concentration of DAPI is normally 4 μg/mL.
3. SYBR Green I (*see* **Note 2**): The stock solution provided by the manufacturer is usually 10,000× concentrate, and manufacturers recommend a working concentration of 10×. The stock should first be diluted 100-fold in DMSO (dimethyl sulfoxide—*see* **Note 1**) to give a diluted solution of 100× (e.g.,

50 μL SYBR I in 4.95 mL of DMSO). This 100× solution can be stored in 5 mL aliquots at −20 °C. For use, the appropriate volume of this diluted 100× solution is added to the nuclei isolation buffer to give a final working concentration of 10×.

2.3.2 Isolation Buffers

Isolation buffers must be prepared using either single- or double-distilled water, filtered through a 0.22 μm filter to remove suspended particles, and stored as specified. Most of the buffers remain stable for up to 3 months if appropriately stored (*see* **Notes 3** and **4**). As indicated below, some buffers can be stored for longer by freezing them in aliquots at −20 °C. However, if this is done, then once thawed, the buffer should not be refrozen. The pH of the buffers is adjusted either with 1 M NaOH or 1 N HCl (*see* **Note 5**). Further information about the roles of the different buffer components is given in **Notes 6** and **7**, while additional details of the protocols can be found in the original references cited for each buffer.

1. *LB01 buffer* [33]: 15 mM Tris, 2 mM Na_2EDTA, 0.5 mM spermine.4HCl, 80 mM KCl, 20 mM NaCl, 0.1 % (v/v) Triton X-100. Adjust to pH 7.5. Add β-mercaptoethanol to give a final concentration of 15 mM (*see* **Note 1**). Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.
2. *Tris $MgCl_2$ buffer* [34]: 200 mM Tris, 4 mM $MgCl_2$, 0.5 % (v/v) Triton X-100. Adjust pH to 7.5 and store at 4 °C.
3. *Galbraith buffer* [35]: 45 mM $MgCl_2$, 20 mM MOPS (*see* **Note 1**), 30 mM sodium citrate, 0.1 % (v/v) Triton X-100. Adjust pH to 7.0. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.
4. *General purpose buffer* [36]: 0.5 mM spermine.4HCl, 30 mM sodium citrate, 20 mM MOPS (*see* **Note 1**), 80 mM KCl, 20 mM NaCl, 0.5 % (v/v) Triton X-100. Adjust to pH 7.0. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.
5. *Woody plant buffer* [36]: 200 mM Tris, 4 mM $MgCl_2$, 2 mM Na_2EDTA, 86 mM NaCl, 10 mM sodium metabisulfite, 1 % PVP-10 (*see* **Note 7**), 1 % (v/v) Triton X-100. Adjust to pH 7.5. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.
6. *$MgSO_4$ buffer* [37]: 9.53 mM $MgSO_4$, 47.67 mM KCl, 4.77 mM HEPES, 6.48 mM DTT (*see* **Note 1**), 0.25 % (v/v) Triton X-100. Adjust to pH 8.0. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.
7. *Bino's buffer* [38]: 200 mM mannitol, 10 mM MOPS (*see* **Note 1**), 0.05 % (v/v) Triton X-100, 10 mM KCl, 10 mM NaCl, 2.5 mM DTT (*see* **Note 1**), 10 mM spermine.4HCl, 2.5 mM Na_2EDTA, 0.05 % (w/v) sodium azide (*see* **Note 1**). Adjust to pH 5.8 and store at 4 °C.

8. *De Laat's buffer* [39]: 15 mM HEPES, 1 mM Na_2EDTA, 0.2 % (v/v) Triton X-100, 80 mM KCl, 20 mM NaCl, 15 mM DTT (*see* **Note 1**), 0.5 mM spermine.4HCl, 300 mM sucrose. Adjust to pH 7.0 and store at 4 °C.
9. *Ebihara's buffer* [40]: 50 mM Na_2SO_3, 50 mM Tris, 40 mg/mL PVP-40 (*see* **Note 7**), 140 mM β-mercaptoethanol (*see* **Note 1**). Adjust to pH 7.5 and store at 4 °C.
10. *Seed buffer* [41]: 5 mM $MgCl_2$, 85 mM NaCl, 100 mM Tris, 0.1 % Triton X-100. Adjust to pH 7.0 and store at 4 °C (*see* **Note 8**).
11. *Otto buffer* [42]: Otto I: 100 mM citric acid monohydrate, 0.5 % (v/v) Tween 20 (*see* **Note 9**). Store at 4 °C. Otto II: 400 mM Na_2HPO_4 (*see* **Note 10**). Store at room temperature.
    The fluorochrome (DAPI or PI; see above) can be added to Otto II before adjusting the final volume of the stock solution. If this is done, the buffer should be stored in the dark at room temperature. Alternatively the fluorochrome can be added directly to the sample at **step 10** of Subheading 3.1.2 or **step 7** of Subheading 3.1.3.
12. *Baranyi's buffer* [43]: Baranyi solution I: 100 mM citric acid monohydrate, 0.5 % (v/v) Triton X-100. Store at 4 °C.
    Baranyi solution II: 400 mM Na_2HPO_4, 10 mM sodium citrate, 25 mM sodium sulfate. Store at room temperature.
    The fluorochrome (DAPI or PI; see above) can be added to Baranyi solution II before adjusting the final volume of the stock solution. If this is done, the buffer should be stored in the dark at room temperature. Alternatively the fluorochrome can be added directly to the sample at **step 10** of Subheading 3.1.2 or **step 7** of Subheading 3.1.3.
13. *Mishiba's buffer* [44]: Solution A: (see recipe for Galbraith buffer, i.e., buffer 3 above).
    Solution B: 10 mM Tris, 50 mM sodium citrate, 2 mM $MgCl_2$, 1 % PVP-40 (original recipe used PVP K-30—*see* **Note 7**), 0.1 % (v/v) Triton X-100, 18 mM β-mercaptoethanol (*see* **Note 1**). Adjust to pH 7.5. Store at 4 °C.

3 Methods

3.1 Isolation of Plant Nuclei

Nuclei suspensions can be prepared according to either the one-step protocol (Subheading 3.1.1) or the two-step protocol (Subheading 3.1.2). The one-step protocol works with most plant species and with buffers 1–10 (Subheading 2.3.2). However, for some plant groups, the two-step protocol using buffers 11 or 12 (Subheading 2.3.2) will provide histograms with much higher-quality peaks. A simplified version of the two-step protocol using buffers 11, 12, and 13 (Subheading 2.3.2) is given in Subheading 3.1.3.

We recommend (unless specified otherwise) working under cold conditions (i.e., keep all solutions, buffers, and prepared samples waiting for analysis on ice, and do the chopping step in a petri dish resting on a bed of ice). Together this helps to inhibit the negative effect of many cytosolic compounds that may be present (e.g., DNase, phenolics, tannins), and it can be especially helpful when working with recalcitrant samples.

3.1.1 Isolation of Plant Nuclei Using the One-Step Protocol

1. Place a small amount of the selected plant tissue (usually about 1 cm^2 or 20 mg) in a 6 cm petri dish (*see* **Note 11**).
2. Add 1 mL of ice-cold isolation buffer (*see* Subheading 2.3.2, buffers 1–10) to the petri dish (*see* **Note 12**).
3. Chop the tissues in the buffer using a new razor blade or sharp scalpel (*see* **Note 13**).
4. Add another 1 mL of the same ice-cold buffer as used in **step 2** (*see* **Note 14**).
5. Mix the crude suspension by gently shaking the petri dish.
6. Filter the homogenate through a 30–42 μm nylon mesh filter into a labelled flow cytometry tube (*see* **Note 15**). The chopping and filtration processes might result in a reduction in the final volume, especially when working with dried samples. To reduce any critical effect, (1) dried samples can be presoaked in buffer for up to 5–10 min, and (2) filters can also be soaked in buffer prior to filtration.
7. Add the appropriate volume of fluorochrome (Subheading 2.3.1) to the nuclei suspension and vortex gently. For a typical sample which is c. 2 mL, the amount of stock PI added is 100 μL, while for DAPI, 80 μL should be added (*see* **Notes 16** and **17**).
8. Keep samples on ice until ready to be analyzed (*see* **Note 18**).
9. Proceed to analyze the nuclear DNA content, vortexing the sample before putting it on the flow cytometer (follow instructions in Subheading 3.2).

3.1.2 Isolation of Plant Nuclei Using the Two-Step Protocol

This procedure uses either the Otto or Baranyi buffers (buffer 11 or 12, respectively, listed in Subheading 2.3.2).

1. Place a small amount of the selected plant tissue (usually about 1 cm^2 or 20 mg) in a 6 cm petri dish (*see* **Note 11**).
2. Add 1 mL of ice-cold Otto I or ice-cold Baranyi solution I buffer (*see* **Note 12**).
3. Chop the tissues in the buffer using a new razor blade or sharp scalpel (*see* **Note 13**).
4. Mix the crude suspension by gently shaking the petri dish.
5. Filter the homogenate through a 30–42 μm nylon mesh filter into a labelled 1.5 mL tube.

6. Pellet the nuclei by centrifuging at 150 g for 5 min (*see* **Notes 19** and **20**).
7. Carefully remove the supernatant leaving approximately 100 μL of the buffer (*see* **Note 21**).
8. Resuspend the pellet by gently shaking and add a further 100 μL of the buffer used in **step 2** (*see* **Note 22**).
9. Add 1 mL of room temperature buffer, either Otto II or Baranyi solution II (*see* **Note 23**).
10. Add the appropriate volume of the fluorochrome to the nuclei suspension (if it is not already in the buffer—*see* Subheading 2.3.2) and vortex gently. For a typical sample which is c. 1.2 mL, the amount of stock PI added is 60 μL, while for DAPI, 50 μL should be added.
11. Incubate the samples at room temperature for few minutes in the dark (*see* **Note 24**).
12. Proceed to analyze the nuclear DNA content, vortexing the sample before putting it on the flow cytometer (follow instructions in Subheading 3.2).

3.1.3 Isolation of Plant Nuclei Using a Simplified Two-Step Protocol

This procedure uses either Otto, Baranyi, or Mishiba's buffer (buffers 11, 12, or 13, respectively, listed in Subheading 2.3.2).

1. Place a small amount of the selected plant tissue (usually about 1 cm^2 or 20 mg) in a 6 cm petri dish (*see* **Note 11**).
2. Either (1) add 0.5 mL of ice-cold Otto I or ice-cold Baranyi solution I buffer, or (2) add 0.2 mL of ice-cold Mishiba's solution A (*see* **Note 12**).
3. Chop the tissues in the buffer using a new razor blade or sharp scalpel (*see* **Note 13**).
4. Mix the crude suspension by gently shaking the petri dish. If using Mishiba's buffer, incubate for 5 min at room temperature.
5. Either (1) add 2 mL of Otto II or Baranyi solution II buffer, or (2) add 1 mL of Mishiba's solution B.
6. Filter the homogenate through a 30–42 μm nylon mesh filter into a labelled flow cytometry tube (*see* **Note 15**).
7. Add the appropriate volume of the fluorochrome to the nuclei suspension (if it is not already in the buffer—*see* Subheading 2.3.2) and vortex gently.
   (a) For a typical sample using either Otto or Baranyi buffers, the volume is usually c. 2.5 mL; thus, the amount of stock PI added is 125 μL, while for DAPI, 100 μL should be added.
   (b) For a typical sample using Mishiba's buffer, the volume is usually c. 1.2 mL; thus, the amount of stock PI added is 60 μL, while for DAPI, 50 μL should be added.

8. (a) For samples in the Otto or Baranyi buffer, incubate at room temperature for few minutes in the dark (*see* **Note 24**).
   (b) For samples in Mishiba's buffer, incubate at room temperature in the dark for 20 min.
9. Proceed to analyze the nuclear DNA content, vortexing the sample before putting it on the flow cytometer (follow instructions in Subheading 3.2).

3.2 Analysis of the Nuclear DNA Content and DNA Ploidy Level

The flow cytometer allows the measurement of several optical properties of the isolated particles (i.e., nuclei) that move one by one through the flow capillary tube illuminated by a laser beam or mercury light source. Prior to analyzing any plant sample, check that the instrument is properly aligned using fluorescent calibration beads (*see* Subheading 2.2). Subsequently test the linearity of the flow cytometer by running a plant sample (e.g., reference standard) and comparing the ratio between the $4C/2C$ peaks, which ideally should be in the range of 1.98–2.02 *sensu* Doležel et al. [8].

The first step in the analysis of a new target species requires the user to determine its relative nuclear DNA fluorescence. This step is described in Subheading 3.2.1. Based on this information, the user can then proceed either to Subheading 3.2.2 to determine the absolute DNA amount or to Subheading 3.2.3 to determine the DNA ploidy level.

3.2.1 Measurement of the Relative Nuclear DNA Fluorescence of a Sample

1. Load the tube containing the suspension of stained nuclei onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Notes 25** and **26**).
2. Adjust the flow rate to a speed of 15–25 nuclei/s (*see* **Notes 27** and **28**).
3. Once the sample is running through the flow cytometer, a flow histogram with peaks will start to appear. The peak positions can then be adjusted using the instrument gain settings to move the peaks within the histogram (*see* **Notes 29** and **30**). It is also possible to adjust the lower limit threshold so that undesirable low-channel signals (e.g., from cell debris and autofluorescent compounds) are excluded from the histogram. If there is a large amount of cell debris/background fluorescence in the flow histogram, then *see* **Note 31**, while if additional, unexpected peaks appear, then *see* **Note 32**.
4. Measure 5,000 particles (*see* **Note 33**).
5. Use the software provided by the flow cytometer manufacturer to assess the quality of histograms by (1) estimating the proportion of background, (2) checking peak symmetry, and (3) evaluating the peak width, expressed as the coefficient of variation, CV% (=SD of peak/mean channel position of the peak × 100) (*see* **Notes 34** and **35**).
6. Save the histogram if appropriate (*see* **Note 36**).

3.2.2 Measurement of the Absolute Nuclear DNA Content of a Sample Using a Reference Standard

Once the target sample has been run on its own to determine what reference standard to use and what gain the machine should be set at (*see* Subheading 3.2.1), a combined sample which includes both the target species and reference standard can then be prepared and run to determine the absolute nuclear DNA content of the target species.

To ensure the estimate of nuclear DNA content in absolute units is as accurate as possible, FCM researchers have adopted several best practice approaches. These include the following recommendations: (1) three specimen plants are collected per population/species and three independent replicates are processed per sample, or (2) five specimens are collected per population/species and two independent replicates are processed per specimen. (3) Only intercalating fluorochromes (e.g., PI) should be used; base-specific fluorochromes such as DAPI are not suitable for estimating nuclear DNA content.

1. Load the sample which contains a suspension of stained nuclei of both the target species and the selected internal reference standard (based on results obtained in Subheading 3.2.1) onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Notes 25** and **26**).
2. Adjust the flow rate to a speed of 15–25 nuclei/s (*see* **Notes 27** and **28**).
3. Once the sample is running through the flow cytometer, a flow histogram with peaks will start to appear. The peak positions can then be adjusted, if necessary, using the instrument gain settings to move the peaks within the histogram. It is also possible to adjust the lower limit threshold so that undesirable low-channel signals (e.g., from cell debris and autofluorescent compounds) are excluded from the histogram.
4. Check to see if there is any evidence of negative effects caused by the presence of cytosolic compounds which can affect the accuracy of the *C*-value estimation. This is done by comparing the position of the G_1 peak of the reference standard in this combined sample with its position in a sample containing just the reference standard (*see* Subheading 3.2.1) (N.B. both samples must be run at the same gain).
5. When this situation arises, alternative isolation methods should be tested (*see* **Note 37**); otherwise, proceed to the next step.
6. Measure 5,000 particles (*see* **Note 33**) (in some protocols, 10,000 particles are recommended) and save the data (*see* **Note 36**).
7. Use the software provided by the flow cytometer manufacturer to assess the quality of histograms (*see* **step 5** of Subheading 3.2.1). Assuming the quality of the histograms is suitable (i.e., CVs<3 %) (*see* **Note 34**), also obtain the statistical information for the histogram (i.e., mean peak position).

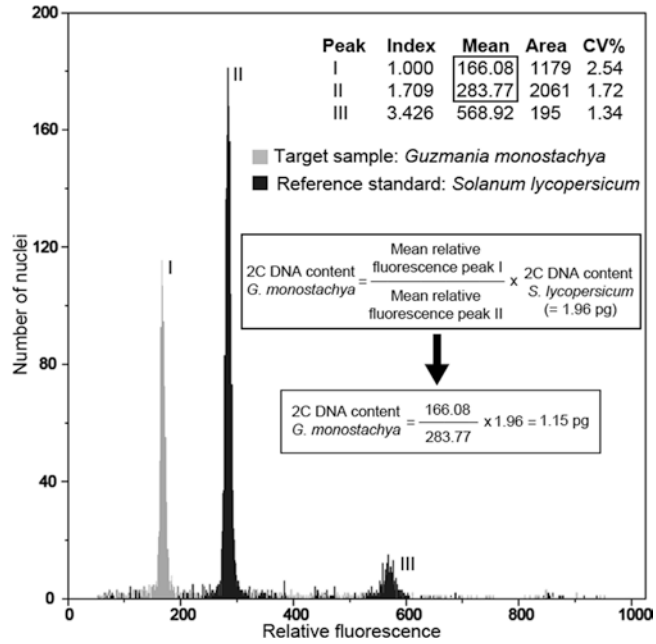


Fig. 3 A typical flow histogram to illustrate how genome size is calculated for the target species *Guzmania monostachia* using *Solanum lycopersicum* as the internal reference standard. Using the output data from the flow cytometer software, the mean relative fluorescence of the G_1 peak of *G. monostachia* (*grey peak* labelled I, i.e., 166.08) is divided by that of the mean G_1 peak of the standard *S. lycopersicum* (*black peak* labelled II, i.e., 283.77). This ratio is then multiplied by the 2C DNA content of *S. lycopersicum* to give the 2*C*-value of *G. monostachia*. To convert between pg and Mbp, use the conversion factor 1 pg = 978 Mbp Dolezel et al. [1]. (N.B. peak III is the G_2 peak of the reference standard.)

8. Calculate the nuclear DNA amount (2*C*-value) of the target plant in each replicate as follows (*see* **Notes 38** and **39**):

$$2C\text{ DNA content target (pg)} = \frac{\text{target sample mean } G_1 \text{ peak}}{\text{standard sample mean } G_1 \text{ peak}} \times 2C\text{ DNA content standard (pg)}$$

For an illustrative sample histogram output and calculation, *see* Fig. 3.

9. Calculate the mean nuclear DNA content and the standard deviation for the species (including all specimens and replicates) (*see* **Note 40**). (N.B. to convert between picograms (pg) and megabase pairs (Mbp) use: 1 pg = 978 Mbp [1]).

3.2.3 Measurement of the Relative Nuclear DNA Content of a Sample Using a Reference Standard to Determine DNA Ploidy Level

Among the multiple uses of FCM, DNA ploidy estimation is becoming highly popular as it allows the rapid screening of multiple samples. The protocol described below is optimized to work at either the species level or within species complexes.

1. Load the sample which contains a suspension of stained nuclei of both the target species of unknown ploidy and either a reference sample comprising a species of known ploidy (i.e., karyologically determined) or another internal standard (in that case, as mentioned above, the ploidy level of at least one target sample must be karyologically determined) (*see* **Note 41**) onto the flow cytometer sample port, and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Note 25**).
2. Perform **steps 2–3** (Subheading 3.2.2).
3. Measure at least 3,000 particles (*see* **Note 42**) and save the data (*see* **Note 36**).
4. Use the software provided by the flow cytometer manufacturer to obtain the statistical information for the histogram (e.g., peak position and ratio, area, CV%).
5. Calculate the relative nuclear DNA amount (DNA ploidy) of the target plant as follows:
   (a) If the reference sample used (with known ploidy) is the same species as the target sample, a perfect overlapping of G_1 peaks will indicate they both have the same ploidy.
   (b) If multiple peaks appear, then calculate the ploidy level using the following formula:

$$\text{Target sample ploidy} = \frac{\text{target sample mean } G_1 \text{ peak}}{\text{standard sample mean } G_1 \text{ peak}} \times \text{reference sample ploidy.}$$

   (c) If one of the cultivars listed in Table 1 is used as the reference standard, ploidy levels can be inferred by means of the ratio between the G_1 peaks of both the standard and the target samples (keeping in mind that the chromosome number of at least one of the target samples must be known). An example of the FCM analysis of ploidy level in the genus *Sorbus* (Rosaceae) is given in Fig. 4.

4 Notes

1. Many of the chemicals that are used in FCM are considered hazardous, and so suitable protective equipment (i.e., gloves, lab coat, fume cupboard) should be used to avoid health risks, and manufacturer's safety recommendations should be followed when using them. For example:
   MOPS (3-morpholino-propanesulfonique acid) and DTT (dithiothreitol) may cause irritation to the eyes, respiratory system, and skin.

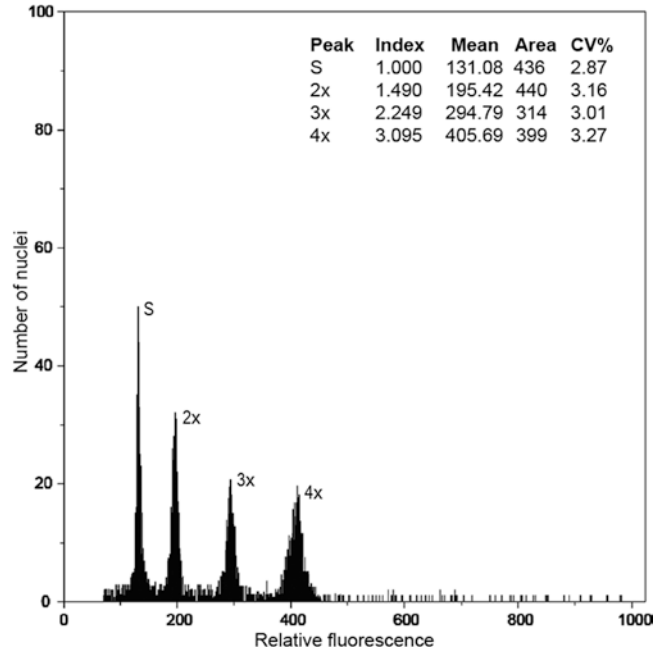


Fig. 4 Flow cytometric ploidy analysis in *Sorbus*. DNA ploidy was assessed in different species of *Sorbus* using the internal reference standard (*Oryza sativa*). Diploid *Sorbus aria* (whose chromosome number has been counted) was used as a reference to uncover higher ploidy levels in related species of unknown ploidy by determining the ratio between the peaks of these *Sorbus* species and the internal reference standard. S = G_1 peak of the internal standard (*Oryza sativa*); 2x = G_1 peak of the chromosomally determined diploid *S. aria*; 3x = G_1 peak of the triploid *S. saxicola*; 4x = G_1 peak of the tetraploid *S. rupicola*

β-mercaptoethanol is very hazardous and can be fatal if inhaled, swallowed, or absorbed through skin contact.

PI is a potential mutagen and may cause irritation to the eyes, respiratory system, and skin.

DAPI is a potential carcinogen and may cause irritation to the eyes, respiratory system, and skin.

DMSO (dimethyl sulfoxide) itself is not considered as a hazardous substance but in contact with other potentially toxic chemicals might enhance their absorption through the skin.

2. PI and SYBR Green I are intercalating fluorescent dyes that bind to double-stranded DNA and RNA with no base preference, whereas DAPI is a fluorescent dye that binds preferentially to AT-rich DNA.
3. Some buffers might precipitate after a while if they have not been stored at the appropriate temperature or when poor quality water has been used. It is therefore recommended that buffers are stored at the specified temperature given in

Subheading 2.3.2 and, wherever possible, that high-quality single- or double-distilled water is used.

4. If the isolation buffer becomes cloudy, changes color, or contains suspended particles, it suggests that the buffer has been stored incorrectly or that the storage time has been exceeded. In either case, this can result in fungi or bacteria growing in the buffer. If this has happened, then new isolation buffer needs to be prepared and stored as indicated (*see* Subheading 2.3.2). Unused buffer should be discarded after 3 months. It is also strongly recommended to prepare small volumes (e.g., 200 mL) so that the stocks are as fresh as possible.
5. The pH of the isolation buffers must be above 4 in order for PI to stain the DNA; most are around a neutral pH. For protocols using either Otto buffer (*see* buffer 11, Subheading 2.3.2) or Baranyi buffer (*see* buffer 12, Subheading 2.3.2), the nuclei are isolated in a citric acid solution which is acidic (i.e., Otto I or Baranyi solution I). The pH is then raised to neutral by the addition of a basic solution containing Na_2HPO_4 (i.e., Otto II or Baranyi solution II) to ensure optimum staining of the DNA when the fluorochrome is added.
6. Isolation buffers contain a number of different components which ensure that not only are sufficient numbers of nuclei released from the cytoplasm but also that the DNA is protected from degradation and binds the fluorochrome quantitatively. Typically isolation buffers include the following components: (1) organic buffers (e.g., Tris, MOPS, HEPES) which stabilize the pH between 7.0 and 8.0 (depending on the buffer) to enable DNA staining by the fluorochrome; (2) nonionic detergents (e.g., Triton X-100 and Tween 20) to facilitate the release of nuclei and prevent their aggregation; (3) chromatin stabilizers (e.g., spermine, $MgCl_2$, $MgSO_4$) to maintain the integrity of the DNA; (4) chelating agents (e.g., Na_2EDTA (ethylenediaminetetraacetic acid disodium salt), sodium citrate) to bind divalent cations such as Mg^{2+} and Mn^{2+} and hence block DNase activity; and (5) inorganic salts (e.g., KCl, NaCl) to ensure the correct ionic strength of the buffer. Some buffers also include β-mercaptoethanol, DTT, ascorbic acid, and sulphite which act as reducing agents to prevent protein oxidation and PVP (*see* **Note 7** below). For a discussion of the effect of different buffer components in a range of plant species, *see* Loureiro et al. [45] and Greilhuber et al. [46].
7. The polymer PVP is used to reduce the effect of polyphenols and other secondary metabolites such as tannins that are often present in plant tissues and which can inhibit the quantitative staining of DNA by the fluorochrome. Such secondary metabolites may also increase cell debris leading to a significant reduction in the quality of the peaks in the flow histogram (*see* **Notes 31** and **35**). Generally PVP-10 and PVP-40 are used

although in certain cases only PVP-360 was shown to result in decent flow histograms (*see* Fig. 5a).

8. A modified version of this buffer was reported by E Hörandl et al. [47] who also added 6.1 mM sodium citrate to the buffer.
9. It is *essential* that the cell culture-tested grade of Tween 20 from Sigma-Aldrich (cat. no. P2287) is used. Tween 20 for molecular biology (Sigma cat. no. P9416) is not suitable for FCM.
10. Dissolving $Na_2HPO_4.12H_2O$ can be speeded up by heating the solution gently.
11. The amount of tissue used needs to be determined empirically, taking into account the amount of nuclei released and the proportion of debris produced. For internal standardization, when needed, also add leaf tissue of the appropriate reference standard species (*see* Table 1).
12. The selection of the most appropriate buffer needs to be determined empirically for each plant group. In many cases, the same buffer works well across a family, while in other cases, different buffers are needed for different genera, or even within a genus.
13. It is very important to use very sharp razor blades or scalpels to chop the tissue into a crude suspension while minimizing damage to the nuclei. It is therefore recommended that each razor blade or scalpel is used only once. The chopping must be vigorous, quick, and short to avoid drying of the sample. We recommend empirical adjustments, especially to the chopping intensity, so that optimal numbers of nuclei are released without generating too much cell debris which can lead to high background signal in the flow histogram and low numbers of nuclei in the G_1 peak.
14. The working volume can be modified, but remember that if this is done, then the volume of the fluorochrome added at **step 7** will need to be adjusted accordingly to maintain the appropriate final concentrations.
15. Check carefully that the sample is free of particles after filtration to minimize the possibility of blockages in the flow cytometer.
16. If the samples have become brown/dark just a few minutes after adding the fluorochrome, this is indicative that the sample is undergoing oxidation due to the negative effect of secondary metabolites present in the cytoplasm. To avoid this problem, it is recommended that the buffer is supplemented with reducing agents such as β-mercaptoethanol or DTT. Another option that might help is the addition of PVP-10, PVP-40, or higher molecular weights such as PVP-360 (*see* **Note 7**), which will help improve histogram quality and sample stability, especially

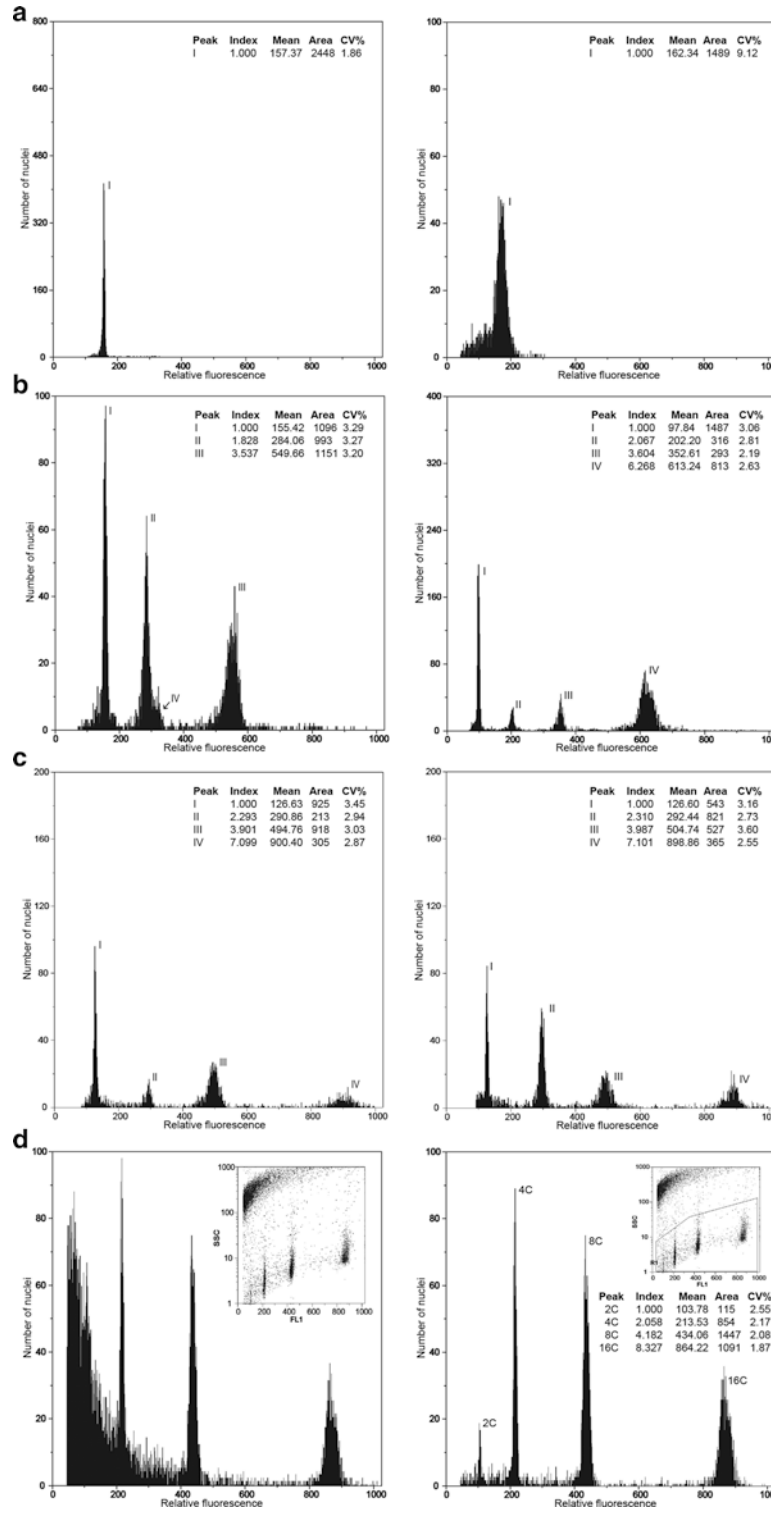


Fig. 5 Troubleshooting problems encountered during flow cytometric analysis of plant material. (**a**) Fluorescence histograms obtained after analysis of isolated nuclei of *Clusia multiflora* (Clusiaceae). Samples in both histograms were

if tannins are present. If the problem persists, then alternative isolation buffers (*see* Subheading 2.3.2) should be tested and the chopping intensity reduced (*see* **Note 13**).

17. Many protocols add RNase (ribonuclease II-A) at 50 μg/mL at this stage when PI is used as the fluorochrome. This is because PI intercalates into double-stranded (ds) nucleic acids so they can stain dsRNA as well as dsDNA. Nevertheless, since RNase is only active between 15 and 70 °C, with an optimal temperature of 60 °C, it can be left out of any protocol that lacks an incubation step within this temperature range. Since the protocols described here do not include such an incubation step, RNase has not been included. Nevertheless, if users want to include an RNase incubation step, a stock RNase solution can be prepared by heating 1 mg/mL RNase to 80 °C for 15 min (to inactivate DNases) and filtering through a 0.22 μm filter. The stock can be stored in 1 mL aliquots at −20 °C.
18. The time between staining (**step 7**) and running the sample on a flow cytometer can vary from a few minutes to up to 1 h. While for some plant samples, a short incubation works fine, for others a longer incubation can give better results. Thus, the incubation time needs to be adjusted empirically for a given plant species to optimize the results.
19. The relative centrifugal speed and time may need to be empirically adjusted.

Fig. 5 (continued) prepared using the same leaf and the same isolation buffer (WPB—36) but supplemented with different types of polyvinylpyrrolidone (PVP) to illustrate the dramatic effect on the quality of the flow histogram (*left*) using PVP-360 and (*right*) using PVP-40. (**b**) Flow histograms of the relative fluorescence in *Dactylorhiza* sp. (Orchidaceae) illustrating the utility of using alternative tissues to leaf samples to estimate nuclear DNA contents. (*Left*) Genome size estimated using pollinia of *Dactylorhiza* sp. and *Solanum lycopersicum* as internal standard [standard: peak I (G_1) and IV (G_2); pollinia: peaks II ($1C-G_1$) and III ($2C-G_2$); calculated 1*C*-value of *Dactylorhiza* sp. = 3.58 pg]. (*Right*) Genome size estimated using leaf tissue of *Dactylorhiza* sp. and *Solanum lycopersicum* as internal standard [standard: peak I (G_1) and II (G_2); *Dactylorhiza* sp. leaf: peaks III ($2C-G_1$) and IV ($4C-G_2$); calculated $2C = 7.06$ pg]. (**c**) Flow histograms of leaf tissue from the orchid *Dracula* sp. (using *Oryza sativa* as internal standard (peak I)), illustrating how different parts of the same leaf can have very different proportions of G_1 and G_2 nuclei. Using a young and actively growing leaf of *Dracula* (c. 1.5 cm long), the apical tip was seen to have a much lower proportion of G_1 nuclei (peak II, left histogram) compared with the basal part of the leaf (peak II, right histogram). (N.B. peaks III and IV correspond to G_2 and partial endopolyploid nuclei, respectively). (**d**) Flow histograms of relative fluorescence in leaf tissue of *Kalanchoe marnieriana* (Crassulaceae) illustrating how poor histograms with much debris (*left*, ungated histogram) can be improved by gating the histogram (*right*) to reveal not only the G_1 nuclei of *K. marnieriana* which was hidden in the debris of the left histogram but also the presence several endopolyploid cycles

20. Samples are stable in Otto I and Baranyi solution I; hence, it is possible to prepare several samples in advance and simultaneously centrifuge them together.
21. It is important to do this step very gently so as not to remove the pelleted nuclei.
22. As samples are stable in Otto I and Baranyi solution I, it is possible to prepare many replicates and store them at either room temperature or 4 °C for up to several hours.
23. The addition of Otto II or Baranyi solution II raises the pH of the sample to c. 7.3 and increases salt concentration. To keep these parameters within a working range, the amount of buffer added at this stage should be about fourfold that of Otto I or Baranyi solution I which now comprises c. 200–250 μL.
24. The optimal incubation time should be adjusted in each case, but short incubation times (e.g., less than 5 min) usually provide the best results because nuclei may not remain stable for a long time after this step.
25. The user can clear the acquisition results as many times as needed until the flow rate becomes stabilized. Do not be tempted to start recording data for analysis until the flow rate has stabilized (usually 0.5–1 min after the start of the run), as this can lead to poor histograms and inaccurate results.
26. If no peaks are appearing in the flow histogram, and assuming that the flow cytometer is properly set up, the peaks are probably off the scale due to an inappropriate gain setting for the sample being analyzed. To locate the peaks, adjust the gain setting of the machine. This can sometimes be done more easily by using the log scale setting of the relative fluorescence (x axis) scale. Once the position of the peak has been located, adjust back to a linear scale to perform analyses. Remember that the gain of the machine should be kept within the range that is recommended by the manufacturer of the flow cytometer to ensure the machine is operating optimally.
27. If the flow rate is slow, there are a number of explanations and possible solutions: (1) this could be a technical problem with the flow cytometer. Any blockage in the flow chamber or in the tubing system can cause a reduction in the number of nuclei recorded. Check that the pressure in the system is within the recommended range for the machine and clean the flow system using either a decontaminant solution or a diluted bleach solution (*see* Subheading 2.2) to wash out any potential blockage. (2) Alternatively, this could be a biological problem caused by the particular plant material being analyzed. The concentration of nuclei in the suspension can vary significantly between samples depending on tissue type, quantity of material used, etc. Hence, the flow rate will need to be adjusted each time a new sample is loaded onto the machine. If the concentration of

nuclei in the sample is low, this will necessitate a high flow rate, and this can result in a broadening of peaks and high CVs (*see* **Notes 34** and **35**). If possible, it is best that this is overcome by preparing a new sample using more material to increase nuclei concentration, rather than running the sample at a high flow rate. When using flow cytometers with a preset sample acquisition rate (e.g., slow, medium, high), we recommend using the slow rate and only increase it if necessary.

Other possible causes of a slow flow rate include the following: inappropriate chopping intensity, the tissue used is not suitable, and/or the isolation buffer selected is not appropriate. Such problems can be overcome by, for example, increasing the amount of tissue used, adjusting the chopping intensity, and testing different types of plant material (*see* Subheading 2.1, and Fig. 5b which illustrates the effect of changing from leaf material to pollinia in the orchid *Dactylorhiza*). Even changing the end of the leaf used for analysis can result in a dramatic change in the proportion of G_1 nuclei released (*see* Fig. 5c). Changing the isolation buffers (*see* Subheading 2.3.2) can also have a large effect, especially if the sample is releasing mucilaginous compounds into the chopping buffer. Indeed, many plants contain mucilage in their cytoplasm, and isolated nuclei may bind to this during the chopping process leading to a low number of released nuclei. Increasing the percentage of detergent (i.e., Triton X-100 up to 4 %) can help but keep in mind that a higher concentration of detergent can also result in higher levels of cell debris and hence poorer quality of flow histograms, so compromise may be necessary.

28. If the flow rate is unstable just after starting acquisition and large numbers of particles are being recorded, even when the flow cytometer is running at a slow speed, this may be due to unstable pressure in the flow cytometer. It can be caused by a number of factors including the presence of suspended particles (e.g., algae) in the sheath fluid and sheath fluid tubes/filters. Check that the pressure is correct and replace sheath fluid, tubes, and filters. If algae become a recurrent problem, 0.02 % sodium azide can be added to the water in the sheath fluid bottle; however, it should be noted that sodium azide is toxic and should be handled appropriately. Alternatively, the sheath fluid bottles can be thoroughly rinsed with domestic bleach (*see* Subheading 2.2) every 2 months, or even more frequently when they are not changed on a daily basis. In addition, many manufacturers recommend that the sheath fluid tubing and filters are replaced every 3 months.

29. Given that most of the measurements will require the use of a reference standard (*see* Subheadings 3.2.2 and 3.2.3), it is strongly recommended that the user knows, in advance, the peak position of a set of reference standards, ranging from small to big genomes (check Table 1 for recommended reference standards). This can be done by adjusting the gain set-

tings so that the G_1 peak of the standard always falls, for example, around channel number 200. Then, when the target sample is run alone for the first time, the user will be able to determine the best reference standard by testing the peak position of the target plant at the different gains selected for the standards. It should be noted that while the G_1 peak is the dominant peak, in many cases, G_2 peaks are present which might interfere with the target sample. Care should therefore be taken to note where the peak positions of the target sample fall in relation to both the G_1 and G_2 peaks of the reference.

The position of the G_1 peaks in the flow histogram of both the target plant and the internal reference standard should be different enough to avoid overlapping peaks. However, ideally, the ratio between the standard and the target plant peaks should not exceed threefold to reduce risk of errors arising due to loss of linearity.

30. If the position of the peak appears to be unstable (i.e., the peak in the histogram builds at a different position each time the acquisition data is cleared), it may suggest the incubation time following the addition of the fluorochrome is insufficient (i.e., **step 8**, Subheading 3.1.1). Check different incubation times to test staining stability. If the problem persists, test alternative isolation buffers. However, when using Otto or Baranyi buffers, the nuclei may be unstable once Otto II or Baranyi solution II has been added (*see* **step 11**, Subheading 3.1.2 or **step 8**, Subheading 3.1.3). For these buffers, increasing the incubation time is only likely to lead to deterioration in the flow histogram quality and unstable peaks.

31. Large amounts of cell debris/background on the flow histogram are a commonly encountered problem (e.g., see histogram on the left of Fig. 5d). There are several explanations and solutions:

    (a) The isolation buffer selected is not appropriate for the sample. Test an alternative isolation buffer (*see* Subheading 2.3.2).

    (b) The tissue selected is not in good condition or optimal for FCM. Test other plant tissues (*see* Subheading 2.1).

    (c) The incubation time (**step 8**, Subheading 3.1.1, **step 11**, Subheading 3.1.2, **step 8**, Subheading 3.1.3) can influence the quality of the flow histogram so try adjusting the incubation time.

    (d) Over-chopping of the sample (*see* **step 3** in Subheadings 3.1.1, 3.1.2, and 3.1.3) can, in some cases, lead to large amounts of background debris in the flow histogram. Reduce chopping intensity and use new sharp razor blades or scalpels for each sample to avoid cell damage.

    (e) If none of the solutions mentioned so far are showing any improvement in the results, then in some cases, gating can be used if the flow cytometer is fitted with a side

scatter detector. In this case, the region of interest is selected in the side scatter vs. forward light histogram so that the flow histogram of relative fluorescence excludes the signals from side scatter. An example of how effective this can be is seen in Fig. 5d.

32. If additional and perhaps unexpected peaks which do not follow an endopolyploid series are present in the flow histogram, then this suggests the presence of contaminants such as insects, insect eggs, and fungi in the plant sample. To avoid this problem, always check the plant material carefully before chopping (using a stereomicroscope if necessary) to ensure there are no contaminating organisms. If endoparasites are suspected, then alternative plant parts will need to be tested.
33. The number of particles that need to be recorded will vary depending on the type of analysis being carried out. Usually 5,000 particles are recorded of estimations of genome size, and this is the recommended number, but for some material, it may not be possible to obtain so many nuclei (e.g., recalcitrant and herbarium material).
34. The CV of a peak is a measure of peak quality and must be kept as low as possible (ideally less than 3 %) and always below 5 %. Higher CVs are not acceptable for publication unless it has been demonstrated that higher quality cannot be achieved after extensive tests with different buffers, incubation times, etc. (e.g., samples rich in polyphenols, old silica-dried samples, and herbarium vouchers).
35. Broad peaks with high and unacceptable CVs are, unfortunately, commonly encountered in the analysis of plant material. There are several possible explanations and solutions. These can broadly be divided into technical and biological sources:

    Technical

    (a) A loss of pressure in the flow cytometer system might result in a reduction of the peak quality. Check that the pressure is correct.

    (b) The instrument might be out of alignment. Align the instrument light source by using calibration beads (*see* Subheading 2.2).

    (c) Broad peaks are produced when the flow rate is too high. Run the samples at a flow rate that is no greater than c. 20 nuclei/s.

    (d) Air bubbles in the flow system can cause peaks with high CVs. Clean the flow chamber as recommended by the manufacturer and take extra care to remove any air bubbles from the filter after the sheath fluid bottle has been refilled. Also make sure that the lid of the sheath fluid bottle is tightly screwed on to seal the system.

(e) As reported by Doležel et al. [8], an obsolete arc lamp used for UV excitation might be a cause for such problem. Replace the lamp and align the instrument.

(f) Weak fluorescence and peaks with large CVs can arise when a sample of DAPI-stained nuclei is analyzed following a sample of PI-stained nuclei. Doležel et al. [8] noted that this situation can arise as a result of fluorochrome interference if the flow cytometer has not been completely cleaned between samples. To avoid this problem, ensure that the machine is thoroughly washed through by running a tube containing a weak solution (1:5 dilution in distilled water) of domestic bleach (do not leave bleach sitting in the system for more than a few minutes), then washing the system thoroughly with distilled water.

Biological

(a) In some cases, the isolation protocol and/or the buffer used is unsuitable for the material being analyzed, and the result can be a poor quality flow histogram. Test alternative isolation buffers (Subheading 2.3.2) and protocols (Subheading 3.1).

(b) Secondary metabolites in the cytoplasm may interfere with the fluorochrome staining of the DNA and lead to an increase in the CVs. Sometimes this can be overcome by supplementing the buffer with reducing agents such as β-mercaptoethanol or DTT. Tannins are also frequent in plants, so the addition of PVP-10/PVP-40 is common to help minimize their effects. PVP-360 has also been shown to be effective and, in certain cases, may work when other PVP types have failed (e.g., *see* Fig. 5a). The effect of secondary metabolites can also be minimized by reducing the chopping intensity and carrying out the nuclei isolation steps on ice and with ice-cold solutions (as recommended in Subheading 3.1).

(c) Some tissues of some plants are just recalcitrant and produce poor results. Test alternative tissues, such as bracts, roots, sepals, seeds, and pollinia (e.g., *see* Fig. 5b) of different parts of the leaf (e.g., *see* Fig. 5c) or try putting the plant in the dark for a few days prior to analysis.

(d) Doležel et al. [8] reported that excluding RNase from the isolation buffer when PI is used to stain DNA can result in increased CVs, especially in tissues with active protein synthesis, such as root tips. Nevertheless, if this is the case, then the protocol should include an incubation step at 37 °C for at least 30 min to ensure the RNase has sufficient time to work (*see* **Note 17** for how to prepare RNase).

(e) The wrong concentration of the DNA fluorochrome can also reduce the quality of the flow histogram, so it is important to check that the fluorochrome solution has been prepared correctly.

36. It is recommended that when a run is saved, the file name should include information on the species analyzed, replicate number, buffer used, and internal reference standard (if applicable). If possible, it is also helpful to get the software to list the instrument settings (e.g., gain and lower limit settings) used for each run. This enables histograms to be compared, if appropriate.
37. If a shift in the position of the G_1 peak of the reference standard is detected, then it is necessary to change the sample preparation. Often, the problem can be solved by changing to another isolation buffer (*see* Subheading 2.3.2). Alternatively, the addition of various compounds can sometimes eliminate the problem, e.g., the addition of 3 % PVP (*see* **Note 7** and Fig. 5a) to bind to polyphenolics or adding DTT and β-mercaptoethanol which are good reducing agents. In addition, the problem can sometimes be overcome by using different plant material such as roots, stems, bracts, or seeds (e.g., *see* Fig. 5b, c).
38. For accurate nuclear DNA amount estimations, it is recommended that the number of particles in both the target and the reference standard G_1 peaks should be similar.
39. Some plant breeding material, pollen, and the gametophyte stage of bryophyte groups (i.e., mosses, liverworts, and hornworts) are haploid. In such cases, the first peak of the target sample in a flow histogram (G_1) will correspond to the 1*C* rather than the 2*C*-value.
40. Technical factors should not account for more than 2–3 % of the variation between different estimates for the same species, although for some material (e.g., recalcitrant tissues), this type of variation may be greater. Higher levels of variability in *C*-value estimates for a species may reflect intraspecific variation due to the presence of chromosomal instabilities (e.g., B chromosomes, supernumerary segments, or aneuploidy) or taxonomic heterogeneity in the samples analyzed (*see* ref. 48 for further discussion on intraspecific variation).
41. (1) Wherever possible, it is recommended that the "reference" sample of known ploidy is at the lowest ploidy level known for a given species/complex (i.e., diploids). (2) If investigating ploidy levels within a species, the reference sample can be a sample of this species whose ploidy has been karyologically determined (e.g., a diploid sample). (3) Following the recommendations of Doležel et al. [8], the nuclear DNA may be stained with PI or DAPI, although the latter option may result in higher-quality histograms. The use of DAPI is also recommended to detect aneuploid specimens.
42. For ploidy level estimations, it is not necessary to measure as many nuclei, so data for a lower number of particles can be collected.

References

1. Doležel J, Bartoš J, Voglmayr H et al (2003) Nuclear DNA content and genome size of trout and human. Cytometry A 51A:127–128
2. Ogur M, Erickson RO, Rosen GU et al (1951) Nucleic acids in relation to cell division in *Lilium longiflorum*. Exp Cell Res 2:73–89
3. Bennett MD, Leitch IJ (2012) Plant DNA C-values database (release 6.0, Dec. 2012). http://data.kew.org/cvalues/
4. Kron P, Suda J, Husband BC (2007) Applications of flow cytometry to evolutionary and population biology. Annu Rev Ecol Evol Syst 38:847–876
5. Loureiro J, Travnicek P, Rauchova J et al (2010) The use of flow cytometry in the biosystematics, ecology and population biology of homoploid plants. Preslia 82:3–21
6. Suda J, Kron P, Husband BC et al (2007) Flow cytometry and ploidy: applications in plant systematics, ecology and evolutionary biology. In: Doležel J, Greilhuber J, Suda J (eds) Flow cytometry with plants cells. Wiley-VCH, Weinheim, pp 103–130
7. Leus L, Van Laere K, Dewitte A et al (2009) Flow cytometry for plant breeding. Acta Hortic 836:221–226
8. Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc 2:2233–2244
9. Suda J, Krahulcova A, Travnicek P et al (2006) Ploidy level versus DNA ploidy level: an appeal for consistent terminology. Taxon 55:447–450
10. Noirot M, Barre P, Louarn J et al (2002) Consequences of stoichiometric error on nuclear DNA content evaluation in *Coffea liberica* var. *dewevrei* using DAPI and propidium iodide. Ann Bot 89:385–389
11. Noirot M, Barre P, Louarn J et al (2000) Nucleus-cytosol interactions: a source of stoichiometric error in flow cytometric estimation of nuclear DNA content in plants. Ann Bot 86:309–316
12. Noirot M, Barre P, Duperray C et al (2003) Effects of caffeine and chlorogenic acid on propidium iodide accessibility to DNA: consequences on genome size evaluation in coffee tree. Ann Bot 92:259–264
13. Noirot M, Barre P, Duperray C et al (2005) Investigation on the causes of stoichiometric error in genome size estimation using heat experiments. Consequences on data interpretation. Ann Bot 95:111–118
14. Price HJ, Hodnett G, Johnston JS (2000) Sunflower (*Helianthus annuus*) leaves contain compounds that reduce nuclear propidium iodide fluorescence. Ann Bot 86:929–934
15. Loureiro J, Rodriguez E, Doležel J et al (2006) Flow cytometric and microscopic analysis of the effect of tannic acid on plant nuclei and estimation of DNA content. Ann Bot 98:515–527
16. Bennett MD, Price HJ, Johnston JS (2008) Anthocyanin inhibits propidium iodide DNA fluorescence in *Euphorbia pulcherrima*: implications for genome size variation and flow cytometry. Ann Bot 101:777–790
17. Loureiro J, Suda J, Doležel J (2007) FLOWer: a plant DNA flow cytometry database. In: Dolezel J, Greilhuber J, Suda J (eds) Flow cytometry with plant cells. Wiley-VCH, Weinheim, pp 423–438
18. Loureiro J, Rodriguez E, Santos C et al (2008) FLOWer: a plant DNA flow cytometry database (release 1.0, May 2008). http://flower.web.ua.pt/
19. Garnatje T, Canela MÁ, Garcia S et al (2011) GSAD: a genome size in the Asteraceae database. Cytometry A 79A:401–404
20. Greilhuber J, Doležel J, Lysak MA et al (2005) The origin, evolution and proposed stabilization of the terms "Genome size" and "C-value" to describe nuclear DNA contents. Ann Bot 95:255–260
21. Leitch IJ, Kahandawala I, Suda J et al (2009) Genome size diversity in orchids: consequences and evolution. Ann Bot 104:469–481
22. Torrell M, Valles J (2001) Genome size in 21 *Artemisia* L. species (Asteraceae, Anthemideae): systematic, evolutionary, and ecological implications. Genome 44: 231–238
23. Garcia S, Sanz M, Garnatje T et al (2004) Variation of DNA amount in 47 populations of the subtribe Artemisiinae and related taxa (Asteraceae, Anthemideae): karyological, ecological, and systematic implications. Genome 47:1004–1014
24. Leitch IJ, Bennett MD (2004) Genome downsizing in polyploid plants. Biol J Linn Soc Lond 82:651–663
25. Lysák MA, Lexer C (2006) Towards the era of comparative evolutionary genomics in Brassicaceae. Plant Syst Evol 259:175–198
26. Poggio L, Burghardt AD, Hunziker JH (1989) Nuclear DNA variation in diploid and polyploid taxa of *Larrea* (Zygophyllaceae). Heredity 63:321–328
27. Sliwinska E, Pisarczyk I, Pawlik A et al (2009) Measuring genome size of desert plants using dry seeds. Botany 87:127–135
28. Sliwinska E, Zielinska E, Jedrzejczyk I (2005) Are seeds suitable for flow cytometric estimation of plant genome size? Cytometry A 64A:72–79
29. Kron P, Husband BC (2012) Using flow cytometry to estimate pollen DNA content: improved

methodology and applications. Ann Bot 110:1067–1078. doi:10.1093/aob/mcs1167
30. Suda J, Travnicek P (2006) Reliable DNA ploidy determination in dehydrated tissues of vascular plants by DAPI flow cytometry: new prospects for plant research. Cytometry A 69A:273–280
31. Bainard JD, Husband BC, Baldwin SJ et al (2011) The effects of rapid desiccation on estimates of plant genome size. Chromosome Res 19:825–842
32. Kolář F, Lučanová M, Těšitel J et al (2012) Glycerol-treated nuclear suspensions: an efficient preservation method for flow cytometric analysis of plant samples. Chromosome Res 20:303–315
33. Doležel J, Binarova P, Lucretti S (1989) Analysis of nuclear DNA content in plant cells by flow cytometry. Biol Plant 31:113–120
34. Pfosser M, Amon A, Lelley T et al (1995) Evaluation of sensitivity of flow-cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat-rye addition lines. Cytometry 21:387–393
35. Galbraith DW, Harkins KR, Maddox JM et al (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science 220:1049–1051
36. Loureiro J, Rodriguez E, Doležel J et al (2007) Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. Ann Bot 100:875–888
37. Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218
38. Bino RJ, Lanteri S, Verhoeven HA et al (1993) Flow cytometric determination of nuclear replication stages in seed tissues. Ann Bot 72: 181–187
39. De Laat AMM, Blaas J (1984) Flow cytometric characterization and sorting of plant chromosomes. Theor Appl Genet 67:463–467
40. Ebihara A, Ishikawa H, Matsumoto S et al (2005) Nuclear DNA, chloroplast DNA, and ploidy analysis clarified biological complexity of the *Vandenboschia radicans* complex (Hymenophyllaceae) in Japan and adjacent areas. Am J Bot 92:1535–1547
41. Matzk F, Meister A, Brutovská R et al (2001) Reconstruction of reproductive diversity in *Hypericum perforatum* L. opens novel strategies to manage apomixis. Plant J 26:275–282
42. Otto F (1992) Preparation and staining of cells for high-resolution DNA analysis. In: Radbruch A (ed) Flow cytometry and cell sorting. Springer, Berlin, pp 101–104
43. Baranyi M, Greilhuber J (1995) Flow cytometric analysis of genome size variation in cultivated and wild *Pisum sativum* (Fabaceae). Plant Syst Evol 194:231–239
44. Mishiba KI, Ando T, Mii M et al (2000) Nuclear DNA content as an index character discriminating taxa in the genus *Petunia* sensu Jussieu (Solanaceae). Ann Bot 85:665–673
45. Loureiro J, Rodriguez E, Doležel J et al (2006) Comparison of four nuclear isolation buffers for plant DNA flow cytometry. Ann Bot 98:679–689
46. Greilhuber J, Temsch EM, Loureiro J (2007) Nuclear DNA content measurement. In: Doležel J, Greilhuber J, Suda J (eds) Flow cytometry with plant cells. Wiley-VCH, Weinheim, pp 67–102
47. Hörandl E, Dobes C, Suda J et al (2011) Apomixis is not prevalent in subnival to nival plants of the European Alps. Ann Bot 108:381–390
48. Greilhuber J, Leitch IJ (2013) Genome size and the phenotype. In: Leitch IJ, Doležel J, Wendel JF, Greilhuber J (eds) Plant genome diversity, vol 2, physical structure, behaviour and evolution of plant genomes. Springer, Wein, pp 323–344
49. Bennett MD, Smith JB (1991) Nuclear DNA amounts in angiosperms. Philos Trans R Soc Lond Ser B 334:309–345
50. Doležel J, Sgorbati S, Lucretti S (1992) Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. Physiol Plant 85:625–631
51. Doležel J, Dolezelova M, Novak FJ (1994) Flow cytometric estimation of nuclear DNA amount in diploid bananas (*Musa acuminata* and *M. balbisiana*). Biol Plant 36:351–357
52. Marie D, Brown SC (1993) A cytometric exercise in plant DNA histograms, with 2C values for 70 species. Biol Cell 78:41–51
53. Obermayer R, Leitch IJ, Hanson L et al (2002) Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. Ann Bot 90:209–217
54. Lysák MA, Doležel J (1998) Estimation of nuclear DNA content in *Sesleria* (Poaceae). Caryologia 52:123–132
55. Doležel J, Greilhuber J, Lucretti S et al (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. Ann Bot 82(Suppl A):17–26

Chapter 15

Molecular Cytogenetics (FISH and Fluorochrome Banding): Resolving Species Relationships and Genome Organization

Sonja Siljak-Yakovlev, Fatima Pustahija, Vedrana Vicic, and Odile Robin

Abstract

Fluorochrome banding (chromomycin, Hoechst, and DAPI) and fluorescence in situ hybridization (FISH) are excellent molecular cytogenetic tools providing various possibilities in the study of chromosomal evolution and genome organization. The constitutive heterochromatin and rRNA genes are the most widely used FISH markers. The rDNA is organized into two distinct gene families (18S-5.8S-26S and 5S) whose number and location vary within the complex of closely related species. Therefore, they are widely used as chromosomal landmarks to provide valuable evidence concerning genome evolution at chromosomal levels.

Key words Chromomycin, *Crepis*, DAPI, Fluorescence in situ hybridization (FISH), Fluorochrome banding, Hoechst, *Pinus*, rRNA genes

1 Introduction

Molecular cytogenetics provides new possibilities in the study of chromosomal evolution and genome organization which also contributes to a better characterization of the karyotype. Fluorochrome banding and fluorescence in situ hybridization (FISH) are excellent tools for chromosome identification in studies of chromosome evolution and genome organization and also to reveal the relationships between different taxa. These molecular cytogenetic approaches have been widely used for karyotyping, e.g., in *Arabidopsis thaliana* [1], *Medicago truncatula* [2], *Picea abies*, and *P. omorika* [3], and for studying evolutionary relationships within many genera, e.g., *Hypochaeris* [4, 5], *Quercus* [6], *Lilium* [7], *Nicotiana* [8], and *Pinus* [9, 10].

Before the development of fluorochrome banding, Giemsa C-banding has been used to reveal constitutive heterochromatin (highly repetitive DNA sequences which remain condensed during the whole cell cycle). This heterochromatin can be GC or AT-rich or "neutral." The most widely used base-specific fluorochrome chromomycin A3 is a fluorescent stain that binds strongly to GC-rich

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_15,

regions in DNA. DAPI (4',6-diamidino-2-phenylindole) or Hoechst (bisbenzimide 33258), on the other hand, is specific for AT-rich DNA. Comparative patterns of fluorochrome banding may be useful not only in identifying homologous chromosomes but also in revealing phylogenetic relationships among species [11–13].

In the case of two closely related species of the genus *Crepis* [*C. praemorsa* (L.) Tausch and *C. incarnata* (Wulf.) Tauch.] with the same chromosome number and almost identical karyotypes, banding techniques revealed a high intra-chromosomal differentiation between two species (Fig. 1). All constitutive heterochromatin in these two species, revealed after Giemsa C-banding, represents AT-rich DNA regions [14]. However, in *C. praemorsa* heterochromatic regions are limited only to centromeres and nucleolar organizer region (NOR). In *C. incarnata* this type of heterochromatin was abundant forming the large telomeric and intercalary bands. The AT-rich DNA regions are consequently GC-poor and present low fluorescent intensity with appropriate fluorochrome (see negative bands on chromomycin-stained chromosomes, Fig. 1d). Before these results, obtained by chromosome banding, in *Flora Europaea* the *C. incarnata* has been considered only as subspecies [*C. praemorsa* subsp. *dinarica* (Beck) Hayek, synonym = *C. incarnata*] [15]. This and numerous other studies demonstrate the usefulness of fluorochrome banding in resolving systematic and phylogenetic relationships between closely related taxonomic entities and point out the high implication of heterochromatin during differentiation of *C. incarnata* (endemic mountain species from Alps) from *C. praemorsa* (ancestral species from Euro-Asiatic plains with a large geographical repartition). In addition to this study, the reproductive isolation has been also detected which confirmed the specific level of these two taxa [16, 17].

Fluorescence in situ hybridization is a 30-year-old molecular cytogenetic tool that has developed continuously. Schwarzacher and Heslop-Harrison [18] provided the most accurately documented data and protocols concerning FISH techniques in plants.

In eukaryotes, rRNA genes present the most widely used FISH markers. They are organized into two distinct gene families. The first family of rRNA genes, encoding for 18S, 5.8S, and 26S ribosomal RNA (45S rDNA), occurs as tandem arrays at one or several specific regions on chromosomes. The 45S rDNA loci consist of tandemly repeated units of the 18S, 5.8S, and 26S rDNA, internal transcribed (ITS1 and ITS2) sequences, and intergenic (IGS) spacers. These genes are highly conserved and the chromosomal segment harboring them is known as a nucleolar organizing region (NOR). The second family is presented by 5S rRNA genes, also highly conserved and widely used as molecular cytogenetic markers.

Due to their high copy number, both families of rRNA genes are easily and reproducibly detectable on chromosomes and constitute suitable landmarks for chromosome identification.

The number and location of rDNA vary within the complex of closely related species; therefore, it could be used as a chromosomal landmark to provide valuable evidence concerning genome

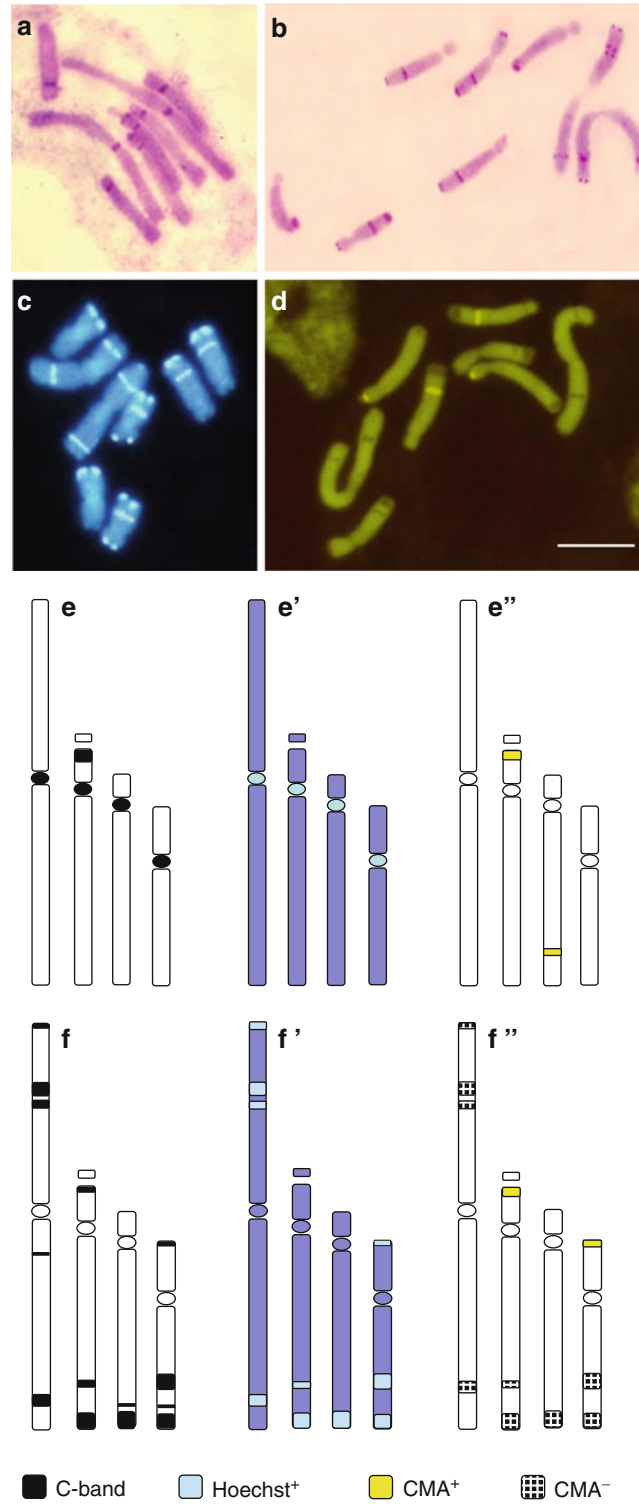


Fig. 1 *Crepis praemorsa*: Giemsa C-banding (**a**), idiogram showing C-bands (**e**), Hoechst (**e**′) and CMA bandings (**e**″). *C. incarnata*: C-banding (**b** and **f**), Hoechst (**c** and **f**′), and CMA bandings (**d** and **f**″). Bar = 10 μm

evolution at chromosomal levels. The rDNAs can change rapidly both in copy number and chromosome distribution, and rDNA transposition or dispersion in plant genomes is frequently observed [19–22]. These rearrangements are generally in correlation with species differentiation and speciation, and FISH analysis of rDNA is a good tool to detect chromosome variations.

The following example demonstrates the use of fluorochrome banding and FISH to detect small structural chromosomal differences even at the level of intraspecific taxonomic categories.

The genus *Pinus*, and Pinaceae family in general, is characterized by the same chromosome number ($2n = 24$) and conserved karyotypes with all metacentric chromosome pairs, except one submetacentric. In such cases, when karyotyping based on morphometric analysis is difficult, the comparative patterns of fluorochrome banding and FISH experiment may be useful not only in identifying homologous chromosomes but also for revealing phylogenetic relationships among taxa.

Thus, in the study of *Pinus nigra* Arnold subspecies [*Pinus nigra* subsp. *laricio* (Poiret) Maire and *Pinus nigra* subsp. *dalmatica* (Visiani) Franco], molecular cytogenetic tools revealed the unsuspected differences in heterochromatin and rDNA organization [9]. DAPI staining after FISH displayed a high number of signals (Figs. 2 and 3). The number of CMA bands was 26 in ssp.

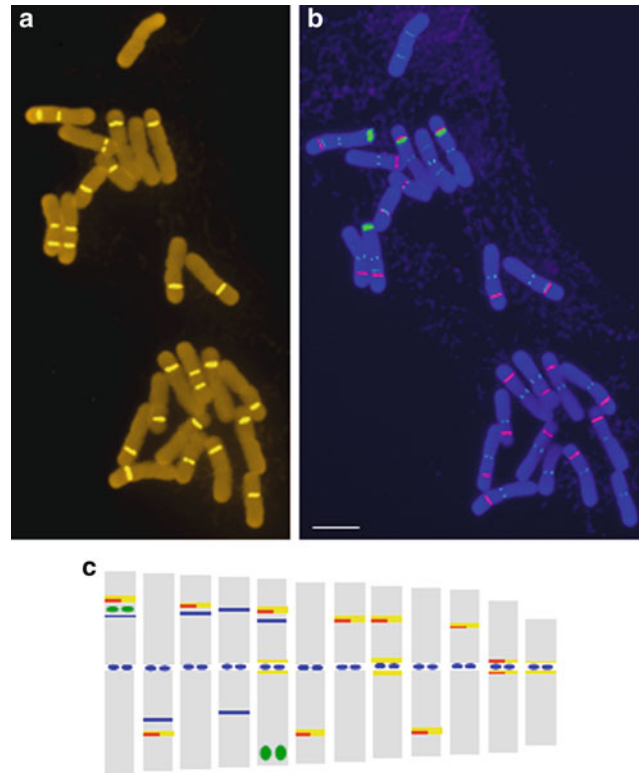


Fig. 2 *Pinus nigra* subsp. *laricio*: CMA-chromomycin banding (**a**); FISH (**b**); corresponding haploid idiogram showing 10 18S and 2 5S rDNA loci, 13 CMA, and 18 DAPI bands (**c**). Bar = 10 μm

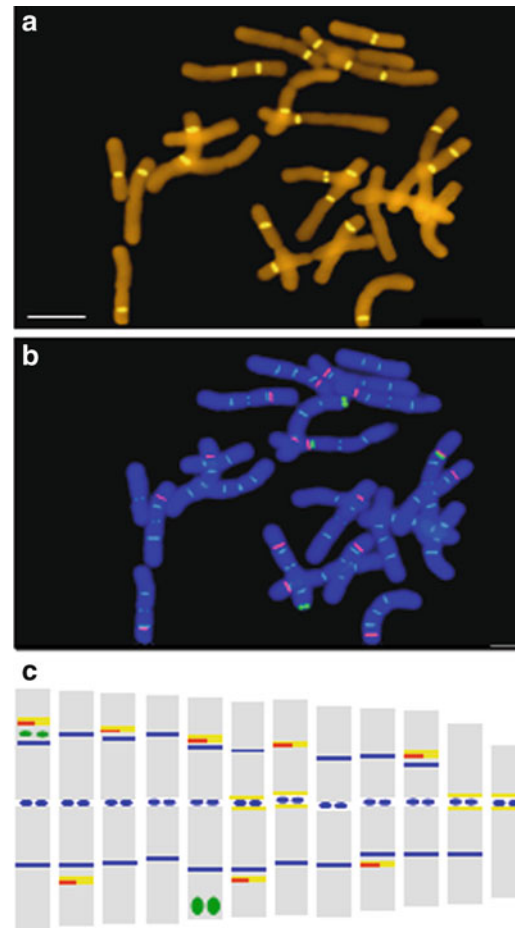


Fig. 3 *Pinus nigra* subsp. *dalmatica*: CMA (**a**); FISH (**b**); haploid idiogram showing 8 18S and 2 5S rDNA loci, 12 CMA, and 32 DAPI bands (**c**). Bar = 10 μm

laricio (Fig. 2a) and 24 in ssp. *dalmatica* (Fig. 3a) with slightly different positions. Since all the centromeres were DAPI positive, the differences were reflected in the number of intercalary DAPI bands. They were distributed either on one or both chromosome arms. Two DAPI patterns were evident: the first with a lower number of signals (36 in ssp. *laricio*) and the second with a higher number of bands (64 in ssp. *dalmatica*; Figs. 2b and 3b, respectively). The number and position of 5S rRNA genes were the same, but the number of 18S rDNA loci was 10 for ssp. *laricio* and 8 for ssp. *dalmatica* (Figs. 2b, c and 3b, c, respectively).

Therefore, the molecular cytogenetic analysis can unequivocally reveal the subtle chromosomal changes even between low taxonomic categories, and by combining with phytogeography and ecology of representatives of a complex of related species, it becomes possible to determine processes of species differentiation and evolution and the phylogenetic relationships between taxa.

2 Materials

Use sterile ultrapure water and analytical grade reagents for preparing solutions. Prepare and store all reagents at room temperature, unless indicated otherwise. For long-term storage the stock solutions can be aliquoted and stored at ≤–20 °C. For short-term storage the solutions can be kept at 2–6 °C, protected from light if necessary. Care should be taken in handling and disposal of dyes and all waste materials, according to applicable local regulations.

2.1 Reagents for Pretreatments and Root Tips Fixations

1. 0.05 % (m/v) *aqueous colchicine solution*: resolve 0.05 g colchicine in 100 mL water.
2. 0.002 M *8-hydroxyquinoline*: resolve 0.029 g 8-hydroxyquinoline in 100 mL water (*see* **Note 1**).
3. *Carnoy I*: freshly prepared 3:1 (v/v) ethanol:glacial acetic acid (*see* **Note 2**).
4. *Carnoy II*: freshly prepared 6:3:1 (v/v) ethanol:chloroform: glacial acetic acid (*see* **Note 3**).

2.2 Buffers

1. *Citrate buffer 0.01 M*, pH 4.6: solution A: 0.1 M citric acid. Solution B: 0.1 M trisodium citrate, pH 4.6. Mix 25.5 mL solution A + 24.5 mL solution B, adjust volume to 100 mL with ddH_2O. Store at –20 °C.
2. *Citrate buffer 0.05 M*, pH 4.6: solution A: 0.5 M citric acid. Solution B: 0.5 M trisodium citrate, pH 4.6. Mix 25.5 mL solution A + 24.5 mL solution B, adjust volume to 100 mL with ddH_2O. Store at –20 °C.
3. *McIlvaine buffer pH 5.5*: solution A: 0.l M citric acid. Solution B: 0.2 M dibasic sodium phosphate. Mix 21.6 mL of A and 28.4 mL of B, adjust volume to 200 mL with ddH_2O. Store at –20 °C.
4. *McIlvaine buffer pH 7.0*: solution A: 0.l M citric acid. Solution B: 0.2 M dibasic sodium phosphate. Mix 18 mL of A and 82 mL of B, diluted to a total of 200 mL with ddH_2O. Store at –20 °C.
5. *McIlvaine buffer pH 7.0 + Mg^{2+}*: prepare a 5 mM Mg^{2+} solution by diluting 0.123 g of $MgSO_4 \cdot 7H_2O$ in 100 mL of McIlvaine buffer pH 7.0 (*see* **Note 4**). Store at –20 °C.

2.3 Enzyme Mixture

1. *Hydrolytic enzymes mixture*: 4 % cellulase "Onozuka" RS (Yakult Pharmaceutical Co.), 1 % pectolyase Y23 (Seishin Pharmaceutical Co.), and 4 % hemicellulase (Sigma-Aldrich) in 0.05 M citrate buffer. Store mixture at –20 °C (*see* **Note 5**).

2.4 Fluorochrome Banding

1. *Chromomycin A3* (CMA): dissolve 0.02 g of CMA in 100 mL of McIlvaine buffer pH 7.0 + Mg^{2+}. Store at –20 °C, protected from light.
   0.05 % (m/v) methyl green dissolved in pH 5.5 McIlvaine buffer. Store at 4 °C.
2. *Hoechst 33258* [Ho; bisbenzimide 33258; 2-[2-(4-Hydroxyphenyl)-6-benzimidazolyl]-6-(1-methyl-4-piperazyl)-benzimidazoletrihydrochloride]: to prepare a stock solution, dissolve 1 mg Ho in 100 mL ddH_2O and store at –20 °C, protected from light. Work solution: dilute 1 mL of Ho stock solution with 4 mL of McIlvaine pH 5.5.
3. *DAPI* (4',6-diamidino-2-phenylindole): stock solution of 2 μg/mL in McIlvaine's buffer pH 7.0. Prepare a working solution of 0.1 mg/mL, aliquot, and store at –20 °C, protected from light.
4. *Glycerol antifade solution*: Citifluor AF_2 (Agar Scientific Oxford Instruments, Stanstead, UK) or manually prepared glycerol solution (McIlvaine pH 7.0 + Mg^{2+}: glycerol = 1:1, v/v).

2.5 FISH (Fluorescence In Situ Hybridization)

1. *RNAse A stock solution*: prepare a stock solution dissolving 10 mg of RNAse in 1 mL of 10 mM Tris–HCl, pH 8.0. Boil for 15 min and allow to cool. Store at –20 °C in aliquots. Prior to use dilute 100× in 2× SSC.
2. *0.01 M HCl.*
3. *Pepsin stock solution:* prepare a 0.1 mg/mL solution by dissolving in 0.01 M HCl. Aliquot and store at –20 °C.
4. *Proteinase K*: prepare a 1 mg/mL stock solution by dissolving in ddH_2O. Store at –20 °C. Prior to use dilute 100× in 2× SSC.
5. *Formamide deionized.*
6. *20× SSC* (saline-sodium citrate buffer): 3 M sodium chloride, 0.3 M sodium citrate tribasic dihydrate, adjusted to pH 7.0 with 1 M HCl, autoclaved, and stored at room temperature. For use in the hybridization mix store at –20 °C.
7. *2× SSC*: dilute 20× SSC 1:10 with ddH_2O.
8. *0.1× SSC*: dilute 13 mL of 2× SSC with 237 mL ddH_2O.
9. *4× SSCT*: dilute 100 mL of 20× SSC with 400 mL ddH_2O. Add 1 mL of Tween 20.
10. *Tween 20.*
11. *Dextran sulfate* (DS): dissolve 50 g of DS in 100 mL of sterile ddH_2O. Store in aliquots at –20 °C.

12. *Sodium dodecyl sulfate* (SDS): dissolve 1 g of SDS in sterile 10 mL ddH_2O. Store in aliquots at –20 °C.
13. *Salmon sperm DNA solution* (SS): concentration 10.5 ± 0.5 mg/mL. Store in aliquots at –20 °C.
14. *Blocking buffer* [bovine serum albumin (BSA) solution]: dissolve 0.1 g of BSA in 2 mL 4× SSCT (*see* **Note 6**). Store at –20 °C.
15. *Antibody buffer* [anti-digoxigenin-fluorescein, Fab fragments (ADF) mixture]: for one slide mix 49.3 μL of blocking buffer with 0.7 μL of ADF (*see* **Note 7**).
16. *Antifade mounting medium*: use Vectashield mounting medium with DAPI (Vector Laboratories, Peterborough, UK).

3 Methods

3.1 Pretreatment and Fixation of Root Tips

Carry out all procedures at room temperature unless otherwise specified.

1. Immerse root tip in colchicine solution for 2–6 h at room temperature (large chromosomes) or 8-hydroxyquinoline solution for 2–4 h at 16 °C (small chromosomes).
2. Fix root tips in Carnoy I or Carnoy II solution for 15–30 min at room temperature and leave in fixative for 24–48 h at 4 °C (*see* **Note 8**).

3.2 Preparation of Protoplasts

Following is the technique of Geber and Schweizer [23] with minor modifications.

1. Thaw enzyme mixture at 37 °C and transfer it to either a watch glass in a Petri dish or a 1.5 mL centrifuge tube (*see* **Note 9**).
2. Wash fixed root tips in 0.05 M citrate buffer for 10 min and then digest in the enzymatic mixture at 37 °C for 10–60 min (depending on root size, *see* **Note 10**).
3. Transfer meristems to a drop of 45 % acetic acid on a clean slide. Place cover slip and apply gentle pressure to spread the chromosomes. Tapping with needle tweezers on top of the cover slip may improve chromosome spreading.

3.3 Cover Slips Removing

Following is the technique of Conger and Fairchild [24] with minor modifications.

1. Rapidly freeze preparations below –70 °C using liquid nitrogen or CO_2, or by placing slides on dry ice or on a metal plate in a –80 °C freezer (*see* **Note 11**).
2. Remove cover slips quickly while frozen, using a razor blade, and rinse briefly in absolute ethanol.
3. Take out slides, air dry, and store at room temperature for a couple of days till the next step (staining).

3.4 Chromomycin Banding

Following are the modified techniques of Schweizer [25] and Kondo and Hizume [26], and the technique of Siljak-Yakovlev et al. [3].

1. Prepare air-dried slides with cover slips removed as described in Subheading 3.3.
2. Prepare McIlvaine buffers (pH 5.5; pH 7.0; pH 7.0 + Mg^{2+}) and CMA working solution.
3. Add a few drops of McIlvaine buffer pH 7.0 + Mg^{2+} to the slides and incubate for 15 min. Gently shake off slides.
4. Apply 80 μL of CMA working solution onto the slides and gently cover with plastic cover slips (cut from autoclavable waste bags) avoiding formation of air bubbles. Incubate for 60–90 min in the dark.
5. Carefully remove the plastic cover slips with tweezers and wash briefly with McIlvaine buffer pH 7.0.
6. Counterstain with methyl green for 7 min in the dark.
7. Wash slides briefly with McIlvaine buffer pH 5.5.
8. Mount preparations in glycerol antifade solution.
9. Store slides in the dark. For long-term conservation store at 4 °C.
10. Observe under an epifluorescence microscope with appropriate filters.

3.5 Hoechst Banding

Following are the techniques of Martin and Hesemann [27].

1. Prepare air-dried slides with cover slips removed as described in Subheading 3.3.
2. Melt McIlvaine buffer pH 5.5 and Ho work solution.
3. Rehydrate slides by incubating successively in 70, 50, and 30 % ethanol series and in ddH_2O for 5 min.
4. Add a few drops of McIlvaine buffer pH 5.5 to the slides and incubate for 10 min.
5. Gently shake off slide and apply 80–100 μL of Ho working solution to the slide for 2 min. Cover with a plastic cover slip, avoiding air bubbles, and protect from light.
6. Carefully remove the plastic cover slip with tweezers and wash briefly with McIlvaine buffer pH 5.5.
7. Apply McIlvaine buffer pH 5.5 to the whole slides and incubate for 15 min.
8. Gently shake off slides. Add a few drops of ddH_2O to the slides and incubate for 15 min.
9. Shake the water off, dry the slides, and mount it in glycerol antifade solution.
10. Store slides in the dark. For long-term conservation store at 4 °C.
11. Observe under an epifluorescence microscope with appropriate filters.

3.6 Destaining Slides After Fluorochrome Bandings

1. Immerse slides in the Carnoy I in staining dish until cover slips float off.
2. Successively dehydrate slides in ice-cold ethanol series (70, 90, and 100 % ethanol) for 5 min each (*see* **Note 12**).
3. Dry the slides for a couple of days in a vertical position in a closed plastic box to prevent accumulation of dust.

3.7 FISH

Following is the technique of Heslop-Harrison et al. [28] with minor modifications by Siljak-Yakovlev et al. [3].

Day One

1. Prepare a humid chamber using a plastic box with moistened paper tissues in the bottom. Warm up to 37 °C.
2. Add 200 μL of RNase A working solution to each slide, cover with plastic cover slips, and incubate in a humid chamber at 37 °C for 1 h.
3. Carefully remove the plastic cover slips with tweezers and wash slides in a Coplin jar in 2× SSC twice for 5 min.
4. Briefly rinse slides in a 0.01 M HCl solution.
5. Incubate slides with 80–100 μL of pepsin working solution for 10–15 min at 37 °C (*see* **Note 13**).
6. Rinse slides in deionized H_2O for 2 min.
7. Wash slides in 2× SSC two times for 5 min.
8. Facultative step: Denaturation in 50 or 70 % (for Gymnosperms) formamide, 2 min at 70 °C. Rinse slides in 2× SSC for 5 min.
9. Dehydrate slides in an ethanol series: 70, 90, and 100 % (−20 °C), 3 min each.
10. Air dry the slides for 1–2 h.
11. Prepare 50 μL of hybridization buffer consisting of 50 % formamide, 10 % dextran sulfate, 0.6 % sodium dodecyl sulfate, 1.5 μL salmon sperm, and 20× SSC. Per 50 μL of hybridization buffer, add 0.5–2 μL of 18S–26S DNA probe (40 ng/μL) and 0.5–2 μL of 5S DNA probe (50 ng/μL) to obtain a hybridization mix (*see* **Note 14**).
12. Denature the probe by incubating the hybridization mixture in a water bath (or a heat block) at 72 °C for 10 min and transfer immediately on ice for a minimum of 5 min (*see* **Note 15**).
13. Add 50 μL of hybridization mixture to each slide and cover with plastic cover slips. Place slides in a plastic box and incubate in a water bath at 72 °C for 10 min (*see* **Note 16**).
14. Place the box to another water bath set at 55 °C for 5 min.
15. Place slides in a humid chamber and incubate over night at 37 °C (*see* **Notes 17** and **18**).

Day Two

1. Preheat the buffers (0.1× SSC, 2× SSC, 4× SSCT) in a water bath to 42 °C.
2. Carefully remove the plastic cover slips with tweezers and immerse slides in a Coplin jar with 2× SSC buffer for 3 min at room temperature.
3. Wash slides twice in 2× SSC for 5 min at 42 °C.
4. Facultative step to reduce background: wash slides in 20 % formamide at 42 °C, two times for 5 min.
5. Wash slides in 0.1× SSC for 5 min at 42 °C.
6. Wash slides in 2× SSC for 5 min at 42 °C.
7. Wash slides in 4× SSCT for 5 min at 42 °C (*see* **Note 19**).
8. Blocking: apply 100 µL of blocking buffer on the slides, cover with plastic cover slips, and incubate for 5 min at room temperature, protect from light. Carefully remove the plastic cover slips.
9. Antibody detection: apply 50 µL of antibody buffer on the slides, cover with plastic cover slips, and incubate at 37 °C for 1 h in a preheated plastic humid chamber.
10. Carefully remove the plastic cover slips and immerse slides in 4× SSCT buffer three times for 5 min.
11. Shake the buffer off, dry the slides, and counterstain with final antifade mounting medium with DAPI. Leave to stand for 5–10 min and remove surplus medium using paper tissue.
12. Store slides in the dark at 4 °C.
13. Observe under an epifluorescence microscope with appropriate filters.

3.8 Modified FISH Protocol

In this paragraph we present a modified and much shorter version of our standard FISH protocol, which we already used and verified for some genera (e.g., *Crepis*, *Iris*, *Narcissus*, *Quercus*, and *Triticum*).

Day One

1. Prepare a humid chamber by placing moistened paper tissues on the bottom of a plastic box and preheat to 37 °C.
2. Add 200 µL of RNase A working solution on each slide, cover with plastic cover slips, and incubate in a humid chamber at 37 °C for 1 h.
3. Immerse slides in a Coplin jar with 2× SSC buffer and wash twice for 5 min. The plastic cover slips will float off the slides during the first wash.
4. Add 50 µL of proteinase K working solution to the slides and incubate for 15 min at 37 °C (*see* **Note 13**).

5. Wash slides in 2× SSC for 5 min.
6. Dehydrate slides in an ethanol series: 70, 90, 100 % (−20 °C), 3 min each.
7. Air dry the slides for 1–2 h.
8. Prepare a modified hybridization buffer consisting of 50 % formamide, 10 % dextran sulfate, 2× SSC, and 50 mM NaH_2PO_4 pH = 7.0 (*see* **Note 20**). Per 50 µL of hybridization buffer, add 0.5–2 of 18S–26S DNA probe (40 ng/µL) and 0.5–2 µL of 5S DNA probe (50 ng/µL) to obtain a hybridization mix (*see* **Note 14**).
9. Add 50 µL of hybridization mix per slide and cover with a cover slips. Place the slides in a water bath set at 85 °C and incubate for 6 min (*see* **Note 16**).
10. Transfer the slides to a humid chamber and hybridize overnight (16–20 h) at 37 °C.

Day Two

1. Preheat the buffers (0.1× SSC, 2× SSC, 4× SSCT) in a bath at 42 °C.
2. Immerse slides in a Coplin jar with 2× SSC buffer and wash for 5 min at room temperature.
3. Wash slides twice in 2× SSC for 5 min at 42 °C.
4. Wash slides in 0.1× SSC for 5 min at 42 °C.
5. Wash slides in 2× SSC for 5 min at 42 °C.
6. Wash slides in 4× SSCT for 5 min at 42 °C.
7. Wash slides 5 min in 4× SSCT at room temperature.
8. Blocking: apply 100 µL of blocking buffer on the slides, cover with plastic cover slips, and incubate in a humid chamber for 30 min at room temperature, protected from light.
9. Antibody detection: dilute antibody stock solution (200 µg/mL) 1:75 with blocking buffer. Gently lift the plastic cover slips and apply 25 µL of this dilution per slide. Cover with plastic cover slips and incubate in a humid chamber for 1 h at 37 °C.
10. Immerse slides in a Coplin jar with 4× SSCT buffer and wash twice for 5 min at room temperature.
11. Gently shake off excess buffer and counterstain with antifade mounting medium with DAPI. Remove surplus of medium using paper tissue, after 5–10 min. Alternatively, counterstain slides with 0.5 µg/µL DAPI for 8 min. After a brief wash in 2× SSC, apply the antifade solution (AF_2) and cover with a cover glass. Remove excess medium using a paper tissue.
12. Place slides in a dark place, at 4 °C. Observe under an epifluorescence microscope with appropriate filters.

3.9 Destaining Slides After FISH

1. Immerse slides in the 2× SSC in staining dish, until cover slips do not lapse.
2. Dehydrate slides in a graded icy ethanol series (70, 90, and 100 %) for 5 min (*see* **Note 12**).
3. Dry the slides in vertical position in closed plastic boxes to avoid dust for a couple of days.
4. Re-start new FISH experiment on the same slides from step 9 (standard protocol) or 6 (modified protocol) on Day One.

4 Notes

1. Store at 4 °C, in a dark glass bottle, not longer than 2 months.
2. It is necessary to use fresh solutions to minimize ester formation, stop mitosis, and preserve chromosome structure integrity.
3. This solution is recommended for oily and waxy tissues to increase the penetration ability of the fixative.
4. It is possible to use $MgCl_2$ instead of $MgSO_4$: add 0.1017 g of $MgCl_2 \cdot 6H_2O$.
5. Proposed enzyme composition and concentrations may require modification for different plant species.
6. Put powder in the buffer and keep at 37 °C for a couple of minutes (without shaking) for easier and faster dissolution.
7. Detection step is not needed if FISH is done with directly labelled probes.
8. For long-term preservation, keep material in the Carnoy fixative (4 °C) for a few days and then transfer it to 70 % ethanol fixative and store at 4 °C or −20 °C.
9. In case of large chromosomes and low number of available root tips, avoid centrifugation protocol. Enzyme mixtures can be reused several times in which case digestion time might need to be slightly increased after each round of use.
10. Exposure time of meristems to enzyme mixture depends on tissue thickness. It is necessary to verify the homogeneity and successive decomposition of meristems of analyzed species: material should be soft and break up easily for optimal time.
11. When using a freezer, preparations need to stay at −80 °C at least 24 h to avoid chromosome loosing during cover slips removing.
12. Store alcohol solutions at −20 °C.
13. Incubation in pepsin and proteinase K should be prolonged in case of larger amounts of cytoplasm on the slides.
14. Calculate the required amount of ddH_2O to the final volume of 50 μL/slide. The probes should be added last and the hybridization mix should be homogenized by vortexing.

15. Rapid cooling of the hybridization mix prevents reannealing of the probe.
16. The exact temperature and duration of treatment varies between species and should be experimentally determined if not already published.
17. It is important to prevent moisture loss by evaporation. However, too much moisture can lead to condensation on the slides which can result in poorly hybridized slides.
18. Duration of hybridization should be prolonged for gymnosperms to up to 48 h.
19. During this step thaw blocking buffer.
20. Hybridization buffer can be prepared in excess volume and stored at −20 °C.

References

1. Murata M, Heslop-Harrison JS, Motoyoushi F (1997) Physical mapping of the 5S ribosomal RNA genes in *Arabidopsis thaliana* by multicolor fluorescence *in situ* hybridization with cosmid clones. Plant J 12(1):31–37
2. Cerbah M, Kevei Z, Siljak-Yakovlev S et al (1998) rDNA organization and heterochromatin pattern in *Medicago truncatula*. Cytogenet Cell Genet 81:141
3. Siljak-Yakovlev S, Cerbah M, Couland J et al (2002) Nuclear DNA content, base composition, heterochromatin and rDNA in *Picea omorika* and *Picea abies*. Theor Appl Genet 104:505–512
4. Cerbah M, Coulaud J, Siljak-Yakovlev S (1998) rDNA organization and evolutionary relationships in the genus *Hypochaeris* (Asteraceae). J Hered 89:312–318
5. Weiss-Schneeweiss H, Tremetsberger K, Schneeweiss GM, Parker JS, Stuessy TF (2008) Karyotype diversification and evolution in diploid and polyploid South American *Hypochaeris* (Asteraceae) inferred from rDNA localization and genetic fingerprint data. Ann Bot 101: 909–918
6. Zoldos V, Papes D, Cerbah M et al (1999) Molecular-cytogenetic studies of ribosomal genes and heterochromatin reveal conserved genome organization among eleven *Quercus* species. Theor Appl Genet 99:969–977
7. Siljak-Yakovlev S, Peccenini S, Muratovic E et al (2003) Chromosomal differentiation and genome size in three European mountain *Lilium* species. Plant Syst Evol 236:165–173
8. Lim KY, Matyásek R, Lichtenstein CP et al (2000) Molecular cytogenetic analyses and phylogenetic studies in the *Nicotiana* section Tomentosae. Chromosoma 109:245–258
9. Bogunic F, Siljak-Yakovlev S, Muratovic E et al (2011) Different karyotype patterns among allopatric *Pinus nigra* (Pinaceae) populations revealed by molecular cytogenetics. Plant Biol 13:194–200
10. Bogunic F, Siljak-Yakovlev S, Muratovic E et al (2011) Molecular cytogenetics and flow cytometry reveal conserved genome organization in *Pinus mugo* and *P. uncinata*. Ann For Sci 68(1):179–187
11. Hizume M, Aria M, Tanaka A (1990) Chromosome banding in the genus *Pinus*. III. Fluorescent banding pattern of *P. luchuensis* and its relationships among the Japanese diploxylon pines. Bot Mag Tokyo 103:103–111
12. Bogunic F, Muratovic E, Siljak-Yakovlev S (2006) Chromosomal differentiation of *Pinus heldreichii* and *Pinus nigra*. Ann For Sci 63:267–274
13. Muratovic E, Robin O, Bogunic F et al (2010) Speciation of European lilies from *Liriotypus* section based on karyotype evolution. Taxon 59:165–175
14. Godelle B, Cartier D, Marie D et al (1993) Heterochromatin study demonstrating the non-linearity of fluorometry useful for calculating genomic base composition. Cytometry 14:618–626
15. Sell PD (1976) Crepis. In: Tutin TG et al (eds) Flora Europaea 4. Cambridge University Press, Cambridge, pp 344–357
16. Siljak-Yakovlev S, Cartier D (1986) Heterochromatin patterns in some taxa of *Crepis praemorsa* complex. Caryologia 39:27–32
17. Cartier D, Siljak-Yakovlev S (1992) Cytogenetics study of the F1 hybrids between *Crepis dinarica* and *Crepis froelichiana*. Plant Syst Evol 182:29–34

18. Schwarzacher T, Heslop-Harrison P (2000) Practical *in situ* hybridization, 2nd edn. BIOS, Oxford, UK
19. Schubert I, Wobus U (1985) *In situ* hybridization confirms jumping nucleolus organizing regions in *Allium*. Chromosoma 92: 143–148
20. Raina SN, Mukai Y (1999) Detection of a variable number of 18S-5.8S-26S and 5S ribosomal DNA loci by fluorescent *in situ* hybridization in diploid and tetraploid *Arachis* species. Genome 42:52–59
21. Raskina O, Belyayev A, Nevo E (2004) Quantum speciation in *Aegilops*: molecular cytogenetic evidence from rDNA cluster variability in natural populations. Proc Natl Acad Sci USA 101:14818–14823
22. Datson PM, Murray BG (2006) Ribosomal DNA locus evolution in *Nemesia*: transposition rather than structural rearrangement as the key mechanism? Chromosome Res 14: 845–857
23. Geber G, Schweizer D (1988) Cytochemical heterochromatin differentiation in *Sinapis alba* (Cruciferae) using a simple air-drying technique for producing chromosome spreads. Plant Syst Evol 158:97–106
24. Conger AD, Fairchild LM (1953) A quick freeze method for making smear slides. Stain Technol 28:281–283
25. Schweiser D (1976) Reverse fluorescent chromosome banding with chromomycin and DAPI. Chromosoma 58:307–324
26. Kondo T, Hizume M (1982) Banding for the chromosomes of *Cryptomeria japonica* D. Don. J Jpn For Soc 64:356–358
27. Martin J, Hesemann CU (1988) Evaluation of improved Giemsa C- and fluorochrome banding techniques in rye chromosome. Heredity 6:459–467
28. Heslop-Harrison LS, Schwarzacher T, Anamthawat-Jonsson K et al (1991) *In situ* hybridization with automated chromosome denaturation. Techniques 3:109–116

Chapter 16

GISH: Resolving Interspecific and Intergeneric Hybrids

Nathalie Piperidis

Abstract

Genomic in situ hybridization (GISH) is an invaluable cytogenetic technique which enables the visualization of whole genomes in hybrids and polyploidy taxa. Total genomic DNA from one or two different species/genome is used as a probe, labeled with a fluorochrome and directly detected on mitotic chromosomes from root-tip meristems. In sugarcane we were able to characterize interspecific hybrids of two closely related species as well as intergeneric hybrids of two closely related genera.

Key words GISH, Fluorochrome, Interspecific, Intergeneric, Genome

1 Introduction

1.1 Genomic In Situ Hybridization (GISH)

GISH was derived from fluorescent in situ hybridization (FISH) techniques developed in the early 1980s by biomedical researchers and it was eventually applied to plant chromosomes in the late 1980s. GISH was first demonstrated in synthetic *Hordeum chilense* × *Secale africanum* [1] and also used to track artificial introgression of chromosomes in wide crosses [2]. The challenges faced by plant chromosome researchers are mainly based on the fact that plants have cell walls and cytoplasmic debris and more condensed chromosomes at privilege stage than in the mammalian cells. GISH is a powerful tool and can be used, for example, to distinguish the genome of one parent from the other by preferential labeling of the genome of one parent. It can also be used to detect alien chromosome(s) in addition lines or alien species in recipient parent, for example. GISH is extremely useful to identify parental chromosomes in interspecific or intergeneric hybrids, to test the origin of natural amphiploids, to track down the introgression of alien chromosomes, or to test the occurrence of exchange between the genomes involved [3, 4]. Multicolor GISH is allowing simultaneous discrimination of multiple genome and identification of diploid progenitor in allopolyploids. GISH requires genomic DNA labeled directly with a fluorochrome or with a hapten capable of

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_16,

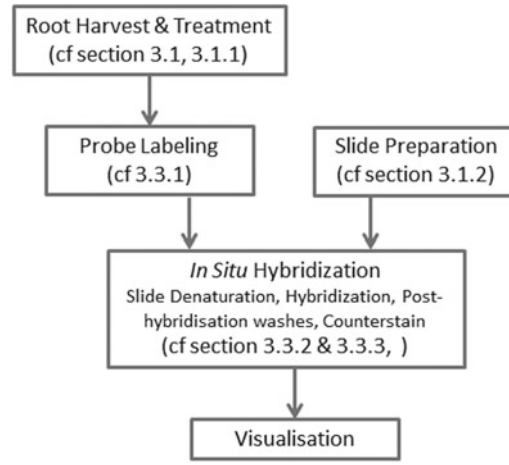


Fig. 1 Overview of the GISH procedure

indirect association with fluorochromes. Nucleic acid of the fluoroprobe(s) will then provide an assay through complementarity pairing with nucleotides of the target DNA on a slide. Fluorochromes provide the ability to visualize in situ homologous regions to the probe within the cellular structure using a fluorescence microscope. Digital camera coupled to the microscope allows the capture of permanent images of the fluorescent patterns on the chromosomes. Figure 1 represents the outline of the procedure.

1.2 Example of Application in Sugarcane

Although classical cytological studies in sugarcane [5] allowed a better understanding of the sugarcane genome, molecular cytogenetic methods lead to important breakthroughs, not only revealing the level of the complexity of modern sugarcane cultivars but also unraveling the taxonomy of the *Saccharum* genus. Modern sugarcane cultivar is one of the most difficult species to work with on a genetic and molecular level. Sugarcane species are considered to have one of the most complicated genome studied. Chromosome numbers were determined, uncovering highly polyploid and, frequently, aneuploid members in this genus [6]. The genome of modern cultivar results in the hybridization of two species of *Saccharum*, the noble cane *S. officinarum* and the wild species *S. spontaneum* which was also revealed by GISH studies. In the past 15 years, molecular cytogenetic techniques have been proven to be a very efficient tool to better understand this complex genome and reveal outcomes that classical molecular markers alone could not. These techniques proved to be particularly relevant to refine our understanding of the genome structure of sugarcane and its taxonomy [7, 8]. In our laboratory we used GISH to characterize interspecific hybrids to taxonomic reclassification of atypical *S. officinarum* as well as intergeneric hybrids involving two different genomes, *Saccharum* and *Erianthus*.

1.2.1 Interspecific Hybrid Between S. officinarum and S. spontaneum

Since the original classification of *Saccharum* species, taxonomy within the genus *Saccharum* has been controversial. *S. officinarum* is known to have $2n=80$ chromosomes; therefore, clones with more than 80 chromosomes should be classified as hybrids. However, Irvine [9] has debated this and suggested that clones that fit the botanical description for *S. officinarum* with more than 80 chromosomes should remain in this classification. GISH studies have contributed to understanding the taxonomic status and relationships of species and clones within the *Saccharum* genus. We used GISH to verify the taxonomic reclassification of atypical *S. officinarum* with more than 80 chromosomes revealed by flow cytometry [7]. GISH results of atypical *S. officinarum* clone Muntok Java are presented in Fig. 2a. Genomic DNA from *S. officinarum* was labeled in "red" with the Alexa Fluor 594-5-dUTP or Rhodamine 5-d-UTP and genomic DNA from *S. spontaneum*

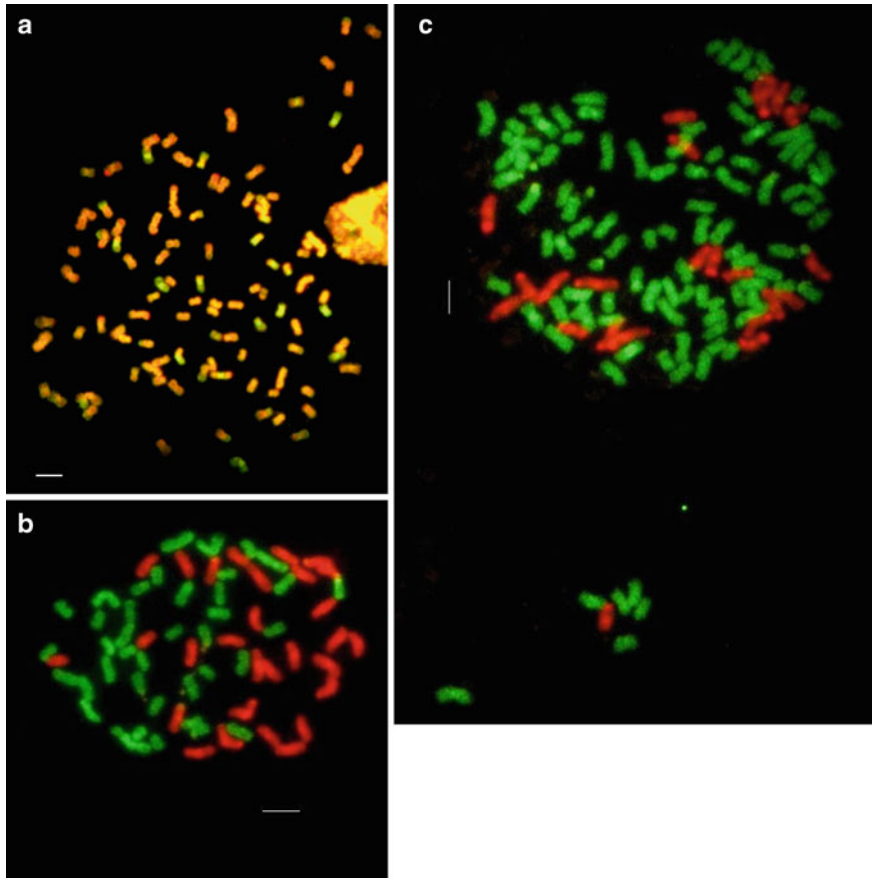


Fig. 2 (**a**) Interspecific chromosome composition of an atypical *S. officinarum* revealed by GISH using total genomic DNA for *S. officinarum* (in *orange*) and total genomic DNA from *S. spontaneum* (in *green*); recombined chromosomes appeared in *both color*. B. C. Intergeneric chromosome composition of an F1 (**b**) and a BC1 (**c**) revealed by GISH using total genomic DNA from *S. officinarum* (in *green*) and *E. arundinaceus* (in *red*)

was labeled in "green" with the Alexa Fluor 488-5-dUTP or Fluorescein 12-d-UTP. Both species are relatively closely related; therefore, *S. officinarum* chromosomes appear "orange" while *S. spontaneum* chromosomes appear "yellow-green," and recombined chromosomes from both species can also be visualized.

1.2.2 Intergeneric Hybrid Between S. officinarum and E. arundinaceus

For our intergeneric GISH characterization, *Erianthus arundinaceus* was labeled in "red" with the Alexa Fluor 594-5-dUTP or Rhodamine 5-d-UTP while the other genera *S. officinarum* was labeled in "green" with the Alexa Fluor 488-5-dUTP or Fluorescein 12-d-UTP [8]. GISH of an F1 and back-cross 1 (BC1) between the 2 genomes are presented in Fig. 2b, c. In these intergeneric hybrids, *E. arundinaceus* appear "red" and the *S. officinarum* chromosomes appear "green" as the two species are not as closely related than in the interspecific hybrids. The fluorochrome colors do not overlap as the genome has minimal cross-hybridization.

2 Materials

Prepare all stock solutions using deionized distilled water (ddH_2O) and chemicals with the highest grade available. For most steps in DNA handling, it is essential that ddH_2O is autoclaved for at least 20 min at 130 °C in order to destroy any DNase activity and ensure sterility. All stock solutions have to be stored at room temperature (RT) unless stated otherwise.

2.1 Equipment

Besides the laboratory standard equipment, few specialized items are needed.

1. CCD (charge-couple device) camera with image capture and processing software.
2. Coplin jars.
3. Epifluorescence/light microscope.
4. Heating plate with magnetic stirrer.
5. Hot plate with digital temperature control for slide warming.
6. Refrigerated centrifuge.

2.2 Stock Solutions Stored at Room Temperature

1. Dextran sulfate (DS): Dissolve 5 g of DS to a final volume of 10 mL of ddH_2O. Stir slowly until dissolved; it could take up to 24 h for the DS to be completely dissolved.
2. 2× SSC (saline sodium citrate) pH 7.0: Dilute 100 mL of 20× SSC pH 7.0 in 900 mL for a final volume of 1,000 mL.

2.3 Stock Solutions Stored at 4 °C

1. Antifade for mounting slide: Vectashield mounting media with DAPI (*see* **Note 1**).
2. 0.25 N HCl: Always work under fume hood, measure 195.56 mL of ddH_2O, and then add 4.44 mL of pure HCl (*see* **Note 1**).

3. 0.04 % 8-Hydroxiquinoline: Add 40 mg of 8-Hydroxiquinoline to 100 mL of ddH_2O. Place on a stirrer at RT for several hours. Store at 4 °C up to 1 year (*see* **Note 2**).
4. 3 M NaOAc (CH3COONa) pH 5.2: Dissolve 40.81 g of sodium acetate×$3H_2O$ in 30 mL ddH_2O, titrate pH to 5.2 with glacial acetic acid, and dilute with ddH_2O to a final volume of 100 mL.
5. TE buffer pH 8.0 (10 mM Tris–HCl pH 8.0, 1 mM Na_2EDTA): Dilute 20 µL of 1 M Tris–HCl pH 8.0 + 4 µL of 500 mM Na_2EDTA pH 8.0 with 1,976 µL ddH_2O.

2.4 Stock Solutions Stored at –20 °C

1. BIOPRIME DNA labeling system for Random Priming labeling.
2. 1 µg/µL carrier DNA Sheared Salmon sperm DNA (SS DNA): Mix 10 mg of DNA with 10 mL of TE pH 8.0. Shear in autoclave for 5 min, denature for 10 min in boiling water, and then place on ice. Aliquot and store.
3. Deionized formamide (FA): (*see* **Note 1**). Work under the fume hood. Add 5 g of ion exchange resin for each 100 mL formamide, cover with aluminum, and stir for 30–60 min. Filter twice with Whatman No. 1. Aliquot in 1 mL tubes as well as in 20 mL tubes and store. Deionize all formamide when a new bottle is opened. Do not keep FA after opening.
4. Digestion citrate buffer: Add 1.47 g of trisodium citrate-dihydrate ($Na_3C_6H_5O_7$×$2H_2O$) + 1.05 g of citric acid-monohydrate ($C_6H_8O_7$×H_2O) + 2.8 g of KCl and add ddH_2O up to 500 mL. Adjust pH to 4.5, aliquot, and store.
5. Digestion enzyme solution: 0.25 g (5 % final concentration) of cellulase Onozuka R-10 (Yakult Honsha Co. Ltd., Japan) + 0.05 g (1 % final concentration) of pectolyase Y-23 (Seishin Pharmaceutical Ltd., Japan) +5 mL of digestion citrate buffer. Place on stirrer at RT for 1 h. Aliquot into microtubes and freeze.
6. Ethanol series: Prepare three solutions at 70, 95, and 100 % ethanol in three Coplin jars and place at –20 °C.
7. 70 % FA/2× SSC: 35 mL FA + 15 mL of 2× SSC (*see* **Note 1**).
8. Fluorochromes: 1 mM F-x-dUTP (*see* **Note 3**): ChromaTide Alexa 594-5-dUTP; ChromaTide Alexa 488-5-dUTP; Fluorescein-12-dUTP; Rhodamine-5-dUTP.
9. dNTP for Random Priming: dATP, dCTP, dGTP, dTTP (100 mM). Dilute all of the individual dNTP at 10 mM final concentration (10 µL of dNTP + 90 µL of ddH_2O).
10. 10x dNTP Fluorochromes (10 mM) mix: On ice add 5 µL of dATP, dGTP, dCTP + 2.5 µL of dTTP + 25 µL dUTP-Alexa + 7.5 µL of ddH_2O. Keep at –20 °C for up to 6 months.

11. 10x dNTP Fluorescein and/or Rhodamine mix: On ice add 5 µL of dATP, dGTP, dCTP + 3.25 µL of dTTP + 17.5 µL dUTP-Alexa + 14.25 µL of ddH_2O. Keep at −20 °C for at least 6 months.
12. 1 % RNase A in 10 mM Tris–HCl pH 7.5, 15 mM NaCl (DNase-free): Dissolve 10 mg of RNase A + 10 µL of 1 M Tris–HCl pH 7.5 + 3 µL of NaCl in 987 µL ddH_2O; incubate in boiling water bath for 15 min, cool slowly, and store in aliquots.

3 Methods

3.1 Root Pretreatment and Slide Preparation

Root-tip meristems are the most commonly used plant tissues used in cytogenetic methods for preparing mitotic chromosomes as they contain cells in active division. Most mitotic chromosome preparations are made from root-tip meristems. Plants are grown in a glass house in 20 L pots with a mixture of 50/50 vermiculate (coarse grade)/perlite (grade 3) with regular and sufficient application of water and nutrients (*see* **Note 4**). Root-tip collection includes a pretreatment in order to arrest as many metaphase cells as possible, a fixative treatment, and then the roots can be stored in a 70 % ethanol solution at 4 °C. For species with low mitotic index, it could be important to estimate the time of the day where best mitotic index slides are obtained. It is usually recommended to set up an assay where quality/mitotic index of slides is recorded in function of the collection time. The harvesting should be conducted by 1/2 h periods over an 8 h day. For example, in sugarcane, we harvest roots between 10:30 h and 11:00 h during the optimal growth period days of October to December [7].

3.1.1 Root Treatment

1. Approximately 0.5 cm of roots are harvested with fine forceps and placed directly in 5 mL bottles containing 0.04 % 8-Hydroxyquinoline for 4 h at RT to arrest cells at metaphase (*see* **Note 1**).
2. Fix in freshly made 3:1 (3 volumes of 100 % ethanol to 1 volume of glacial acetic acid) fixative for 72 h at RT.
3. Store roots in 70 % ethanol at 4 °C until ready for use.

3.1.2 Slide Preparation (See **Note 5**)

1. Rinse roots twice in ddH_2O for 10 min at RT.
2. Hydrolyze roots in 0.25 N HCl for 10 min at RT.
3. Rinse roots in ddH_2O for 10 min at RT.
4. Place roots in digestion citrate buffer for 10 min at RT.
5. Cut the distal 1–1.5 mm of the root tip with a fine scalpel; blot away excess moisture with filter paper.

6. Digest root tips in digestion enzyme solution for 90–180 min in a water bath at 37 °C. The length of time will vary with species and/or size of the root tips.
7. Carefully remove root tips from tubes and place in ddH_2O in a watch glass for at least 20 min at room temperature. Time must be optimized and the root cap must be removed to avoid high background.
8. Use a Pasteur pipette to carefully remove one root tip and place it on a pre-cleaned slide (*see* **Note 6**).
9. Add one or two drops of freshly made 3:1 fixative, immediately break apart the tip, and spread it with a pair of fine forceps (*see* **Note 7**).
10. Air-dry and store overnight in a desiccator (37 °C).

3.2 RNase A Treatment

Prior any GISH experiment, slides are screened to select the one with the best mitotic chromosome cells. Therefore, to avoid disappointment and reduce the cost of GISH if you work with species prominent to low mitotic index (e.g., as in sugarcane), we recommend to only hybridize slides with good mitotic preparations, i.e., with at least five "complete" 2*n* cells. We also recommend taking note of the coordinates of the good mitotic cells for tracking purposes to be able to find the good cell when capturing images. We are also delimiting the hybridization area (with a diamond pen on the back of the slide) for a targeted and more efficient use of the hybridization buffer.

1. Thaw RNase A on ice prior use and remove an 8-μL aliquot.
2. Dilute (1/100) the thawed RNase A aliquot on ice: 8 μL of 1 % RNase A + 80 μL of 20× SSC + 712 μL ddH_2O.
3. Add 50–100 μL of diluted RNase A onto one slide, cover with a plastic cover slip (*see* **Notes 8** and **9**), and incubate in a humidified incubation chamber (*see* **Note 10**) for 45 min at 37 °C.
4. Rinse slides in a Coplin jar in 2× SSC for 10 min at RT.

3.3 GISH Experiment

The method described here for GISH experiment involves Random Priming labeling methods with direct fluorochrome. This method is the preferred method in our laboratory as it is very simple and reliable in order to acquire relatively quick and efficient results. There are alternative options to perform GISH in plants. Different methods such as Nick Translation (NT) labeling with different types of haptens are extensively described in Zhang and Friebe [10]. One of the most common methods for GISH is NT labeling with Biotine and/or Digoxygenin. These haptens will have to be detected and amplified in order to visualize the fluorescent signal.

3.3.1 Probe Labeling by Random Priming

Random Priming achieves best result with good quality DNA. A mixture of different combination of hexamers, octamers, or nonamers is annealed randomly to denatured DNA. The annealed small oligonucleotides will then act as primers and allow the synthesis of the complementary DNA strand by the Pol1 fragment of the Klenow enzyme (Pol1 has a DNA polymerase activity as well as exo-nucleasic activity $3' \rightarrow 5'$). Labeled DNA will consist of a mixture of double- and single-stranded fragments. We used the kit BIOPRIME DNA labeling system with the "green" and "red" Alexa fluorochromes (F-x-dUTP) or with Fluorescein 12-d-UTP and Rhodamine 5-d-UTP (*see* **Notes 11** and **12**).

1. On ice, firstly dilute 1 μg of genomic DNA to a volume of 19 μL ddH_2O and add 20 μL of Random Primers (from the kit). Denature the 39 μL in boiling water for 6 min and stand on ice for 15 min.
2. Finally add 10 μL of 10× dNTP mix + 1 μL of Klenow enzyme. Mix gently, centrifuge briefly, and incubate in a water bath from 5 h to overnight at 37 °C. Longer incubation times usually increase product yield.
3. Add 5 μL of stop buffer.
4. Removal of unincorporated nucleotides and primers is not essential but can be performed by adding 1/10 volume of 3 M NaOAc and 2.5 volumes of 100 % ethanol and centrifuging at 15,000 × *g* for 30 min at 4 °C. Discard supernatant and add 250 mL of 70 % ethanol; centrifuge for 15 min. Discard supernatant carefully. Air-dry tubes for 5 min and resuspend in 20 μL of TE at 37 °C for 5 min. The concentration of the probe should be around 40–50 ng/μL (*see* **Note 12**).
5. The fluorescence of fluorochrome-labeled probes can be estimated by a spot test as follows. Spot 1 μL of fluorochrome-labeled probe onto a small piece of nylon membrane, let air-dry for approximately 10 min, and then examine the fluorescence intensity under a fluorescence microscope with a suitable filter.
6. Probes can be stored at –20 °C.

3.3.2 Slide Denaturation

Chromosomes are denatured by placing slides on a hot plate at 80 °C in order to be ready for the in situ hybridization.

1. Set a hot plate at 80 °C for at least 30 min prior to the denaturation process. We use a digital hot plate for better temperature accuracy (*see* **Note 13**).
2. Apply 200 μL of a solution of 70 % FA/2× SSC, apply cover slip, and place on hot plate for 3 min at 80 °C (*see* **Note 1**). Denaturation time has to be optimized according to the species.

3. Remove cover slip (*see* **Note 8**) and rinse slides in a Coplin jar standing in ice with 2× SSC (at –20 °C) for 3 min.
4. Dehydrate slides through an ethanol series of 70 %, 95 %, and finally 100 % at –20 °C, for 5 min each step.
5. Air-dry vertically (*see* **Note 14**).

3.3.3 In Situ Hybridization

1. Prepare 50 µL of hybridization buffer (HB) per slide: 25 µL FA, 10 µL DS, 5 µL 20× SSC, 1.5 µL SS DNA, 80–100 ng of each DNA probe, and make up to a final volume of 50 µL with ddH_2O (*see* **Notes 3** and **15**).
2. Denature the HB for 10 min in boiling water and then place on ice for at least 15 min.
3. Deposit 50 µL of the HB on dried slide, and cover with a plastic cover slip. Avoid bubbles.
4. Place slides in a humidified incubation chamber (*see* **Note 10**) overnight at 37 °C.
5. Prepare three Coplin jars with 2× SSC, 0.5× SSC, and 0.1× SSC in a 42 °C water bath for stringency washes.
6. Remove cover slips with a squirt of 2× SSC, and wash slides in the 2× SSC for 10 min, then in the 0.5× SSC for 10 min, and finally wash with agitation in the last Coplin jar (0.1× SSC) for another 10 min at RT (*see* **Note 16**).
1. Drain slightly one slide at a time without letting it dry. Counterstain the slides with a drop of antifade vectashield mounting media with DAPI (*see* **Notes 15** and **17**). Cover with a cover glass. Seal cover slip with transparent nail polish. Dry slides horizontally in a slide holder protected from light. Observe under fluorescence microscope with appropriate filter. Store slides horizontally in the dark at 4 °C.

4 Notes

1. Some chemicals especially HCl, FA, DAPI, and glacial acetic acid are dangerous and should be handled with extreme caution. Some product such as FA become even more dangerous when heated, so always follow good laboratory practice and use the fume hood when required.
2. 8-Hydroxyquinoline is sensitive to light. It is therefore important to store the solution in the dark in a bottle covered with aluminum foil. At least ½ h before using the solution, it is best to place the bottle on a stirrer. Finally just before root collection, fill up 5 mL bottles and keep bottles in a box away from the light to ensure a good efficiency of the active product.

3. Fluorochromes will photo-bleached if exposed to light for long periods of time. During probe labeling preparation, it is recommended to work with a benchtop lamp.
4. Ensure that at least for 4 h prior harvesting, the roots are not being watered; they will be more accessible if the pots are not soaked. Good size roots are collected approximately every 3 weeks; if roots are not growing properly, use nutrients specialized in root growth.
5. Root treatment for the slide preparation can be performed in the bottle and the storage solution is removed completely with plastic pipettes. If there is more than one clone/species to be treated, we use a microplate with 24 wells. We treat 2–15 roots from four different species per plate. Each line has 6 wells with 5 in use containing, respectively, ddH_2O, ddH_2O, HCl, ddH_2O, and digestion buffer. Roots are handled carefully with tweezers in each bath for the ten required minutes. Root tips/sample are then cut and grouped by size before being set up in digestion enzyme solution. After 90 min the first lots of tips are placed in the same microplate which only contains ddH_2O. The remaining tips are left in the water bath until ready to be spread.
6. Slides are placed in Coplin jars with 100 % ethanol and dried just before use with Kimwipes. Excess water is removed with a homemade micro Pasteur pipette firstly and then we use the folded Kimwipes used to pre-clean the slide. The Kimwipes with residual ethanol will suck the remaining water around the root.
7. If chromosomes on the slide have too much cytoplasm/too much cell wall debris, make sure that the root cap was removed before spreading as it increases the quality of the slide preparation. The root cap does not normally detach itself from the tip and tweezers are most of the time required to remove the cap at this stage without damaging the tip itself. The digested cap-free tip has to be spread evenly on a 32 mm x 40 mm surface of the slide to concentrate the metaphasic chromosomes to a small area. Avoid spreading twice in the same localization.
8. We use precuts of autoclave bags for plastic cover slip as they handle high temperature well and also as it seems that they do not trap too many bubbles.
9. To remove the cover slip, a squirt bottle of 2× SSC is recommended.
10. Our humidified incubation chamber consists of a large Petri dish lined with paper at the bottom and soaked with water. The slides are set on plexiglass stick or bended Pasteur pipette so they are not directly in contact with the water.

11. To ensure better result during ethanol precipitation, we are using 100 % ethanol at –20 °C, and after adding acetate sodium and ethanol, we leave the tube at –20 °C for 2 h or at –80 °C for 15 min. We also use a refrigerated centrifuge with a swinging bucket as the pellet of DNA would be precipitated at the bottom of the tube. We also preferably use screw cap tubes. After ethanol precipitation DNA pellets labeled with a red fluorochrome are usually readily seen by the eyes whereas DNA labeled with the green fluorochrome are usually of a pale shade of yellow and could be not so easy to see. Before resuspending the probe, make sure that all the ethanol has been removed from the tube. Centrifuging tubes for another minute at $10{,}000 \times g$ can get rid of the excess ethanol as residual ethanol in probe and slides could result in higher background signal.
12. If your slides present no, weak, or patchy hybridization, it is often the results of labeling problems. Check the quality of the DNA before labeling as good quality DNA will give a better probe and the length of the probe is also essential.
    Also check the expiry date of the enzymes and dUTPs being used.
13. If you encounter a poor signal from the probe as well as from the counterstain and chromosome morphology appears abnormal, try denaturing the slide a little less than the recommended 3 min and ensure that the temperature of the hot plate is $\leq$ to 80 °C.
14. The slides can be put in an oven at 37 °C to reduce drying time. Residual ethanol in slide can cause higher background signal.
15. It is recommended to avoid as much as possible exposure to light during the entire procedure (labeling, hybridization, post-hybridization wash, image capture). The laboratory should be entirely dark except from a benchtop lamp. When capturing images, be as quick as possible because each exposure to fluorescent light will remove energy from the fluorochrome and therefore decrease its intensity.

    If the hybridization signal is poor, the concentration of the probe used during hybridization might be too low. Try different concentrations of the probe but ensure that the concentration of the probe after precipitation has not been overestimated. Also make sure that the hybridization solution was mixed thoroughly as the DS solution is very viscous. It is possible to use a special piston-pipette or a normal pipette with a cutoff pipette tip to slowly mix the solution up and down.

    Finally, ensure that no bubbles remain between the slide and cover slip after adding the hybridization solution. If bubbles appear, use fine tweezers to lift up and down the cover slip to carefully remove them.

16. Post-hybridization washes are very important to remove unattached probe and therefore reduce the background signal. Ensure that washes are performed according to the procedure.
17. After applying a drop of mounting media, we apply gentle pressure on the glass cover slip in order to remove excess media. We used a layer of Kimwipes directly on the cover slip and three layers of Whatman paper. Always apply pressure with the thumb when slides are placed on a flat surface in order to prevent breaking the slide and/or cover slip.

Acknowledgements

This work was financially supported by BSES Ltd and also previously by the Australian Centre for International Agricultural Research and the Co-operative Research Centre for Sugar Industry Innovation through Biotechnology.

References

1. Le HT, Armstrong KC, Miki B (1989) Detection of rye DNA in wheat-rye hybrids and wheat translocation stocks using total genomic DNA as a probe. Plant Mol Bio Rep 7:150–158
2. Schwarzacher T, Leitch AR, Heslop-Harrison JS (1989) In situ localization of parental genomes in a wide hybrids. Ann Bot 64: 315–324
3. Jiang J, Gill BS (1994) Different species-specific chromosome translocations in *Triticum timopheevii* and *T. turgidum* support the diphyletic origin of polyploid wheats. Chromosome Res 2(1):59–64
4. Jiang J, Gill BS (2006) Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. Genome 49:1057–1068
5. Sreenivasan TV, Ahloowalia BS, Heinz DJ (1987) Cytogenetics. In: Heinz DJ (ed) Sugarcane improvement through breeding. Elsevier, New York, pp 211–253
6. D'Hont A, Ison D, Alix K, Roux C, Glaszmann J-C (1998) Determination of basic chromosome number in the genus *Saccharum* by physical mapping of ribosomal RNA genes. Genome 41:221–225
7. Piperidis N, Chen JW, Deng HH, Wang LP, Jackson P, Piperidis G (2010) GISH characterization of *Erianthus arundinaceus* chromosomes in three generations of sugarcane intergeneric hybrids. Genome 53:331–336
8. Piperidis G, Piperidis N, D'Hont A (2010) Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Mol Gen Genomics 284:65–73. doi:10.1007/s00438-010-0546-3
9. Irvine JE (1999) *Saccharum* species as horticultural classes. Theor Appl Genet 98:186–194
10. Zhang P, Friebe B (2009) FISH on plant chromosomes. In: Liehr T (ed.) Fluorescence in situ hybridization (FISH)—application guide. Springer Protocol VI, pp 365–394, doi: 10.1007/978-3-540-70581-9_32

Chapter 17

On the Relevance of Molecular Tools for Taxonomic Revision in Malvales, Malvaceae s.l., and Dombeyoideae

Timothée Le Péchon and Luc D.B. Gigord

Abstract

In this article, we present an overview of changes to the taxonomy of Malvales. In traditional classifications, this order was variously circumscribed as including four main families (i.e., Malvaceae, Bombacaceae, Sterculiaceae, and Tiliaceae, also known now as "Core Malvales"), but major disagreements existed between different taxonomic treatments. Contributions from molecular data, new morpho-anatomical data, and progress in methodological approaches have recently led to a new broader concept of this order (namely, "expanded Malvales"). Now, expanded Malvales includes ten families (Neuradaceae, Thymelaeaceae, Sphaerosepalaceae, Bixaceae, Cistaceae, Sarcolaenaceae, Dipterocarpaceae, Cytinaceae, Muntingiaceae, Malvaceae s.l.) distributed among seven monophyletic lineages. All these families were previously considered to have malvalean affinities in some traditional treatments, except the holoparasitic and highly modified Cytinaceae. Although molecular evidence has clarified the Malvales taxonomy, the phylogenetic positions of Sarcolaenaceae, Thymelaeaceae, and Sphaerosepalaceae are still controversial and need new analyses focusing specifically on these families to assess their phylogenetic placement and their morphological evolution.

In a phylogenetic context, molecular data combined with recent examination of morphological characters supported the hypothesis of a common origin of "core Malvales." However, these analyses also showed that the former families but Malvaceae s.s. were paraphyletic or polyphyletic. As a consequence, recent taxonomic treatments grouped taxa formerly included in "Core Malvales" in a broader concept of Malvaceae s.l. Additionally, the intrafamilial taxonomy has been deeply modified, and in its present circumscription, Malvaceae includes nine subfamilies (Grewioideae, Byttnerioideae, Sterculioideae, Dombeyoideae, Brownlowioideae, Tilioideae, Bombacoideae, Malvoideae, Helicteroideae) in two main lineages. Phylogenetic studies on subfamilial rearrangements have focused on the relationships between emblematic taxa such as Bombacoideae and Malvoideae (which form together the /Malvatheca lineage). However, our understanding of the phylogenetic relationships among and within taxa of the other subfamilies (e.g., Dombeyoideae, Tilioideae, and Sterculioideae) has not followed at the same pace. Despite recent investigations, the relationships between the subfamilies of Malvaceae s.l. remain controversial. As an example of these taxonomic issues, we review the systematic studies on Dombeyoideae, with special emphasis on taxa endemic to the Mascarene archipelago (Indian Ocean). Recent investigations have shown that several island endemic genera such as *Trochetia*, *Ruizia*, and *Astiria* (endemic to the Mascarenes) are nested within the mega-genus *Dombeya*. Consequently, the current taxonomy of this genus does not match the phylogeny and should be modified. Therefore, we propose three possible taxonomic schemes as part of an ongoing revision of the Mascarene Dombeyoideae. However, these taxonomic rearrangements should only be made after a broader study of the diversity in Madagascar and adjacent areas. This broader approach shall avoid possibly multiple and contradictory taxonomic revisions of restricted regions if they were each studied in isolation.

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_17,

Key words Taxonomy, Systematics, Molecular data, Morphology, Anatomy, Phylogeny, Core Malvales, *Dombeya*, Mascarenes

1 Introduction

During the two last decades, plant systematics have been deeply modified and improved by the increasing use of molecular data as source of phylogenetic signals. The technical advances made in DNA sequencing and in computational power have allowed the analysis of large quantities of molecular data to reconstruct phylogenetic relationships among living organisms. Through the Angiosperm Phylogeny Group (APG; [1–3]), the taxonomic circumscription of many traditional taxa has been significantly improved by taking into account their phylogenetic relationships. Malvales are one of the striking examples illustrating such deep taxonomic changes.

In this book chapter, we provide an overview of the recent delimitations of this order. We also present taxonomic modifications for the most diverse taxa within Malvales (i.e., Malvaceae). Finally, we discuss the implication of molecular taxonomy at the species level for Dombeyoideae (newly circumscribed taxon included in Malvaceae s.l.) with special emphasis on the species endemic to the Mascarene archipelago (Indian Ocean).

2 Evolution of the Classification Within the Malvales

2.1 Traditional View of the Classification

Malvales were traditionally circumscribed as four main families—referred to by some authors as "core Malvales"—namely, Bombacaceae (ca. 250 species), Malvaceae (ca. 1,500 species), Tiliaceae (ca. 400 species), and Sterculiaceae (ca. 1,000 species; [4]). These four taxa, mainly tropical, include economically important species such as cotton (*Gossypium* spp.), kola nut (*Cola nitida* (Vent.) Schott & Endl.), or cocoa (*Theobroma cacao* L.) and numerous ornamental species in various genera (*Dombeya* Cav., *Hibiscus* L., *Malva* L., *Tilia* L.). In addition to the "core Malvales," groups such as Elaeocarpaceae, Bixaceae, and Cistaceae were reported to have malvalean affinities [5–8]. The circumscription of this order was therefore not clear. However, numerous botanists (except Hutchinson [9]) recognized the unity of Malvales on the basis of close relationships between the four families of the "core Malvales" [5–8]. Several morphological features have been proposed as diagnostic criteria for Malvalean affinities (e.g., the presence of mucilage cells, stellate trichomes, a stratified phloem in a wedge shape, mucilage canals and cavities, malvoid teeth on the leaves, and cyclopropenoid seed oils; [4, 10–13]). Nevertheless, the distribution of these characters appeared erratic throughout

Table 1
Evolution in the traditional classifications of families considered as currently including within the Malvales

	APG III [3]	Kubitzki and Chase [12]	Takhtajan [6]	Thorne [7]	Dahlgren [8]	Cronquist [5]
Bixaceae	+	+	– (Cistales)	– (Violales)	+	– (Violales)
Bombacaceae	[+]	[+]	+	+	+	+
Cistaceae	+	+	– (Cistales)	– (Violales)	+	– (Violales)
Cochlospermaceae	[+]	+	– (Cistales)	– (Violales)	+	– (Violales)
Cytinaceae	+	–	– (Rafflesiales)	– (Rafflesiales)	– (Rafflesiales)	– (Rafflesiales)
Diegodendraceae	[+]	+	+	+	?	– (Theales)
Dipterocarpaceae	+	+	+	+	+	– (Theales)
Malvaceae s.s.	[+]	[+]	+	+	+	+
Monotaceae	[+]	[+]	+	+	[+]	– (Theales)
Muntingiaceae	+	+	[+]	[+]	[+]	[+]
Neuradaceae	+	+	– (Rosales)	– (Rosales)	– (Rosales)	– (Rosales)
Sarcolaenaceae	+	+	+	+	+	– (Theales)
Sphaerosepalaceae	+	+	+	+	+	– (Theales)
Sterculiaceae	[+]	[+]	+	+	+	+
Thymelaeaceae	+	+	– (Thymelaeales)	– (Euphorbiales)	– (Thymelaeales)	– (Myrtales)
Tiliaceae	[+]	[+]	+	+	+	+

this order. Morphological features alone were not sufficient to define unambiguously the composition of this order around the "core Malvales." As a consequence, the intra-ordinal relationships in Malvales varied according to the author considered (Table 1).

2.2 Molecular Data: Upheavals of the Classification

The first broad survey of angiosperm phylogeny using the chloroplast gene *rbcL* [13] suggested close relationships between the traditional "core Malvales" and the Dipterocarpaceae. The monophyly and the taxonomic delimitations of this clade (namely, the "expanded" Malvales, Fig. 1) were confirmed and refined by recent investigations focused on Malvales [10, 14–16] as well as in broader taxonomic surveys [17–20]. In its modern circumscription [3], "expanded" Malvales is a sister group of Brassicales with moderate support [20]. This order includes ten families (ca. 338 genera, [21]) in seven well-supported lineages (Fig. 1; [10, 14, 16, 18, 21]).

In APGIII [3], Bixaceae were broadly circumscribed by including Cochlospermaceae and Diegodendraceae. This family is highly supported by molecular data [10, 14–16]. Despite several common morphological features [12, 22, 23], this circumscription of Bixaceae was never proposed in the past traditional treatments (Table 1).

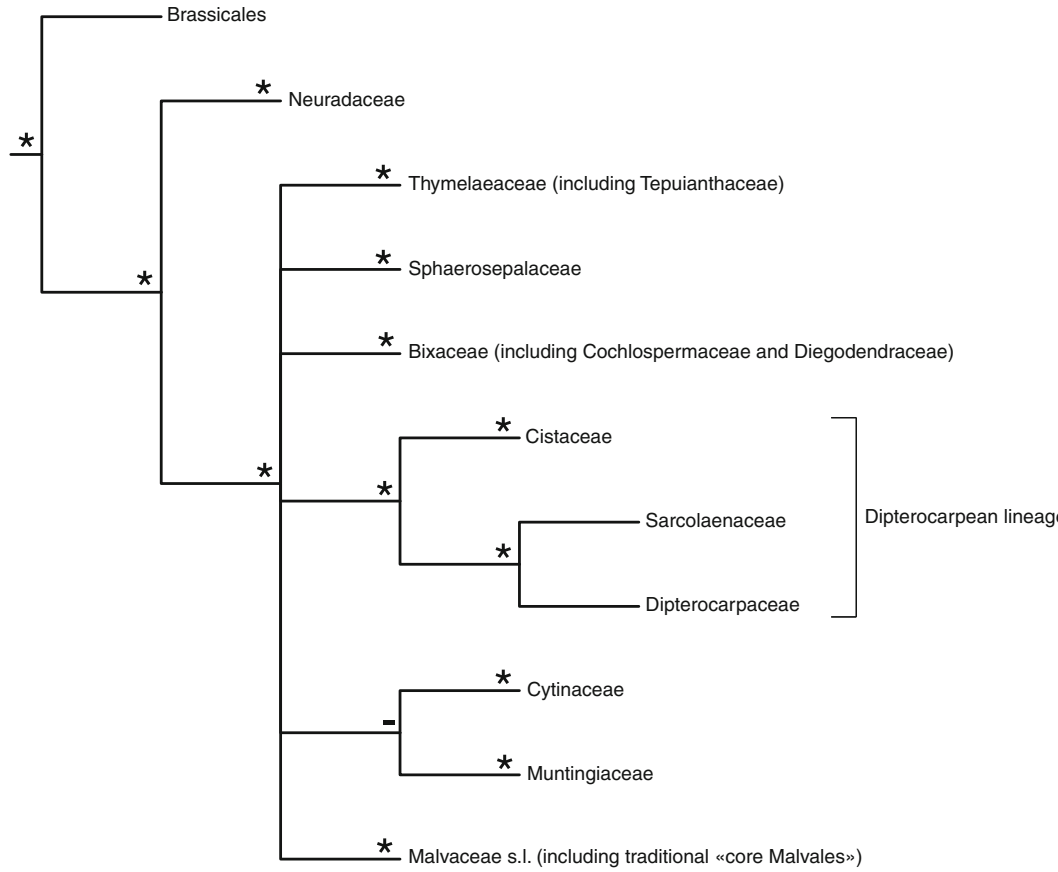


Fig. 1 Summary of the phylogenetic relationships within "expanded Malvales." Nodes indicated by a "-" are moderately supported by the different molecular analyses. Nodes indicated by a "*" are strongly supported by molecular evidence

This may be the consequence of poor morphological knowledge of the monogeneric and monospecific family Diegodendraceae. The phylogenetic position of Bixaceae within "expanded" Malvales is ambiguous and not strongly supported [10, 14, 17, 20]. Although recent studies [16, 20] found similar phylogenetic placement (i.e., sister group of Dipterocarpean lineage), this result is not supported in Nickrent et al. [16]. Additionally, the sampling is not representative of "expanded" Malvales in the broad survey of Soltis et al. [20].

Frequently associated with Bixaceae and considered as related to Malvales, Cistaceae were, however, not included in Malvales (except by Dahlgren [8]; Table 1). Contrary to traditional treatments, molecular data strongly support a sister position of the Dipterocarpean clade (Fig. 1; [10, 14, 16, 18, 20]). These close relationships are confirmed by several morphological features (e.g., vestured pits and bixoid chalazal region) that link Cistaceae to the Dipterocarpean lineage and Bixaceae [22, 24].

Cytinaceae were traditionally included in Rafflesiales (Table 1). Members of this order consisted of holoparasitic plants difficult to ally with green plants because of extreme reduction and/or loss of morphological features [16, 25]. Molecular evidence showed that Rafflesiales were composed of four independent lineages [25]. One of these four groups, the Cytinaceae, was placed in Malvales as the sister group of Muntingiaceae with strong support by Nickrent [16]. Despite the extreme morphological modifications, several characters link Cytinaceae to Muntingiaceae and more generally to Malvales (e.g., trichome types and fruit morphology; [16]).

Dipterocarpaceae were frequently considered as belonging to Malvales (Table 1). Many morphological and anatomical characters (e.g., trichome anatomy, structure of connective appendages, appearance and placement of mucilage cells) strongly suggest close affinities between Dipterocarpaceae, Sarcolaenaceae, and "core" Malvales [12, 26, 27]. According to Maguire and Ashton [28] and Ashton [26], Dipterocarpaceae were composed of three groups: Diptercarpoideae, Monotoideae, and Pakaraimoideae. Kostermans [29, 30] argued that the two latter subfamilies should be included in Tiliaceae. Molecular phylogenies confirm the malvalean affinities of Dipterocarpaceae, validating Ashton's hypothesis (Fig. 1; [14, 15, 18]). This constitutes a group with Sarcolaenaceae and Cistaceae, the Dipterocarpean lineage [10, 12], strongly supported by molecular data. However, phylogenetic relationships between Sarcolaenaceae and the monophyly of Dipterocarpaceae (including Diptercarpoideae, Monotoideae, and Pakaraimoideae) remain unclear. According to Kubitzki and Chase [12] and Ducoussu et al. [31], Sarcolaenaceae are nested within Dipterocarpaceae, which leads to paraphyly of this family. Phylogenetic relationships inferred from sequences of the chloroplast gene *rbcL* give another view by positioning Sarcolaenaceae as the sister group of Dipterocarpaceae ([10, 27, 32]; Fig. 1). However, neither hypothesis is supported by any statistical index (i.e., bootstrap or posterior probabilities).

Muntingiaceae are composed of three monospecific genera (*Muntingia* L., *Neotessmannia* Burret, and *Dicraspidia* Standl.) with strong malvalean affinities (Table 1; [15, 33]). According to Bayer and Kubitzki [34], this family could also include *Petenaea* Lundell. (placed in *Incertae sedis*). These genera were frequently placed in Tiliaceae [6, 9]. Benn and Lemke [35] suggested close relationships between *Muntingia*, *Neotessmannia*, and *Dicraspidia* and placed them in the same tribe Neotessmannieae (Tiliaceae). From molecular data, *Dicraspidia* and *Muntingia* are included in a monophyletic group with high support [14] and constitute a distinct lineage within the "expanded" Malvales. Based on the sequences of two chloroplast genes (*rbcL* and *atpB*) and morphological features, Bayer et al. [15, 33] have defined the family Muntingiaceae, which includes *Muntingia*, *Dicraspidia*, and *Neotessmannia*. Evidence from anatomical features also confirms

the Malvalean affinities of this family [36]. The phylogenetic placement of *Petenaea* as the sister group of Muntingiaceae is not supported and therefore appears unclear. Further investigations are required for validating the inclusion of *Petenaea* in Muntingiaceae [34]. The position of Muntingiaceae is still controversial (i.e., without statistical support), but generally, this family is related to the Dipterocarpean lineage [10, 12, 14, 15, 18]. A recent molecular investigation using four DNA regions has proposed a different phylogenetic placement of Muntingiaceae by supporting a sister relationship with the Cytinaceae holoparasitic group [16]. However, improved sampling within Muntingiaceae is needed to definitively assess this result.

Many morphological characters provide evidences placing Neuradaceae near or in Rosales ([10, 37]; Table 1). For example, on the basis of floral development, Ronse Decraene and Smets [38] suggested the genus *Neurada* L. should be included in Rosaceae. Hubert [39] argued for a placement of *Neurada* within Malvales and provided two potential synapomorphies for "expanded" Malvales (i.e., seed coat anatomy and the presence of cyclopropenoid fatty acid in seeds). Despite the apparent divergence of morphology between Neuradaceae and the "expanded" Malvales, molecular data [10, 14–16, 18, 20] strongly support malvalean affinity for this family. This taxon is placed as the sister group of the whole expanded Malvales with strong statistical support.

The systematic position of Sarcolaenaceae was ambiguous, and it was variously referred to as Theales, Ochnales, or Malvales (Table 1). Morphological and anatomical features such as the presence of mucilage cells, the stratified secondary phloem [40], and the occurrence of cyclopropanoid fatty acids in the seeds [41] strongly suggest malvalean affinities. According to Maguire et al. [42] and Ashton [26], Sarcolaenaceae and Dipterocarpaceae are closely related and share several important morphological (e.g., imbricate calyx) and wood anatomical characters. Molecular evidence from different DNA regions [10, 14–16] clearly supports placement in the Dipterocarpean lineage together with Cistaceae and Dipterocarpaceae. This clade is also characterized by the bixoid chalazal region of the seed coat [22]. The phylogenetic distinction between two subfamilies of Dipterocarpaceae (i.e., Pakaraimoideae and Monotoideae) and Sarcolaenaceae is not supported by molecular data [27, 31, 32]. New analyses focusing on these relationships are still needed to assess the taxonomy of the Dipterocarpean lineage.

Although Hutchinson [9] and Cronquist [5] placed Sphaerosepalaceae within Theales or Ochnales, respectively, this family is more likely included in Malvales [6–8]. This placement is confirmed by anatomical characters, which link Spaherosepalaceae to Diegodendraceae [43]. Molecular phylogenetic approaches also support malvalean affinities for Sphaerosepalaceae. However, the precise position of this family within Malvales is ambiguous.

In the analyses of Fay et al. [14] and Bayer et al. [15], Sphaerosepalaceae are the sister group of Thymelaeaceae. However, recent molecular investigations [10, 16, 17] placed Spaherosepalaceae close to Bixaceae (sensu APG III; 3). Both hypotheses concerning the phylogenetic placement of Sphaerosepalaceae are weakly supported. In the light of these two hypotheses, Horn [44] conducted a comprehensive study focused on the comparative vegetative and reproductive morphology and anatomy of Sphaerosepalaceae. Bixaceae and Sphaerosepalaceae share several morpho-anatomical characters (e.g., floral monosymmetry, single series of stamen trunk bundles, and a well-developed bixoid chalazal), which suggest close affinities between these two families. On the other hand, several features such as leaf architecture, calyx vasculature, and endostomal micropyles clearly support the second hypothesis (i.e., close relationships between Sphaerosepalaceae and Thymelaeaceae). Currently, both hypotheses should be considered and new analyses including molecular and morphological characters must be undertaken to resolve the position of Sphaerosepalaceae within Malvales.

Thymelaeaceae is monophyletic and highly supported by molecular data [10, 14–16, 18, 50]. However, the placement of Thymelaeaceae remained controversial for a long time. Depending on the interpretation of several floral characters (i.e., the gynoecium, the nature of the floral tube, and the interpretation of the petallike structures), various taxonomic positions have been proposed for this family ([45–47]; Table 1). Molecular systematics clearly place Thymelaeaceae within "expanded Malvales." This taxonomic position is supported by the occurrence of cyclopropenoid fatty acids, the exotegemic tegument in seeds, and the features of the secondary phloem [10, 44, 48, 49]. The malvalean affinities of Thymelaeaceae may have been masked by the number of highly derived features in flowers and pollen [44]. Based on both molecular and morphological evidence, Wurdack and Horn [51] showed that the monogeneric Tepuianthaceae is the sister group of Thymelaeaceae. This result is confirmed by several potential morphological synapomorphies. Consequently, the Angiosperm Phylogeny Group [3] adopted an expanded concept of Thymelaeaceae including members of Thymelaeaceae and Tepuianthaceae. Within Malvales, first analyses viewed Thymelaeaceae as the sister group of Sphaerosepalaceae [14, 15] or sister to a clade including both Bixaceae and Dipterocarpean lineages [10]. For Soltis et al. [18], Thymelaeaceae are placed in an unresolved position in Malvales, whereas Savolainen et al. [17] and Nickrent [16] found Thymelaeaceae as the sister group of Muntingiaceae. The last broad survey of angiosperm phylogeny [20] found Thymelaeaceae as sister to Malvaceae. However, none of these hypotheses is supported by statistical analyses.

The last family, Malvaceae s.l. or the "core" Malvales, is monophyletic and includes taxa previously placed in the four traditional families of Malvales: Tiliaceae, Malvaceae s.s.., Bombacaceae, and

Sterculiaceae. The taxonomy of Malvaceae s.l. has been so deeply modified by molecular data that we will focus in the next section on the taxonomic changes within this family.

Molecular evidence has undoubtedly clarified understanding of the Malvales taxonomy. Traditionally limited to four main families, this order currently includes ten families and seven lineages. However, understanding of the evolution of morphological features still remains difficult due to important unresolved relationships in the phylogeny of Malvales. Besides, the phylogenetic placement of several malvalean families is still ambiguous (e.g., the position of Sarcolaenaceae, the monophyly of Dipterocarpaceae, the position of Sphaerosepalaceae, the position of Thymelaeaceae). Only a few studies have focused mainly on the phylogenetic relationships among Malvales [10, 11, 14–16], and the different hypotheses were not really tested because of reduced sampling. New molecular analyses using new markers and focused on Malvales are needed to assess the taxonomic composition, the phylogenetic placement, and the morphological evolution of this order.

3 New Classification of the Core Malvales or Malvaceae s.l.: Contribution of Molecular Data

3.1 Traditional View and Issues in Delimitating Families

As already mentioned, Sterculiaceae, Bombacaceae, Malvaceae s.s., and Tiliaceae were often grouped together in Malvales, constituting the "core" Malvales. Morphological features used for delimiting the different taxa were too scanty and difficult to interpret. As a consequence, the taxonomic boundaries among these families remained unclear. For example, Tiliaceae and Sterculiaceae were mainly distinguished by the degree of fusion of free part of filaments in the androecium [5, 6, 9]. Hutchinson [9] stated: "It will be observed that there is no tangible difference between the two families […] except in the stamens, which in Tiliaceae are numerous, free or very shortly connate at the base, whilst in Sterculiaceae they are united high up or if free and single or in separate bundles then they are opposite to the petals. […] Although the two families are here maintained, a future monographer may combine them into one and probably recast the rather unsatisfactory tribes." Similar issues can be pointed out between Sterculiaceae and Bombacaceae and also between Malvaceae s.s. and Bombacaceae [15].

Families of the "core Malvales" were viewed as a grade of evolution, from the "primitive" Tiliaceae to the "evolved" Malvaceae s.s. As with many "primitive taxa", Tiliaceae often included "aberrant" genera because they presented both plesiomorphic and apomorphic characters (e.g., *Schoutenia* Korth., *Nesogordonia* Baill., *Burretiodendron* Rehder). Consequently, such taxa were difficult to place and move in the traditional classification according to the author's point of view (Table 2).

Table 2
Evolution of the classification of selected genera within the Malvaceae s.l.

Genus	Hutchinson [9]	Takhtajan [6]	Bayer et al. [15]	Whitelock et al. [57], Nyffeler et al. [59], Baum et al. [69]
	Sterculiaceae	Sterculiaceae	Byttneroideae	/Byttneriina Byttneroideae
Theobroma	Theobromeae	Theobromeae	Theobromeae	
Byttneria	Byttnerieae	Byttnerieae	Byttnerieae	
Ayenia	Byttnerieae	Byttnerieae	Byttnerieae	
Melochia	Hermannieae	Hermannieae	Hermannieae	
Leptonychia	Theobromeae	Theobromeae	Incertae sedis	
	Tiliaceae	Tiliaceae	Grewioideae	/Byttneriina Grewioideae
Entelea	Enteleeae	Enteleeae		
Corchorus	Enteleeae	Corchoreae		
Sparmannia	Sparmannieae	Sparmannieae		
Apeiba	Apeibeae	Apeibeae		
Triumfetta	Triumfetteae	Triumfetteae		
Grewia	Grewieae	Grewieae		
Tilia	Tilieae	Tilieae	Tilioideae	/Malvadendrina Tilioideae
Mortoniodendron	Enteleeae	Enteleeae	Incertae sedis	
	Sterculiaceae			
Craigia	Lasiopetaleae	Craigieae	Incertae sedis	
		Sterculiaceae	Helicteroideae	/Malvadendrina Helicteroideae
Triplochiton	Tarrietieae	Triplochitoneae		/Helictereae
Reevesia	Helictereae	Helictereteae		
Helicteres	Helictereae	Helictereteae		
	Bombacaceae	Bombacaceae		
Durio	Durioneae	Durioneae	Incertae sedis	/Durioneae
	Tiliaceae	Tiliaceae	Brownlowoideae	/Malvadendrina Brownlowoideae
Berrya	Berryeae	Berryeae		
Pentace	Berryeae	Berryeae		

(continued)

Table 2
(continued)

Genus	Hutchinson [9]	Takhtajan [6]	Bayer et al. [15]	Whitelock et al. [57], Nyffeler et al. [59], Baum et al. [69]
Brownlowia	Brownlowieae	Brownlowieae		
	Sterculiaceae	Sterculiaceae	Dombeyoideae	/Malvadendrina Dombeyoideae
Dombeya	Dombeyeae	Dombeyeae		
Pterospermum	Helictereae	Helictereteae		
Nesogordonia	Helmiopsideae	Helmiopsideae	Incertae sedis	
	Tiliaceae	Tiliaceae		
Schoutenia	Tilieae	Tilieae		
	Sterculiaceae	Sterculiaceae	Sterculioideae	/Malvadendrina Sterculioideae
Sterculia	Sterculieae	Sterculieae		
Cola	Sterculieae	Sterculieae		
Hildegardia	Tarrietieae	Tarrietieae		
Fremontodendron	Fremontieae	Fremontodendreae	Bombacoideae	/Malvadendrina / Malvatheca Bombacoideae
	Bombacaceae	Bombacaceae		
Pachira	Adansonieae	Bomcaceae		
Bombax	Adansonieae	Bomcaceae		
Adansonia	Adansonieae	Bomcaceae		
	Tiliaceae	Tiliaceae		
Pentaplaris	Brownlowieae	Brownlowieae		
	Bombacaceae	Bombacaceae		/Malvadendrina / Malvatheca
Ochroma	Mantisieae	Mantisieae		
Mantisia	Mantisieae	Mantisieae		/Malvadendrina / Malvatheca Malvoideae
Phragmotheca	Mantisieae	Mantisieae		
	Malvaceae	Malvaceae	Malvoideae	
Hibiscus	Hiscisceae	Hibisceae		
Gossypium	Hiscisceae	Hibisceae		
Lavatera	Malveae	Malveae		

3.2 First Phylogenetic Analysis from Judd and Manchester [11]

Despite these taxonomic problems, the traditional circumscription of "core Malvales" was widely accepted. Using morphological, anatomical, palynological, and chemical features, Judd and Manchester [11] published the first phylogenetic study focused on this group. They established the monophyly of the "core Malvales" prefigured by previous authors [5–8]. However, Judd and Manchester [11] showed that Bombacaceae, Sterculiaceae, and Tiliaceae were polyphyletic and therefore did not form natural lineages. In contrast, Malvaceae s.s. was identified as monophyletic and supported by one synapomorphy: the monothecal anther [11]. As a consequence, these authors argued for the recognition of the "core" Malvales at the familial level, i.e., as Malvaceae s.l. This clade is supported by one unambiguous synapomorphy: the floral nectaries composed of densely packaged multicellular glandular hairs on the sepals. Vogel [52] confirmed the unique anatomy of this feature. However, lack of resolution in the phylogeny hindered the redefinition of taxonomic groups within Malvaceae s.l. This phylogenetic study of Judd and Manchester [11] is the basis of the modern taxonomic treatment of Malvaceae s.l.

3.3 Contribution of Molecular Data: From the Intuition to the Confirmation

Two years after Judd and Manchester [11], the first molecular systematic studies focused on Malvaceae s.l. were conducted in parallel by Alverson et al. [53] and Bayer et al. [15]. In both molecular investigations, the monophyly of the "core" Malvales was confirmed with strong statistical support. Following Judd and Manchester [11], Alverson et al. [53] and Bayer et al. [15] opted for merging all the members of the "core" Malvales into a single family: "expanded" Malvaceae or Malvaceae s.l. Three morphological apomorphies, i.e., the specialization within the inflorescences termed "bicolor unit" [54, 55], the floral nectaries, and the tile cells, also provided evidence for the broad circumscription of the Malvaceae [12, 15, 53]. Malvaceae s.l. are divided into two strongly supported lineages (named by Baum et al. [56] following the phylogenetic code of nomenclature) including nine subfamilies (Fig. 2; [15, 53, 57–59, 61]). We use "/" before a name to indicate those names that have been defined following the phylogenetic nomenclature.

3.3.1 The /Byttneriina Lineage (Fig. 2; [53, 56])

This well-supported clade is the sister group (with moderate support) of the remaining taxa of Malvaceae s.l. The Byttneriina lineage is composed of two subfamilies (detailed below) including taxa of former Sterculiaceae and Tiliaceae.

3.3.2 Byttnerioideae (Fig. 2)

All the taxa of this subfamily previously belonged to Sterculiaceae (Table 2). Based on molecular data [15, 57, 58], Byttnerioideae includes four tribes: Theobromeae, Byttnerieae, Lasiopetaleae, and Hermannieae. While this clade was traditionally diagnosed by its peculiar cucullate petals, the petal evolution traced on the

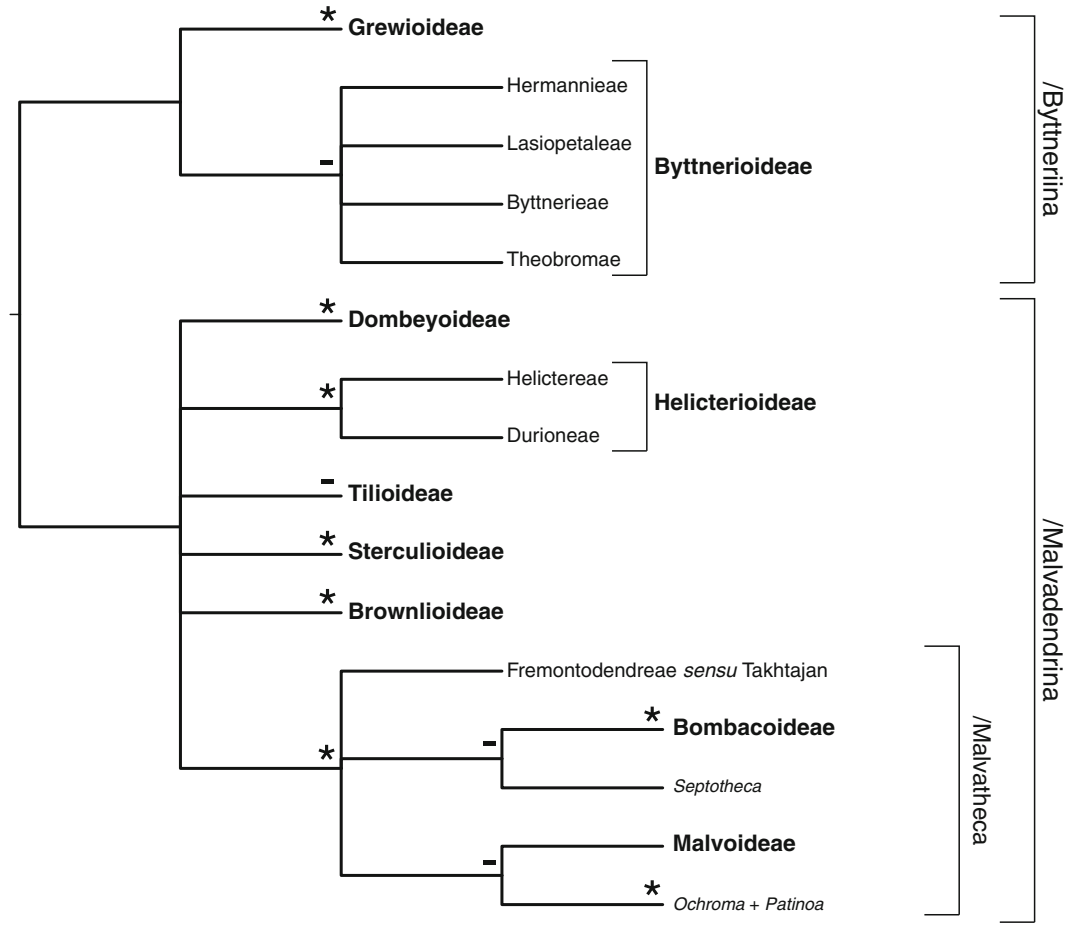


Fig. 2 Molecular phylogenetic relationships within Malvaceae s.l. Subfamilies are indicated in *bold*. Names preceded by a "/" have been defined following phylogenetic nomenclature. *See* Fig. 1 for node legends

molecular phylogeny highlights the significant homoplasy of this character [15, 57]. As a consequence, morpho-anatomical approaches as well as developmental studies must be undertaken to define new potential synapomorphy.

3.3.3 Grewioideae (Fig. 2)

Taxa of Grewioideae were formerly included in Tiliaceae on the basis of the large number of stamens and the free part of the filaments (Fig. 2; Table 2). Grewioideae is highly supported by molecular data [10, 15, 58]. However, its phylogenetic placement varies depending on the molecular markers. Based on *atpB* and *rbcL* [15], Grewioideae are nested in Byttnerioideae, but with low statistical support. Results from the plastid gene *ndhF* clearly show a sister relationship between the two subfamilies of the /Byttneriina lineage [53, 57, 58]. This subfamily is also supported by several morphological features notwithstanding polymorphic features

such as the shape of the pollen, the position of the nectaries, and the origin of the stamen during floral development [15, 60, 61].

3.3.4 The/Malvadendrina Lineage (Fig. 2; [53, 56])

This lineage is found in all molecular systematic studies on Malvaceae s.l., with strong statistical support [15, 53, 57–59]. A unique 21-bp deletion in the chloroplast gene *ndhF* provides additional evidence for the monophyly of this group [53]. The /Malvadendrina lineage includes all of the former members of Bombacaceae and Malvaceae s.s. together with several tribes of Sterculiaceae (e.g., Dombeyeae, Sterculieae) and Tiliaceae (Tilieae; Table 2). In all molecular analyses [15, 53, 59], these taxa are distributed in seven subfamilies within six lineages [59, 61]. Nyffeler et al. [59] noticed inconsistencies between the phylogenetic signal of *rbcL* and those of the three other genes (*matK*, *ndhF*, and *atpB*). Consequently, phylogenetic relationships between the different clades remain unresolved and problematic. The latest phylogenetic hypotheses need to be assessed by additional molecular and/or morphological data.

3.3.5 Dombeyoideae (Fig. 2)

This subfamily is strongly supported by molecular and morphological characters [15, 53, 59, 62]. This group includes representatives of former sterculoid tribes (mainly Dombeyeae and Helmiopsideae) but also several genera traditionally placed in Tiliaceae, such as *Burretiodendron* and *Schoutenia* (Table 2). Dombeyoideae also includes the enigmatic and morphologically "aberrant" genus *Nesogordonia*. Indeed, this subfamily can be diagnosed by the presence of bifid cotyledons and spinose pollen [15, 61], while *Nesogordonia* presents simple cotyledons and smooth pollen [63]. The phylogenetic position of Dombeyoideae within the /Malvatheca lineage is unclear. Depending on the DNA region used, molecular analyses placed Dombeyoideae in three different positions: (1) as the sister group of a clade including Brownlowioideae and Sterculioideae [15], (2) closely related to Tilioideae in a sister relationship [53], (3) or in a basal position in the /Malvandendrina lineage (Fig. 2; [59]). However, these different relationships are only weakly or moderately supported by molecular data.

3.3.6 Helicteroideae (Fig. 2)

Helicteroideae are consistently well supported by molecular data [15, 53, 59, 64]. This subfamily consists of two distinct lineages: the Helicteroideae s.s. ([61], or /Helictereae sensu [64]) which mainly includes taxa (e.g., *Helicteres* L., *Reevesia* Lindl.) from Helictereae (*sensu* Hutchinson [9] and Takhtajan [6], former Sterculiaceae) and Durioneae [61, 64] which is composed of taxa (*Durio* Adans., *Neesia* Blume) from the Bombacaceae. While the alliance of Helicteroideae s.s. and Durioneae is strongly supported by molecular data, the morphological link between the two lineages is difficult to identify and needs further investigation [15, 59, 64]. The phylogenetic position of Helicteroideae is still equivocal within

/Malvadendrina. This subfamily varies between sister relationships with the remaining taxa of the /Malvadendrina [15, 53] or branching directly after Dombeyoideae as sister to Tilioideae-Brownlowioideae-Sterculioideae-/Malvatheca [59]. However, both hypotheses are weakly supported by molecular data.

3.3.7 Tilioideae (Fig. 2)

Traditionally, Tiliaceae included ca. 400 species in many genera such as *Tilia* L., *Schoutenia*, *Burretiodendron*, *Pentaplaris* L.O. Williams & Standl., *Grewia* L., and *Brownlowia* Roxb. In the modern taxonomic treatment of "core" Malvales, the Tilioideae is drastically narrowed and only includes three genera: *Tilia*, *Mortoniodendron* Standl. & Steyerm, and *Craigia* W.W.Sm. & W.E.Evans. Sister relationships between *Craigia* and *Tilia* are well supported by molecular data from the chloroplast gene *ndhF* [53]. Based on three chloroplast markers (*ndhF*, *matK*, *atpB*), Nyffeler et al. [59] clearly place *Mortoniodendron* as sister to the *Tilia-Craigia* clade. In addition to the molecular evidence, these three genera share a crucial morphological feature, the "tilioid" pollen [11, 15]. The flower organization also links *Craigia* and *Tilia* [15, 61]. Phylogenetic position of Tilioideae within /Malvandendrina is unclear and needs further investigation for assessing their placement [59].

3.3.8 Sterculioideae (Fig. 2)

The monophyly of Sterculioideae is well supported by molecular studies [15, 53, 59, 65]. According to Bayer et al. [15], Wilkie et al. [65], and Bayer [61], this subfamily includes genera traditionally placed in Sterculieae (former Sterculiaceae; Table 2). In addition to molecular characters, this group can be diagnosed by several morpho-anatomical features: the apetalous flowers, the petaloid sepals, the secondary apocarpy, and the presence of sheath cells in the rays [11, 15, 65–67]. Using the plastid marker *ndhF*, Wilkie et al. [65] identify four clades (i.e., *Sterculia* L., *Heritiera* Aiton, *Cola* Schott & Endl., and *Brachychiton* Schott & Endl.) within Sterculioideae. However, phylogenetic relationships between these different monophyletic groups remain unresolved. Similarly, Sterculioideae are not clearly placed within /Malvadendrina. Based on the molecular dataset, this subfamily is sister either to Brownlowioideae [15] or to /Malvatheca [59]. Neither of these two hypotheses is supported by statistical analysis. Based on developmental investigations, Sterculioideae share with /Malvatheca similar androecium development [68].

3.3.9 Brownlowioideae (Fig. 2)

Brownlowioideae contain taxa of the former Tiliaceae (i.e., the tribes Berryeae and Brownlowieae). This subfamily is supported by both molecular data and morphological characters [15, 53, 59, 64]. Brownlowioideae are characterized by their campanulate epicalyx composed of fused sepals, the apocarpous gynoecium, and their special arrangement of staminal thecae (i.e., divergent at the

base but convergent at the top of the connective; [15, 61]). Morphological and molecular data are congruent and support a clear division in two distinct lineages. The first clade includes genera with inner staminodes (e.g., *Pentace*, *Brownlowia*), while the second monophyletic group is composed of genera that have only fertile stamens.

The phylogenetic position of Brownlowioideae remains unclear within /Malvadendrina. This subfamily is more likely placed as sister to Sterculioideae [15, 59]. These two taxa share the fused sepals, but they differ in the shape of their pollen (tilioid-shaped pollen for Brownlowioideae versus Grewia-type pollen for Sterculioideae, [15, 60, 61]). The alternative sister relationships branch Brownlowioideae to a clade containing Sterculioideae and /Malvatheca.

3.3.10 The /Malvatheca Lineage (Fig. 2)

In former taxonomic classifications, Bombacaceae and Malvaceae s.s. were reported as closely related, and boundaries between the two taxa were difficult to perceive. Recent molecular studies have confirmed the phylogenetic link between these families by regrouping bombacoid and malvoid genera together within the same lineage, named /Malvatheca [53, 56]. This clade is strongly supported in most analyses using molecular data [53, 59, 69]. The highly modified anthers were often reported as a synapomorphy [59]. However, this character seems to be homoplasious and may have evolved at least twice independently in /Malvatheca (Bombacoideae and Malvoideae [68]). Recent developmental investigations [68, 70] show that the androecial development is a synapomorphy of this clade and unifies /Malvatheca. Concerning the sister clade of this lineage, the most likely hypothesis places the Sterculioideae as the sister group of /Malvatheca based on several common morphological features and molecular characters [59].

Molecular taxonomic studies in /Malvatheca [59, 69, 71] have identified five major monophyletic groups. The first lineage, Bombacoideae (*sensu* [69]), corresponds to the palmately compound-leaved Bombacaceae [15, 53, 69, 71]. In former classifications, Durionae was included in Bombacaceae on the basis of similar anther structure [59]. However, molecular evidence excluded Durionae from Bombacoideae and placed this tribe in Helicteroideae with strong support [15, 53, 59, 64]. Moreover, developmental approaches confirmed the exclusion of Durionae by demonstrating the different origins of the anthers [70].

The second lineage, Malvoideae, mainly corresponds to the former family Malvaceae s.s. and includes the tribes Hibisceae and Gossypieae. Molecular evidence [53, 69, 71] strongly supports the inclusion of the tribe Matiseae (*Matisia* Bompl., *Phragmotheca* Cuatrec., and *Quararibea* Aubl., but excluding *Ochroma* Sw., formerly in Bombacaceae; Table 2) and *Pentaplaris* (formerly in

Tiliaceae [9]; Table 2). These genera share several morphological characters such as the globose and spinose pollen (*Pentaplaris*) and the usually palmately veined leaves (for Matiseae). However, as noticed by Alverson et al. [53], such features are not unique in Malvaceae s.l.; other subfamilies also present these characters (e.g., Dombeyoideae have a similar shape of pollen). As argued by von Balthazar et al. [68, 70], the simplification of vascular system represents a potential synapomorphy and thus supports the new circumscription of Malvoideae.

In the light of molecular evidence, the phylogenetic position of the third lineage composed of two genera (i.e., *Ochroma* and *Patinoa* Cuatrec., formerly in Bombacaceae) remains controversial (Table 2). *Ochroma* and *Patinoa* were considered as belonging to Bombacoideae [15, 61], but studies using different chloroplast genes found that these two genera constitute an independent lineage early divergent in /Malvatheca [53, 59, 69]. The last molecular taxonomic study [71] focused on Bombacoideae gives an accurate view of the phylogenetic relationships of *Ochroma* and *Patinoa*: these are placed at the basal position of Malvoideae with moderate statistical support. This result is corroborated by other morphological and developmental characters [68, 70]. Contrary to Malvoideae and Bombacoideae, *Ochroma* and *Patinoa* are characterized by sessile, often elongate anthers, which appear pleisiomorphic in /Malvatheca [68]. Similarly, vegetative and reproductive features such as simple leaves and petaloid-margined calyces might be pleisiomorphic [53].

In the traditional taxonomy, Fremontodendraceae were included in Sterculiaceae based on the apetalous flowers with colorful calyces and stamens with two theca and four locules ([53, 69]; Table 2). However, based on chloroplast markers, Fremontodendraceae are unambiguously included in /Malvatheca [15, 53, 59, 69]. Baum et al. [69] argued for a placing Fremontodendraceae in an early diverging position, which is strongly supported by a single non-homoplasious deletion in a conserved region of *matK* [69]. Developmental investigations [68] provide other evidence for strengthening this hypothesis.

The genus *Septotheca* Ulbr. constitutes the fifth lineage within /Malvatheca. Placed by Baum et al. [69] as the sister group to the rest of Bombacoideae, this result is not supported by statistical analysis. Recent molecular investigations [71] and developmental studies [68] have found this same phylogenetic position but without statistical support [71].

Recent molecular phylogenetic analyses give an accurate view of the relationships within Malvaceae s.l. This family appears clearly monophyletic supported both by molecular and morphological data. Nine main subfamilies within two lineages have been identified. The main monophyletic groups (i.e., subfamilies and lineages)

are supported by molecular data. However, higher phylogenetic relationships of Malvaceae s.l. are still controversial and lack support from morphology or molecules. Moreover, while the phylogeny of taxa included in the /Malvatheca lineage has been widely investigated, the understanding of relationships among and within the taxa of Tilioideae, Sterculioideae, and Dombeyoideae has not followed at the same pace. As a consequence, phylogenetic relationships within and around these taxa are still largely unknown and are under investigation, and their taxonomy clearly needs further clarification. As an example, we discuss in the next section past and recent studies on the Dombeyoideae with special focus on the taxa endemic to the Mascarenes. Using these endemic species, we will illustrate three possible taxonomic schemes consequent to recent molecular studies.

4 General Presentation of the Dombeyoideae

The number of species in Dombeyoideae (ca. 350 species; [61]) is relatively low in comparison to other subfamilies of Malvaceae s.l., such as Malvoideae (ca. 1,730 species) or Grewioideae (ca. 700 species). However, this subfamily includes one of the most diversified genera in Malvaceae s.l., *Dombeya* (ca. 215 species, [72]). Strikingly, *Dombeya* also represents almost 2 % of Malagasy plant diversity [73] and more than 2 % of the diversity of species endemic to the Mascarene archipelago. Dombeyoideae are distributed in South Asia and in continental Africa, but the center of diversity is undoubtedly the Southwest Indian Ocean biodiversity hotspot, which includes Madagascar and the nearby oceanic archipelagos. The three young volcanic islands of the Mascarenes (Mauritius, 8–10 My; La Réunion, 2–3 My; Rodrigues 8–10 My; [74]) are home to 24 species included in four genera: *Trochetia* DC., *Ruizia* Cav., *Astiria* Lindl.—all endemic to the archipelago—and *Dombeya* Cav. (Fig 3, [75–78]). Besides their high diversity in the Mascarenes, Dombeyoideae have a strong ecological importance in numerous diverse vegetation communities. Species occur in most of the habitats found in the archipelago, from the dry environments to the tropical mountain cloud forests [79, 80] in which *Dombeya* species are particularly abundant and are the main contributors of the canopy. These forests are given the local name “colored tree forest” due to the large range of leaf coloration among *Dombeya* species. Many species are considered as critically endangered, endangered, or vulnerable ([77]; *see* Table 3 for details regarding IUCN evaluation). For these different reasons, the Dombeyoideae and more specifically the genera *Ruizia* and *Trochetia* symbolize both the richness and the weakness of the Mascarene flora.

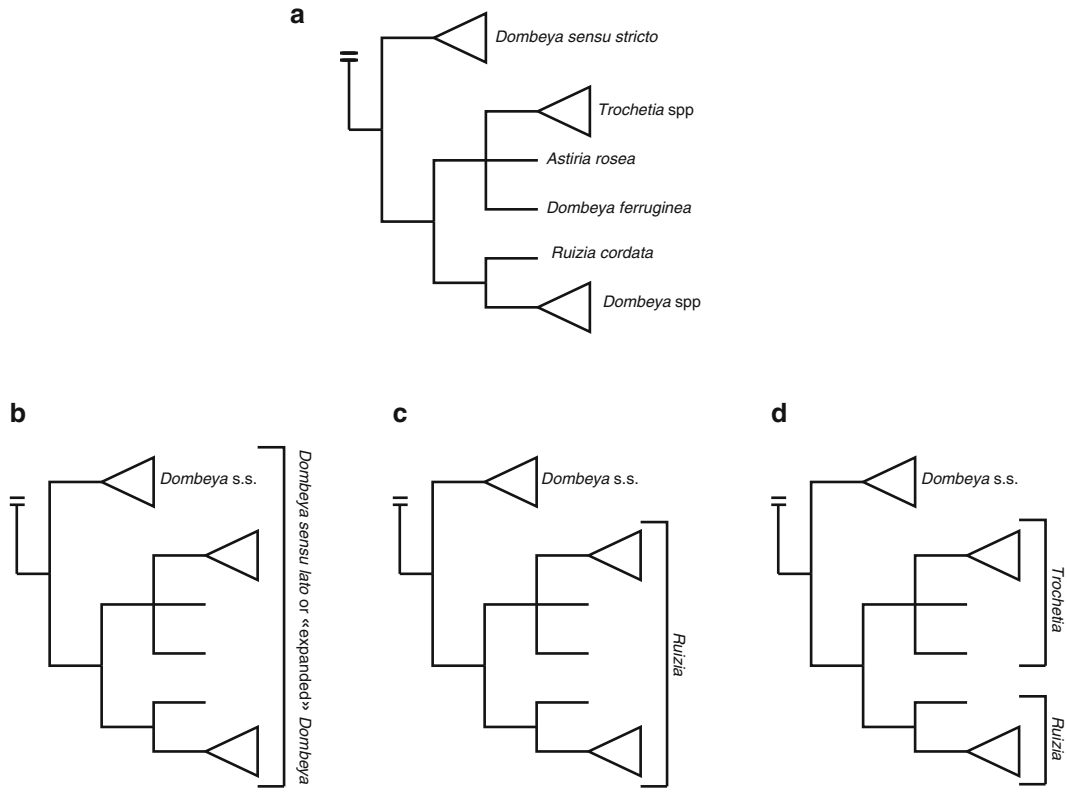


Fig. 3 Illustration of the alternative taxonomic schemes discussed for Dombeyoideae of the Mascarene region

5 Past Taxonomy Versus Phylogenetic Systematics of the Mascarene Dombeyoideae

Taxonomic studies of the Mascarene Dombeyoid taxa started in the eighteenth century and are still in progress [75, 77, 78, 81–86]. There are three traditional treatments of the Dombeyoideae in the Mascarenes but with major disagreements between them (Table 4). Jacob de Cordemoy [85] described 27 species placed in four genera (*Astiria*, *Dombeya*, *Ruizia*, and *Trochetia*). Arènes [86] focused on the species of *Dombeya* and did not recognize the endemic genera *Astiria* and *Trochetia*. This author placed the species of these two latter genera in *Dombeya* (sections *Trochetia* and *Dombeya*), these two genera being therefore considered as synonyms of *Dombeya*. Arènes [86] placed his taxonomic study of the Mascarene species in the broader framework of the flora of Madagascar [87] and he did not accept the sections described by Jacob de Cordemoy [85]. As a consequence, Arenes [86] recognized 27 species placed in five sections of *Dombeya* (Table 4). More recently, Friedmann [75] adopted a broad taxon circumscription

Table 3
IUCN Red List for Dombeyoideae native taxa of the Mascarene archipelago [77]

	Taxa	IUCN criteria	Taxa endemism
1	*Astiria rosea* Lind.	EX	Ma
2	*Dombeya populnea* (Cav.) Baker	CR	Re Ma
3	*Dombeya mauritiana* Friedmann	CR	Ma
4	*Dombeya sevathianii* Le Péchon & Baider	CR	Ma
5	*Dombeya rodriguesiana* Friedmann	CR	Ro
6	*Ruizia cordata* Cav.	CR	Re
7	*Trochetia boutoniana* Friedmann	CR	Ma
8	*Trochetia parviflora* Bojer	CR	Ma
10	*Dombeya acutangula* Cav. subsp. *acutangula* var. *palmata* Arènes	CR	Re
11	*Dombeya acutangula* subsp. *rosea* Friedmann	CR	Ma
12	*Dombeya elegans* var. *virescens* Cordem.	EN	Re
13	*Dombeya blattiolens* Frapp. ex Cordem.	EN	Re
14	*Dombeya ferruginea* Cav. subsp. *ferruginea*	EN	
15	*Dombeya umbellata* Cav.	EN	Re
16	*Dombeya acutangula* subsp. *acutangula* var. *acutangula* Arènes	VU	Re Ro
17	*Dombeya delislei* Arènes	VU	Re
18	*Trochetia blackburniana* DC.	VU	Ma
19	*Trochetia triflora* DC.	VU	Ma
20	*Trochetia uniflora* DC.	VU	Ma
21	*Trochetia granulata* Cordem.	NT	Re
22	*Dombeya ciliata* Cordem.	LC	Re
23	*Dombeya elegans* Cordem. subsp. *elegans*	LC	Re
24	*Dombeya ferruginea* Cav. subsp. *borbonica* Friedmann	LC	Re
25	*Dombeya ficulnea* Baill.	LC	Re
26	*Dombeya formosa* Le Péchon & Pausé	LC	Re
27	*Dombeya pilosa* Cordem.	LC	Re
28	*Dombeya punctata* Cav.	LC	Re
29	*Dombeya reclinata* Cordem.	LC	Re

UICN criteria: *EX* extinct, *CR* critically endangered, *EN* endangered, *VU* vulnerable, *NT* nearly threatened, *LC* low concern. Taxa endemism: *Re* Réunion, *Ma* Mauritius, *Ro* Rodrigues

Table 4
Comparison between different classifications of the Mascarene Dombeyoideae

	Genus	Subgenus	Section	Species number
Jacob de Cordemoy [85]	*Astiria*			1
	Dombeya	*Dombeya*	*Assonia*	2
			Eudombeya	17
			Dombeyella	2
	Ruizia			1
	Trochetia			4
Arènes [86]	*Dombeya*	*Dombeya*	*Astrapeae*	1
			Trochetia	6
			Assonia	1
			Capricornua	3
			Dombeya	17
Friedmann [75]	*Astiria*			1
	Dombeya	*Dombeya*		14
	Ruizia			1
	Trochetia			6

and considered many morphological variations to be infraspecific polymorphism. He accepted 22 species included in four genera: *Astiria*, *Dombeya*, *Ruizia*, and *Trochetia*. He placed numerous species in synonymy (e.g., *D. tiliioides* Arènes as a synonym of *D. delislei* Arènes; *D. lancea* Cordem. and *D. obovata* Cordem. as synonyms of *D. punctata* Cav.; *D. pilosa* var. *globigera* Cordem., *D. pilosa* var. *amplifolia* Cordem., *D. hispides* Arènes, and *D. baillonia-ana* Arènes as synonyms of *D. pilosa* Cordem.; *Dombeya triflora* Arènes and *D. serrata* Arènes as synonyms of *Trochetia triflora* DC.). Conflicting views between these three taxonomic treatments highlight problems in defining the morphological delimitation of the Mascarene genera. For example, *Astiria* only differs from *Dombeya* by the absence of staminodes within the androecium. However, the number of fertile stamens is variable in several *Dombeya* species (e.g., *Dombeya ferruginea* and *Dombeya sevathia-nii*, [76]). Molecular and morphological phylogenetic analyses [88, 89] have been conducted to clarify the taxonomy of Dombeyoideae from the Mascarenes. In both investigations, taxa from the archipelago are polyphyletic and included in four different lineages. In both analyses, using different sources of

Table 5
List of molecular markers used for studying the phylogeny of the Malvales, Malvaceae s.l.

	Genome	Gene expression	Studies using the gene	Taxonomic levels
rbcL	Chloroplast	Coding	[10, 13–18, 20, 27, 31, 32, 46, 50, 51, 58, 59, 62]	Order, family, subfamily
atpb	Chloroplast	Coding	[15–18, 20, 50, 51, 59, 62]	Angiosperm, order, family, subfamily
ndhF	Chloroplast	Coding	[16, 20, 53, 57–59, 62, 64, 65, 69]	Angiosperm, order, family, subfamily
trnK/matK	Chloroplast	Noncoding/coding	[20, 59, 69, 71]	Angiosperm, order, family
trnL-F	Chloroplast	Noncoding	[46, 71]	Family, species
trnQ-rps16	Chloroplast	Noncoding	[89]	Species
Rpl16 intron	Chloroplast	Noncoding	[89]	Species
psbM-trnD	Chloroplast	Noncoding	[89]	Species
18S rDNA	Nuclear	Coding	[16, 18, 20, 50, 51]	Angiosperm, order
26S rDNA	Nuclear	Coding	[20, 50]	Angiosperm
ITS	Nuclear	Noncoding	[64, 71, 89]	Genera, species

phylogenetic signal, the taxonomy of the Mascarene Dombeyoideae is congruent apart from the position of the genus *Trochetia* and the placement of *Dombeya rodriguesiana* [88, 89]. *Astiria* and *Ruizia* (based on morphological characters, [88]) or *Trochetia* and *Ruizia* (based on molecular characters, [89]) are nested in *Dombeya*. Therefore, *Dombeya* is paraphyletic in relation to *Astiria*, *Ruizia*, and *Trochetia* (Figure 3) and needs further study to confirm the status of the putative endemic genera from the Mascarenes.

6 Future Taxonomic Treatments: Different Possibilities to Revise the Taxonomy of Dombeyoideae from the Mascarenes

Dombeya is non-monophyletic in its current delimitation (Fig. 3a; [73, 88, 89]). Therefore, a new circumscription following the monophyly criterion is needed. As there is no strict guideline to define a rank such as an order, family, or genus, several possibilities can be contemplated for a taxonomic revision. Besides taxonomists should minimize nomenclatural changes compared to previous classifications and should name taxa based on clades that are morphologically consistent (i.e., supported by a synapomorphy or by a combination of apomorphic features). Depending on the group, several subjective criteria could be considered. Some groups have also an anthropological and societal emblematic significance.

For instance, *Trochetia* is the national emblem of the Republic of Mauritius. In 1992, *T. boutoniana* was declared the National flower to celebrate gaining the status of the Republic. Similarly, *Ruizia cordata* Cav. is one of the most famous threatened species in La Réunion and symbolizes the highly endangered flora of the dry environments of the island. As a consequence, this nonscientific dimension should be taken into account by taxonomists when proposing taxonomic changes. To illustrate these criteria and their consequences for taxonomy, we propose three possible classifications for the Dombeyoideae of the Mascarenes (Fig. 3). The first possible taxonomic scheme expands the limits of *Dombeya* to include all the species of *Trochetia* and the two monotypic genera *Ruizia* and *Astiria* (Fig. 3b). In such a scheme, the number of changes is minimized and this taxonomy corresponds to the Arènes' treatment of the taxonomy of Mascarene taxa [86]. However, in this broad circumscription, *Dombeya* is difficult to diagnose using morphological characters alone (Table 4). Furthermore, this treatment strongly modifies the traditional treatments [5, 6, 9, 75, 84, 85] which recognized *Ruizia*, *Astiria*, *Trochetia,* and *Dombeya* (except Arènes [86]). Alternatively, to avoid the recognition of *Dombeya* s.l. (i.e., an "expanded" *Dombeya*), the second possible classification is to identify the clade including *Ruizia*, *Astiria*, *Dombeya,* and *Trochetia* as a single generic-ranked taxon (Fig. 3c). *Ruizia* was published earlier than *Astiria* and *Trochetia* [81], so *Ruizia* should therefore be adopted as the name of this monophyletic group. This solution highlights the island endemic status of this group. However, this treatment will imply the highest number of taxonomic changes. Besides, *Ruizia* is a monotypic genus and the name does not reflect the diversity of this clade, which is mainly concentrated within *Trochetia*. The third possible taxonomy is to recognize both subclades (i.e., *Trochetia* and *Ruizia* Fig. 3d) at the generic rank. *Trochetia* [82] has priority over *Astiria* [83]. As a consequence, *Trochetia* should be redefined to refer to the plant species included in this monophyletic group. The clade "*Ruizia*" is only composed of *Ruizia* and *Dombeya* species. All these taxa should be renamed with the generic name *Ruizia*. In this taxonomic configuration, this treatment implies nine taxonomic changes and conserves both the names *Ruizia* and *Trochetia*, the most emblematic genera endemic to the Mascarene. The new circumscription of *Ruizia* modifies the former morphological descriptions, but this monophyletic group can be diagnosed by the heteromorphic juvenile leaves, the particular epicalyx, and the reduced or absent staminode. In contrast, the broad circumscription of *Trochetia* is morphologically heterogeneous. Indeed, *D. ferruginea* has a divergent morphology in comparison to *Trochetia* (most notably in inflorescence architecture and size of flowers).

These three taxonomic treatments are a selection among many taxonomic possibilities. There is no absolute criterion to determine

which one is better compared to the others. However, a combination of subjective criteria (i.e., depending on the taxon) can help a taxonomist to determine the most satisfactory treatment given the evidence available. The Dombeyoideae of the Mascarenes represent a limited fraction of the global diversity of this subfamily, which is mainly found in Madagascar. With more than 200 Malagasy species, the current circumscription of *Dombeya* is still unclear. Similarly to *Ruizia*, *Astiria*, and *Trochetia*, several genera endemic to Madagascar with a low species number, such as *Helmiopsis* H. Perr. and *Helmiopsiella* Arènes, are also nested in *Dombeya* [73, 89]. As a consequence, any new taxonomic treatment of the Mascarene taxa should be put on hold until they have been included in a broader revision of the Malagasy Dombeyoideae. This approach will avoid multiple and contradictory taxonomic revisions of smaller regions.

7 Conclusion

The history of the Malvales classification is a perfect illustration of the past two decades great advances in the field of systematic botany. The taxonomy of this group has been deeply modified and improved by the extensive use of molecular characters for reconstructing evolutionary relationships (Table 5). "Expanded" Malvales are currently circumscribed to include 10 families, which are confirmed by all of the phylogenetic analyses. All of these families were previously considered to have malvalean affinities by various previous workers (except the holoparasitic and highly modified Cytinaceae). Although the delimitation of the Malvales is now widely accepted, the intra-ordinal relationships remain problematic. As a consequence, morphological evolution is still unclear in Malvales. Similar issues are reported in the most emblematic taxa of Malvales, the Malvaceae s.l. In its present circumscription, this family includes all of the former families of the "core Malvales." From both molecular and morphological evidence, the Malvaceae is strongly supported and monophyletic. If the taxonomic composition is quite similar to the previous "core Malvales," delimitation within the family is completely modified. Malvaceae includes nine subfamilies in two main lineages. Despite recent investigations based on four different markers, the relationships between the different subfamilies of Malvaceae s.l. are poorly resolved. While the phylogeny of several emblematic subfamilies such as Bombacoideae, Malvoideae, and Helicteroideae is well understood, other taxa (e.g., Tilioideae, Brownlowioideae, and Dombeyoideae) are still poorly known. As a consequence, these subfamilies need to be revised and the morphological delimitations of their genera assessed. The Mascarene Dombeyoideae exemplified the substantial taxonomic changes, which have to be done for taking into account the recent progress of the molecular taxonomy. While 20

years ago the first development of molecular sequencing was considered as a revolution for systematics, we can observe a new upheaval approaching with the development of next-generation sequencing technologies [90], which study not just one or several regions within a genome but the genome as a whole.

Acknowledgements

This research is included in the BACOMAR Project (http://mahots.univ-reunion.fr/) supported by the University de La Réunion, the Région Réunion, the European Community, and the Ministère Français de la Recherche. Funding from the Open Laboratory of Ecological Restoration and Biodiversity Conservation of Chengdu Institute of Biology, Chinese Academy of Sciences (CAS), to Li-Bing Zhang and a CAS Research Fellowship for International Young Researchers (Grant n°31150110463) provided support to T.L.P. during the writing. We are grateful to Dr. Li-Bing Zhang and Dr. Karen Wilson for their comments on the previous version of the manuscript. We warmly thank David Caron for the illustrations and Pierre Gigord for his moral support.

References

1. Angiosperm Phylogeny Group (1998) An ordinal classification for the families of flowering plants. Ann MO Bot Gard 85:531–553
2. Angiosperm Phylogeny Group II (2003) Classification for the orders and families of flowering plants: APG II. Bot J Linn Soc 141:399–436
3. Angiosperm Phylogenetic Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121
4. Judd WS, Campbell CS, Kellogg EA et al (2008) Plant systematics: a phylogenetic approach, 3rd edn. Sinauer Associates, Sunderland
5. Cronquist A (1988) The evolution and classification of flowering plants, 2nd edn. The New York Botanical Garden, New York
6. Takhtajan A (1997) Diversity and classification of flowering plants. Columbia University Press, New York
7. Thorne RF (1992) An updated phylogenetic classification of the flowering plants. Aliso 13:365–389
8. Dahlgren G (1989) The last Dahlgrenogram: system of classification of the dicotyledons. In: Tan K, Mill RR, Elias TS (eds) Plant taxonomy, phytogeography and related subjects. Edinburgh Univ. Press, Edinburgh, pp 249–260
9. Hutchinson J (1967) The families of flowering plants. Dicotyledons, 2nd edn. Oxford University Press, London
10. Alverson WS, Karol KG, Baum DA et al (1998) Circumscription of the Malvales and relationships to other Rosidae: evidence from rbcL sequence data. Am J Bot 85:876–887
11. Judd WS, Manchester SR (1997) Circumscription of Malvaceae (Malvales) as determined by a preliminary cladistic analysis of morphological, anatomical, and chemical characters. Brittonia 49:384–405
12. Kubitzki K, Chase MW (2003) Introduction to Malvales. In: Kubitzki K, Bayer C (eds) V Flowering plants. Dicotyledons Malvales Capparales and non-betalain Caryophyllales. Springer, Berlin, pp 12–20
13. Chase MW, Soltis DE, Olmstead RG et al (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann MO Bot Gard 80:528–548, 550–580
14. Fay MF, Bayer C, Alverson WS et al (1998) Plastid rbcL sequence data indicate a close affinity between Diegodendron and Bixa. Taxon 47:43–50
15. Bayer C, Fay MF, De Bruijn AY et al (1999) Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: a combined analysis of plastid atpB

and rbcL DNA sequences. Bot J Linn Soc 129: 267–303
16. Nickrent DL (2007) Cytinaceae are sister to Muntingiaceae (Malvales). Taxon 56:1129–1135
17. Savolainen V, Chase MW, Hoot SB et al (2000) Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences. Syst Biol 49:306–362
18. Soltis DE, Soltis PS, Chase MW et al (2000) Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. Bot J Linn Soc 133:381–461
19. Alford MH (2006) Gerrardinaceae: a new family of African flowering plants unresolved among Brassicales, Huerteales, Malvales, and Sapindales. Taxon 55:959–964
20. Soltis DE, Smith SA, Cellinese N et al (2011) Angiosperm phylogeny : 17 genes, 640 taxa. Am J Bot 98:704–730
21. Stevens PF (2001 onwards) Angiosperm Phylogeny Website. Version 9, June 2008
22. Nandi OI (1998) Ovule and seed anatomy of Cistaceae and related Malvanae. Plant Syst Evol 209:239–264
23. Ronse Decraene LP (1989) Floral development of Cochlospermum tinctorium and Bixa orellana with special emphasis on the androecium. Am J Bot 76:1344–1359
24. Jansen S, Baas P, Smets E (2000) Vestured pits in Malvales s.l.: a character with taxonomic significance hidden in the secondary xylem. Taxon 49:169–182
25. Nickrent DL, Blarer A, Qiu Y-L et al (2004) Phylogenetic inference in Rafflesiales: the influence of rate heterogeneity and horizontal gene transfer. BMC Evol Biol 4:1–17
26. Ashton PS (1982) Dipterocarpaceae. In: Van Steenis CGGJ (ed) Flora Malesiana, Series 1, Spermatophyta, vol 9. Martinus Nijhoff Publishers, The Hague, pp 237–552
27. Morton CM, Dayanandan S, Dissanayake D (1999) Phylogeny and biosystematics of Pseudomonotes (Dipterocarpaceae) based on molecular and morphological data. Plant Syst Evol 216:197–205
28. Maguire B, Ashton PS (1980) Pakaraimaea dipterocarpacea II. Taxon 29:225–231
29. Kostermans AJGH (1978) Pakaraimaea dipterocarpacea Maguire & Ashton Belongs to Tiliaceae and Not to Dipterocarpaceae. Taxon 27:357–359
30. Kostermans AJGH (1985) Family status for the Monotoideae Gilg and the Pakaraimoideae Ashton, Maguire and de Zeeuw (Dipterocarpaceae). Taxon 34:426–435
31. Ducousso M, Béna G, Bourgeois C et al (2004) The last common ancestor of Sarcolaenaceae and Asian dipterocarp trees was ectomycorrhizal before the India–Madagascar separation, about 88 million years ago. Mol Ecol 13:231–236
32. Dayanandan S, Ashton PS, Williams SM et al (1999) Phylogeny of the tropical tree family Dipterocarpaceae based on nucleotide sequences of the chloroplast *rbcL* gene. Am J Bot 86:1182–1190
33. Bayer C, Chase MW, Fay MF (1998) Muntingiaceae, a new family of dicotyledons with Malvalean affinities. Taxon 47:37–42
34. Bayer C (2003) Muntingiaceae. In: Kubitzki K (ed) The families and genera of vascular plants. Springer, Berlin, pp 315–319
35. Benn SJ, Lemke DE (1991) Taxonomy of *Neotessmannieae* (Tiliaceae). Am J Bot 78(Suppl):166–167
36. Carlquist S (2005) Wood and bark anatomy of Muntingiaceae: a phylogenetic comparison within Malvales s.l. Brittonia 57:59–67
37. Bayer C (2003) Neuradaceae. In: Kubitzki K (ed) The families and genera of vascular plants. Springer, Berlin, pp 325–328
38. Ronse Decraene LP, Smets E (1995) The floral development of Neurada procumbens L. (Neuradaceae). Acta Bot Neerl 44:439–451
39. Hubert H (1993) Neurada, eine Gattung der Malvales. Sendtnera 1:7–10
40. Den Outer RW, Vooren AP (1980) Bark anatomy of some Sarcolaenaceae and Rhopalocarpaceae and their systematic position. Meded Land Wagen 81:1–15
41. Gaydou EM, Ramanoelina AR (1993) A survey of the Sarcolaenaceae for cyclopropene fatty acids. Phytochemistry 22:1725–1728
42. Maguire B, Ashton PS (1977) Systematic, geography and phyletic considerations. Taxon 26:343–368
43. Dickison WC (1988) Xylem anatomy of Diegodendron humbertii. IAWA Bull NS 9:332–336
44. Horn JW (2004) The morphology and relationships of the Sphaerosepalaceae (Malvales). Bot J Linn Soc 144:1–40
45. Beaumont AJB (2010) Systematic studies in *Gnidia* L. (Thymelaeaceae). Ph.D. Thesis, University of Kwazulu-Natal, South Africa
46. Van der Bank M, Fay MF, Chase MW (2002) Molecular phylogenetics of Thymelaeaceae with particular reference to African and Australian genera. Taxon 51:329–339
47. Herbet BE (2003) Thymelaeaceae. In: Kubitzki K, Bayer C (eds) The families and genera of vascular plants. Dicotyledons. Malvales, Capparales, and non-betalain Caryophyllales. Springer, Berlin, pp 373–396
48. Corner EJH (1976) The seeds of dicotyledons, vol 1 & 2. Cambridge University Press, Cambridge
49. Domke W (1934) Unterschungen über die systematische und geographische Gliederung der Thymelaeaceen nebst einer Neubeschreibung ihrer Gattungen. Bibl Bot 111:1–151

50. Soltis DE, Gitzendanner MA, Soltis PS (2007) A 567-taxon data set for angiosperms: the challenges posed by Bayesian analyses of large data sets. Int J Plant Sci 168:137–157
51. Wurdack KJ, Horn JW (2001) A re-evaluation of the affinities of the Tepuianthaceae: molecular and morphological evidence for placement in Malvales. Botany 2001:151 (abstract)
52. Vogel S (2000) The floral nectaries of Malvaceae sensu lato. A conspectus. Kurtziana 28:155–171
53. Alverson WS, Whitlock BA, Nyffeler R et al (1999) Phylogeny of the core Malvales: evidence from *ndhF* sequence data. Am J Bot 86:1474–1486
54. Bayer C (1998) Synflorescences of Malvaceae. Nord J Bot 18:335–338
55. Bayer C (1999) The bicolor unit—homology and transformation of an inflorescence structure unique to core Malvales. Plant Syst Evol 214:187–198
56. Baum DA, Alverson WS, Nyffeler R (1998) A durian by any other name: taxonomy and nomenclature of the core Malvales. Harv Pap Bot 3:315–330
57. Whitelock BA, Bayer C, Baum DA (2001) Phylogenetic relationships and floral evolution of the Byttnerioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene, ndhF. Syst Bot 26:420–437
58. Whitelock BA, Karol KG, Alverson WS (2003) Chloroplast DNA sequences confirm the placement of the enigmatic *Oceanopapaver* within *Corchorus* (Grewioideae: Malvaceae s.l., Formerly Tiliaceae). Int J Plant Sci 164: 35–41
59. Nyffeler R, Bayer C, Alverson WS et al (2005) Phylogenetic analysis of the Malvadendrina clade (Malvaceae s.l.) based on plastid DNA sequences. Org Div Evol 5:109–123
60. Perveen A, Grafström E, El-Ghazaly G (2004) World Pollen and Spore Flora 23. Malvaceae Adams. P.p. Subfamilies: Grewioideae, Tilioideae, Brownlowioideae. Grana 43: 129–155
61. Bayer C (2003) Malvaceae. In: Kubitzki K (ed) The families and genera of vascular plants. Springer, Berlin, pp 225–311
62. Won H (2009) Phylogenetic position of Corchoropsis Siebold & Zucc. (Malvaceae s.l.) inferred from plastid DNA sequences. J Plant Biol 52:411–416
63. Barnett LC (1988) Systematics of Nesorgordonia Baillon (Sterculiaceae). Ph.D. thesis, University of Texas, Austin
64. Nyffeler R, Baum DA (2000) Phylogenetic relationships of the durians (Bombacaceae-Durioneae or]Malvaceae]Helicteroideae/Durioneae) based on chloroplast and nuclear ribosomal DNA sequences. Plant Syst Evol 224:55–82
65. Wilkie P, Clark A, Pennington TR et al (2006) Phylogenetic relationships within the subfamily Sterculioideae (Malvaceae/Sterculiaceae-Sterculieae) using the chloroplast gene *ndhF*. Syst Bot 31:160–170
66. Chattaway M (1932) The wood of the Sterculiaceae. I. Specialisation of the vertical wood parenchyma within the sub-family Sterculiaceae. New Phytol 31:119–132
67. Chattaway MM (1937) The wood anatomy of the family Sterculiaceae. Philos Trans Roy Soc B 228:313–365
68. von Balthazar M, Schönenberger J, Alverson WS et al (2006) Structure and evolution of the androecium in the Malvatheca clade (Malvaceae s.l.) and implications for Malvaceae and Malvales. Plant Syst Evol 260:171–197
69. Baum DA, DeWitt Smith S, Yen A et al (2004) Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae sensu lato) as inferred from plastid DNA sequences. Am J Bot 91:1863–1871
70. von Balthazar M, Alverson WS, Schönenberger J et al (2004) Comparative floral development and androecium structure in Malvoideae (Malvaceae s.l.). Int J Plant Sci 165:445–473
71. Duarte MC, Esteves GL, Salatino Maria Luiza F et al (2011) Phylogenetic Analyses of *Eriotheca* and Related Genera (Bombacoideae, Malvaceae). Syst Bot 36:690–701
72. Skema C, Dorr LJ (2010) *Dombeya gautieri* (Dombeyaceae), a remarkable new species from Madagascar. Kew Bull 65:305–310
73. Skema C (2012) Toward a new circumscription of *Dombeya* (Malvales: Dombeyaceae): a molecular phylogenetic and morphological study of Dombeya of Madagascar and a new segregate genus, *Andringitra*. Taxon 61:612–628
74. Thébaud C, Warren BH, Strasberg D et al (2009) Mascarene Islands, biology. In: Gillespie RG, Clague DA (eds) Encyclopedia of islands. University of California Press, Berkeley, CA, pp 612–619
75. Friedmann F (1987) Sterculiacées. In: Bosser J, Cadet T, Guého J, Marais W (eds.) Flore des Mascareignes: La Réunion, Maurice, Rodrigues. MSIRI, Port Louis, ORSTOM, Paris et Royal Botanical Garden, Kew, pp 1–50
76. Le Péchon T, Baider C, Gigord LD et al (2011) *Dombeya sevathianii* (Malvaceae): a new critically endangered species endemic to Mauritius (Indian Ocean). Phytotaxa 24:1–10
77. Le Péchon T, Humeau L, Gigord LDB et al (2011) Les Mahots des Mascareignes, Base de Connaissances sur les Mahots des Mascareignes. Université de La Réunion, Saint Denis
78. Le Péchon T, Pausé J-B, Dubuisson J-Y et al (2013) *Dombeya formosa* sp. nov. (Malvaceae s.l.): a New Species Endemic to La Réunion (Indian Ocean) based on morphological and molecular evidence. Syst Bot 38:424–433

79. Cadet T (1980) La végétation de l'île de la Réunion: étude phytoécologique et phytosociologique. Ph.D. thesis, Université Aix-Marseilles, Marseille
80. Blanchard F (2000) Guide des Milieux Naturels. La Réunion-Maurice-Rodrigues, Ulmer, Paris
81. Cavanilles AJ (1787) Tertia dissertatio botanica. In: Monadelphiae classis dissertations decem. Firmin-Didot et Cie, Paris
82. De Candolle AP (1823) Sur quelques genres nouveaux de la famille des Buttnériacées. Mém Mus Hist Nat 10:97–115
83. Lindley J (1844) Astiria rosea. In: Ridgway J (ed.) Edwards's botanical register. London.
84. Baker JB (1877) Flora of Mauritius and the Seychelles. L. Reeve & Co., London
85. Jacob de Cordemoy E (1895) Flore de La Réunion. Klincksieck, Paris
86. Arènes J (1959) Les Dombeya des îles des Mascareignes. Mém Inst Sci Madagascar 9:189–216
87. Arènes J (1959) 131e Famille—Sterculiacées. In: Humbert H (ed) Flore de Madagascar et des Comores. Firmin-Didot et Cie, Paris, pp 1–537
88. Le Péchon T, Cao N, Dubuisson J-Y et al (2009) Systematics of Dombeyoideae (Malvaceae) in the Mascarene archipelago (Indian Ocean) inferred from morphology. Taxon 58:519–531
89. Le Péchon T, Haevermans T, Cruaud C et al (2010) Multiple colonizations from Madagascar and converged acquisition of dioecy in the Mascarene Dombeyoideae (Malvaceae) as inferred from chloroplast and nuclear DNA sequence analyses. Ann Bot Lond 106:343–357
90. Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. Am J Bot 99:312–319

Chapter 18

What Has Molecular Systematics Contributed to Our Knowledge of the Plant Family Proteaceae?

Peter H. Weston

Abstract

Molecular systematics has revolutionized our understanding of the evolution of the Proteaceae. Phylogenetic relationships have been reconstructed down to generic level and below from alignments of chloroplast and nuclear DNA sequences. These trees have enabled the monophyly of all subfamilies, tribes, and subtribes to be rigorously tested and the construction of a new classification of the family at these ranks. Molecular data have also played a major part in testing the monophyly of genera and infrageneric taxa, some of which have been recircumscribed as a result. Molecular trees and chronograms have been used to test numerous previously postulated biogeographic and evolutionary hypotheses, some of which have been modified or abandoned as a result. Hypotheses that have been supported by molecular phylogenetic trees and chronograms include the following: that the proteaceous pattern of repeated disjunct distributions across the southern hemisphere is partly the result of long-distance dispersal; that high proteaceous diversity in south-western Australia and the Cape Floristic Region of South Africa is due to high diversification rates in some clades but is not an evolutionary response to Mediterranean climates; that the sclerophyllous leaves of many shrubby members of the family are not adaptations to dry environments but for protecting mesophyll in brightly illuminated habitats; that deeply encrypted foliar stomata are adaptations for minimizing water loss in dry environments; and that *Protea* originated in the Cape Floristic Region of South Africa and that one of its subclades has greatly expanded its distribution into tropical savannas. Reconstructing phylogeny down to species level is now the main goal of molecular systematists of the Proteaceae. The biggest challenge in achieving this task will be resolving species trees from numerous gene trees in complexes of closely related species.

Key words Proteaceae, Molecular, Systematics, Taxonomy, Phylogeny historical biogeography

1 Introduction

Two British botanists active in the early to mid-nineteenth century, Robert Brown and Joseph Hooker, and two Australians of the late twentieth century, Lawrie Johnson and Barbara Briggs, were largely responsible for proposing the hypotheses that molecular systematists of the Proteaceae have sought to test. In 1810, Brown published the first taxonomic revision of the family down to species level [1], which included an introductory essay on the taxonomy,

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9_18, © Springer Science+Business Media New York 2014

morphology, and biogeography of the family that is a classic text in all of those fields. In that essay, Brown identified systematic and biogeographic patterns that subsequent generations of botanists of the Proteaceae have attempted to explain. He observed that the Proteaceae are almost entirely confined to the southern hemisphere but restricted to its continental landmasses and larger islands; that taxa native to South America had much closer affinities to those of Australia than to those of Africa; that the family is most diverse at warm temperate latitudes, particularly in the Cape Floristic Region of South Africa and in south-western Australia; and that the Proteaceae are almost entirely restricted to low-nutrient soils.

Hooker, in his introductory essays to the flora of New Zealand [2] and flora of Tasmania [3], noted that several taxa within the family formed part of his Antarctic floristic relationship, arguing that such a pattern had to be explained as the remnants of a once continuous Antarctic flora that had been fragmented by geological and climatic changes. Hooker went on later to invoke now-sunken land bridges to account for these broken connections, but the patterns that he identified came very much back into focus when plate tectonic theory seemed to explain them comprehensively in the late 1960s and early 1970s.

Brown's classification recognized taxa at only generic and specific ranks, but other botanists soon converted many of his key leads into named higher taxa. These mostly survived, unchanged, apart from the addition of newly discovered species and genera, until Johnson and Briggs critically reexamined them from a modern evolutionary perspective in a landmark paper [4]. Johnson and Briggs not only erected a strikingly new classification at subfamilial, tribal, subtribal, and generic ranks but also proposed detailed hypotheses of morphological character phylogeny, cytoevolution, historical biogeography, and evolutionary ecology. They based their new classification and evolutionary and biogeographic inferences largely on the results of a numerical phylogenetic analysis of morphological, embryological, phytochemical, and cytological characters derived from both the literature and their own observations. Their tree was not constructed using a "completely rigorous procedure" but by manually modifying the results of their own constrained parsimony algorithm that favored multiple parallel changes over reversals.

Opinions concerning the closest relatives of the Proteaceae varied widely among competing purveyors of angiosperm systems in the 1970s. Johnson and Briggs were confident that the Proteaceae were not closely related to any particular families and that they diverged early from a primitive stock and tentatively suggested that they might be allied to the Rosales as a basis for characterizing a hypothetical ancestor, or "proto-proteacea." However, outgroup comparison was not the only logic that they used for this purpose. They also appealed to general "morphological and

adaptational principles" such as the assumption that reduction in number of parts such as ovules is more likely than their increase.

Johnson and Briggs' classification included five subfamilies, 12 tribes, 28 subtribes, and 74 genera, for a total of 119 higher taxa, of which 41 were monotypic, leaving 78 higher taxa that were open to testing for monophyly (Table 1, Fig. 1). By the mid-1970s, the theory of plate tectonics and associated hypotheses of continental movement had achieved almost universal acceptance among geologists, and Johnson and Briggs invoked it to add a further ten putative clades to Hooker's list of three "well-marked groups" showing austral distributions involving two or more Gondwanic landmasses. They did, however, suggest that some groups might have achieved their disjunct distributions as a result of both vicariance and long-distance dispersal. Johnson and Briggs argued that "proto-proteacea" was a rainforest-dwelling tree with flowers pollinated by nectar-feeding insects, with carpels that each matured into a dehiscent, follicular fruit, which released numerous, winged seeds at maturity. Different lineages had convergently adapted to different environments, moving into sclerophyll woodlands, heathlands, savannas, alpine zones, and even deserts, often acquiring new pollinators such as nectar-feeding birds and mammals, and evolving both dry and fleshy indehiscent fruits—the former moved by wind, water, or gravity and the latter dispersed by vertebrates. They also strongly criticized what they called the "Mediterranean myth"—the notion that the high biodiversity of "Mediterranean hotspots" such as south-western Australia and the Western Cape of South Africa was an evolutionary response to the Mediterranean climates of these regions. Johnson and Briggs's classification and other ideas all required highly resolved, well-supported phylogenetic trees before they could be rigorously tested. That is where molecular systematics came in handy.

2 Molecular Phylogenetic Studies Above Generic Level and Their Implications

The first phylogenetic trees of the Proteaceae produced from alignments of macromolecular sequences were those of Martin and Dowd [5, 6], as part of their broader research program on angiosperm phylogeny and biogeography. This program relied entirely on alignments of sequences of the 40 N-terminal amino acids of the small subunit of ribulose-1,5-bisphosphate carboxylase. This protein is the nucleus-encoded small subunit of the photosynthetic enzyme rubisco, the large subunit of which is encoded by the much more frequently sequenced chloroplast gene *rbcL*. Martin and Dowd sequenced these 40 amino acids from a sample of 14 species of Proteaceae, representing 13 of Johnson and Briggs' genera, nine of their subtribes, and the largest three of their subfamilies. They aligned them under the unproblematic assumption that *rbcS* was

Table 1
Johnson and Briggs's [4] classification of the Proteaceae (columns 1–4) and results of tests of the monophyly of its taxa by molecular systematists from Martin and Dowd [6] to Sauquet et al. [18] (columns 5–11)

Subfamily	Tribe	Subtribe	Genus	Martin and Dowd (1988) [6]	Hoot and Douglas (1998) [11]	Mast (1998) [23]	Barker et al. (2002) [14]	Mast and Givnish (2002) [24]	Weston and Barker (2006) [13]	Sauquet et al. (2009) [18]
Persoonioideae				*Monophyletic*	Paraphyletic				*Monophyletic*	*Monophyletic*
	Bellendeneae		*Bellendena*		Sampled				Sampled	Sampled
	Persoonieae			*Monophyletic*	*Monophyletic*				*Monophyletic*	*Monophyletic*
		Placosperminae	*Placospermum*						Sampled	Sampled
		Persooniinae		*Monophyletic*	*Monophyletic*				*Monophyletic*	*Monophyletic*
			Toronia	Sampled	Sampled		Sampled		Sampled	Sampled
			Persoonia	Sampled	Sampled				Sampled	Sampled
			Garnieria						Sampled	Sampled
			Pycnonia						Sampled	Sampled
			Acidonia						Sampled	Sampled
Proteoideae				Polyphyletic	*Monophyletic*				*Monophyletic*	*Monophyletic*
	Conospermeae				Polyphyletic				Polyphyletic	Polyphyletic
			Symphionema		Sampled		Sampled		Sampled	Sampled
			Agastachys		Sampled				Sampled	Sampled
		Cenarrheninae			Polyphyletic				Polyphyletic	Polyphyletic
			Beauprea		Sampled		Sampled		Sampled	Sampled
			Beaupreopsis							Sampled
			Cenarrhenes		Sampled		Sampled		Sampled	Sampled
		Petrophilinae			Polyphyletic				Polyphyletic	Polyphyletic
			Isopogon	Sampled	Sampled		*Monophyletic*		Sampled	Sampled
			Petrophile		Sampled				Sampled	Sampled
		Stirlingiinae	*Stirlingia*		Sampled		Sampled		Sampled	Sampled
		Conosperminae			*Monophyletic*		*Monophyletic*		*Monophyletic*	*Monophyletic*
			Conospermum		Sampled		Sampled		Sampled	Sampled
			Synaphea		Sampled		Sampled		Sampled	Sampled

		Dilobeinae	*Dilobeia*						Sampled
	Franklandieae				Polyphyletic			Polyphyletic	Polyphyletic
		Adenanthinae	*Adenanthos*	Sampled	Sampled	*Monophyletic*		Sampled	Sampled
		Franklandiinae	*Franklandia*		Sampled			Sampled	Sampled
	Proteeae			Polyphyletic	Polyphyletic			Polyphyletic	Polyphyletic
		Aulacinae			Polyphyletic			Polyphyletic	Polyphyletic
			Aulax		Sampled			Sampled	Sampled
			Leucadendron	Sampled	Sampled	*Monophyletic*		Sampled	Sampled
		Proteinae				Polyphyletic		Polyphyletic	Polyphyletic
			Faurea			Sampled		Sampled	Sampled
			Protea	Sampled		Sampled		Sampled	Sampled
			Diastella			*Monophyletic*		Sampled	Sampled
			Leucospermum			*Monophyletic*		Sampled	Sampled
			Mimetes			Paraphyletic		Sampled	Sampled
			Orothamnus			Sampled		Sampled	Sampled
			Paranomus			Paraphyletic		Sampled	Sampled
			Serruria			*Monophyletic*		Sampled	Sampled
			Sorocephalus			Polyphyletic		Sampled	Sampled
			Spatalla			Polyphyletic		Sampled	Sampled
Carnarvonioideae			*Carnarvonia*		Sampled			Sampled	Sampled
Sphalmioideae			*Sphalmium*		Sampled			Sampled	Sampled
Grevilleoideae					Unresolved			Polyphyletic	Polyphyletic
	Oriteae				*Monophyletic*		Unresolved	*Monophyletic*	*Monophyletic*
			Neorites		Sampled		Sampled	Sampled	Sampled
			Orites		Sampled		Sampled	Sampled	Sampled
	Knightieae			Polyphyletic	Polyphyletic			Polyphyletic	Polyphyletic
		Cardwelliinae	*Cardwellia*		Sampled			Sampled	Sampled
		Knightiinae		Polyphyletic				Polyphyletic	Polyphyletic
			Darlingia	Sampled				Sampled	Sampled
			Eucarpha					Sampled	Sampled
			Knightia	Sampled	Sampled			Sampled	Sampled
	Embothrieae				Paraphyletic		Paraphyletic	Paraphyletic	Paraphyletic
		Buckinghamiinae			Paraphyletic		Paraphyletic	Paraphyletic	Paraphyletic

(continued)

Table 1
(continued)

Subfamily	Tribe	Subtribe	Genus	Martin and Dowd (1988) [6]	Hoot and Douglas (1998) [11]	Mast (1998) [23]	Barker et al. (2002) [14]	Mast and Givnish (2002) [24]	Weston and Barker (2006) [13]	Sauquet et al. (2009) [18]
			Buckinghamia		Sampled			Sampled	Sampled	Sampled
			Opisthiolepis		Sampled			Sampled	Sampled	Sampled
		Embothriinae		*Monophyletic*	*Monophyletic*				*Monophyletic*	*Monophyletic*
			Embothrium	Sampled	Sampled				Sampled	Sampled
			Oreocallis						Sampled	Sampled
			Telopea	Sampled	Sampled				Sampled	Sampled
		Lomatiinae	*Lomatia*	*Monophyletic*	Sampled			Sampled	Sampled	Sampled
		Stenocarpinae							*Monophyletic*	*Monophyletic*
			Stenocarpus		Sampled			Sampled	Sampled	Sampled
			Strangea						Sampled	Sampled
	Grevilleeae					*Monophyletic*			*Monophyletic*	*Monophyletic*
			Finschia						Sampled	Sampled
			Grevillea		Sampled	Sampled			Sampled	Sampled
			Hakea			*Monophyletic*		Sampled	Sampled	Sampled
	Helicieae				Polyphyletic				Polyphyletic	Polyphyletic
		Heliciinae			Polyphyletic				Polyphyletic	Polyphyletic
			Helicia		Sampled				Sampled	Sampled
			Xylomelum		Sampled				Sampled	Sampled
		Hollandaeinae	*Hollandaea*		Sampled				Sampled	Sampled
		Triuniinae	*Triunia*		Sampled				Sampled	Sampled
	Macadamieae				Polyphyletic			Polyphyletic	Polyphyletic	Polyphyletic
		Floydiinae	*Floydia*		Sampled			Sampled	Sampled	Sampled
		Gevuininae			Polyphyletic				Polyphyletic	Polyphyletic
			Euplassa		Sampled				Sampled	Sampled
			Gevuina		Sampled				Sampled	Sampled
			Sleumerodendron						Sampled	Sampled

	Hicksbeachiinae						Polyphyletic	Polyphyletic
		Athertonia					Sampled	Sampled
		Heliciopsis					Sampled	Sampled
		Hicksbeachia					Sampled	Sampled
		Malagasia					Sampled	Sampled
		Virotia					Sampled	Sampled
	Lambertiinae	*Lambertia*		Sampled		Sampled	Sampled	Sampled
	Macadamiinae		Polyphyletic	*Monophyletic*			*Monophyletic*	*Monophyletic*
		Brabejum	Sampled	Sampled			Sampled	Sampled
		Macadamia	Sampled	Sampled	Sampled	Sampled	Sampled	Sampled
		Panopsis		Sampled			Sampled	Sampled
	Roupalinae						Polyphyletic	Polyphyletic
		Kermadecia					Sampled	Sampled
		Roupala				Sampled	Sampled	Sampled
Banksieae				*Monophyletic*	*Monophyletic*	*Monophyletic*	*Monophyletic*	*Monophyletic*
	Musgraveinae			*Monophyletic*	*Monophyletic*		*Monophyletic*	*Monophyletic*
		Austromuellera		Sampled	Sampled		Sampled	Sampled
		Musgravea		Sampled	Sampled	Sampled	Sampled	Sampled
	Banksiinae				*Monophyletic*	*Monophyletic*	*Monophyletic*	*Monophyletic*
		Banksia		Sampled	Paraphyletic	Paraphyletic	Sampled	Sampled
		Dryandra			*Monophyletic*	*Monophyletic*	Sampled	Sampled

"Sampled" indicates that only one species of the higher taxon listed in this row was sampled in the analysis listed at the top of the column. "Monophyletic" indicates that two or more species of the higher taxon in this row were sampled in this analysis and that they formed a clade. "Paraphyletic" and "polyphyletic" indicate that two or more species of the higher taxon in this row were sampled in this analysis and that they formed a paraphyletic or polyphyletic group

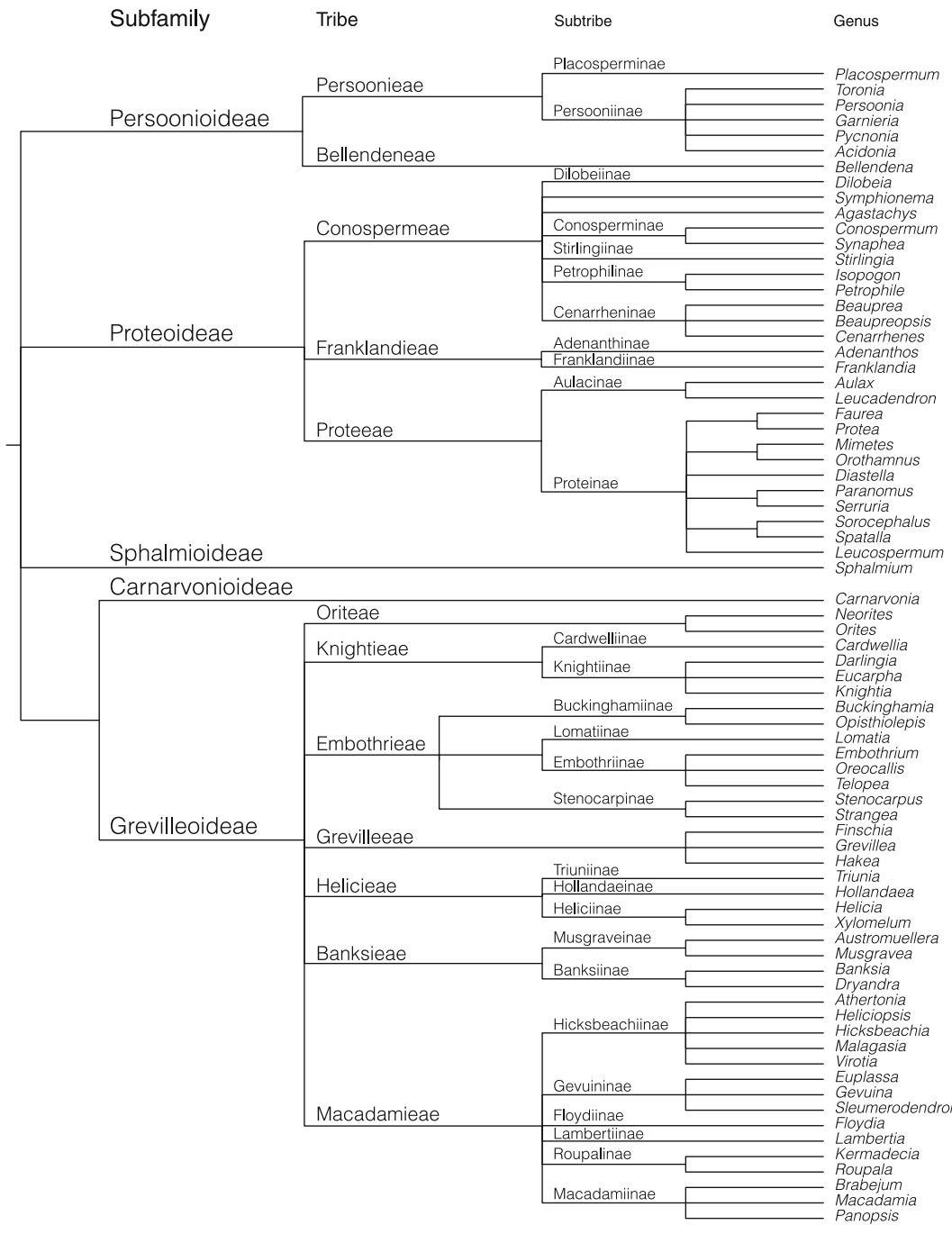


Fig. 1 Johnson and Briggs's [4] phylogenetic tree of the Proteaceae ([4]: Figs. 1 and 2), redrawn as a cladogram. *Names above branches* are Johnson and Briggs's subfamilies, tribes, and subtribes

free of indels and then translated them into a DNA alignment of 120 sites, in which ambiguous third codon positions were "conservatively" translated. Their phylogenetic analysis involved two branch and bound parsimony searches in which members of the subfamily Grevilleoideae were analyzed separately to those of the subfamilies Persoonioideae and Proteoideae, with a hypothetical ancestor for the monocots being used to root both trees and a hypothetical ancestor for the Grevilleoideae being included in the Persoonioideae/Proteoideae analysis. Of the 11 unconstrained clades in their trees, only three were congruent with Johnson and Briggs' classification (*see* Table 1). The congruent taxa were at generic and subtribal levels, and the incongruent ones did not even closely match any groupings found in subsequent studies, suggesting that estimated patristic distances at higher taxonomic levels were saturated with multiple substitutions. Martin and Dowd's analysis was inadequate by today's standards for a number of reasons, most glaring of which was the tiny number of variable sites they analyzed, but it did show that macromolecular sequencing, conducted in ordinary university laboratories, could produce useful phylogenetic data. It thus heralded the coming era of molecular systematics of the Proteaceae.

It did not take long for automated sequencing of PCR-amplified DNA fragments to replace protein-based approaches as the standard source of data for molecular systematic studies above generic level nor for *rbcL* to replace *rbcS* as the first gene of choice for resolving such problems. The first broad analysis of interfamilial angiosperm relationships was based on a parsimony analysis of *rbcL* sequences [7] and placed the Proteaceae as the sister group of the Sabiaceae, near the base of the "eudicot" clade, but without any indication of the level of support for this grouping. This result vaguely confirmed Johnson and Briggs' earlier opinion concerning the taxonomic isolation of the Proteaceae. Other botanists soon came up with more precise estimates of the family's position in the angiosperm tree with the inclusion of data for additional chloroplast and nuclear loci and the use of more sophisticated parsimony search methods. By 1994, new research had tentatively identified the sister group of the Proteaceae as the family Platanaceae [8], and by 2000, the Nelumbonaceae had been confirmed, with moderate support, as the third member of the order Proteales [9]. By 2011, the monophyly of the Proteales, its position branching off one node above the base of the eudicots, and the sister group relationship between Proteaceae and Platanaceae had all received 100 % bootstrap support in a phylogenetic study of a large sample of angiosperms, represented by sequences from all three plant genomes, and the Sabiaceae had tentatively been added as the sister group of the Proteales [10].

Identification of the closest relatives of the Proteaceae was an exciting advance in both our understanding of angiosperm

phylogeny as a whole and in our ability to reconstruct relationships within the family with precision. It also revealed the relevance of fossils of the Platanaceae that would soon prove useful in calibrating molecular chronograms. Ironically, the families that make up the Proteales are so morphologically divergent from one another that Johnson and Briggs would have had a difficult job finding many homologous character states shared by them in constructing their "Proto-Proteacea." *Nelumbo*, in particular, is about as different from the Proteaceae as one can imagine, with its aquatic habitat, herbaceous habit, lack of trichomes, numerous tepals, stamens, and carpels. Only one plausible morphological synapomorphy, the distinctive trichome base, has been identified as shared by the Proteaceae and Platanaceae [11], and none has yet been recognized for the Proteales. But Platanaceae and Nelumbonaceae have proved to be very useful outgroups for molecular phylogenetic analyses within the Proteaceae.

The first authors to test the subfamilies, tribes, and subtribes of Johnson and Briggs rigorously were Hoot and Douglas [12], with their parsimony analysis of an alignment of the chloroplast gene *atpβ* and the *atpβ–rbcL* spacer from representatives of 46 genera (Table 1). This sample included the then recently described *Eidothea*, which had been placed in its own new subfamily, as well as representatives of all five of Johnson and Briggs' subfamilies, all 12 of their tribes, and 23 of their subtribes. A slightly confounding factor was the inclusion of two misidentified sequences (their *Protea* was really a *Leucadendron* and their *Roupala* was *Floydia praealta*) [13]. The tree from their full matrix included 33 unequivocal nodes, of which 29 received some bootstrap support, with 16 earning indices of 95 % or higher. It corroborated one of Johnson and Briggs' subfamilies, the Proteoideae, as monophyletic (including *Eidothea* in its basal radiation), could not resolve relationships between the Carnarvonioideae, Sphalmioideae, and Grevilleoideae, and found the Persoonioideae weakly to be paraphyletic. Of the 9 tribes tested for monophyly, only 2, the Banksieae and Oriteae, were supported, and of 12 subtribes tested, 5 were corroborated. Hoot and Douglas concluded (p. 314) that "future work within the family must include a re-evaluation of taxonomic delimitations and hierarchies." This paper represented the greatest single advance in our understanding of proteaceous phylogeny in at least 23 years.

Two limitations of Hoot and Douglas's analysis were the low number of non-Australian genera included in their sample and the fact that they sequenced only parts of the chloroplast genome. Their representation of the African Proteoideae (Johnson and Briggs' tribe Proteeae) was especially low, including only 2 of the 13 currently recognized genera, *Leucadendron* and *Aulax*. These gaps were rectified to some extent in the parsimony analyses published by Barker et al. [14] of an alignment of ITS nrDNA sequences from a sample of 50 species, which included representatives of 11 of the 12 currently recognized African genera of

subfamily Proteoideae (only *Aulax* was missing). All of their analyses strongly contradicted monophyly of the tribe Proteeae and of its two subtribes (Table 1). Instead, these genera formed two separate clades, all but *Protea* and *Faurea* strongly grouping with the Australian genera *Isopogon* and *Adenanthos*. Barker et al. [14] dubbed the larger of the two African clades the "Cape Clade" and found not only strong support for its monophyly but also significant resolution within it, with *Leucadendron* being shown to be the sister group of the rest of the clade with 100 % bootstrap support and several other groupings of genera also receiving moderate to strong support that was incongruent with clades postulated by Johnson and Briggs [4].

It had become clear that a revision of the suprageneric classification of the Proteaceae was necessary and that a comprehensive molecular phylogenetic analysis should strongly influence the shape of the new system. Weston and Barker [13] undertook this task by synthesizing the results of published molecular phylogenetic analyses (based on sequences for the chloroplast regions *atpβ*, *atpβ–rbcL* spacer, *trnL* intron, *trnL/trnF* spacer, *rp116* intron) along with their own trees derived from analyses of alignments of ITS nrDNA and *rbcL* cpDNA as a phylogenetic supertree, in which components were weighted according to the levels of bootstrap support they had received in the original analyses. They were able to include sequence data for all but two of the currently recognized genera, including all six genera that had been named since 1975. Of 31 of Johnson and Briggs' suprageneric taxa that were open to testing for monophyly (i.e., those that included two or more subtaxa), Weston and Barker's supertree decisively tested all of them (Table 1). Two of Johnson and Briggs' three testable subfamilies, 4 of their 11 testable tribes, and 7 of their 17 testable subtribes were corroborated as monophyletic. On the other hand, 18 higher taxa had been rejected as either para- or polyphyletic, necessitating the abandonment of some names, recircumscription of others, the resurrection of several names from synonymy, and the coining of seven new names. Weston and Barker's new classification included 5 subfamilies, 10 tribes, 19 subtribes, and 79 genera (Table 2). It was not fully resolved, with 12 genera having to be included, *incertae sedis*, at various ranks and a few tribes and subtribes receiving only weak support, but this classification represented a significant progressive step in the history of proteaceous taxonomy.

One might argue that incongruence between trees derived from different data sets identifies a taxonomic problem but does not resolve it. However, where weakly and strongly supported results are found to be incompatible, it is reasonable to prefer the latter. In cases of disagreement between Johnson and Briggs's morphology-based tree and Weston and Barker's molecular supertree, it is reasonable mostly to prefer groupings specified by the molecular tree because:

Table 2
Weston and Barker's [13] classification of the Proteaceae, with levels of support for named clades found by Sauquet et al. [18] indicated in columns 5 (Bayesian posterior probabilities from the BEAST analysis) and 6 (parsimony bootstrap percentages)

Weston and Barker's (2006) classification				Sauquet et al. (2009)	
Subfamily	**Tribe**	**Subtribe**	**Genus**	**PP**	**BS**
Bellendenoideae			*Bellendena*		
Persoonioideae				1.00	100
	Placospermeae		*Placospermum*		
	Persoonieae			1.00	100
			Toronia		
			Garnieria		
			Acidonia		
			Persoonia		
Symphionematoideae				1.00	100
			Symphionema		
			Agastachys		
Proteoideae				1.00	57
	Incertae sedis		*Eidothea*		
	Incertae sedis		*Beauprea*		
	Incertae sedis		*Beaupreopsis*		
	Incertae sedis		*Dilobeia*		
	Incertae sedis		*Cenarrhenes*		
	Incertae sedis		*Franklandia*		
	Conospermeae			1.00	100
		Stirlingiinae			
			Stirlingia		
		Conosperminae		1.00	100
			Conospermum		
			Synaphea		
	Petrophileae			1.00	100
			Petrophile		
			Aulax		
	Proteeae			1.00	100
			Faurea		
			Protea		
	Leucadendreae			1.00	100
		Isopogoninae	*Isopogon*		
		Adenanthinae	*Adenanthos*		
		Leucadendrinae		1.00	99
			Leucadendron		
			Serruria		
			Paranomus		
			Vexatorella		
			Sorocephalus		
			Spatalla		
			Leucospermum		
			Mimetes		
			Diastella		
			Orothamnus		
Grevilleoideae				1.00	100
	Incertae sedis		*Sphalmium*		
	Incertae sedis		*Carnarvonia*		
	Roupaleae			1.00	92

		Incertae sedis	*Megahertzia*		
		Incertae sedis	*Knightia*		
		Incertae sedis	*Eucarpha*		
		Incertae sedis	*Triunia*		
		Roupalinae		1.00	100
			Roupala		
			Neorites		
			Orites		
		Lambertiinae		1.00	100
			Lambertia		
			Xylomelum		
		Floydiinae		1.00	100
			Darlingia		
			Floydia		
		Heliciinae		1.00	100
			Helicia		
			Hollandaea		
	Banksieae			1.00	97
		Musgraveinae		1.00	100
			Austromuellera		
			Musgravea		
		Banksiinae	*Banksia*		
	Embothrieae			1.00	98
		Lomatiinae	*Lomatia*		
		Embothriinae		1.00	100
			Embothrium		
			Oreocallis		
			Alloxylon		
			Telopea		
		Stenocarpinae		>0.50	ns
			Stenocarpus		
			Strangea		
		Hakeinae		1.00	100
			Opisthiolepis		
			Buckinghamia		
			Hakea		
			Grevillea		
			Finschia		
	Macadamieae			0.94	<50
		Macadamiinae[a]		1.00	100
			Macadamia		
			Panopsis		
			Brabejum		
		Malagasiinae		1.00	100
			Malagasia		
			Catalepidia		
		Virotiinae		1.00	99
			Virotia		
			Athertonia		
			Heliciopsis		
		Gevuininae		1.00	100
			Cardwellia		
			Sleumerodendron		
			Euplassa		
			Gevuina		
			Bleasdalea		
			Hicksbeachia		
			Kermadecia		
			Turrillia		

[a]Includes two additional genera named in ([17] Mast et al. 2008): *Lasjia* and *Nothorites*

1. The molecular supertree was based to a large extent on congruence between nuclear and chloroplast trees and thus between biparentally and maternally inherited markers, between which recombination is practically nonexistent. They are therefore close to the ideal of independent sources of phylogenetic evidence. Given the number of taxa sampled, the probability of congruence between the trees due to chance alone is vanishingly small. The morphological tree was based on analysis of one data set.
2. The molecular trees on which the supertree was based were all constructed using an explicit phylogenetic method, maximum parsimony, the merits and weaknesses of which had been discussed extensively in the methodological literature. The morphological tree was constructed by manually modifying the results of a previously unpublished constrained parsimony method that was explained briefly by Johnson and Briggs through a worked example.
3. The molecular tree was based on phylogenetic analyses in which the level of empirical support for groupings was explicitly tested using the parsimony bootstrap. Confidence in the groupings of the morphological tree was subjectively assessed.

Weston and Barker [13] based their new classification almost entirely on their supertree. The only groupings that they chose to disregard were the placements of *Carnarvonia* and *Sphalmium*. These clustered with the tribes Macadamieae and Banksieae, respectively, both with weak bootstrap support in the ITS tree, but both unresolved or untested in the other constituent analyses. Both genera lack the most striking synapomorphy possessed by other members of the subfamily Grevilleoideae, grevilleoid flower pairs, so both genera were included in this subfamily as *taxa incertae sedis*.

A major incentive to conduct molecular systematic research on the Proteaceae has been the prospect of testing Johnson and Briggs's [4] biogeographic hypothesis that most clades showing disjunct, transoceanic distribution patterns are the result of vicariance, mediated by continental drift. Until the late 1990s, cladistic biogeographic methods offered the best way of rigorously doing this, and several of Johnson and Briggs's higher taxa offered the possibility of resolving relationships between taxa endemic to three or more Gondwanic continental blocks. However, Weston and Barker's [13] tree resolved few of these relationships cleanly, with uninformative paralogous patterns tending to dominate intercontinental area relationships. By 2006 the advent of sophisticated new methods for conducting relaxed clock molecular dating analyses offered a different way to test hypotheses of vicariance. Vicariance theory predicts that the estimated age of a putatively vicariant clade should be as old as, or older than, the geological age of the barriers that are responsible for the disjunct distribution.

Three research groups proceeded to use a combination of molecular phylogenetic analysis and paleobotanical evidence to estimate molecular chronograms for higher taxa in the Proteaceae. Barker et al. [15] produced a molecular chronogram based on an alignment of *rbcL* sequences for a sample of 45 species representing 45 genera from across the family, analyzed using the Bayesian methods implemented in MrBayes and Multidivtime. Their results were shocking to those of us who had been comfortable in the belief that long-distance dispersal had almost no role in establishing transoceanic relationships in the Proteaceae [16]. Barker et al. concluded that of eight transoceanic disjunctions of sister groups within their sample, four postdate the breakup of Gondwana and are better explained as the result of long-distance dispersal than vicariance. Especially striking was the 30–50 million year estimate for the age of the tribe Leucadendreae, which spans the Indian Ocean and includes the largest African clade in the family, the subtribe Leucadendrinae. To be consistent with a vicariant history, the Leucadendreae would have to be at least 105 million years old, the estimated timing of final separation of Africa from the remains of Gondwana. However, four disjunctions between Australia and South America were all old enough (>33 million years) to be explicable by fragmentation of ancestors that were formerly widespread from Australia to South America through Antarctica.

Mast et al. [17] used the same phylogenetic methods as Barker et al. but concentrated on the widely distributed tribe Macadamieae sensu Weston and Barker, analyzing an alignment of 7 nuclear and chloroplast genes and a morphological data set of 53 characters scored for a sample of species that included all 16 recognized genera of the tribe. They concluded that Barker et al. [15] had underestimated the role of long-distance dispersal in the Macadamieae, largely as a result of incomplete sampling of genera. While Barker et al. had interpreted the biogeographic history of this tribe as predominantly involving vicariance, with one instance of long-distance dispersal across the Atlantic Ocean from South America to Africa, Mast et al. concluded that no transoceanic disjunctions in the Macadamieae were old enough to have been caused by continental drift. Interestingly, they found a significant correlation between evolutionary origins of indehiscent fruits and reconstructed dispersal events, implying that sealed fruits have facilitated long-distance dispersal, presumably by protecting the embryo from toxic salt water.

Relaxed clock molecular dating has its methodological limitations, the most obvious of which is that fossils can provide only the minimum age of a clade or character, so they are likely mostly to provide underestimates of the ages of clades to which they are assigned. Fossil-calibrated molecular chronograms are therefore more likely to underestimate the ages of clades than overestimate them. A second limitation is that the affinities of many fossils have only been determined by non-phylogenetic methods such as

simple matching and may be inaccurate due to symplesiomorphy or convergence. Mistakes of this kind are likely to result in overestimation of the ages of clades. Thirdly, fossils are preserved parts of dead organisms and are therefore fragmentary and usually offer far fewer morphological characters than living plants, limiting our ability to associate them precisely with nodes of phylogenetic trees. The estimated ages of some transoceanic disjunctions in the Proteaceae are not much lower than the ages of the ocean basins that they span, and these might well represent cases in which vicariance has been falsely rejected because of a general bias towards underestimation of clade ages in molecular chronograms. *Knightia*, a monotypic genus endemic to New Zealand, is a good case in point. Its precise placement near the root of the tribe Roupaleae is not yet known with much confidence, but wherever it belongs, its lineage dates back at least 35–70 million years, consistent with divergence as little as 14 million years after rifting commenced between Zealandia and the remains of Gondwana, 84 million years ago. *Knightia* is thus a reasonable candidate for a plant group that occurs in New Zealand as a result of vicariance and survived widespread inundation of Zealandia during the late Oligocene. Other clades that have been estimated to be almost old enough for vicariance are the African-Australian tribe Petrophileae and the weakly supported clade consisting of the African-Madagascan tribe Proteeae and the New Caledonian genus *Beauprea* [18].

Another research group used molecular dating for an entirely different purpose, to discover whether high species richness in some biodiversity hotspots with Mediterranean climates was the result of a long period of occupancy or high rate of diversification [18, 19]. Sauquet et al. [18] used three different tree-building techniques (parsimony, and the Bayesian phylogenetic algorithms implemented in MrBayes and BEAST), three different relaxed clock dating methods (penalized likelihood, and the Bayesian methods implemented in Multidivtime and BEAST), and a data set that included alignments of 8 DNA regions scored for 79 genera, the most comprehensive molecular sample of proteaceous genera yet assembled. They calibrated their tree using a subset of 25 fossil pollen grains that had been described in the paleobotanical literature but tested their identities using an innovative, parsimony-based method for associating fossils with the nodes of a phylogenetic tree. This study produced a tree (Fig. 2) that was highly congruent with the new classification of Weston and Barker [13] (Tables 1 and 2), necessitating no changes to the circumscriptions of named taxa. Sauquet et al. estimated net diversification rates for every branch in their tree and found that rates were significantly higher in some of the lineages concentrated in Mediterranean biodiversity hotspots (tribes Persoonieae and Leucadendreae and subtribes Hakeinae and Banksiinae) than in those from elsewhere (Fig. 3). However, very slowly diversifying lineages were also found in all areas with

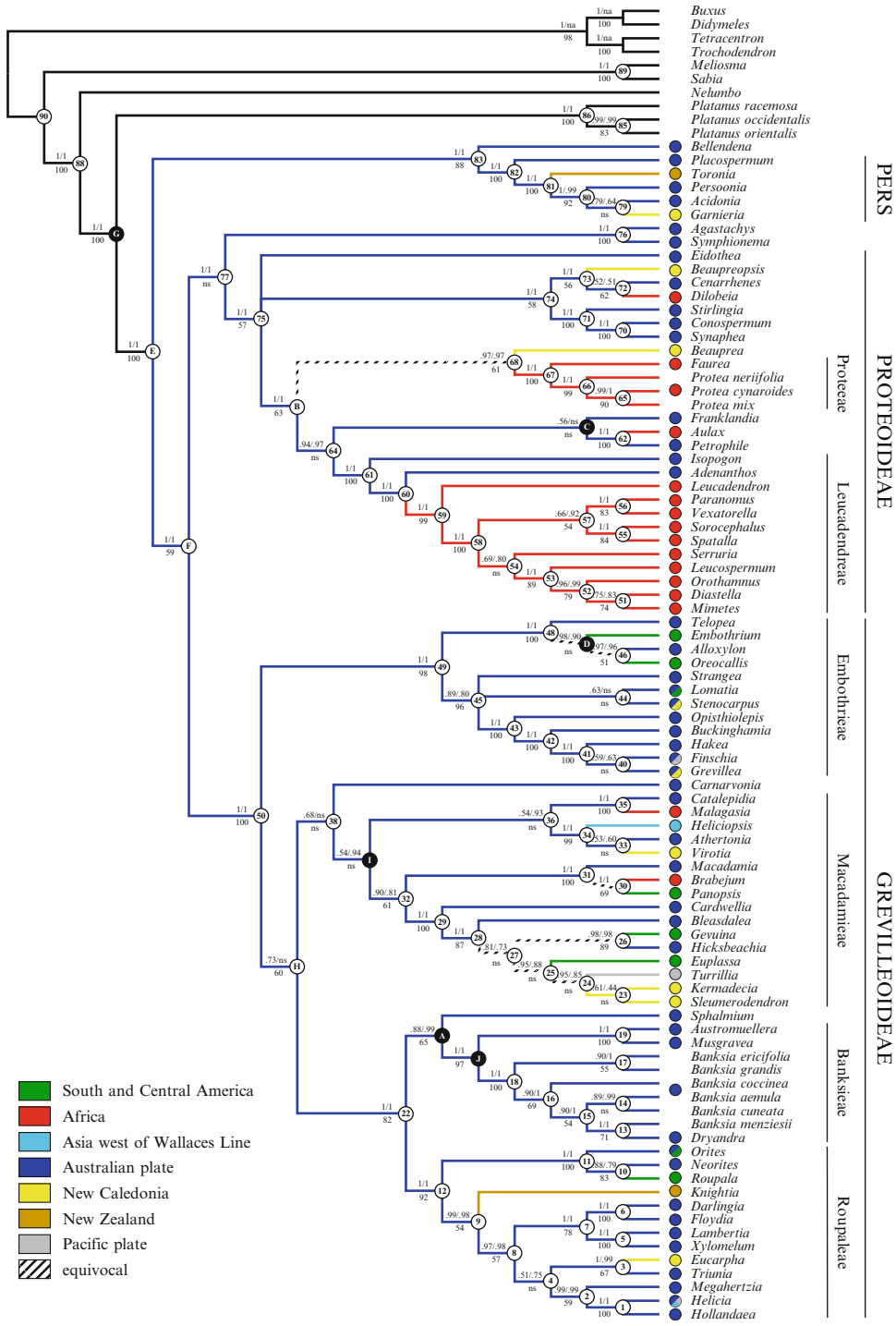


Fig. 2 Majority-rule consensus tree from the unpartitioned Bayesian analysis in MrBayes of Sauquet et al. ([18], figure S1, reproduced with permission of the senior author), with parsimony optimization of biogeographic distribution. Branch support: posterior probabilities from MrBayes (*above*, *left*) and BEAST (*above*, *right*) and parsimony bootstrap values (*below*); ns = not supported (branch collapsing or support <0.5); na = not applicable. Node labels are arbitrary unique numbers. *Plain black dots* identify nodes on which age constraints were applied in the dating analyses. *PERS* Persoonioideae

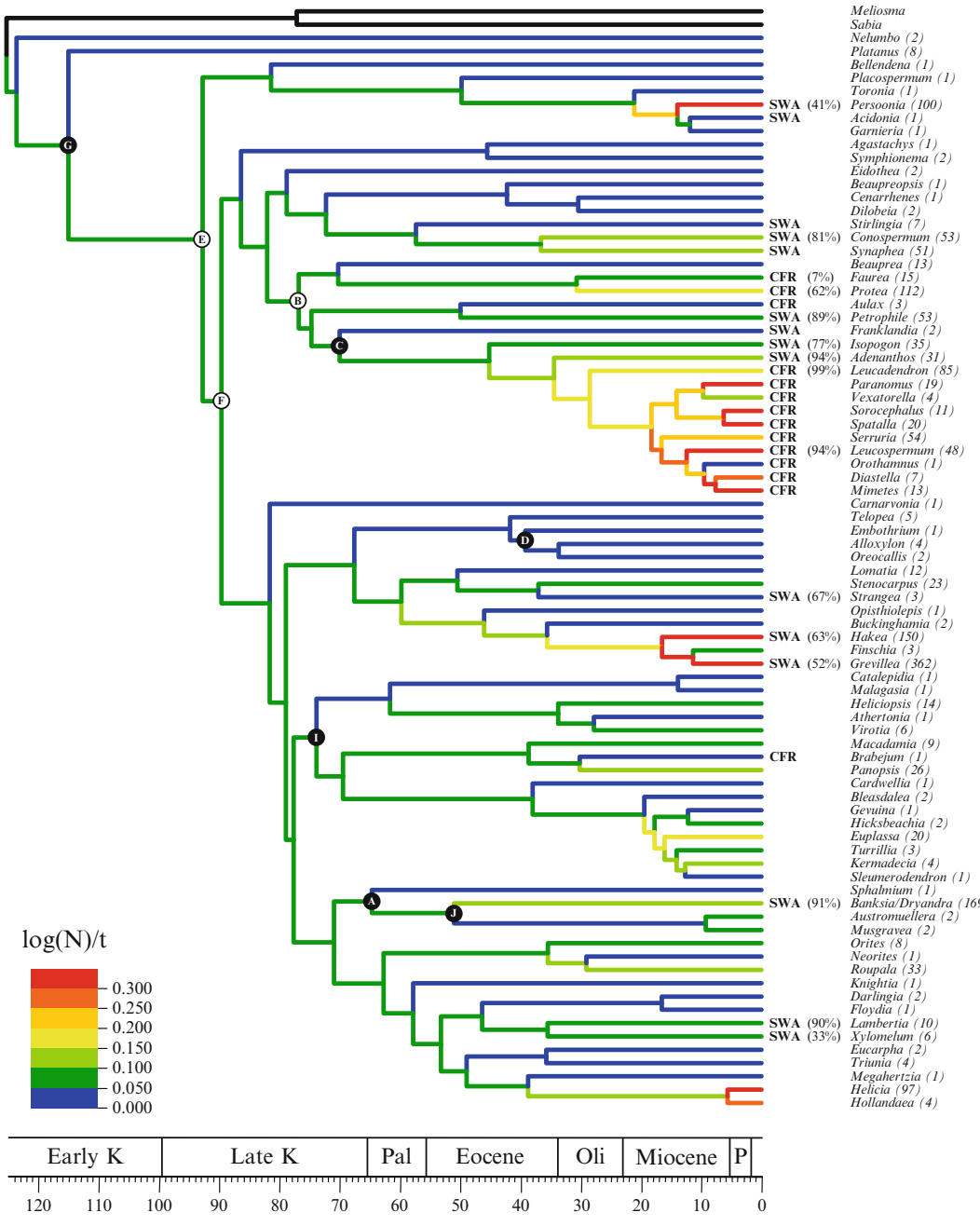


Fig. 3 Molecular chronogram for the Proteaceae, produced using the Bayesian uncorrelated lognormal method (implemented in BEAST) from an alignment of nucleotide sequence data for eight loci ([18], Fig. 1, reproduced with permission of the senior author). Nodes associated with fossil age constraints (as uniform prior distributions) are marked with *black dots. White dots* identify additional, redundant, or uninformative age constraints. Branches are *colored* according to absolute net diversification rate by stem age of their subtending clade. Taxa present in Mediterranean hotspots are identified with either SWA (south-western Australia) or CFR (Cape Floristic Region). Where not endemic to these hotspots, the percentage of hotspot species is mentioned in *brackets*. Total species numbers are indicated in *brackets* after the name of each taxon. Absolute ages are in million years

Mediterranean climates, including the central Chilean biodiversity hotspot, where the Proteaceae are represented by only three species, from two tribes and three subtribes. Sauquet et al. [18] had to agree with Johnson and Briggs [4] that low soil fertility was much more likely than Mediterranean climates to be responsible for the high diversity of Proteaceae in south-western Australia and the Cape Region of South Africa.

The results of the analyses of Sauquet et al. had interesting methodological as well as theoretical implications. Firstly, the three different tree-building algorithms gave remarkably similar results, with parsimony supporting 71 of the 84 nodes found within the Proteaceae by MrBayes. The BEAST tree was even more highly congruent with the MrBayes tree, supporting 80 of its 84 nodes. Of particular interest was the difference between all tree-building methods with respect to their placements of *Carnarvonia*. Parsimony had this genus nested with the tribe Macadamieae, whereas MrBayes found it to be the sister group of this tribe. BEAST, however, placed it as sister group to the rest of the subfamily Grevilleoideae, consistent with Johnson and Briggs's [4] morphology-based tree. This implies that incorporating a relaxed molecular clock into the model assumed by the tree-building algorithm may be a strength rather than a weakness of BEAST. The estimates of clade ages provided by the three different relaxed clock molecular dating analyses were also highly correlated, suggesting that, despite their disparate assumptions, these methods can produce highly congruent and, presumably, robust results. Congruence was low between many of the original identifications of fossil pollen grains reported in the primary literature and those inferred by the authors' parsimony analysis of palynological characters, constrained by the best molecular tree. The fact that most molecular dating analyses rely on calibration fossils that have never been phylogenetically analyzed at all is a reason to be cautiously skeptical of their results.

Molecular systematic studies of suprageneric relationships within the Proteaceae have severely tested the classification and biogeographic hypotheses of Johnson and Briggs [4], found some of their ideas wanting, and in most cases provided well-corroborated replacements for them (Tables 1 and 2). They have also implied that some of Johnson and Briggs's "morphological and adaptational principles" needed critical reexamination. For example, the assumption that reduction in ovule number from multiple ovules per carpel in the most recent common ancestor of Proteaceae to two and then to one has not been supported by the published molecular trees. The most parsimonious reconstruction of the evolution of ovule number in the Proteaceae resolves the ancestral condition for the family as two ovules per carpel, with multiple parallel decreases to one and increases to a range of higher numbers in diverse lineages, possibly also with reversals back to two in

some groups (unpublished data). Similarly, Johnson and Briggs's assumption that changes in carpel orientation from anteroposterior to diagonal would only be possible in actinomorphic flowers is not supported by the reconstruction of ancestral zygomorphic floral symmetry in taxa such as the tribe Embothrieae (unpublished data), in which carpel orientation has clearly been quite labile.

A number of botanists who were not involved in the primary molecular systematic research have used the new phylogenetic knowledge to test other evolutionary hypotheses. Jordan et al. [20], for instance, used a combination of anatomical and environmental data and a supertree constructed from published molecular phylogenies to test competing explanatory hypotheses for the evolution of sclerophylly in the Proteaceae. They found that origins of sclerophyllous anatomical features were significantly correlated with brightly illuminated environments, not with low water availability as predicted by conventional functional explanations for sclerophylly. Sclerophylly was interpreted as an adaptation to protect foliar mesophyll tissue from photo-inhibition, not from desiccation. Some of the same scientists went on to test the association between concealed stomata and dry environments, finding a significant correlation between deep stomatal encryption and aridity [21] but no correlation between shallow encryption and any environmental variable. Crisp et al. [22] used trees from Mast et al. [17] and Sauquet et al. [18], to reconstruct the evolution of biome transitions as part of a broader meta-analysis examining the frequency of such transitions and their association with instances of long-distance dispersal in plant taxa of the southern hemisphere. They found that biome transitions were surprisingly rare across their huge sample extracted from 45 published molecular systematic studies of a wide range of plant groups.

3 Monophyly of Genera and Phylogenetic Relationships Within Them

Most effort in molecular taxonomic research on the Proteaceae so far has concentrated on testing the monophyly of taxa higher than generic rank, but 81 genera are currently in common use in scientific communication and 55 of these comprise two or more species and thus need to be tested for monophyly. The tendency of previous generations of taxonomists to distinguish closely related genera on the basis of alternative states for a single character has provided us with a number of generic pairs and triplets where one or two monophyletic genera are likely to be nested within a paraphyletic residue that has also been treated as a genus. One consequence of the phylogenetic revolution in biological systematics is that the great majority of systematists now regard this practice as a kind of error, of which *Banksia* and *Dryandra* provide a textbook example: *Dryandra* possesses distinctive inflorescence synapomorphies, *Banksia sensu stricto* has no morphological synapomorphies

at all, but *Banksia sensu lato* (i.e., including *Dryandra* as an infrageneric taxon) possesses several synapomorphies of the wood, inflorescence, and fruit [4]. Other generic pairs and triplets that seem likely to illustrate this problem once they have been investigated more rigorously include the following: (a) *Hakea* and *Finschia*, which are both characterized by distinctive morphological synapomorphies of their fruits, but which are likely to be nested within *Grevillea*, which has no obvious synapomorphy of its own; (b) *Strangea*, characterized by woody, serotinous, two-seeded follicles and which is likely to be nested within *Stenocarpus*, with leathery, multiseeded fruits; (c) *Spatalla*, which has distinctively zygomorphic flowers and which seems likely to be nested within actinomorphic flowered *Sorocephalus*; (d) *Orothamnus* and *Diastella*, which both have terminal inflorescences, surrounded by large, colorful, involucral bracts and which are both probably nested within *Mimetes*, with axillary inflorescences and less conspicuous involucral bracts.

The first published molecular taxonomic test of the monophyly and phylogeny of a proteaceous genus was Mast's [23] analysis of *Banksia* and *Dryandra*, followed several years later by papers by Mast and Givnish [24] and Mast et al. [25], based on a larger set of DNA regions and more comprehensive samples of species. The primary aim of this project was to test the morphology-based classification of George [26, 27] and the morphology-based cladogram and cladistic classification of Thiele and Ladiges [28]. Thiele and Ladiges treated *Banksia* and *Dryandra* as sister taxa, as indicated by the existing classification, and proceeded to polarize characters in *Banksia sensu stricto* on the basis of outgroup comparisons with *Dryandra*.

The most dramatic taxonomic finding of the molecular systematic results was the relationship between *Banksia* and *Dryandra*, with the few sampled species of *Dryandra* forming a clade that was deeply nested within a paraphyletic *Banksia*. Thiele and Ladiges's morphological cladogram had been rooted on a strikingly erroneous branch as a consequence of treating *Dryandra* as an outgroup. George's infrageneric classification was similarly disoriented with respect to Mast and Givnish's moderately well-supported and resolved tree. The basal dichotomy in the molecular tree separated the subtribe Banksiinae into two clades characterized by markedly different positioning of stomata on the abaxial leaf surface. One, which has superficial stomata and is comprised of about half of the species of *Banksia sensu stricto*, was given the informal name *Phanerostomata*. The other, which has the stomata deeply encrypted within cavern-like pits and is comprised of *Dryandra* plus the other half of *Banksia sensu stricto*, was dubbed *Cryptostomata*. Thiele and Ladiges's morphological cladogram treated *Phanerostomata* as a clade nested within a paraphyletic *Cryptostomata* and thus could be simply re-rooted on the branch connecting these two parts of the tree. Once re-rooted, substantial parts of the morphological tree

can be seen to be congruent or combinable with the molecular tree. Of George's 14 higher infrageneric taxa that were open to testing, 5 were corroborated as monophyletic, but 7 were found to be polyphyletic, 6 of them moderately to strongly so. Thiele and Ladiges' [28] cladistic classification fared better, with 11 of its higher infrageneric taxa of more than one species being corroborated as monophyletic but 8 rejected as polyphyletic.

Mast and Thiele [29] proceeded to revise the classification of subtribe Banksiinae to recognize only corroborated clades as named taxa by transferring all species of *Dryandra* to *Banksia* and formalizing the names of the two main clades as subgenus *Banksia* (*Cryptostomata*) and subgenus *Spathulatae* (*Phanerostomata*). This nomenclatural change has provoked an interesting sociological phenomenon: a popular revolt among some Australian native plant enthusiasts who do not want to abandon the generic name *Dryandra* [30], actively encouraged by a taxonomist who refuses to abandon those parts of his classification that have been shown to be incompatible with improved knowledge of phylogeny [31].

Other authors have repeatedly reanalyzed the sequence data that Mast's research group assembled, to test hypotheses in biogeography and evolutionary ecology. Crisp and Cook [32], for instance, produced a molecular chronogram for *Banksia* as part of their analysis of the timing of east–west disjunctions in a diverse sample of taxa that are bisected by the arid, calcareous Nullarbor Plain in southern Australia. They found a signature of numerous congruent disjunctions that coincided with the mid-Miocene origin of this edaphic and climatic barrier. Of the two disjunctions evident in *Banksia*, one agreed closely with the congruent signal, while the other was found to be significantly older than it, apparently associated with an earlier inundation of the Nullarbor area.

He et al. [33] also used the data of Mast and Givnish [24] for a relaxed clock molecular dating analysis, with the aim of reconstructing the timing of the evolution of adaptations to fire. They found that two such adaptations, serotinous fruits and retention of dead flower parts on infructescences, were ancestral states in *Banksia*. They concluded that fire has been a selective force in the Australian environment since the origin of the stem lineage of *Banksia*, over 60 million years ago. This conclusion overlooks the point that these features are equally likely to have originated as recently as 45 million years ago, at the most recent "end" of the stem lineage, just before the diversification of extant *Banksia* taxa. Nevertheless, these authors were able to demonstrate the considerable antiquity of fire as a significant selective pressure on the biota of the Australia–Antarctica–South America remnant of Gondwana.

The potential problem of generic paraphyly or polyphyly in the Proteaceae is underscored by the results of a couple of studies that have focused on relationships among genera but which have also tested the monophyly of some genera by including two or more

sampled species in their analyses. Barker et al. [14], for example, tested the monophyly of two Australian and nine African genera, each represented by between two and seven species in their molecular phylogenetic analysis of the tribe Leucadendreae. Although the sampled species of seven genera formed monophyletic groups on their best tree, four were found to be para- or polyphyletic, each with levels of bootstrap support for groupings of some of their species with other genera ranging from weak (60 %) to strong (95 %). No one has yet implemented nomenclatural changes on the basis of these results because of the limited sample of species and the weakness of support for some of the putative groupings that challenged long-accepted circumscriptions of genera.

Mast et al. [17], on the other hand, did make taxonomic changes in response to the results of their analysis of the phylogeny of the tribe Macadamieae (already discussed above). Six genera were represented by multiple species, the most heavily sampled of which was the economically and horticulturally important *Macadamia*, with all nine of its then recognized species included in the analysis and eight of them sequenced for at least one gene. A further three genera with multiple species were comprehensively sampled at the species level. The most surprising result of this analysis was the finding, supported by both morphological and molecular data sets, that *Orites megacarpus* was not really an *Orites* at all and was not even a member of the tribe Roupaleae, to which *Orites* belongs [13], but was instead the sister group of *Panopsis* in the tribe Macadamieae. A new generic name, *Nothorites*, had to be created for this species. *Macadamia* was found to be paraphyletic, with *Panopsis*, *Nothorites*, and *Brabejum* nested within it. This necessitated the creation of another new generic name, *Lasjia*, to take the five tropical species of *Macadamia*. On the positive side, the sampled species of *Panopsis*, *Heliciopsis*, *Virotia*, and *Kermadecia* all formed clades.

Numerous infrageneric classifications in the Proteaceae need to be tested, requiring analyses in which as many extant species within the targeted genera are sampled as possible. Moreover, phylogenetic trees resolved down to species level and below are needed for testing some of the most interesting hypotheses in evolutionary biology and biogeography. Good progress has been made towards achieving this goal in several genera, resulting in the publications on *Banksia* [23–25, 29], discussed above, as well as detailed phylogenetic studies of *Leucadendron* [34], *Protea* [35, 36], and *Hakea* [37]. While most of these authors justified their studies primarily on grounds other than taxonomy, all noted the degree of congruence between their results and existing infrageneric classifications, which varied substantially from genus to genus in line with the richness of both morphological and molecular sources of evidence.

Leucadendron, in which most qualitative morphological characters used by Williams [38] to distinguish higher taxa came from

the fruits, showed poor congruence between the existing classification and the weakly to well-supported components of the maximum parsimony ITS tree of Barker et al. [34]. Eleven of Williams's sections and subsections were tested, and of these, only one was corroborated as monophyletic. Four were found to be para- or polyphyletic with moderate to strong bootstrap support, while evidence for or against monophyly of the remaining six sections and subsections was too meager to list.

The study of *Protea* by Valente et al. [36], which was based on Bayesian analyses of 4 chloroplast and 2 nuclear DNA sequences and 138 AFLPs, showed a higher level of congruence with existing, morphology-based, intuitively constructed classifications than did the molecular tree for *Leucadendron*. Sixteen of Rebelo's [39] 17 informal species groups were tested, and of these, 5 were corroborated as monophyletic. Ten species groups, on the other hand, were found to be polyphyletic, eight of them strongly so, a result that Valente et al. ([36], page 755) surprisingly concluded was "in broad agreement with recent ideas about their taxonomy." The higher level of congruence between the molecular tree and the morphology-based classification in *Protea* than in *Leucadendron* presumably reflects both the broader source of morphological characters that could be recognized in *Protea* and the much larger, more diverse molecular data set.

A well-supported, highly resolved tree for 55 species of *Hakea* was produced by Mast et al. [37] from a combination of 46 morphological character, 4 chloroplast, and 3 nuclear DNA sequences, analyzed using Bayesian and parsimony methods. This tree found mixed results when compared with the informal infrageneric classification of Barker et al. [40]. Of the 19 infrageneric groupings that the tree of Mast et al. tested for monophyly, 7 were corroborated as monophyletic but 10 were found strongly to be polyphyletic and a further 2 paraphyletic. Moreover, Mast et al. found a well-supported basal split in *Hakea* that had not been recognized previously, which separated a clade characterized by obscure leaf venation from one with leaves with prominent, parallel veins. These authors refrained from erecting a new infrageneric classification of *Hakea* on the grounds that morphological evidence was not sufficiently informative to allow them confidently to classify unsequenced taxa.

The biogeographic and evolutionary hypotheses that Barker et al. [34], Valente et al. [36], and Mast et al. [37] sought to test with their trees were both varied and biologically interesting. Both Barker et al. and Valente et al. produced center of origin/dispersal scenarios for their genera, the former using a parsimony-based dispersal–vicariance algorithm (DIVA) and the latter a more sophisticated set of Bayesian probabilistic methods. Both sets of authors concluded that their genera had originated in the Cape Floristic Region, where a Mediterranean climatic regime now predominates, and undergone range expansions into what are now summer–wet regions further north and east. In the case of *Protea*, considerable

range expansion by one clade into tropical Africa, as far north as Eritrea and Guinea, was inferred, accompanied by diversification at a steady rate. This conclusion strongly contradicted earlier speculation (e.g., [4]) that *Protea* had a tropical origin from which it invaded the Western Cape, where it was thought to have rapidly diversified.

Mast et al. [37] eschewed biogeographic reconstruction in favor of retesting the hypothesis of Hanley et al. [41] that bird pollination had evolved multiple times from insect-pollinated ancestors in *Hakea* and that the earlier evolution of cyanogenesis in floral organs had preadapted these lineages to deter potentially destructive pollinators from damaging their flowers. Hanley et al. had built their analysis on the foundation of the morphological tree of Barker et al. [34], which Mast et al. confidently showed had been rooted on a strikingly wrong internode. Their combined morphological/molecular tree was more consistent with multiple changes to insect pollination from a bird-pollinated ancestor in *Hakea*. Floral cyanogenesis in insect-pollinated *Hakea* species was shown to be an example of phylogenetic inertia rather than preadaptation.

These four detailed, generic-level studies have focused on various aspects of the evolution of these taxa but especially on the evolution of adaptations to the drier, more strongly illuminated, more fire-prone environments that came to dominate vast areas of Australia and Africa during the late Cenozoic Era. Encrypted stomata, scleromorphic leaf anatomy, serotinous fruits or infructescences, and post-fire resprouting from lignotubers or epicormic buds have all been shown to have significant, sometimes complex associations with these environmental changes. Perhaps most importantly, however, these papers have laid foundations for publications in evolutionary ecology that were not anticipated by the original authors and also, one presumes, for studies that have yet even to be planned.

4 Molecular Taxonomy at the Species Level and Below

Although numerous surveys of genetic variation have been conducted within variable species and species groups in the Proteaceae, none has claimed an overtly taxonomic aim. These authors saw their studies instead as applications of population genetic analysis in aid of conservation biology (e.g., [42–45]), reproductive biology (e.g., [46–51]), population structure (e.g., [52–54]), and testing modes of speciation (e.g., [55]). Despite not mentioning taxonomy or systematics in their titles and mostly omitting discussion of these subjects in their text, the results of many of these studies have useful taxonomic implications. These have included corroboration of the distinctness or monophyly of existing taxa (e.g., [49, 53, 55]), the demonstration that some previously recognized taxa are para- or polyphyletic (e.g., [53, 55]), the

demonstration of hybridization or introgression between taxa [51, 52], the resolution of phylogenetic relationships between closely related taxa (e.g., [53, 55]), and the resolution of previously unrecognized population structure that could reasonably be used to circumscribe different taxa [42, 45, 46, 48, 54].

Studies of genetic variation at and below the species level have used a variety of sources of genetic data, including allelic variation in allozymes (e.g., [46–48, 52, 53]), restriction fragment length polymorphisms from nuclear and chloroplast loci [42], amplified fragment length polymorphisms (AFLPs, e.g., [43, 44, 51]), and allelic variation at microsatellite loci (e.g., [45, 49, 50, 54]). Most of these studies have used standard taxonomic approaches to characterize population structure, in addition to population genetic indices summarizing degrees of differentiation between populations, levels of gene flow, inbreeding, etc. The taxonomic methods have included hybrid indices [51], multivariate ordination of genetic distances [49, 54], UPGMA clustering of genetic distances [42, 46, 48], and phylogenetic analysis of allele frequencies using Felsenstein's [56] maximum likelihood algorithm implemented in the CONTML program of the PHYLIP package[42, 48, 53, 55]. Some recently published studies [54, 55] have also used newer Bayesian methods based on probabilistic population genetic models for clustering individuals into populations (e.g., STRUCTURE [58]) and for assigning individuals to hybrid categories (e.g., NEWHYBRIDS [59]).

This latter class of analysis has significant advantages over more traditional approaches, and not surprisingly, they have quickly gained popularity. They do not assume the divergent evolutionary model adopted explicitly by phylogenetic analysis and implicitly by UPGMA clustering, an assumption that is likely to be seriously violated below the species level. Unlike CONTML and many genetic distance indices, the newer Bayesian methods do not assume the absence of mutation within the study group, an assumption that is likely to be strongly violated by hypervariable microsatellite loci [57]. Moreover, the assumptions that these methods do make, such as Hardy–Weinberg equilibrium within populations, seem biologically reasonable.

One of these papers [55] warrants closer examination because it dissected a complex data set from several different angles, using some interesting techniques. Prunier and Holsinger analyzed variation in 10 microsatellite loci sampled from 36 populations of six species and two subspecies of the white sugarbushes of the African genus *Protea* [36] with the aim of testing whether this small clade had undergone a recent, explosive radiation. They concluded that the radiation had been recent but gradual and that geographic isolation had played an important role in differentiation of this clade. In doing so, they made some unexpected taxonomic discoveries using both STRUCTURE and CONTML. STRUCTURE was

used to test the distinctness of named taxa and CONTML to test their monophyly, to reconstruct relationships between taxa, and to quantify the relative contributions of molecular divergence between taxa and between populations within taxa. Two of the named taxa were corroborated as distinct clusters by STRUCTURE and as clades by CONTML, but all of the other taxa were found to be problematic, either by showing signatures of sporadic introgression or unidentified sympatry with other taxa in the complex (revealed most effectively by STRUCTURE) or by showing evidence of polyphyly (revealed most effectively by CONTML). The markedly disjunct *Protea mundii* provided the best evidence of polyphyly, with its eastern populations being genetically indistinguishable from the parapatric *P. aurea* subsp. *aurea* and its western populations forming a clade that is sister to their geographic neighbor, *P. lacticolor*. Interestingly, the western populations had been taxonomically suspect for years [55], presumably on the basis of subtle morphological differences that had not been judged to be important enough to warrant taxonomic subdivision. Another notable taxonomic insight was the discovery that a "population" identified a priori as supposedly pure *P. punctata* turned out to be a mixed community of *P. punctata* and sympatric *P. venusta*. This particular finding highlights an improvement that could have been made to the design of the analysis: STRUCTURE could have been used to select out hybrids and reallocate misidentified individuals to circumscribe "pure" taxa prior to phylogenetic analysis, which would probably have avoided some anomalous placements of confounded "populations" on the CONTML tree. Nevertheless, Prunier and Holsinger's paper illustrates well the great potential that molecular systematics has for helping us to resolve in exquisite detail the taxonomy of morphologically difficult species complexes. The main obstacle standing in the way of this potential being realized is the presently high cost of acquiring molecular evidence relative to the cost of morphometric data.

5 The Future

Much work still needs to be done on the molecular systematics of the Proteaceae at all taxonomic levels. Relationships between the subfamilies are mostly still weakly supported as is the monophyly of the subfamily Proteoideae. We are confident about the monophyly of most tribes except the Macadamieae, but interrelationships between tribes are fragile, as are the positions of a number of genera that group weakly with other taxa at this level: *Eidothea*, *Beaupreopsis*, *Cenarrhenes*, *Dilobeia*, *Beauprea*, *Franklandia*, *Carnarvonia*, and *Sphalmium*. The subtribes are well supported with the exception of the Stenocarpinae, but several genera, *Knightia*, *Eucarpha* and *Triunia*, still float uncomfortably about in

the tribe Roupaleae. Relationships between genera within subtribes are often poorly known, and the monophyly of most genera has not even been tested at all. Testing relationships between genera using next-generation DNA sequencing techniques would be a very powerful way to approach these problems. For example, the entire chloroplast genome and nuclear ribosomal RNA repeat unit can be assembled for one representative of each genus using a fraction of a single run on the Illumina sequencing platform [60]. Such a molecular data set, once aligned and phylogenetically analyzed, would have the potential to decisively resolve most, if not all, of the questions outlined above, using two sources of evidence that recombine only occasionally in evolutionary time and thus approach the ideal of independent phylogenetic markers.

Reconstructing phylogenetic relationships within all genera down to species level is our most formidable remaining task. At least five different research groups have made a start on this enormous undertaking, but detailed phylogenetic studies just above the species level show that this will not be a job for the fainthearted. Taxa that are the subject of current and ongoing species-level studies include the subfamilies Persoonioideae, Proteoideae, and Symphionematoideae; subtribes Hakeinae, Heliciinae, and Embothriinae; and the genera *Banksia*, *Lomatia*, and *Orites*.

The main obstacles in the way of achieving these ambitious goals are complexities that one encounters when trying to resolve relationships between very recently differentiated species, most notably low numbers of variable sites, incomplete lineage sorting (also referred to as "deep coalescence"), and introgressive hybridization [61]. The most efficient strategy for pursuing phylogenetic resolution down to species level seems to be to sequence a range of both nuclear and chloroplast genes (the latter analyzed as a concatenated unit) and to assess the resulting gene trees for congruence. If several nuclear genes are phylogenetically congruent with each other and with the chloroplast gene tree, then the strict consensus of those trees is probably an accurate species tree. However, if different gene trees are found to be strongly incongruent, incomplete lineage sorting and/or introgressive hybridization probably needs to be invoked. The methods that have been developed for reconciling different, incongruent gene trees to produce a species tree require the sequencing of such large numbers of loci for large samples of organisms [62] that they are presently viable only for studies demonstrating the feasibility of the approach (e.g., [63]). However, they are unlikely to remain so impractical for long, with the cost of high-throughput DNA sequencing continuing to plummet. In the meantime, the judicious use of Bayesian clustering and phylogenetic analysis of allele frequency data (e.g., [55]) is probably the most cost-effective option for resolving recent radiations with confidence.

6 Conclusions

The development of molecular systematics has been largely responsible for making the last 25 years the most exciting period in the history of plant taxonomy. Few angiosperm taxa above generic level have escaped having their monophyly rigorously tested and their membership recircumscribed if found not to be monophyletic. The history of progress in the Proteaceae parallels that of the angiosperms as a whole. In 1975 when Johnson and Briggs published their phylogenetic tree and classification of the family [4], they knew that the Proteaceae were taxonomically isolated but they had no idea of the family's closest relatives. They were subjectively confident about the monophyly of the Proteaceae and of many of the higher taxa that they recognized within the family, but their confidence could not be quantified given only the analytical tools available then.

Thanks to the development of methods of phylogenetic analysis since the mid-1960s, later paralleled by the spectacular blossoming of new techniques in molecular biology, we now know the phylogenetic position of the Proteaceae with a high degree of confidence. Botanists have been able to test the monophyly of all of Johnson and Briggs's infrafamilial taxa down to generic level as well as the monophyly of 22 of their 55 testable genera and have been able to quantify the degree to which taxa are consistent with the evidence. A new classification above generic level has been published that reflects these substantial improvements in our knowledge.

Discoveries made through molecular systematic analyses at all taxonomic levels have facilitated progress at other levels, although the predominant direction of this information flow has been downwards from order to species. Narrowing down the closest relatives of the Proteaceae to the sequential sister groups Platanaceae and Nelumbonaceae has enabled the resolution of basal relationships within the family and precise circumscription of subfamilies and most of the tribes. This in turn has provided the necessary framework for rooting trees within subtribes and genera. Our improved understanding of relationships within *Banksia* and *Hakea* is a classic example of this process, morphological cladograms for both of which [28, 40] had been rooted at profoundly inappropriate nodes prior to the availability of molecular trees [23, 37]. Conversely, discoveries of generic polyphyly, such as that of *Orites* [17], have sometimes necessitated the recircumscription of tribes and subtribes.

Johnson and Briggs [4] also elaborated a detailed scenario of the evolution of the family in space and time, invoking new knowledge from the earth sciences to explain repeated distributional patterns across the southern hemisphere and evolutionary shifts between biomes. However, the evidence available to them for arbitrating between hypotheses of vicariance and long-distance dispersal amounted to an intuitively assessed "morphological clock":

disjunct taxa were judged to be probably old enough for Gondwanic vicariance if their disjunctions were associated with impressive morphological differences. The pattern of generic distributions, in which several genera are shared by South America and Australia but none by Africa, South America, and Australia, was interpreted as consistent with the temporal sequence of continental separation, in which Africa rifted from the rest of Gondwana long before Australia and South America.

Molecular chronograms have now provided an explicit basis for estimating the minimum ages of disjunct taxa, along with measures of their precision. These estimates have replaced the earlier "eyeballed" assessments, in some cases corroborating earlier conclusions but more often calling them into question. According to both morphological scrutiny and molecular chronograms, *Brabejum* did arrive in Africa and *Faurea* in Madagascar by long-distance dispersal. Joseph Hooker was most probably right in thinking that *Lomatia* and the subtribe Embothriinae are remnants of an Antarctic flora. On the other hand, the tribes Leucadendreae and Macadamieae, although both morphologically diverse, now seem to have acquired their disjunct distributions by long-distance dispersal, not by vicariance as previously thought. Other evolutionary hypotheses have been substantially refined as a result of progress in molecular systematics. Yes, the ancestral biome of the subfamily Grevilleoideae is confidently reconstructed as rainforest but that of the family as a whole might well have been sclerophyllous shrubland, implying multiple transitions back and forth between these biomes in three subfamilies. Yes, insect pollination appears to be ancestral for the family, followed by multiple transitions to vertebrate pollination, but multiple reversals to insect pollination have been confidently reconstructed within *Hakea* and appear likely in other taxa too, contrary to Johnson and Briggs's narrative.

Additional effort needs to be put into resolving the phylogeny of the Proteaceae above generic level because the strength of support for some named higher taxa and many unnamed nodes is still only weak to moderate. However, the focus of newly funded molecular systematic research has shifted to reconstructing relationships within genera down to species level. Achieving this goal will require further improvements in methods of phylogenetic analysis, high-throughput DNA sequencing, and computing infrastructure for managing vast quantities of molecular data. But the prospects for exciting progress on all of these fronts appear to be bright.

Acknowledgements

I am grateful to Barbara Briggs and Maurizio Rossetto for their critical comments on an earlier draft of the manuscript and to Karen Rinkel who kindly drafted Fig. 1.

References

1. Brown R (1810) On the Proteaceae of Jussieu. Trans Linn Soc London 10:15–226
2. Hooker JD (1853) The botany of the Antarctic Voyage of H.M. Discovery Ships *Erebus* and *Terror*, in the years 1839–1843. II Flora Novae-Zelandiae. Lovell Reeve, London
3. Hooker JD (1859) The botany of the Antarctic Voyage of H.M. Discovery Ships *Erebus* and *Terror*, in the years 1839–1843. III Flora Tasmaniae. Reeve, London
4. Johnson LAS, Briggs BG (1975) On the Proteaceae—the evolution and classification of a southern family. Bot J Linn Soc 70:83–182
5. Martin PG, Dowd JM (1984) The study of plant phylogeny using amino acid sequences of ribulose-1,5-bisphosphate carboxylase. IV Proteaceae and Fagaceae and the rate of evolution of the small subunit. Aust J Bot 32: 291–299
6. Martin PG, Dowd JM (1988) A molecular evolutionary clock for angiosperms. Taxon 37:364–377
7. Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michael HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim KJ, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang QY, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann Missouri Bot Gard 80:528–580
8. Drinnan AN, Crane PR, Hoot SB (1994) Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). Pl Syst Evol 8(Suppl):93–122
9. Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress WJ, Nixon KC, Farris JS (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Bot J Linn Soc 133:381–461
10. Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu Y-L, Hilu KW, Davis CD, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS (2011) Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot 98:704–730
11. Carpenter RJ, Hill RS, Jordan GJ (2005) Leaf cuticular morphology links Platanaceae and Proteaceae. Int J Plant Sci 166:843–855
12. Hoot SB, Douglas AW (1998) Phylogeny of the Proteaceae based on *atpB* and *atpB–rbcL* intergenic spacer region sequences. Aust Syst Bot 11
13. Weston PH, Barker NP (2006) A new suprageneric classification of the Proteaceae, with an annotated checklist of genera. Telopea 11:314–344
14. Barker NP, Weston PH, Rourke JP, Reeves G (2002) The relationships of the southern African Proteaceae as elucidated by internal transcribed spacer (ITS) DNA sequence data. Kew Bull 57:867–883
15. Barker NP, Weston PH, Rutschmann F, Sauquet H (2007) Molecular dating of the "Gondwanan" plant family Proteaceae is only partially congruent with the timing of Gondwanan break-up. J Biogeogr 34: 2012–2027
16. Weston PH, Crisp MD (1996) Trans-Pacific biogeographic patterns in the Proteaceae. In: Keast A, Miller SE (eds) The origin and evolution of Pacific Island Biotas, New Guinea to Eastern Polynesia: patterns and processes. SPB Academic Publishing, Amsterdam, pp 215–232
17. Mast AR, Willis CL, Jones EH, Downs KM, Weston PH (2008) A smaller Macadamia from a more vagile tribe: inference of phylogenetic relationships and divergence times in Macadamia and relatives (tribe Macadamieae; Proteaceae). Am J Bot 95:843–870
18. Sauquet H, Weston PH, Anderson CL, Barker NP, Cantrill DJ, Mast AR, Savolainen V (2009) Contrasted patterns of hyperdiversification in Mediterranean hotspots. Proc Natl Acad Sci U S A 106:221–225
19. Sauquet H, Weston PH, Barker NP, Anderson CL, Cantrill DJ, Savolainen V (2009) Using fossils and molecular data to reveal the origins of the Cape proteas (subfamily Proteoideae). Mol Phylogenet Evol 51:31–43
20. Jordan GJ, Dillon RA, Weston PH (2005) Solar radiation as a factor in the evolution of scleromorphic leaf anatomy in Proteaceae. Am J Bot 92:789–796
21. Jordan GJ, Weston PH, Carpenter RJ, Dillon RA, Brodribb TJ (2008) The evolutionary relations of sunken, covered, and encrypted stomata to dry habitats in Proteaceae. Am J Bot 95:521–530
22. Crisp MD, Arroyo MTK, Cook LG, Gandolfo MA, Jordan GJ, McGlone MS, Weston PH, Westoby M, Wilf P, Linder HP (2009) Phylogenetic habitat conservatism on a global scale. Nature (Lond) 458:754–756

23. Mast AR (1998) Molecular systematics of subtribe Banksiinae (*Banksia* and *Dryandra*; Proteaceae) based on cpDNA and nrDNA sequence data: implications for taxonomy and biogeography. Aust Syst Bot 11:321–342
24. Mast AR, Givnish TJ (2002) Historical biogeography and the origin of stomatal distributions in *Banksia* and *Dryandra* (Proteaceae) based on their cpDNA phylogeny. Am J Bot 89:1311–1323
25. Mast AR, Jones EH, Havery SP (2005) An assessment of old and new DNA sequence evidence for the paraphyly of Banksia with respect to Dryandra (Proteaceae). Aust Syst Bot 18:75–88
26. George AS (1981) The genus Banksia L.f. (Proteaceae). Nuytsia 3:239–473
27. George AS (1999) Banksia. In: Wilson A (ed) Flora of Australia, vol 17B, Proteaceae 3 *Hakea* to *Dryandra*. ABRS/CSIRO, Melbourne, pp 175–251
28. Thiele K, Ladiges PY (1996) A cladistic analysis of *Banksia* (Proteaceae). Aust Syst Bot 9:661–733
29. Mast AR, Thiele K (2007) The transfer of *Dryandra* R.Br. to *Banksia* L.f. (Proteaceae). Aust Syst Bot 20:63–71
30. Cavanagh T (2007) So *Dryandra* becomes *Banksia*—what's all the fuss about? Aust Plants (http://anpsa.org.au/APOL2007/aug07-1.html)
31. George AS (2008) You don't have to call *Dryandra Banksia*. Aust Plants (http://anpsa.org.au/APOL2008/sep08-2.html)
32. Crisp MD, Cook LG (2007) A congruent molecular signature of vicariance across multiple plant lineages. Mol Phylogenet Evol 43:1106–1117
33. He T, Lamont BB, Downes KS (2011) *Banksia* born to burn. New Phytol 191:184–196
34. Barker NP, Vanderpoorten A, Morton CM, Rourke JP (2004) Phylogeny, biogeography, and the evolution of life-history traits in *Leucadendron* (Proteaceae). Mol Phylogenet Evol 33:845–860
35. Barraclough TG, Reeves G (2005) The causes of speciation in flowering plant lineages: species-level DNA trees in the African genus *Protea*. In: Bakker FT, Chatrou L, Gravendeel B, Pelser P (eds) Plant species-level systematics: patterns, processes and new applications. Koeltz, Königstein, pp 31–46
36. Valente LM, Reeves G, Schnitzler J, Mason IP, Fay MF, Rebelo TG, Chase MW, Barraclough TG (2009) Diversification of the African genus *Protea* (Proteaceae) in the Cape biodiversity hotspot and beyond: equal rates in different biomes. Evolution 64:745–760
37. Mast AR, Milton EF, Jones EH, Barker RM, Barker WR, Weston PH (2012) Time-calibrated phylogeny of the woody Australian genus *Hakea* (Proteaceae) supports multiple origins of insect-pollination among bird-pollinated ancestors. Am J Bot 99:472–487
38. Williams IJM (1972) A revision of the genus *Leucadendron* (Proteaceae). Contrib Bol Herb 3:1–425
39. Rebelo AG (2001) A field guide to the proteas of southern Africa. Fernwood, South Africa
40. Barker WR, Barker RM, Haegi L (1999) Introduction to *Hakea*. In: Wilson A (ed) Flora of Australia, vol 17B, Proteaceae 3, Hakea to Dryandra. ABRS/CSIRO, Canberra, Australia, pp 1–30
41. Hanley ME, Lamont BB, Armbruster WS (2009) Pollination and plant defence traits covary in Western Australian *Hakeas*. New Phytol 182:251–260
42. Byrne M, Macdonald B, Coates D (1999) Divergence in the chloroplast genome and nuclear rDNA of the rare Western Australian plant *Lambertia orbifolia* Gardner (Proteaceae). Mol Ecol 8:1789–1796
43. Krauss SL (2000) Patterns of mating in *Persoonia mollis* (Proteaceae) revealed by an analysis of paternity using AFLP: implications for conservation. Aust J Bot 48:349–356
44. Rymer PD, Ayre DJ (2006) Does genetic variation and gene flow vary with rarity in obligate seeding *Persoonia* species (Proteaceae)? Conserv Genet 7:919–930
45. Holmes GD, James EA, Hoffmann AA (2008) Divergent levels of genetic variation and ploidy among small populations of the rare shrub, *Grevillea repens*. Conserv Genet 10:827–837. doi:10.1007/S10592-008-9643-9
46. Coates DJ, Sokolowski RES (1992) The mating system and patterns of genetic variation in *Banksia cuneata* A.S.George (Proteaceae). Heredity 69:11–20
47. Krauss SL (1994) Restricted gene flow within the morphologically complex species *Persoonia mollis* (Proteaceae): contrasting evidence from the mating system and pollen dispersal. Heredity 73:142–154. doi:10.1038/hdy.1994.113
48. Coates DJ, Hamley VL (1999) Genetic divergence and the mating system in the endangered and geographically restricted species, *Lambertia orbifolia* Gardner (Proteaceae). Heredity 83:418–427
49. Millar MA, Byrne M, Coates DJ (2010) The maintenance of disparate levels of clonality, genetic diversity and genetic differentiation in disjunct subspecies of the rare *Banksia ionthocarpa*. Mol Ecol 19:4217–4227
50. Llorens TM, Byrne M, Yates CJ, Nistelberger HM, Coates DJ (2012) Evaluating the influence of different aspects of habitat fragmentation on mating patterns and pollen dispersal in the bird-pollinated *Banksia sphaerocarpa* var. *caesia*. Mol Ecol 21:314–328

51. Lamont BB, He T, Enright NJ, Krauss SL, Miller BP (2003) Anthropogenic disturbance promotes hybridization between *Banksia* species by altering their biology. J Evol Biol 16:551–557
52. Krauss SL (1997) Low genetic diversity in *Persoonia mollis* (Proteaceae), a fire-sensitive shrub occurring in a fire-prone habitat. Heredity 78:41–49. doi:10.1038/hdy.1997.5
53. Krauss SL (1998) A phylogeographic analysis of allozyme variation amongst populations of *Persoonia mollis* (Proteaceae). Aust J Bot 46:571–582
54. Rossetto M, Thurlby KAG, Offord CA, Allen CB, Weston PH (2011) The impact of distance and a shifting temperature gradient on genetic connectivity across a heterogeneous landscape. BMC Evol Biol 11:126
55. Prunier R, Holsinger KE (2010) Was it an explosion? Using population genetics to explore the dynamics of a recent radiation within *Protea* (Proteaceae L.). Mol Ecol 19:3968–3980
56. Felsenstein J (1981) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution 35:1229–1242
57. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.
58. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
59. Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160:1217–1229
60. McPherson H, van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M (2013) Capturing chloroplast variation for molecular ecology studies: a simple Next Generation Sequencing approach applied to a rainforest tree. Mol Ecol 13:8
61. Knowles LL (2009) Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst Biol 58:463–467
62. Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a comparison of methods. Syst Biol 60: 126–137
63. Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA (2009) Species trees from highly incongruent gene trees in rice. Syst Biol 58: 489–500

INDEX

A

Adaptor ligation and preamplification
  AFLP 180–181
  microsatellite library 178, 180–182
Allopolyploid 142–144, 325
Amplified fragment length polymorphism (AFLP) 23, 42–44, 48, 72, 76, 77, 82, 192–194, 211–231, 235, 388, 390
Arbitrary primed polymerase chain reaction (AP-PCR) 194, 197, 203, 206

B

Barcode of Life Data Systems 100
BARE1 element 235–237, 244

C

CAPS. *See* Cleaved amplified polymorphic sequences (CAPS)
Cereba (LTR primer design) 243
Character matrix 258–260
Chloroplast DNA (cpDNA) 42, 43, 45, 54, 85–116, 122, 126–128, 375
Chromosome 2, 44, 50, 86, 120, 125, 143, 161–163, 212, 233, 282–284, 294, 305, 309, 310, 312, 313, 316, 321, 325–328, 330–332, 334, 335
Chromosome banding
  DAPI banding 288
  fluorochrome banding 309, 310
  Giemsa C-banding 309–311
  Hoechst banding 317
Cleaved amplified polymorphic sequences (CAPS) 194, 195, 199, 205–206
Cloning
  of PCR fragment
    from retrotransposon 239, 244
    from SCAR 193, 194, 196–197
  in pGEM-T
    microsatellite library 179–188
Clusia multiflora (genome size analysis) 298
Coding DNA 44
Codominant 48, 49, 177, 191, 193, 195, 213
Complete chloroplast sequences 95, 105, 112, 115
Complete mitochondrial sequences 48
Conserved primers (= universal primers) 42, 90, 105, 112, 122, 126, 128, 130–131, 135, 142, 238
Consortium for the Barcode of Life (CBOL) 14, 87, 145
Convergence 25, 45, 380
Copia element 236, 238
Crepis sp. (chromosome banding) 310, 311, 319
Cryptic 22, 24, 27
Cypripedium sp. (genome size analysis) 282–283
Cytogenetics 26, 144, 309–322, 326, 330

D

Dactylorhiza sp. (genome size analysis) 299, 301
DAMD. *See* Directed amplification of minisatellite-region DNA (DAMD)
DArT. *See* Diversity array technology (DArT)
Databases
  for chloroplast plant DNA primers (http://bfw.ac.at/rz/bfwcms.web?dok=4977) 87, 91, 130
  for plant DNA barcoding
    Barcode of Life Data Systems (http://www.boldsystems.org/views/login.php) 100
  for plant DNA sequences
    DNA Database of Japan, DDBJ 263
    EMBL Nucleotide Sequence Database, EMBL-Bank 263
    GenBank (www.ncbi.nlm.nih.gov/genbank/ http://www.ncbi.nlm.nih.gov/genbank/) 91
  for plant genome sizes
    FLOWer database 281
    Plant DNA C-values database 281
Denaturing high-performance liquid chromatography (dHPLC) 86, 88–90, 93–94, 101–104, 108, 109, 114, 116
Destaining slides
  after FISH 321
  after fluorochrome bandings 318
Digestion (= restriction enzymes) 55, 58, 74, 80, 85, 86, 92, 97, 98, 105, 110, 128, 131, 194, 195, 199, 212, 218–219, 225, 235, 236, 241, 242, 247–249, 253, 321, 329, 330, 334
Directed amplification of minisatellite-region DNA (DAMD) 212
Diversity array technology (DArT) 43

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 1115,
DOI 10.1007/978-1-62703-767-9, © Springer Science+Business Media New York 2014

DNA amplification fingerprinting (DAF) 194–195, 197–198, 203–204, 206
DNA barcoding 14, 20–21, 87, 100, 144
DNA extraction
algae (DNA extraction from herbarium specimen) 71, 74, 79
cetyl trimethylammonium bromide (CTAB) extraction 70, 73, 74, 77–79, 81, 83
DNA concentration (= DNA quantification) 58, 64, 107, 110, 131–132, 200, 207, 214, 249, 252
DNA purity (contaminants, impurities) 55, 57, 58, 64, 251
DNeasy Plant Mini Kit (QIagen) 70, 72, 75–77
herbarium specimens (DNA extraction from) 69–83
isolation of plant nuclei (for FCM) 280–282, 284, 285, 288–294, 297, 302
leaf (leaves) sampling for DNA extraction 53–65
lichens (DNA extraction from herbarium specimen) 70, 71, 73–74, 77–79
mosses (DNA extraction from herbarium specimen) 71
mucilaginous tissues (DNA extraction from herbarium specimen rich in) 72, 73, 77–78
mushrooms (DNA extraction from herbarium specimen) 71, 74, 80
Nucleo Spin Plant II kit (Macherey-Nagel) 74, 81
phenolic compounds (DNA extraction from herbarium specimen rich in) 58, 72
plant bulking or DNA pooling for DNA extraction 55
polysaccharides compounds (DNA extraction from herbarium specimen rich in) 57, 71–73, 77, 107, 246
seeds (DNA extraction from herbarium specimen) 72, 73, 76–77
vascular plants/conifers (DNA extraction from herbarium specimen) 70, 71, 76
DNA polymerase 91, 96, 108, 145, 146, 157, 168, 178, 181, 184, 196, 200, 206, 207, 215, 216, 219–221, 225, 226, 228, 240, 241, 247, 248, 251, 332
DNA sequencing
Illumina GoldenGate sequencing 155–158, 163–170
Illumina Infinium HD sequencing 158–159, 170–173
next-generation sequencing (NGS) 24, 25, 90, 94, 105–107, 116, 360
Sanger dideoxy nucleotide sequencing 92
Dombeyoideae (taxonomic revision) 337–360
Dominant 48, 49, 193–195, 204, 213, 302

E

Enzymatic digestion and glass-fiber filtration (EDGF) 74, 80
Erianthus sp. (GISH analysis) 144, 326, 328
Evolutionary rate 44, 45, 47, 125, 126, 144
Extraction of PCR product from gel (SCAR) 193, 194, 196–197, 202, 203, 208

F

FASTA format 95, 161, 266–270
Flow cytometry (FCM) 279–305
Fluorescent *in situ* hybridization (FISH) 144, 309–322, 325

G

Gap extension penalty (GEP) 267, 268
Gap open penalty (GOP) 267, 268
Gel electrophoresis
agarose 58, 59, 91, 96–98, 132, 180, 215, 236, 247, 251
polyacrylamide (PAGE) 89, 92–94, 97, 100–101, 105, 204
Gel staining
ethidium bromide (EB) staining 63, 101, 146, 147, 205
GelRed™ staining 146–148
silver staining 93, 113, 114, 204, 206, 217–218, 224–225
SYBR Gold staining 113, 114
SYBR green staining 113, 114, 241, 242, 251
Gene 23, 24, 29, 30, 44, 47, 85–87, 93, 95, 112, 121–122, 135, 141–144, 154, 155, 161–163, 173, 187, 191, 193, 197, 236, 257, 265, 310, 339, 341, 348–350, 357, 367, 373, 374, 387, 390, 392
Genetic distance 259, 390
Genome size
2C-value 282
1Cx-value 282
holoploid 1C-value 282
monoploid 1C-value 282
standard references for 1C-DNA content 285
Genomic *in situ* hybridization (GISH) 325–326
Gentianella (alignment of chloroplast DNA sequences) 115, 116
Glass-fibre filtration (GF) 74, 80
Guzmania monostachia (genome size analysis) 293
Gypsy element 238

H

Herbarium 5, 7, 8, 20, 21, 25, 27, 69–83, 284, 303
Heteroduplex DNA 86, 89–90, 94, 102
Homoplasy 24–25, 42, 45, 47, 112, 347–348, 351, 410
Hordeum spontaneum (IRAP analysis) 237
Hybrid 48, 49, 143, 144, 192–193, 206, 327–328, 390
Hybridization 22–24, 28, 48, 50, 95, 135, 142, 144, 155–159, 178, 180, 184, 193, 198, 199, 205, 321, 322, 326, 331–333, 335, 336, 389–390, 392

I

Inheritance 48, 126, 191–192, 194, 213
In silico SNP discovery 153–154, 156, 157
Integrative taxonomy 25–28

Internal transcribed spacer (ITS)..... 39, 45, 141–148, 310, 357, 374–375, 378, 387–388
Inter-primer binding site polymorphism (iPBS).....233–253
Inter-retrotransposons amplified polymorphism (IRAP).....41, 233–253
Inter-simple sequence repeats (ISSR)..... 40, 44, 48, 211–231, 237, 243–244
Introgression.....22, 50, 193, 325, 389–390, 392
Intron 44, 86, 87, 111, 112, 115, 122, 128, 357, 375

K

Kalanchoe marnieriana (genome size analysis) 298–299, 304

L

Long interspersed nuclear element (LINE) 234, 238
Long terminal repeats (LTR) 40–41, 43, 234–239, 242–244, 248, 249, 252
Low-copy nuclear genes (LCNG).....42

M

Malvaceae (taxonomic revision).....337–360
Maximum likelihood (ML)..... 49, 259–262, 269, 271, 273, 275, 390
Maximum parsimony (MP)..... 49, 241, 260, 378
Microsatellite (SSR).....24, 25, 39–44, 47–49, 55, 57, 72, 74, 81, 90, 105, 112, 115, 128, 135, 177–181, 183–188, 192–195, 198, 199, 205, 212, 213, 216, 221, 233–235, 237, 243–244, 390
Microsatellite enriched library.....177–188
Minisatellite 39–40, 44, 90, 94, 96, 105, 115, 128
Mitochondrial DNA (mtDNA)..... 43, 45, 106, 111, 121–123, 125–128, 135
ML. *See* Maximum likelihood (ML)
Morphological.....2, 258, 281, 335, 338, 366
MP. *See* Maximum parsimony (MP)
mtDNA. *See* Mitochondrial DNA (mtDNA)

N

Neighbor joining (NJ) 49, 202, 259
Newick format.....271
Non-coding DNA.....39–40, 44, 112
Nuclear DNA.....39–42, 106, 107, 211, 289–294, 299, 305, 388
Nuclear ribosomal ITS.....39, 141–148
Nucleolar organizer regions (NORs).....142, 310

O

Organelle (organellar) DNA 24, 25, 43, 45, 54, 106
Outlier sequence.....261, 266, 267

P

PCR-RFLP (Restriction Fragment Length Polymorphism) (of chloroplast DNA) 89, 92, 94, 96–98, 116, 128, 195
PDR1 retrotransposon.....235
Phylogenetic tree 12, 49, 50, 96, 257–263, 265–267, 269, 271, 273, 275, 367, 372, 380, 393
Phylogeny.....12, 14, 43, 48, 49, 87, 125–126, 141, 142, 257, 258, 263, 271–273, 275, 338, 339, 343, 344, 347–348, 357, 359, 366, 367, 373–374, 385–387, 394
Phylogeography.....90, 125–126
Physaria sp. (genome size analysis) 283–284
Pinus nigra (chromosome banding and FISH analysis) 312–313
Piper nigrum (RAPD amplification and dendrogram) 201, 202
Plant DNA barcode.....144–145
Ploidy 28, 279–305
Plum (AFLP and ISSR amplification).....225
Poaceae (ribosomal ITS alignment).....142, 143
Polymerase chain reaction (PCR)..... 14, 40, 53, 71, 86, 128, 141, 155, 174, 191, 212, 233, 263, 373
Polyploidy..... 50, 282, 283
Positive clone screening using PCR (microsatellite library).....180, 185–186
Preparation of protoplasts.....316
Pretreatment and fixation of root-tips.....314, 316, 330
Primers
  arbitrary primers (RAPD) 193–195, 206, 207
  PBS 18-mers primers 244–246
  primers for chloroplast DNA.....87, 94–96
  primers for microsatellite (for REMAP) 44, 237, 243–244
  retrotransposon LTR primers 40–43, 234–236, 238, 239, 242–244, 248, 252
  universal plant ITS primers 145, 146
Proteaceae (taxonomic revision) 365
Pseudogene.....135, 136, 144
Purification of PCR product before sequencing (mitochondrial DNA).....131, 186
Purification of the preamplification product (microsatellite library).....179, 182–183

R

Random amplified hybridization microsatellites (RAHM) 194, 195, 199, 205
Random amplified microsatellite polymorphism (RAMPO).....194, 195, 198, 205, 208
Random amplified polymorphic DNA (RAPD).....23–24, 42–44, 48, 71, 191–208, 212, 213

Repeated DNA sequences 39–42, 115, 116, 122, 154, 212
Reproducibility 152, 177, 191–194, 207, 212
Restriction enzymes (digestion) 97, 105, 110, 194, 195, 199, 212, 236, 241, 242, 248
Retrotransposon 40, 234,
Retrotransposon-microsatellite amplification polymorphism (REMAP)..... 43, 44, 233–253
Reversion 10, 24–25, 27, 40, 45, 89, 93, 95, 101–102, 108, 130, 132, 133, 137, 145, 146, 148, 155, 198, 204, 206, 234, 238, 366, 383–384, 394
Ribosomal DNA heterogeneity 42
Ribosomal RNA (rDNA)..... 39, 141–148, 310, 392
Rooted phylogenetic tree 266

S

Selective amplification of microsatellite polymorphic loci (SAMPL)..... 42, 44, 211–231
Sequence characterized amplified regions (SCAR).....193, 194, 196–197, 202–203, 208
Sequence-related amplified polymorphism (SRAP) 194, 195, 198, 204, 206
Sequence-specific amplified polymorphism (SSAP)..... 43, 234–236, 239, 253
Short interspersed nuclear element (SINE)..... 40–42, 234, 238, 253
Simple sequence repeats (SSR) (microsatellites).....39, 40, 90, 105, 152, 177, 192, 199, 205, 212, 215–216, 220–221
Single nucleotide polymorphism (SNP)..... 42, 48, 49, 89, 105, 111, 115, 152–157, 160–163, 166, 173, 192
Single primer amplification reaction (SPAR).....212
Softwares
  for analysing and aligning sequence data.....49, 256
    BIOEDIT 99–111
    MEGA5 (ClustalW, Muscle).....267, 268
  for analysing gel images
    Gene Profiler 93
  for Bayesian phylogenetic inference
    BEAST.....275, 376, 380, 381, 383
    MrBayes3 275
  for designing primers
    Amplicon.....96, 102, 103, 194, 195
    Greene SCPrimer 96
    iCODEHOP.....96
    Primer 3 ..... 187, 214, 220
  for determining melting temperature of a double strand DNA
    Melt 103, 130, 247
  for determining the best-fit nucleotide substitution model
    jModelTest2.....269–271
  for DNA homology search
    BLAST.....194, 258–264, 267, 268, 326
  for Illumina GoldenGate Genotyping Assay
    GenomeStudio 157–158, 163–170
  for phylogenetic tree construction with maximum likelihood
    PhyML3 ..... 261–262, 390
  for phylogenetic tree construction with UPGMA/NJ
    MEGA5 ..... 262, 267, 268
    NTSYS-PC 2.01 ..... 201, 202
  for SNP analysis
    autoSNP 154, 155, 157, 160, 163
    AutoSNPdb 154, 155, 157, 160, 163
    SNPServer 154
  for SNP analysis in wheat
    wheatgenome.info..... 157, 161–163
  for T_m calculation and corresponding instructions (http://primerdigital.com/tools/) 252
  for visualization and editing of phylogenetic trees
    FigTree 262, 273–274
Sorbus sp. (genome size analysis)..... 22, 294, 295
Spacer 39, 86, 87, 90, 92, 100, 101, 111, 112, 141, 142, 172, 217, 222, 224, 229, 230, 239, 310, 374, 375
Species concept..... 2, 15, 23, 28–30
Streptophyta sp. (mt DNA cladogram).....126, 127
Sugarcane (*Saccharum* sp.) (GISH analysis) 326–328, 331
Sukkula (IRAP analysis).....244, 249

T

Tandem repeats 39–41, 44, 105, 115, 128, 141, 142
Taxonomy..... 1, 39, 85, 141, 152, 201, 265, 326, 342, 375
Transposable element (TEs).....40, 41, 45, 233, 234, 253
Transposon 40, 233–253
Type I transposable element 40–41
Type II transposable element.....40

U

Unrooted phylogenetic tree 257–258, 266
Unweighted pair grouping with arithmetic mean (UPGMA) 49, 202, 259, 390

V

Vanilla sp. (*rbc*L phylogenetic analysis)..... 262–265, 267, 273, 274

W

Wheat (searching for SNPs in) 154, 157, 160–163