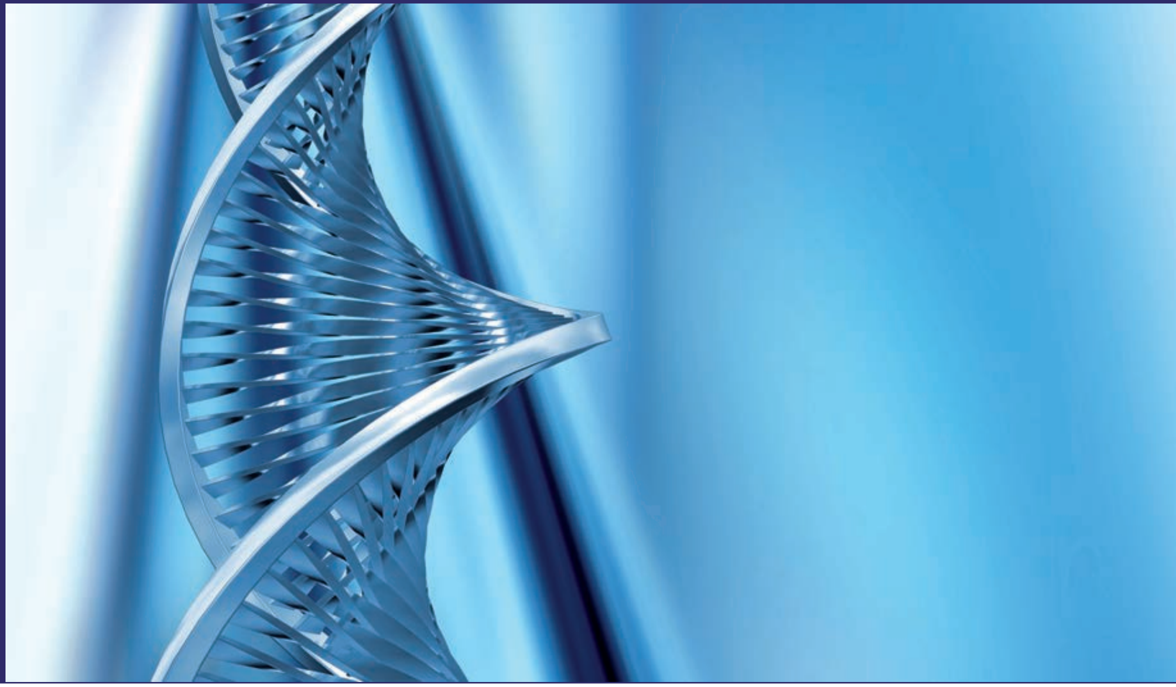
Taxonomy, the science of naming and classifying organisms is the original bioinformatics and a basis for all biology, is fundamentally important in ensuring the quality of life of future human generation on the earth; yet over the past few decades, the teaching interest and research funding in taxonomy have declined. Now taxonomy suddenly became fashionable again due to revolutionary approaches in taxonomy called DNA barcoding. The plant DNA barcoding is now transitioning the epitome of species identification; and thus, ultimately helping in the molecularization of taxonomy, a need of the hour. The 'DNA barcodes' show promise in providing a practical, standardized, species-level identification tool that can be used for biodiversity assessment, life history and ecological studies, and forensic analysis. The most significant scientific information available on plant DNA barcoding technologies have been collected and arranged in order of the different marker systems which describes introductory and application review on plant DNA barcoding and phylogenetics under 18 chapters.

• M. AJMAL ALI, Ph.D. in 2006 from TMBU India, PDF from KRIBB, S. Korea; has published 4 books, 50 papers; supervised 4 Ph.D.; Co-I of KSU funded projects. • G. GYULAI: Ph.D. in 1983, D.Sc. in 2010 from SZIE, Hungary; MTA, FULBRIGHT, DAAD and OECD Fellowship. • F. AL-HEMAID: Ph.D. in 1991 from ABDN, Scotland; has published 45 papers in Systematics.

978-3-659-28095-5

Plant DNA Barcoding and Phylogenetics

Ali, Gyulai, Al-Hemaid



M. Ajmal Ali
G. Gyulai
F. Al-Hemaid

Plant DNA Barcoding and Phylogenetics

LAP LAMBERT Academic Publishing

M. Ajmal Ali
G. Gyulai
F. Al-Hemaid

Plant DNA Barcoding and Phylogenetics

M. Ajmal Ali
G. Gyulai
F. Al-Hemaid

Plant DNA Barcoding and Phylogenetics

LAP LAMBERT Academic Publishing

Impressum / Imprint
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
OmniScriptum GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Printed at: see last page
ISBN: 978-3-659-28095-5

PLANT DNA BARCODING AND PHYLOGENETICS

Editors

M. Ajmal Ali
Department of Botany and Microbiology
College of Science, King Saud University
Riyadh-11451, Saudi Arabia

Gyulai Gábor
Institute of Genetics and Biotechnology
St. István University Gödöllő 2103, Hungary

Fahad Mohammed Abdullrahman Al-Hemaid
Department of Botany and Microbiology
College of Science, King Saud University
Riyadh-11451, Saudi Arabia

Dedicated

to

Sabiha

PREFACE

Taxonomy (-the science of naming and classifying organisms is the original bioinformatics and a basis for all biology) is fundamentally important in ensuring the quality of life of future human generation on the earth; yet over the past few decades, the teaching interest and research funding in taxonomy have declined because of its classical way of practice which lead the discipline many a times to a subject of opinion, and this ultimately gave birth of several taxonomic problems and challenges; therefore, the taxonomists became an endangered race in the era of genomics. Now taxonomy suddenly became fashionable again due to revolutionary approaches in taxonomy called DNA barcoding (-a novel technology to provide rapid, accurate and automated species identifications using short orthologous DNA sequences). In DNA barcoding, complete data set can be obtained from single specimens irrespective to morphological or life stage characters. The core idea of DNA barcoding is based on the fact that the highly conserved stretches of DNA, either coding or noncoding regions vary at very minor degree during the evolution within the species. The sequences suggested to be useful in DNA barcoding include cytoplasmic mitochondrial DNA (e.g. cox1) and chloroplast DNA *(*e.g. rbcL*,* trnL-F*,* matK, ndhF *and* atpB-rbcL*), and* nuclear DNA (ITS and house keeping genes e.g. gapdh)*.* The plant DNA barcoding is now transitioning the epitome of species identification; and thus, ultimately helping in the molecularization of taxonomy, a need of the hour. The 'DNA barcodes' show promise in providing a practical, standardized, species-level identification tool that can be used for biodiversity assessment, life history and ecological studies, forensic analysis and many more. The purpose of this edited book is to collect the most significant scientific information available on PCR-based plant DNA barcoding technologies. The chapters are arranged in order of the different marker systems. The chapters included are from all fields i.e. from DNA extraction to the analysis with molecular probes. This book describes introductive and application review chapters on plant DNA barcoding and phylogenetics under 18 different chapters [which includes (1) Introduction to Plant DNA Barcoding, (2) Molecular Markers for Plant DNA barcoding, (3) Nuclear Sequences in Plant Phylogenetics, (4) Nuclear and Organelle Specific PCR Markers, (5) Maturase K Gene in Plant DNA Barcoding and Phylogenetics, (6) Retrotransposon-based Plant DNA Barcoding, (7) Isoenzymes as Molecular Markers, (8) Applications of Plant DNA Barcoding, (9) Thermocycling in Systematics, (10) DNA Sequencing, (11) Plant DNA Barcoding and Molecular Phylogeny, (12) Plant DNA Barcoding

Methodology: DNA Extraction - Sequencing, (13) In Silico Approach for Phylogenetic Analysis, (14) Life History Barcoding of *Daucus carota*, (15) Cloning and Microsatellite Barcoding of Black Locust, (16) Genetic Diversity Assessment of *Caralluma adscendens*, (17) Barcoding of Transgenes in GM Plants, and (18) Molecular Barcoding of Sex-Linked DNA Markers of Dioecious Plants] which are of great importance to put the essentials of this discipline and protocols into a broader perspective. We hope that the book will be useful to the students of higher studies, botanist, conservation biologist and other those interested in plant DNA barcoding and phylogenetics.

M. Ajmal Ali
Gyulai Gábor
F.M.A. Al-Hemaid

CONTENTS

*****

CONTRIBUTORS

A. Alam: *(Absar Alam) Regional Centre, Central Inland Fisheries Research Institute, 24 Panna Lal Road Allahabad- 211002, Uttar Pradesh, India*

A. Arshi: *(Anfal Arshi) Defence Research & Development Organisation, Defence Institute of Bio-Energy Research, Haldwani-263139, Uttarakhand, India*

A. Bahieldin: *(Ahmed Bahieldin) King Abdulaziz University, Faculty of Science, Department of Biological Sciences, Genomics and Biotechnology Section, Jeddah- 21589, Saudi Arabia; Genetics Department, Faculty of Agriculture, Ain Shams University, Cairo-11241, Egypt*

A. Bhattacharjee: *(Atanu Bhattacharjee) Department of Biotechnology and Bioinformatics, North Eastern Hill University, Shillong-793022, Meghalaya, India*

A. Bittsánszky: *(András Bittsánszky) Hungarian Academy of Sciences, CAR, Plant Protection Institute, Budapest, Hungary 1525*

A. Goyal: *(Arvind Goyal) Department of Botany, University of North Bengal, Siliguri- 734013, West Bengal, India*

A. Kerekes: *(Adrien Kerekes) Institute of Genetics and Biotechnology, Faculty of Agricultural and Environmental Sciences, St István University, Gödöllő 2103, Hungary*

Á. Mendel: *(Ákos Mendel) Institute of Genetics and Biotechnology, Faculty of Agricultural and Environmental Sciences, St István University, Gödöllő 2103, Hungary*

A. Sen: *(Arnab Sen) Department of Botany, University of North Bengal, Siliguri-734013, West Bengal, India*

A.K. Pandey: *(Arun Kumar Pandey) Department of Botany, University of Delhi, Delhi-110007, India*

A.M. Alzohairy: *(Ahmed Mansour Alzohairy) Genetics Department, Faculty of Agriculture, Zagazig University, Zagazig 44511, Egypt*

A.S. Alhomida: *(Abdullah Saleh Alhomida) Department of Biochemistry, College of Science, King Saud University, P.O. Box 2455, Riyadh-11451, Saudi Arabia*

B. Kerti: *(Balázs Kerti) Institute of Genetics and Biotechnology, St. István University, Gödöllő, H-2103, Hungary; Institute of Genetics and Plant Breeding, Corvinus University, Budapest 1118, Hungary*

B. Pandit: *(Bhagirath Pandit) Department of Botany, BSK College (SKMU Dumka), Barharwa-816101, Sahebganj, Jharkhand, India*

C. Lee: *(Changyoung Lee) International Biological Material Research Center, Korea Research Institute of Bioscience and Biotechnology, 111 Gwahangno, Yuseong-gu, Daejeon-305 806, South Korea*

F.M.A. Al-Hemaid: *(Fahad Mohammed Abdullrahman Al-Hemaid) Department of Botany and Microbiology, College of Science, King Saud University, Riyadh-11451, Saudi Arabia*

G. Gyulai: *(Gyulai Gábor) Institute of Genetics and Biotechnology, St. István University, Gödöllő 2103, Hungary*

G. Jahnke: *(Gizella Jahnke) NARIC Research Institute for Viticulture and Enology, H-8261 Badacsonytomaj, Római út 181, Hungary*

H. Elsawy: *(Hany Elsawy) Chemistry Department, Faculty of Science, Tanta University, Tanta, Egypt*

H. Ohm: *(Herb Ohm) Department of Agronomy, Purdue University, 915 West State Street, West Lafayette, IN 47907-2054, USA*

H. Rennenberg: *(Heinz Rennenberg) Albert-Ludwigs-Universitat, Inst. Forstbotanik und Baumphysiologie, Freiburg, 79085, Germany*

H.A. Khan: *(Haseeb Ahmad Khan) Department of Biochemistry, College of Science, King Saud University, Riyadh-11451, Saudi Arabia*

I. Bock: *(István Bock) Agricultural Biotechnology Center, Szent Györgyi A 1, 2100 Gödöllő, Hungary*

J. Lee: *(Joongku Lee) International Biological Material Research Center, Korea Research Institute of Bioscience and Biotechnology, 111 Gwahangno, Yuseong-gu, Daejeon-305 806, South Korea*

J.S.M. Sabir: *(Jamal S.M. Sabir) King Abdulaziz University, Faculty of Science, Department of Biological Sciences, Genomics and Biotechnology Section, Jeddah- 21589, Saudi Arabia*

K. Tóth-Lencsés: *(Kitti Tóth-Lencsés) Institute of Genetics and Biotechnology, Faculty of Agricultural and Environmental Sciences, St István University, Gödöllő 2103, Hungary*

L. Kovács: *(László Kovács) Institute of Genetics and Biotechnology, Faculty of Agricultural and Environmental Sciences, St István University, Gödöllő 2103, Hungary*

L. Waters Jr: *Department of Horticulture, College of Agriculture, Auburn University, Alabama 36849 USA*

L.Y. Murenets: *(Lilja Y. Murenets) Bioorganic Chemistry, Russian Academy of Science, 6 Science Avenue, 142290 Pushchino, Moscow region, Russia*

M. Czakó: *(Mihály Czakó) University of South Carolina, Department of Biological Sciences, Columbia, SC 29208, USA*

M. Rusop: *(Mohamad Rusop) NANO-SciTech Centre (NST), Institute of Science, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia*

***M.A. Ali**: (Mohammad Ajmal Ali) Department of Botany and Microbiology, College of Science, King Saud University, Riyadh- 11451, Saudi Arabia*

***M.M. Mostafa**: (Mahjou M. Mostafa) Genetics and Genetic Engineering Department, Faculty of Agriculture, Benha University, Qalubia, Egypt*

***M.O. Rahman**: (M. Oliur Rahman) Department of Botany, University of Dhaka, Dhaka 1000, Bangladesh*

***M.R. Enan**: (Mohamed Rizk Enan) United Arab Emirates University, College of Science, Biology Department, Al Ain 15551, United Arab Emirates*

***M.R. Naik**: (Manjanaik Ramachandra Naik) Department of P.G. Studies and Research in Applied Botany, Kuvempu University, Jnana Sahyadri, Shankaraghatta- 577451, Karnataka, India*

***M.S. Ola**: (Mohammad Shamsul Ola) Department of Biochemistry, College of Science, King Saud University, P.O. Box 2455, Riyadh- 11451, Saudi Arabia*

***O. Toldi**: (Ottó Toldi) Institute of Genetics and Biotechnology, Faculty of Agricultural and Environmental Sciences, St István University, Gödöllő 2103, Hungary*

***P. Kar**: (Pallab Kar) Department of Botany, University of North Bengal, Siliguri- 734013, West Bengal, India*

***R.K. Jansen**: (Robert K. Jansen) King Abdulaziz University, Faculty of Science, Department of Biological Sciences, Genomics and Biotechnology Section, Jeddah- 21589, Saudi Arabia; Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA*

***R.P. Malone**: (Renée P. Malone) Dublin Institute of Technology, School of Food Science and Environmental Health, Dublin 1, Ireland*

***S. Alrokayan**: (Salman Alrokayan) Research Chair of Targeting and Treatment of Cancer Using Nanoparticles, Department of Biochemistry, College of Science, King Saud University, P.O. Box 2455, Riyadh- 11451, Saudi Arabia*

***S. Edris**: (Sherif Edris) King Abdulaziz University, Faculty of Science, Department of Biological Sciences, Genomics and Biotechnology Section, Jeddah- 21589, Saudi Arabia; Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), Faculty of Medicine, King Abdulaziz University (KAU), Jeddah, Saudi Arabia; Genetics Department, Faculty of Agriculture, Ain Shams University, Cairo- 11241, Egypt*

***S. Karuppusamy**: (Subbiah Karuppusamy) Centre for Botanical Research, Department of Botany, The Madura College (Autonomous), Madurai-625 011, Tamilnadu, India*

***S. Sorvari**: (Seppo Sorvari) Agricultural Research Centre, Institute of Horticulture, Toivonlinnantie 518, FIN-21500 Piikkiö, Finland*

***S.H. Park**: (Sang-Hong Park) Division of Plant Management, National Institute of Ecology, Choongnam, Secheon-gun, Maseo-myeon, Geumgang-ro, 325-813, South Korea*

***S.K. Pandey**: (Shankar Kumar Pandey) Department of Botany, T.M. Bhagalpur University, Bhagalpur- 812007, Bihar, India*

***S.M. Ragheb**: (Suzan M. Ragheb) Department of Biotechnology, The Nile Company for Pharmaceuticals and Chemical Industries, Cairo, Egypt*

***S.Y. Kim**: (Soo-Yong Kim) International Biological Material Research Center, Korea Research Institute of Bioscience and Biotechnology, 111 Gwahangno, Yuseong-gu, Daejeon-305 806, South Korea*

***T. Demku**: (Tamás Demku) Institute of Genetics and Biotechnology, St. Stephanus University, Gödöllő 2103, Hungary*

***T. Kömíves**: (Tamás Kömíves) Hungarian Academy of Sciences, CAR, Plant Protection Institute, Budapest, Hungary 1525*

***Y.L. Krishnamurthy**: (Yelugere Lingyanaik Krishnamurthy) Department of P.G. Studies and Research in Applied Botany, Kuvempu University, Jnana Sahyadri, Shankaraghatta- 577451, Karnataka, India*

***Z. Szabó**: (Zoltán Szabó) Agricultural Biotechnology Center, 2100 Gödöllő, Hungary*

***Z. Tóth**: (Zoltán Tóth) Agricultural Biotechnology Center, 2100 Gödöllő, Hungary*

***Zs. Tóth**: (Zsófia Tóth) Institute of Genetics and Biotechnology, Faculty of Agricultural and Environmental Sciences, St István University, Gödöllő 2103, Hungary*

Plant DNA Barcoding and Phylogenetics, Pages 1-14
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

1 Introduction to Plant DNA Barcoding

M.A. Ali, A.K. Pandey, G. Gyulai, S.H. Park, C. Lee, S.Y. Kim and J. Lee

Introduction

Till date taxonomists have described approximately 1.7 million species, but this figure might be a gross under-estimate of the true biological diversity on Earth (Blaxter 2003; Wilson, 2003). Although taxonomists can identify most organisms with which they are familiar, an ever-growing community requires taxonomic information for a broad range of taxa. DNA barcoding is a novel system designed to provide rapid, accurate, and automatable species identifications by using short, standardized gene regions as internal species tags (Hebert et al., 2005). The genomes of living organisms are analogous to bar-codes. The use of short DNA sequences for biological identifications was first proposed by Paul Herbert and colleagues in 2003. The role of barcodes is to provide a tool to assign unidentified specimens to already characterized species (Hebert et al. 2003). Building upon the idea of the 'universal product code', known as 'barcodes', a few DNA nucleotides (e.g. the sequences of a short DNA fragment) may provide an immediate diagnosis for species. As with commercial barcodes, the use of these 'species barcodes' first require the assembly of a comprehensive library that links barcodes and organisms. DNA barcodes consist of a short sequence of DNA between 400 and 800 base pairs long that can be easily extracted and characterized for all species on this planet. These genetic barcodes will be accessed through a digital library and used to identify unknown plants in the field or garden.

DNA barcoding follows the same principle as does the basic taxonomic practice of associating a name with a specific reference collection in conjunction with a functional understanding of species concepts (i.e., interpreting discontinuities in interspecific variation). In DNA barcoding the complete data can be obtained from single specimens irrespective of sexual morph or life stage. Morphologically indistinguishable taxa can be diagnosed without the need for live material, particular morphs or population measures. Barcode sequences can be generated from type specimens (holotype, paratype or neotype). A specimen barcode can be compared with sequences derived from other molecular taxonomy initiatives. If a close match is found to a named taxon, recourse can be made to traditional monographs and keys to understand the biological properties of the identified MOTU (molecular operational taxonomic units) and their close relatives (Floyd et al. 2002). Molecular phylogenetic analyses can be used to generate testable hypotheses of MOTU interrelatedness. The core idea of DNA barcoding is based on the fact that short pieces of DNA can be found that vary only to a very minor degree within species, such that this variation is much less than between species.

Whether or not actual species can be identified with DNA, the number of distinct DNA sequences in environmental sampling and reconstruction of phylogenetic trees to place these sequences into an evolutionary context have been used in several inventories of cryptic biodiversity (e.g. soil bacteria or marine/freshwater micro-organisms). Initially referred to as DNA typing or profiling, the DNA barcoding initiative has taken this step forward, and several taxa have now been surveyed in their natural habitats using this technique. Such an approach has been particularly useful for marine organisms (Shander and Willassen, 2005), including fishes (Mason, 2003; Ward et al., 2005), soil meiofauna (Blaxter et al., 2004), freshwater meiobenthos (Markmann and Tautz, 2005) and even extinct birds (Lambert et al., 2005). In the rainforests, rapid DNA-based entomological inventories have been performed so efficiently (Monaghan et al., 2005; Smith et al., 2005) that tropical ecologists have been among the most active advocates of DNA barcoding (Janzen, 2004).

Plant DNA barcoding markers

The use of DNA sequences to identify organisms has been proposed as a more efficient approach than traditional taxonomic practices (Blaxter et al, 2004; Tautz et al., 2003). **T**he identification of animal biological diversity by using molecular markers has recently been proposed and demonstrated on a large scale through the use of a short DNA sequence the mitochondrial cytochrome oxidase subunit 1 (cox1, usually referred to as COI in barcoding studies), was proposed to be a good candidate for barcoding animal species (Hebert et al., 2003). The availability of broad-range primers for amplification of mitochondrial COI from diverse invertebrate phyla establishes this gene as a particularly promising target for species identification in animals (Folmer et al., 1994). Plants have relatively little sequence variation in their mitochondrial DNA, perhaps because of hybridization and introgression. A chloroplast gene such as matK (maturase K) or a nuclear gene such as ITS (internal transcribed spacer) may be an effective target for barcoding in plants (Kress et. al., 2005). Kress et al. (2005) have demonstrated the effectiveness of "DNA barcoding" in angiosperms using nrDNA and non-coding cpDNA sequences.

In flowering plants another approach has been put forward. On one hand several plastid loci do discriminate between species, e.g. the trnH-psbA intergenic spacer (Kress et al., 2005) and some more typical phylogenetic markers such as rbcL and trnL-F (Chase et al., 2005), but on the other hand multiple genetic loci might be necessary to account for the common hybridization and polyploidy events in angiosperms. Ribosomal DNA (e.g. ITS in orchids) could be used to complement plastid genes, and shorter low-copy nuclear markers are being discovered that might in the future be used to provide a more sophisticated multiple component barcode for species diagnosis and delimitation (Chase et al., 2005). The sequences used thus for molecular barcoding are the nuclear small subunit ribosomal RNA gene (SSU, also known as 16S in prokaryotes, and 18S in most eukaryotes), the nuclear large-subunit ribosomal RNA gene (LSU, also known as 23S and 28S; the highly variable expansion loops that are flanked by conserved stem sequences are particularly useful), the highly variable internal transcribed spacer section of the ribosomal RNA cistron (ITS, separated by the 5S ribosomal RNA gene into ITS1 and ITS2 regions), the mitochondrial cytochrome *c* oxidase 1 (CO1 or COX1) gene and the chloroplast ribulose bisphosphate carboxylase large subunit (rbcL) gene. Kress et al. (2005) have suggested that the nuclear internal transcribed spacer region and the plastid trnH-psbA intergenic spacer as potentially usable DNA regions for applying barcoding to flowering plants. The internal transcribed spacer is the most commonly sequenced locus used in plant phylogenetic investigations at the species level and shows high levels of interspecific divergence (Pandey and Ali, 2006). The trnH-psbA spacer, although short (≈450-bp), is the most variable plastid region in angiosperms and is easily amplified across a broad range of land plants (Kress et al., 2005).

The primary reason that barcoding has not been applied to plants is that plant mitochondrial genes, because of their low rate of sequence change, are poor candidates for species-level discrimination. The divergence of CO1 coding regions among families of flowering plants has been documented to be only a few base pairs across 1.4 kb of sequence. Furthermore, plants rapidly change their mitochondrial genome structure; thereby precluding the existence of universal intergenic spacers that otherwise would be appropriately variable unique identifiers at the species level. The ITS region has shown broad utility across photosynthetic eukaryotes (with the exception of ferns) and fungi and has been suggested as a possible plant barcode locus. Species-level discrimination and technical ease have been validated in most phylogenetic studies that employ ITS, and a large body of sequence data already exists for this region. An advantage of the ITS region is that it can be amplified in two smaller fragments (ITS1 and ITS2) adjoining the 5.8S locus, which has proven especially useful for degraded samples. The quite conserved 5.8S region in fact contains enough phylogenetic signals for discrimination at the level of orders and phyla, although identification at this taxonomic level is not the concern of barcoding. The 5.8S locus can serve as a critical alignment-free anchor point for search algorithms that make sequence comparisons for both phylogenetic and barcoding purposes. The utility of conserved regions such as 5.8S to generate a pool of nearest neighbors for refined comparisons will be critical for effective database searches, especially when comparing a sequence that has no identical match in a sequence library.

For phylogenetic investigations, the plastid genome has been more readily exploited than the nuclear genome and may offer for plant barcoding what the mitochondrial genome does for animals. It is a uniparentally inherited, nonrecombining, and, in general, structurally stable genome. Universal primers are available for a number of loci and intergenic spacers that are evolving at a variety of rates. The plastid locus most commonly sequenced by plant systematists for phylogenetic purposes is rbcL, followed by the trnL-F intergenic spacer, matK, ndhF, and atpB-rbcL has been suggested as a candidate for plant barcoding, even though it has generally been used to determine evolutionary relationships at the generic level and above. Besides rbcL and atpB, all of the latter plastid loci have been used at the species level with various degrees of success. Most of them (except the trnL-F spacer) require full-length sequences of >1 kb to yield enough sequence length to discriminate species. Most relevant to plant barcoding, no region of the plastid genome has been found to have the high level of variation seen in most animal CO1 barcodes, although a few intergenic spacers have shown more promise than any plastid locus now in general use. Kress et al. (2005) have compared plastid genomes of *Atropa* and *Nicotiana*, and recorded that nine intergenic spacers trnK-rps16, trnH-psbA, rp136-rps8, atpB-rbcL, ycf6-psbM, trnV-atpE, trnC-ycf6, psbM-trnD*,* and trnL-F met the barcode criteria. By comparison, ITS had a much higher divergence value (13.6%) than any of the plastid regions, and rbcL was by far the lowest in divergence (0.83%). Although three spacers (atpB-rbcL, ycf6-psbM, and psbM-trnD) were slightly to moderately longer than our 800-bp cutoff.

Besides ITS, those single-copy nuclear genes or their introns that are gaining prominence in species-level molecular systematics studies (e.g., *leafy, waxy, pistillata, and RPB2*), also have been considered. The significantly greater length of *rbcL* (usually 1,428 bp) causes problems because it is necessary to use four primers for double-stranded sequencing of the entire gene. It has been suggested that the trnH-psbA intergenic spacer is the best plastid option for a DNA barcode sequence that has good priming sites, length, and interspecific variation. In their trials across a diverse set of genera in seven plant families, Kress et al. (2005) reported that three plastid regions (trnH-psbA, rp136-rpf8*,* and trnL-F) ranked highest with respect to amplification success and appropriate sequence length, but trnH-psbA demonstrated nearly three times the percentage sequence divergence of these other two regions. By applying barcode criteria (i.e., length considerations and universality) to the framework of their study, it has been concluded that trnH-psbA has greater potential for species-level discrimination than any other locus (Kress et al., 2005).

Despite this high level of interspecific variation, trnH-psbA has found only limited use in species-level phylogenetic reconstruction because of the short length as well as the difficulty of alignments resulting from a high number of indels (deletions). In contrast with the problems of indels for phylogenetic construction, it is suspected that indels will ultimately enhance the information needed for species identifications, once the appropriate informatics tools for barcoding are developed. Both ITS and trnH-psbA are good starting points for large-scale testing of DNA barcoding across a large sample of angiosperms.

Basic steps in DNA barcoding

DNA barcoding, a new method for the quick identification of any species based on extracting a DNA sequence from a tiny tissue sample of any organism, is now being applied to taxa across the tree of life. As a research tool for taxonomists, DNA barcoding assists in identification by expanding the ability to diagnose species by including all life history stages of an organism. As a biodiversity discovery tool, DNA barcoding helps to flag species that are potentially new to science. As a biological tool, DNA barcoding is being used to address fundamental ecological and evolutionary questions, such as how species in plant communities are assembled. The process of DNA barcoding entails two basic steps: (1) building the DNA barcode library of known species and (2) matching the barcode sequence of the unknown sample against the barcode library for identification. Although DNA barcoding as a methodology has been in use for less than a decade, it has grown exponentially in terms of the number of sequences generated as barcodes as well as its applications (Kress and Erickson, 2012).

DNA is a relatively stable molecule, and can be isolated from museum collections, including specimens preserved in formalin (Fang et al., 2002). The extraction of DNA from specimens in herbarium collections can easily be made. This success may be due to the specimens having been air-dried and in a good state of preservation as evidenced by the generally green appearance of the leaves selected for extraction. Plant voucher specimens vary in how and when they are dried after being pressed. If specimen-drying facilities are not immediately available, especially in humid tropical climates, botanists often treat pressed specimens with ethanol to temporarily preserve them against fungal attack and degradation. Alcohol has been shown to be detrimental to recovering high-quality DNA, although how it will affect the short sequences needed for barcoding is unknown. It is encouraging that museum specimens of insects dried from ethanol storage readily yield CO1 sequences. A more thorough investigation and optimization of methods to extract high-quality barcode DNA from herbarium collections in a high-throughput format will be critical to efficiently build a sequence-database library for plant DNA barcodes. Positive results have been obtained by using well preserved specimens which indicate that the *a priori* selection of apparently under graded plant samples will be an important determinant of success. Fortunately, herbaria often have more than one specimen per species among which to select for successful DNA barcoding.

Recent advances

Global DNA barcoding efforts have resulted in the formation of the Consortium for the Barcode of Life (CBOL). In January 2013, the Barcode of Life Database (BOLD) contained more than 2.7 million specimen records, with 2 million having barcodes belonging to over 170,000 species (Ratnasingham and Hebert, 2007; BOLD Systems, 2013). Smaller databases, containing sequences of specialized groups, also exist [for example, Fungal Database (Crous et al., 2004), Genome Database for Rosaceae, GDR (Jung et al., 2008)].

The main DNA barcoding bodies and resources are (1) Consortium for the Barcode of Life (CBOL) http://www.barcodeoflife.org established in 2004. CBOL promotes DNA barcoding through over 200 member organizations from 50 countries, operates out of the Smithsonian Institution's National Museum of Natural History in Washington, (2) International Barcode of Life (iBOL) http://www.ibol.org Launched in October 2010, iBOL represents a not-for-profit effort to involve both developing and developed countries in the global barcoding effort, establishing commitments and working groups in 25 countries. The Biodiversity Institute of Ontario is the project's scientific hub and its director, (3) The Barcode of Life Datasystems (BOLD) http://www.boldsystems.org. The Barcode of Life Datasystems is an online workbench for DNA barcoders, combines a barcode repository, analytical tools, interface for submission of sequences to GenBank, a species identification tool and connectivity for external web developers and bioinformaticians. The Consortium for the Barcode of Life (CBOL) Plant Working Group (2009) recommended rbcL + matK as a core two-locus combination. However, as these loci encode conserved functional traits, it is not clear whether they provide sufficiently high species resolution. One of the challenges for plant barcoding is the ability to distinguish closely related or recently evolved species.

The classical way of practice of plant taxonomy for the identification of species lead the discipline many a times to a subject of opinion; the plant DNA barcoding is now transitioning the epitome of species identification (Ali et al., 2014). One of the most important uses of the DNA barcoding is in the medicinal plant authentication. Recently ITS, trnH-psbA, rbcL, matK and trnL–trnF gene sequence have successfully been used for DNA barcoding of several plant species. In addition with the above, Chen et al. (2010) tested the discrimination ability of ITS2 in more than 6600 plant samples belonging to 4800 species from 753 distinct genera and found that the rate of successful identification with the ITS2 was 92.7% at the species level. Yao et al. (2010) also evaluated 50,790 plant and 12,221 animal ITS2 sequences downloaded from GenBank, and propose that the ITS2 locus should be used as a universal DNA barcode for identifying plant species and as a complementary locus for CO1 to identify animal species.

Benefits

Traditionally, taxonomic identification has relied upon morphological characters. In the last two decades, molecular tools based on DNA sequences of short standardized gene fragments, termed DNA barcodes, have been developed for species discrimination. The most common DNA barcode used in animals is a fragment of the cytochrome c oxidase (COI) mitochondrial gene, while for plants, two chloroplast gene fragments from the RuBisCo large subunit (rbcL) and maturase K (matK) genes are widely used. Information gathered from DNA barcodes can be used beyond taxonomic studies and will have far-reaching implications across many fields of biology, including ecology (rapid biodiversity assessment and food chain analysis), conservation biology (monitoring of protected species), biosecurity (early identification of invasive pest species), medicine (identification of medically important pathogens and their vectors) and pharmacology (identification of active compounds). However, it is important that the limitations of DNA barcoding are understood and techniques continually adapted and improved as this young science matures (Fisˇer and Buzan, 2014)

DNA barcodes are likely to play a major role in the future of taxonomy. The build-up of DNA databases has great potential for the identification and classification of organisms and for supporting ecological and biodiversity research programmes (Tautz et al., 2002). As a uniform, practical method for species identification, it appears to have broad scientific applications. DNA-based species identification offers enormous potential benefits for the biological scientific community, educators, and the interested public. It will help open the treasury of biological knowledge and increase community interest in conservation biology and understanding of evolution. A rapid and accurate method is now being developed for the quick identification of plant species based on extracting DNA from a tiny tissue sample of a leaf, flower, or fruit.

The direct benefits of DNA barcoding is to make the outputs of systematics available to a large number of end-users by providing standardized and high-tech identification tools, e.g. for biomedicine (parasites and vectors), agriculture (pests), environmental assays and customs (trade in endangered species). It will provide a bio-literacy tool for the general public. DNA based species identification will help open the treasury of biological knowledge, which is currently underused partly because taxonomic expertise for species identification is relatively inaccessible. DNA barcoding will also relieve the enormous burden of identifications from taxonomists, so they can focus on more pertinent duties such as delimiting taxa, resolving their relationships and discovering and describing new species. It will also help in pairing up various life stages of the same species (e.g. seedlings, larvae). The most important aspect of DNA barcoding is that it will facilitate basic biodiversity inventories (Savolainen et al., 2005).

DNA barcoding can be likened to aerial photography, in that it provides an efficient method for mapping the extent of species, though in sample space rather than physical space. The “aerial map” of DNA barcodes will help investigators explore the biological world and make full use of the enormous knowledge that has been built on 250 years of classical taxonomy. As sequencing costs decrease, DNA-based species identification will become available to an increasingly wide community. When costs are low enough, science teachers and backyard naturalists will be able to use DNA barcoding for in depth examination of local ecosystems.

Limitations

DNA-based species identification depends on distinguishing intraspecific from interspecific genetic variation. The ranges of these types of variation are unknown and may differ between groups. It may be difficult to resolve recently diverged species or new species that have arisen through hybridization. There is no universal DNA barcode gene, no single gene that is conserved in all domains of life and exhibits enough sequence divergence for species discrimination. The validity of DNA barcoding therefore depends on establishing reference sequences from taxonomically confirmed specimens. This is likely to be a complex process that will involve cooperation among a diverse group of scientists and institutions.

Sequencing is essentially equally easy for all DNA fragments barring extreme base composition biases, polynucleotide runs and stable secondary structures. However, the ITS region often varies by insertions or deletions within an individual, making sequencing very difficult as two independent sequence types are being analysed simultaneously (Elbadri et al. 2002). ITS sequences are also difficult to align as they tend to evolve by insertion and deletion rather than substitution making the secondary steps of phylogenetic reconstruction problematic. SSU, LSU, COX1 and rbcL are each relatively simple to align and analyse, though exceptions do occur. It may be suggested that any barcoding system should aim to acquire data for at least a nuclear and an organellar gene from single specimens. For specimen-independent, ‘environmental DNA’ based surveys, any target may do, but the universality of SSU and LSU primer sets recommends them. The most common criticism of 18S rDNA, as a source of phylogenetic information, has been that it is not sufficiently variable for phylogenetic reconstruction within the angiosperms and that it is highly prone to insertion and deletion, making sequence alignment difficult. 18S rDNA provides a sufficient number of characters for broad scale phylogenetic reconstruction of the angiosperms.

Where species are simply unknown or no attempts have been made to delimit them, the barcode approach as originally intended would be limited in its applicability. However, it is a widely accepted fact that species, however defined, are variable for most DNA markers including the widely used cox1 gene. Hence, the analogy to commercial barcodes presumes that the variation within these species is smaller than between them.

Barcoding has created some controversy in the taxonomy community (Mallet and Willmott, 2003; Lipscomb et al., 2003; Seberg et al., 2003; DeSalle et al., 2005; Lee, 2004; Ebach and Holdrege, 2005; Will et al, 2005; Gregory, 2005). Traditional taxonomists use multiple morphological traits to delineate species. Today, such traits are increasingly being supplemented with DNA-based information. In contrast, the DNA barcoding identification system is based on what is in essence a single complex character (a portion of one gene, comprising ~650 bp from the first half of the mitochondrial cytochrome c oxidase subunit I gene sometimes called COXI or COI), and barcoding results are therefore seen as being unreliable and prone to errors in identification (Dasmahapatra and Mallet, 2006). Although the mitochondrial cytochrome oxidase subunit I (CO1) is a widely used barcode in a range of animal groups (Hebert et al., 2003), this locus is unsuitable for use in plants due to its low mutation rate (Kress et al., 2005; Cowen et al., 2006; Fazekas et al., 2008). In addition, complex evolutionary processes, such as hybridization and polyploidy, are common in plants, making species boundaries difficult to define (Rieseberg et al., 2006; Fazekas et al., 2009). The number and identity of DNA sequences that should be used for barcoding is a matter of debate (Pennisi, 2007; Ledford, 2008).

In conclusion, methods for identifying species by using short orthologous DNA sequences, known as “DNA barcodes”. In DNA barcoding the complete data can be obtained from single specimens irrespective of sexual morph or life stage. Morphologically indistinguishable taxa can be diagnosed without the need for live material. The core idea of DNA barcoding is based on the fact that short pieces of DNA can be found that vary only to a very minor degree within species, such that this variation is much less than between species. More pragmatically, DNA barcodes have proved useful in biosecurity, e.g. for surveillance of disease vectors (Besansky et al., 2003) and invasive insects (Armstrong and Ball, 2005), as well as for law enforcement and primatology (Lorenz et al., 2005). These “DNA barcodes” show promise in providing a practical, standardized, species-level identification tool that can be used for biodiversity assessment, life history and ecological studies, and forensic analysis.

References

Ali, M.A., Gábor, G., Norbert, H., Balázs, K., Al-Hemaid, F.M.A., Pandey, A.K. and Lee, J. (2014) The changing epitome of species identification- DNA barcoding. Saudi Journal of Biological Sciences 21: 204-231.

Armstrong, K.F. and Ball, S.L. (2005) DNA barcodes for biosecurity: invasive species identification. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1813–1823.

Besansky, N.J., Severson, D.W. and Ferdig, M.T. (2003) DNA barcoding of parasites and invertebrate disease vectors: what you don't know can hurt you. Trends in Parasitology 19: 545–546.

Blaxter, M. (2003) Counting angels with DNA. Nature 421: 122–124.

Blaxter, M., Elsworth, B. and Daub, J. (2004) DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. Transactions of the Royal Society B: Biological Sciences 271: 189–192.

BOLD Systems (2013) BOLD Systems v3. http://www.boldsystems.org/ (Retrieved on 6 May 2014).

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences USA 106: 12794–12797.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.P., Haidar, N. and Savolainen, V. (2005) Land plants and DNA barcodes: short-term and long term goals. Transactions of the Royal Society B: Biological Sciences 360: 1889–1895.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. and Leon, C. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5(1): e8613.

Cowen, R.K., Paris, C.B. and Srinivasan, A. (2006) Scaling of connectivity in marine populations. Science 311 (5760): 522–527.

Crous, P.W., Gams, W., Stalpers, J.A., Robert, V. and Stegehuis, G. (2004) MycoBank: an online initiative to launch mycology into the 21st century. Studies in Mycology 50:19–22.

Dasmahapatra, K.K. and Mallet, J. (2006) DNA barcodes: recent successes and future prospects. Heredity 97 (4): 254–255.

DeSalle, R., Egan, M.G. and Siddall, M. (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1905–1916.

Ebach, M.C., and Holdrege, C. (2005) DNA barcoding is no substitute for taxonomy. Nature 434: 697.

Elbadri, G.A., De Ley, P., Waeyenberge, L., Vierstraete, A., Moens, M. and Vanfleteren, J. (2002) Intraspecific variation in *Radopholus similis* isolates assessed with restriction fragment length polymorphism and DNA sequencing of the internal transcribed spacer region of the ribosomal RNA cistron. International Journal for Parasitology 32: 199–205.

Fang, S.G., Wan, Q.H. and Fijihara, N. (2002) Formalin removal from archival tissue by critical point drying. BioTechniques 33: 604–611.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M. and Barrett, S.C. (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3 (7): e2802.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C., Newmaster, S.G., Hajibabaei, M. and Husband, B.C. (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular Ecology Resources s1: 130–139.

Fisˇer, P.Z. and Buzan, E.V. (2014) 20 years since the introduction of DNA barcoding: from theory to application. Journal of Applied Genetics 55 (1): 43–52.

Floyd, R., Eyualem, A., Papert, A. and Blaxter, M.L. (2002) Molecular barcodes for soil nematode identification. Molecular Ecology 11: 839–850.

Folmer, O., Black, M., Hoeh, W., Lutz, R. and Vrijenhoek, R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3: 294–299.

Gregory, T.R. (2005) DNA barcoding does not compete with taxonomy. Nature 434 1067.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and De Waard, J.R. (2003) Biological identifications through DNA barcodes. Proceedings Biological sciences / The Royal Society 270: 313–321.

Janzen, D.J. (2004) Now is the time. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 359: 731–732.

Jung, S., Staton, M., Lee, T., Blenda, A., Svancara, R., Abbott, A. and Main, D. (2008) GDR (Genome Database for Rosaceae): integrated webdatabase for Rosaceae genomics and genetics data. Nucleic Acids Research 36(Database issue):D1034–D1040.

Kress, W.J. and Erickson, D.L. (2012) DNA barcodes: methods and protocols. Methods in Molecular Biology 858: 3-8.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A. and Janzen, D. H. (2005), Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America 102: 8369–8374.

Lambert, D.M., Baker, A., Huynen, L., Haddrath, O., Hebert, P.D.N. and Millar, C.D. (2005) Is a large-scale DNA-based inventory of ancient life possible? The Journal of heredity 96: 279–284.

Ledford, H. (2008) Botanical identities: DNA barcoding for plants comes a step closer. Nature 451: 616.

Lee, M.S.Y. (2004) The molecularization of taxonomy. Invertebrate Systematics18, 1–66.

Lipscomb, D., Platnick, N. and Wheeler, Q. (2003) The intellectual content of taxonomy: a comment on DNA taxonomy. Trends in Ecology & Evolution 18 (2): 65–66.

Lorenz, J.G., Jackson, W.E., Beck, J.C. and Hanner, R. (2005) The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 360: 1869–1877.

Mallet, J. and Willmott, K. (2003) Taxonomy: renaissance or tower of Babel? Trends in Ecology & Evolution 18 (2): 57–59.

Markmann, M. and Tautz, D. (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 360, 1917–1924.

Mason, B. (2003) Marine surveys sees net gain in number of fish species. Nature 425: 889.

Monaghan, M.T., Balke, M., Gregory, T.R. and Vogler, A.P. (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 360: 1925–1933.

Pandey, A.K. and Ali, M.A. (2006) Molecular Markers in Plant Systematics I: Nuclear sequences. In: Plant Sciences Research in India: Challenges and Prospects (ed. S. Kumar), Botanical survey of India, Dehradun, pp. 21-34.

Pennisi, E. (2007) Taxonomy. Wanted: a barcode for plants. Science 318: 190–191.

Ratnasingham, S. and Hebert, P.D.N. (2007) Bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes 7: 355-364.

Rieseberg, L.H., Wood, T.E. and Baack, E.J. (2006) The nature of plant species. Nature 440: 524–527.

Savolainen, V. and Reeves, G. (2004) A plea for DNA banking. Science 304: 1445.

Savolainen, V., Cowan, R., Vogler, A.P., Roderick, G.K., and Lane, R. (2005), Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 360: 1805-1811.

Schindel, D.E. and Miller, S.E. (2005) DNA barcoding a useful tool for taxonomists. Nature 435: 17.

Seberg, O., Humphries, C.J., Knapp, S., Stevenson, D.W., Petersen, G., Scharff, N. and Andersen, N.M. (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends in Ecology and Evolution 18: 63–65.

Shander, C. and Willassen, E. (2005) What can biological barcoding do for marine biology. Marine Biology Research 1: 79–83.

Smith, M.A., Fisher, B.L. and Hebert, P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 360: 1825–1834.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2003) A plea for DNA taxonomy. Trends in Ecology and Evolution 18: 70–74.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2002) DNA points the way ahead in taxonomy. Nature 418: 479.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. and Hebert, P.D.N. (2005) DNA barcoding Australia’s fish species. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 360: 1847–1857.

Will, K.W., Mishler, B.D. and Wheeler, Q.D. (2005) The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 54 (5): 844–851.

Wilson, E.O. (2003) The encyclopaedia of life. Trends in Ecology and Evolution 18: 77–80.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P. and Chen, S. (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE 5(10): e13102.

*****

Plant DNA Barcoding and Phylogenetics, Pages 15-36
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

2 Molecular Markers for Plant DNA Barcoding

M.R. Enan

Introduction

Traditionally the macroscopic and microscopic identifications are performed to authenticate plant materials at the species level. There are an estimated 300000 plant species in the world (International Union for Conservation of Nature; IUCN, 2012), the accurate classification and identification of this large number of species remains a challenge even for specialist taxonomists. The emergence of DNA barcoding has had a positive impact on biodiversity classification and identification (Gregory, 2005). Several universal systems for molecular systematic analyses were used for lower taxa but were not successfully applied for broader range. The 'Barcode of Life' project aims to create a universal system for a eukaryotic species based on a standard molecular approach. It was initiated in 2003 by researchers at the University of Guelph in Ontario, Canada (http://www.barcoding.si.edu) and promoted in 2004 by the international initiative 'Consortium for the Barcode of Life' (CBOL). The DNA Barcode of Life Data System (BOLD, http://www.boldsystems.org) has progressively been developed since 2004 and was officially established in 2007 (Ratnasingham and Hebert, 2007). This data system enables the storage, analysis and publication of DNA barcode records.

Sample collection and DNA preservation

Total genomic DNA extraction from the collected plant tissue sample is the first step followed by amplification of desired region using barcode primer using PCR. The amplified sequence (amplicon) is then subjected to sequencing in one or both directions. The tools of bioinformatics are then used for the analyses of generated sequences (Figure1).

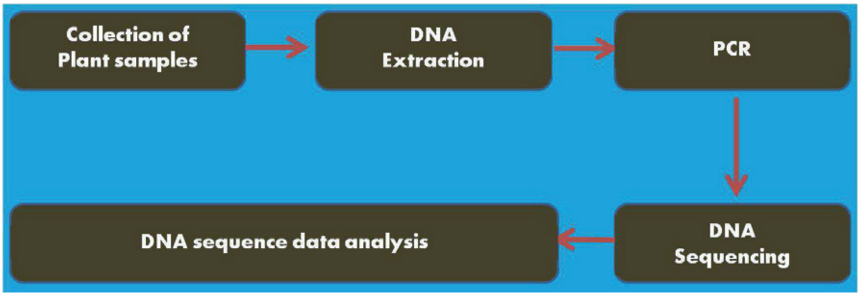


Fig.1: Flow chart demonstrating practical steps involved in plant DNA barcoding.

A close match quickly identifies a species that is already represented in the database. Three methods have been developed to preserve DNA in plant samples collected in the field (Kress, 2004; Gonzalez et al., 2009; Webb et al., 2010). One employs silica gel as a desiccant to rapidly dry the tissue, which reduces degradation in most specimens (Kress, 2004). However, it does not eliminate degradation, and DNA yields are low for some tissues (Condit, 1998). The second method uses a saturated NaCl-CTAB (cetyltrimethylammonium bromide) solution. The high salt partially dehydrates the tissues and the CTAB can complex with nucleic acids, proteins and carbohydrates to slow the degradation processes. However, high degrees of degradation have been noted in some cases with this method, and occasionally low yields of DNA result (Condit, 1998). The third method uses an absorbent paper for preserving the DNA (Webb et al., 2010). Pieces of plant tissues are mashed onto the paper, and then allowed to dry. Almost all methods to extract nucleic acids must be performed in a laboratory (Fine and Ree, 2006; Dick and Kress, 2009). Generally fresh tissues are used for extraction of the nucleic acids, because degradation and other biochemical processes begin immediately after the tissue has been removed from the organism or from its natural substrate. DNA in many species of plants has been detected in dried tissues from months to centuries after the organism has died (Ratnasingham and Hebert, 2007; Dick and Heuertz, 2008). When the samples cannot be effectively sampled, preserved and transported rapidly to the laboratory, then alternatively the laboratory equipment and solutions can be transported to the target specimens in their natural environments in order to extract the DNA (*in situ*), a possible alternative that can minimize degradation and maximize yield. Degradation can be monitored by gel electrophoresis because as the DNA is broken down the higher molecular weight bands become more diffuse and smaller fragments of DNA are seen as increasingly bright smears of fluorescence extending into lower molecular weight regions of the gel. Simultaneously, other degradative processes also occurs, resulting in losses of sequence information (Ratnasingham and Hebert, 2007). The most common changes are losses of bases by hydrolytic attack of the glycosidic bonds. Depurination occurs at a higher rate, but depyrimidization occurs at a lower rate. When these DNAs are used as templates for PCR, approximately 75% of the time, an inaccurate base will be incorporated at those sites, causing a potential loss in sequence accuracy.

Plant DNA barcode primers

In every species, primer information is the most vital in starting the screening of various reported candidate genes towards their suitability. The forward and reverse sequences should be carefully combined (Table 1). Several universal primers for amplifying noncoding spacers of the chloroplast genome have been reported (Demesure et al., 1995). Most of the primers were designed for amplifying spacers between tRNA genes which have been proved variable among species (Demesure et al., 1996). Plant nuclear genes often occur in multiple copies and are highly variable, making the design of universal primers difficult (Yu et al., 2011).

Plant DNA barcode elements

For DNA barcoding to work, sequence variation must be high enough between species so that they can be discriminated from one another; however, it must be low enough within species that a clear threshold between intra- and inter-specific genetic variations can be defined. The two most important traits of DNA barcoding loci are the presence of conserved flanking regions to enable routine amplification across highly different taxa and sufficient internal variability to facilitate species discrimination but with a relatively low level of intra-specific variation.

Additional factors are short length facilitated routine sequencing, even with sub-optimal material, lack of heterozygosity enabling direct polymerase chain reaction followed by sequencing without cloning, ease of alignment that enables the use of character-based data analysis methods, lack of problematic sequence composition, such as regions with several microsatellites, that reduces sequence quality, universal capability to get amplified/sequenced with standardized primers, easy align ability and capability to get recovered easily from herbarium samples and other degraded DNA samples (Hollingsworth et al.,2009).

Types of plant DNA barcode markers

A total of 17 barcode regions (matK, rbcL, ITS, ITS2, psbA-trnH, atpF-atpH, ycf5, psbK-I, psbM-trnD, rps16, coxI, nad1, trnL-F, rpoB, rpoC1, atpF-atpH, rps16) of medicinal plants were reported to aid in the authentication and identification of medicinal plant materials. The majority of barcoding regions stated in the literature were the matK, ITS region, rbcL, and psbA-trnH. Although many studies have searched for a universal plant barcode, none of the available loci work across all species (Chase and Fay, 2009; Chen et al., 2010). The Consortium for the Barcode of Life-Plant Working Group (CBOL) recently recommended the two-locus combination of matK + rbcL as the best plant barcode with a discriminatory efficiency of only 72% (CBOL Plant Working Group, 2009). Taxonomists have suggested that a multi-locus method may be necessary to discriminate plant species (Hebert et al., 2004; Chase et al., 2007; Kress and Erickson, 2007; Erickson et al., 2008; Lahaye et al., 2008; Kane et al., 2012). However, CBOL demonstrated that the use of multiple loci did not clearly improve the species-level discriminatory ability of these techniques (CBOL Plant Working Group, 2009). Researchers have recently proposed the use of the whole-plastid genome sequence in plant identification (Erickson et al., 2008; Sucher and Carles, 2008; Parks et al., 2009; Nock et al., 2011; Yang et al., 2013). However this concept has not yet been universally accepted. One of the main concerns is the high sequencing cost and difficulties involved in obtaining complete plastid genome sequences in comparison to the use of single-locus barcodes. Hollingsworth et al. (2011) argued that the full plastid haplotype is not a good marker because it does not always track species boundaries. To date, it is still unclear whether plastid genomes can be regarded as a suitable barcode.

Table 1: Primer sequences for the candidate genes for barcoding in plants.

Barcode markers	**Primer sequence (5’-3’)**	**Reference**
ITS2 (The second internal transcribed spacer of nuclear ribosomal DNA)	ITS3-F 5'- GCATCGATGAAGAACGTAGC-3' ITS4-R 5'- TCCTCCGCTTATTGATATGC-3'	White et al. (1990)
matK (Maturase coding gene)	matK472F 5'-CCCRTYCATCTGGAAATCTTGGTTC-3' matK1248R 5'-GCTRTRATAATGAGAAAGATTTCTGC-3'	Yu et al. (2011)
	3f-KIM-F 5’-CGTACAGTACTTTTGTGTTTACGAG-3’ 1R KIM-R 5’-ACCCAGTCCATCTGGAAATCTTGGTTC-3’	CBOL Plant Working Group (2009)
	matK_1F 5’-GAACTCGTCGGATGGAGTG-3’ matK_12R 5’-GAGAAATCTTTTTCATTACTACAGTG-3’	Wang et al. (1999)
	matK_2F 5’-CGTACTTTTATGTTTACAGGCTAA-3’ matK_2R 5’-TAAACGATCCTCTCATTCACGA-3’	Wang et al. (1999)
rbcL (Ribulose1,5-biphosphate carboxylase oxygenase large subunit)	rbcL-af 5'- ATGTCACCACAAACAGAAAC-3 rbcL-724r 5'- TCGCATGTACCTGCAGTAGC-3	Kress and Erickson, 2007; Fay et al., 1997
rpoC1 (RNA polymerase γ-subunit gene)	rpoC1-F 5'-GGCAAAGAGGGAAGATTTCG-3 rpoC1-R 5'- CCATAAGCATATCTTGAGTTGG-3	Sass et al. 2007

trnH- psbA (Chloroplast intergenic spacer region)	psbA03_F 5'-GTTATGCATGAACGTAATGCTC-3 trnH-05_R 5'-CGCGCATGGTGGATTCACAATCC-3	Sang et al., 1997; Tate and Simpson, 2003
atpF-atpH (chloroplast intergenic spacer region)	atpF-F 5’-ACTCGCACACACTCCCTTTCC-3’ atpH-R 5’-GCTTTTATGGAAGCTTTAACAAT-3’	Lahaye et al. (2008)
psbK-psbI (chloroplast intergenic spacer region)	psbK-F 5’-TTAGCCTTTGTTTGGCAAG-3’ PsbI-R 5’-AGAGTTTGAGAGTAAGCAT-3’	Lahaye et al. (2008)
accD (Carboxytransferase-β-subunit)	accD-F 5’-AGTATGGGATCCGTAGTAGG-3’ AccD-R 5’-TTTAAAGGATTACGTGGTAC-3’	Sass et al. (2007)
rpoB (RNA polymerase β-subunit gene)	rpoB-F 5’-AAGTGCATTGTTGGAACTGG-3’ RpoB-R 5’-CCGTATGTGAAAAGAAGTATA-3’	Sass et.al. 2007
ndhJ (NADH Dehydrogenase subunit)	ndhj-F 5’-CATAGATCTTTGGGCTTYGA-3’ Ndhj-R 5’-ATAATCCTTACGTAAGGGCC-3’	Sass et.al. 2007
ycf5 (Chloroplast intergenic spacer region)	ycf5-F 5’-GGATTATTAGTCACTCGTTGG-3’ ycf5-R 5’-ACTTACGTGCATCATTAACCA-3’	Sass et.al. 2007

MaturaseK (matK)

The matK coding region is one of the most rapidly evolving regions in chloroplasts and shows a high level of species discrimination among angiosperms (Fazekas et al., 2008; Lahaye et al., 2008).The advantages of this gene are that it is easy to amplify, sequencing and alignment in most land plants and is a good DNA barcoding region for plants at the family and genus levels. Although the matK region is useful to determine species identity and the geographical origin of medicinal herbs, the success rate for the amplification and sequencing of matK region of some plant groups, such as cryptogams, is unsatisfactory and the universality of the amplification primers requires improvement (CBOL Plant Working Group, 2009). However, there are a few reports that some of the barcodes are universally useful for plants, it still remains mandatory to screen out the suited barcode for any new species (Rubinoff et al., 2006; Pennisi, 2007; Ledford, 2008). In general, the genes used in angiosperms are matK*,* rpoC1, rpoB, accD ,YCF5 and ndhJ whereas in non-angiosperms matK, rpoC1, rpoB, accD, and ndhJ are used (http://www.rbgkew.org.uk/barcoding/index.html). With higher potential to identify the variation, easy amplification and alignment, a portion of the plastid matK gene was proposed as a universal DNA barcode for flowering plants (Lahaye et al*.,* 2008). The choice of rbcL+matK as a core barcode was based on the straightforward recovery of the rbcL region and the discriminatory power of the matk region. The matK gene is one of the most rapidly evolving coding sections of the plastid genome (Hilu and Liang, 1997). Studies by Newmaster et al. (2008) in Myristicaceae and Seberg and Petersen (2009) in *Crocus* have confirmed matK and the intergenic spacer trnH-psbA as suitable land plant barcodes. The matK gene has a high evolutionary rate, suitable length and obvious interspecific divergence as well as a low transition/transversion rate (Min and Hickey, 2007; Selvaraj et al., 2008). But the matK is difficult to amplify universally using currently available primer sets. The CBOL Plant Working Group (2009) revealed nearly 90% success rate in amplifying angiosperm DNA using a single primer pair. However, the success was limited in gymnosperms (83%) and much worse in cryptogams (10%) even with multiple primer sets. The matK gene can discriminate more than 90% of species in the Orchidaceae (Kress and Erickson, 2007) but less than 49% in the nutmeg family (Newmaster et al., 2008). Fazekas et al*.* (2008) attempted the identification of 92 species from 32 genera using the matK barcode but only achieved a success rate of 56%. These findings demonstrate that the matK barcode alone is not a suitable universal barcode.

Ribulose 1,5-biphosphate carboxylase oxygenase large subunit (rbcL)

The large subunit of ribulose-bisphosphate carboxylase, rbcL region is a chloroplast gene coding region that has a high amplification success rate in a broad range of flowering plant, gymnosperm and cryptogam species, plus high sequence quality among seven loci tested (CBOL Plant Working Group, 2009). However, the rbcL region showed the lowest divergence (0.83%) among 11 potential barcoding loci tested for the differentiation of two species in Solanaceae, (Kress et al., 2005). Low interspecific variation was also observed between other herbal medicinal materials and their adulterants. However, rbcL sequences evolve slowly and this locus has by far the lowest divergence of plastid genes in flowering plants (Kress et al., 2005). Consequently, it is not suitable at the species level due to its modest discriminatory power (Fazekas et al., 2008; Lahaye et al., 2008; CBOL Plant Working Group, 2009; Chen et al., 2010). Despite these limitations, rbcL was still suggested as one of the best potential candidate plant DNA barcodes based on the straightforward recovery of the gene sequence (Blaxter, 2004; CBOL Plant Working Group, 2009; Hollingsworth et al., 2011). Although rbcL by itself does not meet the desired attributes of a DNA barcoding locus, it is accepted that rbcL in combination with various plastid or nuclear loci can make accurate identifications (Newmaster et al., 2006; Chase et al., 2007; Kress and Erickson, 2007; CBOL Plant Working Group, 2009; Hollingsworth et al., 2009). CBOL demonstrated that the use of seven candidate loci did not significantly improve species-level discriminatory ability compared to rbcL + matK. Thus, the combinations of candidate loci cannot eliminate the inherent deficiencies of current DNA barcoding of plants.

Nuclear barcode marker (ITS)

A variety of loci have been suggested as DNA barcodes for plants, including coding genes and non-coding spacers in the nuclear and plastid genomes (Figure 2). The internal transcribed spacer (ITS) region comprises the ITS1 intergenic spacer, 5.8S rDNA, and the ITS2 intergenic spacer (ITS1-5.8S-ITS2), with size ranging from 400 to 1000 bp in total. This is the most frequently sequenced region for plant phylogenetic studies because of its high species discrimination and technical ease of amplification (Alvarez and wendel, 2003; Kress et al., 2005). Although the ITS region and ITS2 intergenic spacer can help identify herbal medicinal materials by DNA sequencing, these regions sometimes require cloning because of the presence of multiple copies

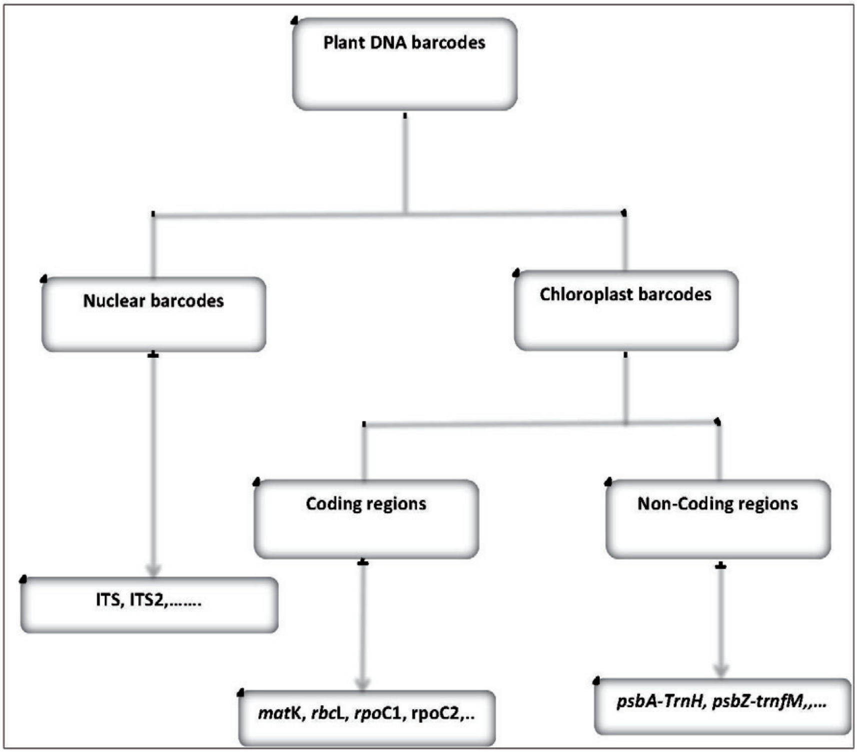


Fig. 2: Schematic illustration of employed DNA barcode markers

-and the problems of secondary structure resulting in poor-quality sequence data (Baldwin et al., 1995; Alvarez and Wendel, 2003). Fungal -contamination is common in herbal medicinal materials that are improperly processed and stored. Fungal ITS sequences are readily amplified using universal primers, generating false-positive PCR results. To overcome this issue, plant-specific primers need to be designed (Zhang et al., 1997; Cullings and Vogler, 1998). The greater discriminatory power of ITS over plastid regions at low taxonomic levels has been widely studied leading to it also being suggested as a plant barcode (Stoeckle, 2003; Kress et al., 2005; Sass et al., 2007), especially in parasitic plants which offer less resolution from plastid barcodes (Hollingsworth et al., 2011). However, CBOL has only regarded ITS as a supplementary locus (CBOL Plant Working Group, 2009). Some limitations prevent it from being a core barcode: incomplete concerted evolution, fungal contamination and difficulties of amplification and sequencing (Hollingsworth et al., 2011). Plant BOL Group recently argued that when direct sequencing was possible, the ITS region should be incorporated into the core barcodes because of higher discriminatory power than plastid barcodes (CBOL Plant Working Group, 2011). To resolve the difficulties involved in sequencing the entire ITS, they suggested ITS2 as a backup because of its conserved sequence characters which reduce amplification and sequencing problems. It was accepted that ITS2 could be used as a novel universal barcode for the identification of a broader range of plant taxa (Chen et al., 2010; Gao et al., 2010a,b; Pang et al., 2010) even from herbarium specimens with degraded DNA (Chiou et al., 2007). Song et al. (2012) recently showed that the ITS2 intra-genomic distances were markedly smaller than those of the intra-specific or inter-specific variants in a wide range of plant families. Internal transcribed spacer regions of nuclear ribosomal DNA (ITS) is commonly recommended based on the facts that these are often highly variable in angiosperms at the generic and species level and divergent copies are often present within single individuals (Kress et al., 2005). Although ITS works well in many plant groups and may be a useful supplementary locus, numerous cases of incomplete concerted evolution and intra-individual variation make it unsuitable as a universal plant barcode.

TrnH-psbA spacer

TrnH-psbA is currently the most widely used plastid DNA barcode marker. The size of the trnH–psbA region of most flowering plants ranges between 340 and 660 bp. This region shows the highest amplification success rate (100%) and discrimination rate (83%) among nine loci tested (Kress et al., 2005; Kress and Erickson, 2007). Therefore, this intergenic spacer appears to be a useful region for the differentiation of medicinal plants from their adulterants. The presence of highly conserved coding sequences on both sides make the design of universal primers feasible (Shaw et al., 2005), with a single primer pair likely to amplify nearly all angiosperms (Shaw et al., 2007). The non-coding intergenic region exhibits most sequence divergence and has high rates of insertion/deletion (Kress and Erickson, 2007). These attributes make trnH-psbA highly suitable as a plant barcode for species discrimination (Kress and Erickson, 2007; Shaw et al., 2007).

Alignment of the trnH-psbA spacer can be highly ambiguous because of its complicated molecular evolution, considerable length variation (Chang et al., 2006), and high rates of insertion/deletion in larger families of angiosperms (Chase et al., 2007). Furthermore, due to the presence of duplicated loci and a pseudogene, the trnH-psbA sequence is much longer >1000 base pairs in some conifers and monocots (Chase et al., 2007; Hollingsworth et al., 2009) while it is exceedingly short, less than 300 bp, in other groups (Kress et al., 2005) and shorter than 100 bp in Bryophytes (Stech and Quandt, 2010). One of the key problems associated with the use of trnH-psbA as a standard barcode is the frequent inversions in some plant lineages, which may lead to large overestimates of genetic divergence and to incorrect phylogenetic assignment (Whitlock et al., 2010). Additionally, because of the premature termination of sequencing reads caused by mononucleotide repeats, longer trnH-psbA regions can be difficult to retrieve without taxon-specific internal sequencing primers designed to obtain high-quality bi-directional sequences (Devey et al., 2009; Ebihara et al., 2010). Shorter trnH-psbA spacers may not have adequate sequence variation for species discrimination. As a consequence, Kress et al. (2005) and Chase et al. (2007), respectively, proposed that trnH-psbA can be used in two-locus or three-locus barcode systems to provide adequate resolution. Kress et al. (2005) also proposed that the trnH-psbA plastid inter-genic spacer region would be a suitable universal barcode for land plants.

Multilocus plant DNA barcoding approaches

Despite extensive efforts to identify a universal plant DNA barcode comparable to CO1 in animals, the task has proved difficult due to the lack of adequate variation within single loci (Kress et al., 2005; Newmaster et al., 2006; Chase et al., 2007; Kress and Erickson, 2007; Sass et al., 2007; Fazekas et al., 2008; Lahaye et al., 2008). Many researchers have suggested that a multi-locus method will be required to obtain adequate species discrimination (Hebert et al., 2004; Kress and Erickson, 2007; Erickson et al., 2008; Kane and Cronk, 2008; Lahaye et al., 2008; CBOL Plant Working Group, 2009; Chase and Fay, 2009). Various combinations of plastid loci have been proposed including rbcL + trnH-psbA (Kress and Erickson, 2007), rpoC1 + rpoB + matK or rpoC1 + matK + trnH-psbA (Chase et al., 2007) and matK + atpF-atpH + psbK-psbI or matK + atpF-atpH + trnH-psbA (Pennisi, 2007). These combined barcodes exhibit higher species discrimination than single-locus approaches. Different research groups have tested different combinations using different taxa while attempting to select a universal barcode, however universal agreement is yet to be reached. Fazekas et al. (2008) compared these barcode combinations using the same large-scale taxonomic samples, but none could identify more than 70% of tested species. de Boer et al. (2014) demonstrated that combining psbA-trnH, rpoC1, and ITS allowed the majority of the market samples to be identified to species level. Taberlet et al. (2007) proposed the chloroplast trnL intron as a potential barcoding candidate gene in angiosperms. Further, three regions atpF-atpH, matK, and psbK-psbI were proposed (Pennisi, 2007). Devey et al. (2009) argued that non-coding regions atpF-atpH and trnH-psbA should be considered as suboptimal barcodes due to occurrence of microsatellites. It was also proposed that, because the plastid genome is evolving so slowly relative to other genomes and shows intra-molecular recombination (Mower et al., 2007), more than one barcode may be necessary to provide enough variation for this technique to work (Newmaster et al., 2006; Chase et al.,2007; Kress and Erickson 2007). Kress and Erickson (2007) proposed to combine the original trnH-psbA barcode with rbcL. This combination is also potential enough to be used as a universal barcode. Hollingsworth et al. (2009) evaluated the seven main candidate plastid regions rpoC1, rpoB, rbcL, matK, trnH -psbA, atpF-atpH and psbK-psbI in three divergent groups of land plants. rpoC1 was the most universal locus and amplified well across all three groups. Chase et al. (2007) proposed to make the universal barcodes with the combination rpoC1+ rpoB+ matK and rpoC1+ matK+ trnH-psbA. Of late, the two-locus combination of rbcL+ matK has been recommended as the core barcode for land plants (CBOL Plant Working Group, 2009). In bryophytes, five loci, rbcL, rpoC1, rps4, trnH-psbA and trnL-trnF were easy to amplify and sequence and showed significant inter-specific genetic variability, making them potentially useful DNA barcodes

Chloroplast genome as a ‘super-barcode’

The first cp-genome was sequenced in 1986 (Shinozaki et al., 1986); by 2012 there were 254 complete plant cp-genomes within public databases, which only accounts for less than 0.01% of total plant species and is still a small number for widespread species identification. The feasibility of using the chloroplast genome (cp-genome) as a ‘super-barcode’ is evaluated, and the concept of a ‘specific barcode’ derived from the comparison between plastid genome sequences from a target group of taxa is presented as an effective option that might be widely applicable to plant identification studies. It has recently been pointed out that the complete cp-genome contained as much variation as the CO1 locus in animals and may be used as a plant barcode (Kane and Cronk, 2008). The complete cp-genome has a conserved sequence ranging from 110 to 160 kbp, greatly exceeding the length of commonly used DNA barcodes and providing more variation to discriminate closely related plants. The cp-genome has been used as a versatile tool for phylogenetics. It can greatly increase resolution at lower taxonomic levels in plant phylogenetic, phylogeographic and population genetic analyses, facilitating the recovery of lineages as monophyletic, and was therefore proposed as a species-level DNA barcode (Parks et al., 2009). Using the cp-genome as a marker circumvents possible issues with gene deletion and low PCR efficiency (Huang et al., 2005). The analysis of this super-barcode also resolves the problems of sequence retrieval usually encountered in traditional barcoding studies. Although sequences from single or multiple chloroplast and nuclear genes have been useful for differentiating species, the cp-genome has been used efficiently to distinguish between closely related species (Parks et al., 2009; Nock et al., 2011), populations (Doorduin et al., 2011) and individuals (Kane et al., 2012; McPherson et al., 2013). This approach is still relatively controversial, Nevertheless plastid-genome-based species classification and identification have been progressively more accepted by taxonomists (Shendure and Ji, 2008; Kumar et al., 2009; Wu et al., 2010; Bayly et al., 2013; Yang et al., 2013). The main challenges of super-barcoding are the establishment of a rich cp-genome database and the reduction of sequencing cost, as well as obtaining a higher quality and quantity of DNA (Kane et al., 2012). As sequencing technology and bioinformatics continue to improve rapidly, complete plastome sequencing will become more popular and may eventually replace Sanger-based DNA barcoding (Bayly et al., 2013; Yang et al., 2013).

Significance of plant DNA barcoding

The main goals of DNA barcoding are to assign unknown specimens to species and to enhance the discovery of new species and facilitate identification, of other organisms with complex or inaccessible morphology (Hebert et al., 2003). In three important situations, relevant species identification must necessarily be molecular-based. First is in the determination of the taxonomic identity of damaged specimens or fragments. The DNA barcoding tool is thus potentially useful in the food industry, diet analyses and in preventing illegal trade and poaching of endangered species. Second, molecular-based identification is necessary when there are no obvious means to match adults with immature specimens. The third case is when morphological traits do not clearly discriminate species especially if species have polymorphic life cycles. DNA barcoding can also be used for a wide range of purposes: to support ownership or intellectual property rights (Stewart, 2005); to reveal cryptic species (Hebert et al., 2004); in forensics to link biological samples to crime scenes (Yoon, 1993; Coyle et al., 2005; Mildenhall, 2006); to support food safety and authenticity of labelling by confirming identity or purity (Galimberti et al., 2012; Huxley-Jones et al., 2012); and in ecological and environmental genomic studies (Valentini et al., 2009).

In summary, the purpose of the DNA barcoding is to rapidly assemble a precise and representative reference library; the reference library will become increasingly useful, enabling the rapid identification of low taxonomic level taxa with specific short-DNA sequences DNA barcoding aims to find a single sequence to identify all species. Yet, no single-locus barcode can achieve the goal. In addition to inadequate variation and low PCR efficiency (often due to sequence variation in the primer binding regions), gene deletion is an important limiting factor for single loci preventing their use as a universal DNA barcode (algae do not contain the matK sequence). Multi-locus markers have been assumed to be more successful in species identification, but studies to date demonstrated that these are also inadequate for universal plant identification. Whole-plastid-based barcodes have shown great potential in species discrimination, especially for closely related taxa. Continuing advances in sequencing technology may make these super-barcodes the method of choice for plant identification.

References

Alvarez, I. and Wendel, J.F. (2003) Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29: 417–434.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S. and Donoghue, M.J. (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Annals of the Missouri Botanical Garden 82: 247–277.

Bayly, M.J., Rigault, P., Spokevicius, A., Ladiges, P.Y., Ades, P.K., Anderson, C., Bossinger, G., Merchant, A., Udovicic, F. and Woodrow, I.E. (2013) Chloroplast genome analysis of Australian eucalypts– *Eucalyptus, Corymbia, Angophora, Allosyncarpia* and *Stockwellia* (Myrtaceae). Molecular Phylogenetics and Evolution 69: 704–716.

Blaxter, M.L. (2004) The promise of a DNA taxonomy. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences 359: 669–679.

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America 106: 12794–12797.

CBOL Plant Working Group (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proceedings of the National Academy of Sciences of the United States of America 108: 19641–19646.

Chang, C.C., Lin, H.C., Lin, I.P., Chow, T.Y., Chen, H.H., Chen, W.H., Cheng, C.H., Lin, C.Y., Liu, S.M. and Chaw, S.M. (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Molecular Biology and Evolution 23: 279–291.

Chase, M.W. and Fay, M.F. (2009) Barcoding of plants and fungi. Science 325: 682–683.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., van den Berg, C., Madrinan, S., Petersen, G., Seberg, O., Jorgsensen, T., Cameron, K.M., Carine, M., Pedersen, N., Hedderson, T.A.J., Conrad, F., Salazar, G.A., Richardson, J.E., Hollingsworth, M.L., Barraclough, T.G., Kelly, L. and Wilkinson, M. (2007) A proposal for a standardised protocol to barcode all land plants. Taxon 56: 295–299.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. and Leon, C. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5: e8613.

Chiou, S.J., Yen, J.H., Fang, C.L., Chen, H.L. and Lin, T.Y. (2007) Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. Planta Medica 73: 1421–1426.

Condit, R. (1998) Tropical forest census plots: methods and results from Barro Colorado Island. Panama and a comparison with other plots. Springer-Verlag, Berlin.

Coyle, H.M., Lee, C., Lin, W., Lee, H.C. and Palmbach, T.M. (2005) Forensic botany: using plant evidence to aid in forensic death investigation. Croatian Medical Journal 46: 606–612.

Cullings, K.W. and Vogler, D.R. (1998) A 5.8S nuclear ribosomal RNA gene sequence database: applications to ecology and evolution. Molecular Ecology 7: 919–923.

de Boer, H.J., Ouarghidi, A., Martin, G., Abbad, A. and Kool, A. (2014) DNA barcoding reveals limited accuracy of identifications based on folk taxonomy. PLoS ONE 9: e84291

Demesure, B., Comps, B. and Petit, R.J. (1996) Chloroplast DNA phylography of the common beeh (*Fagus sylvatica* L.) in Europe. Evolution 50: 1515-2520.

Demesure, B., Sodzi, N. and Petit, R.J. (1995) A set of universal primers for amplification of polymorphic noncoding regions of mitochondrial and chlrorplast DNA in plants. Molecular Evolution 4: 129-131.

Devey, D.S., Chase, M.W. and Clarkson, J.J. (2009) A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. Taxon 58: 7–15.

Dick, C.W. and Heuertz, M. (2008) The complex biogeographic history of a widespread tropical tree species. Evolution 62: 2760–2774

Dick, C.W. and Kress, W.J. (2009) Dissecting tropical plant diversity with forest plots and a molecular toolkit. Bioscience 59:745–755

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T. and Vrieling, K. (2011) The complete chloroplast genome of 17 individuals of pest species jacobaea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Research 18: 93–105.

Ebihara, A., Nitta, J.H. and Ito, M. (2010) Molecular species identification with rich floristic sampling: DNA barcoding the pteridophyte flora of Japan. PLoS ONE 5: e15136.

Erickson, D.L., Spouge, J., Resch, A., Weigt, L.A. and Kress, J.W. (2008) DNA barcoding in land plants: developing standards to quantify and maximize success. Taxon 57: 1304–1316.

Fay, M.F., Swensen, S.M. and Chase, M.W. (1997) Taxonomic affinities of *Medusagyne oppositifolia* (Medusagynaceae). Kew Bulletin 52: 111–120.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M. and Barrett, S.C.H. (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3: e2802.

Fine, P.V.A. and Ree, R.H. (2006) Evidence for a time integrated species-area effect on the latitudinal gradient in tree diversity. The American Naturalist 168:796–804.

Galimberti, A., De Mattia. F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martellos, S. and Labra, M. (2012) DNA barcoding as a new tool for food traceability. Food Research International 50: 55–63.

Gao, T., Yao, H., Song, J.Y., Liu, C., Zhu, Y.J., Ma, X.Y., Pang, X.H., Xu, H.X. and Chen, S.L. (2010a) Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. The Journal of Ethnopharmacology 130: 116-121.

Gao, T., Yao, H., Song, J.Y., Zhu, Y.J., Liu, C. and Chen, S.L. (2010b) Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. BMC Evolutionary Biology 10: 324.

Gonzalez, M.-A., Baraloto, C., Engel, J., Mori, S.A., Pétronelli, P., Riéra, B., Roger, A., Thébaud, C. and Chave, J. (2009) Identification of Amazonian trees with DNA barcodes. PLOS Biology 4:e7483

Gregory, T.R. (2005) DNA barcoding does not compete with taxonomy. Nature 434: 1067.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and De Waard, J.R. (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London Series B 270: 313–321.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. and Hallwachs, W. (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proceedings of the National Academy of Sciences of the United States of America 101: 14812–14817.

Hilu, K.W. and Liang, H. (1997) The matK gene: Sequence variation and application in plant systematics. American Journal of Botany 84:830-839.

Hollingsworth, M.L., Clark, A., Forrest, L.L., Richardson, J., Pennington, T.R., Long, D.G., Cowan, R., Chase, M.W., Gaudeul, M. and Hollingsworth, P.M. (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources 9: 439–457.

Hollingsworth, P.M., Graham, S.W. and Little, D.P. (2011) Choosing and using a plant DNA barcode. PLoS ONE 6: e19254.

Huang, C.Y., Grünheit, N., Ahmadinejad, N., Timmis, J.N. and Martin, W. (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiology 138: 1723–1733.

Huxley-Jones, E., Shaw, J.L., Fletcher, C., Parnell, J. and Watts, P.C. (2012) Use of DNA barcoding to reveal species composition of convenience seafood. Conservation Biology 26: 367–371.

IUCN (2012) The IUCN red list of threatened species, Version 2012.2. Available at http://www.iucnredlist.org Accessed 25.03.2013.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J.Y., Zhang, D., Engels, J.M.M. and Cronk, Q. (2012) Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. American Journal of Botany 99: 320–329.

Kane, N.C. and Cronk, Q. (2008) Botany without borders: barcoding in focus. Molecular Ecology 17: 5175–5176.

Kress, W.J. (2004) Plant floras: how long will they last? A review of flowering plants of the Neotropics. American Journal of Botany 91: 2124–2127

Kress, W.J. and Erickson, D.L. (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS ONE 2: e508.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A. and Janzen, D.H. (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences 102: 8369–8374.

Kumar, S., Hahn, F.M., McMahan, C.M., Cornish, K. and Whalen, M.C. (2009) Comparative analysis of the complete sequence of the plastid genome of *Parthenium argentatum* and identification of DNA barcodes to differentiate *Parthenium* species and lines. BMC Plant Biology 9: 131–142.

Lahaye, R., Van Der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G. and Savolainen, V. (2008) DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences USA 105: 2923–2928.

Ledford H (2008) Botanical identities: DNA barcoding for plants comes a step closer. Nature 415: 616.

McPherson, H., van der Merwe, M., Delaney, S.K., Edwards, M.A., Henry, R.J., McIntosh, E., Rymer, P.D., Milner, M.L., Siow, J. and Rossetto, M. (2013) Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecology 13: 8.

Mildenhall, D. (2006) *Hypericum* pollen determines the presence of burglars at the scene of a crime: an example of forensic palynology. Forensic Science International 163: 231–235.

Min, X.J. and Hickey, D.A. (2007) Assessing the effect of varying sequence length on DNA barcoding of fungi. Molecular Ecology Notes 7: 365–373.

Mower, J., Touzet, P., Gummow, J., Delph, L. and Palmer, J. (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evolutionary Biology 7: 135.

Newmaster, S.G., Fazekas, A.J. and Ragupathy, S. (2006) DNA barcoding in land plants: Evaluation of rbcL in a multigene tiered approach. Canadian Journal of Botany 84: 335–341.

Newmaster, S.G., Fazekas, A.J., Steeves, R.A.D. and Janovec, J. (2008) Testing candidate plant barcode regions in the Myristicaceae. Molecular Ecology Resources 8: 480–490.

Nock, C.J., Waters, D.L., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M. and Henry, R.J. (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal 9: 328–333.

Pang, X.H., Song, J.Y., Zhu, Y.J., Xu, H.X., Huang, L.F. and Chen, S.L. (2010) Applying plant DNA barcodes for Rosaceae species identification. Cladistics 26: 1-6.

Parks, M., Cronn, R. and Liston, A. (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7: 84–100.

Pennisi E (2007) Wanted: a barcode for plants. Science 318: 190–191.

Ratnasingham, S. and Hebert, P.D.N. (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes 7: 355-364.

Rubinoff, D., Cameron, S. and Will, K. (2006) Are plant DNA barcodes a search for the Holy Grail? Trends in Ecology & Evolution 21: 1–2.

Sang, T., Crawford, D.J. and Stuessy, T.F. (1997) Chloroplast DNA phylogeny, reticulate evolution and biogeography of *Paeonia* (Paeoniaceae). American Journal of Botany 84: 1120–1136.

Sass, C., Little, D.P., Stevenson, D.W. and Specht, C.D. (2007) DNA Barcoding in the Cycadales: Testing the potential of proposed barcoding markers for species identification of cycads. PLoS ONE 2: e1154.

Seberg, O. and Petersen, G. (2009) How many loci does it take to DNA barcode a *Crocus*? PLoS ONE 4: e4598.

Selvaraj, D., Sarma, R.K. and Sathishkumar, R. (2008) Phylogenetic analysis of chloroplast matK gene from Zingiberaceae for plant DNA barcoding. Bioinformation 3: 24–27.

Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Siripun, K.C., Winder, C.T., Schilling, E.E. and Small, R.L. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany 92: 142–166.

Shaw, J., Lickey, E.B., Schilling, E.E. and Small, R.L. (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany 94: 275–288.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. Nature Biotechnology 26: 1135–1145.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayshida, N., Matsubayasha, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kata, A., Tohdoh, N., Shimada, H. and Sugiura, M. (1986) The complete nucleotide sequence of the tobacco chloroplast genome. Plant Molecular Biology Reporter 4: 111–148.

Song, J., Shi, L., Li, D., Sun, Y., Niu, Y., Chen, Z., Luo, H., Pang, X., Sun, Z., Liu, C., Lv, A., Deng, Y., Larson-Rabin, Z., Wilkinson, M.

and Chen, S. (2012) Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. PLoS ONE 7: e43971.

Stech, M. and Quandt, D. (2010) 20,000 species and five key markers: the status of molecular bryophyte phylogenetics. Phytotaxa 9: 196–228.

Stewart, C.N. Jr (2005) Monitoring the presence and expression of transgenes in living plants. Trends in Plant Science 10: 390–396.

Stoeckle, M. (2003) Taxonomy, DNA, and the barcode of life. Bioscience 53: 796–797.

Sucher, N.J. and Carles, M.C. (2008) Genome-based approaches to the authentication of medicinal plants. Planta Med 74: 603-623.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. and Willerslev, E. (2007) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. Nucleic Acids Research. 35: e14.

Tate, J.A. and Simpson, B.B. (2003) Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. Systematic Botany 28: 723–737.

Valentini, A., Miquel, C., Nawaz, M.A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J.E. and Taberlet, P. (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Molecular Ecology Resources 9: 51–60.

Wang, X., Tsumura, Y., Yoshimaru, H., Nagasaka, K. and Szmidt, A.E. (1999) Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast rbcL, matK, rpl20-rps18 spacer, and trnV intron sequences. American Journal of Botany 86: 1742–1753.

Webb, C.O., Slik, J.W.F. and Triono, T. (2010) Biodiversity inventory and informatics in Southeast Asia. Biodiversity Conservation 19:955–972

White, T.J., Bruns, T., Lee, S. and Taylor, J.W. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M.A., Gelfand, D.H., Sninsky, J.J., White, T.J. (Eds.) PCR protocols: A guide to methods and applications. New York: Academic Press. 315–322.

Whitlock, B.A., Hale, A.M. and Groff, P.A. (2010) Intraspecific inversions pose a challenge for the trnH-psbA plant DNA barcode. PLoS ONE 5: e11533.

Wu, F.H., Chan, M.T., Liao, D.C., Hsu, C.T., Lee, Y.W., Daniell, H., Duvall, M.R. and Lin, C.S. (2010) Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biology 10: 68–79.

Yang, J.B., Tang, M., Li, H.T., Zhang, Z.R. and Li, D.Z. (2013) Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evolutionary Biology 13: 84.

Yoon, C.K. (1993) Forensic science. Botanical witness for the prosecution. Science 260, 894–895.

Yu, J., Yan, H., Lu, Z. and Zhou, Z. (2011) Screening potential DNA barcode regions of chloroplast coding genome for *Citrus* and its related genera. Scientia Agricultura Sinica 44: 341-348.

Zhang, W., Wendel, J.F. and Clark, L.G. (1997) Bamboozled again! Inadvertent isolation of fungal rDNA sequences from bamboos (Poaceae: Bambusoideae). Molecular Phylogenetics and Evolution 8: 205–217.

*****

Plant DNA Barcoding and Phylogenetics, Pages 37-52
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

3 Nuclear Sequences in Plant Phylogenetics

M.A. Ali, A.K. Pandey, F.M.A. Al-Hemaid, J. Lee, B. Pandit, S.Y. Kim, G. Gyulai and M.O. Rahman

Introduction

Taxonomy, the science of classifying organisms, is basis for all biology. The early taxonomic systems had no theoretical basis; organisms were grouped according to apparent similarity. Since the publication in 1859 of Charles Darwin's 'On the Origin of Species by Means of Natural Selection', however, taxonomy has been based on the accepted propositions of evolutionary descent and relationship. The traditional classification of plants into respective classes, orders, families, genera and species has been based on shared morphologic, cytologic, biochemical and ecologic traits. The development of techniques in molecular biology including those for molecular hybridization, cloning, restriction endonuclease digestions and protein and nucleic acid sequencing have provided many new tools for the investigation of phylogenetic relationships. At the molecular level, the most fundamental comparison possible is of the primary nucleotide sequences of homologous genes in different populations or species (Hamby and Zimmer, 1992). Beginning in the early 1980s and continuing to the present, the use of DNA has represented the "cutting edge" (glamour area) within the entire field of plant systematics. The advancement of PCR based techniques, sequencing technologies and diverse bioinformatics tools for analysis of sequence data has taken phylogenies to a new height (Yang and Rannala, 2012); and therefore, phylogenies have permeated nearly every branch of biology. Our understanding of the relationships among organisms at various levels in the tree of life has been advanced greatly in the last two decades with the aid of DNA molecular systematic techniques and phylogenetic theory (Mitchell and Wen, 2005).

The genome (-term coined in 1920 by Hans Winkler, Professor of Botany at the University of Hamburg, Germany) is the genetic material of an organism, encoded in DNA or RNA for many types of viruses, includes both the genes and the non-coding sequences of the DNA/RNA. Genome size is the total number of DNA base pairs in one copy of a haploid genome. The plant cell contains three different genomes: chloroplast, mitochondrion and nucleus. Plant molecular systematics has relied primarily on the chloroplast genome. This is changing as investigators turn to nuclear gene sequences (ITS, 18S rDNA), often to compare nuclear topologies with existing chloroplast based topologies. An important breakthrough for plant systematist was the study of Baldwin (1992) demonstrating the utility of the internal transcribed spacer of nuclear ribosomal DNA (ITS) for resolving relationships within and among closely related genera in the Asteraceae. Not all regions of the rDNA evolve at the same rate, so, even though some regions are useful for comparisons at or below the genus level, other regions are only useful at the family level or above.

Nuclear ribosomal DNA

Nuclear ribosomal DNA is arranged in tandem repeats in one or a few chromosomal loci. Each repeat consists of a transcribed region that comprises an external transcribed spacer (ETS) followed by the 18S gene, an internal transcribed spacer (ITS-1), the 5.8S gene, a second internal transcribed spacer (ITS-2), and finally the 26S gene (Figure 1). Each such repeat is separated from the next repeat by an intergenic spacer (IGS). Only among closely related species are the chromosomal locations similar. The nuclear genes that code for rRNA are repeated thousands of times within the typical plant genome. In fact they can comprise as much as 10% of the total plant DNA. The most remarkable feature of rDNA is the overall sequence homogeneity among members of the gene family in a given species. The process by which this pattern of intraspecific homogeneity and interspecific heterogeneity is maintained has been called concerted evolution (Zimmer et al., 1980).

Despite the large size of the nuclear genome and the large number and diversity of genes that it includes, most attempts to infer phylogeny with nuclear gene sequences have involved the nuclear ribosomal DNA cistron (rDNA). The approximate lengths of the three coding regions are very similar throughout plants. The 18S gene equals 1,800 bp (Nickrent and Soltis, 1995); the 26S gene equals 3,300 bp (Bult et al., 1995); the 5.8S gene equals 160 bp (Takaiwa et al., 1985). In contrast, the length of the IGS varies considerably (from 1 to 8 kb). This variation in IGS length is the major contributors to the large range of variation in total length of the repeat unit in plants, ranging from approximately 8-15 kb.

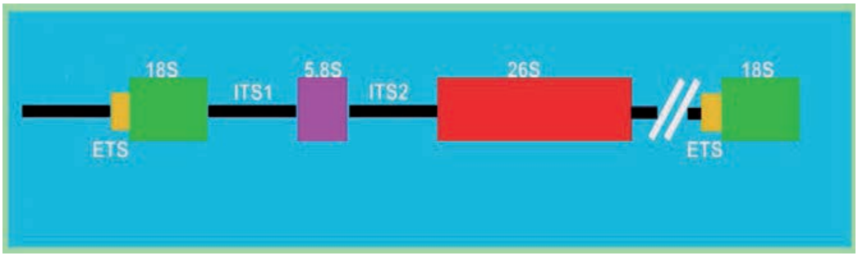


Fig. 1: A typical plant rDNA repeat. ITS-1 and ITS-2 are the two internal transcribed spacer regions. IGS is the intergenic spacer; ETS is the External transcribed spacer.

18S rDNA

18S rDNA sequences have been much more extensively used than 26S rDNA sequences. Although the general taxonomic range of application of the two regions appears to be very similar, the sheer size of the 26S gene (over 3,000 bp) has deterred investigators, particularly with regard to complete sequencing. In contrast the size of 18S rDNA (SSU; approx 1,800 bp) has made it much more amenable to PCR amplification and sequencing. The advantage of obtaining complete, rather than partial approach also provides a database of sequences for the study of molecular evolution. For example, the large data set of complete 18S rDNA sequences obtained by Soltis et al. (1997) afforded the opportunity for the first detailed analyses of molecular evolution of 18S rRNA genes in angiosperms. Only one 18S sequence type is typically found in an angiosperm, but intra individual 18S rDNA variation has been detected in some cases. Many initial studies involved portions of the 18S region; recent studies have typically used the entire 18S gene. The occasional multiple 18S sequence types do not distort phylogenetic relationships is significant in plant groups such as ferns and angiosperms, which are noted for polyploidy.

The 18S gene is a slowly evolving marker and is suitable for inferring phylogenies at higher taxonomic levels such as deep-level phylogeny of angiosperms (Hamby and Zimmer, 1992; Soltis et al., 1997), or among closely related families such as Caryophyllales in combination with other sequence data (Cuénoud et al., 2002). It was also explored for constructing the land plant phylogeny (Soltis et al., 1999). However, the most common criticism of 18S rDNA as a source of phylogenetic information have been that it is not sufficiently variable for phylogenetic reconstruction within the angiosperms and that it is highly prone to insertion and deletion, making sequence alignment difficult. But some studies have demonstrated that 18S rDNA provides a sufficient number of characters for broad scale phylogenetic reconstruction of the angiosperms (Starr et al., 2004; Gaskin et al., 2003; Rossetto et al., 2002; Les et al., 1999).

26S rDNA

The 26S rRNA (LSU; over 3,300 BP) gene is often noted as a candidate for sequencing as either an alternative or a supplement to 18S rDNA. Assessing the phylogenetic utility and molecular evolution of the entire 26S rDNA in plants has been made difficult. In searching for additional base pairs from the nucleus and elucidate higher level relationships, and with the increased use of automated sequencing, investigators have recently developed PCR and sequencing primers for the entire 26S rDNAs. In plants, the 26S is about 3.4 kb long and includes 12 expansion segments (ES) which are more variable (Bult et al., 1995). The overall nucleotide substitution rate of 26S is 1.6-2.2 times higher than that in 18S (Kuzoff et al., 1998). Their comparisons further confirm the higher level of phylogenetic potential of entire 26S rDNA sequences which evolve 1.6 to 2.2 times faster and provide over three times as many parsimony informative characters as does 18S rDNA (Forrest et al., 2005, Gaskin et al., 2003; Starr et al., 2004; Seelanan et al., 1999; Kuzoff et al., 1999; Ashworth, 2000). So far partial 26S sequences have been used among closely related families such as in the Apiales (Chandler and Plunkett, 2004), for determining phylogenetic position of isolated families (Neyland, 2002), or sometimes within a family e.g., in Celastraceae (Simmons et al., 2001).

ITS sequence of nrDNA and 5.8S gene

The nuclear genes that code for rRNA are repeated thousands of times within the typical plant genome. In fact they can comprise as much as 10% of the total plant DNA. The most remarkable feature of rDNA is the overall sequence homogeneity among members of the gene family in a given species. The process by which this pattern of intraspecific homogeneity and interspecific heterogeneity is maintained has been called concerted evolution (Zimmer et al., 1980). One of the remarkable properties of nrDNA (including ITS) genes is that their paralogs within individuals are quite homogenous, resulting from concerted evolution. The underlying molecular processes are presumed to involve unequal crossing over (Smith, 1976) and gene conversion (Nagylaki, 1984). nrDNA paralogs display polymorphisms in individuals where concerted evolution is incomplete, for example in cases where hybridization is involved (Muir et al., 2001), or where concerted evolution cannot act between paralogs effectively when they are dispersed on non-homologous chromosomes in the genome (Wei and Wang, 2004). Recently, multiple divergent ITS paralogs within individuals have been observed in several plant groups (Harpke and Peterson, 2006; Grimm and Denk, 2007; Ochieng et al., 2007; Zheng et al., 2008), which suggest incomplete concerted evolution across the repeats. Among divergent rDNA paralogs, non-functional pseudogenes are prominent, and many studies have demonstrated the existence of pseudogenes in plant genomes, where concerted evolution of nrDNA is incomplete. The pseudogenes are characterized by a higher relative substitution rate, an increased AT content, and lower secondary structure stability (Alvarez and Wendel, 2003). The pseudogenes assumed to have escaped from functional constraints, have accumulated many mutations and can cluster randomly across phylogenetic trees due to long-branch attraction (LBA), which confounds attempts to recover correct phylogenetic species relationships (e.g. Kita and Ito, 2000; Mayol and Rossello, 2001). ITS pseudogenes can potentially be useful for phylogenetic analyses of closely related species, when the functional paralogs provide too low variation (Ochieng et al., 2007).

Moreover, it is widely accepted that in the process of concerted evolution a single mutation can be fixed in a relatively short time period due to unequal crossing over or gene conversion. These homogenization processes have been described as molecular drive. The coding regions show little sequence divergence among closely related species, whereas the spacer regions exhibit higher rates of variability. Therefore, nuclear ribosomal ITS sequence data have a great potential to resolve plant phylogenies at various intrafamiliar levels in angiosperms. Despite the large size of the nuclear genome, most attempts to infer phylogeny with nuclear gene sequences have involved the nuclear ribosomal DNA cistron (rDNA). The internal transcribed spacer (ITS) regions of 18S-26S nuclear rDNA have become a major focus of comparative sequencing at the specific and generic levels in angiosperms. The nuclear ribosomal ITS region including the 5.8S gene has been the most widely used molecular marker at the interspecific and intergeneric levels in plants (Wen and Zimmer, 1996).

Since the first report of the utility of the nrDNA ITS sequences in plants (Baldwin, 1992), it is being extensively used for phylogenetic studies, molecular discrimination of raw drug material and DNA barcoding (Baldwin, 1995; Chen et al., 2010). The nrDNA ITS sequences possesses a number of valuable characteristics, such as the availability of conserved regions for designing universal primers (White et al. 1990), the ease of its amplification, short length (with ITS1 200-300 bases long, ITS2 180-240 bases, and 5.8S ca. 160 bp in flowering plants) and sufficient sequence variation which can easily able to distinguish even very closely related species (Yao et al., 2010). Additionally, the ITS2 shows significant sequence variability at the species level or lower (Coleman, 2003, 2007, 2009; Schultz et al., 2005; 2006; Thornhillet al., 2007). The availability of structural information of ITS2 permits analysis even at higher taxonomic level too (Coleman, 2003, 2007, 2009; Aguilar and Sanchez, 2007; Schultz and Wolf, 2009; Keller et al., 2010). Chen et al. (2010) proposed that ITS2 has potential for use as a standard DNA barcode to identify medicinal plants. The ITS2 region has also been shown to be applicable in discriminating among a wide range of plants genera and families e.g. Asteraceae, Rutaceae, Rosaceae and Araliaceae (Gao et al., 2010; Liu et al., 2012a,b; Luo et al., 2010; Pang et al., 2011; Yao et al., 2010). Besides plants, the ITS2 sequence also has potential for use in barcoding of animals (Yao et al., 2010). The secondary structure of ITS2 are conserved and possesses sufficient variation in primary sequences as well as secondary structure, which also provides useful biological information for alignment; therefore, the ITS2 sequences is also used as molecular morphological characteristics for species identification (Coleman, 2007; Schultz et al., 2005; Koetschan et al., 2010). Moreover, analyses of ITS2 sequences along with secondary structure results into more robust phylogeny (Keller et al., 2008). Therefore, owing to enormous phylogenetic significance, the nrDNA ITS gene is now a day considered as better than its reputation (Wolf and Schultz, 2009).

One of the remarkable properties of nrDNA (including ITS) genes is that their paralogs within individuals are quite homogenous, resulting from concerted evolution. The underlying molecular processes are presumed to involve unequal crossing over (Smith, 1976) and gene conversion (Nagylaki, 1984). The nrDNA paralogs display polymorphisms in individuals where concerted evolution is incomplete, for example in cases where hybridization is involved (Muir et al., 2001), or where concerted evolution cannot act between paralogs effectively when they are dispersed on non-homologous chromosomes in the genome (Wei and Wang, 2004). Recently, multiple divergent ITS paralogs within individuals have been observed in several plant groups (Harpke and Peterson, 2006; Grimm and Denk, 2007; Ochieng et al., 2007; Zheng et al., 2008), which suggest incomplete concerted evolution across the repeats. Among divergent rDNA paralogs, non-functional pseudogenes are prominent, and many studies have demonstrated the existence of pseudogenes in plant genomes, where concerted evolution of nrDNA is incomplete. Pseudogenes are characterized by a higher relative substitution rate, an increased AT content, and lower secondary structure stability (Alvarez and Wendel, 2003). Pseudogenes, assumed to have escaped from functional constraints, have accumulated many mutations and can cluster randomly across phylogenetic trees due to long-branch attraction (LBA), which confounds attempts to recover correct phylogenetic species relationships (e.g. Kita and Ito, 2000; Mayol and Rossello, 2001). ITS pseudogenes can potentially be useful for phylogenetic analyses of closely related species, when the functional paralogs provide too low variation (Ochieng et al., 2007).

External Transcribed Spacer Region (ETS)

The external transcribed spacer (ETS) region (especially the 3'end of the 5'-ETS adjacent to 18S) lies in the intergenic spacer region separating the repetitive 18S-5.8S-26S ribosomal gene blocks from each other. There are two ETS sites: the 3’ and 5’ prime parts which are bordering the 18S and 26S exons. This region is transcribed and plays a role in the ribosome transcription (Linder et al., 2000; Houseley et al., 2007; Granneman and Baserga, 2005, Azuma et al., 2006). The transcription termination site in the 3’ETS- as the transcription initiation site in the 5’ETS-is highly variable in plants. In the recent years a great progress was made surrounding the external transcribed spacers, revealing interesting new features about the region. The homogenization process of concerted evolution is the operating force to eliminate the different repeat types of ETS found within the genome of a single individual. However, concerted evolution is a well known and specific feature of multigene families such as the rDNA locus -the rate and even its accuracy is not well known. In general, the whole process of concerted evolution enhances the sequence similarity between multiply arrays of ITS and ETS. Since its first application by Baldwin and Markos (1998) several analyses successfully adopted this marker as a valuable phylogenetic tool. Sequence comparisons of the rDNA external transcribed spacer (ETS) indicated that it represents an even more valuable instrument for the phylogenetic analysis than ITS (King et al., 1993). The ETS has been used in phylogenetic analysis of families Asteraceae (Granneman and Baserga, 2005, Azuma et al., 2006), Fabaceae (Chandler et al. 2001) and Myrtaceae (Wright et al., 2001). The 5'ETS is more frequently used in phylogenetic studies, than the 3' part. The length of the 5'end ETS range from 425 to 575 bp (McMullen et al., 1986; Schiebel et al., 1989; Tremousaygue et al, 1992; Cordesse et al., 1993) making it easily sequenced. There are less sequences available for ETS compared to ITS.

Low-Copy Nuclear Sequences

The single copy or unique sequence or low-copy nuclear sequences are referred to DNA or nuclear genome of plants consists of certain DNA sequences that are present once per genome. The lengths of single copy sequences in plant genomes usually vary from 200 to several thousand bp. Single or low-copy nuclear genes have great potential to elucidate phylogenetic relationships of plants (Mort and Crawford, 2004; Mort et al., 2004, Schluter et al., 2005). Low-copy nuclear genes in plants are a rich source of phylogenetic information. They hold a great potential to improve the robustness of phylogenetic reconstruction at all taxonomic levels, especially where universal markers such as cpDNA and nrDNA are unable to generate strong phylogenetic hypotheses. Low-copy nuclear genes, however, remain underused in plant phylogenetic studies due to practical and theoretical complications in unraveling the evolutionary dynamics of nuclear gene families. The lack of the universal markers or universal PCR primers of low-copy nuclear genes has also hampered their phylogenetic utility. It has recently become clear that low-copy nuclear genes are particularly helpful in resolving close interspecific relationships and in reconstructing allopolyploidization in plants. Gene markers that are widely, if not universally, useful have begun to emerge. Although utilizing low-copy nuclear genes usually requires extra laboratory work such as designing PCR primers, PCR-cloning, and/or Southern blotting, rapid accumulation of gene sequences in the databases and advances in cloning techniques have continued to make such studies more feasible (Sang, 2002). The advantages of nuclear genes include the availability of many genes, their overall faster rate of evolution, and their biparental inheritance (Small et al., 2004). They also present practical difficulties such as complications associated with discerning orthologues from paralogues, concerted evolution, and recombination among paralogous sequences.

References

Ackerfield, J.R. and J. Wen (2003) Evolution of *Hedera* (the ivy genus, Araliaceae): insights from chloroplast DNA data. International Journal of Plant Sciences 164: 593-602.

Aguilar, C. and Sanchez, J.A. (2007) Phylogenetic hypotheses of gorgoniid octocorals according to ITS2 and their predicted RNA secondary structures. Molecular Phylogenetics and Evolution 43: 774–786.

Álvarez, I. and Wendel, J.F. (2003) Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29: 417–434.

Ashworth, V. (2000) Phylogenetic relationships in Phoradendreae (Viscaceae) inferred from three regions of the nuclear ribosomal cistron. I. Major lineages and paraphyly of Phoradendron. Systematic Botany 25(2): 349-370.

Avise, J.C. (1994) Molecular markers, natural history and evolution. Chapman & Hall, New York.

Azuma, M., Toyama, R., Laver, E. and Dawid, I.B. (2006) Perturbation of rRNA synthesis in the bap28 mutation leads to poptosis mediated by p53 in zebrafish central nervous system. The Journal of Biological Chemistry 281(19):13309–13316.

Baldwin, B.G. (1992) Phylogentic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. Molecular Phylogenetics and Evolution 1: 3-16.

Baldwin, B.G. and Markos, S. (1998) Phylogenetic utility of the external transcribed spacer (ETS) of 18S–26S rDNA: congruence of ETS and ITS trees of Calycadenia (Compositae). Molecular Phylogenetics and Evolution 10: 449–463.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S. and Donoghue, M.J. (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Annals of the Missouri Botanical Garden 82: 247-277.

Bult, C.J., Sweere, J.A., and Zimmer, E.A. (1995) Cryptic sequence simplicity, nucleotide composition bias, and molecular coevolution in the large subunit of ribosomal DNA in plants: Implications for phylogenetic analyses. Annals of the Missouri Botanical Garden 82: 235-246.

Chandler, G.T. and Plunkett, G.M. (2004) Evolution in Apiales: Nuclear and chloroplast markers together in (almost) perfect harmony. Botanical Journal of the Linnean Society 144: 123-147.

Chandler, G.T., Bayer, R.J. and Crisp, M.D. (2001) A molecular phylogeny of the endemic Australian genus Gastrolobium

(Fabaceae: Mirbelieae) and allied genera using chloroplast and nuclear markers. American Journal of Botany 88:1675–1687.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. and Leon, C. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5(1): e8613.

Coleman, A.W. (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. Trends in Genetics 19:370–375.

Coleman, A.W. (2007) Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. Nucleic Acids Research 35:3322–3329.

Coleman, A.W. (2009) Is there a molecular key to the level of ‘‘biological species’’ in eukaryotes? A DNA guide. Molecular Phylogenetics and Evolution 50:197–203.

Cordesse, F., Cooke, R., Tremousaygue, D., Greellet, F. and Delseny, M. (1993) Fine structure and evolution of the rDNA intergenic spacer in rice and other cereals. Journal of molecular evolution 36: 369–379.

Cuénoud, P., Savolainen, V., Chatrou, L.W., Powell, M., Grayer, R.J. and Chase, M.W. (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. American Journal of Botany 89:132-144.

Forrest, L.L., Hughes, M. and Hollingsworth, P.M. (2005) A phylogeny of *Begonia* using nuclear ribosomal sequence data and morphological characters. Systematic Botany 39(3): 671-682.

Gao, T., Yao, H., Song, J., Zhu, Y., Liu, C. and Chen, S. (2010) Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. BMC Evolutionary Biology 10: 324.

Gaskin, J.F. and Schaal, B.A. (2003) Molecular phylogentic investigation of U. S. invasive *Tamari. S*ystematic Botany 28(1): 86-95.

Granneman, S. and Baserga, S.J. (2005) Crosstalk in gene expression: coupling and co-regulation of rDNA transcription, pre-ribosome assembley and pre-rRNA processing. Current opinion in cell biology 17: 281–286.

Grimm, G.W. and Denk, T. (2007) ITS evolution in *Platanus* (Platanaceae): homoeologues, pseudogenes and ancient hybridization. Annals of Botany 101: 403–419.

Hamby, R.K. and Zimmer, E.A. (1992) Ribosomal RNA as a phylogenetic tool in plant systematics. In: Soltis, P.S., Soltis, D.E. and Doyle, J.J. (Eds.), Molecular systematics of plants. Chapman & Hall, New York. pp. 50-91.

Harpke, D. and Peterson, A. (2006) Non-concerted ITS evolution in *Mammillaria* (Cactaceae). Molecular Phylogenetics and Evolution 41: 579–593.

Houseley, J., Kotivic, K., Hage, A.E. and Tollervey, D. (2007) Trf4 targets ncRNAs from telomeric and rDNA spacer regions and functions in rDNA copy number control. EMBO J 26: 4996–5006

Keller, A., Forster, F., Muller, T., Dandekar, T., Schultz, J. and Wolf, M. (2010) Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. Biology Direct 5:4.

Keller, A., Schleicher, T., Förster, F., Ruderisch, B., Dandekar, T., Müller, T. and Wolf, M. (2008) ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales). BMC Evolutionary Biology 8: 218.

King, K., Torres, R.A., Zentgraf, U. and Hembleben, V. (1993) Molecular evolution of the intergeneric spacer in the nuclear ribosomal RNA genes of Cucurbitaceae. Journal of Molecular Evolution 36: 144–183.

Kita, Y. and Ito, M. (2000) Nuclear ribosomal ITS sequences and phylogeny in East Asian *Aconitum* subgenus *Aconitum* (Ranunculaceae), with special reference to extensive polymorphism in individual plants. Plant Systematics and Evolution 225: 1–13.

Koetschan, C., Forster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Müller, T., Wolf, M. and Schultz, J. (2010) The ITS2 database III-sequences and structures for phylogeny. Nucleic Acids Research 38:D275-D279.

Kuzoff, R.K., Soltis, D.E., Hufford, L. and Soltis, P.S. (1999) Phylogentic relationships within *Lithophragma* (Saxifragaceae) hybridization, allopolyploidy, and ovary diversification. Systematic Botany 24(4): 598-615.

Kuzoff, R.K., Sweere, J.A., Soltis, D.E., Soltis, P.S. and Zimmer, E.A. (1998) The phylogenetic potential of entire 26S rDNA sequences in plants. Molecular Biology and Evolution 15: 251-263.

Les, D.H., Schheider, E.L., Padgett, D.J., Soltis, P.S., Soltis, D.E. and Zanis, M. (1999) Phylogeny, classification and floral evolution of water lilies (Nympheaceae; Nymphaeales): a synthesis of non-molecular rbcL, matK and 18S rDNA data. Systematic Botany 24(1): 28-46.

Linder, C.R., Goertzen, L.R., Heuval, B.V., Francisco-Ortega, J. and Jansen, R.K. (2000) The complete external transcribed spacer of 18S-26S rDNA: amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. Molecular Phylogenetics and Evolution 14: 285-303.

Liu, Z., Zeng, X., Yang, D., Chu, G., Yuan, Z. and Chen, S. (2012b) Applying DNA barcodes for identification of plant species in the family Araliaceae. Gene 499(1):76-80.

Liu, Z., Zeng, X., Yang, D., Ren, G., Chu, G., Yuan, Z., Luo, K., Xiao, P. and Chen, S. (2012a) Identification of medicinal vines by ITS2 using complementary discrimination methods. Journal of Ethnopharmacology 141(1):242-249.

Luo, K., Chen, S., Chen, K., Song, J., Yao, H., Ma, X., Zhu, Y., Pang, X., Yu, H., Li, X. and Liu, Z. (2010) Assessment of candidate plant DNA barcodes using the Rutaceae family. Science China Life Sciences 53(6):701-708.

Matthias, W. and Schultz, J. (2009) ITS better than its reputation *Science* (E-Letter, 10 December 2009).

Mayol, M. and Rosselló, J.A. (2001) Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus*. Molecular Phylogenetics and Evolution 19: 167–176.

McMullen, M.D., Hunter, B., Phillips, R.L. and Rubenstein, I. (1986) The structure of the maize ribosomal DNA spacer region. Nucleic Acids Research 14: 4953–4968.

Mitchell, A. and Wen, J. (2005) Phylogeny of *Brassaiopsis* (Araliaceae) in Asia based on nuclear ITS and 5S- NTS DNA Sequences. Systematic Botany 30(4): 872-886.

Mort, M.E. and Crawford, D.J. (2004) The continuing search: low copy nuclear sequences for lower level plant molecular phylogeny. Taxon 53(2): 257-261.

Muir, G., Fleming, C.C., Schlötterer, C. (2001) Three divergent rDNA clusters predate the species divergence in *Quercus petraea* (Matt.) Liebl. and *Quercus robur* L.. Molecular Biology and Evolution 18: 112–119.

Nagylaki, T. (1984) The evolution of multigene families under intrachromosomal gene conversion. Genetics 106: 529–548.

Neyland, R. (2002) A phylogeny inferred from large-subunit (26S) ribosomal DNA sequences suggests the family Dasypogonaceae is closely aligned with Restionaceae allies. Australian Systematic Botany 15: 749-754.

Nickrent, D.L. and Soltis, D.E. (1995) A comparison of angiosperm phylogenies from nuclear 18S rDNA and rbcL sequences. Annals of the Missouri Botanical Garden 82: 208-234.

Ochieng, J.W., Henry, R.J., Baverstock, P.R., Steane, D.A. and Shepherd, M. (2007) Nuclear ribosomal pseudogenes resolve a corroborated monophyly of the eucalypt genus *Corymbia* despite misleading hypotheses at functional ITS paralogs. Molecular Phylogenetics and Evolution 44: 752–764.

Pang, X., Song, J., Zhu, Y., Xu, H., Huang, L. and Chen, S. (2010) Applying plant DNA barcodes for Rosaceae species identification. Cladistics 27(2):165-170.

Rossetto, M., Jackes, B.R., Scott, K.D. and Henry, R.J. (2002). Is the genus *Cissus* (Vitaceae) monophyletic? Evidence from plastid and nuclear ribosomal DNA. Systematic Botany 27(3): 522-533.

Sang T. (2002) Utility of Low-Copy Nuclear Gene Sequences in Plant Phylogenetics. Critical Reviews in Biochemistry and Molecular Biology, 37(3):121–147.

Schiebel, K., von Waldburg, G., Gerstner, J. and Hemleben, V. (1989) Termination of transcription of ribosomal RNA genes of mung bean occurs within a 175 bp repetitive element of the spacer region. Molecular & General Genetics 218:302–307.

Schluter, P.M., Stuessy, T.F. and Paulus, H.F. (2005) Making the first step: practical considerations for the isolation of low copy nuclear sequence markers. Taxon 54 (3): 766-770.

Schultz, J. and Wolf, M. (2009) ITS2 sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics. Molecular Phylogenetics and Evolution 52:520-523.

Schultz, J., Maisel, S., Gerlach, D., Muller, T. and Wolf, M. (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA 11:361-364.

Schultz, J., Muller, T., Achtziger, M., Seibel, P.N., Dandekar, T. and Wolf, M. (2006) The internal transcribed spacer 2 database - a web server for (not only) low level phylogenetic analyses. Nucleic Acids Research 34:W704-W707.

Seelanan, T., Brubaker, C.L., Stewart, J., Craven, L.A. and Wendef, J.F. (1999) Molecular systematics of Australian *Gossypium* section *Grandicalyx* (Malvaceae). Systematic Botany 24(2):183-208.

Simmons, M.P., Savolainen, V., Clevinger, C.C., Archer, R.H. and Davis, J.I. (2001) Phylogeny of the Celastraceae inferred from 26S nuclear ribosomal DNA, phytochrome B, rbcL, atpB, and morphology. Molecular Phylogenetics Evolution 19: 353-366.

Small, R.L., Cronn, R.C. and Wendel, J.F. (2004) Use of nuclear genes for phylogeny reconstruction in plants. Australian Systematic Botany 17: 145-170.

Smith, G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191: 528–535.

Soltis, D. E., Soltis, P.S., Nickrent, D.L., Johnson, L.A., Hahn, W.J., Hoot, S.B., Sweere, J.A., Kuzoff, R.K., Kron, K.A., Chase, M.W., Swensen, S.M., Zimmer, E.A., Chaw, S.M., Gillespie, L.J., Kress, W.J. and Sytsma, K.J. (1997) Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Annals of the Missouri Botanical Garden 84: 1-49.

Soltis, P.S., Soltis, D.E., Wolf, P.G., Nickrent, D.L., Chaw, S.-M. and Chapman, R.L. (1999) The phylogeny of land plants inferred from 18S rDNA sequences: pushing the limits of rDNA signal? Molecular Biology and Evolution 16: 1774-1784.

Starr, J.R. Harris, S.A. and Simpson, D.A. (2004) Phylogeny of the unispicate taxa in Cyperaceae tribe Cariceae I: generic relationships and evolutionary scenarios. Systematic Botany 29(3) 528-544.

Takaiwa, F., Oono, K. and Sugiura, M. (1985) Nucleotide sequences of the 17S-25S spacer region from rice rDNA. Plant Molecular Biology 4: 355-364.

Thornhill, D.J., Lajeunesse, T.C. and Santos, S.R. (2007) Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. Molecular Ecology 16: 5326-5340.

Tremousaygue, D., Laudie, M., Grellet, F. and Delseny, M. (1992) The *Brassica oleracea* rDNA spacer revisited. Plant Molecular Biology 18:1013–1018.

Wei, X.X. and Wang, X.Q. (2004) Recolonization and radiation in *Larix* (Pinaceae), evidence from nuclear ribosomal DNA paralogs. Molecular Ecology 13: 3115–3123.

Wen, J. and Zimmer, E.A. (1996) Phylogeny and biogeography of *Panax* L. (the ginseng genus, Araliaceae): Inference from ITS sequences of nuclear ribosomal DNA. Molecular Phylogenetics and Evolution 6: 167-177.

White, T.J., Bruns, T., Lee, S. and Taylor, J.W. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M.A., Gelfand, D.H., Sninsky, J.J., White, T.J. (Eds.) PCR protocols: A guide to methods and applications. New York: Academic Press. 315–322.

Wright, S.D., Yong, C.G., Wichman, S.R., Dawson, J.W. and Gardner, R.C. (2001) Stepping stones to Hawaii: a trans-equatorial dispersal pathway for Metrosidos (Myrtaceae) inferred from rDNA (ITS ? ETS). Journal of Biogeography 28:769–774

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution 14: 717–724.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P. and Chen, S. (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE 5(10): e13102.

Zheng, X.Y., Cai, D.Y., Yao, L.H. and Teng, Y.W. (2008) Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. Molecular Phylogenetics and Evolution 48: 892–903.

Zimmer, E.A., Martin, S.L., Beverly, S.M., Kan, W. and Wilson, A.C. (1980) Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc. Nat. Acad. Sci. USA 77: 2158-2162.

*****

Plant DNA Barcoding and Phylogenetics, Pages 53-78
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

4 Nuclear and Organelle Specific PCR Markers

A.M. Alzohairy, G. Gyulai, H. Ohm, Z. Szabó,
S.M. Ragheb, M.A. Ali, H. Elsawy and A. Bahieldin

Introduction

The molecular markers targets different regions of the genomes either at coding or non-coding loci. Next to the biochemical (i.e. protein) markers such as ‘Izoenzyme Polymorphism’ (Hunter and Merkert, 1957), there are genetic (i.e. DNA) markers (Gyulai et al., 1997; Schulman, 2007; El-Domyati et al., 2011), such as Restriction Fragment Length Polymorphism (Grodzicker et al., 1974) and the numerous PCR-based (Saiki et al., 1985; Mullis and Faloona, 1987) marker systems. In the genome, these locus specific markers, either dominant or codominant, and either linked to genes or QTLs (Quantitative Trait Loci), are still important and feasible tools to compare robust number of individuals, species and populations. Technically, these markers can be divided into five technical generations (See Table 1) i.e. First Generation Markers [e.g. Polymerase Chain Reaction (Saiki et al., 1985), Restriction Fragment Length Polymorphism (Grodzicker et al., 1974)], Second Generation Markers [e.g. Inter Simple Sequence Repeats (Zietkiewicz et al., 1994), Randomly Amplified Polymorphic DNA (Williams et al., 1990)], Third Generation Markers [e.g. Amplified Fragment Length Polymorphism (Vos et al., 1995), Triple RAPD by using three primers, or more (Mansour et al., 2008)], New Generation Markers [(e.g. Recursive

Table 1: PCR-based DNA marker methods in sections and alphabetical orders (Gyulai et al., 1997; Maheswaran, 2004; Glenn, 2011).

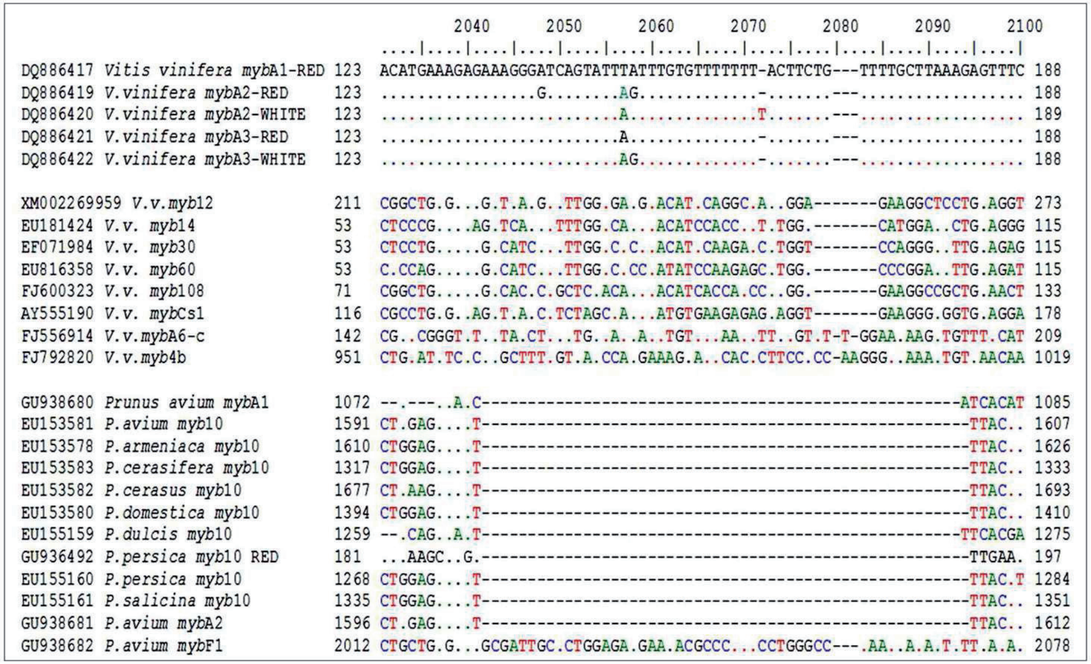
Acronym	Methods	References
(A) First Generation Markers		
ASO	Allele Specific Oligonucleotides	Saiki et al. (1986)
AS-PCR	Allele Specific Polymerase chain Reaction	Landegren et al. (1988)
OP	Oligonucleotide Polymorphism	Beckmann (1988)
PCR	Polymerase Chain Reaction	Saiki et al. (1985)
SSCP	Single Stranded Conformational Polymorphism	Orita et al. (1989)
STS	Sequence Tagged Site	Olsen et al. (1989)
VNTR	Variable Number Tandem Repeats	Jeffreys et al. (1985)
(B) Second Generation Markers		
AP-PCR	Arbitrarily Primed Polymerase Chain Reaction	Welsh and McClelland (1990)
ARMS	Amplification Refractory Mutation System PCR	Newton et al. (1989)
CAPS	Cleaved Amplified Polymorphic Sequence	Akopyanz et al. (1992)
DOP-PCR	Degenerate Oligonucleotides Primer - PCR	Telenius (1992)
ISJ-PCR	Intron-Exon Splice Junction PCR	Weining and Langridge (1991)
ISSR	Inter Simple Sequence Repeats	Zietkiewicz et al. (1994)
MAAP	Multiple Arbitrary Amplicon Profiling	Caetano-Anolles et al. (1993)
RAPD	Randomly Amplified Polymorphic DNA	Williams et al. (1990)
Double-RAPD	RAPD by using two primers	Klein-Lankhorst et al. (1991)
RLGS	Restriction Landmark Genome Scanning	Hatada et al. (1991)
SAMPL	Selective Ampl. MicroSatellite Polymorph. Loci	Morgante and Vogel (1994)
SCAR	Sequence Characterized Amplified Region	Paran and Michelmore (1993)
SSR	Simple Sequence Repeats	Akkaya et al. (1992)

***Table 1. cont*.**

STMS	Sequence Tagged Micro Satellite Sites	Beckmann and Soller (1990)
Tetra-PCR	Allele specific amplification by tetra-primer PCR	Ye et al. (1992)
(C) Third Generation Markers		
AFLP	Amplified Fragment Length Polymorphism	Vos et al. (1995)
ASAP	Allele Specific Associated Primers	Gu et al. (1995)
CFLB	Cleavage Fragment Length Polymorphism	Brow (1996)
DAMD-PCR	Directed Ampl. of Mini Satellite DNA-PCR	Bebeli et al. (1997)
IMP	Inter-MITE Polymorphism	Chang et al. (2001)
IRAP	Inter- Retrotransposon Amplified Polymorphism	Kalendar et al. (1999)
ISTR	Inverse Sequence-Tagged Repeats	Rohde (1996)
MITE	Miniature Inverted-Repeat Transposable Element	Casa et al. (2000)
qRT-PCR	quantitative Real Time PCR	Heid et al. (1996)
RBIP	Retrotransposon Based Insertional Polymorphism	Flavell et al. (1998)
REMAP	Retrotransposon-MicroSatellite Ampl. Polym.	Kalender et al. (1999)
R-ISSR	Combinations of RAPD-ISSR and RAPD-SSR	Ye et al. (2005)
R-PCR	Restricted-PCR	Puskás and Bottka (1995)
RT-PCR	Real-Time PCR	Higuchi et al. (1993)
SNP	Single Nucleotide Polymorphisms	Jordan and Humphries (1994)
SRAP	Sequence Related Ampl. Polymorphism	Li and Quiros (2001)
SSAP	Sequence Specific Ampl. Polymorphism	Waugh et al. (1997)
TE-AFLP	Three Endonuclease AFLP	Van der Wurff et al. (2000)
Triple-RAPD	Triple RAPD by using three primers, or more	Mansour et al. (2008)
(D) New Generation Markers		
DArT	Diversity ARrays Technology	Jaccoud et al. (2001)
KASP	Kbioscience Allele-Specific Polym. Assay	Uitdewilligen et al. (2013)

Table 1. cont.

MSAP	Methylation Sensitive Ampl. Polymorphism	Baurens et al. (2003)
RGF	Recursive Genome Function	Pellionisz (2008)
sRNA-qRT-PCR	Small RNA qRT-PCR	Varkonyi-G and Hellens (2010)
(E) Genome Sequencing (First and New Generations)		
AFFYMETRIX	DNA and RNA Microarrays / Chip	Fodor et al. (1991, 2007)
ddNTPs	Dideoxynucleotide Sequencing (ABI)	Sanger et al. (1980)
ILLUMINA / SOLEXA	The first Short Read Sequencer / bridgePCR	in: Bentley (2006)
Ion Torrent	Proton sequencing (Portable sequencer) / emPCR	Pennisi (2010)
NanoPorSeq (?)	Nanopore genome sequencer	Hayden (2012)
ROCHE454	Pyrosequencing (the 1st Next Gen. Seq.) / emPCR	Ronaghi et al. (1996, 1999)
RT-SEQ (SMRT)	Single Molecule Real Time seq./ PACBIOSCI	http://pacbiodevnet.com
SOLiD/ABI	Seq. by Oligonucl. Ligation and Detection / emPCR	in: Tang et al. (2009)
StarLight (?)	Single-molecule sequencing with quantum dots	in: Glenn et al. (2011)



```
                                                 2040      2050      2060      2070      2080      2090      2100
                                            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
DQ886417 Vitis vinifera mybA1-RED 123  ACATGAAAGAGAAAGGGATCAGTATTTATTTGTGTTTTTTT-ACTTCTG---TTTTGCTTAAAGAGTTTC 188
DQ886419 V.vinifera mybA2-RED     123  .................G........AG.............-.......---.................. 188
DQ886420 V.vinifera mybA2-WHITE   123  ..........................A..............T.......---.................. 189
DQ886421 V.vinifera mybA3-RED     123  ..........................A..............-.......---.................. 188
DQ886422 V.vinifera mybA3-WHITE   123  ..........................AG.............-.......---.................. 188

XM002269959 V.v.myb12             211  CGGCTG.G...G.T.A.G..TTGG.GA.G.ACAT.CAGGC.A..GGA-------GAAGGCTCCTG.AGGT 273
EU181424 V.v. myb14               53   CTCCCG....AG.TCA...TTTGG.CA...ACATCCACC..T.TGG.-------CATGGA..CTG.AGGG 115
EF071984 V.v. myb30               53   CTCCTG.....G.CATC...TTGG.C.C..ACAT.CAAGA.C.TGGT-------CCAGGG..TTG.AGAG 115
EU816358 V.v. myb60               53   C.CCAG.....G.CATC...TTGG.C.CC.ATATCCAAGAGC.TGG.-------CCCGGA..TTG.AGAT 115
FJ600323 V.v. myb108              71   CGGCTG.....G.CAC.C.GCTC.ACA...ACATCACCA.CC..GG.-------GAAGGCCGCTG.AACT 133
AY555190 V.v. mybCs1              116  CGCCTG.G..AG.T.A.C.TCTAGC.A...ATGTGAAGAGAG.AGGT-------GAAGGG.GGTG.AGGA 178
FJ556914 V.v.mybA6-c              142  CG..CGGGT.T..TA.CT...TG..A..A..TGT...AA..TT..GT.T-T-GGAA.AAG.TGTTT.CAT 209
FJ792820 V.v.myb4b                951  CTG.AT.TC.C..GCTTT.GT.A.CCA.GAAAG.A..CAC.CTTCC.CC-AAGGG..AAA.TGT.AACAA 1019

GU938680 Prunus avium mybA1       1072 --.---..A.C----------------------------------------------------ATCACAT 1085
EU153581 P.avium myb10            1591 CT.GAG....T-----------------------------------------------------TTAC.. 1607
EU153578 P.armeniaca myb10        1610 CTGGAG....T-----------------------------------------------------TTAC.. 1626
EU153583 P.cerasifera myb10       1317 CTGGAG....T-----------------------------------------------------TTAC.. 1333
EU153582 P.cerasus myb10          1677 CT.AAG....T-----------------------------------------------------TTAC.. 1693
EU153580 P.domestica myb10        1394 CTGGAG....T-----------------------------------------------------TTAC.. 1410
EU155159 P.dulcis myb10           1259 --.CAG..A.T----------------------------------------------------TTCACGA 1275
GU936492 P.persica myb10 RED      181  ...AAGC..G.-----------------------------------------------------TTGAA. 197
EU155160 P.persica myb10          1268 CTGGAG....T-----------------------------------------------------TTAC.T 1284
EU155161 P.salicina myb10         1335 CTGGAG....T-----------------------------------------------------TTAC.. 1351
GU938681 P.avium mybA2            1596 CT.GAG....T-----------------------------------------------------TTAC.. 1612
GU938682 P.avium mybF1            2012 CTGCTG.G...GCGATTGC.CTGGAGA.GAA.ACGCCC...CCTGGGCC---.AA..A.A.T.TT.A.A. 2078
```

Fig. 1: Sequence alignments (70 nt) of the MybR2R3 TF gene (transcription factor genes of nuDNA), which play roles in the fruit color development including *Vitis* (Vitaceae) and *Prunus* (Rosaceae) species. Sequences were downloaded from the NCBI server (Altschul et al., 1997) and GGB (Grape Genome Browser) http://www.genoscope.cns.fr), following the sequence alignments by BioEdit (Hall, 1999). Consensus nucleotides (.), deletions (-), SNPs (color letters) and accession numbers (NCBI) are indicated.

Genome Function (Pellionisz, 2008)] and Genome Sequencing or First and New Generations [(e.g. Nanopore genome sequencer (Hayden, 2012)]. Before utilization, all of these techniques need primer design including analyses for hairpin, self and heterodimer formations and suitable annealing temperatures (Alzohairy et al., 2014a,b). The primer specificity should be confirmed by tools of basic local alignment search tool (Figure 1) (Altschul et al., 1997) and MSA (Multiple Sequence Alignments) *in silico* with the software programs of BioEdit (Hall, 1999), MULTALIN (Combet et al., 2000), CLUSTAL W (Thompson et al., 1994), MEGA4 (Tamura et al., 2007), and FastPCR (Kalendar et al., 2009). Useful servers are also available, such as National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL), and Integrated DNA Technologies (idtDNA). The *in-silico* PCR (http://insilico.ehu.es/PCR/) is also available for assessment of primers used.

PCR markers target for coding DNA regions

- **RT-PCR (Reverse Transcriptase PCR) and qRT-PCR (Quantitative Real-Time PCR):** The combinations of RTase enzymes (Baltimore, 1970; Temin and Mizutani, 1970), which can transcribe cDNA from RNA copy; and the real-time PCR (RT-PCR) (Higuchi et al., 1993), which can detect the levels of amplifying fragments PCR-cycle bay cycles; led to the development of qRT-PCR (quantitative RT-PCR) (Holland et al., 1991; Heid et al., 1996). This technique replaced the Northern blot analysis (Alwine et al., 1977), which was developed for mRNA quantification (Tang et al., 2009), basically following the DNA blot techniques of Southern blot analysis (Southern, 1975). In qRT-PCR, the isolated RNA templates of either mRNA or small RNA (Varkonyi-Gasic and Hellens, 2010) are first converted to complementary DNA (cDNA) using a RTase enzymes, and then the transcribed cDNA is used as a template for regular PCR. The fragment analysis may conducted by end-point detection on agarose gel (i.e. semi quantitative PCR) (Bittsánszky et al., 2006; Alzohairy et al., 2012) (Figure 2a), or by qPCR equipments (Livak and Schmittgen, 2001; Gyulai et al., 2012a). To follow the amplifying qRT-PCR products cycles by cycles (Figure 2b) DNA staying fluorescent dyes (EtBr, EvaGeen, SybrGreed etc.) are applied either in single or combined (e.g. TaqMan; LifeTechnologies) forms (Freeman et al., 1999).

- **Nested-PCR and Nested qRT-PCR:** Nested PCR involves two subsequent uniplex PCR reactions, in which the first PCR product (i.e. the nest) is used for the second set of primers, which amplifies a secondary target site within the first PCR product. One of the principal of this method is that if a wrong locus were amplified first by mistake the probability is very low to amplify it by the second time with the second primer pair. The other advantage is in the case of very low concentration of target sequence (e.g. viral infections). Based on qRT-PCR, nested qRT-PCR was found unique diagnostic tool to detect RNA viruses in the Human genome with the resolution rate of single tumor cell of 10^6 white blood cells (Drobyski et al., 1994).

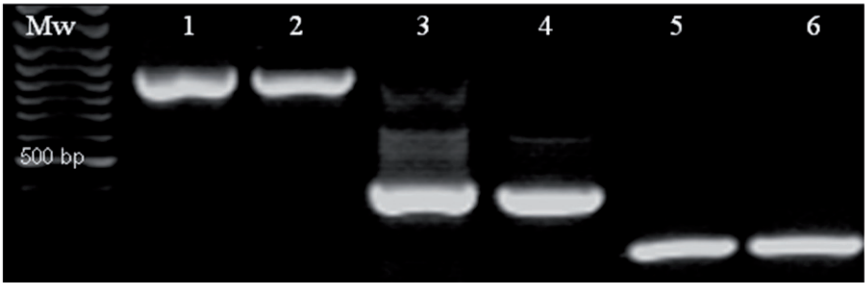


Fig. 2a: Samples of semi quantitative RT-PCR (Real Time Polymerase Chain Reaction) (Corbett Research, Co) with end-point detection on agarose gel (1.2 %). Gene expression levels of the constitutively expressed gene 26SrRNA (1, 2) for standard, and the nuclear encoded gene rbcS (3, 4), and the stress responsive gene gst (5, 6) were determined in the cDNA samples of two poplar (*Populus x canescens*) clones. Bands were quantified by a densitometer (ChemiImagerTM 4400, Alpha Innotech Inc., San Leandro, CA, USA) and calculated by a computer program (Phoretix 1D Advanced, Nonlinear Dynamics, Ltd., Newcastle upon Tyne, UK) (Bittsánszky et al., 2006).

- **Multiplex PCR:** Compared to uniplex PCR, which amplifies single nucleic acid sequence of the genome, multiplex PCR (mPCR) is performed with more than one primer pairs in the same reaction mix, which amplify more than one target sequences. Several mPCR assays were suggested by research groups for microbiological quality control of food (Park et al*.,* 2006), water (Kong et al*.,* 2002), clinical samples and pharmaceutical raw materials and products (Ragheb et al., 2012). An alternative PCR strategy can be applied by using gradient thermocyclers, which allow the use of primers of different annealing temperatures (T_{ann}) for simultaneous amplification of different targets in the same run (Erlich et al., 1991; Don et al., 1991).

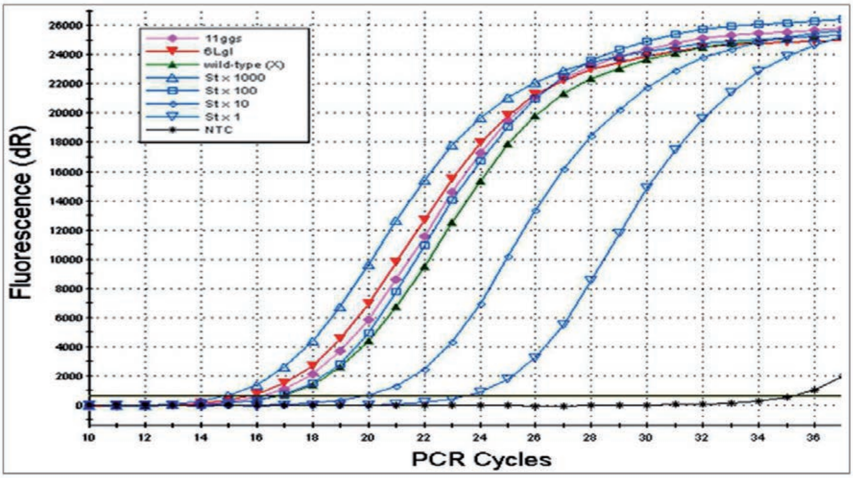


Fig. 2b: Sample of qRT-PCR (quantitative Real Time Polymerase Chain Reaction) measurements (Corbett Research, Co) of the expression of plant gsh1 genes in three poplar (*Populus* x *canescens*) clones (11ggs, 6Lgl, and Wild Type), and compared to the concentration series of control DNAs (1 to 1000 times dilution series) and NTC (Non Template Control). Relative fluorescence (dR) and PCR cycle numbers are indicated (Bittsánszky et al., 2006; Gyulai et al., 2012a).

- **ARMS-PCR:** The tetra-primer Amplification Refractory Mutation System PCR (ARMS-PCR) is a Multiplex type PCR, which provides fast assays for SNP analysis (single polymorphic loci) (locus stands for nucleotide) coupled with sequencing or melting point analysis (Newton et al., 1989). Through the combinations of two outer primers and two allele-specific inner primers the barcoding (i.e. genotyping) requires only regular PCR and fragment separations by electrophoresis (Ye et al., 1992).

- **Simple Sequence Repeats (SSRs):** In all prokaryote and eukaryote genomes SSRs (syn.: microsatellites) are found universally with core SSRs of 1 to 6 nucleotides (Gupta et al., 1994). The length polymorphism between individuals occurs due to the change in the number of core repeats (Jarne and Lagoda, 1996; Szabó et al., 2005; Gyulai et al., 2006; Tóth et al., 2007). Dinucleotide core repeats like (CA)n and (GA)n are the most abundant repeats. In humans, (CA)n repeat occurs once in every 30 kb. PCR primer pairs are designed to the sequences of flanking regions of the microsatellites (Dakin and Avise, 2004), and after the PCR amplification is followed by visualization in agarose or polyacrylamide gels (Figure 3a,b,c). SSR provides co-dominant and highly reproducible markers. Microsatellite markers were found very useful for population genetics, variety identification and protection, monitoring of seed purity and hybrid quality, gene tagging, germplasm evaluation, genome mapping and phylogenetic studies (Lavin et al., 2003) with or without bootstrap analysis (Figure 4) (Dakin and Avise, 2004; Kaukinen et al., 2004; Alzohairy et al., 2012, 2013; Gyulai et al., 2014). In the case of hierarchical cluster analysis the Maximum Likelihood (ML) method was suggested to be the most comprehensive method (Hillis et al., 1994).

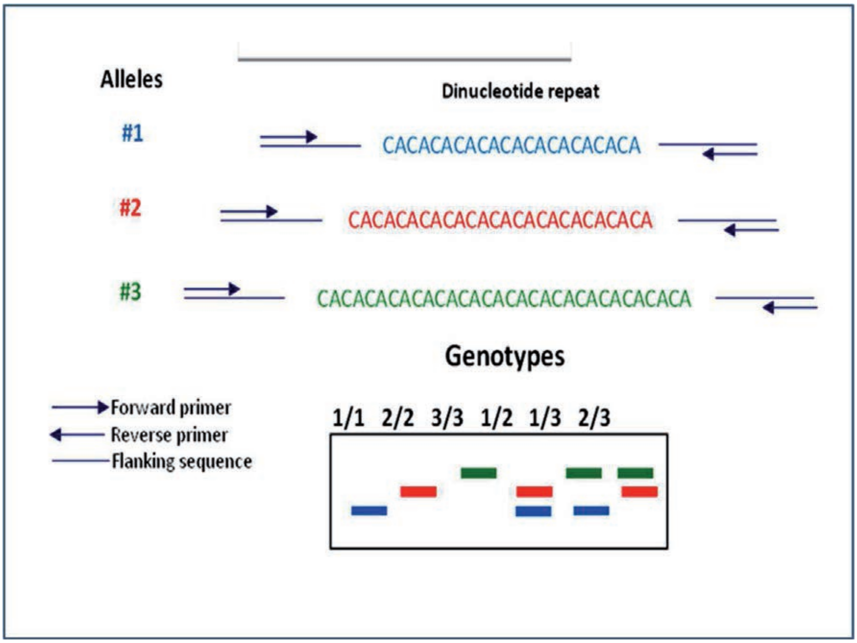


Fig. 3a: Principle of the microsatellite (SSR) based barcoding in the case of a (CA)n repeat.

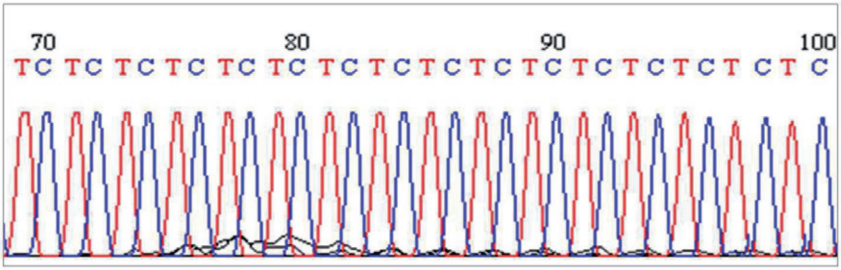


Fig. 3b: Sample of the fluorogram of an SSR (TC)n repeat (68 to 100 nt) of watermelon (*Citrullus lanatus*) microsatellite amplified and sequenced by ABI PRISM 3100 Genetic Analyzer (LifeTechnologies) (Szabó et al., 2005; Gyulai et al., 2011b).

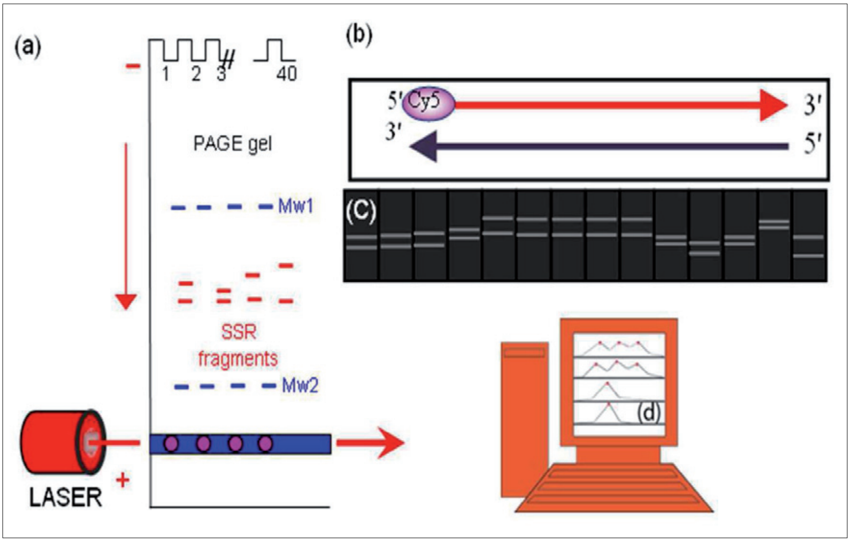


Fig. 3c: Schematic flow chart of the ALF-SSR fragment separation by using ALF method (Automatic Laser Fluorometer; Amersham Bioscience, Uppsala, Sweden - AP Budapest, Hungary) (a). One strand of the PCR primer pair is labelled by Cy5 fluorescent die (b). The outputs of gel image (c) and the real fluorograms (d) are shown. DNA loading lanes (1-40), electric current (poles + and -), Mw standards (Mw1 and Mw2), DNA fragments, and the data analyzing computer program (ALFwin Fragment Analyser Version 1.00) are indicated (Tóth et al., 2007; Gyulai et al., 2006, 2011a,b).

- **ALF-SSR:** The SSR analysis can be automated and multiplexing by labelling the primers with fluorescent dye (Figure 2c). A very sensitive and automated method of ALF-SSR PCR (Automatic Laser Fluorometer). In this method, one primer of each primer pair is labelled with Cy5 (Phosphoramidite), a cyanine type fluorescent dye, at the 5'-end (Sigma) according to Röder et al. (1998). The Cy5 labelled ALF-SSR fragments are excited by helium/neon microlaser at 643 nm, and the emitted fluorescent signal of the Cy5 is detected at 667 nm by ALFexpress II DNA Analyser (Amersham Bioscience, Uppsala, Sweden - AP Budapest, Hungary). The ALF-SSR fluorograms are analyzed by computer program of ALFwin Fragment Analyser Version 1.00 (Pharmacia – Amersham Bioscience, Uppsala, Sweden. AP-Hungary, Budapest) according to Röder et al. (1998), Huang et al. (2002), and Gyulai et al. (2011a,b; 2012b).

- **Inter-simple sequence repeats (ISSR)**: ISSR amplifies (Zietkiewicz et al., 1994) target DNA regions between located in two identical microsatellite repeats (SSRs) (Figure 5). ISSR was used extensively for producing molecular markers in crops and horticultural plants (Lágler et al., 2005; Youssef et al., 2010).

- **RAPD-ISSR (R-ISSR):** R-PCR combines RAPD and ISSR primer pairs in PCR reactions, which is able to reveal new genomic loci that could not be detected with either primer system alone in maize genome (Ye et al., 2005). Further combinations of the different marker systems also suggest new possibilities in DNA barcoding.

- **R-PCR (Restricted PCR)**: For reducing nonspecific PCR amplifications, which is caused by mispriming during PCR reactions, besides the standard pair of primers, 3'-dideoxy-terminated competitor oligonucleotides were applied in the PCR reactions (Puskás and Bottka, 1995). By this way an enhanced specificity of target site amplification was achieved. The competitor oligonucleotides act by masking possible sites of nonspecific primer-template interaction, thus excluding undesired PCR extensions. This technique is generally applicable when highly degenerate primers are used (Puskás and Bottka, 1995).

- **DGGE-RAPD (Denaturing Gradient Gel Electrophoresis - RAPD)**: The detection of DNA polymorphism in self-pollinating species was found to be difficult. To facilitate, DGGE (Denaturing Gradient Gel Electrophoresis) was used for RAPD analysis. In DGGE gel (12% acryalmide in TAE buffer with denaturant gradient, 10-50%, of 7 M urea and 40% formamide) the two alleles of a locus (if different) run separately. This method greatly improved the detection of reproducible DNA polymorphism among closely related plant species and lines. It was used first to estimate pedigree relationships among plant materials in wheat (*Triticum*), barley (*Hordeum*) and oat (*Avena*) (Dweikat et al., 1993). DDGE-RAPD was also found highly discriminative for the identification of barley (*Hordeum*) cultivars with different pedigree (Bahieldin et al., 2006), and proved that DGGE-RAPD is a superior method for detecting DNA polymorphism when compared to RFLP, agarose-RAPD, or polyacrylamide-RAPD methods (Figure 6).

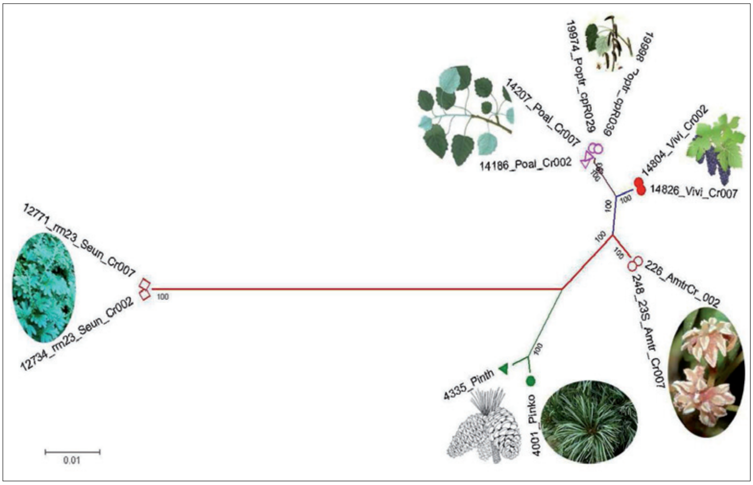


Fig. 4: Bootstrap (1000 x replicates) radial phylogram of 23S rRNS gene of cpDNA of grapevine (*Vitis vinifera*) compared to the evolutionarily first land plant of ferns *Selaginella uncinata* (*Seun*); the ancient higher plant of gymnosperm *Pinus thunbergii (*Pinth) and *Pinus koraiensis* (Pinko); the fist angiosperms plant *Amborella trichopoda* (*Amtr*) of the basal angyosperm ANITA (*Amborella, Nymphaeales, Illiciales, Trimeniaceae* and *Austrobaileya*) group, which plant's (a dwarf tree) by xylotomy still resembles to gymnosperms but by flowers do it to angiosperms. *Vitis vinifera* (Vivi) and two poplars (Poal - *Populus alba* and Poptr - *P. trichocarpa*) are also included. cpDNA sequences (<2830 nt) were downloaded from CGD (Chloroplast Genome Database; http://chloroplast.cbio.psu.edu/index.html). Sequences were aligned by BioEdit (Hall, 1999), following ML phylogram edition (Maximum Likelihood, Hillis et al., 1994) by using MEGA4 program (Tamura et al., 2007). Genetic distance (scale 0.01) is indicated.

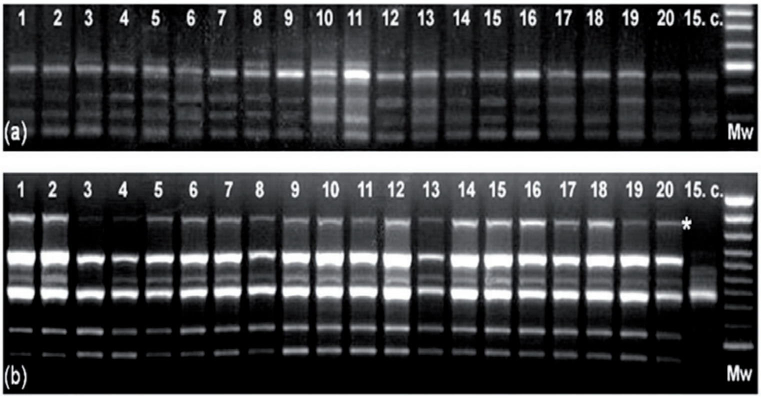


Fig. 5: Samples of ISSR analysis with monomorphic (a) and polymorphic (b) band patterns on agarose gel (0.8 %), which were amplified by primers of FV835 [(AG)$_8$YC] (a), and FV811 [(GA)$_8$C] (b), in common millet (*P. miliaceum*) cultivars (1 to 20) compared to an ancient medieval (15th CENT.) sample (indicated as 15.c). Mw – 100 bp DNA ladder. Asterisk indicates the missing ISSR fragment (Lágler et al., 2005).

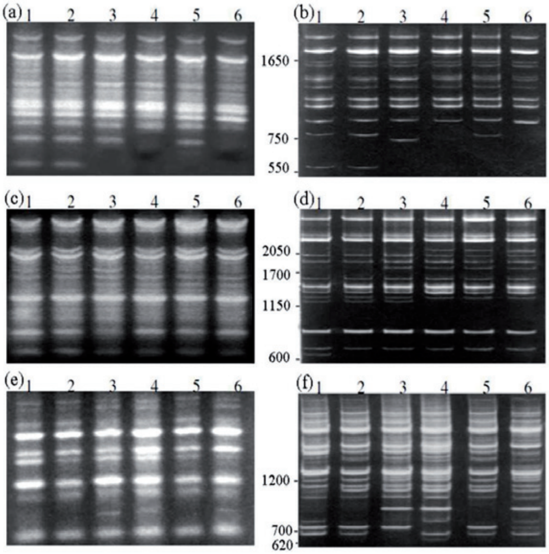


Fig. 6: Comparison of fragment patterns of agarose-RAPD (a), (c) and (e); and DGGE-RAPD (b), (d) and (f) generated by Operon RAPD primers of OP-A12 (a) and (b), OP-B09 (c) and (d); and OP-B15 (e) and (f) used for six barley (*Avena sativa*) cultivars (Bahieldin et al., 2006).

PCR markers target random sites of the genome

- **Single-, Double-, and Triple-Primed RAPDs:** RAPD (Random Amplified Polymorphic DNA) assay is one of the earliest and widely used PCR-assay using single primer of arbitrary nucleotide sequence (Williams et al., 1990). The potential of the original RAPD assay (Gyulai et al., 2000) was further increased by combining two and three primers (Figure 7) in the same PCR reaction (Mansour el al., 2008).

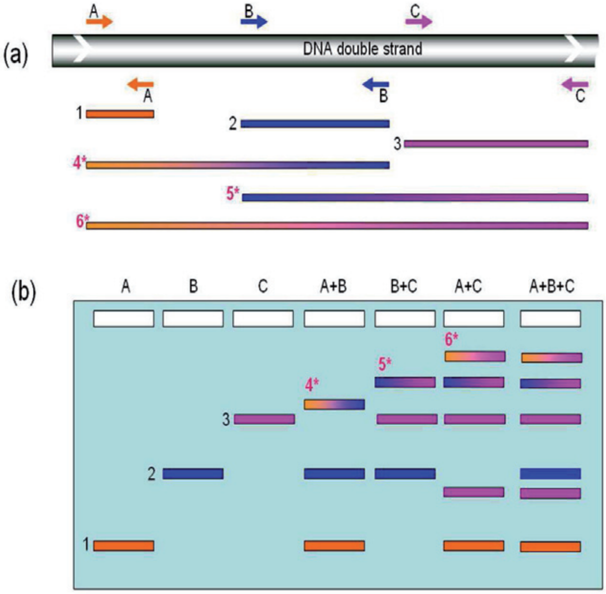


Fig. 7: Principles of the Single- (A), Double- (A, B), and Triple- (A, B, and C) primed RAPD-PCRs (a) with indications of hypothetical band patterns (b). Only a single locus (band) per primers is indicated. Bands amplified with double and triple primers are indicated with asterisk (Mansour et al., 2008).

- **Retrotransposon based PCR markers:** Retrotransposons (RTs) are major constituents of most eukaryotic genomes; they are ubiquitous, dispersed throughout the genome, and their abundance correlates with the host genome sizes, which provides unique possibilities for molecular barcoding (Ali et al., 2014).

Organelle (Chloroplast and Mitochondria) specific PCR markers

DNAs of chloroplast (cpDNA) and mitochondria (mtDNA) have been used very frequently in plant systematic and phylogenetic studies (Ali et al., 2014). Both organelle DNAs are circular molecules ranging in size of 120 Kbp to 500 Kbp (Figure 8), with unique exception of green alga *Floydiella terrestris* with huge cpDNA of 521.168 bp (NCBI# NC_014346); and *Cucumis melo* (Alverson et al., 2010) with giant mtDNA (2,900,000 bp) (Gyulai et al., 2012b; Ali et al., 2014).

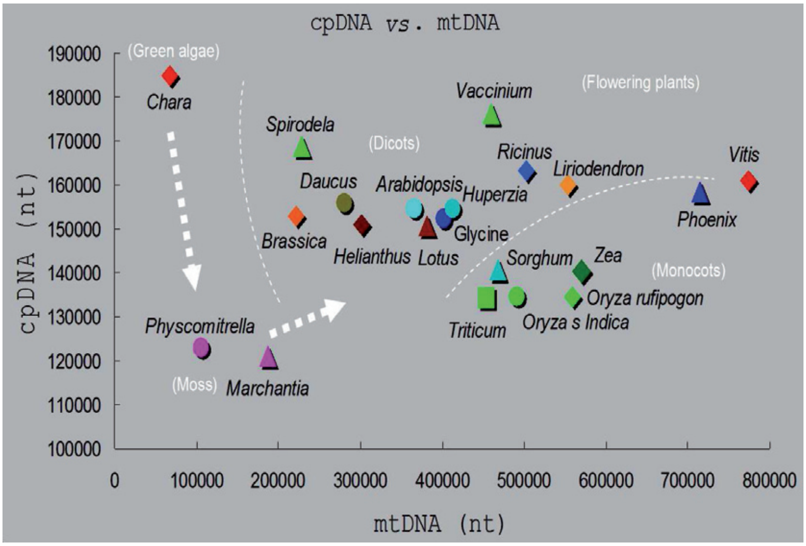


Fig. 8: Size correlations between cpDNA and mtDNA genomes show a shift from green algae (*Chara vulgaris*) with high cpDNA/mtDNA ratio (2.73) through mosses of *Physcomitrella patens* (1.16) and *Marchantia polymorpha* (0.65) towards flowering plants of dicots to monocots with exception of *Vitis.* The decreasing ratio of cpDNA/mtDNA indicates an enlargement of mtDNA during the evolution*: Spirodela polyrhiza* (0,74); *Brassica napus* (0,69); *Daucus carota* (0,55); *Helianthus annuus* (0,50); *Arabidopsis thaliana* (0,42); *Lotus japonicus* (0,40); *Vaccinium macrocarpon* (0,38); *Glycine max* (0,38); *Vigna radiata* (0,38); *Huperzia lucidula* (0,37); *Ricinus communis* (0,32); *Sorghum bicolor* (0,30); *Triticum aestivum* (0,29*); Liriodendron tulipifera* (0,29); *Oryza sativa var. japonica* (0,274); *Oryza sativa var. indica* (0,273); *Zea mays* (0,25); *Oryza rufipogon* (0,24); *Phoenix dactylifera* (0,22); and *Vitis vinifera* (0,21). NCBI (Altschul et al., 1997) data were plotted by XY plot of Microsoft Windows Xcel program (Ali et al., 2014).

There are about 100 functional genes encoded in the chloroplast genomes, which contains, with few exceptions (IRL – IRless), two duplicate regions of inverted repeats (IR) in reverse orientation (from 10 to 76 kb). They divide the chloroplast genome into large (LSC) and small single-copy (SSC) regions. The structural organization of chloroplast genome is highly conserved, i.e., relatively free of large deletions, insertions, transpositions, inversions and SNPs (single nucleotide polymorphism), which make it advantageous for phylogenetic studies. Chloroplast DNA is abundant (generally, 50 chloroplasts are in plant cells, and a single chloroplast have 50 cpDNA copy, which results 2.500 cpDNA copy per cell) compared nuclear DNA (generally 2n).

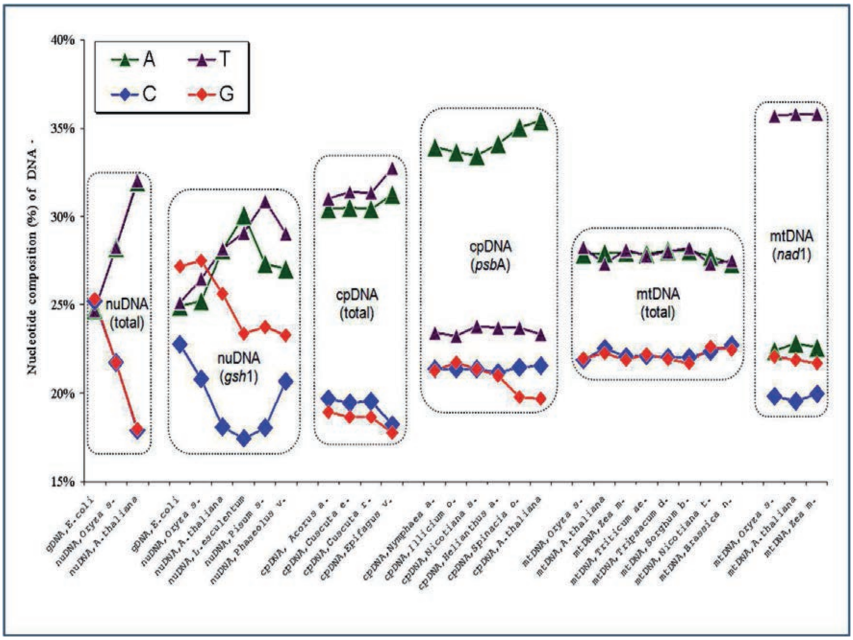


Fig. 9: Nucleotide compositions of DNA sequences. Total genomes (including coding and non-coding regions) (nuDNA, gDNA, cpDNA, and mtDNA) are compared to coding gene sequences (*gsh*1; *psb*A and *nad*1). Accession numbers are NC_003070; NC_008402.1; NC_003076.5; X03954; AJ508916; Y09944; AF017983; AF128455; AF128454; NC_010093.1; NC_009963.1; NC_009766.1; NC_001568.1; NC_006050; EF380354; AB237912; DQ383815; AJ400848; AP000423; NC_007886.1; NC001284.2; NC_007982.1; NC_007579.1; NC_008362.1; NC_008360.1; NC_006581.1; NC_008285; NC_007886; NC_001284; NC_007982 of NCBI (Altschul et al., 1997).

Organelle DNAs are usually uniparentally inherited (maternally in angiosperms and paternally in gymnosperms, in general, with some exceptions) (Neale et al*.,* 1989), which facilitate to determine the maternal parent in hybrids and allopolyploids. Some chloroplast regions like psbA-trnH spacer, and rps16 intron gene evolve relatively rapidly. There are a number of noncoding cpDNA regions which are also useful target of study such as the intergenic spacer of atpB-rbcL (reviewed by Ali et al., 2014). Due to the evolutionary high AT-content of prokaryotic cpDNA and mtDNA of higher plants compared to nuDNA (Figure 9) lower temperatures are used in the PCR reactions (Demesure et al., 1995; Dane et al., 2004).

Total Genome Barcoding (i.e. Sequencing)

From the new generation markers (Pellionisz, 2008) through the total genome sequencing such as the first ddNTP technology followed by Chip technologis (Fodor et al., 1991, 2007), the second- and new generation technologies to the newest proton- (Pennisi, 2010) and nanopore sequencing (Hayden, 2012), which provide the total genome sequences.

References

Akkaya, M.S., Bhagwat, A.A. and Cregan, P.B. (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132: 1131-1139.

Akopyanz, N., Bukanov, N.O., Westblom, T.U. and Berg, D.E. (1992) PCR-based RFLP analysis of DNA sequence diversity in the gastric pathogen Helicobacter pylori. Nucleic Acids Research 20: 6221-6225.

Ali, M.A., Gyulai, G., Hidvégi, N., Kerti, B., Al Hemaid, F.M.A., Pandey, A.K. and Lee, J. (2014) The changing epitome of species identification - DNA barcoding. Saudi Journal of Biological Sciences 21: 204–231.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipmand, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K. and Palmer, J.D. (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Molecular Biology and Evolution 27: 1436–1448.

Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proceedings of the National Academy of Sciences USA 74 (12): 5350–5354.

Alzohairy, A.M., Gyulai, G., Jansen, R.K. and Bahieldin, A. (2013) Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid 69: 1–15.

Alzohairy, A.M., Gyulai, G., Ramadan, M.F., Edris, S., Sabir, J.S.M., Jansen, R.K. and Bahieldin, A. (2014a) Retrotransposon-based molecular markers for assessment of genomic diversity. Functional Plant Biology 41: 781–789.

Alzohairy, A.M., Sabir, J.S.M., Gyulai, G., Younis, R., Jansen, R. and Bahieldin, A. (2014b) Environmental Stress Activation of Plant LTR-Retrotransposons. Functional Plant Biology 41: 557–567.

Alzohairy, A.M., Yousef, A.A., Edris, S.S., Kerti, B., Gyulai, G. and Bahieldin, A. (2012) Detection of long terminal repeat (LTR) retrotransposons reactivation induced by in vitro environmental stresses in barley (*Hordeum vulgare*) via reverse transcription-quantitative polymerase chain reaction (RT-qPCR). Life Science Journal 9: 5019–5026.

Bahieldin, A., Ahmed, I.A., Gad El-Karim, Gh.A., Eissa, H.F., Mahfouz, H.T. and Saleh, O.M. (2006) DGGE-RAPD analysis as a useful tool for cultivar identification. African Journal of Biotechnology 5: 566–569.

Baltimore, D. (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature 226: 1209–1211.

Baurens, P., Bonnot, F., Bienvenu, D., Causse, S., and Legavre, T (2003) Using SD-AFLP and MSAP to assess CCGG methylation in the banana genome. Plant Molecular Biology Reporter 21: 339–348.

Bebeli, P.J., Zhou, Z., Somers, D.J. and Gustafson, J.P. (1997) PCR primed with mini satellite core sequences yields DNA fingerprinting probes in wheat. Theoretical and Applied Genetics 95: 276–283.

Beckmann, J.S. (1988) Oligonucleotide polymorphisms: A new tool for genomic genetics. Bio/Technology 6: 161–164.

Beckmann, J.S. and Soller, M. (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged micro satellite sites. Bio/Technology 8: 930–932.

Bentley, D.R. (2006) Whole-genome re-sequencing. Current Opinion in Genetics & Development 16: 545–552.

Bittsánszky, A., Gyulai, G., Humphreys, M., Gullner, G., Csintalan, Zs., Kiss, J., Szabó, Z., Lágler, R., Tóth, Z., Rennenberg, H. and Kőmíves, T. (2006) RT-PCR analysis and stress response capacity of transgenic *gsh*I-poplar clones (*Populus* x *canescens*) in response to paraquat exposure. Z Naturforschung C, Journal of Bioscience 61: 699–730.

Brow, M.A., Oldenburg, M.C., Lyamichev, V., Heisler, L.M., Lyamicheva, N., Hall, J.G., Eagan, N.J., Olive, D.M., Smith, L.M., Fors, L. and Dahlberg, J.E. (1996) Differentiation of bacterial 16S rRNA genes and intergenic regions and *Mycobacterium tuberculosis* katG genes by structure-specific endonuclease cleavage. Journal of Clinical Microbiology. 34: 3129–3137.

Caetano-Anollés, G., Bassam, B.J. and Gresshoff, P.M. (1993) Enhanced detection of polymorphic DNA by multiple arbitrary amplicon profiling of endonuclease-digested DNA: identification of markers tightly linked to the super nodulation locus in soybean. Molecular Genetics and Genomics 241: 57–64.

Casa, A.M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S. and Wessler, S.R. (2000) The MITE family Heartbreaker (Hbr): Molecular markers in maize. Proceedings of the National Academy of Sciences USA 97: 10083–10089.

Chang, R.Y., O'Donoughue, L.S. and Bureau, T.E. (2001) Inter-MITE polymorphisms (IMP): a high throughput transposon-based genome mapping and fingerprinting approach. Theoretical and Applied Genetics 102: 773781.

Combet, C., Blanchet, C., Geourjon, C. and Deléage, G. (2000) NPS@: network protein sequence analysis. Trends in Biochemical Sciences 25: 147–150.

Dakin, E.E. and Avise, J.C. (2004) Microsatellite null alleles in parentage analysis. Heredity 93: 504–509.

Dane, F., Lang, P. and Bakhtiyarova, R. (2004) Comparative analysis of chloroplast DNA variability in wild and cultivated *Citrullus* species. Theoretical and Applied Genetics 108: 958–966.

Demesure, B., Sodzi, N. and Petit, R.J. (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Molecular Ecology 4: 129–131.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K. and Mattick, J.S. (1991) Touchdown PCR to circumvent spurious priming during gene amplification. Nucleic Acids Research 19: 4008.

Drobyski, W.R., Knox, K.K., Majewski, D. and Carrigan, D.R. (1994) Fatal encephalitis due to variant B human herpesvirus 6 infection in a bone marrow transplant recipient. The New England Journal of Medicine 330: 1356–1360.

Dweikat, I., Mackenzie, S., Levy, M. and Ohm, H. (1993) Pedigree assessment using RAPDDGGE in cereal crop species. Theoretical and Applied Genetics 85: 497–505.

El-Domyati, F.M., Younis, R.A.A, Edris, S., Mansour, A., Sabir, J. and Bahieldin, A. (2011) Molecular markers associated with genetic diversity of some medicinal plants in Sinai. Journal of Medicinal Plants Research 4: 200–210.

Erlich, H.A., Gelfand, D. and Sninsky, J.J. (1991) Recent advances in the polymerase chain reaction. Science 252: 1643–1651.

Flavell, A.J., Knox, M.R., Pearce, S.R. and Ellis, T.H.N. (1998) Retro transposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. The Plant Journal 16: 643–650.

Fodor, A.A., Tickle, T.L. and Richardson, C. (2007) Towards the uniform distribution of null P values on Affymetrix microarrays. Genome Biology 8: R69.

Fodor, A.A., Tickle, T.L. and Richardson, C. (2007) Towards the uniform distribution of null P values on Affymetrix microarrays. Genome Biology 8: R69.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. and Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251: 767–73.

Freeman, W.M., Walker, S.J. and Vrana, K.E. (1999) Quantitative RT-PCR: pitfalls and potential. BioTechniques 26: 124–125.

Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759–769.

Grodzicker, T., Williams, J., Sharp, P. and Sambrook, J. (1974) Physical mapping of temperaturesensitive mutations of adenoviruses. Cold Spring Harbor Symposia on Quantitative Biology 39: 439–446.

Gu, W.K., Weeden, N.F., Yu, J. and Wallace, D.H. (1995) Large-scale, cost-effective screening of PCR products in marker-assited selection applications. Theoretical and Applied Genetics 91: 465–470.

Gupta, M., Chyi, Y.S., Romero-Severson, J. and Owen, J.L. (1994) Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theoretical Applied Genetics 89: 998–1006

Gyulai, G., Bittsánszky, A., Gullner, G., Heltai, Gy., Pilinszky, K., Molnár, E. and Kömíves, T. (2012a) Gene reactivation induced by DNA demethylation in Wild Type and 35S-*gsh*I-*rbc*S transgenic poplars (*Populus x canescens*). Novel plant sources for phytoremediation. Journal of Chemical Science and Technology 1: 9–13.

Gyulai, G., Bittsánszky, A., Szabó, Z., Waters Jr, L., Gullner, G., Kampfl, Gy., Heltai, Gy. and Kőmíves, T. (2014) Phytoextraction potential of wild type and 35S-gshI transgenic poplar trees (*Populus × canescens*) for environmental pollutants herbicide paraquat, salt sodium, zinc sulfate and nitric oxide *in vitro*. International Journal of Phytoremediation 16: 379–396.

Gyulai, G., Dweikat, I., Janovszky, J., Ohm, H. and Sharma, H. (1997) Application of ISSR/SSR-PCR for genome analysis of *Agropyron*, *Bromus*, and *Agropyron x Bromus*. In: Z. Staszewski, W. Mlyniec, R. Osinski (Eds.) Ecological aspects of breeding fodder crops and amenity grasses, PBAI, Radzikow, Poland, pp. 306–312.

Gyulai, G., Gémesné, J.A., Sági, Zs., Venczel, G., Pintér, P., Kristóf, Z., Törjék, O., Bottka, S., Kiss, J. and Zatykó, L. (2000) Doubled haploid development and PCR-analysis of F_1 hybrid derived DH-R_2 paprika (*Capsicum annuum L.*) lines. Journal of Plant Physiology 156: 168–174.

Gyulai, G., Humphreys, M., Lágler, R., Szabó, Z., Tóth, Z., Bittsánszky, A. and Gyulai, F. (2006) Seed remains of common millet from the 4th (Mongolia) and 15th (Hungary) centuries: AFLP, SSR and mtDNA sequence recoveries. Seed Science Research 16: 179–191.

Gyulai, G., Láposi, R., Rennenberg, H., Veres, A., Herschbach, C., Fábián, Gy. and Waters Jr, L. (2011b) Conservation genetics (1710 – 2010) - Cloning of living fossils: Micropropagation of the oldest Hungarian black locust tree (*Robinia pseudoacacia*) planted in 1710 (Bábolna, Hungary). In: G. Gyulai (Ed.) Plant Archaeogenetics. Nova Sci Publisher Inc., New York, USA. pp. 117–127.

Gyulai, G., Szabó, Z., Wichmann, B., Bittsánszky, A., Waters Jr, L., Tóth, Z. and Dane, F. (2012b) Conservation genetics - Heat Map analysis of nuSSRs of aDNA of archaeological watermelons (Cucurbitaceae, *Citrullus lanatus*) compared to current varieties. Genes, Genomes and Genomics 6 (SI1): 86–96.

Gyulai, G., Tóth, Z. and Bittsánszky, A. (2011a) Flesh color reconstruction from aDNAs of *Citrullus* seeds from the 13th, 15th, and 19th cents (Hungary). In: G. Gyulai (Ed.) Plant Archaeogenetics. Nova Sci Publisher Inc., New York, USA. pp. 69–87.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98.

Hatada, I., Hayashizaki, Y., Hirotsune, S., Komatsubara, H. and Mukai, T. (1991) A genome scanning method for higher organism using restriction sites as landmarks. Proceedings of the National Academy of Sciences USA 88: 397–400.

Hayden, E.C. (2012) Nanopore genome sequencer makes its debut. Technique promises it will produce a human genome in 15 minutes. Nature (doi:10.1038/nature.2012.10051)

Heid, C.A., Stevens, J., Livak, K.J. and Williams, P.M. (1996) Real time quantitative PCR. Genome Research 6: 986–993.

Higuchi, R., Fockler, C., Dollinger, G. and Watson, R. (1993) Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Biotechnology 11: 1026–1030.

Hillis, D.M., Huelsenbeck, J.P. and Swofford, D.L. (1994) Hobgoblin of phylogenetics? Nature 369: 363–364.

Holland, P.M., Abramson, R.D., Watson, R. and Gelfand, D.H. (1991) Detection of specific polymerase chain reaction product by utilizing the 5'-3' exonuclease activity of *Thermus aquaticus* DNA polymerase. Proceedings of the National Academy of Sciences USA 88: 7276–7280.

Huang, X.Q., Börner, A., Röder, M.S. and Ganal, M.W. (2002) Assessing genetic diversity of wheat (*Triticum aestivum* L.) germplasm using microsatellite markers. Theoretical and Applied Genetics 105: 99–707.

Hunter, R.L. and Merkert, C.L. (1957) Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. Science 125: 1294–1295.

Jaccoud, D., Peng, K., Feinstein, D. and Kilian, A. (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research 29(4): E25.

Jarne, P. and Lagoda, P.J.L. (1996) Microsatellites, from molecules to populations and back. Trends in Ecology and Evolution 11: 424–429.

Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) Hyper variable mini satellite regions in human DNA. Nature 314: 67–73.

Jordan, S.A. and Humphries, P. (1994) Single nucleotide polymorphism in exon 2 of the BCP gene on 7q31-q35. Human Molecular Genetics 3: 1915

Kalendar, R., Grob, T., Regina, M., Suoniemi, A. and Schulman, A.H. (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theoretical and Applied Genetics 98: 704–711.

Kalendar, R., Lee, D. and Schulman, A.H. (2009) FastPCR Software for PCR Primer and Probe Design and Repeat Search. Genes, Genomes and Genomics 3: 1–14.

Kaukinen, K.H., Supernault, K.J. and Miller, K.M. (2004) Enrichment of tetranucleotide microsatellite loci from invertebrate species. Journal of Shellfish Research 23: 621.

Klein-Lankhorst, R.M., Vermunt, A., Weide, R., Liharska, T. and Zabel, P. (1991) Isolation of molecular markers for tomato (*L. esculentum*) using random amplified polymorphic DNA (RAPD). Theoretical and Applied Genetics 83: 108–14.

Kong, R.Y., Lee, S.K., Law, T.W., Law, S.H. and Wu, R.S. (2002) Rapid detection of six types of bacterial pathogens in marine waters by multiplex PCR. Water Research 36: 2802–2812.

Lágler, R., Gyulai, G., Humphreys, M., Szabó, Z., Horváth, L., Bittsánszky, A., Kiss, J. and Holly, L. (2005) Morphological and molecular analysis of common millet (*P. miliaceum*) cultivars compared to an aDNA sample from the 15th century (Hungary). Euphytica 146: 77–85.

Landegren, U., Kaiser, R., Sanders, J. and Hood, L. (1988) DNA diagnostics. Molecular techniques and automation. Science 241: 1077–1080.

Lavin, M., Wojciechowski, M.F., Gasson, P., Hughes, C.E. and Wheeler, E. (2003) Phylogeny of robinioid legumes (Fabaceae) revisited: *Coursetia* and *Gliricidia* recircumscribed, and a biogeographical appraisal of the Caribbean endemics. Systematic Botany 28: 387–409.

Li, G. and Quiros, C.F. (2001) Sequence-relatedamplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in Brassica. Theoretical and Applied Genetics 103: 455–461.

Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method*.* Methods 25: 402–408.

Mansour, A., Ismail, O.M. and Mohei EL-Din, S.M. (2008) Diversity assessments among Mango (*Mangifera indica* L.) cultivars in Egypt using ISSR and three-primer based RAPD fingerprints. *The African* Journal of Plant Science and Biotechnology 2: 87–92.

Morgante, M. and Vogel, J. (1994) Compound micro satellite primers for the detection of genetic polymorphisms. U.S. Patent Appl., 08/326456.

Mullis, K.B. and Faloona, F.A. (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Methods in Enzymology 155: 335–350.

Neale, D.B., Marshall, K.A. and Sederoff, R.R. (1989) Chloroplast and mitochondrial DNA are paternally inherited in *Sequoia sempervirens* D. Don Endl. Proceedings of the National Academy of Sciences USA 86: 9347–9349.

Newton, C.R., Graham, A., Heptinstall, L.E., Powell, S.J., Summers, C., Kalsheker, N., Smith, J.C. and Markham, A.F. (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Research 17: 2503–2516.

Olsen, M., Hood, L., Cantor, C. and Botstein, D. (1989) A common language for physical mapping of the human genome. Science 245: 1434–1435.

Orita, M., Suzuki, Y., Sekiya, T. and Hayashi, K. (1989) Rapid and sensitive detection of point mutations and DNA polymorphisms using polymerase chain reaction. Genomics 5: 874–879.

Paran, I. and Michelmore, R.W. (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theoretical and Applied Genetics 85: 985–993.

Park, Y.S., Lee, S.R. and Kim, Y.G. (2006) Detection of *Escherichia coli* O157:H7, *Salmonella* spp., *Staphylococcus aureus* and *Listeria monocytogenes in* kimchi by multiplex polymerase chain reaction (mPCR). The Journal of Microbiology 44: 92–97.

Pellionisz, A.J. (2008) The principle of recursive genome function. Cerebellum 7: 348–359.

Pennisi, E. (2010) Semiconductors inspire new sequencing technologies. Science 327(5970): 1190.

Puskás, L.G. and Bottka, S. (1995) Reduction of mispriming in amplification reactions with restricted PCR. Genome Research 5: 309–11.

Ragheb, S.M., Yassin, A.S. and Amin, M.A. (2012) The application of uniplex, duplex, and multiplex PCR for the absence of specified microorganism testing of pharmaceutical excipients and drug products. PDA Journal of Pharmaceutical Science and Technology 66: 307–17.

Röder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.H., Leroy, P. and Ganal, M.W. (1998) A microsatellite map of wheat. Genetics 149: 007–2023.

Rohde, W. (1996) Inverse sequence-tagged repeat (ISTR) analysis, a novel and universal PCR-based technique for genome analysis in the plant and animal kingdom. Journal of Genetics and Breeding 50: 249–261.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. Analytical Biochemistry 242: 84–89.

Ronaghi, M., Nygren, M., Lundeberg, J. and Nyren, P. (1999) Analyses of secondary structures in DNA by pyrosequencing. Analytical Biochemistry 267: 65–71.

Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1986) Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature 324: 163–166.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. Science 230(4732): 1350–1354.

Sanger, F., Coulson, A.R., Barrell, B.G, Smith, A.J.H. and Roe, B.A. (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. Journal of Molecular Biology 143: 161–178.

Schulman, A.H. (2007) Molecular markers to assess genetic diversity. Euphytica 158: 313–321.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology 98: 503–517.

Szabó, Z., Gyulai, G., Humphreys, M., Horváth, L., Bittsánszky, A. and Lágler, R. (2005) Genetic variation of melon (*C. melo*) compared to an extinct landrace from the Middle Ages (Hungary) I. rDNA, SSR and SNP analysis of 47 cultivars. Euphytica 146: 87–94.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. and Surani, M.A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6: 377–82.

Telenius, H., Carter N.P., Bebb, C.E., Nordenskjold, M., Ponder, B.J. and Tunnacliffe, A. (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 13: 718–725.

Temin, H.M. and Mizutani, S. (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226: 1211–1213.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680.

Tóth, Z., Gyulai, G., Horváth, L. and Szabó, Z. (2007) Watermelon (*Citrullus lanatus*) production in Hungary from the Middle Ages. Hungarian Agricultural Research 2007/4: 14–19.

Uitdewilligen, J.G., Wolters, A.M., D'hoop, B.B., Borm, T.J., Visser, R.G., and van Eck, H.J. (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One 8(5): e62355.

van der Wurff, A.W.G., Chan, Y.L., van Straalen, N.M. and Schouten, J. (2000) TE-AFLP: combining rapidity and robustness in DNA fingerprinting. Nucleic Acids Research 28: 105–109.

Varkonyi-Gasic, E. and Hellens, R.P. (2010) qRT-PCR of small RNAs. Methods in Molecular Biology 631: 109–22.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M. (1995) AFLP:

a new technique for DNA fingerprinting. Nucleic Acids Research 23(21): 4407–4414.

Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Kumar, A., Thomas, B.T. and Powell, W. (1997) Genetic distribution of BARE-1 retro transposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular Genetics and Genomics 253: 687–694.

Weining, S. and Langridge, P. (1991) Identification and mapping of polymorphisms in cereals based on polymerase chain reaction. Theoretical and Applied Genetics 82: 209–216.

Welsh, J. and McClelland, M. (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Research 18: 7213–7218.

Williams, J.G.K., Kubelik, A.R., Livak, K.L., Rafalski, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18: 6531–6535.

Ye, C., Yu, Z., Kong, F., Wu, S. and Wang, B. (2005) R-ISSR as a new tool for genomic fingerprinting, mapping, and gene tagging. Plant Molecular Biology Reporter 23: 167–177.

Ye, S., Humphries, S. and Green, F. (1992) Allele specific amplification by tetra-primer PCR. Nucleic Acids Research 20: 1152.

Youssef, M.A., Mansour, A. and Solliman, S.S. (2010) Molecular markers for new promising drought tolerant lines of rice under drought stress via RAPD-PCR and ISSR markers. The Journal of American Science 6: 355–363.

Zietkiewicz, E., Rafalski, A. and Labuda, D. (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20: 176–183.

*****

Plant DNA Barcoding and Phylogenetics, Pages 79-90
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

5 Maturase K Gene in Plant DNA Barcoding and Phylogenetics

P. Kar, A. Goyal and A. Sen

Introduction

DNA barcoding is a tool which is used for the identification of unknown plant and animal species by using a short DNA sequences. In 2003, Paul Hebert and his co-workers from University of Guelph, Canada, discovered and proposed the term DNA barcoding and brought a new dimension in the scientific community. The Consortium for the Barcode of Life (CBOL) plant-working group recommended the 2-locus combination of rbcL and matK as the standard plant barcodes based on assessments of recoverability, sequence quality and levels of species discrimination. Molecular based method such as DNA barcoding is a modern and innovative technique which can explore the evolution as well as the genetic relatedness of plants. Various workers have been studied the plant evolution, and able to solve various anomalies in the taxonomic levels by using the chloroplast gene such as matK and rbcL. The matK gene has two unique features that emphasize its importance in molecular biology and evolution. It is characterized by its fast evolutionary rate and putative function as a group II intron maturase. matK is a chloroplast-encoded gene nested between the 5' and 3' exons of trnK, tRNA-lysine (Sugita et al., 1985) in the large single copy region of the chloroplast genome. The mode and tempo of matK evolution is distinct from other chloroplast genes. Rate of nucleotide substitution in matK is three times higher than that of the large subunit of Rubisco (rbcL) and six fold higher at the amino acid substitution rate (Johnson and Soltis, 1994; Olmstead and Palmer, 1994), establishing it as a fast- or rapidly-evolving gene. This high nucleotide and amino acid substitution rates provides high phylogenetic signal for resolving evolutionary relationships among plants at all taxonomic levels (Hilu and Liang, 1997; Soltis and Soltis, 1998; Hilu et al., 2003). The resolution achieved with sequences of matK is equivalent to using up eleven other genes combined (Hilu and Liang, 1997; Hilu et al., 2003). In addition to the high rate of substitution, matK also displays varying number and size of indels (insertions and deletions) (Hilu and Liang, 1997; Whitten et al., 2000; Hilu et al., 2003). Most indels identified in matK have been found in multiples of three, conserving the reading frame (Hilu and Liang, 1997; Hilu and Alice, 1999; Whitten et al., 2000; Hilu et al., 2003). However, the presence of indels, high substitution rates, and premature stop codons found in some plant families (Kores et al., 2000; Kugita et al., 2003) raises the question of whether a gene, with such features, is capable of maintaining stable protein structure and function. matK is the only gene found in the chloroplast genome of higher plants that contains this putative maturase domain (Neuhaus and Link, 1987). There are 16 group II introns nested within 15 chloroplast genes (Kohchi et al., 1988; Ems et al., 1995; Maier et al., 1995), which would require a maturase for intron splicing and proper protein translation. Maturases are splicing factors that aid in splicing and folding group II introns (Vogel et al., 1997). Studies of the white barley mutant albostrians, a chloroplast ribosomal mutant, demonstrated that although, some group II introns were processed by an imported nuclear maturase, there were at least six plastid genes (trnK, atpF, trnI, trnA, rpl2, and rps12 cis) with group II introns that would require a chloroplast maturase for splicing (Vogel et al., 1997; Vogel et al., 1999). Western blot analysis indicated that a protein of approximately 60 kDa is produced by the matK gene in barley (Vogel et al., 1999). Identification of a MatK protein product and demonstration of a lack of splicing for some group II introns in the albostrians mutant suggest a potential functional role for MatK as a group II intron maturase in the chloroplast.

Structure of the matK gene

For most land plants, the matK gene is nested between the two exons of trnK, tRNA-lysine. The matK ORF is approximately 1500 bp in most angiosperms (Hilu et al., 1999), corresponding to around 500 amino acids for the translated protein product. The structure of this gene includes indels of various length and number (Hilu and Alice, 1999; Hilu et al., 2003). For example, the *Epifagus* matK gene contains a 200-bp deletion at the 5’ end compared to tobacco matK (Ems et al., 1995). Nucleotide substitution rates are not evenly distributed across the matK ORF, but instead, matK has regions displaying high mutation rates (Hilu and Liang, 1997). The third codon position tends to have a slightly higher mutation rate than the first and second codon position, suggesting neutral or purifying selection in this gene (Young and dePamphilis, 2000). Amino acid sequence analysis revealed a highly conserved region close to the 3’ end of this gene that lacks indels (Hilu and Liang, 1997). This region contains 448 bp, is called domain X and has similarity to a conserved functional domain found in mitochondrial group II intron maturases (Sugita et al., 1985; Neuhaus and Link, 1987).

Group II introns

Group II introns are a large class of self-catalytic ribozymes as well as mobile genetic element found within the genes of all three domains of life. Ribozyme activity (e.g., Self-splicing) can occur under high-salt conditions *in vitro*. However, assistance from proteins is required for *in vivo* splicing. Introns can be classified into one of three groups: I, II, or III. Group I introns are considered primarily self-splicing but an accessory protein factor for intron excision is required in some cases (Saldanha et al., 1993; Geese and Waring, 2001). Group II introns often require a maturase for excision and can only self-spice under non-physiological conditions (Saldanha et al., 1993; Noah and Lambowitz, 2003). Group III introns are a modified form of group II introns (Mohr et al., 1993). Group II introns can be further subdivided based on structural characteristics into IIA and IIB (Michel et al., 1989). The cellular location of group I and group II introns is very similar with both being dispersed in mitochondria and chloroplast genomes. Group I and group II introns can be found in mRNA, tRNA and rRNA genes of each of these organelles (Ferat and Michel, 1993; Cho et al., 1998; Vogel et al., 1999; Bhattacharya et al., 2002; Rudi et al., 2003). Unlike group II introns, group I introns have also been found in nuclear genes (Saldanha et al., 1993). For group I introns, the splicing reaction involves a guanine, as the attacking group, to break the phosphodiester bond on the 5’ side of the intron (Alberts et al., 1994). Group II introns, however, use adenine as the attacking nucleotide to form the lariet structure (Alberts et al., 1994; Kelchner, 2002). Group II introns can be either autocatalytic (its ancestral characteristic), or require splicing factors to form the lariat structure and be spliced out of the RNA transcript (Alberts et al., 1994).

Group II intron evolution

Group II introns have been found in mitochondria and chloroplasts of plants and fungus (Mohr et al., 1993). In addition, this group of introns has also been identified in the proteobacterium *Azotobacter vinelandii* and the cyanobacterium *Calothrix*, bacteria related to the probable ancestor of the mitochondria and chloroplast, respectively (Ferat and Michel, 1993). Thus, group II introns are an ancestral character of organelle evolution. Several group II introns bear an open reading frame (Sugita et al., 1985; Mohr et al., 1993; Moran et al. 1995; Saldanha et al. 1999) which often encodes a reverse-transcriptase/maturase that is capable of transposing the intron into a new location (Mohr et al., 1993; Moran et al., 1995). A study using *Saccharomyces cervisiae* demonstrated this mobility by showing that mitochondrial group II introns can be inserted in cox1 alleles that were originally missing this group of introns (Moran et al., 1995). This mobility provided an indication towards evolution of ORFs in this pool of introns. The ORF not only contained domain X for maturase activity, but also, a reverse-transcriptase (RT)-like domain (Mohr et al, 1993). Phylogenetic analysis of the RT domain of intron encdoed proteins (IEPs) indicated sequence homology to the RT domains of retroviruses, such as HIV-1 (Blocker et al., 2005), and non-long terminal repeat (LTR) retroelements (Mohr et al., 1993; Moran et al., 1995; Blocker et al., 2005). Thus, the group II intron-encoded proteins are evolutionarily related to retroviruses and retroelements known to have display similar patterns of mobility within the genome.

Group II intron maturases

It was discovered that some mitochondrial group II introns contain ORFs encoding their own splicing factors termed ‘maturases’ (Saldanha et al., 1993). Maturases are thought to be required as translated protein for *in vivo* splicing of some group II introns (Mohr et al., 1993). Although the maturases of yeast and *Lactococcus* only process the intron in which they are encoded (Matsuura et al., 2001; Cui et al., 2004; Rambo and Doudna, 2004), at least two maturases, CRS2 and MatK, can splice several different introns (Liere and Link, 1995; Jenkins et al., 1997; Vogel et al., 1999; Ostheimer et al., 2003). Both of these maturases are thought to function in the chloroplast (Liere and Link, 1995; Jenkins et al., 1997; Vogel et al., 1999; Ostheimer et al., 2003).

Sequence analysis of 34 intron-encoded ORFs identified three domains that are generally maintained in most maturases: a reverse transcriptase (RT) domain, domain X, and a zinc finger-like region (Mohr et al., 1993). The RT domain is thought to be an ancestral character, remnant from the origin of these introns as non-LTR retrotransposons (Mohr et al., 1993; Moran et al., 1995). The RT domain is active in certain maturases (Moran et al., 1995; Matsuura et al., 1997; Saldanha et al., 1999; Wank et al., 1999). The zinc finger-like domain comprises the core of the DNA endonuclease activity of these maturases (Moran et al., 1995), while the maturase activity is retained in domain X (Mohr et al., 1993). All three regions of the maturase enzyme are thought to act in concert to achieve mobility in group II intron (Saldanha et al., 1999; Singh et al., 2002; Rambo and Doudna, 2004). However, only the RT domain and domain X are required for the splicing activity (Cui et al., 2004; Rambo and Doudna, 2004). A general mechanism for this mobility/maturase activity is through maturase domain that binds to the group II intron and consequently, folds the intron to form a lariat structure by bringing the attacking adenine to the 5' end of the intron (Mohr et al., 1993; Saldanha et al., 1999; Kelchner, 2002). This results in splicing the intron lariat structure out of the precursor RNA. The maturase then remains bound to the excised RNA to form a ribonucleoprotein particle (RNP) (Saldanha et al., 1999). Next, the DNA endonuclease domain creates a double-strand break at the target insertion site (Saldanha et al., 1999). Once the break is formed, the reverse transcriptase domain is activated to integrate the excised group II intron into a new site by DNA-primed reverse transcription (Saldanha et al., 1999). Although the RT and DNA endonuclease activity have been well studied in these introns, the maturase activity is less well understood.

Group II intron maturases have been studied primarily in the *Lactococcus* LtrA maturase protein (Matsuura et al., 1997; Wank et al., 1999; Singh et al., 2002; Noah and Lambowitz, 2003), yeast mitochondrial maturases (Moran et al., 1994) and a few nuclear-encoded maturases (Jenkins et al., 1997; Mohr and Lambowitz, 2003). A mechanism of splicing has been defined for the LtrA maturase (Matsuura et al., 2001; Singh et al. 2002; Rambo and Doudna, 2004), and preliminary research has indicated aspects of nuclear-encoded maturase function (Jenkins et al., 1997; Ostheimer et al., 2003; Ostheimer et al., 2006). However, studies on mitochondrial maturases have not defined a mechanism of group II intron processing. The mechanism of bacterial maturase LtrA is the most defined, and shown to be influenced by magnesium concentration (Matsuura et al., 1997; Noah and Lambowitz 2003). LtrA binds to a high affinity region on the group II intron referred to as DIVa (Matsuura et al., 2001; Singh et al., 2002). This region is also the site of the ORF for the maturase in the intron (Matsuura et al., 2001; Singh et al., 2002; Rambo and Doudna, 2004). Once bound, the protein interacts with other conserved regions in the intron to form the final lariat structure for excision (Matsuura et al., 2001; Singh et al., 2002).

The nuclear-encoded maturase CRS2 is transported to the chloroplast where it processes nine out of the 10 chloroplast-encoded group IIB introns (Ostheimer et al., 2003). CRS2 forms a complex with CAF1 and CAF2 for binding and processing group IIB introns (Ostheimer et al., 2003). However, no other details of the splicing mechanism have been defined. CRS1 is a nuclear-encoded chloroplast maturase that acts only on the group IIA intron of atpF in the chloroplast (Till et al., 2001). However, the group IIA intron of atpF also requires an additional, yet to be identified, factor from the chloroplast for complete excision (Jenkins et al., 1997). Since the protein product of matK is the only putative group II intron maturase encoded in the choroplast genome (Neuhaus and Link, 1987), it can be hypothesized that the additional chloroplast-encoded factor for intron excision in atpF is MatK.

Significance in evolutionary studies

Evolutionary studies in plants utilize several methodologies in order to obtain the most clearly defined robust phylogenetic trees. Molecular sequence data has revolutionized evolutionary studies and enhanced the resolution of phylogenetic trees immensely. Genes used in plant systematics display different trends of evolution. Slow-evolving genes, such as rbcL and atpB, have high sequence conservation among plant groups. This high sequence conservation allows a good resolution that has been confined to the family level, but cannot solve the intricacies below this level (Hilu and Liang, 1997; Goldman et al., 2001). Fast-evolving genes, such as matK, provide enough revenues for evolutionary analysis at the family level and below (Hilu and Liang, 1997; Goldman et al., 2001). The matK gene is considered to be fast-evolving due to the fact that it has a high rate of substitution and more variable sites compared to other genes (Olmstead and Palmer, 1994; Johnson and Soltis, 1995; Hilu and Liang, 1997; Soltis and Soltis, 1998). The matK ORF is not homogenous in rate of nucleotide substitution but instead contains regions of varying rates of substitution (Johnson and Soltis, 1995; Hilu and Liang, 1997). One of the conserved regions in matK is the putative functional domain X (Hilu and Liang, 1997).

In phylogenetic analysis, phylogenetically informative characters are those which are variable and not the product of homoplasy (parallel evolution). However, these characters are not so variable that alignment between specific taxonomic levels can be accomplished. matK provides many informative characters in regions that do not have excessive variability nor excessive conserved sequence and can be aligned to determine evolutionary relationships from the species to the divisional or even higher taxonomic levels (Johnson and Soltis, 1995; Hilu and Liang, 1997; Hilu et al., 1999, 2003). matK has been useful for determining phylogenetic histories for several plant taxa including the Saxifragaceae (Johnson and Soltis, 1995), Orchidaceae (Kores et al., 2000, 2001; Whitten et al., 2000; Goldman et al., 2001), the asterids (Bremer et al., 2002), as well as across all angiosperms taxa (Hilu et al., 2003). Phylogenetic studies using matK have produced more robust trees than had previously been determined using multiple genes (Hilu et al., 2003). Despite this extensive use of the matK gene for phylogenetic analysis, some disputes still remain pertaining to its expression and functionality in the chloroplast genome, reinforcing that matK is just a pseudogene (Kores et al., 2000; Whitten et al., 2000; Goldman et al., 2001; Kores et al., 2001). Nonetheless, researchers also utilized matK for some of their phylogenetic studies. matK has been considered as a pseudogene because they contain stop codons within the ORF, bear indels that create frame-shift mutations, and display an equal level of substitution for all three codon positions (Kores et al., 2000, 2001). Pseudogenes can fall into two categories: genes that are not transcribed, and genes that are transcribed but contain premature stop codons and produce truncated, non-functional proteins (Mighell et al., 2000; Balakirev and Ayala, 2003). The stop codons found within the matK ORF of members of the Orchidaceae may place matK in the second category of pseudogene that produces a truncated protein. However, this result would depend on the reading frame translated. Contrary to some of the findings of the matK gene sequence from the orchids, sequence analysis from nine representative species across the plant kingdom demonstrated that the indels within the matK gene occurred in multiples of three, conserving the matK reading frame (Ems et al., 1995). Additionally, frame-shift mutations found in the 3’ region of matK of the Poaceae, which could also alter or destroy the reading frame; appear to be limited to the very 3’ region of this gene, not affecting the functionality of domain X (Hilu and Alice, 1999). Thus, the ORF of matK appears to be intact and maintained in these plant species (Ems et al., 1995; Hilu and Alice, 1999). Further, the presence of the matK gene without trnK retained in the residual chloroplast genome of Epifagus (Ems et al., 1995) and the report of a protein product for MatK in extracts from barley (Vogel et al., 1999) support that this gene is translated into an essential functional protein product in the chloroplast genome. In 2008, Selvaraj et al., (2008) used the chloroplast matK gene sequences from GenBank to evaluate the generic and species oriented variations and phylogenetic relationships among the members of the family Zingiberaceae. They proved that of the 47 genera representing the family Zingiberaceae, five genus *Afromonum*, *Alpinia*, *Globba*, *Curcuma* and *Zingiber* showed polyphylogeny. They thus suggested that matK gene is a good candidate for DNA barcoding of Zingiberaceae family members.

In conclusion, the chloroplast gene maturase K (matK), proposed to bear group II introns, is one of the most variable coding genes of angiosperms. It is one of the most promising candidates for barcode analysis in land plants. Being a coding region, the matK has very high evolutionary rate and thus finds its application in phylogenetic reconstructions at high taxonomic levels, such as Order or Family, and sometimes also at low taxonomic levels, such as genus or species.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) The Cell. Garland Publishing, Inc., New York.pp. 56-57, 376-377.

Balakirev, E.S. and Ayala, F.J. (2003) Pseudogenes: Are they "junk" or funcitonal DNA? Annual Review of Genetics 37: 123-151.

Bhattacharya, D., Simon, D., Huang, J., Cannone, J.J. and Gutell R.R. (2002) The exon context and distribution of Euascomycetes rRNA spliceosomal intron. BMC Evolutionary Biology 3:7.

Blocker, F.J.H., Mohr, G., Conlan, L.H., Qi, L., Belfort, M. and Lambowitz, A.M. (2005) Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA 11: 14-28.

Bremer, B., Bremer, K., Heidar, N., Erixon, P., Olmstead, R.G., Anderberg, A.A., Kallersjo, M. and Barkhordarian, E. (2002) Phylogenetics of asterids based on 3 coding and 3 noncoding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Molecular Phylogenetics and Evolution 24: 274-301.

Cho, Y., Qiu, Y.-L., Kuhlman, P. and Palmer, J.D. (1998) Explosive invasion of plant mitochondria by a group I intron. Proceedings of the National Academy of Science USA 81: 1991-1995.

Cui, X., Matsuura, M., Wang, Q., Ma, H. and Lambowitz, A.M. (2004) A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to promote RNA splicing. Journal of Molecular Biology 340: 211-231.

Ems, S.C., Morden, C.W., Dixon, C.K., Wolfe, K.H., dePamphilis, C.W. and Palmer, J.D. (1995) Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant Epifagus virginiana. Plant Molecular Biology 29:721-733.

Ferat, J.L. and Michel, F. (1993) Group II self-splicing intron in bacteria. Nature 364:358-361.

Geese, W.J. and Waring, R.B. (2001) A comprehensive characterization of a group IB intron and its encoded maturase reveals that protein-assisted splicing requires an almost intact intron RNA. Journal of Molecular Biology 308: 609-622.

Goldman, D.H., Freudenstein, J.V., Kores, P.J., Molvray, M., Jarrell, D.C., Whitten, W.M., Cameron, K.M., Jansen, R.K. and Chase, M.W. (2001) Phylogenetics of Arethuseae (Orchidaceae) based on plastid matK and rbcL sequences. Systematic Botany 26: 670-695.

Hilu, K.W. and Liang, H. (1997) The matK gene: sequence variation and application in plant systematics. American Journal of Botany 84: 830-839.

Hilu, K.W. and Alice, L.A. (1999). Evolutionary implications of matK indels in Poaceae. American Journal of Botany 86:1735-1741.

Hilu, K.W., Alice, L.A. and Liang, H. (1999) Phylogeny of Poaceae inferred from matK sequences. Annals of the Missouri Botanical Gardens 86: 835-851.

Hilu, K.W., Borsch, T., Muller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, L.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A.B., Jens, G.R., Campbell, C.S. and Chatrou, L.W. (2003) Angiosperm phylogeny based on matK sequence information. American Journal of Botany 90: 1758-1776.

Jenkins, B.D., Khulhanek, D.J. and Barkan, A. (1997) Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classess with distinct requirements for splicing factors. The Plant Cell 9: 283-296.

Johnson, L.A. and Soltis, D.E. (1994) matK DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. Systematic Botany 19: 143-156.

Johnson, L.A. and Soltis, D.E. (1995) Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using matK sequences. Annals of the Missouri Botanical Gardens 82: 149-175.

Kelchner, S.A. (2002) Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. American Journal of Botany 89: 1651-1669.

Kohchi, T., Umesono, K., Ogura, Y., Komine, Y., Nakahigashi, K., Komano, T., Yamada, Y., Ozeki, H. and Ohyama, K. (1988) A nicked group II intron and trans-splicing in liverwort, *Marchantia* polymporpha. Nucleic Acids Research 16: 10025-10036.

Kores, P. J., Molvray, M., Weston, P. H., Hopper, S. D., Brown, A. P., Cameron, K.M., and Chase, M.W. (2001) A phylogenetic analysis of Diurideae (Orchidaceae) based on plastid DNA sequence data. American Journal of Botany 88:1903-1914.

Kores, P.J., Weston, P.H., Molvray, M. and Chase, M.W. (2000) Phylogenetic relationships within the Diurideae (Orchidaceae):

Inferences from plastid matK DNA sequences. In: Wilson, K.L. and Morrison D.A. (eds.), Monocots: Systematics and Evolution. CSIRO Publishin, Collingwood, Victoria Australia, pp. 449-455.

Kugita, M., Kaneko, A., Yamamoto, Y., Takeya, Y., Matsumoto, T. and Yoshinaga, K. (2003) The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the land plants. Nucleic Acids Research 31: 716-721.

Liere, K. and Link, G. (1995) RNA-binding activity of the matK protein encoded by the chloroplast trnK intron from mustard (*Sinapis alba* L.). Nucleic Acids Research 23: 917-921.

Maier, R.M., Neckermann, K., Igloi, G. and Kossel, H. (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. Journal of Molecular Biology 251: 614-628.

Matsuura, M., Noah, J.W. and Lambowitz, A.M. (2001) Mechanism of maturase-promoted group II intron splicing. The EMBO Journal 20: 7259-7270.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., Cavanagh, S., Dunny, G.M., Belfort, M. and Lambowitz, A.M. (1997) A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. Genes & Development 11: 2910-2924.

Michel, F., Umesono, K. and Ozeki, H. (1989) Comparative and functional anatomy of group II catalytic introns-a review. Gene 82: 5-30.

Mighell, A.J., Smith, N.R., Robinson, P.A. and Markham, A.F. (2000) Vertebrate pseudogenes. FEBS Letters 468: 109-114.

Mohr, G. and Lambowitz, A.M. (2003) Putative proteins related to group II intron reverse transcriptase/maturases are encoded by nuclear genes in higher plants. Nucleic Acids Research 31: 647-652.

Mohr, G., Perlman, P.S. and Lambowitz, A.M. (1993) Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Research 21: 4991-4997.

Moran, J.V., Mecklenburg, K.L., Sass, P., Belcher, S.M., Mahnke, D., Lewin, A., and Perlman, P.S. (1994) Splicing defective mutants of the COX1 gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Research 22: 2057-2064.

Moran, J.V., Zimmerly, S., Eskes, R., Kennell, J.C., Lambowitz, A.M., Butow, R.A. and Perlman, P.S. (1995) Mobile group II introns of yeast mitochondrial DNA are novel site specific retroelements. Molecular and Cellular Biology 15: 2828-2838.

Neuhaus, H. and Link, G. (1987) The chloroplast tRNA 157Lys (UUU) gene from mustard (*Sinapsis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. Current Genetics 11: 251-257.

Noah, J.W. and Lambowitz, A.M. (2003) Effects of maturase binding and Mg2+ concentration of group II intron RNA folding investigated by UV-cross-linking. Biochemistry 42: 12466-12480.

Olmstead, R.G. and Palmer J.D. (1994) Chloroplast DNA systematics: a review of methods and data analysis. American Journal of Botany 81: 1205-1224.

Ostheimer, G.J., Rojas, M., Hadivassiliou, H. and Barkan, A. (2006) Formation of the CRS2-CAF2 group II intron splicing complex is mediated by a 22 amino acid motif in the C-terminal region of CAF2. Journal of Biological Chemistry 281 (8): 4732–4738.

Ostheimer, G.J., Williams-Carrier, R., Belcher, S., Osborne, E., Gierke, J. and Barkan, A. (2003) Group II intron splicing factors derived by diversification of an ancient RNA-binding domain. EMBO Journal 22: 3919-3929.

Rambo, R.P. and Doudna, J.A. (2004) Assembly of an active group II intron-maturase complex by protein dimerization. Biochemistry 43: 6486-6497.

Rudi, K., Fossheim, T. and Jakobsen, K.S. (2003) Nested evolution of a tRNA (Leu) (UAA) group I intron by both horizontal intron transfer and recombination of the entire tRNA locus. Journal of Bacteriology 184: 666-671.

Saldanha, R., Chen, B., Wank, H., Matsuura, M., Edwards, J. and Lambowitz, A.M. (1999) RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry 38: 9069-9083.

Saldanha, R., Mohr, G., Belfort, M. and Lambowitz, A.M. (1993) Group I and group II introns. FASEB Journal 7:15-24.

Selvaraj, D., Sarma, R.K. and Staishkumar, R. (2008). Phylogenetic analysis of chloroplast matK gene from Zingiberaceae for plant DNA barcoding. Bioinformation 3(1): 24-27.

Singh, R.N., Saldanha, R.J., D'Souza, L.M. and Lambowitz, A.M. (2002) Binding of a group II intron-encoded reverse transcriptase/maturase to its high affinity intron RNA binding site involves sequence-specific recognition and autoregulates translation. Journal of Molecular Biology 318: 287-303.

Soltis, D.E. and Soltis, P.S. (1998) Choosing an approach and an appropriate gene for phylogenetic analysis. In: Soltis, D.E., Soltis, P.S., Doyle, J.J. (Eds.) Molecular Systematics of Plants II: DNA Sequencing. Kluwer Academic Publishers, Boston. pp. 2-31.

Sugita, M., Shinozaki, K. and Sugiura, M. (1985) Tobacco chloroplast tRNALys (UUU) gene contains a 2.5-kilobase-pair intron: an open reading frame and a conserved boundary sequence in the intron. Proceedings of the National Academy of Science, USA 82: 3557-3561.

Till, B., Schmitz-Linneweber, C., Williams-Carrier, R. and Barkan, A. (2001) CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin. RNA 7: 1227-1238.

Vogel, J., Borner, T. and Hess W. (1999) Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Research 27: 3866-3874.

Vogel, J., Hubschmann, T., Borner, T. and Hess, W.R. (1997) Splicing and intron-internal RNA editing of trnK-matK transcripts in barley plastids: support for MatK as an essential splicing factor. Journal of Molecular Biology 270: 179-187.

Wank, H., San Flippo, J., Singh, R.N., Matsuura, M. and Lambowitz, A.M. (1999) A reverse transcriptase/ maturase promotes splicing by binding at its own coding segment in a group II intron RNA. Molecular Cell 4: 239-250.

Whitten, W.M., Williams, N.H. and Chase, M.W. (2000) Subtribal and generic relationships of Maxillarieae (Orchidaceae) with emphasis on Stanhopeinae: combined molecular evidence. American Journal of Botany 87: 1842-1856.

Young, N.D. and dePamphilis, C.W. (2000) Purifying selection detected in the plastid gene matK and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Molecular Biology and Evolution 17:1933-1941.

*****

Plant DNA Barcoding and Phylogenetics, Pages 91-106
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

6 Retrotransposon-Based Plant DNA Barcoding

A.M. Alzohairy, G. Gyulai, M.M. Mostafa, S. Edris, J.S.M. Sabir, R.K. Jansen, Z. Tóth and A. Bahieldin

Introduction

Retrotransposons (RTs) are major components of most eukaryotic genomes; they are ubiquitous, dispersed throughout the genome, and their abundance correlates with the host genome sizes. Copy-and-paste life style of the RTs consists of three molecular steps, which involve transcription of an RNA copy from the genomic RT, followed by reverse transcription to cDNA, and finally a reintegration event into a new locus of the genome. This process leads to new genomic insertions without excision of the original RT. The target sites of insertions are relatively random and independent for different plant taxa; however, some elements cluster together in ‘repeat seas’ or have a tendency to cluster around the chromosome centromers and telomeres. The structure and copy number of retrotransposon families are strongly influenced by the evolutionary history of the host genome. Molecular barcoding of RTs play an essential role in all fields of genetics and genomics, and represent a powerful tool for molecular barcoding. To detect RT polymorphisms, marker systems generally rely on the amplification of sequences between the ends of the retrotransposon, such as long terminal repeats (LTR) of LTR-retrotransposons (LTR-RT) and the flanking genomic DNA.

Interspersed repetitive DNA sequences comprise a large fraction of the eukaryotic genomes. They predominantly consist of transposable elements (TEs) with two main families, Retrotransposons (Class I) and DNA transposons (Class II) (McClintock, 1984). Retrotransposons (RTs) are the most abundant class of TEs (IHGSC, 2001; Feschotte et al., 2002; Sabot and Schulman, 2006; Kalendar and Schulman, 2006).

There are two major groups of RTs based on the presence vs. absence of long terminal repeats (LTRs), LTR-retrotransposons (LTR-RTs) and non-LTR-retrotransposons. LTR-RTs comprise two main subgroups, copia (with high copy number) and gypsy (with high transposing activity) (Figure 1). Both, copia and gypsy LTR-RTs, carry regulatory sequences of gene promoters such as CAAT box (e.g., CCATT), TATA box (e.g., TGGCTATAAATAG), transcription start (e.g., CCCATGG), polyadenylation signal (e.g., AATAAG), and polyadenylation start (e.g., TAGT) (Ramallo et al., 2008). All these domains are required for replication and integration of RTs (Sabot and Schulman, 2006; Mansour, 2008). The large internal domain of the LTR-RTs encodes the structural proteins of the virus-like particle, which encapsulate the RNA copy of the RT, and the enzymes Reverse Transcriptase and Integrase (Figure 1). The process is called transposition.

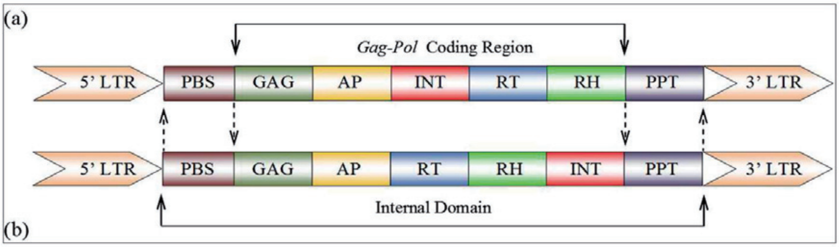


Fig. 1: Schematic representation of structural differences between Copia (a) and Gypsy (b) LTR-RT families. 5`LTR - 5`-end of long terminal repeat; PBS - primer binding site; GAG - group-specific antigen (syn.: capsid protein gene); AP - aspartic protease gene; INT - integrase gene; RT - reverse transcriptase gene; RH - ribonuclease-H gene; 3`LTR - 3`-end of the long terminal repeat; PPT - polypurine tract (Alzohairy et al., 2014b).

There are three further non-autonomous, short derivative, recombinant LTR-RTs, LARD (Large Retrotransposon Derivatives), TRIM (Terminal Repeat Retrotransposon in Miniature) and solo-LTR (sequence carrying 5’ and 3’ LTRs only) (Xiong and Eickbush, 1990; Havecker et al., 2004; Jurka et al., 2007). The size of LTR-RTs varies from long (e.g., Bare1 copia LTR-RT at 13,271 bp, NCBI Z17327) to short (e.g., Bare1 copia solo-LTR-RT at 3,130 bp, NCBI AB014756; and the truncated RLC_Lara copia RT; at 735 bp, NCBI EF067844; TREP2298). In plants, LTR-RTs are more plentiful and active than non-LTR-RTs (AGI, *Arabidopsis* Genome Initiative, 2000; Rice Chromosome 10 Sequencing Consortium, 2003; Alzohairy et al., 2012; 2013; 2014a,b). Due to the induction of chromosome recombinational processes during the meiotic prophase, active retrotransposons tend to lose their activity due to sequence breakage (Mansour, 2007**;** 2008; 2009; Alzohairy et al., 2012; 2013; 2014a,b).

Utilization of retrotransposons as molecular markers

Molecular barcoding methods (Schulman, 2007) based on RTs rely on PCR, and detect large portions of the genome (Kalendar et al., 1999; Kalendar and Schulman, 2006; Venturi et al., 2006; Branco et al., 2007; Sanz et al., 2007; Mansour, 2008; Mansour et al., 2010; Poczai et al., 2013). Barcode and marker systems based on different RTs show different levels of resolution and can be chosen to fit the identification of a given genome (Leigh et al., 2003; Queen et al., 2004; Ashalatha et al., 2005; Chadha and Gopalakrishna, 2005; Tam et al., 2005; Teo et al., 2005; Brik et al., 2006; Kalendar and Schulman, 2006). Retrotransposon-based markers follow Mendelian inheritance with high levels of genetic variability (Manninen et al., 2000; Huo et al., 2009).

Three different orientations of RTs are possible (i.e., head-to-head, tail-to-tail, and head-to-tail), either at a single locus, or inserted next to or within each other (nested RTs). This feature increases the variation available for revealing polymorphism within and among species. If the RT sequence and the adjacent genomic sequences are known, then all types of PCR-based molecular techniques can detect RT polymorphisms.

As the new cDNA copy of RT integrate into a new locus of the genome the old copy persist in the genome across generations, and the variation between ancestral and derived RT loci can be revealed (Mansour, 2008). The presence of a given retrotransposon suggests its orthologue integration, while the absence indicates the plesiomorphic condition prior to integration (Kalendar, 2011). The presence vs. absence of RTs can be utilized to construct phylogenetic trees of species due to the distribution of retrotransposons across organisms. This is the reason that RTs have been suggested to provide powerful phylogenetic markers with little if any homoplasy (Shedlock and Okada, 2000, Schulman, 2007).

Primer design for detection LTR-RTs

The LTR sequences are chosen to minimize the size of the target to be amplified. A primer facing outward from the 5’ LTR will necessarily face inward to a 3’ LTR of a neighboring LTR-RT, because the LTRs are direct repeats. The long sequences of LTR may also interfere with the production of amplicons within the size range of standard PCR. The conservative regions of LTR sequences are also used for designing inverted primers for Long-PCR, which can be used for cloning entire RTs and also for IRAP, REMAP and SSAP techniques.

IRAP primers are designed for using single or double primers

In REMAP, one primer is designed from the LTR and another from a nearby simple sequence repeats (microsatellites, syn.: SSRs). RBIP can detect both the presence and absence of the RT insertion using three primers to generate single-locus codominant markers. In SSAP, two primers are designed to produce amplification between RTs and adaptors ligated to a restriction site (usually MseI or PstI). In IPBS, primers are designed to match and amplify the conserved regions of the primer binding sequences (PBS). One or two primers can be used depending on the desired output of the experiments.

Retrotransposon-Based Insertion Polymorphism (RBIP)

RBIP (Flavell et al., 1998) detects retrotransposon insertions using a primer flanking the insertion site of the genome and another primer binding to the retrotransposon (Figure 2).

The basic RBIP was developed for high-throughput applications by replacing gel electrophoresis with hybridization to a filter, and was developed by studying the PDR1 retrotransposon in *Pisum sativum* (Flavell et al., 1998). One of the disadvantages of this method is that it is more expensive and technically demanding compared to other methods. The method also allows the dot blot approach to be scaled down to microarrays with the attendant advantages in throughput using sensitive oligo-based hybridization to spotted PCR products (Flavell et al., 1998). RBIP requires information on the sequences of the 5’ and 3’ flanking regions of the retrotransposon insertions. One limitation of RBIP is due to size range of standard PCR (about 3-5 Kbp).

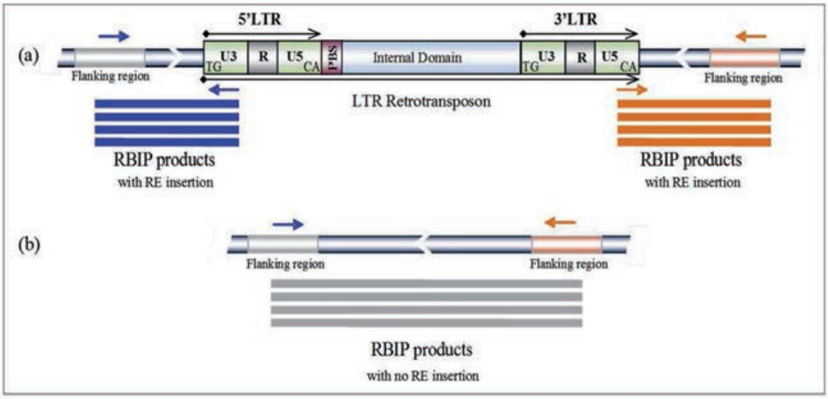


Fig. 2: RBIP (Retrotransposon-Based Insertion Polymorphism; Flavell et al., 1998) detects the presence (a) or absence (b) of retrotransposons in the host genome. Amplification takes place between retrotransposons (3` or 5` LTRs) and proximal flanking region of the genome. The alternative reaction takes place between the primers from the left and right flanks, which is inhibited in the full (RT-occupied) site by the length of retrotransposon, and able to amplify the shorter, empty (RT-unoccupied) site. (Primers are indicated as color arrows) (Alzohairy et al., 2014b).

By using three primers, RBIP can detect both the presence and absence of the TE insertion and generates single-locus codominant markers. RBIP can also generate a dominant marker when only two flanking primers are used (Ribaut and Hoisington, 1998). When RBIP detects the occupied and unoccupied RT sites together, the products blotted onto membrane are probed with a locus-specific probe. Empty sites are usually scored by amplification between the left and right flanks of the presumptive integration site with primers specific to both flanking regions. This method can detect genomic polymorphisms by using standard agarose gel electrophoresis, or by hybridization, which is more useful for automated and high throughput analysis. RBIP was successfully used to generate molecular barcodes to examine the evolutionary history among *Pisum* species (Vershinin et al., 2003; Jing et al., 2005).

Retrotransposon-Microsatellite Amplified Polymorphism (REMAP)

REMAP (Kalendar et al., 1999) combines primers (Figure 3) to RTs and locus-specific simple sequence repeats (SSRs) of the genome (Kalendar and Schulman, 2006; Mansour, 2008; Kalendar, 2011). This technique is applicable when SSR locates near the retrotransposons (Tsumura et al., 1996; Mansour, 2008; Kalendar, 2011). Amplification between retrotransposon and a nearby SSR requires neither digestion with restriction enzymes nor adaptor ligation to generate the marker bands. This protocol can be completed in 1-2 days (Kalendar and Schulman, 2006; Mansour, 2008, Kalendar, 2011) and has been used to measure diversity, similarity and cladistic relationships in many genotypes of *Oryza sativa* (Branco et al., 2007), rice pathogens (*Magnaporthe grisea*) (Chadha and Gopalakrishna, 2005), *Spartina* sp. (Baumel et al., 2002) and *Avena sativa* (Tanhuanpää et al., 2007).

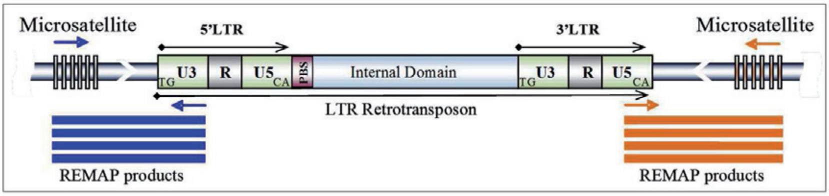


Fig. 3: REMAP (REtrotransposon-Microsatellite Amplified Polymorphism; Kalendar et al., 1999) amplifies genomic DNA stretches between LTRs of the LTR-RT and a nearby microsatellite (vertical bars). (Primers are indicated as color arrows) (Alzohairy et al., 2014b).

Sequence-Specific Amplified Polymorphism (SSAP)

SSAP (Waugh et al., 1997) analysis (Figure 4) was one of the first retrotransposon-based barcoding methods relying on AFLP (amplified fragment length polymorphism) (Vos et al., 1995). SSAP utilized the BARE-1 LTR-RT for molecular barcoding (Waugh et al., 1997) using one primer matching the end of an RT (e.g., 3’ LTR) and the other matching an AFLP-like restriction site (usually MseI or PstI) adaptor. Primer pairs contains two or three selective nucleotides of MseI or PstI (or any restriction enzyme) adaptor primers and one nt selective nucleotide of either ^{32}P- or fluorescently-labeled retrotransposon-specific primers (Ellis et al., 1998).

SSAP primers are often designed to the LTR region, but could also match to an internal sequence of the RT, like the polypurine tract (PPT), which is found internal to the 3'-LTR of retrotransposons (Ellis et al., 1998). Non-selective primers could also be used when restriction enzymes have a long recognition site sequence, or when the copy number of the RTs is low. The number of selected bases may be increased in the case of high-copy-number families. The use of single or double enzyme digestions (or infrequent cutting enzymes) allows the survey of all insertion sites for a given RT, and can be considered as a variant of anchored PCR. The quality of SSAP pattern depends on the SSAP primers used. Primers that give highly polymorphic, clear, and reproducible SSAP banding patterns are candidate primers for subsequent work. Amplified fragments are commonly separated on 6% polyacrylamide sequencing gels and visualized by autoradiograph.

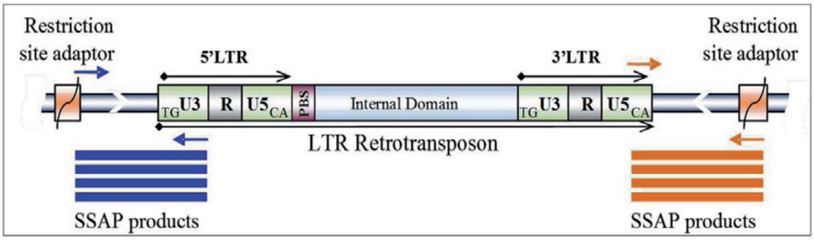


Fig. 4: SSAP (Sequence-Specific Amplified Polymorphism; Waugh et al*.,* 1997) amplifies sequence region between the retrotransposon and a restriction site anchored by an adaptor. Primers (color arrows) used for amplification match the adaptor (broken box) and the retrotransposon (in the LTR box, *e.g.,* U3', R, and U5'). (Alzohairy et al., 2014b).

SSAP usually displays a higher level of polymorphism as compared to AFLP (Ellis et al., 1998; Nagy et al., 2006; Syed et al., 2006; Venturi et al., 2006), and has been extensively used in *Hordeum vulgare* (Leigh et al., 2003), *Triticum* spp. (Queen et al., 2004), *Aegilops* spp. (Nagy et al., 2006), *Avena sativa* (Yu and Wise, 2000), *Malus domestica* (Venturi et al., 2006), *Cynara cardunculus* (Lanteri et al., 2006), *Lactuca sativa* (Syed et al., 2006), *Pisum sativum* and other Fabaceae (Ellis et al., 1998; Jing et al*.,* 2005), *Capsicum annuum* and *Solanum lycopersicum* (Tam et al., 2005) and *Ipomoea batatas* (Tahara et al., 2004).

SSAP was also used for cladistic molecular barcodes to resolve evolutionary history in *Nicotiana* (Petit et al., 2007), *Vicia* (Sanz et al*.,* 2007), *Oryza* (Gao et al., 2004), *Triticum* (Queen et al., 2004) and *Zea* (García-Martínez and Martínez-Izquierdo, 2003).

Inter-Retrotransposons Amplified Polymorphisms (IRAP)

There are many techniques that are based on inter-repeat amplification polymorphism such as REMAP (Kalendar et al., 1999 ; Kalendar and Schulman, 2006), inter-MITE amplification, and IRAP (Kalendar et al., 1999) (Figure 5). IRAP is based on the fact that retrotransposons generally cluster together in 'repeat seas' surrounding 'genome islands', and may be nested within each other (Kalendar et al., 1999; Mansour, 2008). By this way, IRAP detects insertional polymorphisms of retrotransposons by amplifying the DNA sequences of two neighboring retroelements such as LTR-RTs and SINE-like sequences (Kalendar et al., 1999).

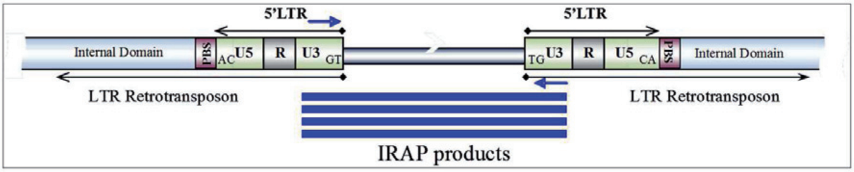


Fig. 5: IRAP (Inter-Retrotransposons Amplified Polymorphisms; Kalendar et al., 1999) amplifies genomic DNA stretches between abundant dispersed repeats, such as the LTRs of two LTR-RTs. The primers (color arrows) point outwards from the LTRs of LTR-RTs to amplify region between two LTR-RTs (Alzohairy et al., 2014b).

IRAP does not require restriction enzyme digestion or ligation (Kalendar and Schulman, 2006; Mansour, 2008; Kalendar, 2011). Different retrotransposon insertions increase the number of sites amplified and sizes of inter-RT fragments, which can be used as marker to detect genotype polymorphism.

One or two PCR primers can be used for IRAP. The primers should be pointing outwards from the LTRs of RT to amplify the region between two RTs (Kalendar, 2011). The two primers could be designed to either the same or different RT families. IRAP can be also carried out with a single primer, which matches either the 5' or 3' end of the LTR but oriented away from the LTR itself. The copy number of RTs, size and insertion pattern can affect the complexity of the fingerprinting pattern (Mansour, 2008; Mansour et al., 2010). The pattern obtained with two primers does not likely represent simply the sum of the products obtained with each primer individually. In the case of retrotransposons dispersed within the genome, IRAP produces too many fragments to give good resolution on gels, or no products because target amplification sites are too far apart to generate amplicons. Yet, IRAP overcomes some of the drawbacks of other techniques. Unlike SSAP, IRAP does not require either radioactivity or fluorescent labeling of primers. The method was used widely for BARE-1 RT of the *Hordeum vulgare* genome to measure diversity of genotypes (Kalendar et al., 1999; Manninen et al., 2000, 2006). IRAP was also used for barcoding of genotypes of *Oryza sativa* (Branco et al., 2007), *Musa* (Ashalatha et al., 2005; Teo et al., 2005), *Brassica* (Tatout et al., 1999), *Spartina* (Baumel et al., 2002), *Triticum* (Boyko et al., 2002) and *Solanum* (Mansour et al., 2010).

Inter Primer Binding Sequence (IPBS)

IPBS method (Kalendar et al., 2010) is frequently used for displaying retrotransposon polymorphisms (Figure 6). The need for sequence information to design IPBS primers is the case in all RT-based molecular barcoding techniques. IPBS tends to overcome this problem (Kalendar et al., 2010) as the primer binding sequence (PBS) is part of the internal domain of retrotransposons. IPBS utilizes the highly conserved regions of PBS site for tRNAs (Kalendar et al., 2010). While the process of reverse transcription is conserved among all retroviruses, the specific tRNA capture varies for different retroviruses and retrotransposons. Thus, the IPBS amplification method can be useful for all retroviruses that contain conservative PBS sites for tRNAi^{Met}, tRNA^{Lys}, tRNA^{Pro}, tRNA^{Trp}, tRNA^{Asn}, tRNA^{Ser}, tRNA^{Arg}, tRNA^{Phe}, tRNA^{Leu} or tRNA^{Gln} (Kalendar et al., 2010). As in plant species RTs are nested, mixed, inverted or truncated in the geneome, RTs can be easily amplified using conservative PBS primers. PCR amplification occurs between two nested PBSs of two neighboring LTR-RTs (Figure 6).

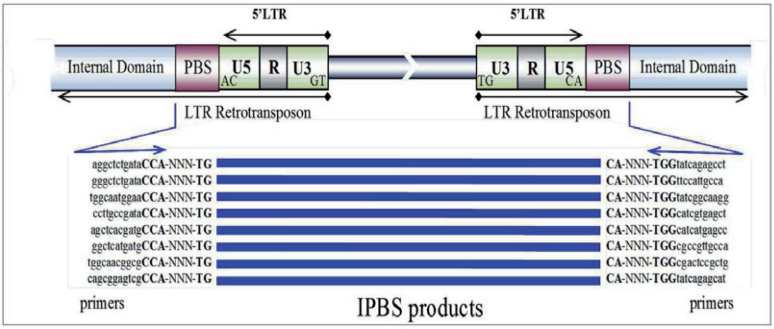


Fig. 6: IPBS (Inter Primer Binding Sequence; Kalendar et al., 2010) method utilizes the conserved sequence of PBS of LTR-RTs for screening retrotraposososns. Sequences shown are conserved regions of PBS used for primer (color arrows) design (Alzohairy et al., 2014b).

PBS sequences can also be used for detecting other retrotransposons when the retrotransposon density is high within the genome (Kalendar, 2011). Retrotransposon movements and recombinations can also be monitored because new inserts or recombinations will be polymorphic, which will appear only in plant lines in which the insertions or recombinations have taken place.

In conclusion, several retrotransposon-based molecular barcoding systems were developed based on PCR amplifications between sequences of RTs and the flanking DNA of the host genome (Kalendar and Schulman, 2006). These marker systems were found to be highly effective tools for tracking transpositions and diversities of RTs, and determining phylogenetic relationships of plant taxa (Hamdi et al*.,* 1999; Shedlock and Okada, 2000). Many reports also suggest that the differences in genome size observed in the plant kingdom are related to variations of RTs activity and consequently their content, which suggests that RTs play important roles in the evolution of genome sizes (Vitte and Panaud, 2005; Alzohairy et al., 2012; 2013; 2014a,b). Other studies used LTR-RT barcoding detected the effects of environmental stresses on the re-activation of retrotransposons and hence their genetic diversity (reviewed in Alzohairy et al., 2014a). Many applications were also reported for study of phylogeny, genetic diversity and the functional analyses of genes using LTR-RT based barcoding (Waugh et al., 1997; Flavell et al., 1998; Kalendar and Schulman, 2006; Mansour, 2008; Roos et al., 2004).

References

AGI (2000) *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796-815.

Alzohairy, A.M., Gyulai, G., Jansen, R.K. and Bahieldin, A. (2013) Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid 69: 1-15.

Alzohairy, A.M., Gyulai, G., Ramadan, M.F., Edris, S., Sabir, J.S.M., Jansen, R.K., Eissa, H.F. and Bahieldin, A. (2014b) Retrotransposon-based molecular markers for assessment of genomic diversity. Functional Plant Biology 41(8) 781-789.

Alzohairy, A.M., Sabir, J.S.M., Gyulai, G., Younis, R., Jansen, R.K. and Bahieldin, A. (2014a) Environmental stress activation of plant LTR-retrotransposons. Functional Plant Biology 41(6): 557-567.

Alzohairy, A.M., Yousef, M.A., Edris, S.S., Kerti, B. and Gyulai, G. (2012) Detection of long terminal repeat (LTR) retrotransposons reactivation induced by in vitro environmental stresses in barley (*Hordeum vulgare*) via reverse transcription-quantitative polymerase chain reaction (RT-qPCR). Life Science Journal 9: 5019-5026.

Ashalatha, S.N., Teo, C.H., Schwarzacher, T. and Heslop-Harrison, J.S. (2005) Genome classification of banana cultivars from South India using IRAP markers. Euphytica 144: 285-290.

Baumel, A., Ainouche, M., Kalendar, R. and Schulman, A.H. (2002) Inter-retrotransposon amplified polymorphism (IRAP), and retotransposon-microsatellite amplified polymorphism (REMAP) in populations of the young allopolyploid species *Spartina angelica* Hubbard (Poaceae). Molecular Biology and Evolution 19: 1218-1227.

Boyko, E., Kalendar, R., Korzun, V., Gill, B. and Schulman, A.H. (2002) Combined mapping of *Aegilops tauschii* by retrotransposon, microsatellite, and gene markers. Plant Molecular Biology 48: 767-790.

Branco, C.J.S., Vieira, E.A., Malone, G., Kopp, M.M., Malone, E., Bernardes, A., Mistura, C.C., Carvalho, F.I.F. and Oliveira, C.A. (2007) IRAP and REMAP assessments of genetic similarity in rice (*Oryza sativa*). Journal of Applied Genetics 48: 107-113.

Brik, A.F., Kalendar, R.N., Stratula, O.R., Sivolap and Yu, M. (2006) IRAP and REMAP analyses of Barley (*Hordeum vulgare*) varieties of Odessa breeding. Cytology and Genetics 3: 24-33.

Chadha, S. and Gopalakrishna, T. (2005) Retrotransposon-microsatellite amplified polymorphism (REMAP) markers for genetic diversity assessment of the rice (*Oryza sativa*) blast pathogen (*Magnaporthe grisea*). Genome 48: 943-945.

Ellis, T.H.N., Poyser, S.J., Knox, M.R., Vershinin, A.V. and Ambrose, M.J. (1998) Polymorphism of insertion sites of *Ty*1-copia class retrotransposon insertion site polymorphism for linkage and diversity analysis in pea. Molecular & General Genetics 260: 9-19.

Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. Nature reviews. Genetics 3: 329-341.

Flavell, A.J., Knox, M.R., Pearce, S.R. and Ellis, T.H.N. (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. The Plant Journal: for Cell and Molecular Biology 16: 643-650.

Gao, L., McCarthy, E.M., Ganko, E.W. and McDonald, J.F. (2004) Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice (*Oryza sativa*) genome sequences. BMC Genomics 2: 18.

García-Martínez, J. and Martínez-Izquierdo, J.A. (2003) Study on the evolution of the grande retrotransposon in the *Zea* genus. Molecular Biology and Evolution 20: 831-841.

Hamdi, H., Nishio, H., Zielinski, R., and Dugaiczyk, A. (1999) Origin and phylogenetic distribution of *Alu* DNA repeats: irreversible events in the evolution of primates. Journal of Molecular Biology 289: 861-871.

Havecker, E.R., Gao X. and Voytas, D.F. (2004) The diversity of LTR retrotransposons. Genome Biology 5: 225.

Huo, H., Conner, J.A. and Ozias-Akins, P. (2009) Genetic mapping of the apospory-specific genomic region in *Pennisetum squamulatum* using retrotransposon-based molecular markers. Theoretical and Applied Genetics 119: 199-212.

IHGSC (International Human Genome Sequencing Consortium) (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Jing, R., Knox, M.R., Lee, J.M., Vershinin, A.V., Ambrose, M., Ellis, T.H.N. and Flavell, A.J. (2005) Insertional polymorphism and antiquity of PDR1 retrotransposon insertions in *Pisum* species. Genetics 171: 741-752.

Jurka, J., Kapitonov, V., Kohany, O. and Jurka, M.I.V. (2007) Repetitive sequences in complex genomes: structure and evolution. Annual Review of Genomics and Human Genetics 8: 241-259.

Kalendar, R. (2011) The use of retrotransposon-based molecular markers to analyze genetic diversity. Field and Vegetable Crops Research 48: 261–274.

Kalendar, R. and Schulman, H.A. (2006) IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nature Protocols 1: 2478-2484.

Kalendar, R., Antonius, K., Smykal, P. and Schulman, A.H. (2010) iPBS: A universal method for DNA fingerprinting and retrotransposon isolation. Theoretical and Applied Genetics 121: 1419-1430.

Kalendar, R., Grob, T., Regina, M., Suomeni, A. and Schulman, A. (1999) IRAP and REMAP two new retrotransposon-based DNA fingerprinting techniques. Theoretical and Applied Genetics 98: 704-711.

Lanteri, S., Acquadro, A., Comino, C., Mauro, R., Mauromicale, G. and Portis, E. (2006) A first linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus* L.) based on AFLP, S-SAP, M-AFLP and microsatellite markers. Theoretical and Applied Genetics 112: 1532-1542.

Leigh, F., Kalendar, R., Lea, V., Lee, D., Donini, P. and Schulman, A.H. (2003) Comparison of the utility of barley (*Hordeum vulgare*) retrotransposon families for genetic analysis by molecular marker techniques. Molecular Genetics and Genomics 269: 464-474.

Manninen, O., Kalendar, R., Robinson, J. and Schulman, A.H. (2000) Application of BARE-1 retrotransposons markers to the mapping of a major resistance gene for net blotch in carley (*Hordeum vulgare*). Molecular & General Genetics 264: 325-334.

Manninen, O.M., Jalli, M., Kalendar, R., Schulman, A., Afanasenko, O. and Robinson, J. (2006) Mapping of major spot-type and net-type netblotch resistance genes in the Ethiopian barley (*Hordeum vulgare*) line CI 9819. Genome 49: 1564-1571.

Mansour, A. (2007) Epigenetic activation of genomic retrotransposon. Journal of Cell and Molecular Biology 6: 99-107.

Mansour, A. (2008) Utilization of genomic retrotransposon as cladistic molecular markers. Journal of Cell and Molecular Biology 7: 17-28.

Mansour, A. (2009) Water deficit induction of *Copia* and *Gypsy* genomic retrotransposons. Plant Stress 3: 33-39.

Mansour, A., Jaime, A., da Silva, T., Edris, S. and Younis, R.A.A., (2010) Comparative assessment of genetic diversity in some tomato cultivars using IRAP, ISSR and RAPD molecular markers. Genes, Genomes and Genomics 4(1):, 41-47.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science 226: 792-801.

Nagy, E.D., Molnar, I., Schneider, A., Kovacs, G. and Molnar-Lang, M. (2006) Characterization of chromosome-specific S-SAP markers and their use in studying genetic diversity in *Aegilops* species. Genome 49: 289-296.

Petit, M., Lim, K.Y., Julio, E., Poncet, C., Dorlhac de Borne, F., Kovarik, A., Leitch, A.R., Grandbastien, M.A. and Mhiri, C. (2007) Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). Molecular Genetics and Genomics 278: 1-15.

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J.P.T. and Hyvönen, J. (2013) Advances in plant gene-targeted and functional markers: a review. Plant Methods 9: 6.

Queen, R.A., Gribbon, B.M., James, C., Jack, P. and Flavell, A.J. (2004) Retrotransposon based molecular markers for linkage and genetic diversity analysis in wheat. Molecular Genetics and Genomics 271: 91-97.

Ramallo, E., Kalendar, R., Schulman, A.H. and Martinez-Izquierdo, J.A. (2008) *Reme*1, a *Copia* retrotransposon in melon, is transcriptionally induced by UV light. Plant Molecular Biology 66: 137-150.

Ribaut J.-M. and Hoisington D.A. (1998) Marker assisted selection: new tools and strategies. Trends in Plant Science 3: 236-239.

Roos, C., Schmitz, J. and Zischler, H. (2004) Primate jumping genes elucidate strepsirrhine phylogeny. Proceedings of the National Academy of Sciences of the USA 101: 10650-10654.

Sabot, F. and Schulman, A.H. (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97: 381-388.

Sanz, A.M., Gonzalez, S.G., Syed, N.H., Suso, M.J., Saldaña, C.C. and Flavell, A.J. (2007) Genetic diversity analysis in *Vicia* species using retrotransposon-based SSAP markers. Molecular Genetics and Genomics 278: 433-441.

Schulman A.H. (2007) Molecular markers to assess genetic diversity. Euphytica 158: 313-321.

Shedlock, A.M. and Okada, N. (2000) SINE insertions: Powerful tools for molecular systematics. BioEssays 22: 148-160.

Syed, N.H., Sørensen, A.P., Antonise, R., van de Wiel, C., van der Linden, C.G., van't Westende, W., Hooftman, D.A., den Nijs, H.C. and Flavell, A.J. (2006) A detailed linkage map of lettuce based on SSAP, AFLP and NBS markers. Theoretical and Applied Genetics 112: 517-527.

Tahara, M., Aoki, T., Suzuka, S., Yamashita, H., Tanaka, M., Matsunaga, S. and Kokumai, S. (2004) Isolation of an active element from a high-copy-number family of retrotransposons in the sweet potato genome. Molecular Genetics and Genomics 272: 116-127.

Tam, S.M., Mhiri, C., Vogelaar, A., Kerkveld, M., Pearce, S.R. and Grandbastien, M.A. (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theoretical and Applied Genetics 110: 819-831.

Tanhuanpää, P., Kalendar, R., Schulman, A.H. and Kiviharju, E. (2007) A major gene for grain cadmium accumulation in oat_(*Avena sativa* L.). Genome 50: 588-594.

Tatout, C., Warwick, S., Lenoir, A. and Deragon, J.-M. (1999) Sine insertions as clade markers for wild *Crucifer* species. Molecular Biology and Evolution 16: 1614-1621.

Teo, C.H., Tan, S.H., Ho, C.L., Faridah, Q.Z., Othman, Y.R., Heslop-Harrison, J.S., Kalendar, R. and Schulman, A.H. (2005) Genome constitution and classification using retrotransposon-based markers in the orphan crop banana. Journal of Plant Biology 48: 96-105.

The-Rice-Chromosome-10-Sequencing-Consortium (2003): In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* 300(5625):1566-1569.

Tsumura, Y., Ohba, K. and Strauss, S.H. (1996) Diversity and inheritance of inter-simple sequence repeat polymorphisms in Douglas-fir (*Pseudotsuga menziesii*) and sugi (*Cryptomeria japonica*). Theoretical and Applied Genetics 92: 40-45.

Venturi, S., Dondini, L., Donini, P. and Sansavini, S. (2006) Retrotransposon characterisation and fingerprinting of apple clones by S-SAP markers. Theoretical and Applied Genetics 112: 440-444.

Vershinin, A.V., Alnutt, T.R., Knox, M.R., Ambrose, M.R. and Ellis, T.H.N. (2003) Transposable elements reveal the impact of introgression, rather than transposition, in *Pisum* diversity, evolution and domestication. Molecular Biology and Evolution 20: 2067-2075.

Vitte, C. and Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenetic and Genome Research 110: 91-107.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. and Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 11: 4407–4414.

Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Kumar, A., Thomas, B.T. and Powell, W. (1997) Genetic distribution of BARE-1 retrotransposable elements in the barley (*Hordeum vulgare*) genome revealed by sequence-specific amplification polymorphisms (S-SAP). Molecular & General Genetics 253, 687-694.

Xiong, Y. and Eickbush, T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO Journal 9: 3353-3362.

Yu, G.-X. and Wise, R.P. (2000) An anchored AFLP- and retrotransposon-based map of diploid *Avena*. Genome 43: 736-749.

*****

Plant DNA Barcoding and Phylogenetics, Pages 107-120
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

7 Isoenzymes as Molecular Markers

G. Jahnke

Introduction

Molecular markers are biomolecules i.e. proteins or DNAs which varies among the different individuals of a population. The molecular markers generally have no apparent effect on phenotypes, but they can be detected by different molecular methods, and are used for different purposes in genetics and breeding (Figure 1). Isoenzymes or isozymes are enzymes that differs in electrophoretic feature (multiple forms), but catalyse the same biochemical reaction. The different electrophoretic attribute can be traced back to different amino acid sequence (primary structure), which causes different size, shape and/or charge. Different amino acid sequences in an enzyme cannot cause different electrophoretic feature in all the time. In other words, isoenzymes are different, but do the same work. The term “isozyme” was introduced by Market and Moller in 1959 to describe different molecular forms of enzymes with the same substrate specificity. The differences in primary structure can cause differences in secondary and tertiary structure of protein influence the size, shape and charge of it. In most of the cases such information are not available, but electrophoretic analyses can distinguish, and genetic analyses can presume the genetic origin (Hajósné Novák, 1999).

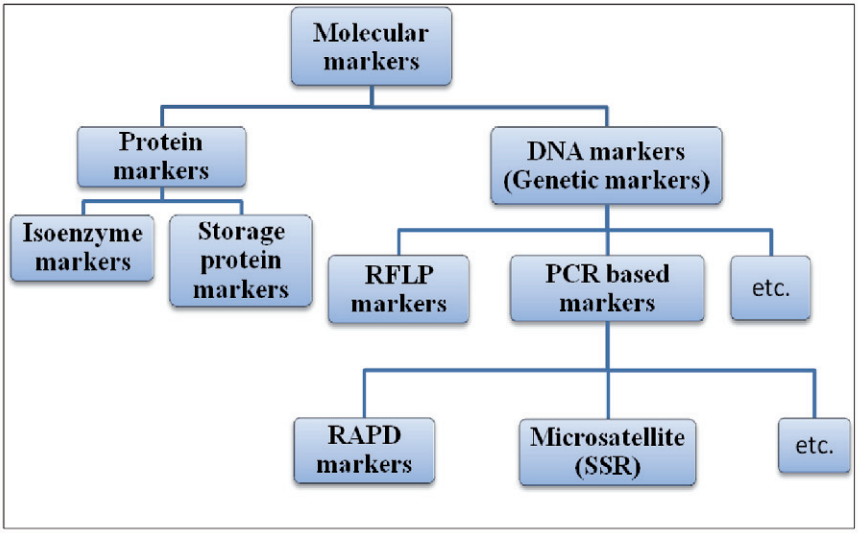


Fig. 1: The different types of molecular markers.

Salient features of isoenzymes

Developmental stage specificity: Isoenzymes are developmental stage specific, as from the same organism and tissue, different isoenzymes can be find in different developmental stages. For example differences in isozyme expression based on stage of development were detected in phosphoglucomutase (PGM) and 6-phosphogluconate dehydrogenase (PGD) systems in peach (*Prunus persica*) seeds (Figure 2) during stratification (Mowrey and Werner, 1990).

Tissue specificity: In the different tissues in an organism, different isoenzymes can present at the same time. For example in tobacco (*Nicotiana tabacum*) leaf, root, pith, and callus tissues express different peroxidise isoenzymes. Root tissue expresses all of the detectable isozymes, whereas each of the other tissues examined expressed a different subset of these isozymes (Lagrimini and Rothstein, 1987).

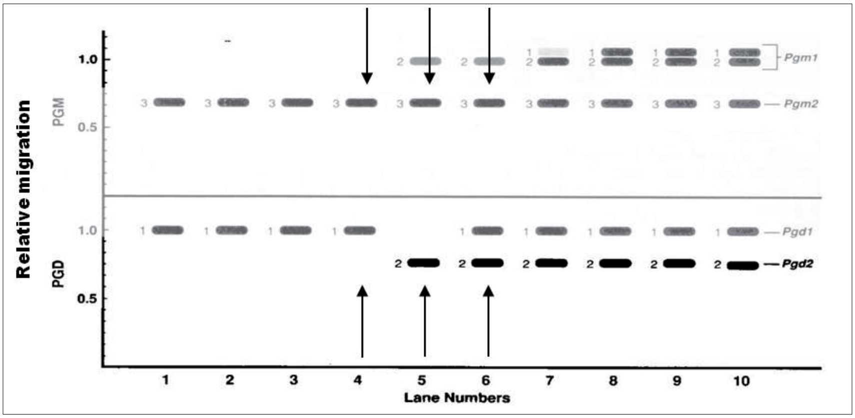


Fig. 2: Diagrammatic representation of banding patterns observed for phosphoglucomutase (PGM) and 6-phosphogluconate dehydrogenase (PGD) in dry (1) and imbibed seed (2); seed after 1 week (3), 1.5 months (4), 2.5 months (5), and 3 months (6) of stratification; cotyledon tissue (7) from 1-month-old seedling; leaf tissue of l-month- (8) and 3-month-old (9) seedlings; and leaf tissue of adult plant (10). Numbers in parentheses refer to lane numbers (Mowrey and Werner, 1990)

Cultivar specificity: From the different cultivars or varieties of on species, different isoenzymes can express. This can ensure cultivar identification in cultivated plants with high isoenzyme polymorphism such as grape (*Vitis vinifera*).

Origin, structure and groups of isoenzymes

- **Multilocus isoenzymes:** Multiple forms of an enzyme can be coded by different genes (loci). During the evolution, these multiple loci can be formed by gene duplication, which can be caused by unequal crossing-over (Figure 3.). The other possible way of evolution to generate multilocus isoenzymes is, that the mutation of originally different genes eventuate similar catalytic functions.

- **Allelic isoenzymes (allozymes):** The allelic isoenzymes or allozymes are coded by one locus. Different mutation effects in the affected locus could cause different alleles of isoenzymes during the evolution.

- **Secondary isoenzymes:** The secondary isoenzymes are coded by the same allele of the same locus, enzymes are modified during the translation. They can play important role in gene regulation.

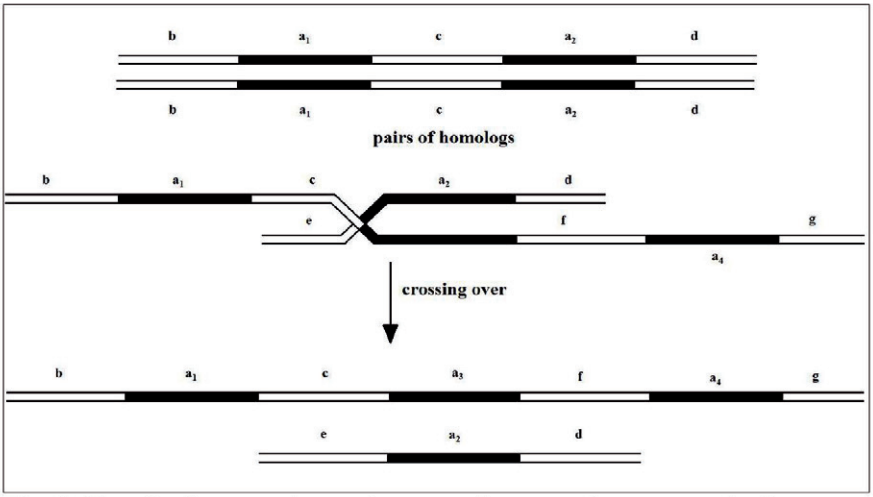


Fig. 3: Result of unequal crossing-over (increase in gene copies), a_1, a_2, a_3, a_4 homolog sequences. (Hajósné Novák, 1999)

Detection of isoenzymes

Isoenzymes can be separated based on their different electrophoretic features- by gel electrophoresis and isoelectric focusing; and identify based on their same (or similar) catalytic features, and by special histochemical staining protocols. All of the isoenzymes show a unique pattern in the gel called as zymogram (Figure 4). The interpretation of isozyme zymograms is shown Figure 5.

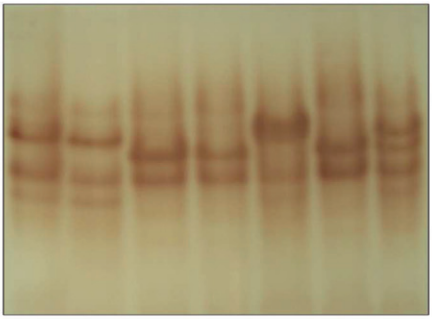


Fig. 4: Zymograms of cathecol oxidase isoenzymes of 7 grapevine (*Vitis vinifera*) cultivars in polyacrylamide gels.

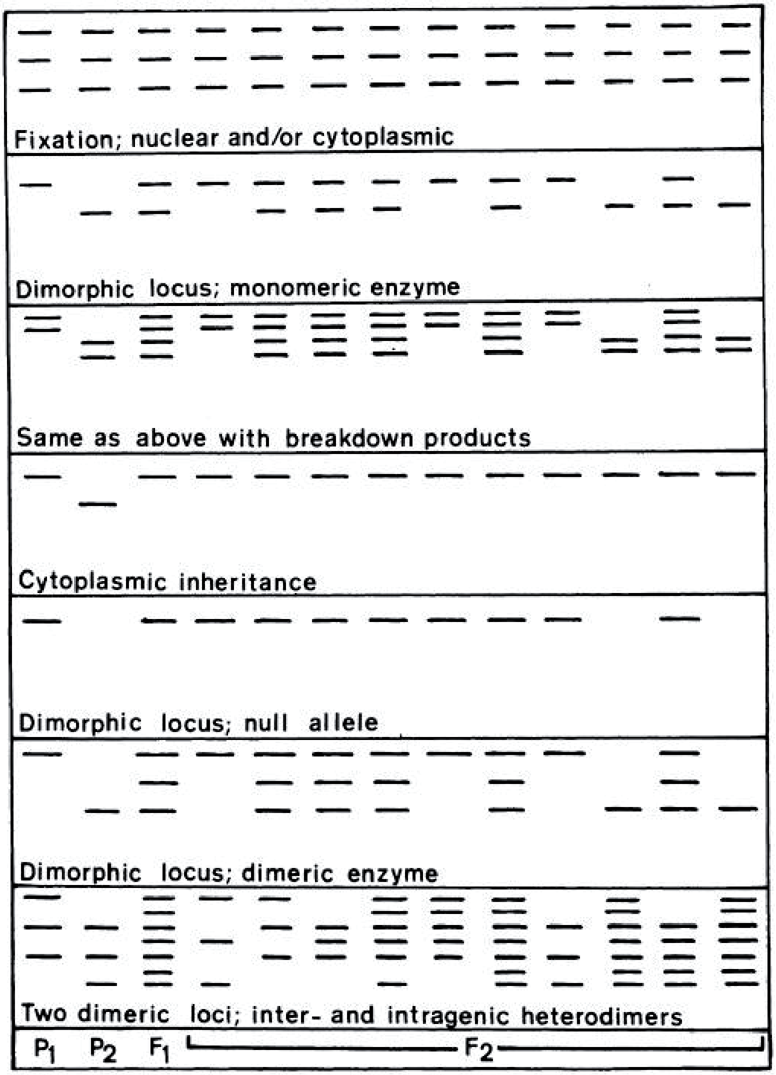

Fixation; nuclear and/or cytoplasmic
Dimorphic locus; monomeric enzyme
Same as above with breakdown products
Cytoplasmic inheritance
Dimorphic locus; null allele
Dimorphic locus; dimeric enzyme
Two dimeric loci; inter- and intragenic heterodimers
P1 P2 F1 F2

Fig. 5: Genetic interpretations of electrophoretic variations. (Shields et al., 1983)

Gel electrophoresis: Gel electrophoresis is a method for separation and analysis of macromolecules (DNA, RNA and proteins) and their fragments, based on their size and charge. The separation power is the difference between the voltages of the two ends of the gel. Movements of proteins in the electric field are effected by their size, shape and charge. It is important to keep activity of the enzymes during the electrophoresis, which is sometimes difficult and needs optimisation. In isoenzyme analyses horizontal (starch) or vertical (polyakrylamide) gel electrophoresis are used. Advantages of starch gel electrophoreses are: easier preparing procedure, both of positive and negative migration of the samples are possible, as the samples starts from the middle of the gel (Figure 6). Different enzymes can be analysed in one time, because the gel can be cut into thin slices after electrophoresis. Disadvantage is the lower resolution, compared by polyacrylamide gel electrophoresis.

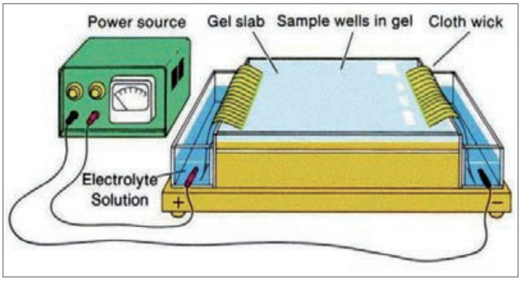


Fig. 6: Apparatus setup starch gel electrophoresis (Source: http://www.cas.miamioh.edu/~wilsonkg/old/gene2005/gene/mendexceptns/f4p3.jpg)

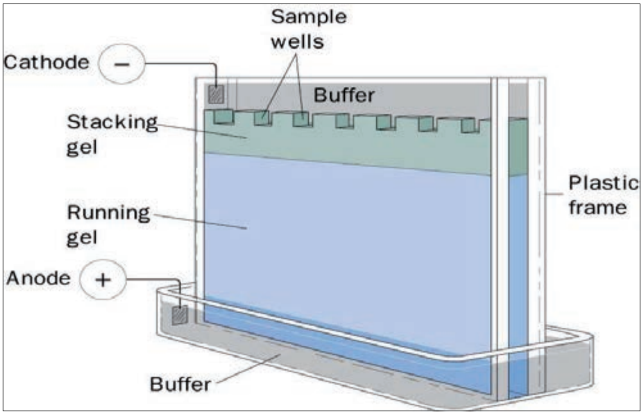


Fig. 7: Schematic view of polyacrylamide gel electrophoresis (Source: http://www.siumed.edu/~bbartholomew/course_material/protein_methods.htm)

Acrylamide is a synthetic hydrophilic material, which polymerase into gel in the presence of bisacrylamide and free radicals. This gel has a homogenous structure; therefore the resolution of the gel is high. Disadvantages of polyacrylamide gel electrophoresis are that the chemicals and the gels are toxic and/or carcinogen and proteins can migrate only in one direction (Figure 7). To precisely identify isoenzyme variants, one can calculate the relative mobility (Rf value) of an isoenzyme band as the movement of the band through a gel relative to the dye front (Figure 8).

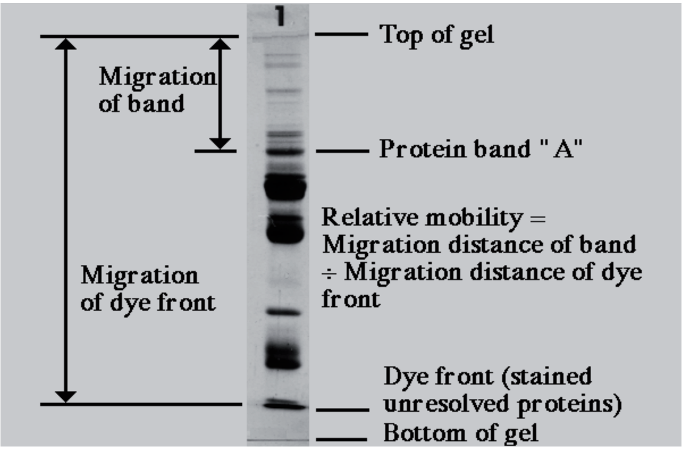

1
Top of gel
Migration of band
Protein band "A"
Relative mobility = Migration distance of band ÷ Migration distance of dye front
Migration of dye front
Dye front (stained unresolved proteins)
Bottom of gel

Fig. 8: Calculation of relative mobility. (Source: http://www.ruf.rice.edu/~bioslabs/studies/sds-page/rf.html)

Isoelectric focusing: Isoelectric focusing (IEF) is an electrophoretic technique for the separation of proteins based on their isoelectric point (pI or IEP). The pI is the pH at which a protein has no net charge and thus, does not migrate further in an electric field (Figure 9).

The separation of proteins during isoelectric focusing is supervened in a gel, where a pH gradient was formed before electrophoresis. The pH gradient can be formed by adding an ampholyte solution into the gel mixture before polymerisation. It needs practice and exactitude. IEF gels are made of acrylamide or agarose. The most important advantage of isoelectric focusing is that it can identify isoenzyme variants by their pI value, which is more precise method than the calculation of Rf values in the case of standard gel electrophoresis techniques.

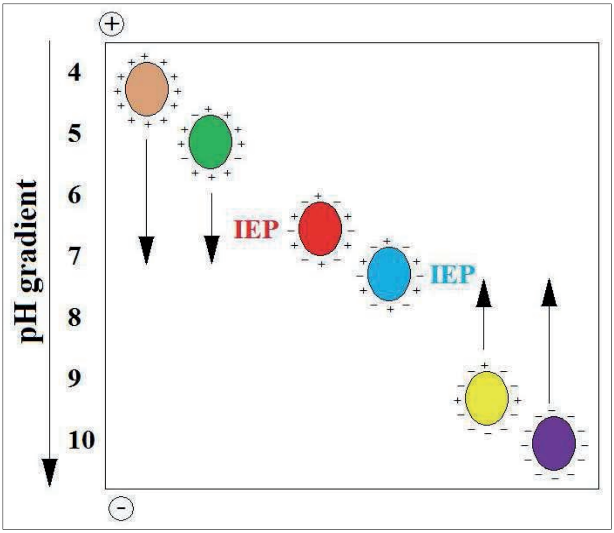
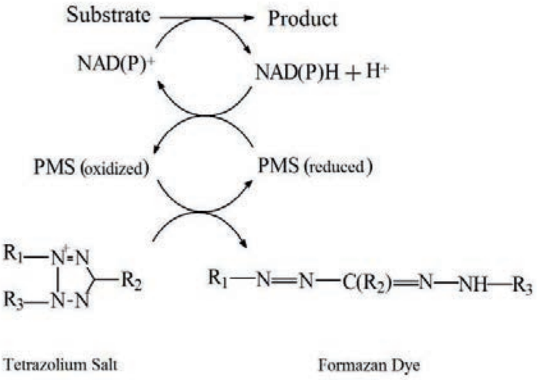


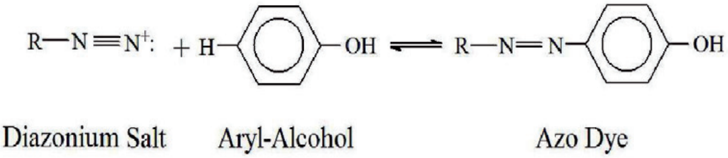
Fig. 9: Isoelectric point of proteins

Histochemical staining: Isoenzymes as proteins are separable by electrophoresis and are detectable as enzymes by enzyme specific staining. The basis of the staining protocols is the catalytic activity of isoenzymes: they make specific product(s) from specific substrate(s) (Vallejos, 1983).

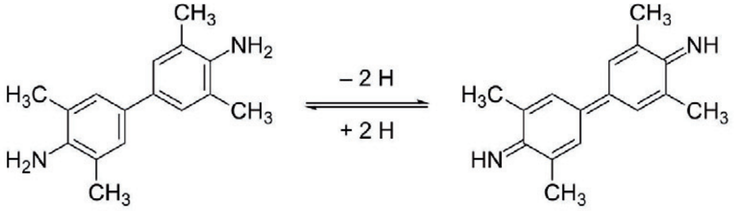
- **The tetrazolium system:** This is a very widespread staining protocol; which can stain enzymes as for example: 6-phosphogluconate dehydrogenase or phosphoglucomutase by this system. This system is based on this reaction chain:



- **The diazonium system:** Enzymes as for example acid phophatase, esterases can be stained by the diazonium system. During the reaction catalysed by the enzyme, aryl-alcohol is produced. The aryl-alcohol react with diazonium salt, producing azo dye:



- **The redox system:** During the reduction of the substrate, the 3,3',5,5'-tetramethylbenzidine (TMBZ) oxidise to 3,3',5,5'-tetramethylbenzidine diimine, which is an insoluble dye.



Application of isoenzyme analyses in genetics and plant breeding

Isoenzyme analyses in grapevine (*Vitis vinifera*) genetics: A special isoenzyme zymogram is characteristic for the *Vitis vinifera* proles pontica cultivars. The genetic diversity of Hungarian grapevine cultivars with ioenzyme markers were investigated by Jahnke et al. (2009). The isoenzyme patterns of 4 enzyme systems (catechol-oxidase, glutamate- oxalacetate-transaminase, acid phosphatase and peroxidase) of 48 grapevine (*Vitis vinifera*) varieties were analysed. The results with CO, GOT, AcP and PER enzymes were reproducible and the zymograms obtained from the woody stems were independent from the time of sampling during the dormant period of the grape (Figure 10). Based on the isoenzyme patterns of these 4 enzymes most of the investigated varieties (40/48) were identified.

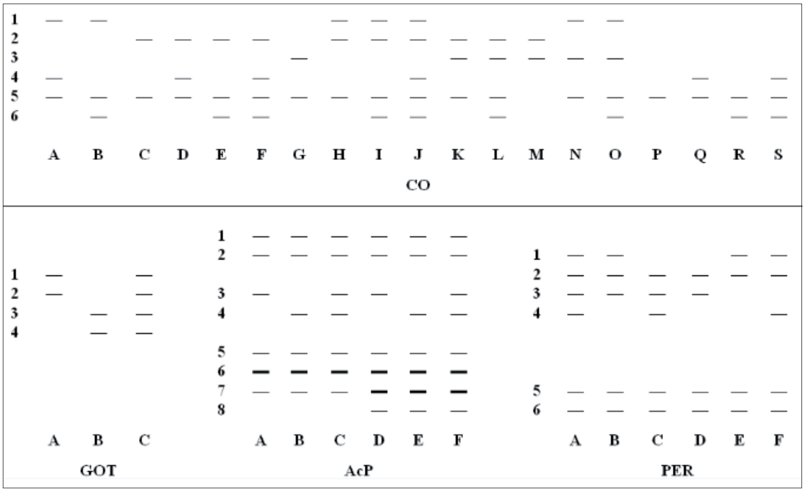


Fig. 10: Characteristic interpretative zymograms observed for CO, GOT, AcP and PER enzymes. The letters mark the different types of isoenzyme patterns, while numbers refer to the number of different isoenzyme bands (Jahnke et al., 2009.)

The ampelographical (morphological) characters used so far to describe *Vitis vinifera* cultivars significantly vary with different environmental conditions. Negrul (1968) divided the *Vitis vinifera* cultivars to so-called proleses (ecogeographical groups) based on stabile morphological features, and geographical spread. A total of three proleses were recognised: proles orientalis, proles occidentalis and proles pontica. A correlation was found between the isoenzyme patterns and the classification to proles of the varieties. It was established, that while the varieties of the proles pontica differed from those of the proles orientalis and occidentalis, the two latter groups could have not been differentiated from each other. This results support the presence of ecogeographical groups (Figure 11).

Acid phosphatase-I isoenzyme as molecular marker for MAS in nematode resistance breeding of tomato (*Lycopersicon esculentum*): *Meloidogyne incognita* is a nematode (roundworm) in the family Heteroderidae. It is commonly called the "southern root-knot nematode" or the "cotton root-knot nematode". This parasitic roundworm has worldwide distribution and numerous hosts. It is an important plant parasite classified in parasitology as a root-knot nematode, as it prefers to attack the root of its host plant. When *M. incognita* attacks the roots of plants, it sets up a feeding location, where it deforms the normal root cells and establishes giant cells. The roots become gnarled or nodulated, forming galls, hence the term "root-knot" nematode. *M. incognita* is a serious pest of tomato (*Lycopersicon esculentum*) as well (Figure 12).

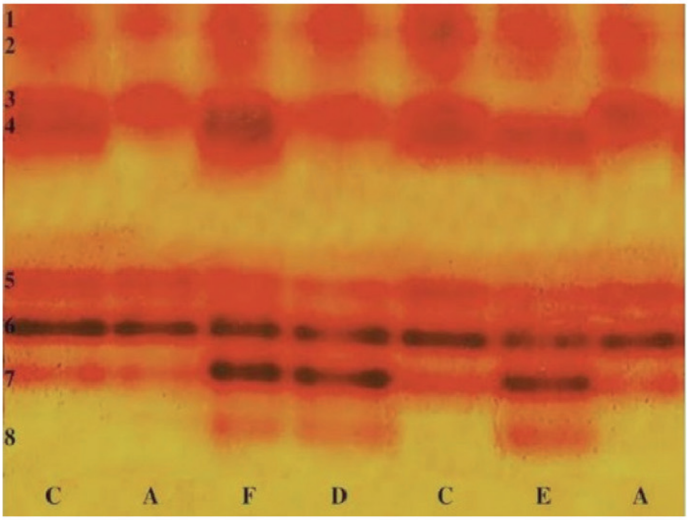


Fig. 11: Isoenzyme gel photo for AcP.

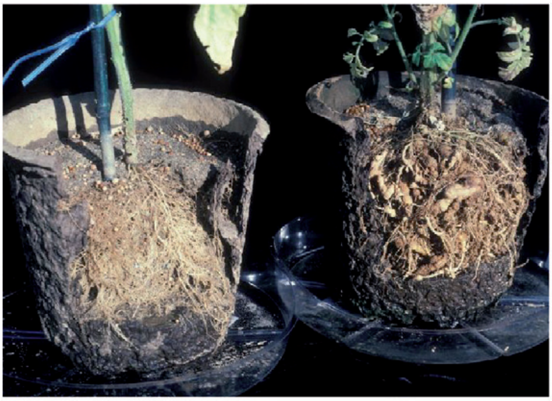


Fig. 12: Healthy tomato (left) and root knot (right) (Source: http://www.forestryimages.org/browse/detail.cfm?imgnum=1570801)

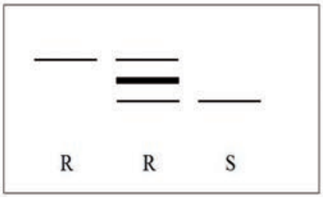


Fig. 13: Aps-1 zymograms of resistant (R) and susceptible (S) tomato plants

Nine resistant processing tomato (*Lycopersicon esculentum*) cultivars and advanced lines were compared with four susceptible cultivars in 1,3-dichloropropene-fumigated and nontreated plots on Meloidogyne incognita-infested sites over 3 years. Yield of all resistant genotypes grown in non treated and nematicide-treated plots did not differ and was greater than yield of susceptible genotypes. *M. incognita* initial soil population densities caused 39.3-56.5% yield suppressions of susceptible genotypes. Nematode injury to susceptible plants usually caused both fruit soluble solids content and pH to increase significantly. Only trace nematode reproduction occurred on resistant genotypes in nontreated plots, whereas large population density increases occurred on susceptible genotypes.

The use of MAS (Marker Assisted Selection) in tomato breeding begin in 1974, Rick and Fobes (1974) found the “Mi” –a dominant resistance gene against the cotton root-knot nematode (*Meloidogyne incognita*) close linked to the isoenzyme gene Aps-1. This enzyme in tomato is a dimer of 2 subunits, therefore the zymograms of heterozygotes show three bands – one each in the parental positions, and a band in the intermediate position (Figure 13). The segregation showed that either very close linkage between Aps-1 and Mi or pleiotropy is involved. This isozyme marker still is being used in tomato breeding for selecting for nematode resistance.

References

Hajósné Novák, M. (Ed. 1999) Genetikai variabilitás a növénynemesítlésben. (Genetic variability in plant breeding-in Hungarian) Mezőgazda Kiadó Budapest, 1999. P.143.

Jahnke, G., Májer, J., Lakatos, A., Györffyné Molnár, J., Deák, E., Stefanovits-Bányai, É. and Varga, P. (2009) Isoenzyme and microsatellite analysis of *Vitis vinifera* L. varieties in the Hungarian grape germplasm. Scientia Horticulturae 120: 213–221.

Lagrimini, L.M. and Rothstein, S. (1987) Tissue Specificity of Tobacco Peroxidase Isozymes and their Induction by wounding and tobacco mosaic virus infection. Plant Physiology 84: 438-442.

Mowrey, B.D. and Werner, D.J. (1990) Developmental Specific Isozyme Expression in Peach. HortScience 25(2): 219-222.

Negrul, A.M. (1968) Questions of origin and breeding of the grape vine on a genetical basis. Genetika (Genetics) 4 (3): 84-97.

Rick, C.M. and Fobes, J. (1974) Association of an allozyme with nematode resistance. Tomato Genetics Cooperative Report 24: 25.

Shields, C.R., Orton, T.J. and Stuber, W. (1983) An outline of general resource needs and procedures for the electrophoretic separation of active enzymes from plant tissue. In: Tanksley, S. D., Orton, T.J. (Eds.) Isozymes in Plant Genetics and Breeding. Part A. Amsterdam: Elsevier Science Publishers B.V. pp. 443-467.

Vallejos, C.E. (1983) Enzyme activity staining. In: Tanksley, S.D., T.J. Orton (Eds.) Isozymes in Plant Genetics and Breeding. Part A. Amsterdam: Elsevier Science Publishers B.V. 469-516. p

*****

Plant DNA Barcoding and Phylogenetics, Pages 121-142
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

8 Applications of Plant DNA Barcoding

M.A. Ali, G. Gyulai, A.K. Pandey, A. Arshi and S.K. Pandey

Introduction

The plant DNA barcoding offers taxonomists the opportunity to greatly expand, and eventually complete, a global inventory of life's diversity. DNA barcoding is of great utility to users of taxonomy. It provides more rapid progress then the traditional taxonomic work (Gregory, 2005). The plant DNA barcoding allows taxonomists to rapidly sort specimens by highlighting divergent taxa that may represent new species. The advocates of DNA barcoding says that it have revitalize biological collections and speed up species identification and inventories (Gregory, 2005; Schindel and Miller, 2005); however the opponents argue that it will destroy traditional systematics and turn it into a service industry (Ebach and Holdrege, 2005; Seberg et al., 2003). Once fully developed, DNA barcoding will have the potential to completely revolutionize our knowledge of diversity of living organisms and our relationship to nature. By harnessing technological advances in electronics and genetics, DNA barcoding will help many people quickly and cheaply recognize known species and retrieve information about them, and will speed discovery of thousands of species yet to be named.

Initiating plant DNA barcoding studies

The DNA sequence data and plant DNA barcoding are well on the way to being accepted as the global standard for species identification; however, such development is still limited in use. With the rich biological resources in many developing countries and many excellent taxonomists who are intimately familiar with the regional flora and interesting systematic questions, more plant DNA barcoding and molecular systematic studies by colleagues from developing countries should advance our understanding of the tree of life at the global scale and offer opportunities to address many new evolutionary questions as well (Ali and Choudhary, 2011). Plant DNA barcoding and molecular systematic research requires more equipment for data collection and analysis. It is technically more expensive than the classical, morphological and anatomical studies, but perhaps affordable. There is need to harness the mountains of DNA data being generated in modern laboratories and also to use the data from deep morphology in systematic (Wen and Pandey, 2005). DNA barcodes are likely to play a major role in the future of taxonomy. The build-up of DNA databases has great potential for the identification and classification of organisms and for supporting ecological and biodiversity research programs (Tautz et al., 2003). As a uniform, practical method for species identification, it appears to have broad scientific applications. DNA-based identification of species offers enormous potential for the biological scientific community, educators, and the interested general public. It will help open the treasury of biological knowledge and increase community interest in conservation biology and understanding of evolution.

Barcoding of life

The reported success of using the barcoding region in distinguishing species from a range of taxa and to reveal cryptic species is remarkable. However, it is known that species identification based on a single DNA sequence will always produce some erroneous results. Efforts should therefore be made to develop nuclear barcodes to complement the barcoding region currently in use. As the advantages and limitations of barcoding become apparent, it is clear that taxonomic approaches integrating DNA sequencing, morphology and ecological studies will achieve maximum efficiency at species identification (Dasmahapatra and Mallet, 2006). The urgency of creating tissue banks has been well recognized (Savolainen and Reeves, 2004; Lorenz et al., 2005), and solutions for linking DNA samples with taxonomic vouchers are being developed for all sorts of organisms. Barcoding of life will have to be both integrative and integrated with other taxonomic initiatives such as the Global Taxonomic Initiative of the Convention on Biological Diversity (www.biodiv.org), and the Global Biodiversity Information Facility (www.gbif.org). Finally, by barcoding of life, 'Life Barcoders' will identify species linked via the World Wide Web to other kinds of biodiversity data such as images, usage and conservation status. The direct benefits of DNA barcoding is to make the outputs of systematics available to a large number of end-users by providing standardized and high-tech identification tools, e.g. for biomedicine (parasites and vectors), agriculture (pests), environmental assays and customs (trade in endangered species).

Future perspectives of DNA barcoding

The future perspective of DNA barcoding will be to provide a bio-literacy tool for the general public and it will also help in opening to the treasury of biological knowledge, which is currently underused partly because of the weak taxonomic expertise for species identification. DNA barcoding will also relieve the enormous burden of identifications of taxonomists, so they can focus on more pertinent to discovering and describing the new species. The most important aspect of DNA barcoding is that it will facilitate basic biodiversity inventories (Savolainen et al., 2005; Lahaye et al., 2008). DNA barcoding can be likened to aerial photography, in that it provides an efficient method for mapping the extant species, though in sample space rather than physical space. The "aerial map" of DNA barcodes will help investigators explore the biological world and make full use of the enormous knowledge that has been built on 250 years of classical taxonomy. As sequencing costs decrease, DNA-based species identification will become available to an increasingly wide scientific community. When costs are low enough, researchers, teachers and naturalists will be able to use DNA barcoding in depth for examination of local ecosystems. As DNA barcodes are applicable to all life stages, it is also useful in cases where e.g. larval stages are difficult to identify with traditional methods of butterflies (Janzen et al., 2005) or amphibians (Vences et al., 2005), and insects in which several casts have different 'unrelated' morphologies (Smith et al., 2005). However, DNA barcoding is applied only in conjunction with classical approaches based on morphology.

The main reasons of DNA barcoding is works with fragments, works for all stages of life, unmasks look-alikes, reduces ambiguity, makes expertise go further, democratizes access, opens the way for an electronic handheld field guide, the life barcoder, sprouts new leaves on the tree of life, demonstrates value of collections, speeds writing the encyclopedia of life. Barcoding links biological identification to advancing frontiers in DNA sequencing, miniaturization in electronics, and computerized information storage. Integrating those links will lead to portable desktop devices and ultimately to hand-held barcoders. A handheld barcoder, such as the one envisioned here, would have many uses. Promoting technology development of portable devices for field use will be a major goal of this initiative (http://barcoding.si.edu/PDF/TenReasonsBarcoding.pdf).

Authentication of raw herbal material

A common problem with raw drug trade has been the admixtures with morphologically allied and geographically co-occurring species (Nair et al., 1983; Bisset, 1984; Sunita, 1992; Khatoon et al., 2006; Mitra and Kannan, 2007). Over 80% of the medicinal plants for raw drug trade are predominantly collected from the wild by local farmers or collectors, who often rely only on their experience in identifying the species being collected. Services of specialists like taxonomists are rarely availed for authentication. Thus, it is not uncommon to find admixtures of related/allied species and infrequently also for unrelated genera. Among the reasons attributed for species admixtures are the apparent confusion in vernacular names between indigenous systems of medicine and local dialects, non availability of authentic plant, similarity in morphological features, etc. The possibility of admixtures is particularly high when the species in question co-occurs with morphologically similar species. Frequently, admixtures could also be deliberate due to adulteration (Mitra and Kannan, 2007). The consequences of species admixtures can range from reducing the efficacy of the drug to lowering the trade value (Wieniawski, 2001; Song et al., 2009). Efforts have been made to accurately identify medicinal plants (Jayasinghe et al., 2009). Besides conventional methods including examination of wood anatomy and morpho-taxonomical keys, several-DNA-based methods have been developed to resolve these problems (Sucher and Carles, 2008). With the advent of DNA barcode tools, attempts are being made to use several candidate barcode regions to identify species (Ali et al., 2014).

Since, the adulteration of herbal material in its trading is a common problem; therefore, authentication of raw herbal material is one of the most important requirements needed by the pharmaceutical companies for quality control of the drug obtained from the medicinal plants. There are variety of methods which are based on morphological, biochemical or histological characteristics employed for the accurate identification of medicinal plants in order to ensure the purity, quality and safety of the drugs, but the results obtained from these method are not always reproducible because these characteristic changes under different environmental conditions; however, in contrast to aforesaid methods, the DNA-based methods for authentication of medicinal plants are considered to be more reliable for fresh as well as dried samples particularly for those medicinal plants in which variation within species and the divergence among species is difficult to understand. For example, identification of species of *Panax* is one of the challenges due to occurrence of high level of morphological variation within the genus as well as even within the population; hence, the adulteration is common in raw material trading of ginseng which ultimately reduces the efficacy of the drug obtained from it. There is also a reliable and practical method for species identification of *Panax* is lacking. Various techniques are in use or have been tried for the purpose of identification of *Panax* species such as metabolic chemicals profiling resolved by high performance liquid chromatography (Chan et al., 2000), molecular markers such as random amplified polymorphic DNA (RAPD) and microsatellite markers (Ngan et al., 1999; Hon et al., 2003), peptide nucleic acid microarray (Lee et al., 2010), pyrosequencing (Leem, et al., 2005); however, the techniques used so far have suffered from low efficiency, reproducibility and reliability. Species identification based on DNA sequences is a method of high efficiency, reproducibility and reliability. The screening for candidate DNA barcoding loci in *Panax* demonstrated that the combination of psbA-trnH and ITS is suitable for its identification (Zuo et al., 2011). Similarly, ITS, trnH-psbA, rbcL, matK and trnL–trnF gene sequences have successfully been used for DNA barcoding of several plant species (Table 1), though, the success rates of psbA-trnH remain much lower at the species level (Chen et al., 2010).

Moreover, a total of 17 barcode regions (matK, rbcL, ITS, ITS2, psbA-trnH, atpF-atpH, ycf5, psbK-I, psbM trnD, rps16, coxI, nad1, trnL-F, rpoB, rpoC1, atpF-atpH, rps16) of medicinal plants were reported to aid in the authentication and identification of medicinal plant materials. Besides using known genomic regions, other PCR-based methods have been applied to develop markers that help with the authentication and identification of medicinal plant material: RAPD, RFLP, microsatellites, ISSRs, SNPs, and ARMS. SCAR markers have been developed from RAPD, ISSR and a variety of genomic regions (Techen et al., 2014). In addition with the above, Chen et al. (2010) tested the discrimination ability of ITS2 in more than 6600 plant samples belonging to 4800 species from 753 distinct genera (see the link for reference: http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0008613.s008) and found that the rate of successful identification with the ITS2 was 92.7% at the species level. Yao et al. (2010) also evaluated 50,790 plants and 12,221 animal-

Table1: DNA barcoding studies for the identification of herbal medicinal materials.

Taxon	Family	DNA region	Reference
Achyranthes bidentata	Amaranthaceae	ITS	Wang et al., (2004)
Aconitum species	Ranunculaceae	ITS	Luo and Yang (2008); Zhang et al. (2010b)
Adenophora lobophylla	Campanulaceae	ITS	Ge et al. (1997)
Alpinia species	Zingiberaceae	ITS	Zhao et al. (2000, 2001)
Amomum species	Zingiberaceae	ITS	Pan et al. (2001); Zhou et al. (2002)
Angelica sinensis	Apiaceae	ITS	Ji et al. (2002); Zhang et al. (2003); Zhao et al. (2006)
Angelica species	Apiaceae	5S	Mizukami (1995, 1997)
Aquilaria sinensis	Thymelaeaceae	ITS	Shen et al. (2008); Niu et al. (2010)
Arctium lappa	Asteraceae	ITS	Liu et al. (2010)
Arisaema species	Araceae	rbcL	Kondo et al. (1998)
Aristolochia species	Aristolochiaceae	trnH-psbA	Li et al. (2010)
Artemisia species	Asteraceae	trnH-psbA	Liu and Ji (2009)
Astragalus species	Fabaceae	ITS	Dong et al. (2003)
Atractylodes species	Asteraceae	ITS	Shiba et al. (2006)
Atractylodes species	Asteraceae	trnL–trnF	Ge et al. (2007)
Belamcanda chinensis	Iridaceae	rbcL	Qin et al. (2003)
Bupleurum species	Apiaceae	ITS	Xie et al. (2006); Yang et al. (2007); Xie et al. (2009)
Changium smyrnioides	Apiaceae	ITS	Tao et al. (2008)
Chuanminshen violaceum	Apiaceae	ITS	Tao et al. (2008)
Cinnamomum species	Lauraceae	trnL–trnF	Kojoma et al. (2002)
Citrus grandis	Rutaceae	trnH-psbA	Su et al. (2010)
Citrus medica	Rutaceae	ITS	Gao et al. (2007)
Cnidium monnieri	Apiaceae	ITS	Cai et al. (2000)

Cnidium monnieri	Apiaceae	matK	Cao et al. (2001)
Cnidium officinale	Apiaceae	rbcL	Kondo et al. (1996)
Cnidium officinale	Apiaceae	matK	Liu et al. (2002)
Codonopsis tangshen	Campanulaceae	ITS	Luo et al. (2010)
Crocus sativus	Iridaceae	ITS	Mao et al. (2007); Che et al. (2007)
Cynanchum species	Apocynaceae	ITS	Zhang et al. (2010a)
Dendrobium chrysanthum	Orchidaceae	ITS	Xu et al. (2001)
Dendrobium nobile	Orchidaceae	ITS	Ge et al. (2008)
Dendrobium officinale	Orchidaceae	ITS	Ding et al. (2002a)
Dendrobium species	Orchidaceae	ITS	Lau et al. (2001); Ding et al. (2002a,b,c); Xu et al. (2006)
Dendrobium species	Orchidaceae	trnH-psbA	Yao et al. (2009)
Dendrobium species	Orchidaceae	rbcL	Asahina et al. (2010)
Dendrobium species	Orchidaceae	matK	Teng et al. (2002); Asahina et al. (2010)
Dioscorea species	Dioscoreaceae	ITS	Wang et al. (2007)
Dryopteris crassirhizoma	Dryopteridaceae	rbcL	Zhao et al. (2007)
Ephedra species	Ephedraceae	ITS	Guo et al. (2006)
Epimedium species	Berberidaceae	5S	Sun et al. (2004)
Eucommia ulmoides	Eucommiaceae	ITS	Ma et al. (2004)
Euphorbia species	Euphorbiaceae	ITS	Jiang et al. (2005)
Fritillaria species	Liliaceae	5S	Cai et al. (1999)
Gentiana dahurica	Gentianaceae	ITS	Ji et al. (2003b)
Glycyrrhiza species	Fabaceae	rbcL	Hayashi et al. (1998, 2000, 2005)
Gynostemma pentaphyllum	Cucurbitaceae	ITS	Jiang et al. (2009)
Hedyotis diffusa	Rubiaceae	ITS	Hao et al. (2004); Liu and Hao (2005)
Hypericum perforatum	Hypericaceae	ITS	Howard et al. (2009)
Ligusticum chuanxiong	Apiaceae	ITS	Liu et al. (2002)

Ligusticum chuanxiong	Apiaceae	matK	Liu et al. (2002)
Liriope species	Asparagaceae	ITS	Huang et al. (2009)
Lonicera japonica	Caprifoliaceae	5S	Li et al. (2001)
Lycium barbarum	Solanaceae	ITS	Shi et al. (2008)
Mitragyna speciosa	Rubiaceae	ITS	Sukrong et al. (2007)
Morinda officinalis	Rubiaceae	ITS	Ding and Fang (2005)
Nelumbo nucifera	Nelumbonaceae	ITS	Lin et al. (2007)
Ophiopogon japonicus	Asparagaceae	ITS	Huang et al. (2009)
Panax ginseng	Araliaceae	ITS	Ma et al. (2000)
Panax notoginseng	Araliaceae	matK	Fushimi et al. (2000); Zhang et al. (2006)
Panax species	Araliaceae	ITS	Ngan et al. (1999)
Panax species	Araliaceae	matK	Zhu et al. (2003)
Panax vietnamensis	Araliaceae	matK	Komatsu et al. (2001)
Paris species	Melanthiaceae	trnH-psbA	Yang et al. (2010)
Polygonum multiflorum	Polygonaceae	ITS	Zhang and Shi (2007)
Polygonum multiflorum	Polygonaceae	matK	Yan et al. (2008)
Polygonum tinctorium	Polygonaceae	ITS	Song et al. (2009)
Pseudostellaria heterophylla	Caryophyllaceae	ITS	Yu et al. (2003); Zhu et al. (2007)
Pueraria species	Fabaceae	ITS	Zeng et al. (2003); Sun et al. (2007)
Pueraria species	Fabaceae	5S	Sun et al. (2007)
Rheum palmatum	Polygonaceae	ITS	Zhang et al. (2003); Ji et al. (2003a)
Rheum species	Apiaceae	matK	Yang et al. (2004)
Rhodiola alsia	Crassulaceae	ITS	Gao et al. (2009)
Sabia parviflora	Sabiaceae	trnH-psbA	Sui et al. (2010)
Sabia parviflora	Sabiaceae	rbcL	Sui et al. (2010)
Sabia parviflora	Sabiaceae	matK	Sui et al. (2010)
Salvia miltiorrhiza	Lamiaceae	ITS	Wang and Wang (2005)

Saussurea lappa	Asteraceae	ITS	Chen et al. (2008)
Saussurea lappa	Asteraceae	5S	Chen et al. (2008)
Saussurea medusa	Asteraceae	ITS	Liu et al. (2001b)
Schisandra chinensis	Schisandraceae	ITS	Gao et al. (2003)
Species in Polygonaceae	Polygonaceae	trnH-psbA	Song et al. (2009)
Species in Polygonaceae	Polygonaceae	rbcL	Song et al. (2009)
Stellaria media	Caryophyllaceae	ITS	Zhao et al. (2009)
Stellaria media	Caryophyllaceae	trnL–trnF	Zhao et al. (2009)
Stemona tuberosa	Stemonaceae	trnH-psbA	Vongsak et al. (2008)
Stemona tuberose	Stemonaceae	ITS	Jiang et al. (2006)
Swertia mussotii	Gentianaceae	ITS	Liu et al. (2001a)
Swertia mussotii	Gentianaceae	5S	Yu et al. (2008)
Tripterygium wilfordii	Celastraceae	ITS	Law et al. (2010)
Tripterygium wilfordii	Celastraceae	5S	Law et al. (2010)
Verbena officinalis	Verbenaceae	ITS	Ruzicka et al. (2009)

-ITS2 sequences downloaded from GenBank, and propose that the ITS2 locus should be used as a universal DNA barcode for identifying plant species and as a complementary locus for CO1 to identify animal species.

Plant DNA barcoding and conservation of biodiversity

Molecular markers are increasingly used for screening of germplasm to study genetic diversity, identify redundancies in the collections (Rao, 2004). Sustainable utilization of plant genetic resources is essential to meet the demand for future food and health security. Despite the tradition of systematic biology as the science of diversity, systematics has until recently contributed relatively little to the theory and practice of conservation biology. The four areas in which systematics could contribute to the conservation of rare plant species are: (i) species concepts, (ii) the identification of lineages worthy of conservation, (iii) the setting of conservation priorities, and (iv) the effects of hybridization on the biology and conservation of rare species. Species concepts that incorporate history and reflect phylogeny ultimately is more useful for preserving biodiversity. Phylogenetic analyses involving conspecific populations often reveal multiple lineages that may warrant protection as evolutionarily distinct units. Phylogenetic information provides the tools for inferring relationships among organisms and, in conjunction with biogeography, for identifying those areas that harbour many actively speciation groups. Hybridization may lead to the extinction of a rare species, but in other cases, ironically, artificial hybridization with a more widespread congener may be the only way to preserve the gene pool of a rare species (Soltis and Gitzendanner, 1999).

References

Ali, M.A. and Choudhary, R.K. (2011) India needs more plant taxonomists. Nature 471: (7336) 37-37.

Ali, M.A., Gyulai, G., Hidvégi, N., Kerti, B., Al Hemaid, F.M.A., Pandey, A.K. and Lee, J. (2014) The changing epitome of species identification - DNA barcoding. Saudi Journal of Biological Sciences 21: 204–231.

Asahina, H., Shinozaki, J., Masuda, K., Morimitsu, Y. and Satake, M. (2010) Identification of medicinal *Dendrobium* species by phylogenetic analyses using *matK* and *rbcL* sequences. Journal of Natural Medicines 64: 133-138.

Bisset, W.G. (1984) Herbal Drugs and Phytopharmaceuticals. CRC Press, London.

Cai, J.N., Zhou, K.Y., Xu, L.S., Wang, Z.T., Shen, X., Wang, Y.Q. and Li, X.B. (2000) Ribosomal DNA ITS sequence analyses of *Cnidium monnieri* from different geographical origin in China. Acta Pharmaceutica Sinica 35: 56-59.

Cai, Z.H., Li, P., Dong, T.T. and Tsim, K.W. (1999) Molecular diversity of *5S-rRNA* spacer domain in *Fritillaria* species revealed by PCR analysis. Planta Medica 65: 360-364.

Cao, H., Cai, J.N., Liu, Y.P., Wang, Z.T. and Xu, L.S. (2001) Correlative analysis between geographical distribution and nucleotide sequence of chloroplast *matK* gene of *Cnidium monnieri* fruit in China. Chinese Pharmaceutical Journal 36: 373-376.

Chan, T.W.D., But, P.P.H., Cheng, S.W., Kwok, I.M.Y., Lau, F.W. and Xu, H.X. (2000) Differentiation and authentication of *Panax ginseng*, *Panax quinquefolius* and ginseng products by the use of HPLC/MS. Analytical chemistry 72:1281–1287.

Che, J., Tang, L., Liu, Y.J., He, W. and Chen, F. (2007) Molecular identity of *Crocus sativus* and its misused substitutes by ITS sequence. China Journal of Chinese Materia Medica 32: 668-671.

Chen, F., Chan, H.Y., Wong, K.L., Wang, J., Yu, M.T., But, P.P.H. and Shaw, P.C. (2008) Authentication of *Saussurea lappa*, an endangered medicinal material, by ITS DNA and *5S rRNA* sequencing. Planta Medica 74: 889-892.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. and Leon, C. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5(1): e8613.

Dasmahapatra, K.K. and Mallet, J. (2006) DNA barcodes: recent successes and future prospects. Heredity 97(4): 254-255.

Ding, P. and Fang. Q. (2005) Ribosomal DNA-ITS sequence analysis and molecular identification of *Morinda officinalis* and its counterfeit species. Chinese Traditional and Herbal Drugs 36: 908-911.

Ding, X., Xu, L., Wang, Z., Zhou, K., Xum H. and Wang, Y. (2002a) Authentication of stems of *Dendrobium officinale* by rDNA ITS region sequences. Planta Medica 68: 191-192.

Ding, X.Y., Wang, Z.T., Xu, H., Xu, L.S. and Zhou, K.Y. (2002b) Database establishment of the whole rDNA ITS region of *Dendrobium* species of “Fengdou” and authentication by analysis of their sequences. Acta Pharmaceutica Sinica 37: 567-573.

Ding, X.Y., Wang, Z.T., Xu, L.S., Xu, H., Zhou, K.Y. and Shi, G.X. (2002c) Study on sequence difference and SNP phenomenon of rDNA ITS region in F type and H type population of *Dendrobium officinale*. China Journal of Chinese Materia Medica 27: 85-89.

Dong, T.T., Ma, X.Q., Clarke, C., Song, Z.H., Ji, Z.N., Lo, C.K. and Tsim, K.W. (2003) Phylogeny of *Astragalus* in China: molecular evidence from the DNA sequences of *5S rRNA* spacer, ITS, and 18S rRNA. Journal of Agricultural and Food Chemistry 51: 6709-6714.

Ebach, M.C. and Holdrege, C. (2005) DNA barcoding is no substitute for taxonomy. Nature 434: 697.

Fushimi, H., Komatsu, K., Namba, T. and Isobe, M. (2000) Genetic heterogeneity of ribosomal RNA gene and *matK* gene in *Panax notoginseng*. Planta Medica 66: 659-661.

Gao, J.P., Wang, Y.H., Qiao, C.F. and Chen, D.F. (2003) Ribosomal DNA ITS sequences analysis of the Chinese crude drug fructus schisandrae sphenantherae and fruits of *Schisandra viridis*. China Journal of Chinese Materia Medica 28: 706–710.

Gao, Q.B., Zhang, D.J., Duan, Y.Z., Zhang, F.Q. and Chen, S.L. (2009) Preliminary studies on ITS sequences of nrDNA from *Rhodiola alsia*. Journal of Anhui Agricultural Sciences 37: 49-50.

Gao, X.X., Xhen, X.Y. and Luo, Y.S. (2007) Preliminary study on rDNA ITS sequencing and characteristics of *Citrus medica* L. var. *sarcodactylis* (Noot.) Swingle. Lishizen Medicine and Material Medica Research 5: 162-163.

Ge, S., Schaal, B.A. and Hong, D.Y. (1997) A reevaluation of the status of *A. lobophylla* based on ITS sequence, with reference to the utility of ITS sequence in *Adenophora*. Acta Pharmaceutica Sinica 35: 385-395.

Ge, X.J., Xiao, K., Li, X.Q. and Tang, Y.P. (2008) Analysis of Chishui *Dendrobium nobile* based on rDNA ITS variation. Journal of Jiangsu University (Medicinal Edition) 18: 510-512.

Ge, Y.F., Hang, Y.Y., Xia, B. and Wei, Y.L. (2007) Sequencing of *trnL*-F and analysis of interspecific genetic relationship of five medicinal species in *Atractylodes* DC. Journal of Plant Resources and Environment 16: 12-16.

Gregory, T.R. (2005) DNA barcoding does not compete with taxonomy. Nature 434: 1067.

Guo, Y., Tsuruga. A., Yamaguchi, S., Oba, K., Iwai, K., Sekita, S. and Mizukami, H. (2006) Sequence analysis of chloroplast *chlB* gene of medicinal *Ephedra* species and its application to authentication of *Ephedra* Herb. Biological and Pharmaceutical Bulletin 29: 1207-1211.

Hao, M.G., Liu, Z.Q. and Wang, J.L. (2004) Application of the sequences of rDNA ITS to identify Chinese drug *Hedyotis diffusa*. Journal of Anhui Normal University (Natural Science) 27: 188-191.

Hayashi, H., Hosono, N., Kondo, M., Hiraoka, N. and Ikeshiro, Y. (1998) Phylogenetic relationship of *Glycyrrhiza* plants based on *rbcL* sequences. Biological and Pharmaceutical Bulletin 21:782-783.

Hayashi, H., Hosono, N., Kondo, M., Hiraoka, N., Ikeshiro, Y., Shibanom M., Kusano, G., Yamamoto, H., Tanaka, T. and Inoue, K. (2000) Phylogenetic relationship of six *Glycyrrhiza* species based on *rbcL* sequences and chemical constituents. Biological and Pharmaceutical Bulletin 23: 602-606.

Hayashi, H., Miwa, E. and Inoue, K. (2005) Phylogenetic relationship of *Glycyrrhiza lepidota*, American licorice, in genus *Glycyrrhiza* based on *rbcL* sequences and chemical constituents. Biological and Pharmaceutical Bulletin 28(1):161-164.

Hon, C.C., Chow, Y.C., Zeng, F.Y. and Leung, F.C.C. (2003) Genetic authentication of ginseng and other traditional Chinese medicine. Acta Pharmacologica Sinica 24: 841-846.

Howard, C., Bremner, P.D., Fowler, M.R., Isodo, B, Scott, N.W. and Slater, A. (2009) Molecular identification of *Hypericum perforatum* by PCR amplification of the ITS and *5.8S rDNA* region. Planta Medica 75: 864-869.

Huang, Y.J, Chen, J.Y., Su, H.J., Huang, Y.Z. and Wan, X.F. (2009) rDNA ITS of *Radix ophiopogonis* and *Radix liriopes* from different regions. Fujian Journal of Agricultural Sciences 24: 508-512.

Janzen, D.H., Hajibabaei, M., Burns, J.M., Hallwachs, W., Remigio, E. and Hebert, P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philosophical Transactions of the Royal Society B: Biological 360: 1835-1845.

Jayasinghe, R.L. Niu, H., Coram, T.E., Kong, S., Kaganovitch, J., Xue, C.C.L., Li, C.G. and Pang, E.C.K. (2009) Effectiveness of an innovative prototype subtracted diversity array (SDA) for fingerprinting plant species of medicinal importance. Planta Medica 75: 1180-1185.

Ji, K.P., Li, X.H., Li, Y.D. and Zhang, X.L. (2003a) Identification of *Rheum palmatum* L (Dahuang) by method of measuring internal transcribed spacer regions of rRNA gene. World Science and Technology-Modernization of Traditional Chinese Medicine 4: 44-47.

Ji, K.P., Li, Y.D., Zhang, X.L. and Li, X.H. (2002) Identification by measuring internal transcribed spacer regions of rRNA gene in *Radix Angelicae sinensis*. Chinese Traditional and Herbal Drugs 34: 66-69.

Ji, K.P., Zhang, X.L., Liu, L.S., Lu, Q.Y. and Cheng, C. (2003b) Primary study on measuring the internal transcribed spacer 1 regions of rRNA Genein seeds of *Gentiana dahurica*. China Journal of Chinese Materia Medica 28: 313-316.

Jiang, J.H., Meng, N., Cao, X.Y., Zhou, S.B. and Dai, C.C. (2005) ITS sequence analysis on medicinal plants of *Euphorbia* L. in Anhui and Jiangsu Provinces. Chinese Traditional and Herbal Drugs 36: 900-902.

Jiang, L.Y., Guo, Z.G., Wang, C. and Zhao, G.F. (2009) ITS sequence analysis of *Gynostemma pentaphyllum* from different habitats in China. Chinese Traditional and Herbal Drugs 40: 1123-1127.

Jiang, R.W., Hon, P.M., Xu, Y.T., Chan, Y.M., Xu, H.X., Shaw, P.C. and But, P.P.H. (2006) Isolation and chemotaxonomic significance of tuberostemospironine-type alkaloids from *Stemona tuberosa*. Phytochemistry 67: 52-57.

Khatoon, S., Rai, V., Rawat, A.K.S. and Mehrotra, S. (2006) Comparative pharmacognostic studies of three *Phyllanthus* species. Journal of Ethnopharmacology 104: 79-86.

Kojoma, M., Kurihara, K., Yamada, K., Sekita, S., Satake, M. and Iida, O. (2002) Genetic identification of cinnamon (*Cinnamomum* spp.) based on the *trnL–trnF* chloroplast DNA. Planta Medica 68: 94-96.

Komatsu, K., Zhu, S., Fushimi, H., Qui, T.K., Cai, S. and Kadota, S. (2001) Phylogenetic analysis based on 18S rRNA gene and *matK* gene sequences of *Panax vietnamensis* and five related species. Planta Medica 67: 461-465.

Kondo, K., Terabayashi, S. and Higuchi, M.I. (1998) Discrimination between *Banxia* and *Tiannanxing* based on *rbcL* sequences. Natural Medicine 52: 253-258.

Kondo, K., Terabayashi, S. and Okada, M. (1996) Phylogenetic relationship of medicinally important *Cnidium officinale* and Japanese Apiaceae based on *rbcL* sequences. Journal of Plant Research 109: 21-27.

Lahaye, R., van der Bank. M., Bogarin. D., Warner. J., Pupulin. F., Gigot. G., Maurin. O., Duthoit. S., Barraclough. T.G. and Savolainen, V. (2008) DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences USA 105: 2923-2928.

Lau, D.T., Shaw, P.C., Wang, J. and But, P.P.H. (2001) Authentication of medicinal *Dendrobium* species by the internal transcribed spacer of ribosomal DNA. Planta Medica 67: 456-460.

Law, S.K., Simmons, M.P., Techen, N., Khan, I.A., He, M.F., Shaw, P.C. and But, P.P.H. (2010) Molecular analyses of the Chinese herb Leigongteng (*Tripterygium wilfordii* Hook.f.). Phytochemistry 72: 21-26.

Lee, J.W., Bang, K.H., Choi, J.J., Chung, J.W., Lee, J.H., Jo, I.H., Seo, A.Y., Kim, Y.C., Kim, O.T., Cha, S.W. (2010). Development of peptide nucleic acid (PNA) microarray for identification of *Panax* species based on the nuclear ribosomal internal transcribed spacer (ITS) and 5.8S rDNA regions. Genes & Genomics 32(5): 463-468.

Leem, K., Kim, S.C., Yang, C.H. and Seo, J. (2005) Genetic identification of *Panax ginseng* and *Panax quinquefolius* by pyrosequencing methods. Bioscience, Biotechnology, and Biochemistry 69(9): 1771-1773.

Li, M., Ling, K.H., Lam, H., Shaw, P.C., Cheng, L., Techen, N., Khan, I.A., Chang, Y.S. and But, P.P.H. (2010) *Cardiocrinum* seeds as a replacement for *Aristolochia* fruits in treating cough. Journal of Ethnopharmacology 130: 429-432.

Li, P., Cai, Z.H. and Xing, J.B. (2001) Preliminary attempt to identify geoherbalism of Flos Lonicerae by sequence divergence of *5S-rRNA*gene spacer region. Chinese Traditional and Herbal Drugs 32: 834-837.

Lin, S., Zheng, W.W., Wu, J.Z., Zhou, L.J. and Song, Y.N. (2007) PCR, clone and sequence analysis of rDNA-ITS of *Nelumbo nucifera* from different geographical origins in China. China Journal of Chinese Materia Medica 32: 671-675.

Liu, J.Q., Chen, Z.D., Liao, Z.X. and Lu, A.M. (2001a) A comparison of the ITS sequences of the Tibetan medicine “Zangyinchen”: *Swertia mussotii* and its adulterant species. Acta Pharmaceutica Sinica 36: 67-70.

Liu, J.Q., Chen, Z.D. and Lu, A.M. (2001b) Comparison on internal transcribed spacers (ITS) sequences of Tibetan medicine *Saussurea medusa* and its easily confusable species. Chinese Traditional and Herbal Drugs 32: 443-445.

Liu, T. and Ji, Y.H. (2009) *Psb*A–*trnH* sequence analysis from chloroplast on medicinal plants of *Artemisia*. Chinese Agricultural Science Bulletin 25: 46-49.

Liu, Y.N., Xu, L. and Dou, D.Q. (2010) PCR amplification, cloning and sequence analysis of rDNA ITS of *Arctium lappa* from different geographical origins in China. Chinese Traditional and Herbal Drugs 33: 26-28.

Liu, Y.P., Cao, H., Han, G.R., Fushimi, H. and Komatsu, K. (2002) *Mat*K and ITS nucleotide sequencing of crude drug Chuanxiong and phylogenetic relationship between their species from China and Japan. Acta Pharmaceutica Sinica 37: 63-68.

Liu, Z.Q. and Hao, M.G. (2005) Identification of *Hedyotis diffusa* using rDNA ITS sequence. Shaanxi Journal of Traditional Chinese Medicine 26: 167-169.

Lorenz, J.G., Jackson, W.E., Beck, J.C. and Hanner, R. (2005) The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. Philosophical Transactions of the Royal Society B: Biological 360: 1869–1877.

Luo, H.B., Zhang, Z., Wu, Z.H. and Yun, D.P. (2010) Analysis of nrDNA ITS gene sequence of famous-region drug *Codonopsis tangshen* Oliv. Journal of Anhui Agricultural Sciences 38: 4427- 4428.

Luo, Y. and Yang, Q.E. (2008) Study on ITS sequences of herbal medicine of Chuanwu and Caowu. Chinese Pharmaceutical Journal 43: 820-823.

Ma, X.J., Wang, X.Q., Xiao, P.G. and Hong, D.Y. (2000) Comparison of ITS sequences between wild ginseng DNA and garden ginseng DNA. China Journal of Chinese Materia Medica 25: 206-209.

Ma, Y.H., Yang, J.A., Jia, W.Z. and Ye, G.S. (2004) Sequence analysis of ITS of nuclear ribosomal DNA (nrDNA) of *Eucommia ulmoides* from different geographical origin in China. Journal of Northwest Forestry University 19: 16-19.

Mao, S.G., Luo, Y.M., Shen, J. and Ding, X.Y. (2007) Authentication of *Crocus sativus* L. and its adulterants by rDNA ITS sequences and allele-specific PCR. Journal of Nanjing Normal University (Natural Science Edition) 30: 89-92.

Mitra, S.K. and Kannan, R. (2007) A Note on unintentional adulterations in ayurvedic herbs. Ethnobotanical Leaflets 11: 11-15.

Mizukami, H. (1995) Amplification and sequence of a *5S-rRNA* gene spacer region from the crude drug “*Angelica* root”. Biological and Pharmaceutical Bulletin 18: 1299-1301.

Mizukami, H., Hao, B.S. and Tanaka, T. (1997) Nucleotide sequence of *5S-rDNA* intergenic spacer region in *Angelica acutiloba*. Natural Medicine 51: 376-378.

Nair, V.K., Yoganarasimhan, K.R., Murthy, K. and Shantha, T.R. (1983) Studies on some South Indian market samples of ayurvedic drugs II. Ancient Science of Life 3: 60-66.

Ngan, F., Shaw, P., But, P. and Wang, J. (1999) Molecular authentication of *Panax* species. Phytochemistry 50: 787-791.

Niu, X.L., Ji, K.P. and Lu, G.Q. (2010) Preliminary studies on identification of *Aquilaria sinensis* (L.) Gilg by the PCR product of rDNA ITS sequencing. Guangdong Agricultural Sciences 37: 167-169.

Pan, H.X., Hunag, F., Wang, P.X., Zhou, L.J., Cao, L.Y. and Liang, R.Y. (2001) Identification of *Amomum villosum*, *Anomum vllosun* var. *xanthioides* and *Amomum longiligulare* on ITS-1 Sequence. Chinese Traditional and Herbal Drugs 24: 481-482.

Qin, N.J., Huang, Y., Yang, G., Xu, L.S. and Zhou, K.Y. (2003) *RbcL* sequence analysis of *Belamcanda chinensis* and related medicinal plants of *Iris*. Acta Pharmaceutica Sinica 38: 147-152.

Rao, N.K. (2004) Plant genetic resources: advancing conservation and use through biotechnology. African Journal of Biotechnology 3 (2): 136-145.

Ruzicka, J., Lukas, B., Merza, L., Gohler, I., Abel, G., Popp, M. and Novak, J. (2009) Identification of *Verbena officinalis* based on ITS sequence analysis and RAPD-derived molecular markers. Planta Medica 75: 1271-1276.

Savolainen, V., Cowan, R.S., Vogler, A.P., Roderick, G.K. and Lane, R. (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1805-1811.

Savolainen, V. and Reeves, G. (2004) A plea for DNA banking. Science 304: 1445.

Schindel, D.E. and Miller, S.E. (2005) DNA barcoding a useful tool for taxonomists. Nature 435: 17.

Seberg, O., Humphries, C.J., Knapp, S., Stevenson, D.W., Petersen, G., Scharff, N. and Andersen, N.M. (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends in Ecology & Evolution 18: 63–65.

Shen, Y.J., Yan, P., Zhao, X., Pang, Q.H. and Zhao, S.J. (2008) Application of ISSR marker and ITS sequence to investigation of genetic variation of *Aquilaria sinensis*. Journal of South China University of Technology (Natural Science Edition) 36: 128-132.

Shi, Z.G., An, W. and Zhao, Y.L. (2008) Genetic polymorphism of 18 *Lycium barbarum* resources based on nrDNA ITS sequences. Journal of Anhui Agricultural Sciences 49: 10379-0380.

Shiba, M., Kondo, K., Miki, E., Yamaji, H., Morota, T., Terabayashi, S., Takedam S., Sasaki, H., Miyamoto, K. and Aburada, M. (2006) Identification of medicinal *Atractylodes* based on ITS sequences of nrDNA. Biological and Pharmaceutical Bulletin 29: 315-320.

Smith, M.A., Fisher, B.L. and Hebert, P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1825–1834.

Soltis, P.S. and Gitzendanner, M.A. (1999) Molecular systematics and the conservation of rare species. Conservation Biology 13: 471-483.

Song, J., Yao, H., Li, Y., Li, X., Lin, Y., Liu, C., Han, J., Xie, C. and Chen, S. (2009) Authentication of the family Polygonaceae in Chinese pharmacopoeia by DNA barcoding technique. Journal of Ethnopharmacology 24: 434-439.

Su, C., Wong, K.L., But, P.P.H., Su, W.W. and Shaw, P.C. (2010) Molecular authentication of the Chinese herb Huajuhong and related medicinal material by DNA sequencing and ISSR marker. Journal of Food and Drug Analysis 18: 161-170.

Sucher, N.J. and Carles, M.C. (2008) Genome-based approaches to the authentication of medicinal plants. Planta Medica 74: 603-623.

Sui, X.Y., Hung, Y., Tan, Y., Guo, Y. and Long, C.L. (2010) Molecular authentication of the ethnomedicinal plant *Sabia parviflora* and its adulterants by DNA barcoding technique. Planta Medica 77(5): 492-496.

Sukrong, S., Zhu, S., Ruangrungsi, N., Phadungcharoen, T., Palanuvej, C. and Komatsu, K. (2007) Molecular analysis of the genus *Mitragyna* existing in Thailand based on rDNA ITS sequences and its application to identify a narcotic species: *Mitragyna speciosa*. Biological and Pharmaceutical Bulletin 30: 1284-1288.

Sun, Y., Fung, K.P., Leung, P.C., Shi, D. and Shaw, P.C. (2004) Characterization of medicinal *Epimedium* species by *5S rRNA* gene spacer sequencing. Planta Medica 70: 287-288.

Sun, Y., Shaw, P.C. and Fung, K.P. (2007) Molecular authentication of *Radix puerariae lobatae* and *Radix puerariae thomsonii* by ITS and *5S rRNA* spacer sequencing. Biological and Pharmaceutical Bulletin 30: 173-175.

Sunita, G. (1992) Substitute and adulterant plants. Periodical Experts Book Agency, New Delhi.

Tao, X.Y., Gui, X.Q., Fu, C.X. and Qiu, Y.X. (2008) Analysis of genetic differentiation and phylogenetic relationship between *Changium smyrnioides* and *Chuanminshen violaceum* using molecular markers and ITS sequences. Journal of Zhejiang University (Agriculture & Life Sciences) 34: 473-481.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2003) A plea for DNA taxonomy. Trends in Ecology & Evolution 18(2): 70-74.

Techen, N., Parveen, I., Panm Z. and Khan, I.A. (2014) DNA barcoding of medicinal plant material for identification. Current Opinion in Biotechnology 25: 103-110.

Teng, Y.F., Wu, X.J., Xu, H., Wang, Z.T., Yu, G.D. and Xu, L.S. (2002) A comparison of *matK* sequences between Herba Dendrobii (Shihu) and its adulterant species. Journal of China Pharmaceutical University 33: 280-283.

Vences, M., Thomas, M., Bonett, R.M. and Vieites, D.R. (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. Philosophical Transactions of the Royal Society B: Biological 360: 1859-1868.

Vongsak, B., Kengtong, S., Vajrodaya, S. and Sukrong, S. (2008) Sequencing analysis of the medicinal plant *Stemona tuberose* and five related species existing in Thailand based on *trnH– psbA* chloroplast DNA. Planta Medica 74: 1764-1766.

Wang, D., Bai, Y. and Chen, Z.H. (2007) Comparison of ribosomal DNA ITS sequence of *Rhizoma dioscoreae* in different geographical regions. Lishizen Medicine and Materia Medica Research 18: 54-55.

Wang, H. and Wang, Q. (2005) Analysis of rDNA ITS sequences of Radix et Rhizoma Salviae Miltiorrhizae and plants of *Salvia* L. Acta Pharmaceutica Sinica 36: 105-109.

Wang, S.M., Liang, S.W., Zhou, K.Y., Liu, Z.Q., Feng, W.S. and Wu, M.X. (2004) Ribosomal rDNA ITS sequence analysis of root of *Achyranthes bidentata*. Acta Pharmaceutica Sinica 35: 559- 562.

Wen, J. and Pandey, A.K. (2005) Initiating DNA molecular systematic studies in a developing country. In: Pandey, A.K, Wen, J., Dogra, J.V.V. (Eds.) Plant Taxonomy: Advances and Relevance. CBS Publishers & Distributors, New Delhi, India, pp. 31-43.

Wieniawski, W. (2001) Risk assessment as an element of drug control. WHO Drug Information 15: 7–11.

Xie, H., Huo, K.K., Chao, Z. and Pan, S.L. (2009) Identification of crude drugs from Chinese medicinal plants of the genus *Bupleurum* using ribosomal DNA ITS sequences. Planta Medica 75: 89-93.

Xie, H., Zhao, Z.L., Huo, K.K., Wu, B.Y. and Pan, S.L. (2006) ITS sequence of 9 *Bupleurum* species and its application in identification of Chaihu (Radix Buplurei). Journal of Southern Mediclal University 26: 1460-1463.

Xu, H., Li, X.B., Wang, Z.T., Ding, X.Y., Xu, L.S. and Zhou, K.Y. (2001) rDNA ITS sequencing of Herba Dendrobii (hunagcao). Acta Pharmaceutica Sinica 36: 777-783.

Xu, H., Wang, Z., Ding, X., Zhou, K. and Xu, L. (2006) Differentiation of *Dendrobium* species used as “Huangcao Shihu” by rDNA ITS sequence analysis. Planta Medica 72: 89-92.

Yan, P., Pang, Q.H., Jiao, X.W., Zhao, X., Shen, Y.J. and Zhao, S.J. (2008) Genetic variation and identification of cultivated *Fallopia multiflora* and its wild relatives by using chloroplast *matK* and 18S rRNA gene sequences. Planta Medica 74: 1504-1509.

Yang, D.Y., Fushimi, H., Cai, S.Q. and Komatsu, K. (2004) Molecular analysis of *Rheum* species used as Rhei Rhizoma based on the chloroplast *matK* gene sequence and its application for identification. Biological and Pharmaceutical Bulletin 27: 375-383.

Yang, Y., Zhai, Y., Liu, T., Zhang, F. and Ji, Y. (2010) Detection of *Valeriana jatamansi* as an adulterant of medicinal *Paris* by length variation of chloroplast *psbA–trnH* region. Planta Medica 77: 81-97.

Yang, Z.Y., Chao, Z., Huo, K.K., Xie, H., Tian, Z.P. and Pan, S.L. (2007) ITS sequence analysis used for molecular identification of the *Bupleurum* species from northwestern China. Phytomedicine 14: 416-423.

Yao, H., Song, J.Y., Ma, X.Y., Liu, C., Li, Y., Xu, H.X., Han, J.P., Duan, L.S. and Chen, S.L. (2009) Identification of *Dendrobium* species by a candidate DNA barcode sequence: the chloroplast *psbA- trnH* intergenic region. Planta Medica 75: 667-669.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P. and Chen, S. (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE*,* 5(10): e13102.

Yu, M.T., Wong, K.L., Zong, Y.Y., Shaw, P.C. and Che, C.T. (2008) Identification of *Swertia mussotii* and its adulterant *Swertia* species by *5S rRNA* gene spacer. China Journal of Chinese Materia Medica 33: 502-504.

Yu, Y.B., Qin, M.J., Liang, Z.T., Yu, G.D. and Tan, N.H. (2003) Ribosomal DNA ITS sequence comparisons of *Pseudostellaria heterophylla* from different geographical regions. Journal of Plant Resources and Environment 12: 1-5.

Zeng, M., Ma, Y.J., Zheng, S.Q., Xu, J.F. and Di, X.H. (2003) Studies on ribosomal DNA sequence analyses of *Radix puerariae* and its sibling species. Chinese Pharmaceutical Journal 38: 173-175.

Zhang, H.Y. and Shi, X.G. (2007) Analysis of rDNA ITS sequences in root tuber of *Polygonum multiflorum* from various habitats. Chinese Traditional and Herbal Drugs 38: 911-914.

Zhang, N., Yan, B., Xu, X.H. and Xu, L.C. (2010a) Sequence analysis of rDNA-ITS of Baishouwu from different species. China Journal of Chinese Materia Medica 35: 1537-1540.

Zhang, W., Han, Y.L. and Zhu, J.H. (2010b) Analysis of rDNA ITS sequences of plants of *Aconitum* L. and its sibling species. Journal of Biology 27: 50-52.

Zhang, X.L., Ji, K.P., Li, Y.D., Chen, C., Li, X.H. and Liu, L.H. (2003) Studies on establishing rRNA gene map of *Angelica sinensis* and *Rheum palmatum* from Gansu by DNA sequencing. Journal of Chinese Medicinal Materials 26: 481-483

Zhang, Y., Zhang, J.C., Huang, M.H., Yang, M.S. and Cao, H. (2006) Detection of genetic homogeneity of *Panax notoginseng* cultivars by sequencing nuclear 18S rRNA and plastid *matK* genes. Planta Medica 72: 860-862.

Zhao, G.P., Niizeki, M., Ishigawa, L., Qian, S.Q., Zhang, Q. and Ao, J.N. (2006) ITS sequence analysis of Chinese and Japanese medicinal plants of *Angelica* L. Acta Pharmaceutica Sinica 37: 1072-1076.

Zhao, H.G., Zhou, J.J., Cao, S.S., Zheng, Y.H., Shan, Y. and Xia, B. (2009) Analysis of interspecific relationship among *Stellaria media* and its related species based on ITS and *trnL*-F sequence differences. Journal of Plant Resources and Environment 18: 1-5.

Zhao, Z.L., Leng, C.H. and Wang, Z.T. (2007) Identification of *Dryopteris crassirhizoma* and the adulterant species based on cpDNA *rbcL* and translated amino acid sequences. Planta Medica 73: 1230-1233.

Zhao, Z.L., Zhou, K.Y., Dong, H. and Xu, L.S. (2001) Characters of nrDNA ITS region sequences of fruits of *Alpinia galanga* and their adulterants. Planta Medica 67: 381-383.

Zhao, Z.L., Zhou, K.Y., Wang, Z.T., Qin, M.J. and Dong, H. (2000) Determination of nrDNA ITS sequences of Caodoukou (*Alpinia hainanensis* K. Schum.) and Yizhi (*A. oxyphylla* Miq.). Journal of Plant Resources and Environment 9: 38-40.

Zhou, L., Wang, P.X., Huang, F., Cao, L.Y. and Liang, R.Y. (2002) ITS sequence analysis of *Amomum villosum*. Chinese Traditional and Herbal Drugs 33: 72-75.

Zhu, S., Fushimi, H. and Cai, S.Q. (2003) A new variety of genus *Panax* from southern Yunnan, China and its nucleotide sequences of 18S ribosomal RNA gene and *matK* gene. Journal of Japanese Botany 78: 86-94.

Zhu, Y., Qin, M.J., Hang, Y.Y. and Wang, L.Q. (2007) Authentication of *Pseudostellaria heterophylla* and its counterfeit species by analysis of rDNA ITS sequences. Chinese Journal of Natural Medicines 5: 211-215.

Zuo, Y., Chen, Z., Kondo, K., Funamoto, T., Wen, J. and Zhou, S. (2011) DNA barcoding of *Panax* species. Planta Medica 72(2): 182-87.

*****

Plant DNA Barcoding and Phylogenetics, Pages 143-154
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

9 Thermocycling in Systematics

A. Alam

Introduction

Polymerase Chain Reaction (PCR) is a technique of molecular biology to amplify single/few copies of DNA fragment to million copies within hours employing a simple enzymatic reaction that can serve as templates for downstream applications. The PCR technique is one of most extensively used methods for analysing DNA. Almost all reagents or parameters can be changed to suit the need of an individual researcher. This unique combination of flexibility with specificity has led to the advent of PCR as a principal laboratory technology. The word “polymerase” is used because the only enzyme used in this reaction is DNA polymerase. As the products of the first reaction becomes the substrates for the subsequent reaction, and hence the term “chain” and as it is the reaction of various components (DNA template, de-oxynucleotide triphosphates, DNA Polymerase, Mg^{++} ions and buffer solution) justifies the usage of the word “reaction”.

Historical background

Although the process was first described by Kjell Kleppe and H.G. Khorana in 1968, the concept of PCR was first discovered by an American biochemist Dr. Kerry Mullis in 1983. Dr. Kerry Mullis was then working at Cetrus Corporation in Emeryville, CA and reported it in the year 1985. Mullis got the idea of the PCR while driving during the night in the Californian mountains. In 1993, Mullis was awarded the Nobel Prize in Chemistry along with Michael Smith for his work on PCR, a basic technique in biochemistry and molecular biology.

Thermocycles in PCR

Amplification of the specific DNA fragment occurs in three steps, viz. denaturation, annealing and extension and each step is repeated for 35-40 cycles (Figure 1). This is done in an automated thermal cycler. The reaction mixture in tubes is heated and cooled in a very short time. The PCR product increases exponentially as both strands are copied during PCR. For instance, if PCR is initiated with one copy of gene, there will be two copies, after two cycles there will be four copies and so on, and after 35 cycles there will be 2^{35} copies (Figure 2).Too few cycles result in low product yield while too many cycles give non-specific background products.

Denaturation

Doubled stranded DNA is separated into single strands in the denaturation steps to facilitate the annealing of primers. In the denaturation process, the hydrogen bonds that connect the two DNA strands are broken. This is done by heating to a temperature above the melting temperature. Prior to the denaturation cycles, the DNA is often initially denatured for an extended time, usually up to 5-10 minutes, to ensure complete separation of both the template DNA and primers into single strands. PCR works with denaturing temperature of 91-97^{o}C because when nucleic acids are heated in ionic strength lower than 150 mM NaCl, the melting temperature is usually less than 100^{o}C. The enzyme generally used in the PCR is *Taq* polymerase which has half-life of 30 minutes at 95^{o}C. This half-life cannot support more than 35-40 amplification cycles.

Primer annealing

After denaturation, the temperature is lowered during annealing step (about 60 seconds) for the purpose of annealing with primers before the single stranded template DNA binds themselves. Base composition, length and concentration of the primers determine the temperature and the time required for primer annealing, and is usually 5^{o}C below the lowest melting temperature (T_m). Annealing temperature in the range of 55 to 65°C generally yield the best result.

Extension/Elongation

During the elongation step, the DNA polymerase copies the template DNA. The nucleotides, complementary to template are added to the primers on the 3' side of the primer from 5' to 3' side direction of the template. Duration of elongation is determined by length and nucleotide composition of the DNA template and temperature. Elongation temperature depends on the type of DNA polymerase used. The DNA polymerases works well at 72°C and the rate of nucleotide addition at this temperature varies from 35 to100 nucleotides per second which in turn depends on the pH, salt concentration, nature of DNA template and buffer of the PCR reaction mixture. As a thumb rule, 1 minute is considered optimum for an extension of one Kb of DNA product.

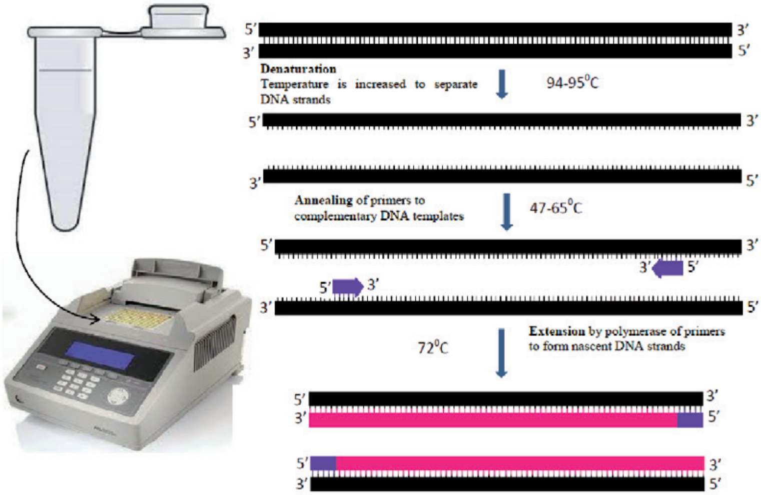

5'
3'
3'
5'
Denaturation
Temperature is increased to separate DNA strands
94-95°C
5'
3'
3'
5'
Annealing of primers to complementary DNA templates
47-65°C
5'
3'
3'
5'
5'
3'
3'
5'
Extension by polymerase of primers to form nascent DNA strands
72°C
5'
3'
3'
5'
5'
3'
3'
5'

Fig. 1: Image showing the different steps in Polymerase Chain Reaction (PCR)

Chemical Components for PCR Reaction

The components for the PCR amplification are the DNA template, a reverse and a forward primer, the thermostable DNA polymerase, four type of de-oxynucleotide triphosphates (dATP, dCTP, dGTP and dTTP) and reaction buffer containing magnesium ion (pH 8.4).

Template DNA

Higher or lower quantity of DNA template gives poor amplification. Hence, tiny amount of DNA template containing 100 to 10000 molecules is considered sufficient in the PCR reaction mixture which is usually 60-80 ng. No proportionate advantage in increasing the starting material to 1μg over 100 ng and ten-fold advantage to 100 ng over 100 ng was observed in amplicon production.

De-oxynucleotide triphosphates (dNTPs)

The dNTPs are the building blocks of DNA which is linked to 3' end of the primer with the help of the thermostable DNA polymerase. Each de-oxynucleotide triphosphates (dATP, dCTP, dGTP and dTTP) are used at concentration of 200 μM (0.2mM). High concentration of dNTPs results in mis-incorporations.

Primers

Two primers, each complementary to the specific target sequence called “forward” and “reverse” primers are the most important components of the PCR reaction. In DNA barcoding usually “Universal” primers are used that target DNA sequences shared by any species containing sequence of interest. The length of the primer ranges from 18 to 30 bases. Primers of greater length are found to result in non-specific priming and mismatch pairing. The concentration ranging from 10 to 50 pmoles is generally used in a PCR reaction. Annealing of the primers to the template DNA depends on melting temperature of primer (T_m).

T_m = [(number of A+T residues) X 2] + (number of G+C residues) X 4] ^{o}C

Thermostable polymerase

Earlier thermo-labile polymerase was used in the PCR which required the addition of enzymes during each cycle because at denaturation step, polymerase activity is destroyed. Then the discovery of *Taq* polymerase isolated from the bacterium (*Thermus aquaticus*) that inhabited hot springs actually revolutionized the PCR technology. This enzyme could withstand high temperatures. However, a recombinant *Taq,* a cloned version of the enzyme is the most widely used in almost all PCR. *Taq* is able to resist temperatures above 90°C. It has a half-life of about 40 min at 95°C. *Taq* DNA polymerase result in PCR product with 'A' overhangs which is exploited to produce TA cloning and TOPO cloning. Major drawback with *Taq* polymarse is its low fidelity that does not have 3' to 5' exonuclease proofreading activity. Because of this, it cannot remove the incorrect added nucleotides. Another thermostable polymerase known as *pfu* posses superior thermostability and proofreading activity compared to other thermostable polymerases. As a result, the PCR products generated by *pfu* polymerase are blunt ended with fewer misincorportaed nucleotides. The concentration of *Taq* DNA polymerase varies from 1 to 2.5 units per 100 µl of the reaction mixture.

PCR buffer

The reaction buffer contains KCL, Tris HCl (pH 8.4), $MgCl_2$. Mg^{2+} serves as essential co-factor for all Type II enzymes including restriction endonucleases and polymerases. It form soluble complex with dNTPs which is essential for dNTP incorporation. Mg^{2+} also stimulates polymerase activity and enhances the melting temperature of the primer/template interaction, thereby stabilizing the duplex interaction. When 200 µM concentrations each of dNTPs is used, about 1.0 to 1.5 mM is usually considered optimum. Low concentration results in low yield and excess of Mg^{++} produces non-specific amplification. Buffer contains 10 mM Tris HCl (pH 8.4) and KCl of up to 50 mM. KCl facilitates primer annealing. Higher concentration of KCl inhibits the DNA polymerase activity. Gelatin or bovine serum and ionic detergents are included which help stabilize the enzyme. PCR reaction buffer is usually supplied as 10X concentrate. Reduction of PCR reaction below 1X seriously hampered the amplicon production.

Volume calculation for PCR reaction

The volume of ingredients required in the PCR reaction is calculated by equation: $\mathbf{C_S \ X \ V_1 = C_F \ X \ V_2}$ (Where, C_s = conc. of the ingredient in the stock solution, C_F = conc. of the ingredient required in the reaction, V_1 = vol. of the ingredient required for the reaction, V_2 = total volume of the reaction)

Setting up the PCR reaction

Sterile 0.2 ml PCR tubes with proper label is taken and kept on PCR rack. PCR components consisting of deionised sterile water (18.77 µl), PCR buffer (2.5 µl), dNTP mix (0.5 µl), primers (0.2 µl of each primer), $MgCl_2$ (1.5 µl), *Taq* polymerase (0.33 µl) and finally template DNA (1 µl) are added in the order as given (Table 1) for 25 µl assay. Tightly close the PCR tubes and contents are mixed by gently tapping with fingers or spinning in mini microfuge tube at 6000 rpm for 20 seconds. The PCR tube then load into PCR machine/ thermal cycler to run the PCR program (Table 2).

Table 1: Final concentration and volume of PCR component in the reaction mixture.

PCR component	Final conc. In the reaction mixture	Volume
Distilled water	As required	18.77 µl
10X PCR Buffer	1X	2.50 µl
10 mM dNTP mix	200 µM	0.50 µl
Forward primer	5 – 10 pmoles	0.20 µl
Reverse primer	5 -10 pmoles	0.20 µl
$MgCl_2$(15mM)	1.5 mM	1.50 µl
Taq Polymerase (6 U/ µl)	1 U	0.33 µl
DNA template (80 ng / µl)	---	1 µl
	Total volume	25 µl

Table 2: A typical PCR program**.**

Steps	Temperature (^{0}C)	Time (seconds)	Cycle (s)
Initial denaturation	95	300	1
Denaturation	94	30	35
Annealing	55	30	35
Extension	72	45	35
Final extension	72	600	1
Hold	4		

Detection of PCR product using Agarose Gel Electrophoresis

Gel electrophoresis is technique of separation of nucleic acids and protein under the influence of an electric field. At a particular pH, the biological molecules exist in solution as electrically charged particles. These charged particles will migrate either to the cathode or anode depending on the nature of charge and their net charge.

When a potential differences (voltage) is applied across the electrodes, it generates potential gradient, E, which is the applied voltage (V) divided by distance (d) between the electrodes (*E = V/d)*. The force that drives a charged molecule towards an electrode is the product of potential gradient and the charge of q coulombs on the particle (*F = Eq)*. However, the frictional force that retards the movement of a charged molecule is dependent on the hydrodynamic size of the molecule and shape of the molecule, the pore size of the electrophoresis medium and the viscosity of the buffer.

The velocity (v) of charged particle in an electric field- **v = Eq / f**.

Electrophoretic mobility (M) can be defined as of an ion can then be defined by the ion's velocity divided by the potential gradient: *M = v / E.* In addition M can be equivalently expressed as the charge of the molecule, *q*, divided by the frictional coefficient, *f (M = q / f).*

Thus when electricity is applied to the medium containing biological molecules, depending on their net charge and size, molecules (nucleic acids/proteins) begin to migrate due to their different electrophoretic mobility resulting in the their (DNA/protein) separation. DNA can be separated either by Agarose gel electrophoresis

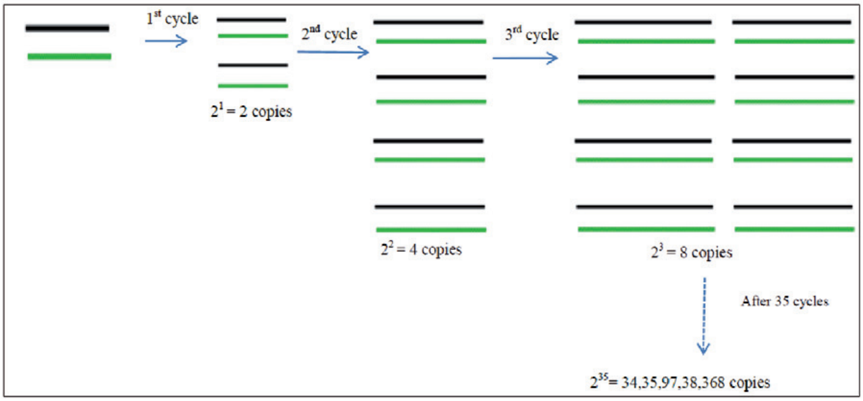


Fig. 2: Image showing the number of copies synthesized in different cycles of PCR.

Preparation of agarose gel

Agarose typically is used at 0.5 to 2% (Table 3). For the preparation of 2% gel, 600 mg of Agarose is dissolved by heating in a flask containing 30 ml TAE or TBE (0.5X). When the solution temperature comes down to 55-60^0 C, 1.2µl of (Ethidium bromide) EtBr is added in the solution.

Thereafter, this solution is poured in the casting plate with adjusted gel comb which is then left at room temperature for around 45 minutes to solidify. After the gel polymerisation, 0.5 X TAE buffer is added in sufficient quantity as running buffer.

Table 3: Recommended agarose gel percentages for resolution of DNA

% Agarose	DNA size (range in bp)
0.75	10,000-15,000
1.0	500-10,000
1.25	300-5000
1.5	200-4000
2.0	100-2500
2.5	50-1000

Loading amplicons in the gel

1.0µl of PCR product is mixed with 5.0 µl of 0.5 x TAE and 1.0 µl of DNA loading dye (0.25% bromophenol blue, 0.25% xylene cyanol, 30% glycerol in water) and loaded into the wells along with suitable DNA marker (may be 100 bp). The amplicons are run in electrophoresis machine at 70V for 30-40 minutes and the band patterns are visualised under UV light and photographed using gel documentation system (Figure 3).

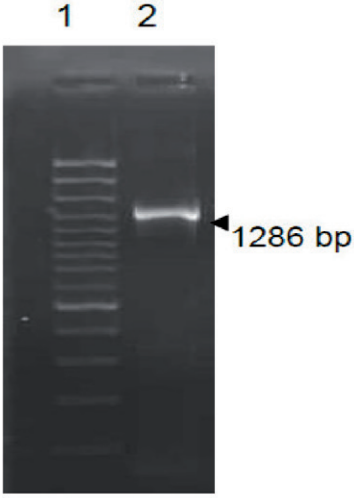


Fig. 3: Lane 1: Gene Ruler^{TM} 100 bp DNA ladder Plus (MBI Fermentas); lane 2: PCR product of 1286 bp.

Applications

PCR is one of the most widely used techniques in molecular biology. It is able to amplify minute quantity of DNA rapidly and specifically to such an extent that the DNA becomes easy to detect, study and used for any desired purpose. PCR is a versatile tool. It can be used in a wide variety of ways. This method of amplifying a rare sequence from a mixture has numerous applications in basic research viz. cloning, gene expression studies, site directed mutagenesis, mutation screening, drug discovery, construction of cDNA library, classification of organism, genotyping, molecular ecology, molecular archaeology, molecular epidemiology and bioinformatics etc. In the applied research, it finds application in the areas of DNA fingerprinting, genetic matching, detection of pathogens, pre-natal diagnosis, gene therapy and many more. Since it is feasible with PCR to execute analyses on extremely small quantity of DNA, it is easy to determine genetic and evolutionary connections between different species.

Polymerase chain reaction is one of basic step in DNA fingerprinting. PCR based DNA markers are RAPD (employs a short arbitrary PCR primer of 10 nucleotides), Amplified Fragment Length Polymorphism (AFLP), microsatellites (Simple Sequence Repeat SSR). PCR products generated often vary in length. This polymorphism can be used in identification and classification of strains, population and higher taxonomic groups. DNA barcoding is a species identification tools that involves the use of a short DNA sequence or sequences from a standardized locus (or loci). In animals it is well established. Herbert (2004) demonstrated that a sequence of a 655 base fragment of 5’ end of the mitochondrial cytochrome c oxidase subunit I (COI) is appropriate for discriminating closely related species across diverse animal Phyla. For the plant species, chloroplast genes, viz. megakaryocyte-associated tyrosine kinase (matK), ribulose-bisphosphate gene (rbcL) and ITS (Internal Transcribed Spacers) region of about 700bp are suitable loci. PCR in combination with DNA sequencing is now playing a central role for studies of the systematics of plant and animal species.

In summary, the PCR (developed in 1983 by Kary Mullis) is now a common and often indispensable technique used in medical and biological research laboratory for a variety of applications (Bartlett and Stirling, 2003). These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases (Saiki et al., 1985, 1988). In 1993, Mullis was awarded the Nobel Prize in Chemistry along with Michael Smith for his work on PCR (see Kary Mullis Nobel Lecture, December 8, 1993, http://www.nobelprize.org/nobel_prizes/chemistry/Laureates/1993/mullis-lecture.html). Polymerase chain reaction (PCR) has been used in research and seems to be both sensitive and specific: commercial kits are now being developed for the amplification, for example AccuPrep® PCR Purification Kit (http://us.bioneer.com/Protocol/AccuPrep%20PCR%20Purification%20Kit.pdf).

Gel electrophoresis is a method for separation and analysis of macromolecules (DNA, RNA and proteins) and their fragments, based on their size and charge. It is used in clinical chemistry to separate proteins by charge and/or size (IEF agarose, essentially size independent) and in biochemistry and molecular biology to separate a mixed population of DNA and RNA fragments by length, to estimate the size of DNA and RNA fragments or to separate proteins by charge (Kryndushkin et al., 2003). Nucleic acid molecules are separated by applying an electric field to move the negatively charged molecules through a matrix of agarose or other substances. Shorter molecules move faster and migrate farther than longer ones because shorter molecules migrate more easily through the pores of the gel. This phenomenon is called sieving (Sambrook and Russel, 2001). Proteins are separated by charge in agarose because the pores of the gel are too large to sieve proteins. Gel electrophoresis can also be used for separation of nanoparticles. Gel electrophoresis uses a gel as an anticonvective medium and/or sieving medium during electrophoresis, the movement of a charged particle in an electrical field. Gels suppress the thermal convection caused by application of the electric field, and can also act as a sieving medium, retarding the passage of molecules; gels can also simply serve to maintain the finished separation, so that a post electrophoresis stain can be applied. DNA gel electrophoresis is usually performed for analytical purposes, often after amplification of DNA via PCR, but may be used as a preparative technique prior to use of other methods such as mass spectrometry, RFLP, PCR, cloning, DNA sequencing, or Southern blotting for further characterization.

References

Bartlett, J.M. and Stirling, D. (2003) A short history of the polymerase chain reaction. Methods in Molecular Biology 226: 3-6.

Kryndushkin, D.S., Alexandrov, I.M., Ter-Avanesyanm, M.D. and Kushnirov, V.V. (2003) "Yeast [PSI+] prion aggregates are formed by small Sup35 polymers fragmented by Hsp104". Journal of Biological Chemistry 278 (49): 49636–43.

Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., Mullis, K. and Erlich, H. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239 (4839): 487–491.

Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H. and Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia". Science 230 (4732): 1350–1354.

Sambrook, J. and Russel, D.W. (2001) Molecular Cloning: A Laboratory Manual 3rd Ed. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY.

*****

Plant DNA Barcoding and Phylogenetics, Pages 155-170
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

10 DNA Sequencing

A. Alam

Introduction

Genes are located on the chromosome, and are made of deoxyribonucleic acid (DNA). A strand of DNA contains numerous genes. All these genes contain complete information needed to code for molecules called proteins. The nucleotide in DNA (Figure 1) consist of a phosphate, a sugar (deoxyribose) and one of the four bases [cytosine (C), thymine (T), adenine (A), guanine (G)].

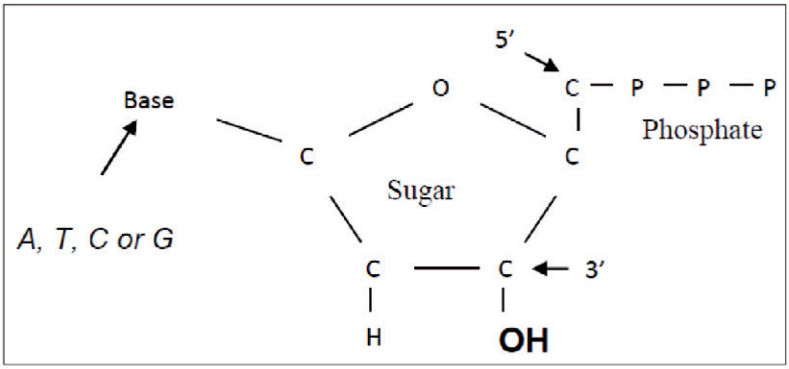


Fig. 1: Depicting the structure of a nucleotide: A basic unit of DNA.

DNA sequence is the detailed description of the order of building blocks or bases. DNA sequencing is the process of determining the precise order of the four nucleotides bases Adenine, Guanine, Thymine and Cytosine in a stretch of DNA. DNA sequencing has proved useful in revealing the kind of genetic information in the form of base sequences that are carried in particular segment of the DNA. The sequence information can be analysed to hunt genes and also to determine the changes in sequences of a gene (mutation). Knowledge of DNA sequences has become indispensable for basic research and other applied fields requiring DNA sequencing data like diagnostics, biotechnology, forensic biology, biological systematics etc. The advent of DNA sequencing tool has revolutionised the biological research and discovery.

Historical background

The double-stranded, helical, complementary, anti-parallel model for DNA was proposed by James Watson and Francis Crick in the year 1953. The technique of radioactive labelling was introduced to Sanger by Chris Anfinsen, a United States biochemist in 1954. During 1961-1963, researchers crack the genetic code linking gene and protein. DNA polymerase which was the first enzyme to make DNA in a test tube was discovered and isolated by Coenberg in 1958. The tRNA was the first nucleic acid molecule sequenced by Holley and coworkers in 1965. Discovery of Type-II restriction enzymes by Hamilton Smith and co-workers in 1970 was an important event in history of the DNA sequencing. Hans Kosel successfully synthesised DNA primer in 1970. In the year 1972, Herbert Boyer discovered the restriction enzyme EcoR1. In 1975, Sanger and Coulson introduced the plus and minus method for DNA sequencing. In the year 1977, DNA sequencing method called Maxam Gilbert was discovered. These two methods were replaced by chain terminator method/Sanger sequencing/dideoxy terminator method. Obtainable sequence length was approximately 100 nucleotides. In 1983, Karry Mullis invented Polymerase Chain Reaction (PCR). Dye terminator sequencing method reported in 1986 used four dideoxynucleotide chain terminators, labelled with different fluorescent dye, permitted sequencing in a single reaction rather than four. The obtainable sequence by this method was around 1000 nucleotides. Leroy Hood and Llyod Smith of California Institute of technology and colleagues declared the first automated DNA sequencing machine in 1986. During 1990s Pyrosequencing was developed by Pal Nyren. Sanger sequencing using dye-terminators was the principal sequencing technique until the introduction of so-called next-generation sequencing technologies beginning in 2005. History of genome sequencing is presented in Table 1.

Table 1: History of genome sequencing

Organism type	Organism	Size	Note
Virus	Bacteriophage MS2	3.5 kb	First sequenced RNA-genome (1976)
Virus	Phage Φ-X174	5.4 kb	First sequenced DNA-genome (1977)
Bacterium	*Haemophilus influenzae*	1.8 Mb	First genome of a living organism sequenced (July 1995)
Yeast	*Saccharomyces cerevisiae*	12.1 Mb	First eukaryotic genome sequenced –completed in 1996
Nematode	*Caenorhabditis elegans*	100 Mb	First multicellular animal genome sequenced (December 1998)
Plant	*Arabidopsis thaliana*	157 Mb	First plant genome sequenced (December 2000)
Insect	*Drosophila melanogaster* (fruit fly)	130 Mb	In the year 2000
Mammal	*Homo sapiens*	3 billion bps	Sequencing completed in the year 2003
Puffer fish	*Takifugu rubripes*	390Mb	First fish genome sequencing completed in the year 2002

Plus and Minus method of DNA sequencing

A particular stretch of the DNA is synthesized using DNA polymerase to generate series of DNA molecules of varying lengths. The unused dNTPs were removed. The DNA synthesis was continued in four pairs of minus and plus reaction mixtures. The minus reaction mixture had three dNTPs while plus reaction mixture had only one dNTPs. DNA so generated was separated by electrophoresis. Then each minus and plus pair were compared to indicate the length of the new polydeoxyribonucleotide (by the mobilities of the bands) and the position at which polymerization had terminated as a result of the absence of the missing dNTP. Employing this method, the genome of ΦX174 bacteriophage was sequenced. The disadvantage of this method was that it was only useful on single stranded DNA, and required both the plus and minus sequences in order to ensure that the result was completely accurate.

Maxam Gilbert method of DNA sequencing

This method requires the splicing of the purified DNA by the restriction endonucleases. The phosphate at 5'end of the cleaved fragment was removed using phosphatase and replaced by radioactive phosphate (^{32}P) employing the enzyme Kinase. This radioactively labeled fragment was again subjected to another restriction endonuclease which further cut the DNA fragment. The DNA fragments were separated by Electrophoresis to separate the two, labeled and unlabeled end sub fragments from each other resulting in sub fragments with one labeled and an unlabeled ends. The DNA sub fragment whose sequence required identified was purified from the gel and separated from its other end labeled sub fragment. The end labeled DNA sub fragments was further divided and placed in four base specific chemical solutions (G, A & G, C and C & T) to generate DNA fragments without bases. They were then treated with reagents that break the DNA fragment at sites from where the bases had been removed resulting in DNA fragments of different length. These four reaction samples were then subjected to electrophoresis with each reaction ran on its own lane. After electrophoresis the gel was removed, dried and subjected to autoradiography. A dark band in each reaction indicated the presence of a base (Figure 2). DNA sequence was then read from the bottom of the gel to the top of the gel. This method did not gain much popularity because of the extensive use of hazardous chemical and technical complexity.

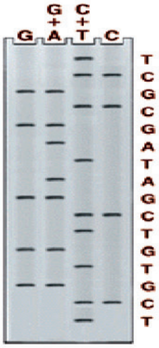


Fig. 2: Separation and detection of ^{32}P-labeled DNA fragments by polyacrylamide gel electrophoresis (PAGE). (Source: http://nationaldiagnostics.com/article_info.pparticles_id/20).

Chain terminator method of DNA sequencing

This method is also known as Sanger or dideoxy sequencing. In short, it is replication of the single stranded DNA fragment to be sequenced using DNA polymerase, Mg^{+2}, dNTPs, ddNTPs and ^{32}P labelled primers. Frederick Sanger was awarded the Nobel Prize in 1980 for developing this technique. Dideoxynucleotide triphosphates (ddNTPs) are used as chain terminators.

The DNA sample is divided into four separate sequencing reactions containing DNA template, all the four nucleotides (dATP, dCTP, dGTP and dTTP) and the DNA polymerase. To each reaction, one of the four dideoxynucleotide triphosphates (ddA, ddC, ddG and ddT) was added, which lack the OH group at the 3′ carbon resulting in DNA fragments of assorted length. The newly synthesized DNA fragments were heat denatured and separated by gel electrophoresis with each of the four reaction run in one of the individual lanes (A, T, G and C). The DNA bands were then visualized by autoradiography or UV light, and then DNA sequence read off directly from the X-ray film or gel image lanes (Figure 3).

Dye terminator sequencing

Over a period, Sager sequencing method witnessed great advances in the technique such as fluorescent labelling, capillary electrophoresis, and general automation. Dye terminator sequencing is a variant of Sanger sequencing in which each of four dideoxynucleotide triphosphates (dATP, ddCTP, ddGTP and ddTTP) is labelled with a fluorescent dye. This permitted sequencing in one reaction rather than four. Sanger sequencing using dye terminators became the dominant sequencing technique until the introduction of so-called next-generation sequencing technologies beginning in 2005.

The sequencing reaction mixture contains the DNA template, Taq polymerase (the enzyme to duplicate DNA), a single primer, normal nucleotide: deoxy-adenosine triphosphate (dATP), deoxy-guanosine triphosphate (dGTP), deoxy-cytidine triphosphate (dCTP), deoxy thymidine triphosphate (dTTP)) and a limited number of fluorescently tagged dideoxynucleotides (**T**, **C**, **G**, **A**), which lack the OH group at the 3′ carbon. The reaction mixture is subjected to polymerase chain reaction with each cycle consisting of denaturation (94-98^{0}C for 20-30 seconds), annealing (50-65^{0}C for 20-30 seconds) and extension (70-75^{0}C). During the elongation, when a dideoxynucleotide is added, the DNA extension is terminated resulting in a gradual build-up of differently sized fragments with many copies of each of the following products:

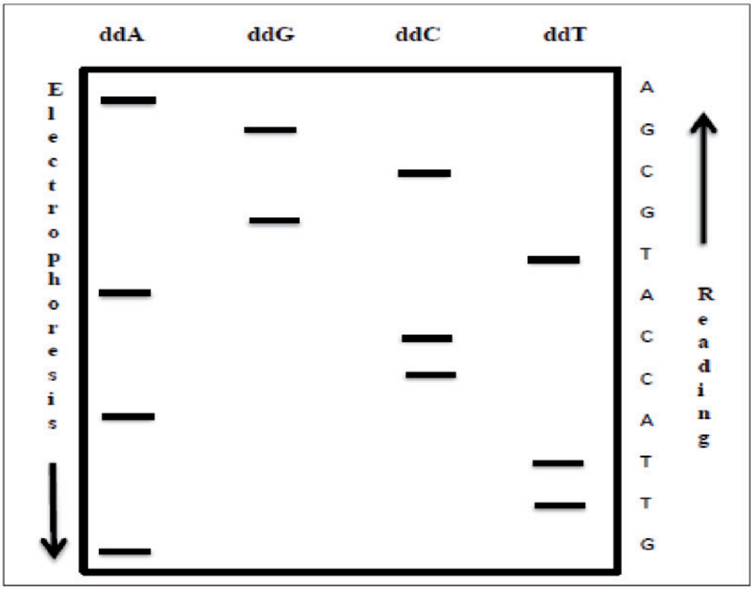


Fig. 3: The extended ^{32}P -labeled primers are separated by PAGE.

The differently sized DNA fragments in the reaction mixture can be size-separated by running on a single lane of a gel (Figure 4) and the inferred nucleotide sequence can then be deduced by a computer. Later, the method was performed with automated sequencing machines that used capillary electrophoresis (CE). Capillary electrophoresis uses a denaturing flowable polymer, has largely replaced the use of gel separation technique (Polyacrylamide Gel Electrophoresis PAGE). During the capillary electrophoresis, the extended product enters the capillary as a result of the electrophoretic injection. A high voltage applied to the buffered sequencing reaction causes the negatively charged fragments to separate according to their size based on their total charge. A machine scans the lane with a laser detection device. The laser beam excites the dye on the fragment to fluoresce at excitation wavelength of 460 nm and emission as fluorescence (emission wavelength of ddATP - 512 nm, ddCTP - 519 nm, ddGTP - 505 nm and ddTTP - 526 nm) from the label conjugated to the ddNTPs can be read by a photocell and recorded on a computer (Figure 5). The key to the colors used in the chromatogram are: **red = T, black = G, blue = C, green = A**. The computer output for a sequencing run consists of chromatogram that can be opened in DNA sequence viewers such as FinchTV, Artemis, BioEdit, Sequence Navigator, DNA etc.

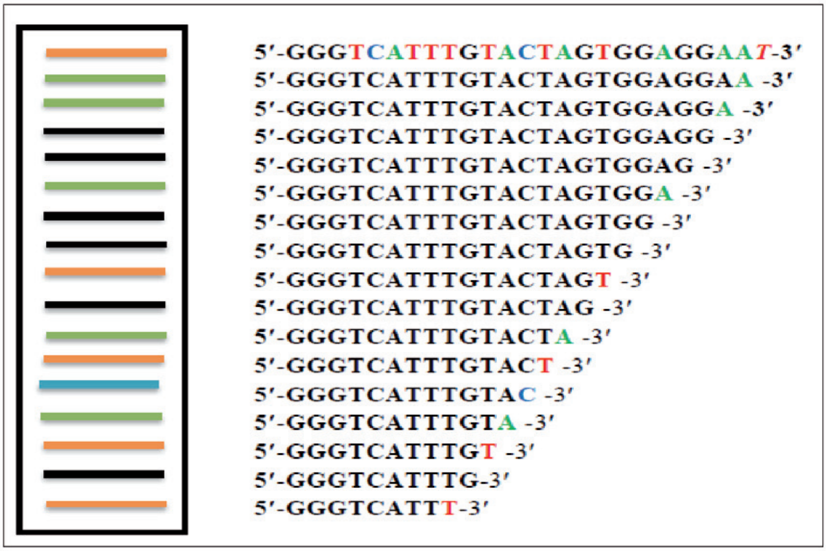

5′-GGGTCATTTGTACTAGTGGAGGAA*T*-3′
5′-GGGTCATTTGTACTAGTGGAGGAA -3′
5′-GGGTCATTTGTACTAGTGGAGGA -3′
5′-GGGTCATTTGTACTAGTGGAGG -3′
5′-GGGTCATTTGTACTAGTGGAG -3′
5′-GGGTCATTTGTACTAGTGGA -3′
5′-GGGTCATTTGTACTAGTGG -3′
5′-GGGTCATTTGTACTAGTG -3′
5′-GGGTCATTTGTACTAGT -3′
5′-GGGTCATTTGTACTAG -3′
5′-GGGTCATTTGTACTA -3′
5′-GGGTCATTTGTACT -3′
5′-GGGTCATTTGTAC -3′
5′-GGGTCATTTGTA -3′
5′-GGGTCATTTGT -3′
5′-GGGTCATTTG-3′
5′-GGGTCATTT-3′

Fig. 4: Differentially sized fragment generated by Sanger sequencing using primer 5′-GGGTCATT-3′.

Shortgun Sanger sequencing methods

It is a method of sequencing a DNA fragment longer than 1000 nucleotide base pairs. The genome was fragmented and cloned into a plasmid vector which is transformed in *E. coli*. For each sequencing reaction, a single colony is picked and the plasmid DNA isolated. The PCR was then performed using the four fluorescently labeled dideoxy nucleotide triphosphate (ddNTPs) generating a ladder of ddNTP terminated, dye labeled product. They were then subjected high resolution gel electrophoretic separation with one of 96 or 384 capillaries with one run of a sequencing instrument. After sequencing individual fragments, the sequences can be reassembled on the basis of their overlapping regions (Figure 6)

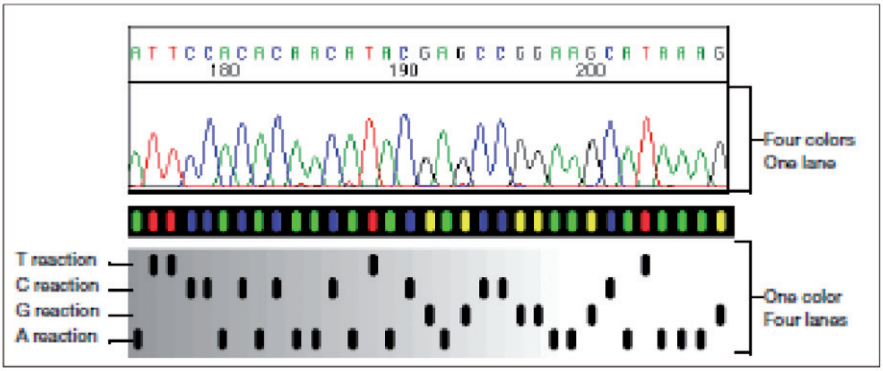


Fig. 5: Fluorescent sequencing compared with radioactive sequencing (Source: Wikipedia).

Primer walking methods of DNA sequencing

First a segment of the genome is sequenced using the primer designed from a known region. Subsequently, a primer is designed to hybridize 3'region, determined in the previous steps. These primers then serve as start point to establish an additional >500 bp of sequence data. Again, the primes are designed from the newly established sequence. The process is repeated and the sequences reassembled on the basis of their overlapping regions.

Expressed sequence tags (EST) sequencing methods

EST has proved to be a useful tool in the identification of active genes in a tissue. mRNA is first isolated and then cloned in a plasmid vector. After cloning in a vector it is followed by transformation in a bacteria, isolation of bacterial, plasmid DNA extraction and DNA sequencing. Each clone is sequenced only once to decipher a “single pass” sequence tags of about 300-800 bp.

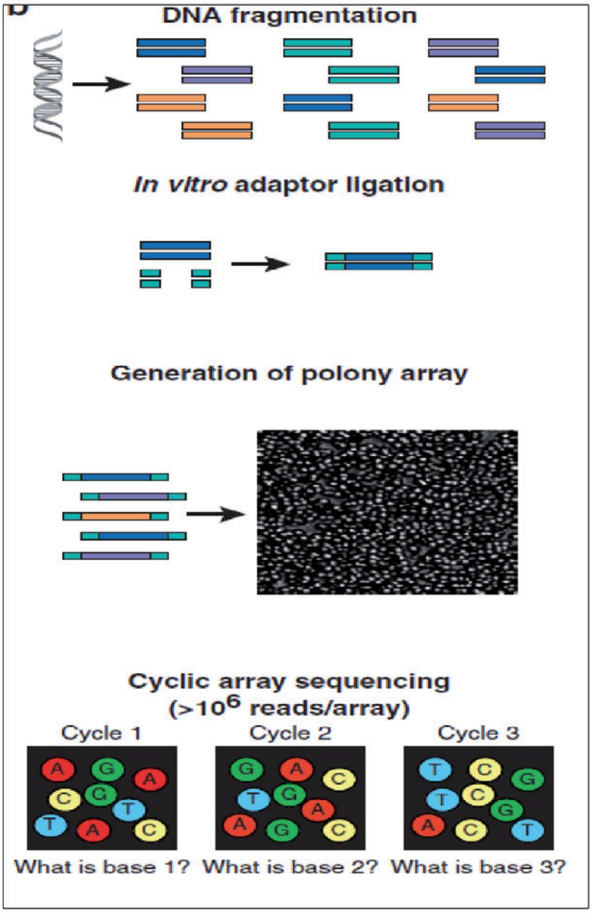


Fig. 6: Workflow of a shortgun DNA sequencing technology (Shendure and Ji, 2008).

Next Generation DNA sequencing technologies

Next generation DNA sequencing techniques are based on immobilization of DNA on to a solid support, cyclic sequencing reaction using fluidic devices and detection of molecular events by imaging. They are highly multiplexed, allowing simultaneous sequencing and analysis of millions of samples. “Second generation” is used in reference to the various implementation of cyclic array sequencing that have recently been realized in commercial products (Table 2) like GS FLX (454 Genome Sequencers, Roche Applied Science; Basel), Genome Analyzer (Solexa, Illumina, San Diego), the SOLiD platform (Applied Biosystems, Foster City, CA, USA), PacBio RS (Pacific Biosciences). They utilize three critical steps:

- **DNA sample preparation (Sequencing library):** It involves the fragmentation of the genomic DNA into short fragments. Adaptors are the linked to the randomly fragmented DNA. Generation of DNA fragments with common or universal nucleic ends is known as “Sequencing library”.

- **Immobilisation:** Short DNA fragments with adaptors are essentially required for the attachment to the solid surface, the site wherein the sequencing reaction takes place. Next generation DNA sequencers except PacBIo require the amplification of sequencing library involving in situ amplification in emulsion or in solution resulting in the generation of clusters of DNA copies.

- **Sequencing:** Sequencing is usually performed employing DNA polymerases synthesis of fluorescent nucleotides or ligation fluorescent oligonucleotides (Figure 7).

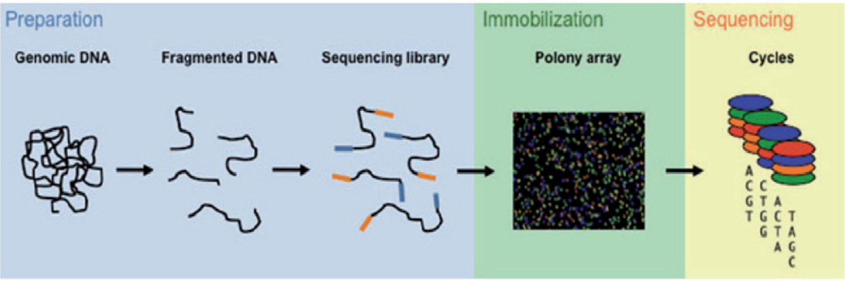


Fig. 7: High-throughput sequencing (Next Generation Sequencing) workflow (Myllykangas, 2012).

Application of Second Generation DNA sequencing Technologies

DNA sequencing is useful in biotechnology research and discovery, diagnostics, and forensics. Next generation DNA sequencing has further-

Table 2: High-throughput sequencing platforms.

	GS FLX	Genome analyser	SOLiD	PacBio
Company	454 Life Sciences, Roche	Solexa, Illumina	Applied Biosystems	Pacific Biosensors
Library construction	(Linear adaptors) Fragment, Mate paired	(Linear adaptors) Fragment, Mate pair, Paired end	(Linear adaptors) Fragment, Mate paired	Bubble adaptors
DNA support	25-35 µm bead immobilised to Pico Titer plate	Flow cell	1µm paramagnetic beads	Zero mode wave guide
Generation feature	Emulsion PCR	Bridge PCR	Emulsion PCR	Single molecule
Sequencing chemistry	Sequencing by synthesis using polymerase (Pyrosequncing) of bead on DNA templates	Sequencing by synthesis using reversible fluorescent dye terminators employing single clonal molecule array	Sequencing by ligation/hybridisation (Octamers with two base end coding) of bead bond DNA templates	Sequencing by synthesis using DNA polymerase

Sequencing reaction surface	High density well plate	8 channel flow cell	Single slide imaged in a panel	Single Molecule Real Time (SMRT) cell
Detection method	chemiluminescence detection/pyrosequencing	Fluorophore labelled reversible terminator nucleotide	Fluorophore labeled oligonucleotide probes	Phospholinked fluorophore labelled nucleotide
Read length	Upto 900 bp	35-50 bp	35-50 bp	200 bp
Throughput	100-400 MB/day	1.0-1.3GB/day	900 MB-1.1 GB/day	< 50 Mb /run (5 hrs)

-enhanced the data collection on DNA sequencing for the prediction and inference on a broad range of biological phenomena. Massive parallel sequencing platforms are widely available thereby reducing the sequencing costs drastically. This has dramatically accelerated the biological and biomedical research, enabling comprehensive analysis of genomes in real sense. Some of the application is mentioned below in Table 2.

Table 2: Application of next generation DNA sequencing (Shendure and Ji, 2008)

Category	**Example of application**
Complete genome	Comprehensive polymorphism and mutation discovery in individual human genome
Reduced representation sequencing	Large scale polymorphism discovery
Targeted genomic sequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commercial flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing, transcript annotation, discovery of transcribed SNPs or somatic mutation
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin Immunoprecipitation sequencing (ChIP-Seq)	Genome mapping of DNA-protein interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

DNA sequence editing and alignment

Initial 30-40 bases and bases inserted after 600-700 good quality read, are not reliable in DNA sequence generated in the form of electropherogram because the peaks are of worst quality and not reliable. Forward primer is absent in forward sequence and reverse primer is absent in reverse sequence. In DNA sequencing, primer is not labeled, so the first nucleotide that extends from primer is incorporated in sequencing PCR using fluorescent dideoxy nucleotide triphosphate (ddNTP). Both forward and reverse strand sequences of each individual are merged in one final sequence called 'contig' or 'consensus sequence'. Reverse strand sequences are inverted and aligned with forward strand sequence. The ambiguities located against the sequencing electropherogram are corrected accordingly. The DNA sequences can be edited with software like DNASTAR, MEGA, Sequence Navigator, BioEdit etc. and aligned using CLustalW/X, PILEUP, PROBCONS, MUSCLE, MAFFT, DIALIGN, POA and SeqMan, Sequence Navigator, BioEdit.

Nucleotide Sequence Database

The nucleotide sequence generated can be submitted in any one of the following Nucleotide Sequence Databases:

- GeneBank: It is maintained by Bethesda, USA
- EMBL (Europen Molecular Biology Laboratory): It is maintained by Cambridge, U. K.
- DDBJ (DNA Data Bank of Japan). It is being maintained by Mishima, Japan.

In conclusions, the progress in DNA sequencing technology has resulted in higher throughput DNA sequencing systems with lower cost. Emerging technologies in DNA sequencing are anticipated to be faster than the current throughput technologies. Semiconductor and Nanopore sequencing are among the emerging techniques in field of DNA sequencing. The reduction in costs of DNA sequencing by several order in magnitude will eventually be useful for individual investigators to purse project that were accessible only to major genomic centers. The massive challenge before the scientific world is how to go extract biologically or clinically meaningful understanding of the huge quantity of data generated.

References

Mahboob, F. (2010) DNA Sequencing: Maxam Gilbert Method. Biotech-Research (http://www.biotecharticles.com/Biotech-Research-Article/DNA-Sequencing Maxam-Gilbert-Method-285.html).

Mahender, S. (2014) Short course on DNA Barcoding of Aquatic Organisms: A Tool for Molecular Taxonomy. In: Quality check of DNA Sequencing Electropherogram. NBFGR, Lucknow. Pp.52-56. http://nationaldiagnostics.com/article_info.pparticles_id/20

Myllykangas, S., Buenrostro, J. and Hanlee, P. Ji. (2012) Overview of sequencing technologies In: Rodríguez-Ezpeleta et al. (Eds.), Bioinformatics for High Throughput Sequencing, 11 DOI 10.1007/978-1-4614-0782-9_2, © Springer Science+Business Media, LLC.

Sanger, F. (1988) Sequences, sequences, and sequences. Annual Review of Biochemistry 57: 1–28

Shendure, J., Mitra, R.D., Varma, C. and Church, G.M. (2004) Advanced sequencing technologies: methods and goals. Nature Reviews Genetics 5: 335–344.

Shendure, J. and Ji, H, (2008) Next generation DNA-sequencing. Nature Biotechnology 26 (10): 1135-1145.

Swerdlow, H., Wu, S.L., Harke, H. and Dovichi, N.J. (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. Journal of Chromatography A 516: 61–67.

Wikipedia: http://en.wikipedia.org/wiki/DNA_sequencing

Wu, R. (1994) Development of the primer-ex- tension approach: a key role in DNA sequencing. Trends in Biochemical Sciences 19: 429-433.

*****

Plant DNA Barcoding and Phylogenetics, Pages 171-200
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

11 Plant DNA Barcoding and Molecular Phylogeny

M.A Ali, A.K. Pandey, G. Gyulai, J. Lee and
F.M.A. Al-Hemaid

Introduction

The science of naming and classifying organisms is the original bioinformatics and a basis for all biology. There are approximately 1.7 million species identified by using morphological (i.e. Linnean) characters including 808 gymnosperm, and 90000 monocots and about 200000 dicots of angiosperms. This number may be a gross under-estimate of the true biological diversity of earth (Blaxter, 2003; Wilson, 2003). The study of taxonomy is fundamentally important in ensuring the quality of life of future human generation on the earth; yet over the past few decades, the teaching and research funding in taxonomy has declined because of its classical way of practice which lead the discipline many a times to a subject of opinion, and this ultimately gave birth of several problems and challenges, and therefore the taxonomists became an endangered race in the era of genomics (Ali et al., 2014). Now taxonomy suddenly became fashionable again due to revolutionary approaches in taxonomy called DNA barcoding (-a novel technology to provide rapid, accurate, and automated species identifications using short orthologous DNA sequences). The DNA barcoding (Hebert et al., 2003, 2004) (syn.: profiling, genotyping) based on highly conserved sequence information provide new tools for systematics (Hebert and Barrett, 2005; Hebert and Gregory, 2005) and phylogeny (Wyman et al., 2004; Leebens-Mack et al., 2005; Marshall, 2005; Jansen et al., 2006; Hansen et al., 2006). DNA barcodes consist of short sequences of DNA between 400 and 800 base pairs that can be routinely amplified by PCR (polymerase chain reaction) and sequenced of the species studied.

DNA barcoding bodies

The main DNA barcoding bodies and resources are (1) Consortium for the Barcode of Life (CBOL) http://www.barcodeoflife.org established in 2004. CBOL promotes DNA barcoding through over 200 member organizations from 50 countries, operates out of the Smithsonian Institution's National Museum of Natural History in Washington, (2) International Barcode of Life (iBOL) http://www.ibol.org launched in October 2010, iBOL represents a not-for-profit effort to involve both developing and developed countries in the global barcoding effort, establishing commitments and working groups in 25 countries. The Biodiversity Institute of Ontario is the project's scientific hub and its director, (3) The Barcode of Life Datasystems (BOLD) (http://www.boldsystems.org) is an online workbench for DNA barcoders, combines a barcode repository, analytical tools, interface for submission of sequences to GenBank, a species identification tool and connectivity for external web developers and bioinformaticians, established in 2005 by the Biodiversity Institute of Ontario. The Consortium for the Barcode of Life (CBOL) Plant Working Group (2009) recommended rbcL + matK as a core two-locus combination. However, as these loci encode conserved functional traits, it is not clear whether they provide sufficiently high species resolution. One of the challenges for plant barcoding is the ability to distinguish closely related or recently evolved species.

GenBank

GenBank (-the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences) have very important role in DNA barcoding. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis. There are several ways to search and retrieve data from GenBank. Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences). Search and align

GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently. The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information (http://www.ncbi.nlm.nih.gov/genbank/).

Molecular phylogeny

The use of DNA or protein sequences to identify organisms was proposed as a more efficient approach than traditional taxonomic practices (Blaxter et al., 2004; Tautz et al., 2003). Dobzhansky (1973) stated that nothing in biology makes sense except in the light of evolution. Phylogeny is in the midst of a renaissance, heralded by the widespread application of new analytical approaches and molecular techniques. Phylogenetic analyses provided insights into relationships at all levels of evolution. The phylogenetic trees now available at all levels of the taxonomic hierarchy for animals and plants, which play a pivotal role in comparative studies in diverse fields from ecology to molecular evolution and comparative genetics (Soltis and Soltis, 2000). The basic DNA nucleotide substitution rate was estimated to be 1.3 x 10-8 (Ma and Bennetzen, 2004) and 6.5 x 10-9 (Gaut et al., 1996) substitution per locus per year in grasses, and it was estimated to 1.5 x 10-8 in Arabidopsis (Koch et al., 2000). A chloroplast gene such as matK (maturase K) or a nuclear gene such as ITS (internal transcribed spacer) may be an effective target for barcoding in plants (Kress et al., 2005; Kress and Erickson, 2008). Kress et al. (2005) have demonstrated the effectiveness of DNA barcoding in angiosperms. Ribosomal DNA (e.g. ITS) could be used to complement of results based on plastid genes, that may provide a more sophisticated multiple component barcode for species diagnosis and delimitation (Chase et al., 2005). Sequences used for molecular barcoding are the nuclear small subunit ribosomal RNA gene (SSU, also known as 16S in prokaryotes, and 18S in most eukaryotes), the nuclear large-subunit ribosomal RNA gene (LSU, also known as 23S and 28S), the highly variable internal transcribed spacer section of the ribosomal RNA cistron (ITS, separated by the 5S ribosomal RNA gene into ITS1 and ITS2 regions), the mitochondrial cytochrome c oxidase 1 (CO1 or cox1) gene and the chloroplast ribulose bisphosphate carboxylase large subunit (rbcL) gene. Kress et al., 2005 have suggested that ITS spacer region and the plastid trnH-psbA have greater potential for species-level discrimination than any other locus, the trnH-psbA combined with rp136-rpf8, and trnL-F ranked the highest amplification success with appropriate sequence length (Kress et al., 2005).

Phylogenetic analyses

In DNA barcoding the sequences of the barcoding region are obtained from various individuals. The resulting sequence data are then used to construct a phylogenetic tree. In such a tree, similar, putatively related individuals are clustered together. The term 'DNA barcode' seems to imply that each species is characterized by a unique sequence, but there is of course considerable genetic variation within each species as well as between species. However, genetic distances between species are usually greater than those within species, so the phylogenetic tree is characterized by clusters of closely related individuals, and each cluster is assumed to represent a separate species (Dasmahapatra and Mallet, 2006).

A diverse array of molecular techniques are available for studying genetic variability, including restriction site analysis, analysis of DNA rearrangements, gene and intron loss, and the dominantly used PCR based techniques followed by DNA sequencing and cladistic analyses of the nuclear genome (nuDNA) and both organelle genomes of mitochondria (mtDNA) and chloroplast (cpDNA) (Martins and Hellwig, 2005; Mitchell and Wen, 2005). The phylogenetic analysis tools used for the phylogeny and DNA barcoding have been summarized in Table 1.

The three commonly used methods for phylogenetic analysis are MP (maximum parsimony), ML (maximum likelihood), and (BI) Bayesian inference. Of them ML (maximum likelihood) was found to be the most discriminative (Hillis et al., 1994).

The maximum parsimony algorithm (Farris, 1970; Swofford et al., 1996) searches for the minimum number of genetic events (e.g. nucleotide substitutions) to infer the shortest possible tree (i.e., the maximally parsimonious tree). Often the analysis generates multiple equally most parsimonious trees. When evolutionary rates are drastically different among the species analyzed, results from parsimony analysis can be misleading (e.g., long-branch attraction; Felsenstein, 1978). Parsimony analysis is most often performed with the computer program PAUP* 4.0 (Swofford, 2002), and MEGA (Tamura et al., 2007, 2011).

The maximum likelihood (ML) method (Felsenstein, 1985; Hillis et al. 1994, 1996) evaluates an evolutionary hypothesis in terms of the probability that the proposed model and the hypothesized history would give rise to the observed data set properly. The topology with the highest maximum probability or likelihood is then chosen. This method may have lower variance than other methods and is thus least affected by sampling error and differential rates of evolution. It can statistically evaluate different tree topologies and use all of the sequence information.

The Bayesian phylogenetic inference is model-based method and was proposed as an alternative to maximum likelihood (Rannala and Yang, 1996; Yang and Rannala, 1997). The computer program MrBayes 3.0 (Huelsenbeck and Ronquist, 2001) performs Bayesian estimation of phylogeny based on the posterior probability distribution of trees, which is approximated using a simulation technique called Markov chain Monte Carlo (or MCMC). MrBayes can combine information from different data partitions or subsets evolving under different stochastic evolutionary models. This allows the user to analyze heterogeneous data sets consisting of different data types, including morphology and nucleotides. Bayesian inference has facilitated the exploration of parameter-rich evolutionary models.

Table 1: A brief of phylogenetic analysis tools used for the phylogeny and DNA barcoding.

Phylogenetic analysis tools	**Description**
ABI to FASTA converter 1.1.2	Automatic ABI to FASTA Converter
abi2xml 1.2	Utility to convert the binary file format from an ABI PRISM TM 377 DNA Sequencer to an xml file
Align-m v2.3	Multiple alignment software
AmplifX 1.7.0	Software for seeking in a collection of primers
AnnHyb v.4.946	A tool for working with and managing nucleotide sequences in multiple formats
ArboDraw 2006	A program for building and displaying phylogenetic trees
Archaeopteryx 0.9813	A Java tool for the visualization of annotated phylogenetic trees

Artemis 15.1.10	A free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation
Assemble2 1.0.0	A graphical tool to construct and study RNA architectures
AssociationViewer 2.0	Java application used to display SNPs in a genetic context
AutoDimer 1.0	Developed to rapidly screen previously selected PCR primers
BAMBE 2.02b	Bayesian Analysis in Molecular Biology and Evolution
Baobab 3.31	An editor for large phylogenetic trees written in Java
Bayesian evolutionary analysis sampling trees (BEAST)	A Bayesian MCMC program for inferring rooted trees under the clock or relaxed-clock models. It can be used to analyse nucleotide and amino acid sequences, as well as morphological data. A suite of programs, such as Tracer and FigTree, are also provided to diagnose, summarize and visualize results. http://beast.bio.ed.ac.uk (Drummond and Rambaut, 2007)
BEAST 1.7.5	Cross-platform program for Bayesian MCMC analysis of molecular sequences
BioCocoa 2.2.2	Code for handling and manipulating biological sequences
BioEdit	A biological sequence alignment editor written for Windows 95/98/NT/2000/XP BioEdit is a fairly comprehensive sequence alignment and analysis tool. BioEdit supports a wide array of file types and offers a simple interface for local BLAST searches http://www.mbio.ncsu.edu/bioedit/bioedit.html (Hall,1999)

BioLign 4.0.6	Tool for alignment, SNP identification, and PHRAP evaluation
Biology Workbench 3.2	A web-based tool for biologists. The WorkBench allows biologists to search many popular protein and nucleic acid sequence databases
BioSeqAnalyzer 1.0 demo	A bioinformatics software tool for analyzing DNA and protein sequences
BLAST+ 2.2.28	The Basic Local Alignment Search Tool
Bosque 1.8	A graphical software to perform phylogenetic analyses
CHROMA1.0	A tool for generating annotated multiple sequence alignments
Circles 0.1.1	Program for inferring RNA secondary structure
Clann 3.2.3	Determining the optimal phylogenetic supertree
CLC Sequence Viewer 7.5	Bioinformatics analyses workbench
ClustalW	Multiple Sequence Alignment *(EBI, United Kingdom).* This provides one with a number of options for data presentation, homology matrices [BLOSUM (Henikoff), PAM (Dayhoff) or GONNET, and presentation of phylogenetic trees (Neighbor-Joining, Phylip or Distance). Other sites offering ClustalW alignment are at the Pasteur Institute, Kyoto University and chEMBLnet.org http://www.ebi.ac.uk/Tools/msa/clustalw2/
CodonCode Aligner 4.2.5	A program for sequence assembly, contig editing, and mutation detection, available for Windows and Mac OS X
COMPONENT 2.0	Program for analysing evolutionary trees

ClustalX	ClustalX is a windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analyzing the results. This program allows to create Neighbor Joining trees with bootstrapping. http://www.clustal.org/ (Thompson et al., 1997)
CodonCode Aligner 4.2.5	A program for sequence assembly, contig editing, and mutation detection, available for Windows and Mac OS X
Convertrix 1.0	A Batch DNA Sample Converter
DAMBE 5.3.48	Data Analysis and Molecular Biology and Evolution software
dbEST	dbEST is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences, or Expressed Sequence Tags, from a number of organisms
Dendroscope 3.2.8	An interactive viewer for large phylogenetic trees
DensiTree 2.1.10	A program for qualitative analysis of sets of trees
DFW 2.51 trail	A compact, easy to use DNA analysis program, ideal for small-scale sequencing projects
DNA Baser 4.7	Software for DNA sequence assembly
DNA Counter v1.0.2	Tool shows the proportions between nucleotides in a DNA sequence
DNA Dragon 1.5.6 build1	DNA Sequence Contig Assembler Software
DNA for Windows	DNA analysis program

DNA Master 5.22.1	DNA sequence editor and analysis package
DNAMAN 8.0 Demo	sequence analysis tools
DNAmend 1.02A	Cloning software
DNAPlotter 1.10	Circular and linear interactive genome visualisation
DNASIS Max trial 3.0	Bioinformatics software
DnaSP 5.10.01	DNA Sequence Polymorphism, is a software package for the analysis of nucleotide polymorphism from aligned DNA sequence data
DNATREE 1.3	A computer program that simulates the branching of an evolutionary tree
DNAux 3.0	Auxiliary DNA Software
DNAWorks 3.0	Automatic oligonucleotide design for PCR-based gene synthesis
DoubleTree 0.7	An application for comparing two trees using coupled interaction
EMBOSS 6.5.7	The European Molecular Biology Open Software Suite
e-PCR 2.3.12	Electronic PCR (e-PCR) is computational procedure that is used to identify sequence tagged sites(STSs), within DNA sequences
Exonerate 2.2	A multiple sequence alignment program
FASTA 36.3.6f	Compares a protein sequence to another protein sequence or to a protein database, or a DNA sequence to another DNA sequence or a DNA library
FASTA/BLAST SCAN 2.4	A program for processing nucleotide sequences alignment made with FASTA and BLAST alignment tools

FastME 2.0.7	A distance based phylogeny reconstruction algorithm
FastPCR 6.5.04	A free software for design PCR primers
FigTree	FigTree is designed as a graphical viewer of phylogenetic trees to display summarized and annotated trees produced by BEAST
figtree 1.4	A graphical viewer of phylogenetic trees
FinchTV 1.5	DNA sequence analysis program
Format Converter v2.2.5	This program takes as input a sequence or sequences (e.g., an alignment) in an unspecified format and converts the sequence(s) to a different user-specified format
gbench 2.7.12	NCBI Genome Workbench
GDA 1.1	Genetic Data Analysis software
Genamics Expression 1.1	A revolutionary new Windows application for DNA and protein sequence analysis
GenBank to FASTA converter 08.11.12	A freeware that can convert GenBank (gb/gbk) file format to FASTA format
Gene Construction Kit 4.0.3 Demo	Allows graphic manipulation of DNA sequences and sophisticated plasmid drawing options
Geneious	Geneious (Alexei Drummond Biomatters Ltd. Auckland, New Zealand) provides an automatically-updating library of genomic and genetic data; for organizing and visualizing data. It provides a fully integrated, visually-advanced toolset for: sequence alignment and phylogenetics; sequence analysis including BLAST; protein structure viewing, NCBI, EMBL, Pubmed auto-find etc. http://www.geneious.com/

Geneious 8.0.3	Genome & proteome research tools
GenescanView1.2	Allows to visualize genescan files (.fsa format from ABI PRISM sequencers) and to view the exact peak size
GeneStudio Pro 2.2.0.0	A modern suite of molecular biology applications for the Windows platform built on our sequence format conversion engine, SeqVerter
Genetic algorithm for rapid likelihood inference (GARLI)	A program that uses genetic algorithms to search for maximum likelihood trees. It includes the GTR + Γ model and special cases and can analyse nucleotide, amino acid and codon sequences (Zwickl, 2006)
GeneTree 1.3	An experimental program for comparing gene and species trees
Genie 3.0	Genealogy Interval Explorer
GENOME EXPLORER 1.0b	A Java front end to many useful bioinformatics tools
GenomeComp 1.3	A DNA sequence comparison tool and graphical user interface (GUI) viewer implemented in Perl/Tk.
GenomePixelizer 2003.10.1	Useful in the detection of duplication events in genomes, tracking the "footprints" of evolution, as well as displaying the genetic maps and other aspects of comparative genetics
GENtle 1.9.4	A software for DNA and amino acid editing
GEODIS 2.6	A program for the calculation of the statistics and associated P-values for the nested clade analysis (NCA) developed by Templeton and collaborators
geWorkbench 2.4.1	Genomics Workbench, a Java-based open-source platform for integrated genomics

G-InforBIO V1.90	e-Workbench for "databasing", comparative genome analysis
goldMINER 2.0.27	Software for automated gene annotation and functional classification
HIV Database	The HIV databases contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials
Hypothesis testing using phylogenies (HYPHY)	A maximum likelihood program for fitting models of molecular evolution. It implements a high-level language that the user can use to specify models and to set up likelihood ratio tests. http://www.hyphy.org (Kosakovsky et al., 2005)
ITS2 Database	The ITS2 Database presents an exhaustive dataset of internal transcribed spacer 2 sequences from NCBI GenBank accurately reannotated. Following an annotation by profile Hidden Markov Models (HMMs), the secondary structure of each sequence is predicted. The ITS2 Database also provides several tools to process ITS2 sequences, including annotation, structural prediction, motif detection and BLAST (Altschul et al., 1997) search on the combined sequence-structure information. Moreover, it integrates trimmed versions of 4SALE (Seibel et al., 2006, 2008) and ProfDistS (Wolf et al., 2008) for multiple sequence-structure alignment calculation and Neighbor Joining (Saitou and Nei, 1987) tree reconstruction. Together they form a coherent analysis pipeline from an initial set of sequences to a phylogeny based on sequence and secondary structure. http://its2.bioapps.biozentrum.uni-wuerzburg.de/
iCE 3.5	Internet Contig Explorer
IGB 8.0.0	Integrated Genome Browser

IGV 2.3.19	A visualization tool to simultaneously integrate and analyze multiple types of genomic data
ISYS v1.35	A dynamic, flexible open source platform for the integration of bioinformatics software tools and databases
JAligner 1.0	Local pairwise sequence alignment software
Jalview 2.8	A multiple alignment editor written entirely in java
jambw 1.1	The Java based Molecular Biologist's Workbench
Jevtrace2 v3.16b	A java implementation of the evolutionary trace method
jMODELTEST 2.1.3	Phylogenetic model averaging
K-Estimator 6.1v	A program to estimate the number of synonymous (Ks) and nonsynonymous substitutions (Ka) per site and the confidence intervals by Monte Carlo simulations
Laj 070222	Tool for viewing and manipulating the output from pairwise alignment programs
Lalnview 3.0	Graphical program for visualizing local alignments between two sequences
Leaphy 1.0	Likelihood estimation algorithms in phylogenetics
LocalMotif 1.0	A software tool for discovering transcription factor binding motifs
loopDloop 2.07b	A tool for drawing RNA secondary structures in molecular biology
MACAW 2.05	Multiple Alignment Construction & Analysis Workbench

MacClade	MacClade is a computer program for phylogenetic analysis written by David Maddison and Wayne Maddison. Its analytical strength is in studies of character evolution. It also provides many tools for entering and editing data and phylogenies, and for producing tree diagrams and charts. http://macclade.org/
MAFFT	MAFFT is a multiple sequence alignment program for unix-like operating systems
MAFFT 7.205	A multiple sequence alignment program
Mauve 2.3.1	Genome Alignment Software
MB6.84	DNA analysis program
MESA 1.9.23	Macroevolutionary Analysis & Simulation
mesquite 2.75	Software for evolutionary biology
ModelPie 1.01	A windows and linux interface for modeltest.
Modeltest	Modeltest is a program that uses hierarchical likelihood ratio tests (hLRT) to compare the fit of the nested GTR (General Time Reversible) family of nucleotide substitution models. Additionally, it calculates the Akaike Information Criterion estimate associated with the likelihood scores. http://darwin.uvigo.es/software/jmodeltest.html (Posada and Crandall, 1998)
Molecular evolutionary genetic analysis (MEGA)	A Windows-based program with a full graphical user interface that can be run under Mac OSX or Linux using Windows emulators. It includes distance, parsimony and likelihood methods of phylogeny reconstruction, although its strength lies in the distance methods. It incorporates the alignment program ClustalW and can retrieve data from GenBank. http://www.megasoftware.net (Tamura et al. 2011)

MrBayes	A Bayesian MCMC program for phylogenetic inference. It includes all of the models of nucleotide, amino acid and codon substitution developed for likelihood analysis. http://mrbayes.net (Huelsenbeck and Ronquist, 2001)
multalin	Multiple sequence alignment by Florence Corpet
NDE 0.5.0	NEXUS Data Editor, is a program to create and edit NEXUS format data files
Neighbor-Joining	Neighbor-Joining method is proposed for reconstructing phylogenetic trees from evolutionary distance data (Saitou and Nei, 1987)
NetPrimer	Primer Analysis Software
Network 4.611	Software for generating evolutionary trees and networks
ngKLAST 2.5	All-in-one KLAST and BLAST workstation
NimbleTree 2.6	Program for making phylogenetic trees starting from sequence data
NJplot 2.4	A tree drawing program able to draw any phylogenetic tree
NoePrimer 3.0	Unique and innovative primer design studio
NSA 3.3	Nucleotide Sequence Analyzer
Oligo 7.58 Demo	Oligo 6.71 Demo researchers in PCR and related technologies software
Oligo Calculator	On line tool to find Length, melting Temperature, %GC content and Molecular Weight of DNA sequence.
ORFprimer 1.6.4.1	Java Application for automatical primer design

PAL 1.51	A Java library for molecular evolution and phylogenetics
PAUP	David Swofford of the School of Computational Science and Information Technology, Florida State University, Tallahassee, Florida has written PAUP* (which originally meant Phylogenetic Analysis Using Parsimony). PAUP*version 4.0beta10 has been released as a provisional version by Sinauer Associates, of Sunderland, Massachusetts. It has Macintosh, PowerMac, Windows, and Unix/OpenVMS versions. PAUP* has many options and close compatibility with MacClade. It includes parsimony, distance matrix, invariants, and maximum likelihood methods and many indices and statistical tests. http://paup.csit.fsu.edu (Swofford, 2002)
PCRTiler 1.42	Automated Design of Tiled and Specific PCR Primer Pairs
Pebble v1.0	Phylogenetics, Evolutionary Biology, and Bioinformatics in a moduLar Environment
PerlPrimer v1.1.21	A free, open-source GUI application that designs primers for PCR
PHASE 2.0	A software package for phylogenetics and sequence evolution
Phred/Phrap/Consed 23.0	DNA Sequence Assembler & Finishing Tools
PHYLIP	A package of programs for inferring phylogenies. PHYLIP is the most widely-distributed phylogeny package, and competes with PAUP to be the one responsible for the largest number of published trees. http://evolution.genetics.washington.edu/phylip.html

Phylodendron 0.8d	Phylogenetic tree drawing
Phylogen 1.1	Implements some straight-forward birth-death models for simulating phylogenies
Phylogenetic analysis by maximum likelihood (PAML)	A collection of programs for estimating parameters and testing hypotheses using likelihood. It is mostly used for tests of positive selection, ancestral reconstruction and molecular clock dating. It is not appropriate for tree searches. http://abacus.gene.ucl.ac.uk/ software (Yang, 2007)
Phylogeny	Parsimony method programs/ Distance matrix method programs / Maximum likelihood method programs / Computation of distance / Manipulation and visualization of phylogenetic tree/ Other programs
Phylogeny.fr	Phylogeny.fr - is a simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences. It includes multiple alignment (MUSCLE, T-Coffee, ClustalW, ProbCons), phylogeny (PhyML, MrBayes, TNT, BioNJ), tree viewer (Drawgram, Drawtree, ATV) and utility programs (e.g. Gblocks to eliminate poorly aligned positions and divergent regions). http://www.phylogeny.fr/ (Dereeper et al., 2008)
PhyloGrapher2003.4.3	A program designed to visualize and study evolutionary relationships within families of homologous genes or proteins (elements)
Phyltools 1.32	A freeware utilities package that works with the phylogenetic inference package Phylip
PhyML	A fast program for searching for the maximum likelihood trees using nucleotide or protein sequence data. http://www.atgc-montpellier. fr/phyml/binaries.php

pknotsRG 1.3	A tool for folding RNA secondary structures
PolnTree 1.0.1.2	An application that allows to build, visualize and customize phylogenetic trees
PowerBlast 1.2.0	Blast local tools
PRAP 2.0b3	Software for Parsimony ratchet analyses with PAUP
Primer 3	Pick primers from a DNA sequence
Primer Premier 6.21 DEMO	A comprehensive primer design tool to design
Primer Prim'er 5.6.0	A PCR primer design tool that completely automates the primer design process
Primo Pro 3.4	Standard PCR, reduces primer dimer and random primering
PriorsEditor 1.0.11	A general workbench for regulatory region analysis and transcription factor binding site discovery
ProfDist 0.9.9 Beta	A tool for the construction of large phylogenetic trees based on profile distances
ProfDistS	Distance based phylogeny on sequence-structure lignments. (Wolf et al., 2008)
ProSeq 3.5	Program for sequence editing and population genetics (mol/evol) analysis
PVT	Phylogenetic Visualization tool
QAlign 2.60.80	Multiple alignment and editor software
RAxML	A fast program for searching for the maximum likelihood trees under the GTR model using nucleotide or amino acid sequences (Stamatakis, 2006)
Readseq	A tool for converting between common sequence file formats.

REAP	An integrated environment for the manipulation and phylogenetic analysis
RNA draw 1.1 b2	An integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows
RnaDv 1_0	A Design and Visualization Tool for RNA Secondary Structure
RnaFamily	A simple software tools that enables to display all secondary structures of a family of RNA molecules
RnaIIViewer 1.0.1	RNA Analysis Visualization Tool
RNApasta 1.01	A utility for collecting statistics from aligned and structurally annotated RNA sequences
RNAstructure 5.6	A Windows program for the prediction and analysis of RNA secondary structure
RnaViz 2.0.3	A user-friendly, portable, GUI program for producing publication-quality secondary structure drawings of RNA molecules
RNAz 2.1	Program for predicting structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments
Savant 2.0.5	A desktop visualization tool for genomic data
SDG Web Primer	Design of PCR or sequencing primers
SeaView 4.4.2	A graphical multiple sequence alignment editor
SeqCorator 2.03	A powerful sketchpad to decorate sequence
Seqool 3.1	A sequence analysis tool
SeqPup 0.9	A biological sequence editor and analysis program

SeqState 1.41	Primer design and sequence statistics for phylogenetic DNA data sets
seqtools v. 8.4.071	Tools for basic and advanced analyses of nucleotide and protein sequences
Sequence Scanner v1.0	The free Sequence Scanner Software enables to view, edit, print and export sequence data generated using the Applied Biosystems Genetic Analyzers
Sequences Alignments and Comparisons	Sequences consensus and Sequences comparaison / Pairwise comparisons / Multiple alignments / Alignments of structures / Alignments display / HMM (Hidden Markov Models)
SequenceViewer 1.0	A graphic tool to visualize DNA sequences
Sequencher	The Premier DNA Sequence Analysis Software for Sanger and NGS Datasets http://www.genecodes.com/
Sequencher 5.2.3 Demo	The industry standard software for DNA sequence analysis
Sequin 12.91	A stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank, EMBL, or DDBJ sequence databases
Sequlator 2013	Free multiple sequence alignment editor
SeqVISTA 1.81	A Graphical Tool for Sequence Feature Visualization and Comparison
SERIAL CLONER 2.6-1	Software for DNA cloning, sequence analysis and visualisation
SnS-Align	Structure and Sequence Alignment Software
SplitsTree 4.13.1	Application for computing evolutionary networks from molecular sequence data

SRS 6.1.3.11	Sequence Retrieval System
SSRHunter 1.3	Simple Sequence Repeat Search tool
SStructView 1.2.2	A Java applet for viewing RNA secondary structures and linking to multiple computational backends
Staden 2.0.0b9	A fully developed set of DNA sequence assembly (Gap4), editing and analysis tools
STRAP 20140609	Multiple Sequence Alignment Interactive Program
TCS 1.21	A Java computer program to estimate gene genealogies including multifurcations and/or reticulations (i.e. networks)
TGGE-STAR	To facilitate the design of PCR primers
Topali 2.5	Statistical and evolutionary analysis of multiple sequence alignments
Tree analysis using new technology (TNT)	A fast parsimony program intended for very large data sets (Goloboff et al., 2008)
TREECON 1.3b Demo	A software package developed primarily for the construction and drawing of phylogenetic trees
TreeExplorer 2.12	Display a phylogenetic tree in several different styles
TreeGraph 2.0.47-206 beta	A graphical editor for phylogenetic trees
TreeJuxtaposer 2.1	Compare phylogenetic evolutionary trees interactively and automatically
TreeMap 3.0b	An experimental program for comparing host and parasite trees
TreeMe 07/2008	A comprehensive phylogenetic tree visualization and manipulation software

tree-puzzle 5.2	A computer program to reconstruct phylogenetic trees from molecular sequence data by maximum likelihood
TreeView	TreeView provides a simple way to view the contents of a NEXUS, PHYLIP, or other format tree
T-REX 4.01	Tree and Reticulogram reconstruction
tRNAscan-SE 1.21	A program for improved detection of transfer RNA genes in genomic sequence
Utopia 1.4.5	Molecular Structure Viewer and Colour Interactive Editor for Multiple Alignments
Vector NTI Advance 11.5.2 Demo	Sequence analysis and data management software
Vienna RNA Package 2.1.6	Programs for the prediction and comparison of RNA secondary structures
vised11.exe	Visual Sequence Editor
Web Apollo 20131122	A genome annotation viewer and editor
WinBlast v.0.2.0	A Windows graphical front-end for NCBI BLAST
WinGene 2.31	An application for Win95/Win98/Win NT for analysis of nucleotide sequences
WINIPAUP 1.0	Paup Windows Interface software
Xplorer 2.4.3	Stand alone Chromatogram Editor with Mutation Detection
XRNA 1.2.0b	A Java based suite of tools for the creation, annotation and display of RNA secondary structure diagrams
YASS v1.14	DNA local alignment tool
YCDMA 3.1.1	Program is designated to manage microsatellite data

Implications of DNA barcoding

Traditional taxonomists use multiple morphological traits to delineate species. Today, such traits are increasingly being supplemented with DNA-based information. In contrast, the DNA barcoding identification system is based on what is in essence a single complex character (a portion of one gene, comprising ~650 bp from the first half of the mitochondrial cytochrome c oxidase subunit I gene sometimes called COXI or COI), and barcoding results are therefore seen as being unreliable and prone to errors in identification (Dasmahapatra and Mallet, 2006). Although the mitochondrial cytochrome oxidase subunit I (CO1) is a widely used barcode in a range of animal groups (Hebert et al., 2003), this locus is unsuitable for use in plants due to its low mutation rate (Kress et al., 2005; Cowen et al., 2006; Fazekas et al., 2008). In addition, complex evolutionary processes, such as hybridization and polyploidy are common in plants, lead the species boundaries difficult to define (Rieseberg et al., 2006; Fazekas et al., 2009).

In DNA barcoding, complete data set can be obtained from single specimens irrespective to morphological or life stage characters. The core idea of DNA barcoding is based on the fact that the highly conserved stretches of DNA, either coding or non coding regions, vary at very minor degree during the evolution within the species. The number and identity of DNA sequences that should be used for barcoding is a matter of debate (Pennisi, 2007; Ledford, 2008). Sequences suggested to be useful in DNA barcoding include cytoplasmic mitochondrial DNA (e.g. cox1) and chloroplast DNA *(*e.g. rbcL*,* trnL-F*,* matK, ndhF*,* and atpB rbcL*), and* nuclear DNA (ITS, and house keeping genes e.g. gapdh). The 'DNA barcodes' show promise in providing a practical, standardized, species-level identification tool that can be used for biodiversity assessment, life history and ecological studies, forensic analysis, and many more. Morphologically distinguishable taxa may not require barcoding; however, subspecies (ssp.), cultivars (cv.), eco- and morphotypes, mutants, species complex and clones can be diagnosed with molecular barcoding. Barcode of a specimen can be compared with sequences derived from other taxa, and in the case of dissimilarities species identity can be determined by molecular phylogenetic analyses based on MOTU, molecular operational taxonomic units (Floyd et al., 2002). DNA barcoding was particularly useful for marine organisms (Shander and Willassen, 2005), including fishes (Mason, 2003; Ward et al., 2005); soil miofauna (Blaxter et al., 2004) and freshwater meiobenthos (Markmann and Tautz, 2005); and extinct birds (Lambert et al., 2005). In the rainforests, rapid DNA-based entomological inventories were so effective (Monaghan et al., 2005; Smith et al., 2005) that tropical ecologists were the most active advocates of DNA barcoding (Janzen, 2004; Janzen et al., 2005). More pragmatically, DNA barcodes have proved to be useful in biosecurity, e.g. for surveillance of disease vectors (Besansky et al., 2003) and invasive insects (Armstrong and Ball, 2005), as well as for law enforcement and primatology (Lorenz et al., 2005). However, DNA barcoding has created some controversy in the taxonomy community (Mallet and Willmott, 2003; Lipscomb et al., 2003; Seberg et al., 2003; DeSalle et al., 2005; Lee, 2004; Ebach and Holdrege, 2005; Will et al, 2005).

References

Ali, M.A., Gyulai, G., Hidvégi, N., Kerti, B., Al-Hemaid, F.M.A., Pandey, A.K. and Lee, J. (2014) The changing epitome of species identification – DNA barcoding. Saudi Journal of Biological Sciences 21: 204–231.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipmanm D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17): 3389-3402.

Armstrong, K.F. and Ball, S.L. (2005) DNA barcodes for biosecurity: invasive species identification. Philosophical Transactions of the Royal Society B: Biological Sciences 360 (1462): 1813-1823.

Besansky, N.J., Severson, D.W. and Ferdig, M.T. (2003) DNA barcoding of parasites and invertebrate disease vectors: what you don’t know can hurt you. Trends in Parasitology 19: 545-546.

Blaxter, M. (2003) Counting angels with DNA. Nature 421: 122-124.

Blaxter, M., Elsworth, B. and Daub, J. (2004) DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. Proceedings. Biological sciences / The Royal Society 271: 189-192.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.P., Haidar, N. and Savolainen, V. (2005) Land plants and DNA barcodes: short-term and long term goals. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1889-1895.

Cowen, R.K., Paris, C.B. and Srinivasan, A. (2006) Scaling of connectivity in marine populations. Science 311(5760): 522-527.

Dasmahapatra, K.K. and Mallet, J. (2006) DNA barcodes: recent successes and future prospects. Heredity 97(4): 254-255.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M. and Gascuel, O. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Research 36: W465-469.

DeSalle, R., Egan, M.G. and Siddall, M. (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1905-1916.

Dobzhansky, T. (1973) Nothing in biology makes sense except in the light of evolution. The American Biology Teacher 35: 125-129.

Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7: 214.

Ebach, M.C. and Holdrege, C. (2005) DNA barcoding is no subsitute for taxonomy. Nature 434: 697.

Farris, J.S. (1970) Methods for computing Wagner trees. Systematic Zoology 19: 83-92.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M. and Barrett, S.C. (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3(7): e2802.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C., Newmaster, S.G., Hajibabaei, M. and Husband, B.C. (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular Ecology Resources s1: 130-139.

Felsenstein, J. (1978) Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology 27: 401-410.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783-791.

Floyd, R., Eyualem, A., Papert, A. and Blaxter, M.L. (2002) Molecular barcodes for soil nematode identification. Molecular Ecology 11: 839-850.

Gaut, B.S, Morton, B.R., McCaig, B.C. and Clegg, M.T. (1996) Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene rbcL. Proceedings of the National Academy of Sciences USA 93: 10274-10279.

Goloboff, P.A., Farris, J.S. and Nixon, K.C. (2008) TNT, a free program for phylogenetic analysis. Cladistics 24: 774-786.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95-98**.**

Hansen, A.K., Gilbert, L.E., Simpson, B.B., Downie, S.R., Cervi, A.C. and Jansen R.K. (2006) Phylogenetic relationships and chromosome evolution in tropical *Passiflora* based on cpDNA *trn*L/*trn*T intergenic spacer sequences and distribution of the chloroplast *rpo*C1 intron. Systematic Botany 31: 138-150.

Hebert, P.D.N. and Barrett., R.D.H. (2005) Reply to the comment by L. Prendini on "Identifying spiders through DNA barcodes. Canadian Journal of Zoology 83: 505–506.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003) Biological Identifications through DNA Barcodes. Proceedings Biological Sciences / The Royal Society 270 (1512): 313-321.

Hebert, P.D.N. and Gregory, T.R. (2005) The promise of DNA barcoding for taxonomy. Systematic Biology 54(5): 852-859.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S. and Francis, C.M. (2004) Identification of birds through DNA barcodes. PLoS Biol. 2 (10): e312.

Hillis, D.M., Huelsenbeck, J.P. and Swofford, D.L. (1994) Hobgoblin of phylogenetics? Nature 369: 363-364.

Hillis, D.M., Moritz, C. and Mable, B.K. (1996) Molecular Systematics, 2nd ed. Sinauer Associates, Sunderland, Massachusetts.

Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

Jansen, R.K., Kaittanis, C., Saski, C., Lee, S-B., Tompkins, J., Alverson, A.J. and Daniell, H. (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among Rosids. *BMC Evolutionary Biology* 6: 32.

Janzen, D. J., 2004. Now is the time. Philosophical Transactions of the Royal Society B: Biological Sciences 359: 731-732.

Janzen, D.H., Hajibabaei, M., Burns, J.M., Hallwachs, W., Remigio, E. and Hebert, P.D.N. (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1835-1845.

Koch, M.A., Haubold, B. and Mitchell-Olds, T. (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis* and related genera (Brassicaceaea). Molecular Biology and Evolution17: 1483-1498.

Kosakovsky, P., Frost, S.D. and Muse, S.V. (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679.

Kress, J.W. and Erickson, D.L. (2008) DNA barcodes: genes, genomics, and bioinformatics. Proceedings of the National Academy of Sciences USA 105: 2761-2762.

Kress, J.W., Wurdack, J.K., Zimmer, E.A., Weigt, A.L. and Janzen, H.D. (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences USA 102: 8369-8374.

Lambert, D.M., Baker, A., Huynen, L., Haddrath, O., Hebert, P.D.N. and Millar, C.D. (2005) Is a large-scale DNA-based inventory of ancient life possible? Journal of Heredity 96: 279-284.

Ledford, H. (2008) Botanical identities: DNA barcoding for plants comes a step closer. Nature 451: 616.

Lee, M.S.Y. (2004) The molecularization of taxonomy. Invertebrate Systematics 18: 1-6.

Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J., Fourcade, M., Chumley, T., Boore, J.L., Jansen, R.K. and dePamphilis, C.W. (2005) Identifying the basal angiosperms in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. Molecular Biology and Evolution 22: 1948-1963.

Lipscomb, D., Platnick, N. and Wheeler. Q. (2003) The intellectual content of taxonomy: a comment on DNA taxonomy. Trends in Ecology & Evolution 18(2): 65-66.

Lorenz, J.G., Jackson, W.E., Beck, J.C. and Hanner, R. (2005) The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1869-1877.

Ma, J. and Benetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences USA 101: 12404-12410.

Mallet, J. and Willmott, K. (2003) Taxonomy: renaissance or tower of Babel? Trends in Ecology & Evolution 18(2): 57-59.

Markmann, M. and Tautz, D. (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1917-1924.

Marshall, E. (2005) Will DNA bar codes breathe life into classification? Science 307: 1037.

Martins, L. and Hellwig, F.H. (2005) Systematic position of the genera *Serratula* and *Klasea* Centaureinae (Cardueae, Asteraceae) inferred from ETS and ITS sequence data and new combination in Klasea. Taxon 54(3): 632-638.

Mason, B. (2003) Marine surveys sees net gain in number of fish species. Nature 425: 889.

Mitchell, A. and Wen, J. (2005) Phylogeny of *Brassaiopsis* (Araliaceae) in Asia based on nuclear ITS and 5S- NTS DNA Sequences. Systematic Botany 30(4): 872-886.

Monaghan, M.T., Balke, M., Gregory, T.R. and Vogler, A.P. (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1925-1933.

Pennisi, E. (2007) Taxonomy. Wanted: a barcode for plants. Science 318 : 190-191.

Posada, D. and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817-818.

Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43: 304-311.

Rieseberg, L.H., Wood, T.E. and Baack, E.J. (2006) The nature of plant species. Nature 440: 524-527.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425.

Seberg, O., Humphries, C.J., Knapp, S., Stevenson, D.W., Petersen, G., Scharff, N. and Andersen, N.M. (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends in Ecology & Evolution 18: 63–65.

Seibel, P., Müller, T., Dandekar, T., Schultz, J. and Wolf, M. (2006) 4SALE - A tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformatics 7: 498.

Seibel, P., Müller, T., Dandekar, T. and Wolf, M. (2008) Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4 SALE. BMC Research Notes 1: 91.

Shander, C. and Willassen, E. (2005) What can biological barcoding do for marine biology. Marine Biology Research 1: 79-83.

Smith, M.A., Fisher, B.L. and Hebert, P.D.N. (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1825-1834.

Soltis, E.D and Soltis, P.S. (2000) Contributions of plant molecular systematics to studies of molecular evolution. Plant Molecular Biology 42: 45-75.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Swofford, D.L. (2002) PAUP: Phylogenetic Analysis using Maximum Parsimony (and other method). Version 4.0b10. Sinauer, Sunderland, Massachusetts.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996) Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics, 2nd edition. Sinauer, Sunderland, Massachusetts, pp 407-514.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596-1599.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution (10): 2731-2739.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. and Vogler, A.P. (2003) A plea for DNA taxonomy. Trends in Ecology & Evolution 18(2): 70-74.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 24: 4876-4882.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. and Hebert, P.D.N. (2005) DNA barcoding Australia’s fish species. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1847-1857.

Will, K.W., Mishler, B.D. and Wheeler, Q.D. (2005) The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 54(5): 844-851.

Wilson, E. O. (2003) The encyclopaedia of life. Trends in Ecology & Evolution 18: 77-80.

Wolf, M., Ruderisch, B., Dandekar, T., Schultz, J. and Müller, T. (2008) ProfDistS: (profile-) distance based phylogeny on sequence-structure alignments. Bioinformatics 24: 2401-2402.

Wyman, S.K., Boore, J.L. and Jansen, R.K. (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586-1591.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution 14: 717-724.

Zwickl, D. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Thesis, University of Texas at Austin.

*****

Plant DNA Barcoding and Phylogenetics, Pages 201-216
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

12 Plant DNA Barcoding Methodology: DNA Extraction - Sequencing

H.A. Khan, A.S. Alhomida, S. Alrokayan, M.S. Ola and M. Rusop

Introduction

DNA sequencing is a process of determining the precise sequential order of the four nucleotides within a DNA molecule. With the advent of fluorescence based sequencing chemistry and automated analyzers, DNA sequencing has become simpler, accurate and economical and currently a routine activity in major laboratories. The gene sequences so obtained are used for various purposes, and barcoding is one of the important applications of DNA sequencing. In this chapter, we have described methods of plant DNA extraction, PCR amplification, sequencing techniques and their application in DNA barcoding.

DNA sequencing provides information about the sequential arrangement of four nucleotide bases, adenine (A), guanine (G), cytosine(C) and thymine (T) present in a molecule of DNA being sequenced. Initial studies were developed for the identification and characterization of clinically important microorganisms that led the path to obtain DNA sequences within a few days (Hultman et al., 1989; Brytting et al., 1992). Sequencing based molecular techniques provide better resolution at intra-genus and above level, while 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 intra-species genotypic variation study (Robinson et al., 1999). This distinction is important to grasp for population studies, particularly when the diversity data are used as a basis for making decisions about conservation of plant resources. For instance, a recent study on Napier grass (*Pennisetum purpureum*) has showed that AFLP is incompatible with RAPD and morphological data hence re-entry of all accessions of Napier grass based on DNA barcoding is suggested as a means to resolve the lingering problems regarding the identity of accessions (Struwig et al., 2009). DNA barcoding is potentially of great value as morphology-based identification pose complexity and need experience to identify. The Consortium for the Barcode of Life (CBOL) plant working group recommended the 2-locus combination of rbcL (Ribulose-1, 5-bisphosphate carboxylase/oxygenase large subunit) and matK (Maturase K) as the standard for barcoding of all land plants based on the assessments of recoverability, sequence quality and levels of species discrimination (CBOL, 2009).

Molecular phylogenies in plants are traditionally based on sequence variation in the chloroplast DNA (cpDNA) (Despres et al., 2003). This approach has proved to be very powerful at the family level through the sequencing of coding regions such as rbcL (Chase et al., 1993). However, low evolutionary rate of these sequences limits the power of cpDNA for the assignment at the genus or species level (Soltis et al., 1993). As a consequence, the relationships among closely related taxa have been inferred using non-coding sequences (Gielly and Taberlet, 1996). However, the potential problems due to gene flow of cpDNA among closely related taxa, as well as the lack of phylogenetic resolution, triggered the development of new approaches based on nuclear DNA. The most common alternative corresponds to the sequencing of the internal transcribed spacer (ITS) of 18S-25S nuclear ribosomal DNA (Baldwin, 1992; Yuan et al., 1996). When both cpDNA and ITS sequencing fail to resolve phylogenies, the amplified fragment length polymorphism (AFLP) approach has the potential to solve such difficulties, particularly among closely related species, or at the intra-specific level (Koopman et al., 2001; Sun, 2001; Zhang et al., 2001). Therefore, integration of recently developed barcoding with the techniques such as RAPD, AFLP, microsatellite and SNP seems to provide better resolution.

DNA Barcoding of Plants

DNA barcoding is a technique for characterizing species of organisms using a short DNA sequence from a standard and agreed-upon position in the genome. DNA barcode sequences are very short relative to the entire genome and they can be obtained reasonably quickly and cheaply

(Kress et al., 2005). Recently, it was determined that about 35% of the plant species that constitute the standing vegetation are vulnerable to elimination because they are not represented in the seed bank of the Red Sea area (Hegazy et al., 2009). Therefore, appropriate measures for the preservation of plant species are urgently needed. Traditionally, subjective methods based on the morphological methods are difficult to apply accurately for discrimination and authentication. Particularly, in case of medicinal plants, the use of chromatographic techniques and marker compounds to standardize botanical preparations is also limited because the medicines have variable sources and chemical complexity, which is affected by growth, storage conditions and harvest times (Joshi et al., 2004; Zhang et al., 2007). Nowadays it is widely accepted that any valid plant barcode will be multi-locus, preferably existing of a conservative coding region like rbcL, in combination with a more rapidly evolving region, which is most likely non-coding (Kress et al., 2009). The success of species-level assignment of plants using Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) with individual barcodes was obtained with matK (99%), followed by trnH-psbA (95%) and then rbcL (75%). Sequence of coding region rbcL and trnL-F as a two-locus DNA barcode was recently successfully used for the identification of NW European ferns, whereas selected locus matK for barcoding did not work satisfactorily (de-Groot, 2011). Use of three-locus DNA barcode resulted in >98% correct identifications of 296 species of woody trees, shrubs and palms (Kress et al., 2009). Recently, we have reported novel barcodes of desert plants using DNA sequencing of rbcL and matK genes (Bafeel et al., 2011, 2012a,b,c,d)

Phylogenetic methods were also applied in recently conducted study of barcoding species (Roy et al., 2010) using each barcode locus taken alone and in combinations to evaluate species recovery. When all the sequences for a given locus were considered, ITS, matK and trnH-psbA were able to form species specific clade only in case of *Berberis pachyacantha*. The clades formed in the trees were mostly mixtures of several species. Therefore, establishing a local barcode data base will be valuable for a broad range of potential ecological applications; including the building of community phylogenies (Kress et al., 2009). Morphological identification is inapplicable when studying population biology. In such cases, barcoding is a very efficient and valuable technique. Already, some ecologists used barcoding approach to identify specific unknown plant sample for practical purposes (Li et al., 2009; Van-de-Wiel et al., 2009). Ongoing development of new primers, improvements on sequencing techniques and a lot of data that have been emerging on barcoding of plants (Soltis et al., 1996; Plunkett et al., 1997; Burgess et al., 2011). Recently, plant diversity below ground was determined using rbcL gene sequences as a core plant DNA barcoding marker (Kesanakurti et al., 2011). New genus was also described based on DNA sequences of the chloroplast matK pseudogene and ITS of the nuclear ribosomal DNA (Tsukaya et al., 2011). The generation of matK sequences for some plant groups has been reported problematic because this part of the chloroplast genome underwent a strong restructuring during the evolution (de-Groot et al., 2011; Duffy et al., 2009). None of the currently existing primer sets are likely suitable for all lineages of land plants (Roy et al., 2010; Li et al., 2009; Hollingsworth et al., 2009) and efforts are now focusing on the development of complex primer assays to achieve reliable amplification and sequencing of matK in land plants.

Plant DNA Extraction

DNA extraction is the first step before proceeding for DNA sequencing. Most of the plant DNA isolation methods including commercial kits require grinding of the plant material in liquid nitrogen. Any tissue immersed in liquid nitrogen instantly becomes brittle solid to facilitate crushing into powder, with an additional advantage of maintaining the tissue at low temperature. However, the grinding step in liquid nitrogen may be omitted for soft, easy-to-grind plant materials (Lin and Ritland 1995). To avoid the problems related with the preservation and use of liquid nitrogen, acid-washed sand or glass powder were used for grinding the leaves of date palm (Ouenzar et al., 1998). DNA has also been extracted using sand from many genera of rain forest plant species (Scott and Playford, 1996). The highly versatile cetyl trimethylammonium bromide (CTAB) method has been used for the extraction of DNA from various plant materials (Doyle and Doyle, 1990). There are three main contaminants associated with plant DNA that can cause considerable difficulties when conducting PCR experiments: polyphenolic compounds, polysaccharides and RNA. Inclusion of sodium chloride (NaCl) with the lysis buffer has been used for removing polysaccharides (Fang et al., 1992). Similarly, polyvinylpyrrolidone (PVP) has been recommended for removal of polyphenolic compounds (Maliyakal et al., 1992). Recently, a combination of NaCl, PVP and LiCl has been used with the CTAB method for the isolation of genomic DNA from coniferous tissues (Barzegari et al., 2010). However, the individual effects of NaCl, PVP and LiCl as well as their typical combinations have not been tested for optimal isolation of genomic DNA from plant tissues. Arif et al. (2010) examined the individual and combined effects of NaCl, PVP and LiCl in conjunction with the basic CTAB protocol for extraction of DNA from date palm leaves. They observed that grinding of date palm leaves with sterile sand and inclusion of NaCl (1.4 M) in the lysis buffer without the costly use of liquid nitrogen, PVP and LiCl, provides a DNA yield of sufficient purity, suitable for PCR amplification and subsequent use. However, for routine DNA extraction from plant tissues, commercial kits are commonly used with the aid of automated nucleic acid extraction system (Figure 1a)

Amplification of barcoding genes using PCR

A simple PCR protocol for the amplification of rbcL and matK genes is given here. A total volume of 30 µL of PCR master mixture contained the following: 15 µL of FideliTaq (USB Corporation, Cleveland, OH) or any other brand of PCR Master Mix, giving a final concentration of 200 µM of each four deoxynucletides and 1.5 mM $MgCl_2$, 1 µM (each) primer, 25-500 ng of genomic DNA of plant sample and the rest was adjusted with sterile distilled water. The primer sequences are as follows: rbcLaF (5´ATG TCA CCA CAA ACA GAG ACT AAA GC 3´); rbcLaR (5´ GTA AAA TCA AGT CCA CCR CG 3´); rbcL1F (5´ ATG TCA CCA CAA ACA GAA AC 3´); rbcL724R (5´ TCG CAT GTA CCT GCA GTA GC 3´); matK2.1F (5´ CCT ATC CAT CTG GAA ATC TTA G 3´); matK5R (5´ GTT CTA GCA CAA GAA AGT CG 3´); matK390F (5´ CGA TCT ATT CAT TCA ATA TTT C3´); matK1326R (5´ TCT AGC ACA CGA AAG TCG AAG T 3´).

PCR amplification is performed using a thermal cycler (Figure 1b) as follows: 95 °C for 1 min, followed by 35 cycles of 95 °C for 30 sec, 50 °C (matK 2.1F-matk 5R) for 30 sec and 68 °C for 1 min, followed by an elongation step at 68°C for 5 min. All the PCR-conditions are the same for all the primer-pair except the annealing temperatures which are as follows: 45^{o}C for matK 390F - matK 1326R, 51^{o}C for the rbcL aF – rbcL aR and 48^{o}C for rbcL 1F – rbcL 724R. A long (20 x 14 cm) 1% agarose gel using 1x TAE buffer containing 0.5 µg/mL ethidium bromide is used for electrophoresis of PCR products. Gel images of the amplified bands are obtained using gel documentation and imaging system (Figure 1c) while their sizes (base pairs) are determined using a marker (standard) such as a100-bp ladder (GE Healthcare). A representative view of gel image showing the amplified bands of rbcL and matK genes is given in Figure 2. The PCR products need to be purified using Qiaquick (Qiagen) or a similar PCR purification kit before proceeding for DNA sequencing.

The optimal setting of PCR conditions plays a crucial role in obtaining desired amplified products and thereby a successful barcoding, especially for plants (Jones et al., 1997). In terms of absolute discriminatory power, promising results occurred in liverworts using rbcL alone (90% species discrimination) (Hollingsworth et al., 2009). These failures appear to arise because some specimens produce poor quality

DNA, although cannot rule out the possibility of the primer mismatch, particularly for matK (Burgess et al., 2011). Some of these attributes of matK may have discouraged researchers from using matK sequences in broad studies such as overall angiosperm relationships. Another reason for infrequent use of matK at broad levels may be that taxon-specific primers are usually required (Hilu et al., 2003). In our recent study, sequencing was 100% successful (16/16 samples) for rbcL when we used both forward and reverse primers however the sequence success was only 75% (12/16 samples) for matK gene (Bafeel et al., 2010d). These findings corroborated with de-Groot et al. (2011) that rbcL amplification is unproblematic and successful for all samples. Recent study (Burgess et al., 2011) on 436 species in 269 genera of land plants of temperate region also showed that sequencing success was highest for rbcL (91.4% of the total samples examined).

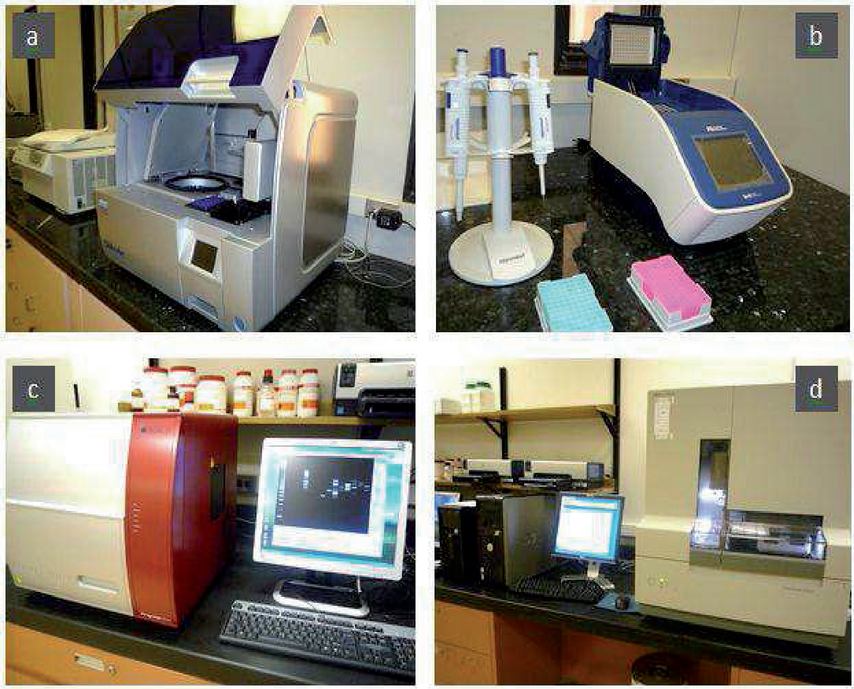


Fig. 1: Instrumentation setup for DNA sequencing (a) automated nucleic acids extraction system, (b) thermal cycler, (c) gel documentation and imaging system and (d) automated DNA sequencer.

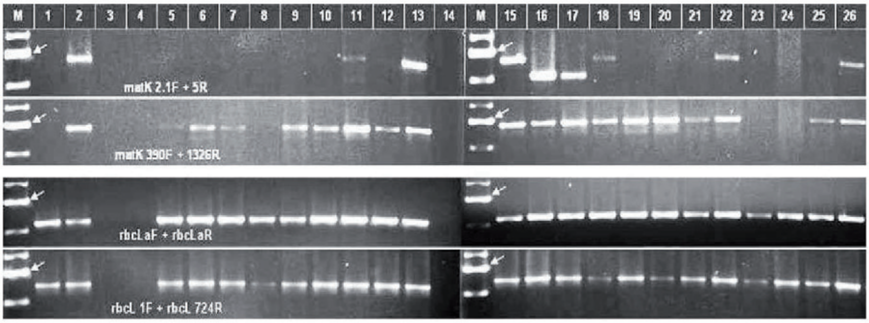


Fig. 2: Agarose get electrophoretogram showing bands of PCR amplified products of rbcL and matK universal primers. M, marker 100 bp ladder; arrows indicate the 800 bp size of the marker; Lanes 1-26 are different plant species.

Sequencing Techniques

Sequencing techniques includes any method or technology that is used to determine the order of the four nucleotides, A, G, C and T in a strand of DNA. The first DNA sequences were obtained in the early 1970s using laborious methods based on radiolabelled reagents and two-dimensional chromatography that was later replaced by polyacrylamide gel electrophoresis. Following the development of fluorescence-based sequencing methods with automated capillary electrophoresis, DNA sequencing has become much easier and faster (Pettersson et al., 2009). In the following text, we summarized both the conventional as well as next generation sequencing techniques.

Conventional Sequencing

The commonly used dye-terminator sequencing technique is the standard method for automated sequencing analysis (Olsvik et al., 1993). The dye-terminator sequencing method, supported by automated medium- to high-throughput DNA sequencers (Figure 1d) is being used for a vast majority of sequencing work. The basic technique related with dye terminator sequencing and phylogenetic analysis is illustrated in Figure 3. Dye-terminator sequencing utilizes labeling of the chain terminator dideoxynucleotides (ddNTPs), which allows sequencing in a single reaction, rather than four reactions as in the previously used labeled-primer method. In dye-terminator sequencing, the four ddNTPs chain terminators are labeled with four different fluorescent dyes, each with a different wavelength of fluorescence emission.

The main advantages of this technique are its robustness, automation and greater accuracy (>98%). However, the limitation of this technique includes dye effects due to differences in the incorporation of the dye-labeled chain terminators into the DNA fragment. Such incorporation of dye may result in unequal peak heights and shapes in the electronic DNA sequence electrophoretogram after capillary electrophoresis. Another limitation is its inability to sequence longer segments though it can reliably sequence up to approximately 900 nucleotide long DNA fragments in a single reaction. The advent of new generation sequencers with solid state chemistry has significantly overcome the inherent problems associated with previous models of sequencers.

The sequencing reaction protocol using BigDye Terminator Cycle Sequencing kit (Applied Biosystems, USA) is simple and straightforward and summarized in the following text. Prepare a 20 µl of reaction mixture by adding 10-50 ng of template DNA (amplified product), 3.2 pmol primer (forward or reverse), 4 µl Terminator Ready Reaction Mix, and 2 µl BigDye Sequencing Buffer to nuclease free distilled water in a PCR tube or microplate well. Mix well the contents and spin briefly. Place the tubes or microplate in a thermal cycler and perform an initial denaturation at 96 °C for 1 min. Repeat the following conditions for 25 cycles: 96 °C for 10 s; 50 °C for 5 s; 60 °C for 4 min and then hold at 4 °C until ready to purify. After purification (removal of unincorporated terminators) of sequencing products using either commercial kits or ethanol/EDTS/sodium acetate precipitation, proceed for injecting the purified products into the automated DNA sequencer. A representative electrophoretogram obtained from the automated genetic analyzer is shown in Figure 4.

Next Generation Sequencing

The new generation sequencing (NGS) is based on non-Sanger based sequencing technology that has been evolving on its promise of sequencing DNA at unprecedented speed, thereby enabling impressivescientific achievements and novel biological applications. Next generation platforms do not rely on Sanger chemistry (Sanger et al., 1997) as did the first generation machines used for the last 30 years (Schuster, 2008). The first of this kind of 2nd generation of sequencing technique appeared in 2005 with the landmark publication of the sequencing-by-synthesis technology developed by 454 Life Sciences (Margulies et al., 2005) based on pyrosequencing (Ronaghi et al., 2006; Nyrén, 2007). Commercial 2nd generation sequencing methods can be distinguished by the role of PCR in library preparation. There are four main platforms; all being amplification-based: (i) Roche 454 GS FLX, (ii) Illumina Genome Analyzer IIx, (iii) ABI SOLiD 3 Plus System and (iv) Polonator G.007 (Lerner and Fleischer, 2010). Common principles of these second generation sequencing techniques are illustrated in Figure 5.

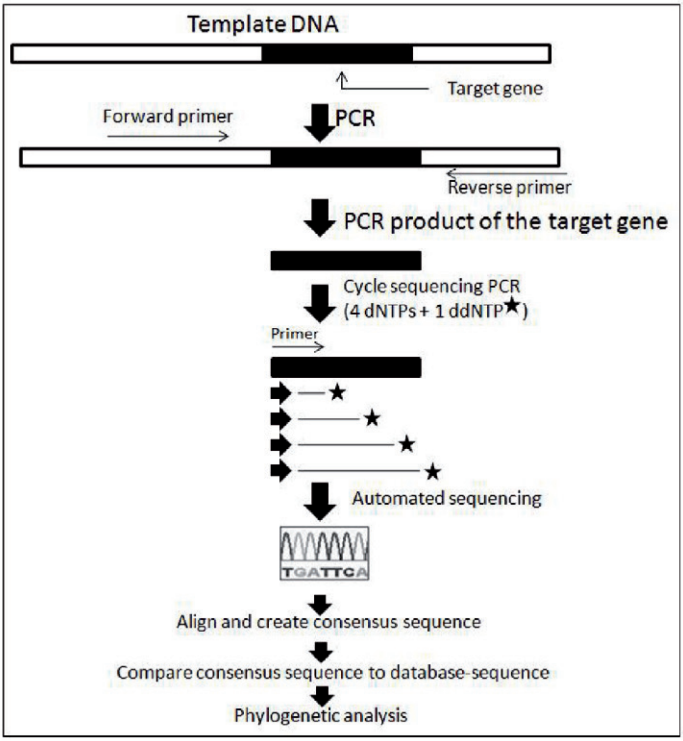


Fig. 3: Schematic presentation of DNA sequencing of a specific gene for application in phylogenetic analysis.

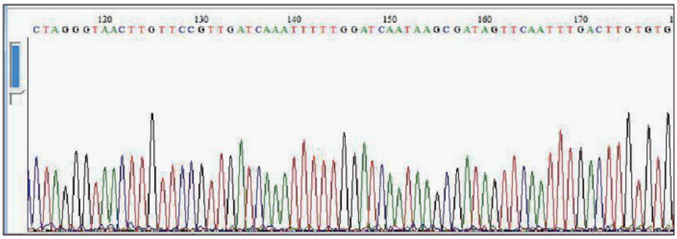


Fig. 4: Electrophoretogram obtained from the sequencing run on the ABI PRISM 3130XL automated sequencer.

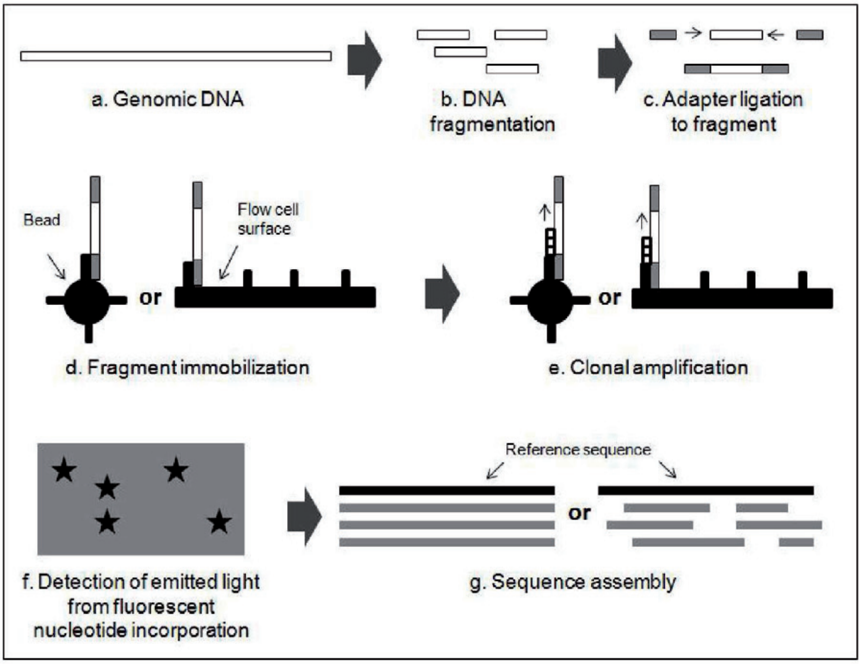

a. Genomic DNA
b. DNA fragmentation
c. Adapter ligation to fragment
Bead
or
Flow cell surface
d. Fragment immobilization
or
e. Clonal amplification
f. Detection of emitted light from fluorescent nucleotide incorporation
Reference sequence
or
g. Sequence assembly

Fig. 5: A flow diagram showing the principles of next-generation sequencing.

These techniques have made it possible to conduct robust population-genetic studies based on complete genomes rather than partial sequences of a single gene. Rapid progress in genome sequences of various plant species through next generation sequencing will further extend our understanding how genotypic variation translates into phenotypic characteristics. A comparative genomic approach is extraordinarily useful for identifying functional loci related to morphological, geographical and physiological variation, and thus next generation sequencing technology will enable us to better understand the process of plant evolution.

In conclusion, DNA sequencing is a process of determining the sequential arrangement of the four nucleotides in a target gene. Recent advancements in the solid state sequencing chemistry have made it possible to accurately sequence even the whole genomes. However, for most purposes, florescence chemistry based automated sequencers are commonly used in most laboratories.

One of the important applications of DNA sequencing is identification of species using barcoding genes. A combination of rbcL and matK genes has been recommended for barcoding of plant species. Although the currently available primers for rbcL and matK perform satisfactorily for most of the plant species, they fail to amplify these segments in certain cases. Thus, there is a need to discover more efficient and robust primers for a broader coverage of plant species.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Arif, I.A., Bakir, M.A., Khan, H.A., Ahamed, A., Al Farhan, A.H., Al Homaidan, A.A., Al Sadoon, M., Bahkali, A.H. and Shobrak, M. (2010) A simple method for DNA extraction from mature leaves of date palm: impact of sand grinding and composition of lysis buffer. International Journal of Molecular Sciences 11: 3149-3157.

Bafeel, S.O., Arif, I.A., Bakir, M.A., Al Homaidan, A.A., Al Farhan, A.H, and Khan, H.A. (2012a) DNA barcoding of arid wild plants using rbcL gene sequences. Genetics and Molecular Research 11: 1934-1941.

Bafeel, S.O., Arif, I.A., Al Homaidan, A.A., Khan, H.A., Ahamed, A. and Bakir, M.A. (2012b) Assessment of DNA Barcoding for the Identification of *Chenopodium murale* L. (Chenopodiaceae). International Journal of Biology 4: 66-74.

Bafeel, S.O., Alaklabi, A., Arif, I.A., Khan, H.A., Al Farhan, A.H., Ahamed, A., Thomas, J. and Bakir, M.A. (2012c) Ribulose-1, 5-biphosphate carboxylase (rbcL) gene sequence and random amplification of polymorphic DNA (RAPD) profile of regionally endangered tree species *Coptosperma graveolens* subsp. *arabicum* (S. Moore) Degreef. Plant Omics Journal 5: 285-290.

Bafeel, S.O., Alaklabi, A., Arif, I.A., Khan, H.A., Al Farhan, A.H., Ahamed, A., Thomas, J. and Bakir, M.A. (2012d) Molecular Characterization of regionally endangered tree species *Mimusops laurifolia* (Forssk.) Friis (Sapotaceae). International Journal of Biology 4: 29-37.

Bafeel, S.O., Arif, I.A., Bakir, M.A., Khan, H.A., Al Farhan, A.H., Al Homaidan, A.A., Ahamed, A. and Thomas J (2011) Comparative evaluation of PCR success with universal primers of maturase K (matK) and ribulose-1,5-bisphosphate carboxylase oxygenase large subunit (rbcL) for barcoding of some arid plants. Plant Omics Journal 4: 195-198.

Baldwin, B.G. (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the Compositae. Molecular Phylogenetics and Evolution 1: 3–16.

Barzegari, A., Vahed, S.Z., Atashpaz, S., Khani, S. and Omidi, Y. (2010) Rapid and simple methodology for isolation of high quality genomic DNA from coniferous tissues (*Taxus baccata*). Molecular Biology Reports 37: 833–837.

Brytting, M., Wahlberg, J., Lundeberg, J., Wahren, B., Uhlen, M. and Sundqvist, V.A. (1992) Variations in the cytomegalovirus major immediate-early gene found by direct genomic sequencing. Journal of Clinical Microbiology 30: 955–960.

Burgess, K.S., Fazekas, A.J., Kesanakurti, R.P., Graham, S.W., Husband, B.C., Newmaster, S.G., Percy, D.M., Hajibabaei, M. and Barrett, S.C.H. (2011) Discriminating plant species in a local temperate flora using the rbcL+ matK DNA barcode. Methods in Ecology and Evolution 2: 333 - 340.

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences USA 106: 12794-12797.

Chase, M.W., Soltis, D.E., Olmstead, R.G., Morgan, D., Les, D.H., Mishler, B.D., Duvall, M.R., Price, R.A., Hills, H.G., Qiu, Y.L., Kron, K.A., Rettig, J.H., Conti, E., Palmer, J.D., Manhart, J.R., Sytsma, K.J., Michaels, H.J., Kress, W.J., Karol, K.G., Clark, W.D., Hedren, M., Gaut, B.S., Jansen, R.K., Kim, K.J., Wimpee, C.F., Smith, J.F., Furnier, G.R., Strauss, S.H., Xiang, Q.Y., Plunkett, G.M., Soltis, P.S., Swensen, S.M., Williams, S.E., Gadek, P.A., Quinn, C.J., Eguiarte, L.E., Golenberg, E., Learn, G.H., Graham, S.W., Barrett, S.C.H., Dayanandan, S. and Albert, V.A. (1993) Phylogenetics of seed plants – an analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Garden 80: 528–580.

de-Groot, G.A. , During, H.J., Maas, J.W., Schneider, H., Vogel, J.C. and Erkens, R.H.J. (2011) Use of rbcL and trnL-F as a two-locus DNA barcode for identification of NW European ferns: an ecological perspective. PLoS ONE 6: e16371.

Despres, L., Gielly, L., Redoutet, B. and Taberlet, P. (2003) Using AFLP to resolve phylogenetic relationships in a morphologically diversified plant species complex when nuclear and chloroplast sequences fail to reveal variability. Molecular Phylogenetics and Evolution 27: 185–196.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of plant DNA from fresh tissue. Focus 12: 13–15.

Duffy, A.M., Kelchner, S.A. and Wolf, P.G. (2009) Conservation of selection on matK following an ancient loss of its flanking intron. Gene 438: 17–25.

Fang, G., Hammar, S. and Rebecca, R. (1992) A quick and inexpensive method for removing polysaccharides from plant genomic DNA. Biotechniques 13: 52–56.

Gielly, L. and Taberlet, P. (1996) A phylogeny of the European gentians inferred from chloroplast trnL UAA intron sequences. Botanical Journal of the Linnean Society 120: 57–75.

Hegazy, A.K., Hammouda, O., Lovettt, D.J. and Gomaa, N.H. (2009) Variations of the germinable soil seed bank along the altitudinal gradient in the northwestern Red Sea region. Acta Ecologica Sinica 2: 20-29.

Hilu, K.W., Borsch, T., Muller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, L.A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T.A.B., Rohwer, J.G., Campbell, C.S., Chatrou, L.W. (2003) Angiosperm phylogeny based on matK sequence information. American Journal of Botany 90: 1758–1776.

Hollingsworth, M.L., Clark, A.A., Forrest, L., Richardson, J., Pennington, R.T., Long, D., Cowan, R., Chase, M.W., Gaudeul, M. and Hollingsworth, P.M. (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources 9: 439–457.

Hultman, T., Stahl, S., Hornes, E. and Uhlen, M. (1989) Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. Nucleic Acids Research 17: 4937–4946.

Jones, C.J., Edwards, K.J., Castaglione, S., Winfield, M.O., Sala, F., van de Wiel, C., Bredemeijer, G., Vosman, B., Matthes, M., Daly, A., Brettschneider, R., Bettini, P., Buiatti, M., Maestri, E., Malcevschi, A., Marmiroli, N., Aert, R., Volckaert, G., Rueda, J., Linacero, R., Vazquez, A. and Karp, A. (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Molecular Breeding 3: 381-390.

Joshi, K., Chavan, P., Warude, D. and Patwardhan, B. (2004) Molecular markers in herbal drug technology. Current Science 87: 159-165.

Kesanakurti, P.R., Fazekas, A.J., Burgess, K.S., Percy, D.M., Newmaster, S.G., Graham, S.W., Barrett, S.C., Hajibabae, I. M. and Husband, B.C. (2011) Spatial patterns of plant diversity below-ground as revealed by DNA barcoding. Molecular Ecology 20: 1289–1302.

Koopman, W.J.M., Zevenbergen, M.J. and Van-den-Berg, R.G. (2001) Species relationships in *Lactuca* S.L. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. American Journal of Botany 88: 1881–1887.

Kress, W.J., Erickson, D.L., Jones, F.A., Swenson, N.G., Perez, R., Sanjur, O. and Bermingham, E. (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences USA 106: 18621–18626.

Kress, W.J., Wurdack, K., Zimmer, E.A., Weigt, L. and Janzen, D.H. (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences USA 102: 8369–8374.

Lerner, H.R.L. and Fleischer, R.C. (2010) Prospects for the use of next-generation sequencing methods in Ornithology. The Auk 127: 4–15.

Li, F.W, Tan, B.C., Buchbender, V., Moran, R.C., Rouhan, G., Wang, C.N. and Quandt, D. (2009) Identifying a mysterious aquatic fern gametophyte. Plant Systematics and Evolution 281: 77–86.

Lin, J.Z. and Ritland, K. (1995) Flower petals allow simpler and better isolation of DNA for plant RAPD analysis. Plant Molecular Biology Reporter 13: 210–213.

Maliyakal, E.J. (1992) An efficient method for isolation of RNA and DNA from plants containing polyphenolics. Nucleic Acids Research 20: 2381.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in open microfabricated high density picoliter reactors. Nature 437: 376–380.

Nyrén, P. (2007) The history of pyrosequencing. Methods in Molecular Biology 373: 1–14.

Olsvik, O., Wahlberg, J., Petterson, B., Uhlén, M., Popovic, T., Wachsmuth, I.K. and Fields, P.I. (1993) Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains. Journal of Clinical Microbiology 31: 22–25.

Ouenzar, B., Hartmann, C., Rode, A. and Benslimane, A. (1998) Date Palm DNA mini-preparation without liquid nitrogen. Plant Molecular Biology Reporter 16: 263–269.

Pettersson, E., Lundeberg, J. and Ahmadian, A. (2009) Generations of sequencing technologies. Genomics 93: 105–111.

Plunkett, G.M., Soltis, D.E. and Soltis, P.S. (1997) Clarification of the relationship between Apiaceae and Araliaceae based on matK and rbcL sequence data. American Journal of Botany 84: 365-580.

Robinson, J.P. and Harris, S.A. (1999) In Which DNA Marker for Which Purpose. Gillet, E.M. (Ed.) Institut für Forstgenetik und Forstpflanzenzüchtung, Universität Göttingen: Göttingen, Germany, pp. 1–27.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (2006) Real-time DNA sequencing using detection of pyrophosphate release. Analytical Biochemistry 242: 84–89.

Roy, S., Tyagi, A., Shukla, V., Kumar, A., Singh, U.M., Chaudhary, L.B., Datt, B., Bag, S.K., Singh, P.K., Nair, N.K., Husain, T. and Tuli, R. (2010) Universal plant DNA barcode loci may not work in complex groups: a case study with Indian *Berberis* Species. PLoS ONE 5: e13674.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences USA 74: 5463–5467.

Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. Nature Methods 5: 16–18.

Scott, K.D. and Playford, J. (1996) DNA lysis technique for PCR in rain forest plant species. Biotechniques 20: 974–978.

Soltis, D.E., Kuzoff, R.K., Conti, E., Gornall, R. and Ferguson, K. (1996) matK and rbcL gene sequence data that *Saxifraga* (Saxifragaceae) is polyphyletic. American Journal of Botany 83: 371-382.

Soltis, D.E., Morgan, D.R., Grable, A., Soltis, P.S. and Kuzoff, R. (1993) Molecular systematics of Saxifragaceae *sensu stricto*. American Journal of Botany 80: 1056–1081.

Struwig, M., Mienie, C.M.S., van den Berg, J., Mucina, L. and Buys, M.H. (2009) AFLPs are incompatible with RAPD and morphological data in *Pennisetum purpureum* (Napier grass). Biochemical Systematics and Ecology 37: 645–652.

Sun, M. (2001) Comparative analysis of phylogenetic relationships of grain amaranths and their wild relatives (*Amaranthus*; amaranthaceae) using internal transcribed spacer, amplified length polymorphism, and double-primer fluorescent intersimple sequence repeat markers. Molecular Phylogenetics and Evolution 21: 372–387.

Tsukaya, H., Nakajima, M. and Okada, H. (2011) *Kalimantanorchis*: A new genus of mycotrophic orchid from West Kalimantan, Borneo. Systematic Botany 36: 49-52.

Van-de-Wiel, C.C.M., Van-der-Schoot, J., Van-Valkenburg, J.L.C.H., Duistermaat, H. and Smulders, M.J.M. (2009) DNA barcoding discriminates the noxious invasive plant species, floating pennywort (*Hydrocotyle ranunculoides* L.f.), from non-invasive relatives. Molecular Ecology Resources 9: 1086–1091.

Yuan, Y.M., Küpfer, P. and Doyle, J. (1996) Infrageneric phylogeny of the genus *Gentiana* (Gentianaceae) inferred from nucleotide sequences of the internal transcribed spacers ITSs of nuclear ribosomal DNA. American Journal of Botany 83: 641–652.

Zhang, L.B., Comes, H.P. and Kadereit, J.W. (2001) Phylogeny and quaternary history of the European montane/alpine endemic *Soldanella* (Primulaceae) based on ITS and AFLP variation. American Journal of Botany 88: 2331–2345.

Zhang, Y.B., Shaw, P.C., Sze, C.W., Wang, Z.T. and Tong, Y. (2007) Molecular authentification of Chinese herbal materials. Journal of Food and Drug Analysis 15: 1-9.

*****

Plant DNA Barcoding and Phylogenetics, Pages 217-226
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

13 *In Silico* Approach for Phylogenetic Analysis

A. Bhattacharjee

Introduction

Phylogenetic analysis of DNA or protein sequences is a vital tool in every field of modern molecular biology today for understanding the evolutionary relationship among different taxa (Kingdoms, Phyla, Classes, Families, Genera and Species etc.). Phylogenetic analysis also plays a significant role in understanding the adaptive evolution at the molecular level (Chandrasekharan et al., 1996; Jermann et al., 1995) and provides answer to the evolutionary pattern of mutigene families (Atchley et al., 1994; Goodwin et al., 1996). This is also an important tool in understanding the mechanism of maintenance of polymorphic alleles in populations (Figueroa et al., 1988). Other equally significant applications of this tool include classification of metagenomic sequences (Brady and Salzberg, 2011), identification of genes, regulatory elements and non coding RNAs in novel genomes (Kellis et al., 2003; Pedersen et al., 2006). With a parallel growth in the experimental sequencing techniques such as Next Generation Sequencing (NGS) technologies, the amount of data generated has increased manifold which necessitates the development of more rapid and robust methods for molecular phylogenetic analysis.

A phylogenetic tree is an estimate of the relationships among taxa and their hypothetical ancestors (Nei and Kumar, 2000). It is essentially a tree containing nodes that are connected by branches. Each branch represents the evolution of genetic lineage through time and each node corresponds to the birth of a new lineage. There are two types of trees which are typically found: e.g. (1) rooted trees- which have different nodes emanating from a single node, and (2) unrooted tree- those which do not originate from one single node) (Figure 1). PAUP (Swofford, 2002), MrBayes (Huelsenbeck and Ronquist, 2001), Bayesian evolutionary analysis sampling trees, BEAST (Drummond and Rambaut, 2007), PhyML (Guindon and Gascuel, 2003), Genetic algorithm for rapid likelihood inference GARLI (Zwickl, 2006), Tree analysis using new technology TNT (Goloboff et al., 2008) and MEGA5 (Tamura et al., 2011) are some phylogenetic programs used for phylogenetic tree construction.

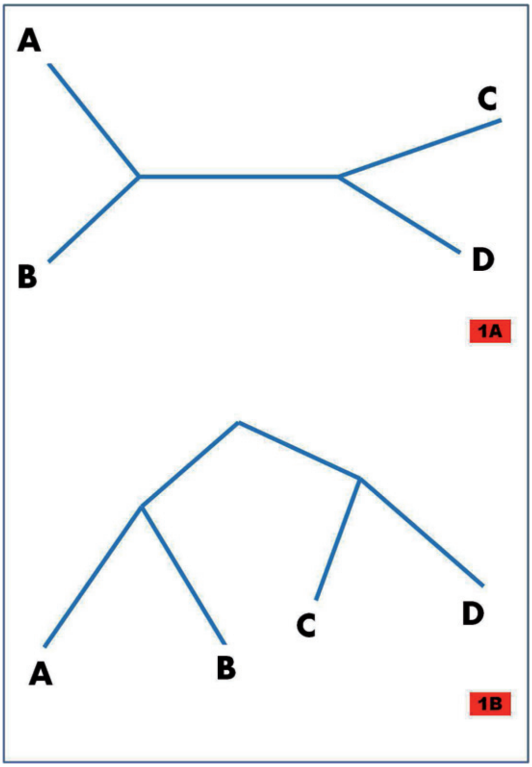


Fig.1: Schematic representation of a) unrooted and b) rooted tree.

Today most phylogenetic tree is constructed using molecular data such as DNA or protein sequences. Construction of a phylogenetic tree typically involves following four different steps:

i. Identification and retrieval of a set of homologous DNA or protein sequences
ii. Multiple sequence alignment of these sequences
iii. Estimation and validation of tree from the aligned sequences
iv. Analysis of the tree

Tree building methods

Phylogenetic tree building methods can be categorized into two types-character based method and distance based method. In character based methods such as Maximum Likelihood, Maximum Parsimony and Bayesian Inference methods etc. simultaneously compare all the sequences taking one character at a time to determine score for each tree. The tree score for Maximum Parsimony is the minimum number of substitution or mutations, posterior probability for Bayesian method and log -likelihood value for Maximum Likelihood method. In distance based methods such as unwieghted pair group method using arithmetic averages (UPGMA), Neighbor Joining and Fitch and Margoliash algorithms etc. the distance between every pair of sequences is calculated and the resulting distance matrix is used to construct the tree.

Distance based methods

Pairwise sequence distance is calculated using Markov chain model of nucleotide substitution such as HKY85 (Hasegawa et al., 1985) model, JC69 model (Jukes and Cantor, 1969), K80 model (Kimura, 1980) etc. Each model differs in its assumptions of the rate and frequency of substitution between any pair of nucleotides.

Least-squares method

It computes the minimum sum of the squared distances between the observed pairwise distances (dij) and estimated pairwise distances (*dˆij*) (Fitch and Margoliash, 1967).

$$Q = \sum_{i=1}^{s} \sum_{i=1}^{s} (\hat{d}_{ij} - d_{ij})^2 \qquad \text{..}\textbf{(1)}$$

Optimizing branch lengths (or *dˆij*) yields score *Q* for the given tree, and the tree with the smallest score is the least squares estimate of the true tree. The minimum evolution method relies on the tree length (which is the sum of branch lengths) instead of *Q* for tree selection (Rzhetsky and Nei, 1992). According to the minimum evolution criterion, shorter trees are more likely to be correct than longer trees. The most widely used distance based method is the neighbor joining method (Saitou and Nei, 1987). It uses a cluster algorithm and operates by beginning with a star tree and successively choosing a pair of taxa to join together until a fully resolved tree is obtained (Figure 2). The taxa to be joined are selected in such a way as to minimize an estimate of tree length (Gascuel and Steel, 2006).

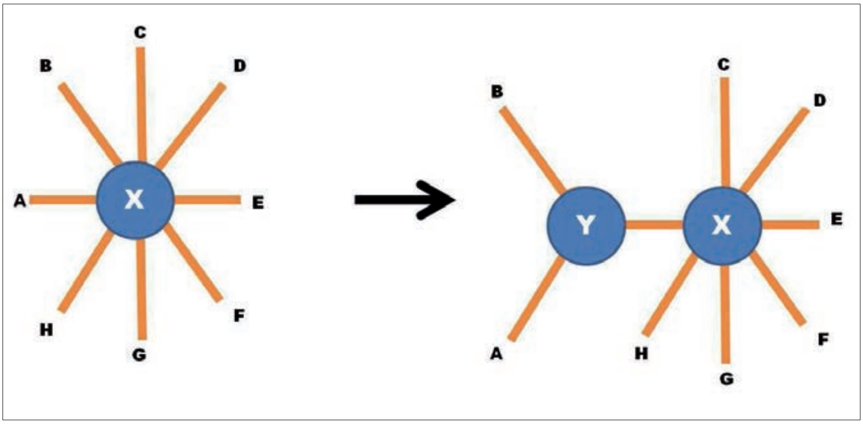


Fig. 2: The Neighbor joining algorithm. It begins from a star tree having eight peripheral nodes (A through H) and a central root X. two nodes A and B are joined together reducing the number of nodes by one at root X. This process is repeated until a fully resolved tree is obtained.

One of the major advantages of distance based method is that they are computationally much faster as it uses cluster algorithm so need not have to compare many trees under optimal criterion. For this reason distance based methods such as Neighbour Joining is useful for large data sets which have low sequence divergence. The disadvantage of this method is its sensitivity to gaps in the alignment (Bruno et al., 2000) and performs poorly for more divergent sequences because of large sampling errors due to large distances.

Maximum Parsimony

This method minimizes the number of changes on a tree by assigning character state to interior nodes on a tree. The character length is the minimum of changes required for that site whereas tree score is the sum of character lengths of all the sites. The maximum parsimony method generates a tree that minimizes the tree score. But not all sites are useful for comparison by maximum parsimony method. For example if the same nucleotide repeats in all the species at a particular site then it will be assigned a character length zero. Such sites are called constant sites. Secondly, in a singleton site, only one species has a distinct nucleotide and all others have same nucleotide is also less significant for the study as character length is always one.

The advantages of this method are its simplicity, amenable to extensive mathematical analysis and easy to understand. The major pitfalls of this method are its lack of explicit assumptions which makes it difficult to incorporate any knowledge of sequence evolution in tree construction. It also suffers from a problem known as Long Branch attraction (Felsenstein, 1978).

Maximum Likelihood

Maximum likelihood was developed by R.A. Fischer in 1920 as a statistical tool to estimate unknown parameters in a model (Aldrich, 1997). Subsequently the first algorithm for Maximum Likelihood analysis of DNA sequences was made by Felsenstein (Felsenstein, 1981). This method has been implemented in the software for producing realistic models of sequence evolution. There are two optimization steps involved in Maximum Likelihood tree estimation are: (1) optimization of branch lengths to evaluate the tree score for each candidate tree, and (2) search in the tree space for Maximum Likelihood tree.

The advantages of this method include its reliability in understanding the sequence evolution and its assumptions are explicit so that it can be evaluated and improved. The drawbacks of this method include poor statistical properties if the model is underspecified and the tree search process is computationally demanding.

Bayesian Method

Bayesian method was introduced in phylogenetic analysis in the late 1990s (Rannala and Yang, 1996; Yang and Rannala, 1997). The early methods rely on molecular clock assumption but gradually more efficient Markov chain Monte Carlo (MCMC) algorithm were developed that eliminate the clock assumption and the release of the program MrBayes (Huelsenbeck and Ronquist, 2001) made the method widely used in molecular systematists. Bayesian inference is based on Bayes's theorem, which can be expressed as

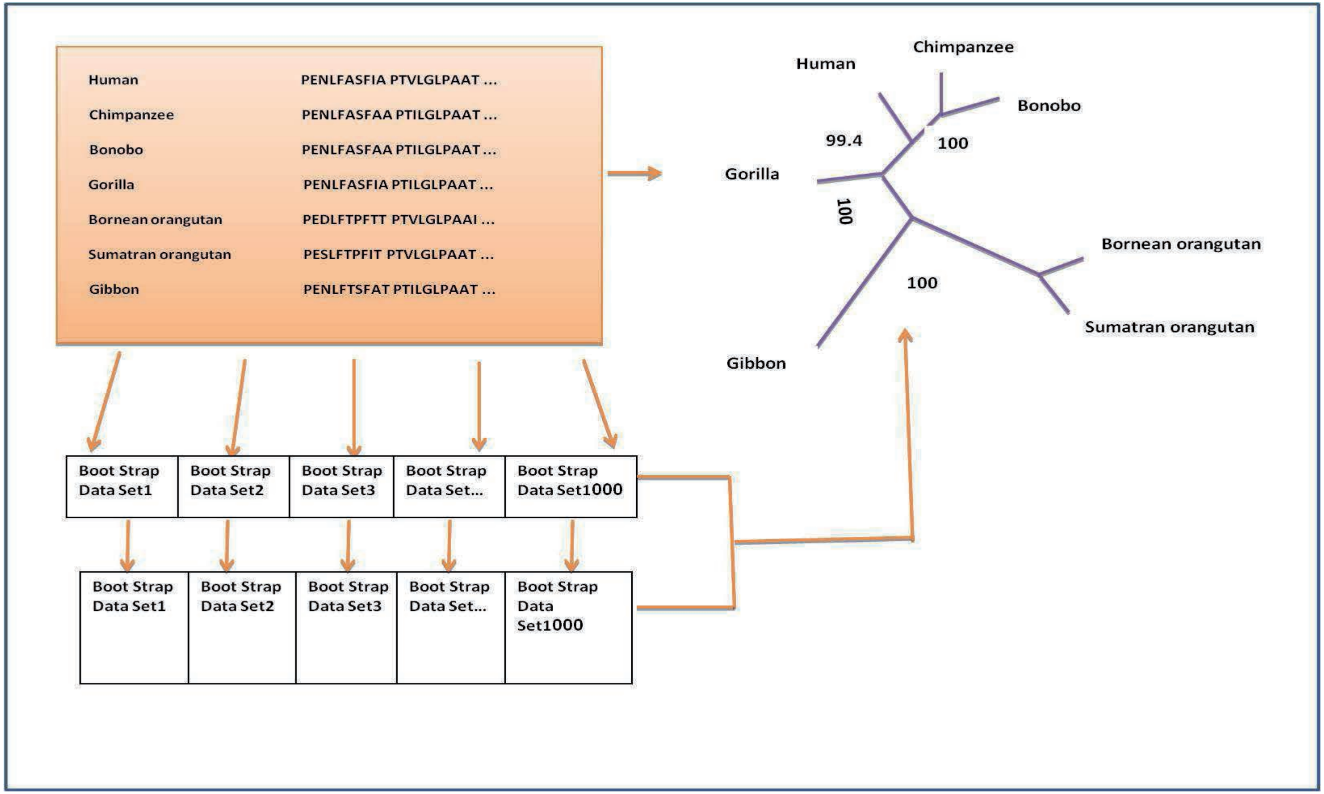
$$\mathrm{P}(T,\theta|D) = \frac{\mathrm{P}(T,\theta)\mathrm{P}(D|\,T,\theta)}{\mathrm{P}(D)} \qquad \text{......................................}\mathbf{(2)}$$

Where P(T,θ) is the prior probability for tree T and parameter θ, P(D|T,θ) is the likelihood or probability of the data given the tree and parameter, and P(T,θ|D) is the posterior probability. The denominator P(D) is a normalizing constant, as its role is to ensure that P(T,θ|D) sums over the trees and integrates over the parameters to one. The theorem states that the posterior is proportional to the prior times the likelihood, or the posterior information is the prior information plus the data information (Yang and Rannala, 2012).

The advantages of this method include use of realistic substitution models, as in maximum likelihood, prior probability allows the incorporation of information or expert knowledge and posterior probabilities for trees and clades have easy interpretations. The drawbacks of this method include Markov chain Monte Carlo (MCMC) uses heavy computation and in large data sets, MCMC convergence and mixing problems can be difficult to locate (Yang and Rannala, 2012).

Felsenstein's Bootstrap test for tree evaluation methods

One of the most widely used method to check the reliability of an expected tree is Felsenstein's (Yang, 1994) bootstrap test (Figure 3). In this test, the reliability of an inferred tree is evaluated using Efron's (Yang, 2006) bootstrap resampling technique. A set of nucleotide or amino acid residues is randomly sampled with replacement from the original data set, and this random set is used for constructing a new phylogenetic tree. This process is repeated several times, and the proportion of replications in which a given sequence cluster-



Human	PENLFASFIA PTVLGLPAAT ...
Chimpanzee	PENLFASFAA PTILGLPAAT ...
Bonobo	PENLFASFAA PTILGLPAAT ...
Gorilla	PENLFASFIA PTILGLPAAT ...
Bornean orangutan	PEDLFTPFTT PTVLGLPAAI ...
Sumatran orangutan	PESLFTPFIT PTVLGLPAAT ...
Gibbon	PENLFTSFAT PTILGLPAAT ...

Fig. 3: Bootstrap method for phylogenetic tree evaluation.

-appears is computed. If this proportion is high (>0.95) for a sequence cluster, this cluster is considered to be statistically significant (Yang and Rannala, 2012).

References

Aldrich, J. (1997) R.A. Fisher and the making of maximum likelihood 1912-1922. Statistical Science 3: 162-176.

Atchley, W.R., Fitch, W.M. and Bronner-Fraser, M. (1994) Molecular evolution of the MyoD family of transcription factors. Proceedings of the National Academy of Sciences USA 91:11522– 11526.

Brady, A. and Salzberg, S. (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nature Methods 8: 367.

Bruno, W.J., Socci, N.D. and Halpern, A.L. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Molecular Biology and Evolution 17: 189–197.

Chandrasekharan, U.M., Sanker. S., Glynias, M.J., Karnik, S.S. and Husain, A. (1996) Angiotensin II-forming activity in a reconstructed ancestral chymase. Science 271: 502–505.

Drummond, A.J. and Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7: 214.

Felsenstein, J. (1978) Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology 27: 401–410.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17: 368–376.

Figueroa, F., Gunther, E. and Klein, J. (1988) MHC polymorphism pre-dating speciation. Nature 335: 265–267.

Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. Science 155: 279–284.

Gascuel, O. and Steel, M. (2006) Neighbor-joining revealed. Molecular Biology and Evolution 23:1997–2000.

Goloboff, P.A., Farris, J.S. and Nixon, K.C. (2008) TNT, a free program for phylogenetic analysis. Cladistics 24: 774–786.

Goodwin, R.L., Baumann, H. and Berger, F.G. (1996) Patterns of divergence during evolution of α1-proteinase inhibitors in mammals. Molecular Biology and Evolution 13:346–358.

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating the human–ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22: 160–174.

Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.

Jermann, R.M., Opitz, J.G., Stackhouse, J. and Benner, S.A. (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374: 57–59

Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In: Munro, H.N. (Ed.) Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.

Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120.

Nei, M. and Kumar, S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W. and Haussler, D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Computational Biology 2: e33.

Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43: 304–311.

Rzhetsky, A. and Nei, M. (1992) A simple method for estimating and testing minimum-evolution trees. Molecular Biology and Evolution 9: 945–967.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406–425.

Swofford, D.L. (2002) PAUP: Phylogenetic Analysis using Maximum Parsimony (and other method). Version 4.0b10. Sinauer, Sunderland, Massachusetts.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 10: 2731–2739.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. Molecular Biology and Evolution 14: 717–724.

Yang, Z. (2006) Computational Molecular Evolution. Oxford University Press, UK, 2006.

Yang, Z. (1994) Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution 39: 105–111.

Yang, Z. and Rannala, B. (2012) Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13: 303–314.

Zwickl, D. J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

*****

Plant DNA Barcoding and Phylogenetics, Pages 227-240
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

14 Life History Barcoding of *Daucus carota*

O. Toldi, S. Sorvari, K. Tóth-Lencsés, L. Kovács,
A. Kerekes, Á. Mendel, Zs. Tóth and G. Gyulai

Introduction

Tools of plant biotechnology have been found to be useful for monitoring the developmental pathways of plant ontogenesis induced *in vitro*. Large-scale micropropagation technologies were developed to obtain huge numbers of true-to-type regenerants by the elimination of somaclonal variants. In practice, however, the genotype identity of the regenerated plants was often doubtful due to the genetic instability of *in vitro* cultured explants (Larkin and Scowcroft, 1981; Karp, 1991; Gyulai et al., 2003). The most frequently reported genetic variation of the somaclones comprised polyploidy (Caligari and Shohet, 1993), aneuploidy (Geier et al., 1992), haploidy (Nuti Ronchi et al., 1992ab), chromosome rearrangements (Karp, 1991), single gene mutations (Müller et al., 1990), gene copy number alterations (Landsmann and Uhrig, 1985), and cytoplasmic genetic variations. Apart from these genetically irreversible mutations, reversible genetic changes were also obtained such as growth regulator-dependent dis- and rearrangements of repetitive DNA sequences, and changes in RAPD patterns. This Chapter describes a combination of methods by which both the reversible and irreversible genetic variations could be detected by means of RAPD and Flow cytometry.

Plant material, culture conditions and life-history sampling

Surface-sterilized and *in vitro*-germinated carrot (*Daucus carota cv*. 'Nantes Duke') seeds were used to induce somatic embryogenesis. Hypocotyl sections of the seedlings were cut and inoculated in liquid MS medium (Murashige and Skoog, 1962) in the presence *vs*. absence of 1.0 mg L^{-1} 2,4-D. The 2,4-D - induced hypocotyls (code H2) were analyzed together with non-induced hypocotyls (code H1) and epicotyls (code C2) as well. The 2,4-D induced hypocotyl segments were removed after a sufficiently dense embryogenic suspension culture was formed in liquid MS medium containing 1.0 mg L^{-1} 2,4-D. Young proembryogenic suspensions (code P1) were incubated for 35 days and the old proembryogenic suspensions (code P2) were incubated for 115 days under the same conditions. To induce embryogenesis and synchronize the cultures, cells taken from a proembryogenic culture (P1) were mechanically fractionated and maintained in growth regulator-free liquid MS medium. Suspensions at the stage of heart-torpedo embryo (code E1) were obtained 15 days after the withdrawal of the 2,4-D. Elongated-torpedo stage embryos (code E2) developed 35 days after the withdrawal. For plant regeneration, the maturated embryo cultures were placed in the dark without subculturing and shaking. The morphologically normal plantlets (code IV2) were selected and potted in the greenhouse as R_0 regenerants for further analysis. The morphologically abnormal plantlets (code IV1) with undetermined polarity and reduced apical dominance were also examined. Besides these morphological alterations, the most important difference between the IV1 and IV2 plantlets were that the IV1 plantlets were unable to develop into plants under *in vitro* and/or *in vivo* conditions (Table 1).

Flow cytometry of nuclear DNA analysis

DNA extraction for RAPD analysis and nucleus isolation for Flow cytometry were carried out from a part of the given culture, while the rest was used for further stages of the tissue culture process (Table 1). Cell nuclei from leaf, epicotyl and hypocotyl tissues were isolated mechanically, as described by Dolezel et al., (1989). Approximately 10 mg tissues were chopped with a scalpel in a glass petri dishes containing 2 ml lysis buffer LB01 (pH 7.5) supplemented with 2 µg/ml DAPI. Settled suspension cells were resuspended in lysis buffer LB02 and left at room temperature for 15 min. The cell nuclei were released from the cells by syringing twice through a 25-gauge needle. Cell fragments were filtered through a 21 µm nylon filter and kept on ice until the analysis. The Flow cytometer was adjusted so that the diploid-like peak of the nuclei isolated from the control carrot plant should be on channel 100. Flow cytometric measurements were performed with a Partec CA-II computerized compact flow cytometer. The amounts of nuclear DNA were calculated on the basis of the areas under the plotted graphs analyzed by SIS Soft Imaging Software.

Table 1: Sampling of carrot (*Daucus carota cv*. 'Nantes Duke') for monitoring tissue culture systems by stage-to-stage development using RAPD and Flow Cytometry. The DNA extraction and cell nuclei isolation were carried out from a part of the cultures, and the rest was used for further tissue culture process. Liquid (LM) and agar solidified (SM) media were used *in vitro* with and without 2,4-D supplementation.

Life-history sample sets	**Treatment types**
N1 - seedling for control	*in vivo*, greenhouse conditions
N2 - seedling for control	
C2 - epicotyls excised	*in vitro*, hormone-free LM
H1 - hypocotyls (non-treated)	
H2 - hypocotyls (treated)	*in vitro*, LM + 1.0 mg/l 2,4-D
P1 - young proembryogenic suspension	
P2 - old proembryogenic suspension	
E1 - young embryo suspension	*in vitro*, hormone-free LM
E2 - old embryo suspension	
IV1 - somatic embryo derived abnormal plantlets	*in vitro*, hormone-free SM
IV2 - somatic embryo derived true-to-type plantlets (S_1-S_{12})	

Total DNA was isolated from approximately 1g of plant materials. The extracted DNA was treated with RNAase A (10 mg/ml) for 1h at 37 °C. The DNA was then extracted again with phenol/chloroform/isoamyl alcohol (25/24/1; v/v/v), following isopropanol precipitation and 70% ethanol wash. The pellet was dissolved in 50 µl sterile ddH_2. The final concentration of DNA (20 ng $µl^{-1}$) was measured by Hoefer TKO-100 fluorimeter.

Forty 10-mer RAPD primers (kits A, B from Operon Technologies, Almeda, California) were used for PCR amplifications. RAPD reactions were performed in a volume of 50 µl containing 100 ng of extracted DNA, 1.0 µM of primer, 200 µM of dATP, dTTP, dGTP, dCTP (Pharmacia), 1/10 volume 10xPCR-buffer (Boehringer Mannheim). Taq DNA polymerase (Boehringer Mannheim) concentration was 3.0 units per sample. For the RAPD cycling conditions, samples were first heated to

94°C for 3 min. before entering a 40-cycle PCR procedure of 94°C for 1 min., 36°C for 2 min. and 72°C for 3 min., followed by 72°C for 5 min. and closed at 4°C. Amplifications were carried out using a PCT-100/60 thermal cycler (MJ Research Inc). Amplified DNA fragments were separated by gel electrophoresis at 100V for 7h in a 1.5-2% agarose gel with a TBE buffer. Gels were stained with ethidium bromide and fragment patterns were photographed under UV light for further analysis. RAPD patterns were analyzed by the *restml* program of the PHYLIP 3.5c software package. The mean band number (mbn) for each sample was calculated.

The RAPD and Flow cytometry techniques have already been used to study the genotype identity of the tissue-culture-originated regenerants (Wyman et al., 1992; Jacq et al., 1993; Isabel et al., 1993; Brown et al., 1993; Mitykó et al., 1995; Taylor et al., 1995; Gyulai et al., 2000). Since among the R_0 plants a wide range of aberrations were found simultaneously, here we found that the combined application of RAPD and Flow cytometry is useful to eliminate the aberrant regenerants. RAPD barcoding was found to detect changes in the DNA sequences (Tingey and del Tufo, 1993), and ploidy level differences was also detected by flow cytometry (Dolezel et al., 1989). The whole process of somatic embryogenesis was also screened to investigate whether it is possible to detect genetic aberrations (Smith and Street, 1974; Nuti Ronchi et al., 1992a,b) either with chromosome counting and DNA-microdensitometry or detecting of changes in the number of the repetitive DNA sequences (Arnholdt-Schmitt, 1995).

RAPD of the somatic embryo originated carrot plants

The genotype identity of 35 R_0 carrot plants regenerated through somatic embryogenesis (Figure 1, lanes S1-S12) was tested in the control of 5 seed-derived carrot plantlets germinated *in vitro* without using plant hormones (samples N1 and N2). For gene modified control, T-DNA-inserted carrot plants (T1 and T2) were used, which were developed by genetic transformation. Tobacco (*Nicotiana tabacum*) plant (C3) as a taxonomically non-related species was also included. PCR amplifications were obtained with all the 40 primers, which resulted in 2-23 bands per primer (Figure 2). In 39 cases, homogenous genotype identity was detected in both the control and the somatic-embryo-originated carrot plants (control tobacco always showed a different band pattern). One primer, the OPB-11 (5-GTAGACCCGT-3), resulted in a different band pattern of carrot somatic embryos (besides the tobacco) (Figure 1-2).

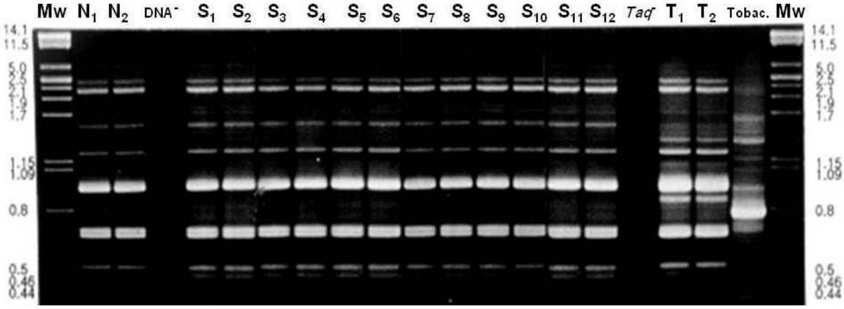


Fig. 1: Samples of RAPD (OPB-11) patterns of carrot (*Daucus carota*) tissue cultures not treated with 2,4-D. Lanes follow the developmental stages (Table 1). Lanes Mw - size marker (Kbp), lambda phage DNA digested with *Pst*I (Kbp). Lanes N_1 and N_2 – carrot samples (Table 1) of a life-history cultures. Lane DNA⁻ - no-template DNA. Lanes S_1 to S_{12} - carrot plantlets regenerated via somatic embryo genesis. Lane Taq⁻ - no-Taq DNA polymerase. Lanes T_1 and T_2 - T-DNA-inserted transgenic carrot controls. Lane Tobac. - tobacco (*Nicotiana*) control plants.

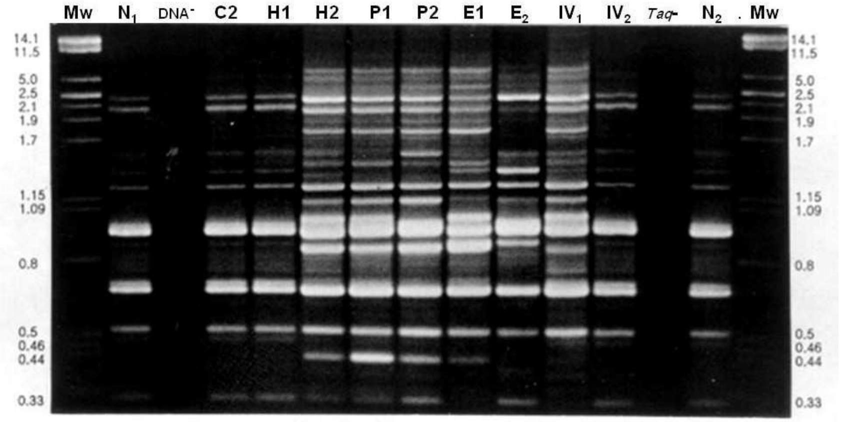


Fig. 2: Samples of RAPD (OPB-11) patterns of stage-to-stage carrot (*Daucus carota*) tissue cultures treated with 2,4-D. Lanes follow the developmental stages (Table 1). Lanes Mw - size marker (Kbp) of lambda phage DNA digested with *Pst*I (Kbp). Lane DNA⁻ - no-template DNA. Lane Taq⁻ - no-Taq DNA polymerase. Lanes N1 and N2 - individual carrot seedlings. Lanes N_1 to IV2 and N_2 – carrot samples of a life-history cultures (Table 1).

Classification of the carrot tissue culture system by RAPD and flow cytometry

Fifteen primers (OPA-01, OPA-08, OPA-12, OPA-16, OPA-17, OPA-19, OPA-20, OPB-01, OPB-02, OPB-03, OPB-05, OPB-11, OPB-12, OPB-16, OPB-18) were randomly chosen from the 40 primers showing homogenous uniformity during the genotype identity test of the 35 R$_0$ plants (Figure 1. S_1-S1$_2$). Both the epicotyl (C2) and the non-treated hypocotyl (H1) culture, as well as the morphologically normal regenerants (IV2) showed band patterns corresponding to the seedling controls (N1 and N2).

During dendrogram analysis, samples of N1, C2, H1, IV2 and N2 incubated in the absence of 2,4-D was chosen to be the root of the similarity tree as a standard (Figure 3). In this sense, the cultures incubated with 2,4-D (H2, P1 and P2) represented the top of the dendrogram. Between these two main sample groups, embryo cultures (E1 and E2) and deviant plantlets (IV1) belonging to an intermediate group were found. Based on the band pattern analysis by the restml program of PHYLIP 3.5c, the cell and tissue samples of somatic embryogenesis of the carrot could be divided into proembryogenic (H2, P1 and P2), embryogenic (E1 and E2) and plant regeneration (IV1 and IV2) phases (Figure 3). We have stated that the mean band number (mbn) of the samples N1, C2, H1, IV2 and N2 was the same 11.33 bands (Table 2).

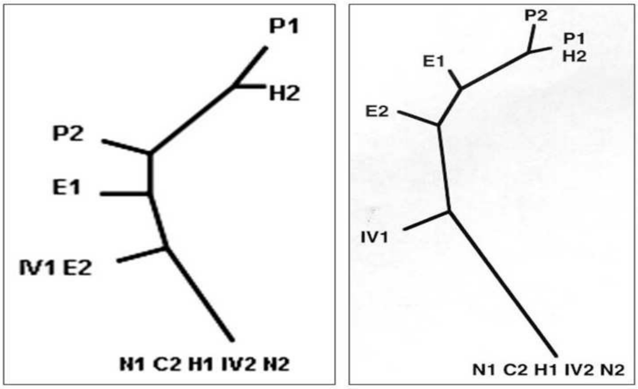


Fig 3: Classification of stage-to-stage fingerprinted carrot (*Daucus carotta*) tissue culture system (Table 1) by restml program of PHYLIP 3.5c. The distances among the samples are proportional to the RAPD pattern differences. The root of the dendrogram was fixed to the samples of N1, C2, H1, IV2 and N2 treated with (a) and without (b) 2,4-D.

Table 2: DNA content (% of analyzed samples) and mean band rearrangements during carrot plant regeneration via somatic embryogenesis. The standard deviations were less than 5% of mean value in each case (*abbreviations - Table 1).

Life-history sample set*	DNA content per sample (% of the analyzed samples)				Mean band number
	haploid-like cells	aneuploid-like cells	diploid-like cells	tetraploid-like cells	
N_1	-	-	100.0	-	11.33
N_2	-	-	100.0	-	11.33
C2	-	-	100.0	-	11.33
H1	-	-	100.0	-	11.33
H2	64.9	24.9	10.2	-	20.00
P1	58.3	30.9	10.8	-	20.00
P2	58.3	30.9	10.8	-	18.66
E1	28.1	45.4	26.5	-	17.00
E2	-	-	88.8	11.2	15.66
IV1	-	-	88.8	11.2	13.66
IV2 (S_1-S_{12})	-	-	100.0	-	11.33

Within the group of samples incubated in the presence of 2,4-D, the mbn corresponds to the H2 and the P1 (20.0 bands) and a little less in the case of P2 proembryogenic cultures (18.66 bands). In the intermediate sample group, the mbn gradually decreased (E1=17.0 bands; E2=15.66 bands; IV1=13.66 bands) as far as the 11.33 band value of *in vitro*-regenerated plantlets (IV2), which corresponded to the parameters of the individual controls (N1 and N2) and the starting tissue cultures (C2 and H1). In each case, the increased mbn characteristic of the 2,4-D-induced (H2, P1 and P2) and the intermediate (E1, E2 and IV1) samples appeared in a way that the band pattern of these samples included all the bands of those of N1, C2, H1, IV2 and N2 samples with some extra bands.

The possible anomalies due to the different ploidy levels were tested by flow cytometry. No DNA content differences were found between the *in vitro*-treated R_0 and the non-treated seed derived plants (Figure 4A). Similar to the controls (N1 and N2), the epicotyl (C2) and the non-induced hypocotyl (H1) cultures showed a diploid-like DNA content (Figure 4A, Table 2). A common characteristic of the cultures incubated with 2,4-D (H2, P1 and P2) was the coexistence of haploid-like (58.3-64.9% of the cells), aneuploid-like (24.9-30.9% of the cells) and diploid-like (10.2-10.8% of the cells) cell lines (Figure 4B-III, Table 2). After the withdrawal of 2,4-D, the haploid-like cell line was selected from the tissue culture (E1=28.1%, E2=0%, IV1=0%, IV2=0%) as well as the aneuploid-like cell line (E1=45.4%, E2=0%, IV1=0%, IV2=0%) (Figure 4B-V, Table 2). At the same time the diploid-like cell line gradually became dominant (E1=26.5%, E2=88.8%, IV1=88.8%). In the case of the E2 and the IV1 cultures 11.2% of the cells were tetraploid-like, which were later turned back to have diploid-like DNA contents in the regenerated plants (IV2). These regenerated plants showed the same DNA contents as the individual control plants (Figure 4A, Table 2).

Conclusions

The dedifferentiated cell cultures of carrot were found to be mixoploid with co-existence of cell lines with different ploidy levels. It was detected by chromosome counting and DNA-microdensitometry in callus, cell suspension and embryo cultures (Smith and Street, 1974; Nuti Ronchi et al., 1992a,b). The genetic basis of this phenomenon is based on the irregular cell-divisions (Nuti Ronchi et al., 1992a,b), and the reversible changes in the proportion of the repetitive DNA sequences (Arnholdt-Schmitt, 1995). In this latter case there was an important observation that, despite the chromosome number of a cultured cells are diploid, the DNA content can be reduced to one third volume at the proembryogenic cultures of carrot due to the effect of 2,4-D (Fujimura and Komamine, 1982) used in tissue culture. This result indicates that the 2n chromosome number does not mean automatically that these samples have also the same DNA contents. Therefore, we have used new terms (haploid-like, aneuploid-like, diploid-like, tetraploid-like DNA contents) for describing the DNA contents of our samples (Table 2). In spite of the above mentioned data, the DNA contents of the regenerated carrots showed a high uniformity and tru-to-type identity (Figure 4A, Table 2), because only the diploid-like cell lines carry *in vitro* regeneration ability (Smith and Street, 1974).

Tissue cultures were monitored by RAPD in black spruce (Isabel et al., 1993), *Triticum tauschii* callus and suspension cultures (Brown et al., 1993) and sugarcane protoplast derived callus (Taylor et al., 1995). Our examinations were focused on the whole life-history of the cultures, which approach was based on the assumption that if the RAPD patterns of the samples were identical (belonging to the same genotype: Figure 1), than the culture-specific RAPD pattern should also be the same (lanes N1 and N2 to lane C2, H1 and IV2) (Figure 2). The mutagenic effect of the growth regulator 2,4-D, and the mixoploidy caused by 2,4-D dependent abnormal cell divisions are well studied (Nuti Ronchi et al., 1992a,b). In the present case, the different RAPD patterns in the proembryogenic (H2, P1 and P2) and embryo suspensions (E1 E2), as well as in the morphologically deviant regenerants (IV1), were caused by the 2,4-D treatment, since other factors of the *in vitro* culture had no effect on these parameters (lanes N1 and N2 to lane C2 and H1) (Figure 2; Table 1).

On the basis of the present studies, a hypothesis was developed about the reversible genetic changes during carrot tissue culture detected by RAPD and flow cytometry. The cell divisions initiated by 2,4-D resulted in mutant cells with different DNA contents consisting of haploid-like, aneuploid-like, diploid-like and tetraploid-like cells. The culture-level RAPD barcoding showed that the 100% diploid-like starting genotype became a mixture of different mutant and DNA containing deviant, de novo genotypes, due to the effect of 2,4-D. The band patterns of these deviant genotypes contained the bands characteristic of the starting genotype and the additional extra bands of the de novo genotypes, in each case. The omitting 2,4-D led to the advantage of the de novo genotype, which resulted in the rearrangement with the original DNA contents and RAPD pattern. Our hypothesis is only one of the possible explanations. Therefore, further verification is necessary for widening the methodological bases of molecular-level analysis of tissue culture systems. Beside the culture-level RAPD and flow cytometry applied, monitoring of methyl-cytosine vs. cytosine ratio (LoSchiavo et al., 1989) as well as the simultaneous measurement of histone H1 kinase concentration and activity may provide new insight to this phenomena.

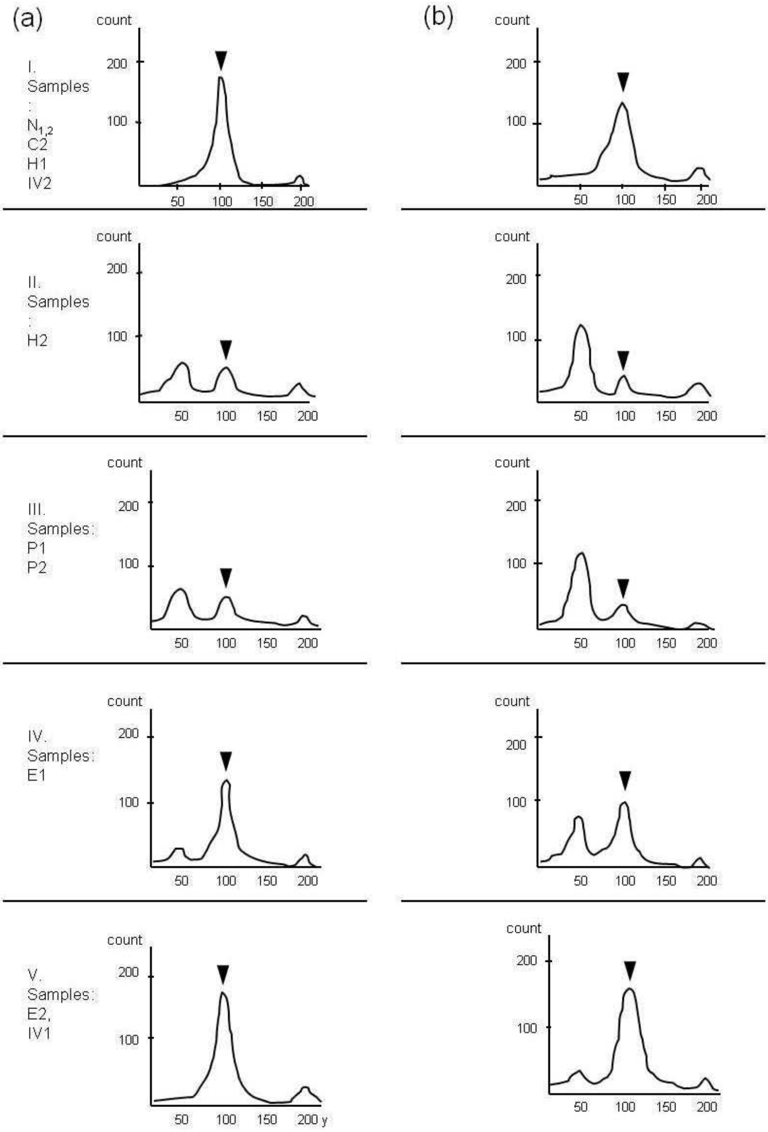


Fig. 4: Flow cytometry plots of the carrot (*Daucus carota*) tissue culture (Table 1) groups (I.-V.) treated with (a) and without 2,4-D. The peak of diploid (2n) DNA volume was adjusted to the channel 100.

In conclusion, we combined two genotype identity checking methods, which can be suitable not only to distinguish the individuals of different genotypes, but also to compare tissue culture systems and methods due to its culture-level application.

References

Arnholdt-Schmitt, B. (1995) Physiological aspects of genome variability in tissue culture. II. Growth phase-dependent quantitative variability of repetitive BstNI fragments of primary cultures of *Daucus carota* L. Theoretical and Applied Genetics 91: 816-823.

Brown, P.T.H., Lange, F.D., Kranz, E. and Lörz, H. (1993) Analysis of single protoplasts and regenerated plants by PCR and RAPD technology. Molecular & General Genetics 237: 311-317.

Caligari, P.D.S. and Shohet, S. (1993) Variability in somatic embryos. In: K. Redenbaugh (Ed.), Synseeds, CRC Press Inc., Boca Raton, Florida. pp. 164-173.

Dolezel, J., Binarova, P. and Lucretti, S. (1989) Analysis of nuclear DNA content in plant cells by flow cytometry, Biologia Plantarum 31/2: 113-120.

Fujimura, T. and Komamine, A. (1982) Molecular aspects of somatic embryogenesis in a synchronous system. In: Fujiwara, A. (ed.). Plant tissue culture 1982: Proceedings, 5th International Congress of Plant Tissue and Cell Culture held at Tokyo and Lake Yamanake, Japan, July 11-16, 1982, pp. 105-106.

Geier, T., Beck, A. and Preil, W. (1992) High uniformity of plants regenerated from cytogenetically variable embryogenic suspension cultures of poinsettia (*Euphorbia pulcherrima* Willd. ex Klotzsch), Plant Cell Report 11: 150-154.

Gyulai, G., Gémesné, J.A., Sági, Zs., Venczel, G., Pintér, P., Kristóf, Z., Törjék, O., Heszky, L., Bottka, S., Kiss, J. and Zatykó, L. (2000) Doubled haploid development and PCR-analysis of F_1 hybrid derived DH-R_2 paprika (*Capsicum annuum L.*) lines. Journal of Plant Physiology 156:168-174.

Gyulai, G., Mester, Z., Kiss, J., Szemán, L., Heszky, L. and Idnurm, A. (2003) Somaclone breeding of reed canarygrass (*Phalaris arundinacea* L). Grass and Forage Science 58: 210-215.

Isabel, N., Tremblay, L., Michaud, M., Tremblay, F.M. and Bousquet, J. (1993) RAPDs as an aid to evaluate the genetic integrity of somatic embryogenesis-derived populations of *Picea mariana* (Mill.) B.S.P. Theoretical and Applied Genetics 86: 81-87.

Jacq, B., Tetu, T., Sangwan, R.S., De Laat, A. and Sangwan-Norreel, B.S. (1993) Efficient production of uniform plants from cotyledon explants of sugarbeet (*Beta vulgaris* L.). Plant Breeding 110: 185-191.

Karp, A. (1991) On the current understanding of somaclonal variation, In: Miflin, B.J. (Ed.), Oxford Surveys of Plant Molecular and Cell Biology, vol 7, University Press, Oxford, 1991, pp. 1-58.

Landsmann, J. and Uhrig, H. (1985) Somaclonal variation in *Solanum tuberosum* detected at the molecular level, Theoretical and Applied Genetics 71: 500-506.

Larkin, P.J. and Scowcroft, W.R. (1981) Somaclonal variation - a novel source of variability from cell cultures for plant improvement, Theoretical and Applied Genetics 60: 197-214.

LoSchiavo, F., Pitto, L., Giuliano, G., Torti, G., Nuti-Ronchi, V., Marazziti, D., Vergara, R., Orselli, S. and Terzi, M. (1989) DNA methylation of embryogenic carrot cell cultures and its variations as caused by mutation, differentiation, hormones and hypomethylating drugs. Theoretical and Applied Genetics 77: 1989, 325-331.

Mitykó, J., Andrásfalvy, A., Csilléry, G. and Fári, M. (1995) Anther-culture response in different genotypes and F1 hybrids of pepper (*Capsicum annuum* L.). Plant Breeding 114:78-80.

Murashige, T. and Skoog, F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. Physiologia Plantarum 15: 473-497.

Müller, E., Brown, P.T.H. and Lörz, H. (1990) DNA variation in tissue-culture-derived rice plants, Theoretical and Applied Genetics 80: 673-679.

Nuti Ronchi, V.N., Giorgetti, L., Tonelli, M. and Martini, G. (1992a) Ploidy reduction and genome segregation in cultured carrot cell lines, I. Prophase chromosome reduction. Plant Cell, Tissue and Organ Culture 30: 107-114.

Nuti Ronchi, V.N., Giorgetti, L., Tonelli, M. and Martini, G. (1992b) Ploidy reduction and genome segregation in cultured carrot cell lines. II. Somatic meiosis. Plant Cell, Tissue and Organ Culture 30: 115-119.

Smith, S.M. and Street, H.E. (1974) The decline of embryogenic potential as callus and suspension cultures of carrot (*Daucus carota* L.) are serially subcultured. Annals of Botany 38: 223-241.

Taylor, P.W.J., Fraser, T.A., Ko, H.L. and Henry, R.J. (1995) RAPD analysis of sugarcane during tissue culture. in: Terzi, M. et al. (Eds.), Current Issues in Plant Molecular and Cellular Biology, Kluwer Academic Publishers, Dordrecht, pp. 241-246.

Tingey, S.V. and del Tufo, J.P. (1993) Genetic analysis with random amplified polymorphic DNA markers. Plant Physiology 101: 349-352.

Wyman, J., Brassard, N., Flipo, D. and Laliberté, S. (1992) Ploidy level stability of callus tissue, axillary and adventitious shoots of *Larix* X *eurolepis* Henry regenerated *in vitro.* Plant Science 85: 189-196.

*****

Plant DNA Barcoding and Phylogenetics, Pages 241-254
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

15 Cloning and Microsatellite Barcoding of Black locust

G. Gyulai, T. Demku and L Waters Jr.

Introduction

Black locust, *Robinia pseudoacacia* (family Fabaceae) is a nitrogen fixing, drought tolerant, hard wood, deciduous and multipurpose honey tree, with extremely hard and rot resistant wood, possess 2n= 20 chromosome numbers (Isely and Peabody, 1984). It is native to the southeastern United States; however, *Robinia* fossils were found in Miocene flora (Middle Badenian; 14.3 to 3.8 million years BP) (Böhme et al*.,* 2007) prior its extinction from Europe.

Fossilized *Robinia*-like trees (*Robinia zirkelii*) *(Platen)* (Müller-Stoll and Mädel), (syn.: *Cercidoxylon / Robinoxylon z.*; *Robinia / Robinoxylon brewerii*; and *Paleo-Robinoxylon z.*) were found in several sites in North America from Late Eocene (Chadronian; 38 – 33.9 million years BP) Nebraska, USA (Wheeler and Landon, 1992). Petrified woods of *Robinia alexanderi* (Webber) and *R. breweri* (Prakash, Barghoorn and Scott) were also identified, but all of these samples were considered today to be *R. pseudoacacia* (Lawrence et al., 1977), which indicates the genome stability of *Robinia* without morphological changes during the last ten millions years of evolution.

Old plant varieties, heirlooms and rare species are threatened by extinction due to reasons of changes in environmental conditions, falling below the minimum viable population and over cultivation. Conservation genetics provides effective tools of micropropagation to produce a large number of identical clones. As the clones develop from somatic meristems or organs the genome (DNA) remains identical in each clone (Bhojwani and Razdan, 1996).

Natural cloning shows the importance and evolutionary success of clonal propagation. The 'Trembling Giant' is a clonal colony of a single male quaking aspen (*Populus tremuloides*), which was determined to be a single living organism with an estimated 80,000 years old age, an estimated total weight of 6 million kg, and an assumed one massive underground root system (Utah, USA). Natural clonal propagation of monocot see grasses (e.g. *Posidonia oceanica* and *Cymodocea nodosa*) also shows the ability to span life time to 100.000 years (Sandoval-Gil et al., 2014).

The small genus *Robinia* comprises a limited numbers of species and hybrids, such as *R. boyntonii, R. elliottii, R. Hartwigii, R. hispida* (bristly locust), *R. kelseyi, R. luxurians, R. nana, R. neomexicana* (New Mexican locust), *R. pseudoacacia* (black locust, false acacia), *R. viscosa* (clammy locust), *R. zirkelii;* and the hybrids: *Robinia × ambigua (R. pseudoacacia × R. viscosa)* (Idaho locust), *Robinia × holdtii (R. neomexicana × R. pseudoacacia, Robinia × longiloba (R. hispida × R. viscosa)* and *Robinia × margarettiae (R. hispida × R. pseudoacacia)*.

The oldest black locust tree of Europe was planted in 1602 by Jean Robin (1550-1629) who introduced it to Europe (see the name *Robinia*, given by Linnaeus), and it is still growing in Paris (Jardin Royal des Plantes Medicinales, France) (Figure 1) and probably transplanted in 1635. The second oldest black locust in Europe was planted in 1662, and grows in Doorwerth, Hollandia (Figure 1). The oldest Hungarian black locust, which may be the third oldest in Europe, was planted in 1710 by Count Szapáry at city Bábolna, and still grows in a good condition (Figure 1). Kew's 'Old Lions' (London, UK) was planted in 1762 (Figure 1). After these pioneering periods, black locust quickly widespread in Europe. In 2012 the spreading area exceeded 464.000 ha in Hungary, which is 24.1% of the total forests (Keresztesi, 1983; Rédei and Veperdi, 2009). The industries use almost every part of the tree; timbers have a high energy values; however the most important profit is the acacia honey.

By leaf type, black locust has at least four pinnate type, the *monophylla (*syn.: *unifolia)*-, *triphylla*, *oligophylla* (*regular*) and the *polyphylla* (*microphylla* up to 25 leaflets) types (Dini-P and Aravanopoulos, 2008). These leaf type varieties showed different hybrid patterns in crossings (Dini-P and Aravanopoulos, 2008). Ancient *'Bábolna-1710*' black locust belongs to the *oligophylla* (*regular*) leaf type (Figure 1), and with shorter, 2 to 3 seeded pods, compared to that of current longer 4-to-8 seeded pods (Gyulai et al., 2011).

Further Legume trees were introduced to Europe such as *Robinia hispida* (in 1743), and *Robinia viscosa* (in 1797) (Földes, 1903; Peabody, 1982). *Gleditsia triacanthos* (hone locust) was introduced to Europe probably at the early 1600s (Putod, 1982; Santamour and McArdle, 1983).

Here we present the use of micropropagation for *in vitro* cloning and microsatellite analysis of the oldest Hungarian black locust tree '*Bábolna-1710*' to maintain its gene pool.

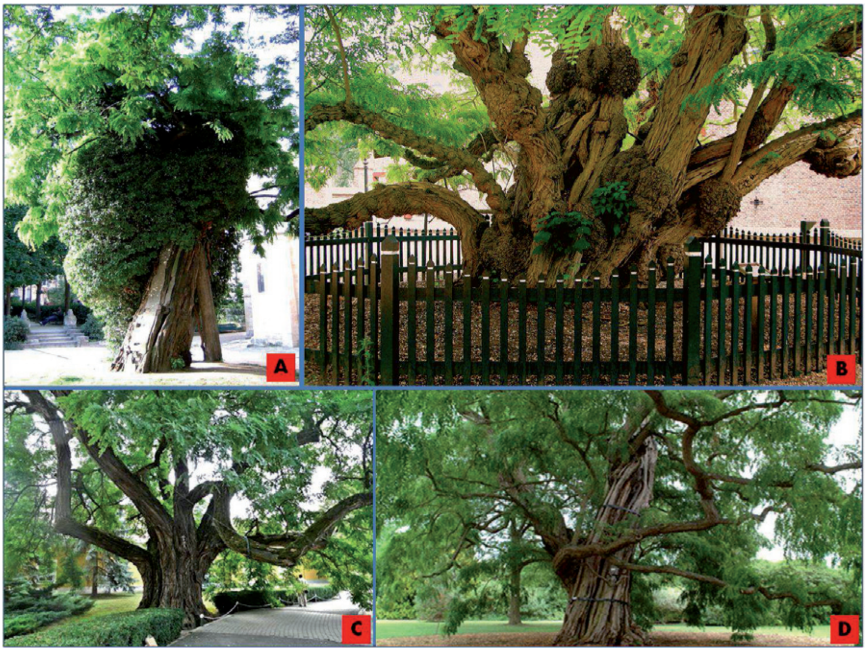


Fig. 1: Samples of old European black locust (*Robinia pseudoacacia*) trees (a) the Royal Herbal Garden Paris, France, planted in 1602, (B) Doorwerth, Hollandia, planted in 1678, (C) Bábolna, Hungary, planted in 1710, (D) Kew's '*Old Lions*', London, UK, planted in 1762. http://www.monumentaltrees.com/en/trees/blacklocust/records/

Cloning

Buds of black locust trees were sampled for micropopagation (Figure 2) and DNA analyses. Shoot apical buds were dissected and processed for aseptic shoot culture following the general tissue culture protocols (Gyulai et al., 1992, 2006). Buds were cleaned, and washed with detergent (3 min), followed by surface sterilization with ethanol (70% v/v) for 1 min and a commercial bleaching agent (8% NaOCl w/v) for 1 min; followed by three rinses with sterile distilled water, and incubated in aseptic tissue culture medium F6 (Gyulai et al., 1992) supplemented with 0.1 mg/L 2,4-dichlorophenoxyacetic acid (2,4-D) and kinetin, respectively.

In vitro clones of the oldest Hungarian black locust (*Robinia pseudoacacia cv. 'Bábolna-1710'*) were sprouting from aseptic buds in three weeks (Figure 2). Several buds produced calli. After further micropropagatiion with nodal segments, clones were rooted on hormone-free medium, transplanted to pots (Shu et al., 2003) and grown in fields (Figure 2). In total more than fifty regenerants were transplanted, however, due to the week rooting capability, about ten fully developed trees survived the fist winter.

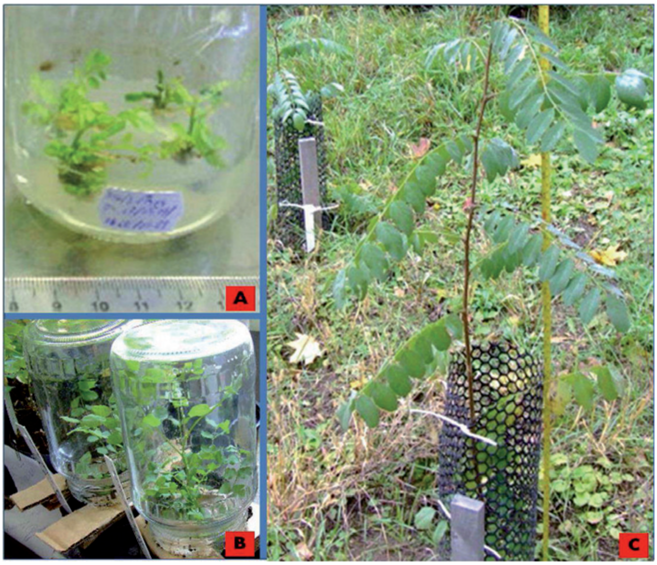


Fig. 2: *In vitro* (A), potted (B) and field grown (C) clones of the oldest Hungarian black locust tree (*Robinia pseudoacacia cv. 'Bábolna-1710'*).

Molecular analysis

Young fresh leaves (0.1g) were ground in an aseptic mortar with liquid nitrogen. DNA was extracted by the CTAB (cethyltrimethylammonium bromide) method according to the modification of Gyulai et al. (2000) followed by an RNase-A treatment (Sigma, R-4875) for 30 min at 37°C in each case. The quality and quantity of extracted DNA (2 μl) was measured by a NanoDrop ND-1000 UV-Vis spectrophotometer (NanoDrop Technologies, Delaware, USA – BioScience, Budapest, Hungary). DNA samples were adjusted to a concentration of 30 ng/μl with ddH_2O and subjected to PCR amplification.

For PCR analysis six loci were amplified (Table 1). Hot Start PCR (Erlich et al., 1991) was combined with Touchdown PCR (Don et al. 1991) using AmpliTaq Gold^{TM} Polymerase. Reactions were carried out in a total volume of 25 µl (containing genomic DNA of 30-50 ng, 1 x PCR buffer (2.5 mM $MgCl_2$), dNTPs (200 µM each), 20 pmol of each primer and 1.0 U of *Taq* polymerase. Touchdown PCR was performed by decreasing the annealing temperature by 1.0 ^oC / per cycle with each of the initial 12 cycles (PE 9700, Applied Biosystems), followed by a ‘touchdown’ annealing temperature for the remaining 25 cycles at 56 ^oC for 30 s with a final cycle of 72 ^oC for 10 min (transgene detection) and held at 4 ^oC. A minimum of three independent DNA preparations of each sample was used. Amplifications were assayed by agarose (1.8 %, SeaKem LE, FMC) gel electrophoresis (Owl system), stained with ethidium bromide (0.5 ng/µl) after running at 80 V in 0.5 x TBE buffer. Each successful reaction with scorable bands was repeated at least twice. Transilluminated gels were analyzed by the ChemiIlmager v 5.5 computer program (Alpha Innotech Corporation - Bio-Science Kft, Budapest, Hungary). A negative control which contained all the necessary PCR components except template DNA was included in the PCR runs.

Microsatellite and ITS fragments were forwarded for ALF analysis using ALF ExpressII (Pharmacia – Amersham, AP-Hungary, Budapest). One strand of each of the SSR primer pairs was labeled with Cy5 dye (Röder et al., 1998; Huang et al., 2002; Gyulai et al., 2006, 2011).

Except *Gunnerales* and *Geraniales* all of the eudicot plant orders include woody species. Order Fabales comprise 20.055 species of 754 genera of 4 families (Fabaceae, Polygalaceae, Quillajaceae, Surianaceae). Of them family Fabaceae (syn. Robinoid legumes; tree legume) comprise the most numerous woody species (240 – 253) of a plant family, divided in three subfamilies (Faboideae, Caeasalpinoideae, and Mimosoideae) (Figure 3) and comprising 11-12 genera (*Hebestigma/ ennea, Gliricidia/Poitea, Olneya, Robinia, Poissonia, Coursetia, Peteria, Genistidium* and *Sphinctospermum*) (Lavin et al., 2003). Genus *Astragalus* of Fabaceae, is also unique being the largest plant genus, which comprises 2,500 species (Sanderson and Wojciechowsk,1996). Legume trees have several endemic genera with unique genome constitutions such as the Caribbean endemic *Hebestigma* and *Poitea*; and the Australian *Castaneospermum australe* (Lavin et al., 2003).

Table 1: SSR primer pair sequences used for nuclear DNA analyses with indications of the expected fragment sizes.

Primer loci	**Sequences**	**References**	**Fragment sizes**
Rops16 $(CT)_{13}$	AACCCTAAAAGCCTCGTTATC TGGCATTTTTTGGAAGACACC	Lian et al. (2004)	195-223 bp
RP035 $(TC)_{15}$	GGAGTGGAATGCATGCTCTCATG TCCAAATGGAAACTCCCTTGAAACAGC	Mishima et al. (2009)	89-112 bp
RP102 $(GA)_{12}$	CCAAATCTCAAAATGTGCTAAGTAGC ACTTGGGCTATGGTATTGCA		205-211 bp
RP200 $(AG)_{23}$	GGTTTCTTTGTTCACCTGCTCTGG ACCTACGTGTCCACGGCTCT		160-198 bp
RP206 $(GT)_{9}$	GCCAAATCCCATTAGATCACAGTTGA AGAAGTTAGACTTACGTGCTGC		222-246 bp
*scu*10 $(CAA)_{6}$	TTCTCCGCCACCTCCTTTTCAC TACCCCCACAACCCTTTTTCCC	Scott et al. (2000)	205 – 274 bp

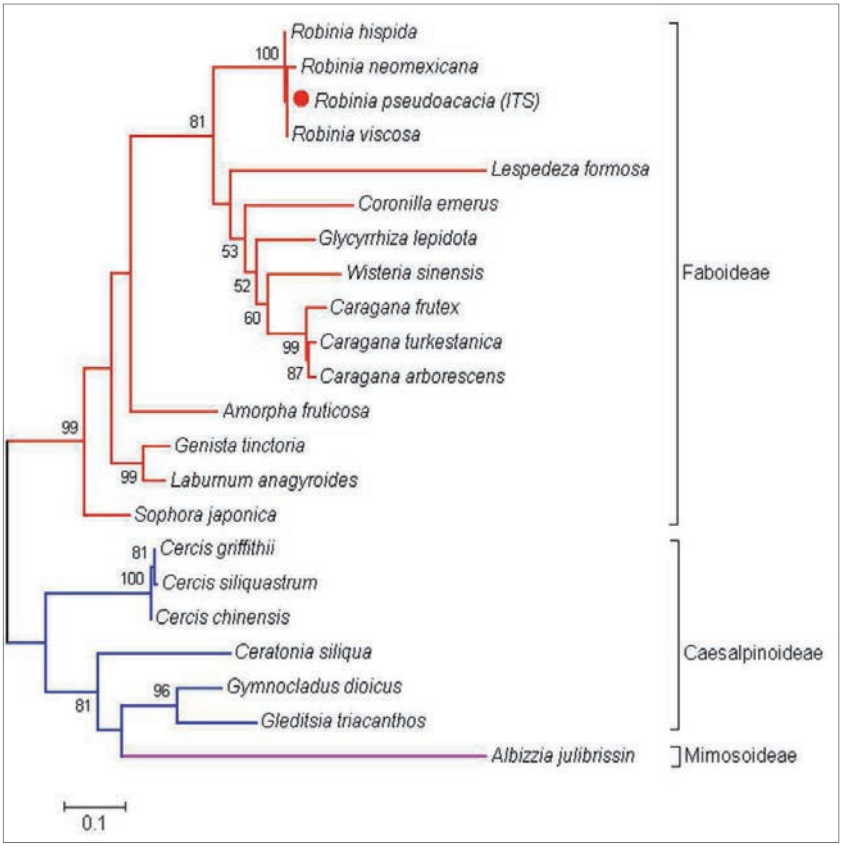


Fig. 3A: *In silico* ITS1-5.8S-ITS2 (ML - Maximum Likelihood; Hillis et al*.,* 1994) phylogram (MEGA4; Tamura et al., 2007) of legume trees including the main four *Robinia* species. NCBI accession numbers are: *Albizia julibrissin* (FJ572041), *Amorpha fruticosa* (AFU59890), *Caragana arborescens* (FJ537262), *Caragana frutex* (FJ537285), *Caragana turkestanica* (FJ537256), *Cerationia siliqua* (AJ245576, AJ245575), *Cercis chinensis* (FJ432284), *Cercis griffithii* (FJ432280), *Cercis siliquastrum* (FJ432281), *Colutea arborescens* (CAU56010, CAU56009), *Hyppocrepis* (*Coronilla) emerus* (AF450240), *Genista tinctoria* (AF330664), *Gleditsia triacanthos* (AF509969), *Glycyrrhiza lepidota* (U50759, U50758), *Gymnoclagus dioica* (AF510032), *Laburnum anagroides* (AY263679), *Lespedeza thunbergii* (GU572199), *Robinia hispida* (AF398819), *Robinia neomexicana* (AF537351), *Robinia pseudoacacia* (AF450153), *Robinia viscosa* (AF398821), *Sophora japonica* (FJ528289), *Wisteria sinensis* (EU424072).

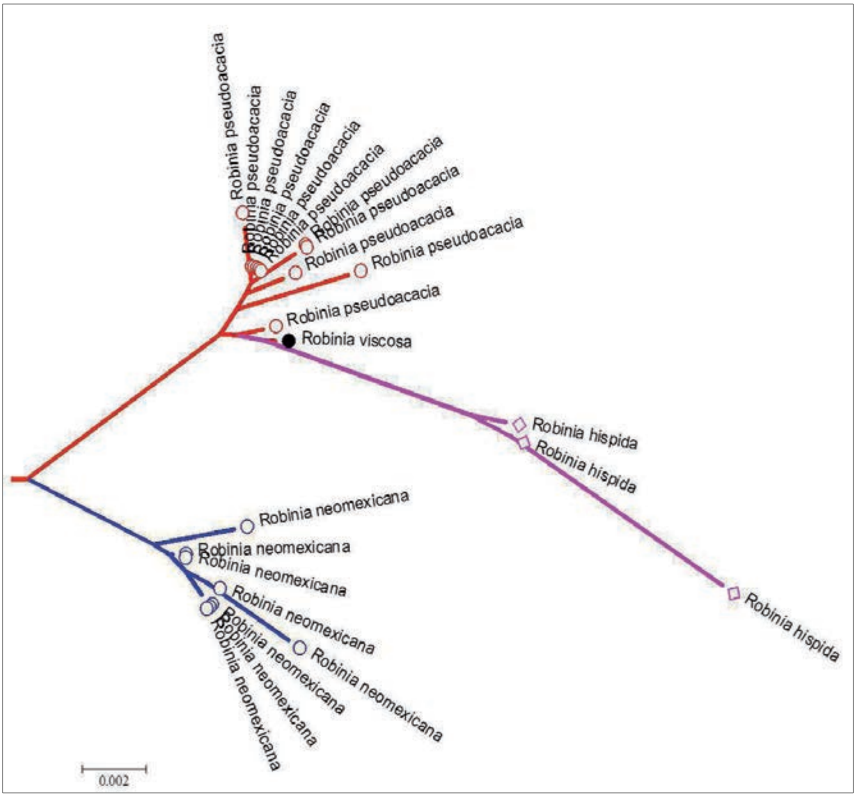


Fig. 3B: *In silico* ITS1-5.8S-ITS2 (ML - Maximum Likelihood; Hillis et al., 1994) phylogram (MEGA4; Tamura et al., 2007) of legume trees grown in Hungary (NCBI, 683 bp). NCBI accession numbers are: *Albizia julibrissin* (FJ572041), *Amorpha fruticosa* (AFU59890), *Caragana arborescens* (FJ537262), *Caragana frutex* (FJ537285), *Caragana turkestanica* (FJ537256), *Cerationia siliqua* (AJ245576, AJ245575), *Cercis chinensis* (FJ432284), *Cercis griffithii* (FJ432280), *Cercis siliquastrum* (FJ432281), *Colutea arborescens* (CAU56010, CAU56009), *Hyppocrepis* (*Coronilla) emerus* (AF450240), *Genista tinctoria* (AF330664), *Gleditsia triacanthos* (AF509969), *Glycyrrhiza lepidota* (U50759, U50758), *Gymnoclagus dioica* (AF510032), *Laburnum anagroides* (AY263679), *Lespedeza thunbergii* (GU572199), *Robinia hispida* (AF398819), *Robinia neomexicana* (AF537351), *Robinia pseuodoacacia* (AF450153), *Robinia viscosa* (AF398821), *Sophora japonica* (FJ528289), *Wisteria sinensis* (EU424072).

The species specific, highly conserved locus of nuclear ITS sequences were highly monomorphic among species of *Robinia* (Figure 3) and clade *Robinieae* (*Coursetia, Cracca, Genistidium, Gliricidia, Hebestigma, Hybosema, Lennea, Olneya, Peteria, Poitea, Robinia, Sphinctospermum*) (*Robinioid* clade) (Lavin et al*.,* 2003) with rare SNPs (Single Nucleotide Polymorphism). ITS sequences were also sequenced from *Robinia*, Hancock Cave, USA (NCBI # AF176391; AF176390 and AF176389). The ITS analysis of the four main *Robinia* species growing in Hungary, indicated that *R. viscosa* is genetically the closest species to *R. pseudoacacia*, and *R. hispida* (Figure 3).

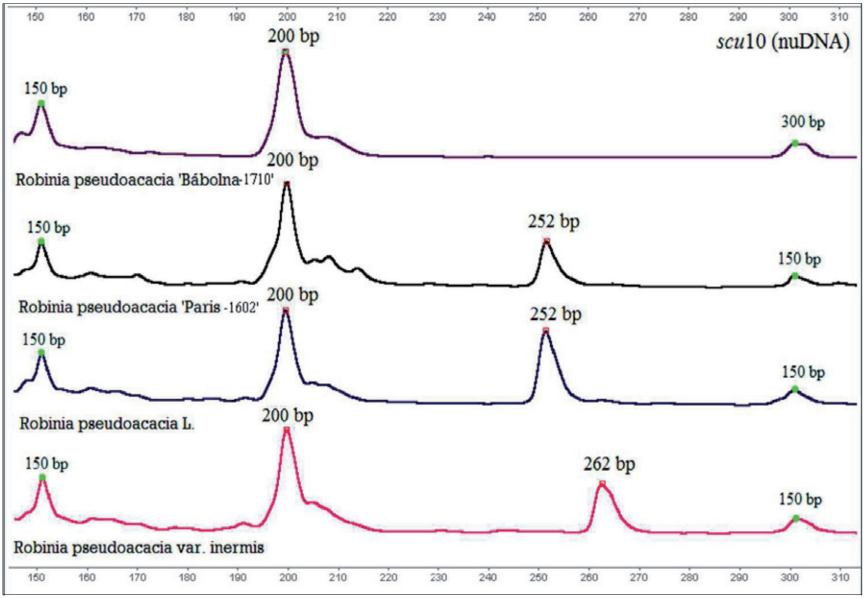


Fig. 4: Molecular marker for thornlessness (*R. p. inermis*) at the nuDNA *scu*10 (262 bp) locus compared tor '*Bábolna-1710*' clone. PCR amplified ALF fluorograms (15-300 nt) of four legume trees were compared. Mw standards (150- and 300 bp), and the amplified DNA fragments (200-, 252-, and 262 bp) are indicated.

In our study presented thornless ('*inermis*') clones were also analyzed and a new molecular barcode for thornlessness was detected at *scu*10 locus (Figure 4-5). Thornlessness is caused mainly epigenetically by epidermis mutation in shoot meristems. Certain thornless varieties of-

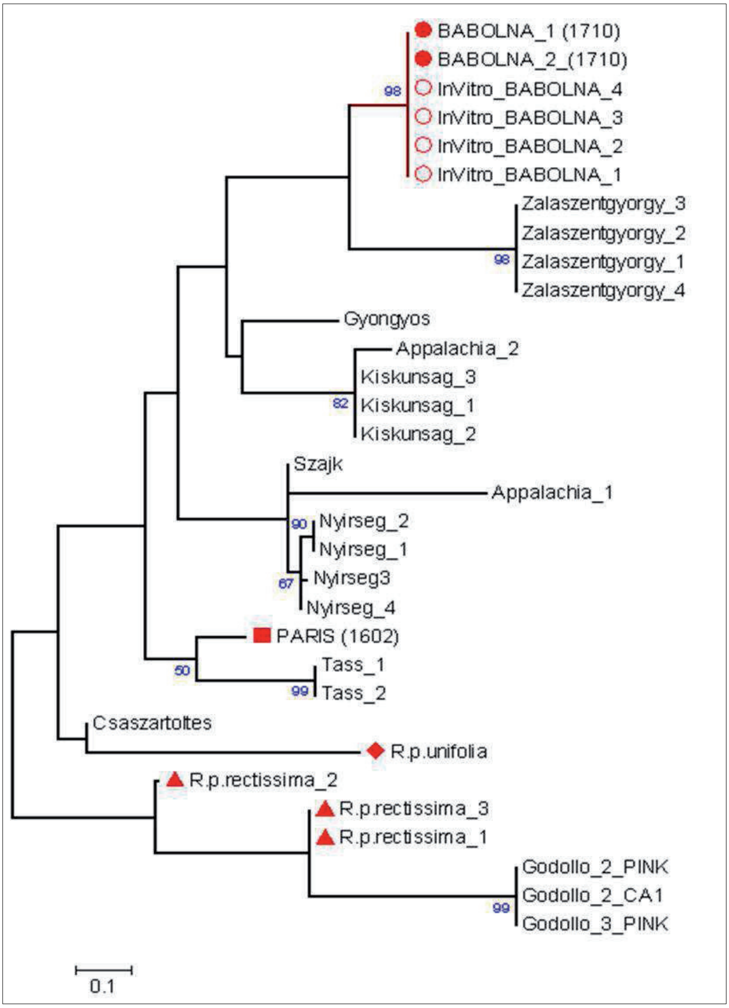


Fig. 5: ML (Maximum Likelihood; Hillis et al., 1994) phylogram (MEGA4; Tamura et al., 2009) of *Robinia pseudoacacia* clones growing in different habitats of Hungary. The oldest Hungarian black locust (*Robinia pseudoacacia cv. 'Bábolna-1710'*; ●) and its *in vitro* clones (○), and further 31 samples from natural black locust populations sampled in Hungary (names of cities are indicated) are compared to the oldest European clone 'Paris-1602' (France) (■), and two subspecies of *R. p. rectissima* (▲) and *R. p. unifolia* (♦) (Hungary). The phylogram was based on the patterns of 170 PCR fragments of 28 alleles at six SSR loci determined of the study presented. Relative genetic distance (scale 0.1) and bootstrap values over 50% are indicated at the nodes.

-blackberry, e.g. the tetraploid periclinal chimera *Rubus laciniatus* Willd. cv. '*thornless evergreen'* (identified in 1939 and improved by gamma irradiation by Jennings in 1984) produce thorny adventitious root suckers (Scott et al., 1957; Hall et al., 1986; Lewers et al., 2008). Thornless (non-chimeral) blackberry plants were obtained from tissue culture (Hall et al., 1986). When these plants were grown to maturity, flowered and hybridized with various thorny and thornless cultivars, the thornless vs. thorny segregation ratios suggested that thornless gene(s) might be dominant over the thorny alleles (Hall et al., 1986). Thornlessness can also be linked to morphological characters like in thornless blackberry(*Rubus fruticosus*) where cotyledon marginal hairs of seedlings are absent. Trees of *Gleditsia triacanthos* of 10 years old or more showed a definite thornless region in the upper and outer shoot growth. When hardwood cuttings for propagation were taken from this thornless area, the scions generally remained thornless (Chase, 1947). In case of black locust, several forms of '*inermis'* were identified such as *Robinia inermis* Jacquin; *Robinia pseudoacacia* f. *inermis* (DC.) Rehder; *Robinia pseudoacacia* subsp. *inermis* (Jacquin) Arcangeli; *Robinia pseudoacacia* var. *inermis* (Ortega) DC; *Robinia pseudoacacia* var. *inermis* DC. Thornless black locust mutants were also developed after gamma irradiation and tissue culture

Microsatellites (SSRs) of *Robinia pesudoacacia* are available at gene banks (e.g. NCBI). The most frequently used markers are: (1) Rops02 (AB075029), (2) Rops04 (AB075030), (3) Rops05 (AB075031), (4) Rops06 (AB075032), (5) Rops08 (AB075033), (6) Rops09 (AB075034), (7) Rops15 (AB120731), (8) Rops16 (AB120732), (9) Rops18 (AB120733), (10) RP032 (AB353934), (11) RP035 (AB353927), (12) RP102 (AB353928), (13) RP106 (AB353929), (14) RP109 (AB353930), (15) RP150 (AB353931), (16) RP200 (AB353933), (17) RP206 (AB353932) and (18) RP211 (AB353935). Of these, here we used five SSR loci of *Rops*16 - $(CT)_{13}$; *RP*035 - $(TC)_{15}$; *RP*102 - $(GA)_{12}$; *RP*200 $(AG)_{23}$; and *RP*206 $(GT)_9$, including *scu*10, an SSR of *Vitis vinifera*. Nuclear SSRs were found useful in *Robinia* genotyping (Lian and Hogetsu, 2002), however certain microsatellites showed high somatic instability revealing differences in the SSR lengths at the locus *Rops*15 among the leaflets of the same pinnate leaf (Lian et al., 2004).

To conclude, clonal propagation of a 301-year-old black locust tree grown in city Bábolna (Hungary) was applied successfully for maintaining a 301-year-old genome. The method might also be applied to other 'living fossils' such as the 11,700-year-old *'King Clone'* creosote bush (*Larrea tridentata*) (Mojave Desert, Lucerne Valley, California, USA); the 9500-year-old *'Old Tjikko'* Norway spruce (*Picea abies*) (Sweden); the 4789-year-old '*Methuselah*' Basin Bristlecone Pine (*Pinus longaeva*) (Wheeler Peak, eastern Nevada, USA), the 4-5000-year-old yew (*Taxus* *baccata*) (Llangernyw, Conwy County Borough, North Wales, UK), and probably be useful to resurrect '*Prometheus'*, the formerly oldest *Pinus* tree (4862-year-old *Pinus longaeva*; USA) cut down in 1964 as a research mistake.

References

Bhojwani, S.S. and Razdan, M.K. (1996) Plant tissue culture: theory and practice, Revised edition Elsevier, vii+767 pp. ISBN 0444816232.

Böhme, M., Bruch, A.A. and Selmeier, A. (2007) The reconstruction of Early and Middle Miocene climate and vegetation in Southern Germany as determined from the fossil wood flora. Palaeogeography, Palaeoclimatology, Palaeoecology 253: 91-114.

Chase, S.B. (1947) Propagation of thornless honeylocust. Journal of Forestry 45:715-722.

Dini-P, O. and Aravanopoulos, F.A. (2008) Artificial hybridization between *Robinia pseudoacacia* L. and *R. pseudoacacia* var. *monophylla* Carr. Forestry 81: 91-101.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K. and Mattick, J.S. (1991) Touchdown PCR to circumvent spurious priming during gene amplification. Nucleic Acids Research 19: 4008.

Erlich, H.A., Gelfand, D. and Sninsky, J.J. (1991) Recent advances in the polymerase chain reaction. Science 252: 1643-1651.

Földes, J. (1903) Adalékok az akác ismeretéhez ('Further data about Robinia'). Erdészeti lapok 1903/I: 63-65.

Gyulai, G., Janovszky, J., Kiss, E., Lelik, L., Csillag, A. and Heszky, L.E. (1992) Callus initiation and plant regeneration from inflorescence primordia of the intergeneric hybrid *Agropyron repens* (L.) Beauv. x *Bromus inermis* Leyss. cv. *nanus* on a modified nutritive medium. Plant Cell Report 11: 266-269*.*

Gyulai, G., Gémesné, J.A., Sági, Zs., Venczel, G., Pintér, P., Kristóf, Z., Törjék, O., Heszky, L., Bottka, S., Kiss, J. and Zatykó, L. (2000) Doubled haploid development and PCR-analysis of F_1 hybrid derived DH-R_2 paprika (*Capsicum annuum L.*) lines. J Plant Physiology 156:168-174.

Gyulai, G., Humphreys, M., Lágler, R., Szabó, Z., Tóth, Z., Bittsánszky, A., Gyulai, F. and Heszky, L. (2006) Seed remains of common millet from the 4th (Mongolia) and 15th (Hungary) centuries: AFLP, SSR and mtDNA sequence recoveries. Seed Science Research 16: 179-191.

Gyulai, G., Láposi, R., Rennenberg, H., Veres, A., Herschbach, C., Fábián, Gy. and Waters Jr, L. (2011) Conservation genetics (1710 – 2010) - Cloning of living fossils: Micropropagation of the oldest Hungarian black locust tree (*Robinia pseudoacacia*) planted in 1710 (Bábolna, Hungary). In: Gyulai, G. (Ed.) Plant Archaeogenetics. Nova Sci Publisher Inc., New York, USA. pp. 117-127.

Hall, H.K., Cohen, D., Skirvin, R.M. (1986) The inheritance of thornlessness from tissue culture-derived '*Thornless evergreen*' blackberry. Euphytica 35: 891-898.

Hillis, D.M., Huelsenbeck, J.P. and Swofford, D.L. (1994) Hobgoblin of phylogenetics? Nature 369: 363–364.

Huang, X.Q., Börner, A., Röder, M.S. and Ganal, M.W. (2002) Assessing genetic diversity of wheat (*Triticum aestivum* L.) germplasm using microsatellite markers. Theoretical and Applied Genetics 105: 699-707.

Iseley, S. and Peabody, F.J. (1984) *Robinia* (Leguminosae: Papilionoidea). Castanea 49: 187-202.

Keresztesi, B. (1983) Breeding and cultivation of black locust (*Robinia pseudoacacia* L.) in. Hungary. Forest Ecology and Management 6: 217-244.

Lavin, M., Wojciechowski, M.F., Gasson, P., Hughes, C.E. and Wheeler, E. (2003) Phylogeny of robinioid legumes (Fabaceae) revisited: *Coursetia* and *Gliricidia* recircumscribed, and a biogeographical appraisal of the Caribbean endemics. Systematic Botany 28: 387–409.

Lawrence, C. Matten, L.C., Gastaldo, R.A. and Lee, M.R. (1977) Fossil *Robinia* wood from the western United States. Review of Palaeobotany and Palynology 24:195-208.

Lewers, K.S., Saski, C.A., Cuthbertson, B.J., Henry, D.C., Staton, M.E., Main, D.S., Dhanaraj, A.L., Rowland, L.J. and Tomkins, J.P. (2008) A blackberry (*Rubus* L.) expressed sequence tag library for the development of simple sequence repeat markers. BMC Plant Biology 8: 69.

Lian, C. and Hogetsu, T. (2002) Development of microsatellite markers in black locust (*Robinia pseudoacacia*) using a dual-supression-PCR technique. Molecular Ecology Notes 2: 211-213.

Lian, C., Oishi, R., Miyashita, N. and Hogetsu, T. (2004) High somatic instability of a microsatellite locus in a clonal tree, *Robinia pseudoacacia*. Theoretical and Applied Genetics 108:836-841.

Mishima, K., Hirao, T., Urano, S., Watanabe, A. and Takata, K. (2009) Isolation and characterization of microsatellite markers from

Robinia pseudoacacia L. Molecular Ecology Resources 9: 850–852.

Peabody, F.J. (1982) A 350-Year-Old American Legume in Paris. Castanea 47: 99-104.

Putod, R. (1982) Les arbres fouragers. Le Févier. *Forêt* Méditerranéenne 4: 33-42.

Rédei, K. and Veperdi, I. (2009) The role of black locust (*Robinia pseudoacacia* L.) in establishment of short-rotation energy plantations in Hungary. International Journal of Horticultural Science 15: 41–44.

Röder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.H., Leroy, P. and Ganal, M.W. (1998) A microsatellite map of wheat. Genetics 149: 2007-2023.

Sanderson, M.J. and Wojciechowsky, M.F. (1996) Diversification rates in a temperate legume clade: Are there so many species of Astragalus (Fabaceae)? American Journal of Botany 83:1488-1502.

Sandoval-Gil, J.M., Ruiz, J.M., Marín-Guirao, L., Bernardeau-Esteller, J. and Sánchez-Lizaso, J.L. (2014) Ecophysiological plasticity of shallow and deep populations of the Mediterranean seagrasses *Posidonia oceanica* and *Cymodocea nodosa* in response to hypersaline stress. Marine Environmental Research 95: 39-61.

Santamour, F.S. and McArdle, A.J. (1983) Checklist of cultivars of honeylocust (*Gleditsia triacanthos* L) of some trees. Journal of Arboriculture 9: 248-252.

Scott, D.H., Darrow, G.M. and Ink, D.P. (1957) *Merton Thornless* as a parent in breeding thornless blackberries. Proceedings of the American Society for Horticultural Science 69: 268–277.

Shu, Q.Y., Liu, G.S., Qi, D.M., Chu, C.C., Li, J. and Li, H.J. (2003) An effective method for axillary bud culture and RAPD analysis of cloned plants in tetraploid black locust. Plant Cell Reports 22:175-180.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599.

Wheeler, E.A. and Landon, J. (1992) Late Eocene (Chadronian) dicotyledonous woods from Nebraska: evolutionary and ecological significance. Review of Palaeobotany and Palynoiogy 74: 267-282.

*****

Plant DNA Barcoding and Phylogenetics, Pages 255-267
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

16 Genetic Diversity Assessment of *Caralluma adscendens*

M.R. Naik, Y.L. Krishnamurthy and S. Karuppusamy

Introduction

Caralluma R.Br (*Sensu lato*) of the family Apocynaceae, is a genus of about 120 species, with a wide distribution in India, Arabia and the Mediterranean area of the world, with more number of species concentrated in Africa (Mabberley, 1993). The genus *Caralluma* was first named by Robert Brown in 1810 to an Indian species with very characteristic elongated flowering succulent stem. The plant is known to have possessed a variety of medicinal properties include carminative, febrifugal, anthelmintic, antirheumatic, antidiabetic, antipyretic, anti-inflammatory, antinociceptive and antioxidant (Karupusamy et al., 2013). Many species of *Caralluma* have been reported as potential appetite suppressants and all these species found esterified polyhydroxy pregnane glycoside (Deepak et al., 1997). *Caralluma adscendens* which is endemic to southern India have been recognized by six infraspecific varieties so far (Karuppusamy et al., 2013). In which *Caralluma adscendens* var. *fimbriata* is distributed in Andhra Pradesh, Karnataka and Tamilnadu region of India, and exhibited diverse morphologic and genetic variations. The taxon *Caralluma adscendens* var. *fimbriata* is consumed as vegetable during famine to decrease appetite and as thirst quencher. Rebecca et al. (2007) have observed the var. *fimbriata* extract on appetite suppressive and reduction of food intake in Indian populations. Radhakrishnan et al. (1999) have proved the *Caralluma adscendens* var. *fimbriata* extract for weight loss in human being.

Caralluma adscendens var. *fimbriata* is highly demanded in pharmaceutical companies for the preparation of number of antiobesity formularies in particularly weight loss pills, syrups and capsules. The company based raw drugs largely as succulent stems collected only from the natural population in southern Peninsular India. Many of the natural population of *C. adscendens* var. *fimbriata* are growing with its related species and varieties in natural habitat. The taxa also have hybridized potential with its coexisting varieties to exhibit number of morphoforms and genetic variations. Due to diverse nature of morphological plasticity, identification and taxonomy is much confusing, and also for extraction active pharmaceutical principles too varied in the populations.

Taxonomy based on morphological characters are supported with various kinds of biochemical (HPLC, HPTLC) and molecular (RFLP, RAPD, ISSR, AFLP) data used to study intra-specific and inter-specific polymorphism (Willium et al*.,* 1990, Welisch and Mc Chelland 1990, Shinde et al., 2007). These techniques have been extensively used for the identification of genotypes in plants and these are convenient method for the detection of polymorphism in the absence of sequence information with a relatively low cost. The related members of Apocynaceae such as *Gymnema sylvestri* R.Br (Siddique et al., 2000) and *Gymnema* germplasm (Nair and Keshavachandran, 2006) have been genetically characterized using molecular markers techniques. The genetic composition of *Caralluma* in different geographical locations needs to be assessed for its efficient conservation and management since these are threatened species. Yet there has been no previous report on the use of RAPD and ISSR method to characterize the genetic diversity of *Caralluma* across the different geographical locations in India.

Variations are most common features among the Asclepiads in particularly succulent Asclepiads like *Caralluma* and *Ceropegia* in India. Even within the species many varieties have described under the *Caralluma adscendens*. They are distributed overlapping manner in the small geographical area of habitat and also naturally inbreeding groups formed many new combinatorial characters in the habitat. *Caralluma adscendens* var. *fimbriata* is pharmaceutically important plant which is highly overexploited from the natural habitat for extraction of natural medicine. There is no any scientific collection methods followed until now for these valuable medicinal plants and there is no availability of any agronomic techniques for cultivation of the plant species. The plant collectors are not knew about the importance of bioresource and sustainable utilization of limited plant sources. Commercial plant collectors collected all the varieties for extraction of medicine but medicinal potentiality found with precise variety var. *fimbriata* only. Keeping these points in view the present investigation was planned to carryout morphologic and genetic variation assessment using RAPD and ISSR markers for the identification of potential germplasm for cultivation and sustainable utilization.

Taxon sampling

Caralluma adscendens comprises seven varieties in southern India and about four varieties are distributed in Karnataka. Among them var. *fimbriata* is highly medicinal which is existing variety of morphofoms in Karnataka. For the identification of potential morphofom population, here we used molecular tools for assessing the diversity. *C. adscendens* var. *fimbriata (Cf)* plant was collected from 12 districts of Karnataka (Table 1), and were identified by taxonomists at Department of Post Graduate Studies and Research in Applied Botany, Kuvempu University and deposited in its depository.

Table 1: *Caralluma adscendens* var. *fimbriata* (*s.str.*) collected from twelve districts of Karnataka

Accession number	Locality
Cf1	Gulbarga university, Gulbarga
Cf2	Halmatti, Bijapur
Cf3	Kudalsangam, Bagalkot
Cf4	Ranibennur, Haveri
Cf5	Mallapura, Gadag
Cf6	Raichur fort, Raichur
Cf7	Darogi, Bellary
Cf8	Kodaganur, Davangere
Cf9	Muddapura, Chitradurga
Cf10	Kadur (Pura), Chickmagalur
Cf11	Arasikere, Hassan
Cf12	Siddarbetta, Tumkur

Genomic DNA Isolation

The total genomic DNA was extracted with the help of a modified protocol based on previous literature (Aras et al*.,* 2005, Ibrahim, 2011, Jabbarzadeh, 2009, Hameed et al*.,* 2004). The extraction was done from fresh phyllodes, sliced into a portion of the plant sample of 50 mg each.

The plant material was washed, dried and weighed accurately before use. The plant material was ground with 600µl 2X cetyl trimethyl ammonium bromide (CTAB) buffer (100 mM Tris HCl at pH 8.0, 20 mM EDTA, 2% CTAB, 1.4 mM NaCl and 0.1% β-mercaptoethanol) with mortar and pestle and incubated at 65ºC for 45 minutes with occasional mixing. 5% PVP was added to the CTAB buffer just before the grounding of plant material. The solution was extracted twice with an equal volume of Chloroform: Isoamyl alcohol (24:1) by centrifugation at 10000rpm for 15 min at room temperature in a microfuge tube. The supernatant was precipitated with ice cold ethanol and the precipitate was collected by centrifugation. The precipitate was washed with 70% ethanol and air dried. The DNA pellet was finally dissolved in 40µl TE buffer 10 mM Tris-HCl, 1 mM EDTA, pH 8, RNase A to remove RNA and stored at -20ºC. The quality and quantity of isolated DNA was determined by Gel Electrophoresis with 0.8% agarose in 1X TAE (Tris Acetate EDTA) buffer, followed by Ethidium bromide staining. The gel was visualized under a UV transilluminator to analyze the amount of DNA isolated by comparing the bands formed with a DNA weight maker (Lambda-Hind III) purchased from Geno Biosciences, Bangalore.

Primer Screening

After a thorough primer screening, ten random decamer primers, corresponding to kit N and O from Operon Technologies (Alameda, California, USA) and five synthesized ISSR primers (M/S Bangalore Genei, Bangalore, India) were finalized for study using PCR techniques to determine the genetic relationships between the samples under study. The primers selected after screening is listed in Table 2. These primers were selected based on their ability to detect distinct, clearly resolved and polymorphic amplified products. To ensure reproducibility, the primers generating no, weak, or complex patterns were discarded.

RAPD and ISSR assay

Reaction conditions for amplification were optimized according to (Williams et al*.,* 1990) Amplification was performed in volume of 25 µl for single reaction. The reaction mixture was composed of 20ng template DNA, 100 µM of each deoxyribonucleotide triphosphate, 20 ng of decanucleotide primer, 1.5mM $MgCl_2$, 1X Taq buffer (10mM Tris-HCl (pH 9.0), 50mM KCl, 0.001% gelatin), and 1 U Taq DNA polymerase (M/S Bangalore Genei, India). Thermal profile conditions were as follows: initial denaturation for 2 min at 94^{0}C, followed by 40 cycles of denaturation at 94^{0}c for 30 seconds, primer annealing at 40^{0}C for 1 min and extension at 72^0 C for 2 min and lastly final extension (12 min) was carried out at 72^0 C. The PCR products were stored at 4^{0}C until further study. The amplified products were subsequently separated on 1.5 % agarose gel with Ethidium bromide staining for about 1 h at 80 volts. The products were visualized under UV transilluminator. PCR products were analyzed using a 1 kb DNA ladder (Geno Biosciences, Bangalore, India).

Data were recorded as presence (1) or absence (0) of band products from the photographic examination. Each amplification fragment was named by the source of the primer, the kit letter or number, the primer number and its approximate size in base pairs. Bands with similar mobility to those detected in the negative control, if any, were not scored. A pair-wise matrix of distances between landraces was determined for the RAPD and ISSR data using the Dice formula in the Free Tree programme. The average of similarity matrices was used to generate a tree by UPGMA (unweighted pair-group method arithmetic average) using NTSYS- PCversion 2.1 (Jaccard, 1908; Rohlf, 2000).

Table 2: List of RAPD and ISSR analysed.

Primer		**Sequence**
RAPD	OPN 11	5’TCGCCGCAAA3’
	OPN 12	5’CACAGACACC3’
	OPN 13	5’AGCGTCACTC3’
	OPN 14	5TCGTGCGGGT’3’
	OPN 15	5’CAGCGACTGT3’
	OPN 16	5’AAGCGACCTG3’
	OPN 17	5’CATTGGGGAG3’
	OPN 18	5’GGTGAGGTCA3’
	OPO 19	5’GTCCGTACTG3’
	OPO 20	5’GGTGCTCCGT3’
ISSR		
	HB10	5’GAGAGAGAGACCGAGAGAGAGACC3’
	HB11	5’GTGTGTGTGTCCGTGTGTGTGTCC3’
	HB12	5’CACCACCACGCCACCACCACGC3’
	844A	5’CTCTCTCTCTCTCTCTACCTCTCTCTCTCTCTCTC3’
	844B	5’CTCTCTCTCTCTCTCTGCCTCTCTCTCTCTCTCTGC3’

Genetic diversity based on RAPD marker

Using ten RAPD primers, a total number of 73 bands were produced, of which 43 were polymorphic bands and a polymorphism of 58.9 per cent was observed (Table 3). The size of the RAPD fragments ranged from 100-3000bp; visualized using UV Trans illuminator. The banding profile by RAPD primers have been shown in Figure 1. The primer OPN20 produced maximum number of fragments (11 fragments) with 7 polymorphic bands. However, the primer OPN 15 produced maximum polymorphism (80 per cent). OPN14 produced the lowest number of amplified bands and OPN16 produced the least polymorphism (37.5%) amongst the primers studied. It had also been well documented that geographical distribution and ecological niches exhibit the different genetic characterizations and had strong effects on the organization of genetic constitution (Loveless and Hamrick, 1984). Pejic et al*.,* (1998) reported that 150 polymorphic bands made it possible for a researcher to reliably estimate genetic similarities among genotypes within the same species. Esposito et al*.,* (2007) found a total of 162 polymorphic bands ranging in size from 400 to 1200 bp, with an average of 23 bands per primer combination. Li and Quiros (2001) in *Brassica oleracea* L., Budak et al. (2004) in Buffalo grass and Ferriol et al. (2004) in *Cucurbita moschata* reported that there were 10-20 polymorphic bands per primer combination.

In the dendrogram generated using RAPD, *C. adscendens* var. *fimbriata* (Cf) collected from Bijapur Cf2 is placed separately in clade I, isolated from the rest 11 species by RAPD molecular approaches. The remaining eleven accessions are positioned in clade II and are differentiated into two clusters I and II by RAPD marker systems. Cluster I contained seven species like Cf6, Cf10, Cf11, Cf12, Cf7, Cf8 and Cf9. Cluster II contained four accessions i.e. Cf1, Cf3, Cf4 and Cf5. These accessions were from Gulbarga (Cf1) and Bagalkot (Cf3) district which were showing genetic relation to plants of Haveri (Cf4) and Gadag (Cf5) where they showed 72 per cent similarity (Figure 2).

The highest dissimilarity value (0.874) was observed between Cf12 and Cf11 followed by Cf12 and Cf10, Cf9 and Cf8, Cf10 and Cf6, Cf12 and Cf9, Cf12 and Cf7, Cf12 and Cf6, Cf7 and Cf6. The lowest dissimilarity was observed between Cf8 and Cf1 (0.586) and Cf8 and Cf2 (0.586). A value of Jaccard coefficient of correlation indicates that Gulbarga (Cf1), Bijapur (Cf2) and Bagalkot (Cf3) are very similar. Genetic diversity of the species with highest dissimilarity will be useful for further program. In the present study, the data showed that *C. adscendens* var. *fimbriata* differentiate genetic races from north Karnataka to south part of Karnataka.

Table 3: Total number of amplified fragments and the number of polymorphic fragments generated by PCR using selected RAPD primers

Primer	Total of bands(a)	Monomorphic bands	Polymorphic bands (b)	(%) of Polymorphism (b/a x 100)
OPN 11	7	3	4	57.14
OPN 12	10	5	5	50
OPN 13	9	4	5	55.5
OPN 14	3	1	2	66.6
OPN 15	5	1	4	80
OPN 16	8	5	3	37.5
OPN 17	5	2	3	60
OPN 18	5	2	3	60
OPO 19	10	3	7	70
OPO 20	11	4	7	63.6
Total	73		43	
Mean value	7.3			58.9

In RAPD, clade I was belongs to plants of Bijapur. Externally features of clade B represented by C. *adscendens* var. *fimbriata* separated into one group rest of 11 districts, were positioned in clade II. Clade B divided into clusters C and D. The cluster C had plant species distributed in 7 districts (Hassan, Bellary, Bagalkot, Davangere, Tumkur, Raichur, Bijapur) and cluster D had three districts (Haveri, Chickmagalur and Chitradurga). In RAPD, clade II again separated in to two groups i.e., cluster I and II. Cluster I containing seven accessions those are of Raichur (Cf6), Chikmagalure (Cf10), Hassan (Cf11), Tumkur (Cf12), Bellary (Cf7), Davangere (Cf8) and Chitradurga (Cf9). Cluster II contained accession collected from Gulbarga (Cf1), Bagalkot (Cf3), Haveri (Cf4) and Gadag (Cf5) plants. These results clearly showed that *C. adscendens* var. *fimbriata* has two stocks of diversity. The plants grow at north interior Karnataka show variation with plants of south Karnataka at molecular level.

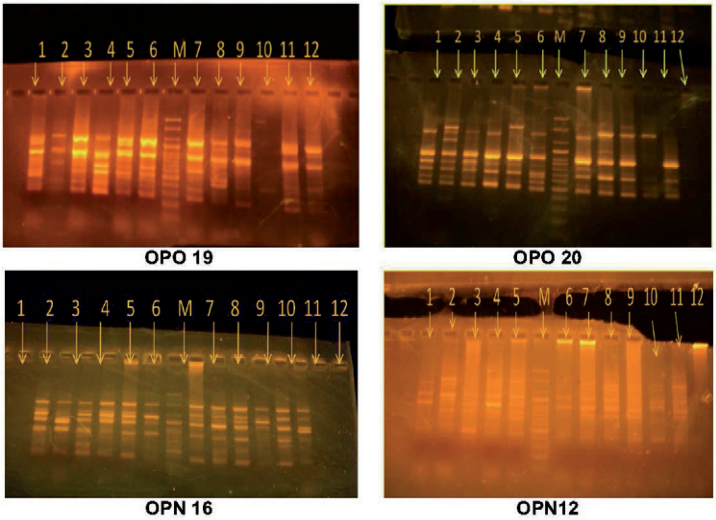

1 2 3 4 5 6 M 7 8 9 10 11 12
OPO 19
1 2 3 4 5 6 M 7 8 9 10 11 12
OPO 20
1 2 3 4 5 6 M 7 8 9 10 11 12
OPN 16
1 2 3 4 5 M 6 7 8 9 10 11 12
OPN12

Fig. 1: The number of bands yielded by primers in RAPD Analysis (M – Marker; 1 - Gulbarga 2 – Bijapura; 3 – Bagalkot; 4 – Yadgiri; 5 – Gadag; 6 – Raichur; 7 – Ballari; 8 – Davanagere; 9 – Chitradurga; 10 – Chikmagalore; 11 - Hassan and 12 – Tumkur)

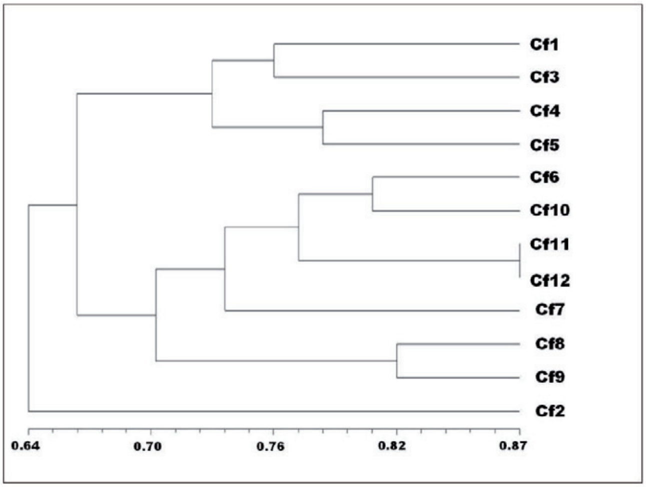


Fig. 2: Dendrogram of 12 accessions of *Caralluma fimbriata* (Cf) using RAPD markers (Cf1 - Gulbarga, Cf2 - Bijapur, Cf3 - Bagalkot, Cf4 - Haveri, Cf5 - Gadag, Cf6 - Raichur, Cf7 - Bellary, Cf8 - Davangere, Cf9 - Chitradurga,Cf10 - Chickmagalur , Cf11 - Hassan and Cf12 - Tumkuri)

Genetic diversity based on ISSR markers

Among ten primers of ISSR, five primers such as HB10, HB11, HB12, 844A, 844B showed good banding pattern. A total of 28 bands were observed with 17 polymorphic bands with 60.7 per cent polymorphism. Molecular weight of the bands ranged between 100-3000bp visualized using 1kb DNA ladder (Figure3). Primers HB10 and 844A showed same number of polymorphic bands. HB11 and HB12 showed 3 polymorphic bands (Table 4). A similarity of polymorphism had previously been reported using ISSR markers in other medicinal and aromatic plants, such as *Artemisia herba-alba* (Mohsen and Ali, 2008), *Changium smyrnioides* (Qiu et al., 2004) and *Tribulus terrestris* (Sarwat et al., 2008).

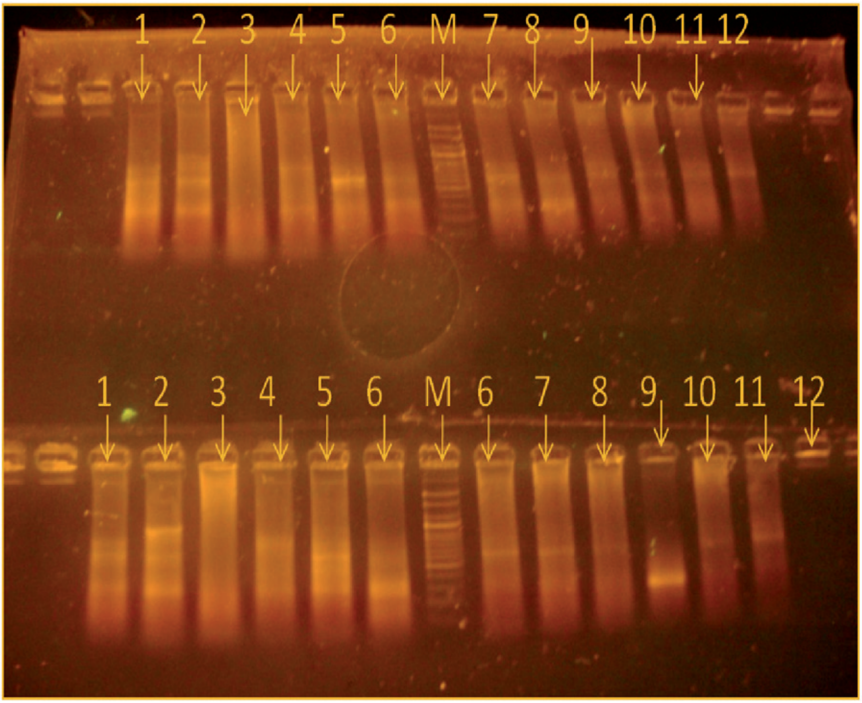


Fig. 3: ISSR analysis (HB11 and HB 12) primer (M – Marker; 1 - Gulbarga 2 – Bijapura; 3 – Bagalkot; 4 – Yadgiri; 5 – Gadag; 6 – Raichur; 7 – Ballari; 8 – Davanagere; 9 – Chitradurga; 10 – Chikmagalore; 11 - Hassan and 12 – Tumkur).

Table 4: Total number of amplified fragments and the number of polymorphic bands generated by PCR using selected ISSR Primers.

Primer	**Total of bands (a)**	**Monomor-phic bands**	**Polymorphic bands (b)**	**Polymorphism (%) (b/a x 100)**
HB10	6	2	4	66.6
HB11	5	2	3	60
HB12	5	2	3	60
844A	6	2	4	66.6
844B	6	3	3	50
Total	**28**	**11**	**17**	
Mean	**5.6**			**60.7**

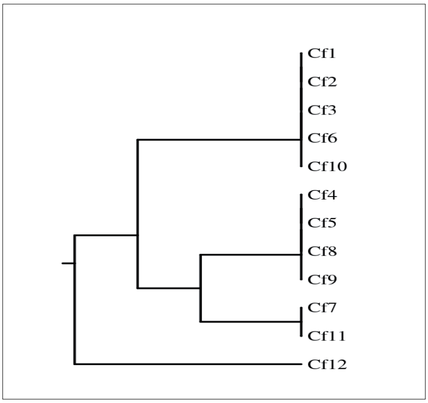


Fig. 4: Dendrogram showing the cluster analysis of 12 accessions of *Caralluma fimbriata* using ISSR markers.

Genetic similarity indices of ISSR markers ranged from 60 per cent to 75 per cent. The banding patterns were representing 12 accessions into two clades A and B. The accessions collected from Tumkur district has shown to be slight different from all other plants and was separated and formed an isolated clade A. Other eleven plants represented clade B was divided into two clusters C and D. Cluster C contains plant samples from Gulbarga, Bijapur, Bagalkot, Raichur and Chickmagalur; with similarity of 75 per cent. Cluster D is again divided into two groups E and F. Group E consists of four plant samples from Haveri, Gadag, Davangere and Chitradurga. Group F consists of two plant samples from Raichur and Hassan are grouped together with a similarity of 70 per cent in the dengrogram Figure 4.

In conclusions, *Caralluma fimbriata* collected from different part of Karnataka were observed for their morphological and genetic diversity analysis by RAPD and ISSR marker techniques using different primers. The banding pattern of both RAPD and ISSR grouped different plants in different clusters. The plant collected from Gulbarga and Bagalkot both were grouped together and showed most similarity compared to other *Caralluma fimbriata* varieties collected from other places of Karnataka. The genetic variability in a gene pool is normally considered as being the major resource available to breeders and conservationists for germplasm improvement programmes.

References

Aras, S., Polat, J.B., Cansaran, D. and Soylemezo Lu, G. (2005) RAPD analysis of genetic relations between Buzgulu grape cultivars (*Vitis vinifera*) grown in different parts of Turkey. Acta Biologica Cracoviensia 47: 77-82.

Budak, H., Shearman, R.C., Parmaksiz, I., Gaussoin, R.E., Riosdan, T.P. and Dweikat, I. (2004) Molecular characterization of buffalograss germplasm using sequence-related amplified polymorphism markers. Theoretical and Applied Genetics 108: 328-334.

Deepak, D., Srivastav, S. and Khare, A. (1997) Pregnane glycosides. Progress in the chemistry of organic natural products. 71: 169-325.

Esposito, M.A., Martin, E.A., Cravero, V.P. and Cointry, E. (2007) Characterization of pea accessions by SRAP’s markers. Scientia Horticulturae 113: 329-335.

Ferriol, M., Pico, B., Fernandex de Cordova, P. and Neuz F. (2004) Molecular diversity of a germplasm collection of squash (*Cucurbita moschata*) determined by SRAP and AFLP markers. Crop Science 44: 653-664.

Hameed, A., Mali, S.A., Iqbal, N., Arshad, R. and Farooq, S. (2004) A rapid (100 min) method for isolating high yield and quality DNA from leaves, roots and coleoptile of wheat (*Triticum aestivum* L.) suitable for apoptotic and other molecular studies. International Journal of Agricultural Biology 6(2): 383-387.

Ibrahim, R.I.H. (2011) A modified CTAB protocol for DNA extraction from young flower petals of some medicinal plant species. Genetic Conservation 10(40): 165-182.

Jabbarzadeh, Z., Khosh-Khui, M., Salehi, H. and Saberivand, A. (2009) Optimization of DNA extraction for ISSR studies in seven important rose species of Iran. American-Eurasian Journal of Sustainable Agriculture 3(4): 639-642.

Jaccard P. (1908) Nouvelles rechearches sur la distribution lorale. Bulletin de La Society.Vaudoise Sciences Naturelles 44: 223-270.

Karuppusamy, S., Ugraiah, A. and Pullaiah, T. (2013) *Caralluma* (*Sensu lato*) - Antiobesity Plants. Regency Publications, New Delhi, India.

Li, G. and Quiros, C.F. (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*. Theoretical and Applied Genetics 103: 455-461.

Loveless, M.D. and Hamrick, J.L. (1984) Ecological determinants of genetic structure in plant populations. Annual Review of Ecology and Systematics 15: 65-95.

Mabberley, D.J. (1993) The Plant Book. Cambridge University Press, Cambridge.

Mohsen, H. and Ali, F. (2008) Study of genetic polymorphism of *Artemisia herba alba* from Tunisia using ISST markers. African Journal Biotechnology 7: 44-50.

Nair, S. and Keshavachandran, R. (2006) Molecular diversity in Chakkarakolli (*Gymnema sylvestre* R. Br) assessed through isozme and RAPD analysis. Journal of Tropical Agriculture 44: 31-36.

Pejic, I., Ajmone-Marsan, P., Morgante, M., Kozumplick, V., Castiglioni, P., Taramino, G. and Motto M. (1998) Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. Theoretical and Applied Genetics 97: 1248-1255.

Qiu, Y.X., Hong, D.Y., Fu, C.X. and Cameron, K.M. (2004) Genetic variation in the endangered and endemic species *Changium smyrnioides* (Apiaceae). Biochemical Systematic Ecology 32: 583–596.

Radhakrishnan, R., Zakaria, M., Aslam, M.V., Liu, X.M., Chan, K. and Habibullah, M. (1999) Antihyperglycemic effects of *Caralluma arabica* in diabetic mice. Journal of Pharmacy and Pharmacology 51: 116.

Rebecca, K., Tony, R., Srinivas, S.K., Mario, V., Rajendran, R. and Anura, V.K. (2007) Effect of *Caralluma fimbriata* extract on appetite food intake and anthropometry in adult Indian men and women. Appetite 48: 338-344.

Rohlf, F.J. (2000) NTSYS-pc: Numerical taxonomy and multivariate analysis system. Version 2.02. Exeter Software, Setauket, New York.

Sarwat, M., Das, S. and Srivastava, P.S. (2008) Analysis of genetic diversity through AFLP, SAMPL, ISSR and RAPD markers in *Tribulus terrestris*, a medicinal herb. Plant Cell Report 27: 519–528.

Shinde, V.M., Dhalwal, K., Mahadik, K.R., Joshi, K.S. and Patwardhan, B.K. (2007) RAPD analysis for determination of components in herbal medicines. Evidenced-based Complementary and Alternative Medicine 4(1): 21-23.

Siddiqui, A.A., Ahmad, A. and Dongra, S. (2000) Development in the Chemistry and Pharmocology of *Gymnema sylvestre*. Journal of Medicinal Aromatic Plant Science 22: 223-231.

Welisch, J. and Mc Chelland. M. (1992) PCR-amplified length polymorphism in tRNA intergenic spaces for categorizing *Staphylococci*. Molecular Microbiology 6: 1673-1680.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalsk, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers Nucleic Acids Research 18: 6531–6535.

*****

Plant DNA Barcoding and Phylogenetics, Pages 269-278
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

17 Barcoding of Transgenes in GM Plants

A. Bittsánszky, G. Gyulai, R.P. Malone, I. Bock, L.Y. Murenets, M. Czakó, T. Kömíves and H. Rennenberg

Introduction

Transgenes represent genetic markers artificially introduced by mankind motivated to improve crops. Detection of the marker in the Genetically Modified Organism individual, and its vegetative or sexual progenies, and monitoring it in test and cultivated populations as well as in exposed non-target organism populations is of fundamental and practical importance. The genetically modified state of an organism, i.e. the presence of the transgene, is verified essentially by DNA barcoding. Selecting the DNA sequence for barcode is straightforward because usually a known sequence is introduced. The introduction of genes, self or foreign, into plants, had prerequisites. The ability to select and or identify desired genotypes in cells, tissues or intact plants laid the foundations for application of genetic transformation of plants. The Biological Research Centre (Szeged, Hungary) can be considered as the *Genius Loci* (Smil, 2001) of the current plant biotechnology since methodologies of plant cell line selections for chloroplast mutants (Maliga et al., 1973; Garab et al., 1974), cell fusion (Dudits et al., 1976), genetic transformation (Márton et al., 1979; Koncz et al., 1984), bacterial nitrogen fixation (Kiss et al., 1980), and artificial chromosomes (Hadlaczky, 2001) were either fundamentally developed or highly improved here.

The first stable higher plant mutant, the antibiotic (i.e. streptomycin, SR) resistant (i.e. mutant) tobacco (SR1) was selected (Maliga et al., 1973) *in vitro*, followed by the selection (Maliga et al., 1975) and identification of SR1A15 (Sváb and Maliga, 1986) the first double mutant of higher plants, the albino (chloroplast) tobacco (Maliga et al., 1975; Páy and Smith, 1988). Later, as the early forms of gene transfer, protoplast cell fusion plants (i.e. cybrids) were developed in several laboratories (Kao and Michayluk, 1974; Melchers and Labib, 1974; Power et al*.,* 1976; Dudits et al*.,* 1977; Medgyesi et al*.,* 1985).

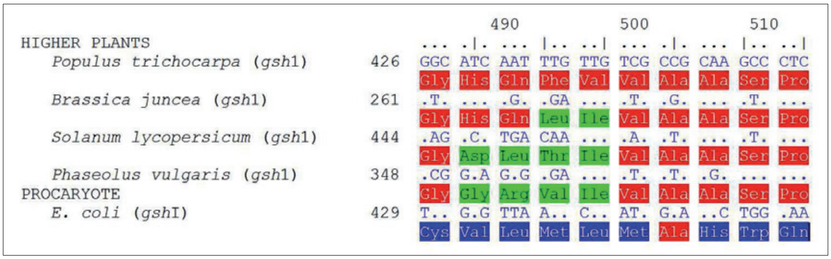
Alternatives to the conventional haploid genome transfer (i.e, pollination), the technologies of gene transfer resulting in stable transgenic crops (i.e. GM - genetically modified, or GMO - genetically modified organism), were developed in four laboratories at the same time in 1983: GM *Nicotiana plumbaginifolia* (resistant to the antibiotic kanamycin) (Bevan et al*.,* 1983), other tobacco lines resistant to kanamycin and methotrexate (a drug used to treat cancer and rheumatoid arthritis) (Herrera-Estrella et al., 1983), GM petunia resistant to kanamycin (Fraley et al*.,* 1983), and GM sunflower transformed by phaseolin gene isolated from bean (Murai et al., 1983).

The first field trial of GM cotton was carried out in 1990, followed by the first FDA-approved (Unites States Food and Drug Administration) transgenic food of Flavr-Savr tomato in 1994 (Bruening and Lyons, 2000). A series of further GM crops were released in 1995, such as the canola oil seed rape (*Brassica napus*) with modified oil composition (Calgene), Bt (*Bacillus thuringiensis*) corn (Ciba-Geigy) resistant to the herbicide bromoxynil (Calgene), Bt cotton (Monsanto), GM soybeans resistant to herbicide glyphosate (Monsanto); virus-resistant squash (Asgrow), and delayed ripening tomatoes (DNAP, Zeneca/Peto and Monsanto) (Conner et al., 2003). Later, a series of woody plants were also bred by genetic transformation (Arisi et al*.,* 1997; Noctor et al*.,* 1998; Bittsánszky et al., 2005; Gyulai et al*.,* 2012, 2014).

Here we present a case study of barcoding (i.e. detecting and monitoring GM plants) the CaMV-35S-*gsh*I poplar (*Populus x canescens*) with techniques useful for both developing GM plants and for anti-GM purposes.

Barcoding of CaMV-35S-*gsh*I transgene in GM poplar (*P. x canescens)*

The phytoextraction and remediative capacity of poplars was improved significantly by genetic transformation of *Populus* x *canescens* (*P. tremula* x *P. alba*) to overexpress the bacterial gene coding for γ-glutamylcysteine synthetase (γ-ECS, EC 3.2.3.3), which is the rate-limiting regulatory enzyme in the biosynthesis of the ubiquitous tripeptide thiol compound glutathione (GSH, γ-L-glutamyl-L-cysteinyl-glycine) (Arisi et al., 1997; Noctor et al., 1998). Here we show how *gsh*I transgene is detected by using *gsh*I-specific PCR primers (Koprivova et al., 2002; Gyulai et al., 2005). The sequence differences between the eukaryotic plant *gsh*1 gene and the prokaryotic *gsh*I transgene of *E. coli* (Figure 1a) made it feasible to design transgene specific PCR primers.



```
                                          490             500             510
HIGHER PLANTS                         ... .|. ... |.. ..| ... .|. ... |.. ..|
   Populus trichocarpa (gsh1)    426  GGC ATC AAT TTG TTG TCG CCG CAA GCC CTC
                                      Gly His Gln Phe Val Val Ala Ala Ser Pro
   Brassica juncea (gsh1)        261  .T. ... .G. .GA ... .T. .G. ... .T. ...
                                      Gly His Gln Leu Ile Val Ala Ala Ser Pro
   Solanum lycopersicum (gsh1)   444  .AG .C. TGA CAA ... .A. .T. ... .T. ...
                                      Gly Asp Leu Thr Ile Val Ala Ala Ser Pro
   Phaseolus vulgaris (gsh1)     348  .CG G.A G.G .GA ... .T. .T. .G. ... ...
PROCARYOTE                            Gly Gly Arg Val Ile Val Ala Ala Ser Pro
   E. coli (gshI)                429  T.. G.G TTA A.. C.. AT. G.A ..C TGG .AA
                                      Cys Val Leu Met Leu Met Ala His Trp Gln
```

Fig. 1a: Sequence diversities of four orthologous plant *gsh*1 genes (samples of 30 nt) and their amino acid translations (10 aa) of GSH proteins (Glutathione Synthase), compared to the non-orthologous prokaryotic *gshI*/GSI of *E. coli*. Synonymous and non-synonymous nucleotide substitutions (first rows of plant species), and the translated (by BioEdit; Hall, 1999) aa changes are indicated in boxes in different colors. The *gsh*1 of poplar, NCBI # EF148665, was downloaded, BLASTed and aligned by NCBI server (Altschul et al., 1997).

- **DNA extraction:** Total DNA samples of 0.1 g leaf tissue in each case were extracted in CTAB (cethyltrimethylammonium bromide) buffer followed by RNase-A (from bovine pancreas, Sigma, R-4875, treatment) for 30 min at 37°C. DNA samples of 10 individuals of each line were pooled in one bulk and subjected to PCR analysis.

- **Multiple sequence alignments for primer design:** Nucleotide sequences of genes *gsh*1 were downloaded from the National Center for Biotechnology Information (NCBI) databases (Altschul et al., 1997). Multiple sequence alignments were applied *in silico* with the software programs BioEdit Sequence Alignment Editor (North Carolina State University, USA) (Hall, 1999), MULTALIN (Combet et al., 2000), CLUSTAL W (Thompson et al., 1994), FastPCR (Kalendar et al., 2009), and computer program MEGA4 (Tamura et al., 2007).

- **Barcoding of the transgene:** The *gsh*I-transgene (*E. coli*, NCBI #X03954) in the transformed poplar clones was amplified by the *gsh*I specific primer 5'-ATCCCGGACGTATCACAGG-3' (position bp. 341-359 in *gsh*I) and its reverse 3'-GATGCACCAAACAGATAAGG-5' (position bp 939-920 in *gsh*I) according to Koprivova et al. (2002) and Gyulai et al. (2005).

  Hot Start PCR was combined with Touchdown PCR by using AmpliTaq Gold^{TM} Polymerase. Reactions were carried out in a total volume of 10 µl, and 25 µl (transgene detection), respectively, containing 50 ng of genomic DNA. For transgene analysis 1 x PCR buffer (2.5 mM $MgCl_2$), dNTPs (200 µM each), 20 pmol of each primer and 0.5 U of *Taq* polymerase were used (Toldi et al., 2002).

  Touchdown PCR was performed by decreasing the annealing temperature from 66^{o}C to 56^{o}C by 0.7^{o}C/ 30s increments per cycle in each of the initial 12 cycles (PE 9700, Applied Biosystems), followed by a 'touchdown' annealing temperature for the remaining 25 cycles at 56^{o}C for 30 s with a final cycle of 60^{o}C for 45 min or 72^{o}C for 10 min (transgene detection) and hold at 4^{o}C. A minimum of three independent DNA preparations of each sample were used. Amplifications were assayed by agarose (1.8 %, SeaKem LE, FMC) gel electrophoresis (Owl system), stained with ethidium bromide (0.5 ng/µl) after running at 80V in 1X TAE buffer (Gyulai et al., 2005). Each successful reaction with scorable bands was repeated at least twice. Transilluminated gels were analyzed by the ChemiImager v 5.5 computer program (Alpha Innotech). A negative control which contained all the necessary PCR components except template DNA was included in the PCR runs.

  Double strand breaks (DSBs) of DNA as the initial events of recombination occur not only in the meiotic but also in the somatic cells (Puchta, 1999), which can cause transgene elimination (Figure 1b). In the present study, the *gsh*I transgene was found to be stably incorporated for at least eighteen years (Arisi et al., 1997; Noctor et al., 1998), in the tested poplar lines (*ggs*11 and *lgl*6), and no transgene elimination or segregation was detected, which can occur during several cycles of micropropagation *in vitro* (Figure 1c). The RT-qPCR analysis (see Chapter 6) confirmed that the transgene was not lost by revealing the high expression levels of the transgene CaMV-35S-*gsh*I in poplar exposed to herbicides (Bittsánszky et al., 2006; Gyulai et al., 2008).

  In conclusions, by means of molecular barcoding the transgenes, either coding or reporter genes, can be detected in the genetically transformed GM plants for both GM and anti-GM purposes. The sequence differences between the foreign gene and the resident genes make it feasible to design GMO-specific barcodes. We should emphasize that as opposed to more involved southern blotting and mapping of transgenes barcoding is simple, cost effective and possibly accessible to the public and organizations through specialized commercial laboratories.

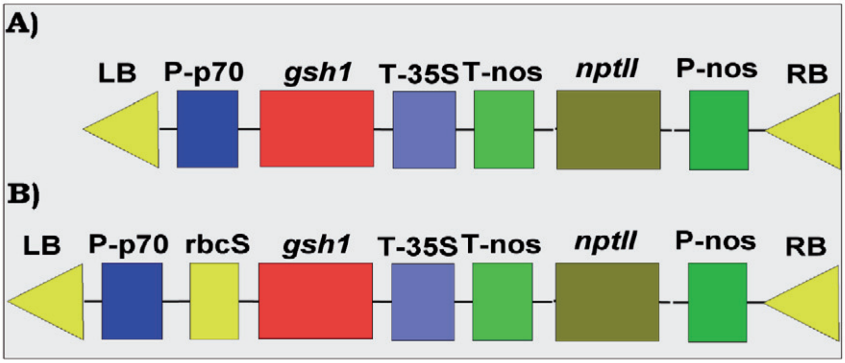


Fig. 1b: Binary vector construct for *Agrobacterium tumefaciens* mediated transformation (Arisi et al., 1997; Noctor et al., 1998) to overexpress the *gsh*I transgene (γ-glutamylcysteine synthetase; syn.: γ-ECS, EC 3.2.3.3) in poplar (*Populus x canescens*) either in the cytosol (*ggs*11): (A) LB, left border; P70 CaMV (Cauliflower mosaic virus) 35S promoter with double-enhancers; T, poly(A); nos, nopaline synthase; nptII, neomycin phosphotransferase; RB, right border; or in the chloroplast (*lgl*6): (B) directed by transit peptide of pea *rbc*S (Koprivova et al., 2002; Gyulai et al., 2005; Bittsánszky et al., 2006).

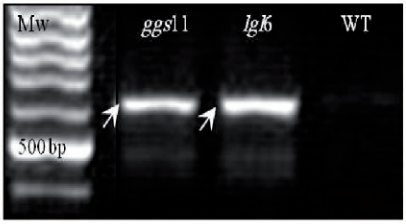


Fig. 1c: PCR detection of the partial sequence (598 bp) (arrows) of the GM (CaMV-35S-*gsh*I-transgene cloned from *E.* coli; NCBI # X03954) poplar (*Populus* x *canescens*) clones of *ggs*11 (cyt-ECS) and *lgl*6 (chl-ECS), and the non-transformed (WT) clone (0.8% agarose gel). Primer pair was 5'-ATCCCGGACGTATCACAG G-3' (positions on 341-359 nt of *gsh*I) and 3'-GATGCACCAAACAGATAA GG-5' (position on 939-920 nt of *gsh*I) (Gyulai et al., 2005).

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipmand, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.

Arisi, A.C.M., Noctor, G., Foyer, C.H. and Jouanin, L. (1997) Modification of thiol contents in poplar *(Populus tremula* x *P. alba*) overexpressing enzymes involved in glutathione synthesis. Planta 203: 362–372.

Bevan, M.W., Flavell, R.B. and Chilton, M.D. (1983) A chimaeric antibiotic resistance gene as a selectable marker for plant cell transformation. Nature 304: 184–187.

Bittsánszky, A., Kőmíves, T., Gullner, G., Gyulai, G., Kiss, J., Heszky, L., Radimszky, L. and Rennenberg, H. (2005) Ability of transgenic poplars with elevated glutathione content to tolerate Zinc (2+) stress. Environment International 31: 251–254.

Bittsánszky, A., Gyulai, G., Humphreys, M., Gullner, G., Csintalan, Zs., Kiss, J., Szabó, Z., Lágler, R., Tóth, Z., Rennenberg, H., Heszky, L. and T Kőmíves (2006) RT-PCR analysis and stress response capacity of transgenic *gsh*I-poplar clones (*Populus* x *canescens*) in response to paraquat exposure. Zeitschrift fur Naturforschung C 61: 699–730.

Bruening, G. and Lyons, J.M. (2000) The case of the FLAVR SAVR tomato. California Agriculture 54: 6–7.

Conner, A.J., Glare, T.R. and Nap, J.P. (2003) The release of genetically modified crops into the environment. Part II. Overview of ecological risk assessment The Plant journal : for cell and molecular biology33: 19–46.

Combet, C., Blanchet, C., Geourjon, C. and Deléage, G. (2000) NPS@: Network Protein Sequence Analysis. Trends in biochemical sciences 25: 147–150.

Dudits, D., Raskó, I., Hadlaczky, G. and de-Faria, A.L. (1976) Fusion of human cells with carrot protoplasts induced by polyethylene glycol. Hereditas 82: 121–124.

Dudits, D., Hadlaczky, G., Levi, E., Fejér, O. an Lázár, G. (1977) Somatic hybridization of *Daucus carota* and *D. cepillifolius* by protoplast fusion Theoretical and Applied Genetics 51: 127–132.

Fraley, R.T., Rogers, S.G., Horsch, .RB., Sanders, P.R., Flick, J.S., Adams, S.P., Bittner, M.L., Brand, L.A., Fink, C.L., Fry, J.S., Galluppi, G.R., Goldberg, S.B., Hoffmann, N.L. and Woo, S.C. (1983) Expression of bacterial genes in plant cells. Proceedings of the National Academy of Sciences USA 80: 4803–4807.

Garab, G., Horváth, G. and Faludi-Daniel, A. (1974) Resolution of the fluorescence bands in greening chloroplasts of maize. Biochemical and Biophysical Research Communications 56: 1004-1009.

Gyulai, G., Humphreys, M., Bittsánszky, A., Skøt, K., Kiss, J., Skøt, L., Gullner, G., Heywood, S., Szabó, Z., Lovatt, A., Radimszky, L., Roderick, H., Abberton, M., Rennenberg, H., Kőmíves, T. and Heszky, L. (2005) AFLP analysis and improved phytoextraction capacity of transgenic *gsh*I-poplar clones (*Populus canescens* L.) *in vitro*. Zeitschrift für Naturforschung C 60: 300–306.

Gyulai, G., Tóth, Z., Bittsánszky, A., Szabó, Z., Gullner, G., Kiss, J., Kőmíves, T. and Heszky, L. (2008) Gene up-regulation by DNA demethylation in 35S-*gsh*I-transgenic poplars (*Populus* x *canescens*). In: Wolf, T. and Koch, J. (Eds.) Genetically Modified Plants: New Research Trends. Nova Science Publisher, Inc. USA, Chapter 8, pp. 173–191.

Gyulai, G., Bittsánszky, A., Gullner, G., Heltai, Gy., Pilinszky, K., Molnár, E. and Kömíves, T. (2012) Gene reactivation induced by DNA demethylation in Wild Type and 35S-*gsh*I-*rbc*S transgenic poplars (*Populus x canescens*). Novel plant sources for phytoremediation. Journal of Chemical Science and Technology 1: 9–13.

Gyulai, G., Bittsánszky, A., Szabó, Z., Waters Jr, L., Gullner, G., Kampfl, Gy., Heltai, Gy. and Kőmíves, T. (2014) Phytoextraction potential of wild type and 35S-gshI transgenic poplar trees (*Populus* × *canescens*) for environmental pollutants herbicide paraquat, salt sodium, zinc sulfate and nitric oxide *in vitro*. International journal of phytoremediation 16: 379–396.

Hadlaczky, G. (2001) Satellite DNA-based artificial chromosomes for use in gene therapy. Current Opinion in Molecular Therapeutics 3: 125–32.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98.

Herrera-Estrella, L., Depicker, A., van Montagu, M. and Schell, J. (1983) Expression of chimaeric genes transferred into plant cells using a Ti-plasmid-derived vector. Nature 303: 209–213.

Kalendar, R., Lee, D. and Schulman, A.H. (2009) FastPCR Software for PCR Primer and Probe Design and Repeat Search. Genes, Genomes and Genomics 3: 1–14.

Kao, K.N. and Michayluk, M.R. (1974) A method for high frequency intergeneric fusion of plant protoplasts. Planta 115: 355–367.

Kiss, G.B., Dobo, K., Dusha, I., Breznovits, A., Orosz, L., Vincze, E. and Kondorosi, A. (1980) Isolation and characterization of an R-prime plasmid from *Rhizobium meliloti*. Journal of Bacteriology141: 121–128.

Koncz, C., Kreztzaler, F., Kálmán, Z. and Schell, J. (1984) A simple method to transfer, integrate and study expression of foreign genes, such as chicken ovalbumin and alpha-actin in plant tumors. The EMBO Journal 3: 1029-1037.

Koprivova, A., Kopriva, S., Jager, D., Will, B., Jouanin, L. and Rennenberg, H. (2002) Evaluation of transgenic poplars over-expressing enzymes of glutathione synthesis for phytoremediation of cadmium. Plant Biology 4: 664–670.

Maliga, P., Sz-Breznovits, A. and Márton, L. (1973) Streptomycin-resistant plants from callus culture of haploid tobacco. Nature New Biology 244: 29–30.

Maliga, P., Sz-Breznovits, A., Márton, L. and Joo, F. (1975) Non-Mendelian streptomycin-resistant tobacco mutant with altered chloroplasts and mitochondria. Nature 255: 401–402.

Márton, L., Wullems, G.J., Molendijk, L. and Schilperoort, R.A. (1979) *In Vitro* transformation of cultured cells from *N. tabacum* by *Agrobacterium tumefaciens*. Nature 277: 129–131.

Medgyesi, P., Fejes, E. and Maliga, P. (1985) Interspecific chloroplast recombination in a *Nicotiana* somatic hybrid. Proceedings of the National Academy of Sciences USA 82: 6960–6964.

Melchers, G. and Labib, G. (1974) Somatic hybridization of plants by fusion of protoplasts. I. Selection of light resistant hybrids of "haploid" light sensitive varieties of tobacco. Molecular Genetics and Genomics135: 277–294.

Murai, N., Sutton, D.W., Murray, M.G., Slightom, J.L., Merlo, D.J., Reichert, N.A., Sengupta-Gopalan, C., Stock, C.A., Barker, R.F., Kemp, J.D. and Hall, T.C. (1983) Phaseolin gene from bean is expressed after transfer to sunflower via tumor-inducing plasmid vectors. Science 222: 476–482.

Noctor, G., Arisi, A.C.M., Jouanin, L., and Foyer, C.H. (1998) Manipulation of glutathione and amino acid biosynthesis in the chloroplast. Plant Physiology 118: 471–482.

Páy, A. and Smith, M.A. (1988) A rapid method for purification of organelles for DNA isolation: self-generated percol gradients. Pant Cell Reports 7: 96-99.

Power, J.B., Frearson, E.M., Hayward, C., George, D., Evans, P.K., Berry, S.F. and Cocking, E.C. (1976) Somatic hybridisation of *Petunia hybrida* and *P. parodii*. Nature 263: 500–502.

Puchta, H. (1999) Double-strand break-induced recombination between ectopic homologous sequences in somatic plant cells. Genetics 152: 1173–1181.

Smil, V. (2001) Genius loci - The twentieth century was made in Budapest. Nature 409: 21.

Sváb, Z. and Maliga, P. (1986) *Nicotiana tabacum* mutants with chloroplast encoded streptomycin resistance and pigment deficiency. *Theoretical* and Applied Genetics 72: 637–643.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680.

Toldi, O., Tóth, S., Ponyi, T., Scott, P. (2002) An effective and reproducible transformation protocol for the model resurrection plant *Craterostigma plantagineum* Hochst. Plant Cell Reports 21: 63–69.

*****

Plant DNA Barcoding and Phylogenetics, Pages 279-298
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)
© 2015 Lambert Academic Publishing, Germany

18 Molecular Barcoding of Sex-Linked DNA Markers of Dioecious Plants

G. Gyulai, B. Kerti and M.A. Ali

Introduction

Dioecious plant species account for 6% (14,620) of the total about 240,000 flowering plant species, of both gymnosperms and angiosperms, in 7.1% (959 of 13,500) of genera and 43% (157 of 365) of families (Renner and Ricklefs, 1995; Cronquist, 1988). None of the ancient gymnosperm specis (all the extant 860 'coniferous-like' species) has bisexual (syn.: hermaphrodite) flower (Gyulai et al., 2012), they have only unisexual either staminate (male ♂) or pistillate (female ♀) flower that develop either on the same plant, i.e. monoecious species (e.g. *Pinus*), or on separate plants, i.e. dioecious species (e.g. *Taxus, Ginkgo*) Linkies et al., 2010). This indicates the ancestral property of the unisexual flowers, however, angiosperms with bisexual flowers tend to 'mutate back' during the evolution towards monoecy and dioecy (i.e. sexual dimorphism) coupled with sex chromosome development to promote outcrossing and for escaping inbreeding (McClung, 1901).

By selecting sex-linked DNA markers, several dioecious crops gained huge agronomical importance, e.g. all-male/supermale *Asparagus* has higher yield then females, male *Taxus* bears red berries for ornamental use, female *Phoenix* produces dates, and female *Cannabis* gives flower used for brewing beer, etc. Here, we overview those of PCR based molecular markers which are used to determine male- and female individuals of dioecious plants.

Plant dioecy

Similar to animals, the ratio between individuals of male vs. female - in plants they are pistillate (♀) and staminate (♂) - of dioecious plants is not necessarily 50% : 50%, as male and female individuals are differentially affected by biotic and abiotic environmental stresses, which bias the equilibrium expectation of 1:1 sex ratio (Zhao et al., 2012). Several studies have indicated that gender-specific physiological responses of dioecious plants are different to environmental stresses, like rising air temperature and CO_2 concentration of global climate change. These may lead to dramatic changes in sex ratio and consequently the ranges of distribution and growth pattern in the environment (Tognetti, 2012). In general, male plants show higher level of photosynthetic carbon fixation, and consequently higher fresh mass production (Wang and Curtis, 2001; Xu et al., 2008). Male *Salix arctica* also had a significantly higher photosynthetic rate than females at elevated CO_2 temperature (Jones et al., 1999). Male *Populus tremuloides* trees also showed higher photosynthesis activity than females throughout the growing season, regardless of CO_2 concentrations (Wang and Curtis, 2001). Male *Populus cathayana* trees showed higher drought and chilling tolerance then females (Zhang et al. 2011). The ratio of female to male *Populus tremuloides* trees on the Front Range in Colorado was 1.27 below 2450 m of elevation, but was only 0.56 above 2900 m (Grant and Mitton, 1979), which indicates more cold tolerance of male trees.

The growth form, clonality, fleshy fruits, pollen and seed dispersal vectors, and the possession of sex chromosomes also affect sex-ratio and result in male-biased flowering ratio, which is twice as much as female-biased ratios (Field et al., 2013). Male-bias was found to associate with long-lived growth forms of trees, biotic seed dispersal, and fleshy fruits. Female-bias associated with clonality, especially for herbaceous species, and abiotic pollen dispersal. In addition, there are species with extreme female-biased sex ratios, e.g. seagrasses of *Phyllospadix scouleri* and *P. serrulatus* shows 90% female bias (Shelton, 2010). Sex ratio of *Phyllospadix* species were found even among seedlings but became female-biased at later life stages, indicating that sex ratio is driven by male-biased mortality (Shelton, 2010).

Some dioecious plants, similar to animals, show differences not only in DNA fragments (i.e. markers) but also in sex chromosoemes. Plant sex chromosomes are not as rare as previously assumed (Charlesworth, 2002). By families of angiosperms, Amaranthaceae, Aracaceae, Asparagaceae, Cannabaceae, Caricaceae, Caryophyllaceae, Cucurbitaceae, Polygonaceae, Rosaceae and Santalaceae comprise species with sex chromosomes (Field et al., 2013). Evolutionary, plant sex chromosomes may evolved some million years ago, e.g. in *Silene,* it might have developed about 5–10 Mya ago, however, in papaya (*Carica papaya*) it might have evolved more recently. To compare, sex chromosomes in mammals may developed 300 Mya ago (Lahn and Page, 1999), and were discovered only about a century ago in human (XX diploid homomorphic sex-chromosomes for females, and XY diploid heteromorphic sex chromosome for males) (McClung, 1901). To compare, flowering plants may began appearing in fossil records at about of 124,6 million year ago (Mya) (Sun et al., 2002), however angiosperms show about 158-179 MYA evolution based on DNA sequence data (Wikström et al., 2001).

Heteromorphic sex chromosomes for males (XY) and homomorphic for females (XX) were described, for e.g*. Cannabis sativa, Humulus lupulus* and *H. japonicus, Silene latifolia S. dioica and S. indica* and *Coccinia indica*. Some of them have a dosage compensation system, e.g. in *Humulus*. Homomorphic sex chromosomes were characterized cytologically in *Actinidia deliciosa, A. chinensis; Asparagus officinalis, Antennaria dioica, Carica papaya, Vasconcellea species, Silene otites, Spinacia oleracea, Bryonia multiflora, Ecballium elaterium, Dioscorea tokoro, Thalictrum species, Fragaria species* and in *Vitis* (Ming et al., 2007). Female-bias mainly occurs in species with sex chromosomes and there is some evidence for a greater degree of bias in those with heteromorphic sex chromosomes (Field et al., 2013).

There are plant species with more complex sex pattern including gynodioecious vs. gynomonoecious, androdioecious vs. andromonoecious, and tridioecious vs. trimonoecious plants (Ainsworth 2000; Heikrujam et al., 2015). In the case of *Asparagus* (2n=20) there are four different individuals (i.e. genotypes) in a population, the homogametic females (XX), the heterogametic males (XY), the andromonoecious males (also XY), and the supermales (YY) (Ii et al., 2012). The crossing experiments revealed that when an andromonoecious (XY) plant is selved (XY x XY), it gives a segregation of 1 XX (with berries) : 2 XY : 1 YY; and, when a supermale (YY) plant is crossed with a female (XX) plant (YY x XX) only male (XY) plants are found in the progeny - both indicate a single dominant Mendelian gene segregation (Reamon-Büttner et al., 1998). For cultivation, male and supermale plants are desired because of the higher yield and longevity. Breeders use difficult testcrosses to identify these genopytes, however molecular barcoding by PCR technology (Table 1) provides powerful tools to identify all genotypes at seedling stage (Reamon-Büttner et al., 1998). Unlike the time consuming cytological sex determinations, PCR based markers highly facilitate sex determination in plants (Table 1).

Sex determination in common sea-buckthorn (*Hippophae rhamnoides*)

Sea-buckthorns are deciduous shrubs of *Elaeagnaceae* family, and comprising about seven species. Two of them became intensively studied and cultivated, one of them *Hippophae rhamnoides*, grows in Eurasia, the other, *H. salicifolia* is native to Himalaya. Recently, common sea-buckthorn is cultivated in Europe intensively for the nutritious berries, also used as medicine. As general, seedlings (i.e. planst developed from seeds) grow faster than the cuttings (i.e. rooted branches, or root sprouts), so the growers and nurseries prefer seedlings which need female-linked marker determination to exclude male individuals at seedling stage to reduce growing costs. Here we present a case study for the identification of gender-specific DNA marker for *H. rhamnoides* by using RAPD primer OPC-20 (Figure 1), and PCR conditions according to Sharma et al. (2010) (Table 1), however the amplified male-specific DNA fragment was about 500 bp (Figure 1). After a thorough sex determination of the seedlings, and by excluding 90% of males showing the extra RAPD ♂OPC-20_{500bp} DNA band (Figure 1) a highly productive female (90%) sea-buckthorn (*H. rhamnoides*) plantation can be developed. The more proper method using SCAR (Table 1) markers (Korekar et al., 2012) are in progress.

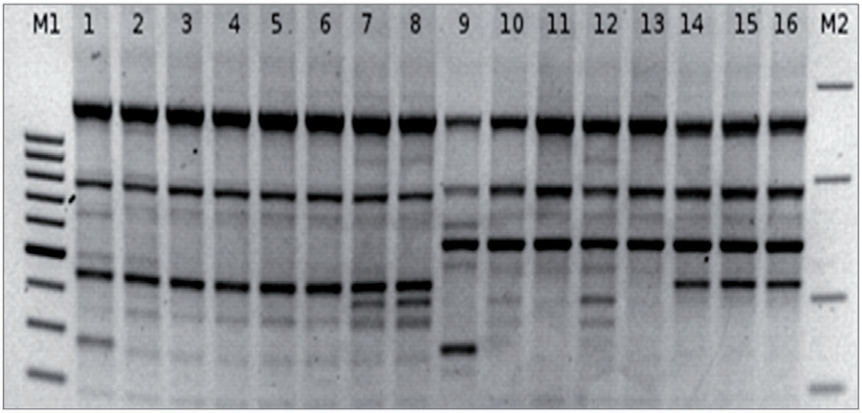


Fig. 1: Male-specific RAPD OPC-20 DNA marker (500 bp) detected in common sea-buckthorn (*Hippophae rhamnoides ssp. Carpathica*) population by using the method of Sharma et al. (2010). Females (lanes 1-8) and males (lane 9-16) are identified. M1 - Mw 100 bp DNA size marker. M2 - FastRuler Low Ramge DNA ruler. Arrowhead indicates the male specific OPC/20-911 bp DNA band in each male plant.

Table 1: Selected list of sex-linked (male/staminate/♂ *vs.* female/pistillate/♀) PCR based DNA markers of dioecious plants (Zhang et al., 1998; Milewicz and Sawicki, 2013; Heikrujam et al., 2015).

Species	**Primer types**	**Primer sequences** (5'–3')	**Fragments size (bp)**	Gender (♂ or ♀)	**References** (# NCBI)
Actinidia deliciosa	RAPD	CAGGCCCTTC (OPA-01)	1031	♀	Shirkot et al. (2002)
		AATCGGGCTG (OPA-02)	2000		
		GTGACGTAGG (OPA-08)	700		
		CAATCGCCGT (OPA-11)	2800		
		AGCCAGCGAA (OPA-16)	3000		
		GTTTCGCTCC (OPB-01)	2000		
		GATGACCGCC (OPC-05)	350	♂	
		CTCACGTTGG (OPN-01)	600		
Asparagus officinalis	RAPD	TTCACGGTGG (T35R54)(XY-YY)	1600		li et al. (2012)
		GACGGATCAG (OPC-15)	~360,~900		Jiang and Sink (1997)
	RAPD-SCAR	SCC15 (n.i.)	~600		
Aucuba japonica	RAPD-SCAR	SCAR-OPA10-424 SCAR-OPN11-1095			Maki (2009)
Borassus flabellifer	RAPD	GGTCCCTGAC (OPA-06)	600		George et al. (2007)
Calamus simplicifolius	RAPD	TCTCGCCTAC (OPAD-3)	500		Yang et al. (2005)
	RAPD-SCAR	TCTCGCCTACCTTTTACCA TCTCGCCTACAGGAACAACA/CsMale1	500		Li et al. (2010)

Carica papaya	RAPD	ACGGCGTATG (OPE-19)	2180	♀	Niroshini et al. (2000)
		GAGGATCCCT (OPF-02)	800	♂	Parasnis et al. (2000)
	ISSR	(GATA)4 (not in PCR)	5kb		Parasnis et al. (1999)
		(GACA)4	n.i.	♀	Gangopadhyay et al. (2007)
Cucumis melo	male sterility	ACCACGAGTGTCGAGAAGAA ACCACGAGTGAGGGATCTTC	644	ms-3	Park et al. (2004)
Cycas circinalis	RAPD	GTTTCGCTCC (OPB-01)	686	♂	Gangopadhyay et al. (2007) (#DQ386640)
		TGCGCCCTTC (OPB-05)	2048	♀	
	RAPD-SCAR	n.i.	255	♂	
Garcinia morella	RAPD	CAGCGACTGT (OPN-15)	634		Joseph et al. (2014) (#KJ809108*)
	RAPD-SCAR*	CAGCGACTGTTGGCGGAATG* AAAACTATGTATGTCAGCGAC	634		
Ginkgo biloba	RAPD-SCAR	CTGCTGGGACACAGTACAGAGTTTG GGGTTGTCGCCAAGGTTAT (S10)	571		Liao et al. (2009)
		CTGCTGGGACTTATAGGTCTTACTG AGATCCTATCACTGATCCGAAACAA	688	♀	

Hippophae rhamnoides	RAPD	CATCCGTGCT (OPD-15)	600	♂	Persson and Nybom (1998)
		ACTTCGCCAC (OPC-20)	911		Sharma et al. (2010)
		AATCGGGCTG (OPA-04)*	1164	♀	Korekar et al. (2012) (#JQ284019*) (#JQ284020**)
		CAAGGGCAGA (OPT-06)**	868		
	RAPD-SCAR	TATGAGCTCTCGACTGACAGCCA* CTGTTGTCCGAGATGACGCGT; HRX1	470		
		AAGTGTGGCCACCGTCGTAAGA; HRX2 ACCGTGTCGATGCACTGTGTATAG**	386		
Hippophae salicifolia	RAPD	TTGGTACCCC (OPF-11)	1190		Rana et al. (2009)
	RAPD-SCAR	TATGAGCTCTCGACTGACAGCCA* CTGTTGTCCGAGATGACGCGT; HRX1	470		Chawla et al. (2014)
Hippophae tibetana	RAPD-SCAR	TATGAGCTCTCGACTGACAGCCA* CTGTTGTCCGAGATGACGCGT; HRX1	470		Chawla et al. (2014)
Humulus lupulus	RAPD	GAAACGGGTG (OPA-07)	1700	♂	Polley et al. (1997)
		TGAGCCTCAC (OPJ-09)	1200		
		GGCGAAGGTT (OPU-08)	1400		
	ISSR	[AC]8YG*	700		Danilova and Karlov (2006)
		[CA]8GT**	500		
	ISSR-STS	GGGACTCGGTAACACAGAAAGGCA* AGCCCCACCTACACCACGACAACC	542		
		CAGTGTTTCTCTCGGGTTCTCTTG** AACCACACATAATTCCCATCTTGC	387		

Melandrium album (syn.: Silene latifolia)	RAPD	CACCGTATCC (OPD-12)	800, 980	🟩	Zhang et al. (1998)
			980	♀	Mulcahy et al. (1992)
	RAPD-SCAR (OPB-07)	GGTGACGCAGTTGTGGAGATG GGTGACGCAGACCCAAATTAT	750	♂	Zhang et al. (1998)
	RAPD-SCAR (OPD-05)	TGAGCGGACACGGGTGGGGC TGAGCGGACATTGTGAGGTTACCTCC	135		
	RAPD-SCAR (OPD-12)	TTCCCTCCTCCTTTCTCTCTC TAGAAGAAGATGGGTGATTTGG	800		
	RAPD-SCAR (OPK-02)	GCAAATGGGTTTAGTGTAGTGGTT GTCTCCGCAATTATCACACTAAGT	805		
	RAPD-SCAR (OPQ-14)	GGACGCTTCATGACCCATTTACTC GGACGCTTCAGCGGGCGGGATT	700		

	RAPD-SCAR (OPX-11)	GGAGCCTCAGGGATTAGAAAGCCT GGAGCCTCAGTACTAATAACATCA	400	🟩	
	RAPD-SACR (OPX-18)	GACTAGGTGGGATCGGCTG GACTAGGTGGCCATACTAGGA	1000	🟩	
Melandrium rubrum (syn.: Silene dioica)	RAPD	GGGTAACGCC (OPA-09)	590	🟩	Di Stilio et al. (1998)
			810	🟨 ♀	
	RAPD-SCAR (OPD-12)	TTCCCTCCTCCTTTCTCTCTC TAGAAGAAGATGGGTGATTTGG	800	🟩 ♂	Zhang et al. (1998)
	RAPD-SCAR (OPX-11)	GGAGCCTCAGGGATTAGAAAGCCT GGAGCCTCAGTACTAATAACATCA	400	🟩	
Melndrium diclinis (syn.: Silene diclinis)	RAPD-SCAR (OPB-07)	GGTGACGCAGTTGTGGAGATG GGTGACGCAGACCCAAATTAT	750	🟩	
	RAPD-SCAR (OPD-05)	TGAGCGGACACGGGTGGGGC TGAGCGGACATTGTGAGGTTACCTCC	135	🟩	

	RAPD-SACR (OPX-18)	GACTAGGTGGGATCGGCTG GACTAGGTGGCCATACTAGGA	1000		
Momordica dioica	RAPD	GGGTTCGGAA (OPA-15)	1500	♂	Patil et al. (2012)
	RAPD-SCAR (OPA-15)	CCGAACCCTAGAGAATAGCAAG TTCCGAACCCAGCCCGCTC	1500		
Myristica fragrans	RAPD	GAGTCTCAGG (OPE-11)	416	♀	Shibu et al. (2000)
Phoenix dactylifera	RAPD	GTGATCGCAG (OPA-10) TCGGCGATAG (OPA-12) GGTCTACACC (OPD-10)	490 750 800		Younis et al. (2008)
		GGTCTACACC (OPD-10)	370,675	♂	
	RAPD-SCAR (OPA-02)	TTTTGGGCTTGTCTAGCATC GTTCTGCAAAATTAAAGAGAAAAGGT	406, 354		Dhawan et al. (2013) (#JN123357)
			354	♀	
Piper longum	DD-SCAR	TTTGTATAATCAATAATCTGTGG* AAGCTTTGGTCCGGGGAGTCC	232		Manoj et al. (2008)
		ATTTCTAGGCACCATTGATGG** AAGTTTACTCTTTGAACTTGA	210	♂	

Pistacia atlantica Pistacia khinjuk	RAPD	BC_{1200} (n.i.)	1200	♀	Esfandiyari et al. (2012)
	RAPD-SCAR	GTCGTAGATGAAAACACC TAATAGAAGCCATAGA	300		
Pistacia chinensis	RAPD	GTTTCGCTCC (OPB-01)	473		Sun et al. (2014)
	RAPD-SCAR (OPY-01)	CCTGGTTGCTTGTGTTGATTAG GAGTGTCATCAAGCCATCTGTC	636		
Pistacia vera	RAPD	CCTCCAGTGT (OPO-08)	945		Hormaza et al. (1994)
Populus tomentosa	RAPD	ACCCGGTCAAC (OPD-20)	1800	♂	Hou et al. (2009)
Populus trichocarpa†	SSSR (BPTTG60) (BPTGG82) (BPCA90)	CAAGAACTCAGACATGATCAGATC CTTTGCACGTTAATAAGGAGACTG CTTGAAGAGCGAAAACTCAGCAG CTCTAAATCCAAGGTTGGTTACC CCTAGCCTTCATTCTCATTCAGC GGTTGCTAGTCAGCTTCTTACC	n.i.		Pakull et al. (2011)
Pseudocalliergon trifarium (moss)	ISSR-SCAR	GGATTGATATTGGCATTGAGT TGGAATGTCACATTGTTTAGGA	159	♀	Korpelainen et al. (2008)

Rumex nivalis	AFLP-SCAR	GTTAGAATAATCTATTTCATTTGCC TTCACCTATATCGATGACC / *Rniv*Y	150	♂	Stehlik and Blattner (2004)
Salix viminalis	RAPD	CTAGAGGCCG (UBC-354)	560	♀	Alstrom-R. et al. (1998)
	RAPD	CTAGAGGCCG (UBC-354)	549		Gunter et al. (2003)
	RAPD-SACR (UBC-354)	GAGAGGGAGGGAGATTTAAG* CGCCGTAGCAGATTGTTAATCAC	520		
	RAPD	TCACCACGGT (OPAE-08)	1300		
	RAPD-SACR (OPAE-08)	TGGTTAGGTGTCGTGATGGA** CAATCCACAATGCTTTTGA	780		
	AFLP-SCAR	CACCGAGGCATTGGAGATAAAC CACTTCTTGGATTTCTTCCCACC	n.i.		Semerikov et al. (2003)
Simmondsia chinensis	RAPD	AGGAGTCGGA (OPAL-20)	460		Hosseini et al. (2011)
		GGGCCACTCA (OPT-01)	680	♂	
	ISSR	(AG)8T (UBC807)	1200		Sharma et al. (2008)††

Trichosanthus dioica	RAPD	GTCCCGACGA (OPC-07)	567	♀	Singh et al. (2002)
	ISSR	(GTGC)4	808	♂	Nanda et al. (2013) †††

Abbreviations: DD - Differential Display (ddRT-PCR). ISSR - Inter Simple Sequence Repeats. OP - Operon Technologies Inc. RAPD - Random Aamplified Polymorphic DNA. SCAR - Sequence Characterized Amplified Region. SSR - Simple Sequence Repeats. STS - Sequence Tagged Site. *n.i.* - not indicated in the paper. † - also in *Populus alba* - P. *tremuloides* - *T. termula* x *P. tremuloides*; and all of the sequence stretch of *P. trichocarpa* genome (Tuskan et al., 2006) from 8.074.724 to 8.746.182 bp but not in *P. nigra* (Pakull et al., 2011). †† - also validated by Heikrujam et al. (2014). ††† - accessions from #JX678996 to #JX679001.

Acknowledgements

The study was funded in part by the grant of Research Centre of Excellence (17586-4/2013/TUDPOL), MKK, Szent István University, Gödöllő, Hungary.

References

Ainsworth, C. (2000) Boys and girls come out to play: The molecular biology of dioecious plants. Annals of Botany 86: 211–221.

Alstrom-Rapaport, C., Lascoux, M., Wang, Y.C., Roberts, G. and Tuskan, G.A. (1998) Identification of a RAPD marker linked to sex determination in the basket willow (*Salix viminalis* L.). Journal of Heredity 89: 44–49

Charlesworth, D. (2002) Plant sex determination and sex chromosomes. Heredity 88: 94–101.

Chawla, A., Kant, A., Stobdan, T., Srivastava, R.B., Chauhana, R.S. (2014) Cross-species application of sex linked markers in *H. salicifolia* and *H. tibetana*. Scientia Horticulturae 170: 281–283.

Cronquist, A. (1988) The evolution and classification of flowering plants, 2nd ed. Columbia University Press, New York, NY.

Danilova, T.V. and Karlov, G.I. (2006) Application of inter simple sequence repeat Euphytica 151: 15–21.

Dhawan, C., Kharb, P., Sharma, R., Uppal, S. and Aggarwal, R.K. (2013) Development of male-specific SCAR marker in date palm (*Phoenix dactylifera* L.). Tree Genetics & Genomes 9: 1143–1150.

Di Stilio, V.S., Kesseli, R.V. and Mulcahy, D.L. (1998) A pseudoautosomal random amplified polymorphic DNA marker for the sex chromosomes of *Silene dioica*. Genetics 149: 2057-2062.

Esfandiyari, B., Davarynejad, G.H., Shahriari, F., Kiani, M. and Mathe, A. (2012) Data to the sex determination in *Pistacia* species using molecular markers. Euphytica 185: 227–231.

Field, D.L., Pickup, M. and Barrett, S.C.H. (2013) Comparative analyses of sex-ratio variation in dioecious flowering plants. Evolution 67: 661–672.

Gangopadhyay, G., Roy, S.K., Ghose, K., Poddar, R., Bandyopadhyay, T., Basu, D. and Mukherjee, K.K. (2007) Sex detection of *Carica papaya* and *Cycas circinalis* in pre- flowering stage by ISSR and RAPD. Current Science 92(4): 524- 526.

George, J., Karun, A., Manimekalai, R., Rajesh, M.K. and Remya, P. (2007) Identification of RAPD markers linked to sex determination in palmyrah (*Borassus flabellifer* L.). Current Science 93(8): 1075-1077.

Grant, M.C. and Mitton, J.B. (1979) Elevational gradients in adult sex ratios and sexual differentiation in vegetative growth rates of *Populus tremuloides* Michx. Evolution 33: 914–918.

Gunter, L.E., Roberts, G.T., Lee, K., Laminer, F.W. and Tuskan, G.A. (2003) The development of two flanking markers linked to a sex determination locus in *Salix viminalis* L. Journal of Heredity 94: 185–189.

Gyulai, G., Bittsánszky, A., Malone, P.R., Király, K., Gullner, G. and Kőmíves, T. (2012) Transzgénikus nyárfák (35*S*-*gsh*I-11*ggs* és 35S-*rbcS*-*gsh*I-6*lgl*) alkalmazása a fitoremediációban I,II. Zöld Biotechnológia (ed. Dudits D) 8/3-4: 9–13; 8/5-6: 9–14.

Heikrujam, M., Sharma, K., Kumar, J. and Agrawal, V. (2014) Validation of male sex-specific UBC-8071200 ISSR marker and its conversion into sequence tagged sites marker in Jojoba: a high precision oil yielding dioecious shrub. Plant Breeding 133: 666–671.

Heikrujam, M., Sharma, K., Prasad, M. and Agrawal, V. (2015) Review on different mechanisms of sex determination and sex-linked molecular markers in dioecious crops: a current update. Euphytica 201: 161–194.

Hormaza, J.I., Dollo, L. and Polito, V.S. (1994) Identification of a RAPD marker linked to sex determination in *Pistacia vera* using bulked segregant analysis. Theoretical and Applied Genetics 89: 9–13.

Hosseini, F.S., Hassani, H.S., Arrin, M.J., Baghizadeh, A. and Mohammadi-Nejab, G. (2011) Sex determination of Jojoba (*Simmondsia hinensis* cv. Arizona) by random amplified polymorphic DNA (RAPD) molecular marker. African Journal of Biotechnology 10: 470–474.

Hou, W., Fan, J., Zhou, F. and Zhao, S. (2009) RAPD markers related to sex locus in *Populus tomentosa*. Frontiers of Forestry in China 4(2): 223–226.

Ii, Y., Uragami, A., Uno, Y., Kanechi, M. and Inagaki, N. (2012) RAPD based analysis of differences between male and female genotypes of *Asparagus officinalis*. Horticultural Science (Prague) 39: 33–37.

Jiang, C. and Sink, K.C. (1997) RAPD and SCAR markers linked to the sex expression locus M in *Asparagus*. Euphytica 94: 329–333.

Jones, M.H., MacDonald, S.E. and Henry, G.H.R. (1999) Sex- and habitat-specific responses of a high arctic willow, *Salix arctica*, to experimental climate change. Oikos 87: 129–138.

Joseph, K.S., Murthy, H.N. and Ravishankar, K.V. (2014) Development of male-specific SCAR marker in *Garcinia morella* (Gaertn.) Desr. Journal of Genetics 93: 875–878.

Korekar, G., Sharma, R.K., Kumar, R., Meenu, Bisht, N.C., Srivastava, R.B., Ahuja, P.S. and Stobdan, T. (2012) Identification and validation of sex-linked SCAR markers in dioecious *Hippophae rhamnoides* L. (Elaeagnaceae). Biotechnology letters 34: 973–978

Korpelainen, H., Bisang, I., Hedenäs, L. and Kolehmainen, J. (2008) The first sex-specific molecular marker discovered in the moss *Pseudocalliergon trifarium*. The Journal of Heredity 99(6): 581–587.

Lahn, B.T. and Page, D.C. (1999) Four evolutionary strata on the human X chromosome. Science 286: 964–967.

Li, M., Yang, H., Li, F., Yang, F., Yin, G. and Gan, S. (2010) A malespecific SCAR marker in *Calamus simplicifolius*, a dioecious rattan species endemic to China. Molecular Breeding 25: 549–551.

Liao, L., Liu, J., Dai, Y., Li, Q., Xie, M., Chen, Q., Yin, H., Qiu, G. and Liu, X. (2009) Development and application of SCAR markers for sex identification in the dioecious species *Ginkgo biloba* L. Euphytica 169: 49-55.

Linkies, A., Graeber, K., Knight, Ch. and Leubner-Metzger, G. (2010) The evolution of seeds. New Phytologist 186: 817–831.

Maki, M. (2009) Development of SCAR markers for sex determination in the dioecious shrub *Aucuba japonica* (Cornaceae). Genome 52: 231–237.

Manoj, P., Banerjee, N.S. and Ravichandran, P. (2008) Development of sex specific molecular markers in dioecious *Piper longum* L. plants by differential display. Journal of Theoretical and Applied Information Technology 4: 459–465.

McClung, C.E. (1901) Notes on the accessory chromosome. Anatomischer Anzeiger 20: 220–226.

Milewicz, M. and Sawicki, J. (2013) Sex-linked markers in dioecious plants. Plant Omics Journal 6(2): 144-149.

Ming, R., Wang, J., Moore, P.H. and Paterson, A.H. (2007) Sex chromosomes in flowering plants. American Journal of Botany 94: 141–150.

Mulcahy, D.L., Weeden, N.F., Kesseli, R. and Caroll, S.B. (1992) DNA probes for the Y-chromosomes of *Silene latifolia*, a dioecious angiosperm. Sexual Plant Reproduction 5: 86–88.

Nanda, S., Kar, B., Nayak, S., Jha, S., Joshi, R.K. (2013) Development of an ISSR based STS marker for sex identification in pointed gourd (*Trichosanthes dioica* Roxb.). Scientia Horticulturae 150: 11–15.

Niroshini, E., Everard, J.M.D.T., Karunanayake, E.H. and Tirimanne, T.L.S. (2000) Sex specific random amplified DNA (RAPD) markers in *Carica papaya* L. Tropical Agricultural Research12: 41–49.

Pakull, B., Groppe, K., Mecucci, F., Gaudet, M., Sabatti, M. and Fladung, M. (2011) Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a *Populus tremula* x *Populus* *tremuloides* cross. Canadian Journal of Forest Research 2: 245–253.

Parasnis, A.S., Gupta, V.S., Tamhankar, S.A. and Ranjekar, P.K. (2000) A highly reliable sex diagnostic PCR assay for mass screening of papaya seedling. Molecular Breeding 6: 337–344.

Parasnis, A.S., Ramakrishna, W., Chowdari, K.V., Gupta, V.S. and Ranjekar, P.K. (1999) Microsatellite (GATA)*n* reveals sex specific differences in papaya. Theoretical and Applied Genetics 99: 1047–1052.

Park, S.O., Crosby, K.M., Hunag, R. and Mirkov, T.E. (2004) Identification and confirmation of RAPD and SCAR markers linked to the *ms*-3 gene controlling male sterility in melon (*Cucumis melo* L.). Journal of the American Society for Horticultural Science 129: 819-825.

Patil, C.G., Baratakke, R.C. and Sandigwad, A.M. (2012) Development of a RAPD-based SCAR marker for sex identification in *Momordica dioica* Roxb. Isreal Journal of Plant Science 60: 457–465.

Persson, H.A. and Nybom, H. (1998) Genetic sex determination and RAPD marker segregation in the dioecious species sea buckthorn (*Hippophae rhamnoides* L.). Hereditas 129:45–51

Polley, A., Seigner, E. and Ganal, M.W. (1997) Identification of sex in hop (*Humulus lupulus* L.) using molecular markers. Genome 40: 351–357.

Rana, S., Shirkot, P. and Yadav, M.C. (2009) A female sex associated randomly amplified polymorphic DNA marker in dioecious *Hippophae salicifolia*. Genes Genome Genomics 3: 96–101.

Reamon-Büttner, S.M., Schondelmaier, J. and Jung, C. (1998) AFLP markers tightly linked to the sex locus in *Asparagus officinalis* L. Molecular Breeding 4: 91–98.

Renner, S.S. and Ricklefs, R.E. (1995) Dioecy and its correlates in the flowering plants. American Journal of Botany 82: 596–606.

Semerikov, V., Lagercrantz, U., Tsarouhas, V., Ronnberg-Wastljung, A., Alstrom-Rapaport, C. and Lascoux, M. (2003) Genetic mapping of sex-linked markers in *Salix viminalis* L. Heredity 91: 293–299.

Sharma, A., Zinta, G., Rana, S. and Shirko, P. (2010) Molecular identification of sex in *Hippophae rhamnoides* L. Using isozyme and RAPD markers. Forestry Studies in China 12: 62–66.

Sharma, K., Agrawal, V., Prasad, M., Gupta, S., Kumar, R. and Prasad, M. (2008) ISSR marker-assisted selection of male and female plants in a promising dioecious crop, jojoba (*Simmondsia chinensis*). Plant Biotechnology Report 2: 239–243.

Shelton, A.O. (2010) The origin of female-biased sex ratios in intertidal seagrasses (*Phyllospadix* spp.). Ecology 91: 1380–1390.

Shibu, M.P., Ravishanker, K.V., Anand, L., Ganeshaiah, K.N. and Shaanker, R.U. (2000) Identification of sex-specific DNA markers in the dioecious tree, nutmeg (*Myristica fragrans* Houtt.). PGR Newsletter 121: 59–61.

Shirkot, P., Sharma, D.R. and Mohapatra, T. (2002) Molecular identification of sex in *Actinidia deliciosa* var. *deliciosa* by RAPD markers. Scientia Horticulturae 94: 33–39.

Singh, M., Kumar, S., Singh, A.K., Ram, D., Kalloo, G. (2002) Female sex associated RAPD marker in pointed gourd (*Trichosanthes dioica* Roxb.). Curr Sci 82: 131–132.

Stehlik, I. and Blattner, F.R. (2004) Sex-specific SCAR markers in the dioecious plant *Rumex nivalis* (*Polygonaceae*) and implications for the evolution of sex chromosomes. Theoretical and Applied Genetics 108: 238- 242.

Sun, G., Ji, Q., Dilcher, D.L., Zheng, S., Nixon, K.C. and Wang, X. (2002) Archaefructaceae, a new basal angiosperm family. Science 296: 899–903.

Sun, Q., Yang, X. and Li, R. (2014) SCAR marker for sex identification of *Pistacia chinensis* Bunge (Anacardiaceae). Genetics and Molecular Research 13: 1395-1401.

Tognetti, R. (2012) Adaptation to climate change of dioecious plants: does gender balance matter? Tree Physiology 32: 1321–1324.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg ,G., Van de Peer, Y. and

Rokhsar, D.. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793): 1596-604.

Wang, X. and Curtis, P.S. (2001) Gender-specific responses of *Populus tremuloides* to atmospheric CO_2 enrichment. New Phytologist 150: 675–684.

Wikström, N., Savolainen, N. and Chase, M.W. (2001) Evolution of the angiosperm: calibrating the family tree. Proceedings of the Royal Society of London. Series B, Biological sciences 268: 2211–2220.

Xu, X., Yang, F., Xiao, X., Zhang, S., Korpelainen. H. and Li, C. (2008) Sex-specific responses of *Populus cathayana* to drought and elevated temperatures. Plant, Cell & Environment 31: 850–860.

Yang, H., Gan, S., Yin, G. and Hu, H. (2005) Identification of random amplified polymorphic DNA markers linked to sex determination in *Calamus simplicifolius* C.F. Wei. Journal of Integrative Plant Biology 47: 1249–1253.

Younis, R.A.A., Ismail, O.M. and Soliman, S.S. (2008) Identification of sex-specific DNA markers for datepalm (*Phoenix dactylifera* L.) using RAPD and ISSR techniques. Research Journal of Agriculture and Biological Sciences 4: 278–284

Zhang, S., Jiang, H., Peng, S., Korpelainen, H. and Li, C. (2011) Sex-related differences in morphological, physiological, and ultrastructural responses of *Populus cathayana* to chilling. Journal of Experimental Botany 62: 675–686.

Zhang, Y.H., Di Stilio, V.S., Rehman, F., Avery, A., Mulcahy, D. and Kesseli, R. (1998) Y chromosome specific markes and the evolution of dioecy in the genus *Silene*. Genome 41: 141-147.

Zhao, H., Li, Y., Zhang, X., Korpelainen, H. and Li, C. (2012) Sex-related and stage dependent source-to-sink transition in *Populus cathayana* grown at elevated CO_2 and elevated temperature. Tree Physiology 32: 1325–1338.

*****