Molecular Markers and DNA Barcoding in Moringa

9

N. Manikanda Boopathi and M. Raveendran

Abstract

Evolving novel crop cultivars with improved agronomic performance and increased yield and quality traits using traditional breeding approaches is a time-consuming, labor intensive, and elaborate process. Recent developments in molecular marker technologies and sophisticated tools in genetic mapping and quantitative trait loci localization in the plant genome, offered unique and promising avenues in molecular breeding. It speeds up the routine breeding process by precisely introgressing and evaluating the breeding materials. Since the first introduction of restriction fragment length polymorphism (RFLP), several types of molecular makers (mostly random DNA markers, which are phenotypically neutral and intricately linked to the gene of interest) have been developed and successfully employed in crop improvement program during the past few decades. An overview of different kinds of molecular markers and their applications is provided in this chapter, besides providing a snapshot on the molecular markers and its applications in Moringa. This chapter also provides particulars on basics of DNA barcoding, genes that are suitable for Moringa DNA barcoding, and its applications in Moringa trade.

9.1 Prelude on Molecular Markers

Evolving elite cultivars in crop plants including Moringa has been attempted using different conventional breeding strategies such as introduction, selection, hybridization, mutagenesis, development of synthetic and composite varieties. In fact, all the cultivars released in Moringa have been developed using the conventional breeding methods (see Chap. 5), which involves evaluation of traits (such as pod yield and quality) under natural conditions (famer's field) and/or greenhouse conditions and selecting the best performing lines. However, such selection process is highly influenced by the soil and other input resources and prevailing climatic conditions.

Recent advances in molecular markers and genetic maps have shown their potentials in fast and accurate recovery of the target traits in several field crops, and hence, it is anticipated that employing marker technology will also speedup the Moringa crop improvement process. Use of markers in breeding program has long history and it started with the employment of morphological traits (such as seed shape and color) as first-generation markers. However, owing to their limited number and adverse effects of

N. M. Boopathi (✉) · M. Raveendran
Department of Plant Biotechnology, Centre for Plant Molecular Biology & Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India
e-mail: nmboopathi@tnau.ac.in

N. M. Boopathi et al. (eds.), *The Moringa Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-80956-0_9

environment factors on expression of those traits, isozymes or biochemical markers were alternatively used as a second-generation markers to indirectly select the plant breeding materials. Soon, it was comprehended that isozymes were also not supportive for such breeding efforts owing to the facts that they are also limited in number, unstable during experimental assays, and possessing poor genetic relationship with the target traits (i.e., all the agronomically important traits are not always related with appropriate isozymes). As research intensified in deoxyribose nucleic acids (DNA) and ribo nucleic acids (RNA), the third-generation markers viz., DNA and RNA markers became popular in plant breeding, and they almost replaced the previous generation markers.

The key reasons for employing third-generation markers in plant breeding are:

(1) Identification of molecular markers are not involved in any pleiotropic and epistatic effect on the phenotype and they are unbiased in phenotypic reactions.
(2) Segregation and inheritance of DNA markers are not affected by the environment.
(3) Assays that used to capture DNA marker data are robust and can be automated.
(4) DNA or molecular markers are stable and detectable in all parts of plants irrespective of their growth and development and not affected by any factors.

Though there are different types of molecular markers, in order to realize their full potentials in plant breeding program, a perfect molecular marker for the plant breeding program should have the following characteristics:

(1) It should generate highest level of polymorphism even between closely related plant species (which are frequently used in plant breeding program due to their compatibility during hybridization) and be evenly distributed throughout the genome.
(2) The marker should be independent and reliable and should generate multiple alleles.
(3) The experiment used to assay the selected marker should be simple, swift, and cheap.
(4) The marker assay should require little tissue and/or DNA samples.
(5) Development of the marker assay should not depend on genomic information of an organism.
(6) The identified polymorphic marker should be linked to large numbers of agronomic- and yield-related traits.
(7) It should offer adequate resolution (minimum distance) to the identified quantitative trait loci (QTL; see Sect. 9.1.2).

On the other hand, there is no molecular marker that can be used as an ideal marker as each marker class has its own advantages and disadvantages (Table 9.1). Assay that used to detect each marker class as well as their applications largely differ from each other at their abundance in the given genome, polymorphic information content between the investigated parental lines, specific ability to detect the target QTLs, reproducibility across the laboratories, and ease of the assay and their affordability. Depending on the need, several modifications in the assay techniques can be made, leading to a next generation of advanced molecular markers.

9.1.1 Types of DNA-Based Markers

Specific fragment of the DNA/RNA that has the ability to distinguish two or more individuals of the crop plants is called as a molecular marker. Usually, in a typical gel-based assay, all the molecular markers produce a band-like structure in the lane that belong to each accessions, and the banding pattern (which is generated by the length differences of alleles of the each accessions; Fig. 9.1) is used to differentiate the individuals. The publication of Botstein and his team in 1980 on restriction fragment length polymorphism (RFLP) was the first report that described the molecular marker and its utilization in the construction of genetic maps.

For simple understanding, marker techniques are grouped into two categories: (1) non-PCR-based or hybridization-based techniques (e.g.,

Table 9.1 Summary of applications and restrictions of different classes of molecular markers when they are employed in plant breeding program

Type of marker[a]	Advantage	Limitations
Protein markers	1. Rated as technically simpler and lesser expensive markers when compared with other marker classes 2. Markers are codominantly inherited	1. Proteins are unstable and it is difficult to differentiate the assay results and experimental errors 2. As less than 30% of the nucleotide variations lead to changes in protein mass (and hence changes in protein fragment sizes), majority of the isozymes cannot be detected as polymorphic marker, as they result into similar banding pattern 3. Further, the polyploidy nature of majority of the cultivated plants interfere with the interpretation of allozymes patterns
RFLP	1. Extremely specific to the given locus 2. Inherited in true Mendelian fashion 3. Show highest reproducibility among the marker classes. 4. Shown to possess adequate genomic abundance	1. Large quantity of pure DNA is required for the assay 2. Need prior genomic sequence information to develop probe for the assay 3. Involves radioactive and toxic reagents in the assay 4. Require skilled manpower and cannot be automated
AFLP	1. Among the marker classes, shown to generate highest number of Multi-locus markers (can produce 50–100 polymorphic fragments between closely related genotypes per assay)	1. Though it can be semi-automated, needs heavy investment and elaborate procedure, and 2. Require skilled workforce
RAPD	1. Need minimum laboratory infrastructure 2. Identified as simplest assay technique 3. As a startup in genome research, it can be applied to any genome	1. Reproducibility across the laboratory is the frequent problem associated with this marker due to the fact that a single variation in the experimental procedure will lead to produce errors. However, multiple replications to confirm the polymorphism will avoid this limitation 2. RAPD markers are dominantly inherited
SSR	1. Simple assay technique with minimum requirement of infrastructure 2. As the markers are codominantly inherited, it is easy to identify the heterozygotes from the homozygote parents 3. Regarded as the breeder's marker of choice due to their extensive effectiveness	1. Largely depends on the repeat motif identification, which require prior genomic information of the given species 2. SSR have relatively poor discriminative power among the closely related accessions
SNP	1. Automated and large-scale genotyping assay methods assist in the detection of highest numbers of informative markers 2. Variations in the coding regions may or may not change the amino acid sequence, however, it can be effectively linked to the trait of interest 3. SNPs are employed in fine mapping by developing high-density genetic maps which is useful for positional cloning (a process by which the underlying gene is cloned and sequenced)	1. Largely depend on the genomic information 2. Require relatively higher investment for initial infrastructure development

[a]See Sect. 9.1.1 for expansion of each marker class

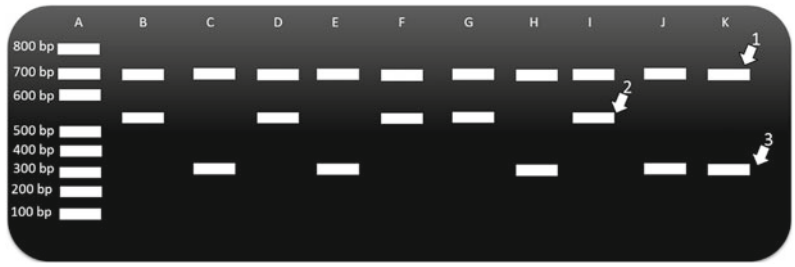

Fig. 9.1 A schematic illustration of molecular markers and its usage in differentiating the genetic variation among the plant accessions. A—marker ladder lane, which is used to denote the approximate size of the DNA fragments that moves in the gel. Lanes B to K represents different Moringa ecotypes. Arrow heads with numerals denotes different alleles produced by the same marker (say for example, RAPD) that can be able to differentiate the ecotypes. In this illustration, Allele 1, is common for all the ecotypes and hence said to be monomorphic. Whereas, Allele 2 and Allele 3 are said to be polymorphic as they have different banding pattern among the Moringa ecotypes. Allele 2 can be used to differentiate the ecotypes B, D, E, G, and I from the investigated ecotypes B–K. Allele 3 can be used to differentiate the ecotypes C, E, H, J, and K from the investigated ecotypes B–K

RFLP) and (2) PCR-based techniques (e.g., SSR and RAPD).

Novel molecular marker systems were developed after the invention of polymerase chain reaction (PCR) technology due to its minimalism in experimental procedure and superior chance of getting polymorphism. Another major breakthrough is the use of synthetic random primers for PCR amplification (and hence such marker class development does not require prior sequence information) for PCR analysis which facilitated the introduction of various types of molecular markers for a range of applications in crop improvement.

PCR-based marker development are further grouped into two subcategories: (1) arbitrarily primed PCR-based techniques or randomly targeted sequence amplification techniques (e.g., Random amplified polymorphic DNA (RAPD), Amplified fragment length polymorphism (AFLP)), and (2) specific sequence-focused PCR techniques (such as microsatellite (or short tandem repeats or simple sequences repeats, SSR)-based marker technique, Inter Simple Sequence Repeats (ISSRs), Randomly Amplified Microsatellite Polymorphisms (RAMP; which is a combination of RAPD and ISSR), Cleaved Amplified Polymorphic Sequences (CAPS), Sequence Characterized Amplified Regions (SCAR), Sequence-Related Amplified Polymorphism (SRAP), Target Region Amplification Polymorphism (TRAP), Single Strand Conformation Polymorphism (SSCP), Inter-Retrotransposon Amplified Polymorphism (IRAP) and REtrotransposon-Microsatellite Amplified Polymorphism (REMAP), Retrotransposon-Based Insertion Polymorphism (RBIP), Single Nucleotide Polymorphism (SNP)).

As science advances in genomics and transcriptomics, those data have also been employed to identify novel marker classes. For example, molecular markers were developed by PCR amplification of expressed region of the genome (exons) and they are defined as genic molecular markers. Such markers are functional portions of the given genome and they are mainly developed by using the complementary DNA or Expressed Sequence Tags (cDNA/EST) datasets. Further, huge amount of small RNA sequences have also been generated due to the rapid progress in next-generation DNA sequencing which led to develop microRNA-specific novel functional markers at the DNA level. Detailed discussion on the marker assays for different kinds of molecular markers and troubleshooting measures can be found at Boopathi (2020).

The important inherent properties of molecular markers (Tanksley 1983) that distinguish them from morphological markers are:

1. Morphological markers required whole plant assay whereas molecular markers can be determined at tissue or cellular levels.
2. A relatively lesser number of informative morphological markers are available (which also require application of exogenous mutagens); however, in nature, abundant informative alleles are available at the molecular level.
3. It has also been noticed that the given morphological marker very often linked with undesirable phenotypes. In contrast, alternate alleles of molecular markers rarely reported to be associated with deleterious effects.
4. Majority of the molecular markers segregate codominantly, and hence, both homozygous and heterozygous alleles can be distinguished easily; whereas dominant segregation of morphological markers limit their utility in many breeding applications.
5. Unlimited numbers of molecular markers with fewer epistatic or pleiotropic effects can be explicitly scored in the segregating generations, which is not at all possible with morphological markers.

9.1.2 Applications of Molecular Markers

Applications of molecular markers have been realized in analyzing genetic diversity, identifying molecular phylogeny and evolutionary relationships, detecting duplications in germplasm collections, identifying close relatives, fingerprinting varieties/cultivars/hybrids, mapping genetic loci and tagging genes for management and improvement of crops, conserving endangered germplasm species, discriminating male and female flowers, etc.

Codominant molecular markers were primarily employed to fix the hybrids. Dominant markers cannot distinguish the allelic difference of a gene when it is in heterozygous conditions whereas codominant marker can differentiate the allelic difference. Therefore, codominant markers are used to categorize the hybrids simply and effortlessly from their homozygous parents. The markers coming under the codominant category are protein markers (such as Allozymes and Isozymes) and DNA-based single locus markers (such as RFLP, SSR, and SNPs). RAPD, ISSR, and AFLP are called as dominant markers.

Markers with increased reliability and functional information can be developed using bioinformatic tools when the sequences from the genic regions are used. As these markers are developed from genic regions, any polymorphism detected in the individuals, varying for the given trait, will facilitate in exploring the underlying biological function. Besides, such markers can be used as potential candidates for comparative genomic studies and offer an accurate estimate of functional diversity. The data developed from these studies would be useful to understand the molecular mechanisms involved in acclimatization of plant species to diverse environmental conditions.

Large variations in reproductive systems such as hermaphrodite, monoecious, and dioecious have been recognized in flowering plants. Identification of gender-specific plants, at the seedling stage itself, will significantly reduce the time, effort, and resources required to sustain the female plants in the field, to guarantee greater yields. Hence, sex-specific molecular markers have shown its applications in the crop breeding program with greater precision.

In view of its immediate potential applications of molecular markers in Moringa, only two applications are described in detail in the below subsections.

9.1.2.1 Marker-Assisted Selection

Molecular markers have shown to play different roles in plant breeding program, and the immediate application of DNA-based molecular markers in Moringa is developing high-resolution genetic (or linkage) maps and detecting molecular markers tightly linked with genes or quantitative trait loci (QTLs; which comprises a set of genes) underlying agriculturally

important quantitative traits (such as disease and insect resistance, abiotic stress tolerance and pod or leaf yield, and other economic quality related characteristics). Comprehensive procedures for constructing linkage map and analysis of QTLs in plants have been described by Boopathi (2020). Linkage mapping and QTL analysis have provided opportunities for indirect selection of economically important traits using the linked markers instead of the trait itself. Such selection process is known as marker-assisted selection (MAS).

MAS has been considered to be superior over desirable trait-based selection due to the following advantages:

i. Rapid and simple selection of the desirable traits even without the involvement of selection environment (e.g., for the selection of certain disease resistance traits, MAS is considered to be quicker, cheaper, and more efficient than the routine phenotype-based selection; besides selection can be performed even when there is no disease incidence).
ii. Some of the economically important traits are expressed and can be scored at the end of the growing season, and further, they are regulated by the unstable developmental process in response to the environmental conditions. Use of molecular markers linked to the trait enables early selection for such traits and breeder's convenience.
iii. Molecular markers also facilitate selection of multiple traits from the same plant population that are grown in the same season and environment. Such effort is not possible through conventional breeding. For example, selection of drought and submergence tolerance traits requires evaluation of plant population under different seasons and environments. On the other hand, plants with drought and submergence tolerance can be selected using the respective markers linked to these traits without conducting the field evaluation trials.
iv. Without performing any progeny testing for target traits such as disease resistance, selection for recessive traits can be done with molecular markers.
v. Use of foreground, background, and recombinant markers in MAS enables rapid removal of undesirable genetic backgrounds as well as plant materials with undesirable gene combinations in segregating populations during the early stage of plant breeding program, and thus, greatly reduces the breeding cycle besides retaining the most useful plant materials.
vi. Selection of progenies using molecular markers greatly reduces the problems associated with linkage drag, i.e., it efficiently breaks the linkage between favorable and unfavorable alleles.
vii. The traits that show low heritability and exhibit high genotype x environment interactions (and hence such traits cannot be selected efficiently due to environmental effect during conventional selection), can be selected effortlessly using molecular markers.
viii. As the MAS is independent of seasons and environments, multiple numbers of selections can be done in a year. Thus, MAS has more pronounced applications in speed breeding.

Owing to the above advantages, molecular markers are currently being employed extensively in both public and private sector crop breeding programs to genetically improve the target traits in the given crop.

On the other side, although molecular markers have been identified and found to be associated with several agriculturally important traits, they are not always consistent or having their immediate application in plant breeding programs. Main confronts in assemblage of markers to the trait of interest for efficient MAS is identifying markers that are closely linked, inexpensive, and simply assayed. However, majority of the identified markers for MAS are unsuccessful in proving their efficiency as these identified

markers are not either tightly close to or not part of the gene itself. Even tiny genetic distance identified between the markers and the genes of interest results in genetic recombination, and ultimately, leads to loss of the trait of interest.

Exchange of genetic elements between the marker and the gene may occur through the process known as genetic recombination during the evolution or plant breeding process. Owing to this genetic recombination, even though the plants are selected with the associated markers, selected plants may not have the desirable gene in the favorable allele combinations. Such plants are called as false positives. Apparently, such increase in false positives among the progenies will lead to decrease in the breeding efficiency. Therefore, validation of identified molecular markers (by confirming its closeness to the target gene either by fine mapping or positional cloning) before implementing the MAS will greatly increase the utility of the given molecular marker in crop improvement programs.

Next difficulty in widespread utilization of MAS in crop improvement programs is the need for identifying consistent marker alleles linked to the target genes across the germplasm and environments. In practice, MAS has been found to be effective when the markers are linked to the trait (s) of interest across the breeding populations. Unfortunately, in majority of the reports, the identified markers were not validated across the global (or at least using the national) breeding populations.

Thus, cautious development and validation of markers are required before implementing them as routine indirect selection criteria. Though the use of molecular markers has shown its widespread applications in field and horticultural crops (Boopathi 2020), their potential applications have not yet been fully realized in Moringa. Those scientific reports that focused on use of molecular markers in Moringa are summarized here under.

9.1.2.2 Documenting Genetic Variation in Moringa Using Molecular Markers

Understanding the molecular mechanisms of different biological process, growth, and development can be revealed by carefully designed experiments that explore the plant genetic variation at molecular level. Though the next-generation sequencing strategies provide the opportunity to sequence the plant genome relatively with simple efforts, there is still some issues in getting the complete plant genome; besides, it would be heavily priced, if all the associated plant kingdom is targeted to sequence. To this end, molecular markers would be an cost-effective and simple approach not only to capture the plant genetic variation by offering necessary genomic landmarks in the given plant species, but also it provides methods to correlate the phenotypes with the identified polymorphic markers.

Studies that focused on ecological, evolutionary, taxonomical, phylogenic, and genetic experiments in plant sciences have been successfully and now being routinely conducted with molecular markers such as RFLP, RAPD, SSR, and AFLP. Their utilities as well as restrictions associated with these markers in unraveling the plant genetic variation have been realized (Boopathi 2020).

Using the basic principles of early generation markers development, advanced and semi-automated techniques (primarily based on the principle of detecting the SNPs) have emerged and being employed to high-throughput analysis of the genetic variation with large numbers of samples in a short span of time. Such advanced methodologies (that detect the variations in the form of SNPs, structural variations (SVs), Insertions and Deletions (InDels)) have evolved to combine useful features of previous marker techniques and also increased the sensitivity and resolution to detect genetic discontinuity and

distinctiveness. Besides, the recent marker development also focuses to capture the variations in other untapped genomic regions such as retrotransposons, mitochondrial, and chloroplast-based microsatellites. Such efforts increased the genome sampling exposure and thus provided extended information on genetic variation. However, it involves initial investment in sequencing the organelle genomes and targeted genomic repeat rich regions in the given plant germplasm.

Alternatively, earlier RAPD and AFLP principles are employed but using cDNA library specifically developed from the given plant tissues that are under various development stages or stressed by different biotic and abiotic factors. Such efforts greatly increase our ability to study different types of gene expression models and reveal molecular mechanisms behind such biological responses which would have practical utility in designing precise plant breeding agenda.

9.2 Molecular Markers in Moringa

The genomic information on Moringa was reported as early as 1937 by mentioning "*Moringa tree is a true diploid with 2n = 28*" (Patel and Narayana 1937). However, complete potentials of molecular markers have not yet been realized in Moringa, and they are mainly explored extensively to document the genetic diversity that exist in the different Moringa germplasm accessions that have been maintained at different institutes.

Among the different classes of markers, majority of the studies employed randomly generated multi-locus markers such as AFLP, RAPD, and ISSRs as these are found to provide significant variation among the investigated Moringa accessions and invariably established that these marker classes are rapid and cost-effective fingerprinting techniques (Table 9.2). For example, Hassanien and Al-Soqeer (2018) assessed the genetic diversity of *Moringa oleifera* and *Moringa peregrina* that were present in Saudi Arabia using ISSR markers, whereas Kumar et al. (2017) found RAPD markers were useful to capture the genetic variations and explore the genetic relationships among seven advanced breeding lines of Moringa.

Another striking point from the published reports is that there is a significant genetic variation existing even among the investigated closely related Moringa accessions. This may be due to out-crossing that occurs frequently in Moringa (Actually, it is a cross-pollinated crop). This implies that those genetically diverse unique accessions would be a valuable resource for further breeding program that focus on novel traits, such as improved nutritional content, and conservation of such genetic variation for the future.

9.2.1 SSR Marker Development in Moringa

The need for development of SSRs in Moringa has been first described in Wu et al. (2010), and they amplified microsatellite loci using the customized enrichment protocol that employed AFLP of nucleotide repeats containing sequences (Zane et al. 2002). As per the protocol, Wu et al. (2010) screened 288 randomly selected clones by colony PCR and identified that 210 of them were microsatellite motifs. Though 209 of these clones were productively sequenced, only 192 clones (67%) were possessing SSRs. Consequently, primers were designed to flank these microsatellites and found that out of 69 pairs of 20-base primers, only 46 pairs successfully amplified target regions. Finally, they reported 20 Moringa SSRs that can offer useful polymorphism among Moringa germplasm. It can be noticed from this study that the strategy employed for development of novel SSRs was a laborious and time-consuming process besides yielding a poor number of polymorphic SSR markers.

Alternatively, to increase the efficiency of reporting large numbers of novel SSRs, various molecular strategies were employed. For example, selective hybridization of microsatellite regions of the genomic libraries was regarded as

Table 9.2 Selected studies that focused to utilize molecular markers for the assessment of genetic variation among the Moringa genotypes

Investigated plant materials	Markers employed	Remarks	References
This study utilized seven different populations (two from Tamil Nadu (South India), one from Ex-Nsanje region (Southern Malawi), and four from Kenya with 20 replications (totally 140 genotypes)	AFLP	If the *M. oleifera* genetic resources were explored from the regions in which they have originated, they show increased chance of documenting novel genetic variation. AFLP can be used to capture maximum number of informative alleles among the marker classes	Muluvi et al. (1999)
Sixteen accessions from Germplasm Bank (BAG) of Embrapa Coastal Tablelands, Sergipe, Brazil	RAPD	Owing to low diversity detected in this investigation, new activities of collection was proposed to increase the diversity	da Silva et al. (2012)
Eight Moringa cultivars of South India	Both random and gene-specific markers (RAPD, ISSR, and cytochrome P_{450}-specific)	Though the investigated cultivars were classified into four sub-clusters, there was no geographical-based grouping. This confirmed active spread of the planting materials. Further, as Moringa is a cross-pollinated crop, there was great rate of gene flow	Saini et al. (2013)
Totally 12 populations comprising 300 genotypes assembled from northern (Himachal Pradesh) and southern (Tamil Nadu) states of India	SSR	SSR-based grouping disagree with geographical distribution. Hence, concluded that there was seed movement and/or high rates of gene flow between the adjacent populations	Ganesan et al. (2014)
Seven advanced breeding lines of Moringa	Different types of markers such as seed protein profile, RAPD, and Cytochrome P450 gene-specific markers	Only RAPD and Cytochrome P450 markers generated considerable amount of polymorphism, 79.68 and 86.44%, respectively	Kumar et al. (2017)
Ninety-seven Moringa accessions assembled from various parts of India	SSRs	Reported high genetic diversity among the investigated accessions	Rajalakshmi et al. (2017)
Tissue cultured Moringa plants	RAMP	RAMP can be used to ensure genetic stability of the different systems of micropropagation of *M. oleifera*	Avila-Treviño et al. (2017)
Moringa oleifera and *M. peregrina*	Morphological markers and ISSRs	ISSR were effective in the characterization of genetic diversity	Hassanien and Al-Soqeer (2018)

(continued)

Table 9.2 (continued)

Investigated plant materials	Markers employed	Remarks	References
Seven genotypes per species of *Moringa oleifera* and *M. peregrina*	Nutritional characteristics and SSR	A strong relationship was found between nutritional and molecular classifications of genotypes	Hassanein (2018)
Ninety-seven *M. oleifera* accessions (2 hybrids and 95 naturally grown traditional varieties)	ISSR and SRAP	These markers are useful for population structure and association studies	Rajalakshmi et al. (2019)
Leaf samples collected from 10 M. *oleifera* plants that were naturally found in Middle Delta, Egypt	Random markers such as ISSR, start codon targeted (SCoT) markers, and barcoding genes such as *ITS2*	ISSR, SCoT, and ITS markers can be employed to effectively identify *M. oleifera* products from its adulterants and similar species	Hassan et al. (2020)
Totally, 23 M. *oleifera* accessions that were collected from Kerala, Tamil Nadu, and Karnataka states of India	Seven different cytochrome P_{450} ($CytP_{450}$) gene-specific markers	$CytP_{450}$ can be used to capture genetic diversity in Moringa	Ravi et al. (2020)

robust, high-throughput, reproducible, and inexpensive method. The steps involved in selective hybridization facilitate securing large-scale microsatellite sequences that are labeled with biotin probes and consequently capturing them using magnetic beads coated with streptavidin. Alternatively, the probes can also be fixed on nitrate filter. This kind of enrichment of SSRs and their isolation have been successfully employed in different crop species including Guava. However, this method is again found to be involved in high demand for infrastructure facilities, skilled manpower, and additional time.

Another novel strategy proposed for large-scale development of SSR is use of next-generation sequencing data for identifying SSR motif and designing primers that flank such motifs. Thanks to the widespread availability of next-generation sequencing facilities, nowadays, draft or complete genome sequences and transcriptomic sequences of several crop species including Moringa (refer Chap. 10 for more details) are available for public use.

Presence of microsatellite markers in those sequences can be easily identified by using several algorithms (for example, the MIcroSAtellite (MISA) identification tool (http://pgrc.ipk-gatersleben.de/misa/misa.html); SSRIT available in Gramene (http://gramene.org/ssrit) and SSR-Locator (www.microsatellite.org/ssr.php)). Such software offer wide range of parameters that can facilitate the rapid and effortless identification of perfect di-, tri-, tetra-, penta-, and hexanucleotide motifs with the customized minimum number of repetitions. While identifying the SSRs, mononucleotide repeats usually are not considered as it may lead to misinterpretation of results.

Subsequently, the primer pairs will be designed by using software such as Primer3 v0.4.0 with a set of adjustable parameters. The following widely employed parameters would be useful to design robust primer pairs: primer size 18–25 bp (optimum 20 bp), PCR product size 100–400 bp (optimum 280 bp), GC percent 45–70 (optimum 50%), Tm 57–63 °C (optimum 60 °C).

The SSRs identified in the above workflow must be validated with the target crop, and it can also be tested among the related species to test verify their transferability among the closely related species. More than thirty thousand such genome-specific and genic SSR markers (Fig. 9.2) were developed at this laboratory using the genomic and transcriptomic sequences of Moringa cultivar PKM1 (article in preparation) for the first time. A set of randomly selected

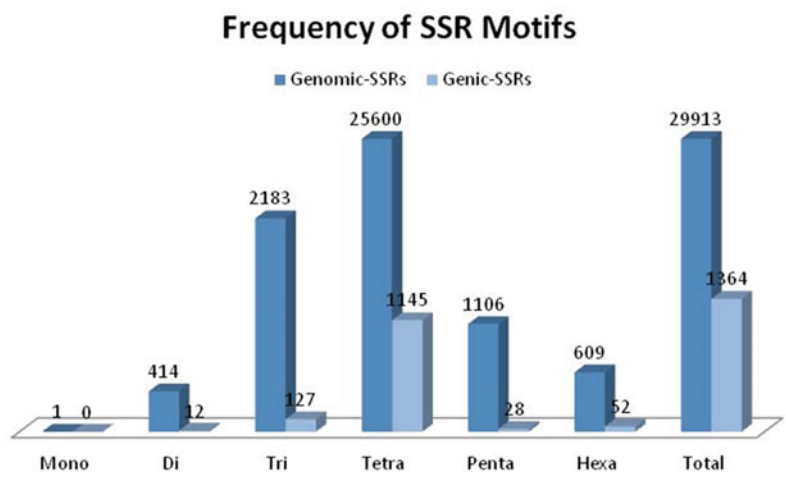

Fig. 9.2 Summary of SSRs identified using resequencing and transcriptomics sequence of Moringa cultivar PKM1

SSRs were validated in Moringa ecotypes and proved their utility in Moringa research. Informative SSR markers identified in this exercise will have several applications such as varietal protection, germplasm management, trait mapping, gene isolation and selections, biotic and/or abiotic stress resistance improvement using molecular breeding strategies.

9.3 Prelude on DNA Barcoding

DNA barcoding is the procedure for identification of crop/animal/microbial species using the variations at nucleotide level that are found in the specific DNA fragments (DNA barcodes). DNA barcoding has been first established in animals using *cytochrome c oxidase subunit 1* (*COI*) gene, and now, it is being used as a standard DNA barcode in animals.

However, owing to the lack of universally accepted DNA barcoding gene(s), it is yet to be established in plants. Use of several plant candidate genes from plastid and nuclear genome, such as *atpF-atpH* spacer, *matK*, *rbcL*, *rpoB*, *rpoC1*, *psbK-psbl* spacer, and *trnH-psbA* spacer, have resulted in that there is no single gene that can be used for DNA barcode in plants. Consequently, Consortium for the Barcode of Life (CBoL)–Plant Working Group recommended a two-gene combination, *rbcL* + *matK,* as the plant DNA barcodes by comparing the sequence quality, reproducibility across the laboratories, relative efficacy of testing, and levels of species discrimination.

Employment of *rbcL* + *matK* as DNA barcodes has been shown to be universal framework in DNA barcoding of land plants and identify and discriminate the closely related species and ecotypes within the species. It is recommended to initiate DNA barcoding of any land plant species with this two-locus standard barcode (*rbcL* + *matk*) as it has reported so far a minimum of 70% species discriminatory power; however, additional DNA barcoding genes need to be explored if superior resolution is required.

Retrieval of better intraspecific than interspecific variation is considered as the ideal qualities of a DNA barcoding region in such a way that highly similar species can be recognized using those unique variations in barcodes. Additional qualities of an ideal DNA barcoding region includes they should be short ($\sim$750 bp), universally amplifiable across all plant taxa using simple procedure, generated sequences should align readily with the reference sequences, and represents low copies of insertion-deletions (InDels).

Key application note of DNA barcoding is to provide a common community resource of DNA sequences which can precisely identify the organisms and clarify the taxonomical classifications. Such barcoding genomic resource shed light on species-level taxonomy among the

individuals having simple and similar morphologies, those with exaggeratedly broad classifications, those that are tiny in size (or powdered for easy of packing) and cannot be handled at morphological level, and those that are inadequately characterized even though they possess genetic diversity due to the complex procedure of morphology-based taxonomy.

DNA barcoding is now widespread among several areas of plant sciences and professionals such as taxonomists, ecologists, conservationists, foresters, agriculturalists, forensic scientists, customs, and quarantine officers as they are routinely employing DNA barcoding for plant characterization. Hollingsworth et al. (2011) have reported that discrimination power of the given DNA barcode can be very much useful and the following are the examples of such instances:

(1) Characterization of genetic diversity of plant species exist in the given geographical regions where immature material and/or plant fragments need to be identified.
(2) Identification of the plant species in the processed and packed plant products (such as diet supplements prepared from herbs, food products, and the components of herbal medicines). Usually, the challenge is to identify the adulteration of different or related species in the target plant product and notify the members of other taxonomic groups rather than given species.
(3) Classifying the unfamiliar plant species with the existing species group when there is no information on the investigated plants. It is also found to be useful in classifying the species rich plant systems where there is poor availability of expertise in taxonomy.
(4) Application in ecological forensics to critically characterize plant roots, seedlings, or cryptic life stages (e.g., fern gametophytes) where DNA barcoding can be used as an affordable practical route to precisely identify the specimens. For example, *rbcL* sequences alone was successfully employed by Kesanakurthi et al. (2011) to characterize 85% of all root samples and provided a detailed report on ecological factors that played important role in the subterranean spatial organization of plant diversity. Similarly, Harmon (2010) used *matK* DNA barcodes to notify unspecific plant species in herbal supplements, and found that herbal supplements of Black Cohosh did not contain the target north American species *Actea raceomosa* and instead it was a mix of Asian species of *Actea*.
(5) Other forensic studies in which DNA barcoding was employed are (i) predicting poached wildlife, (ii) pinpointing forensic relevant fly species in forensic cases, (iii) recognizing species in illegal egg and timber smuggle as well as in ecological forensic studies, (iv) monitoring biological invasions in soil as well as water, and (v) stressing the need for upgradation of taxonomic revision.
(6) Protecting the endangered species in trade is also effectively mediated by DNA barcoding. Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES; http://www.cites.org/eng/disc/species.shtml) has protected more than 29,000 plant species, and it is important to find effective and affordable methods to discriminate CITES-listed plant species from non-CITES-listed plants. To this end, *matK* DNA barcodes were found to be efficient to differentiate traded timber products of *Ramin* (*Gonostylus*) species which are CITES protected (Ogden et al. 2008).

Though DNA barcoding is proved useful in initial identification and categorization of plants at species level, a major limitation of DNA barcoding is it lacks precise definitiveness of the plant species (as that of animals) that can be accepted by the court of law and other regulatory frameworks. In order to realize the full potential of DNA barcoding, it is vitally required to develop an appropriate reference library which warrants production of a database on DNA barcodes from vouchered plant samples that were completely characterized. Several such geographically- and taxon-based projects are being undertaken in the development of DNA barcode database including in an Indian Moringa cultivar,

PKM1 (see below). Besides, sequence of different DNA barcoding genes from different plant species is also stored in GenBank, which could be a valuable and useful resource for taxonomic classifications at molecular level.

However, those data were collected and archived by the individual scientists, and it lacks validation with the voucher information and linking with the passport data of the specimens. This ultimately leads to a difficulty in identifying and removing wrongly identified specimens or irrelevant sequences. Therefore, it is imperative to link the vouchers, sequences, trace files, and other metadata by following BARCODE standards, as it is done in the development of GenBank databases and Barcode of Life Data systems (BOLD).

9.4 DNA Barcoding in Moringa

Even though India has an enormous genetic wealth of *Moringa oleifera* found in varied geographical locations, attempts towards capturing such variations in its fullest potential are limited. Existence of variability in this crop is not only because of adaptation to varied environmental conditions, but also due to the mechanism of cross-pollination. Therefore, such wide genetic variations and hidden special genetic characteristics are yet to be fully explored to document their uniqueness. Though there are several attempts in DNA barcoding of large numbers of agricultural and horticultural crops, it has not yet been explored in Moringa.

Besides documenting genetic variation and its potentials in crop improvement programs in Moringa, DNA Barcoding can also be utilized in authenticating the Moringa leaf products in both regional and global markets. Recent past decades have witnessed restoration of health care with plant-based products, and consequently, there has been an elevated demand for herbal products and drugs in global trade. Particularly, Moringa become an important export and import food commodity as it has been considered as the natural nutrition of the tropics.

Success of herbal drug industry chiefly depends on precise, rapid, and scientifically authenticated identification of the plant products. Conventionally, it has been done with a taxonomic expert, who employs combinations of both traditional and modern tools (such as morphological characteristics and biochemistry-based methods) to identify the plant materials. However, such approach would be elaborate, lengthy, and costlier for bulk identification of plant materials at industrial scale. Besides, trends in high processing of plant products with numerous ingredients may not offer accurate result with the conventional methods. This necessitates speedy, accurate, and affordable tools to screen the huge volume of plant products.

9.4.1 Selection of Appropriate DNA Barcoding Genes

As outlined above, quality control of Moringa supply systems can be effectively managed with the help of DNA barcoding. However, application of DNA barcoding in Moringa critically require seamless protocols for DNA barcoding and fool-proof method of DNA sequence-based identification that are affordable to industry and officially agreed by the competent authorities for their role in regulated procedures. DNA barcoding genes that can be used for herbal plant identification includes *matK, rbcL, trnH-psbA, ITS, trnL-F, 5S-rRNA,* and *18S-rRNA*.

As highlighted in the Sect. 9.3, in order to select a universal DNA barcode gene for plants, the Consortium for the Barcode of Life (CBOL) initiated the formation of a working group and tested the utility of seven leading candidate barcoding genes on the following parameters: universality (ease of amplification and sequencing), sequence quality, and discriminatory power. However, this project finally concluded that there is no single gene that can be used for all the plant species as each gene has its own strengths and weaknesses. Consequently, CBOL Plant Working group suggested two plastid coding regions,

rbcL and *matK*, to initiate the DNA barcoding in plants and when it is required, it can be complemented with additional genes such as *ITS* (Hollingsworth et al. 2011).

The reference for *rbcL* barcode was obtained from complete *Arabidopsis thaliana* plastid genome sequence (gi 7,525,012:54,958–56,397) which targets a 599 bp region at the 5' end of the gene (starting at 1 bp and ends at 599 bp, including primer sites). The *matK* barcode region is also derived from the complete *A. thaliana* plastid genome sequence (gi 7,525,012:2056–3570) which is ~841 bp in length but at the center of the gene (starts at 205 bp and ends at 1046 bp, including primer sites).

Both *rbcL* and *matK* can be easily amplified from the plant genome and particularly, *matK* possess higher discriminatory power and generate adequate variation in even closely related plant species. It is also established elsewhere that among the actively evolving coding regions of plastid genome, *matK* is on the top and it is recognized as the most similar plant analogue to the *CO1* animal barcode. On the other hand, *matK* cannot be easily amplified by PCR using universal primers, particularly, it is the most difficult in non-angiosperms. Contrastingly, *rbcL* can be easily amplified from the plant species and their sequence can be easily aligned with the reference sequence. Though it produces relatively lesser discriminatory power, it can be used as a foundation to create a DNA barcode database for the given plant species.

Further, owing to the fact that these DNA barcoding regions are obtained from coding genes, in silico translation of the obtained sequences can be used to analyze the errors due to editing or assembly, orientation of sequences, and existence of pseudogenes. Such analysis offers the opportunity to assess the diversity in DNA barcodes among taxonomic groups that belongs to different geographical regions (Hollingsworth et al. 2011). At the same time, it should be noted that the discrimination power of *CO1* in animals is far better than *rbcL* + *matK* in plants, and it would be always suggested to employ additional barcoding genes, if greater resolution is required.

9.4.2 DNA Barcoding of Moringa Ecotypes Cultivated in Southern India

Keeping the above points in the workplan, this laboratory has generated DNA barcodes in Moringa using *Internal Transcribed Spacer* (*ITS;* which comprises both *ITS1* and *ITS2*), *rbcL,* and *matK* genes. Such combinations of DNA barcodes clearly provided sufficient discrimination among and even closely related Moringa ecotypes that are prevailing in Southern India. Resolution of *ITS* in discriminating few selected Moringa ecotypes cultivated in Southern part of India is provided in Fig. 9.2. All the *ITS* sequences obtained from these ecotypes were submitted for public use and it is available as GenBank IDs KT737744 to KT737801. When compared with *matK* and *rbcLa*, *ITS* has a greater number of single nucleotide polymorphism as evident in Fig. 9.3.

The phylogenetic tree constructed by *ITS* gene sequences classified all the investigated accessions into two groups (Fig. 9.4). As expected, group I was constituted by outgroup alone and rest of the investigated moringa accessions were grouped as group II. In this *ITS* phylogenetic tree topology, MO 53, MO 54, and MO 52 were grouped as clade I and the remaining accessions comprised as major clade II. The boot strap value was marked at the point of node divergence. The maximum divergence was noticed in node I among MO 53, MO 54, and MO 52. The highest conservation on the sequence of *ITS* was seen between MO 42 and MO 24 followed by MO 36 and MO 47 and MO 25 and MO 12 (Fig. 9.3).

On the other hand, it should be mentioned that several investigated moringa ecotypes cannot be distinguished using the *ITS* sequence. The low discrimination power of *ITS* and *matK* is already reported in *Allium* species and wild potatoes. However, it is suggested that use of more number of barcoding genes such as *rps16*, *trnH-psbA*, *trnL-F*, *trnD-T,* and *rpL32-trnL* can be utilized for better discrimination of very closely related moringa accessions such as those reported above.

Next-generation sequencing (NGS) kindled a transformation in DNA barcoding (especially

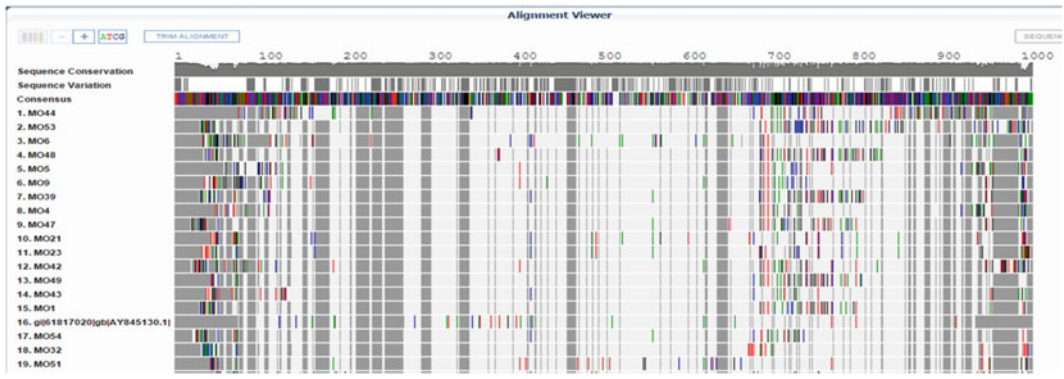

Fig. 9.3 Results of DNA barcoding of investigated moringa ecotypes using DNA subway by employing "MUSCLE" tool utilizing *ITS* gene sequence. Totally 60 genotypes (denoted with a prefix MO) were investigated in this study along with one outgroup and, barcodes of only 19 investigated barcodes were shown below along with the outgroup

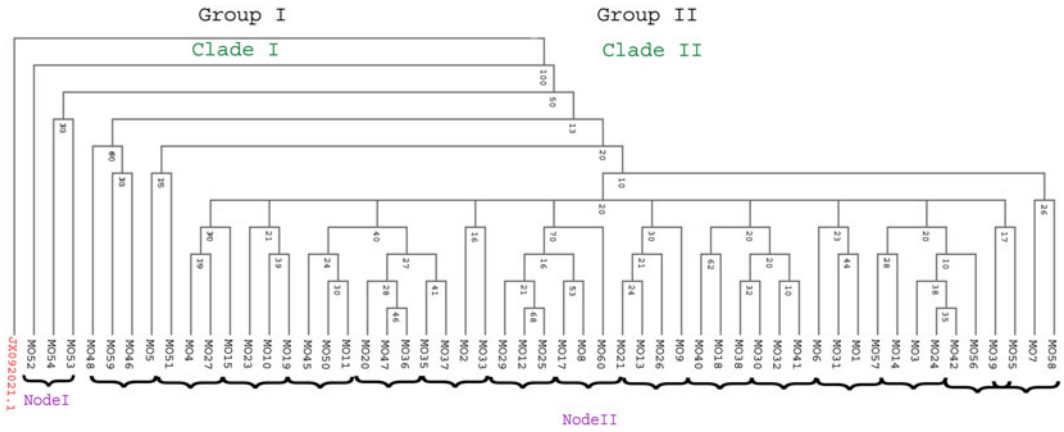

Fig. 9.4 Phylogenetic Tree constructed from the sequences of ITS amplified from the 60 moringa ecotypes grown in Southern India. Sequence of an outgroup (GenBank ID: JX092021.1) has also been included in the analysis to validate the results

metabarcoding) as it produces lengthy average sequence read length. Thus, huge amount of sequencing data can be generated from the several targeted regions of the investigated Moringa samples, and thus, it offers greater power to discriminate the samples at various genomic regions in addition to the above barcoding regions. Hence, the fundamental application of DNA barcodes, identification of unknown plant samples, gets more sophisticated and automated with the help of the NGS tools. This would be more desirable in herbal industry as raw materials used in herbal medicine are obtained directly from the markets as dried or powdered plant parts. Further, combining Moringa DNA barcoding with NGS, metabolomics, transcriptomics, and proteomics will facilitate species discovery, evolution, and conservation of useful variation for future.

References

Avila-Treviño JA, Muñoz-Alemán JM, Pérez-Molphe-Balch E, Rodríguez-Sahagún A, Morales-Domínguez JF (2017) *In vitro* propagation from bud and apex explants of Moringa oleifera and evaluation of the

genetic stability with RAMP marker. S Afr J Bot 108:149–156

Boopathi NM (2020) Marker assisted selection. In: Genetic mapping and marker assisted selection. Springer, Singapore, pp 481–504

da Silva AVC, dos Santos ARF, da Silva Lédo A, Feitosa RB, Almeida CS, da Silva GM, Rangel MSA (2012) Moringa genetic diversity from germplasm bank using RAPD markers. Trop Subtrop Agroecosystems 15(1):31–39

Ganesan SK, Singh R, Choudhury DR, Bharadwaj J, Gupta V, Singode A (2014) Genetic diversity and population structure study of drumstick (Moringa oleifera Lam.) using morphological and SSR markers. Ind Crops Prod 60:316–325

Harmon L (2010) Rare flowers and common herbal supplements get unmasked with plant DNA barcoding (reporting unpublished data of Damon Little and David Baker). Scientific American Observations. http://www.scientificamerican.com/blog/post.cfm?id=rare-flowers-and-common-herbalsupp-2010-04-18

Hassan FA, Ismail IA, Mazrou R, Hassan M (2020) Applicability of inter-simple sequence repeat (ISSR), start codon targeted (SCoT) markers and ITS2 gene sequencing for genetic diversity assessment in Moringa oleifera Lam. J Appl Res Med Aromat Plants 100256

Hassanein AMA (2018) Nutritional, chemical and molecular characterisation of Moringa oleifera Lam. and Moringa peregrina (Forssk.) Fiori genotypes. J Hortic Sci Biotech 93(5):500–509

Hassanein AMA, Al-Soqeer AA (2018) Morphological and genetic diversity of Moringa oleifera and Moringa peregrina genotypes. Hortic Environ Biotechnol 59 (2):251–261

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PloS One 6(5):e19254

Kesanakurthi RP, Fazekas AJ, Burgess KS, Percy DM, Newmaster SG et al (2011) Spatial patterns of plant diversity below ground as revealed by DNA barcoding. Mol Ecol 20:1289–1302

Kumar P, Dolkar R, Manjunatha G, Pallavi HM (2017) Molecular fingerprinting and assessment of genetic variations among advanced breeding lines of Moringa oleifera L. by using seed protein, RAPD and Cytochrome P450 based markers. S Afr J Bot 111:60–67

Muluvi GM, Sprent JI, Soranzo N, Provan J, Odee D, Folkard G, McNicol JW, Powell W (1999) Amplified fragment length polymorphism (AFLP) analysis of genetic variation in Moringa oleifera Lam. Mol Ecol 8 (3):463–470

Ogden R, McGough HN, Cowan RS, Chua L, Groves M et al (2008) SNP based method for the genetic identification of Ramin (Gonystylus spp.) timber and products: applied research meeting CITES enforcement needs. Endanger Species Res 9:255–261

Patel JS, Narayana GV (1937) Chromosome numbers in some economic flowering plants. Curr Sci 5(9):479–479

Rajalakshmi R, Rajalakshmi S, Parida A (2017). Evaluation of the genetic diversity and population structure in drumstick (Moringa oleifera L.) using SSR markers. Current Sci 112(6):1250

Rajalakshmi R, Rajalakshmi S, Parida A (2019) Genetic diversity, population structure and correlation study in Moringa oleifera Lam. using ISSR and SRAP markers. Proc Natl Acad Sci India Sect B Biol Sci 89 (4):1361–1371

Ravi RD, Siril EA, Nair BR (2020) The efficiency of Cytochrome P450 gene-based markers in accessing genetic variability of drumstick (Moringa oleifera Lam.) accessions. Molecul Biol Rep 47 (4):2929–2939

Saini RK, Saad KR, Ravishankar GA, Giridhar P, Shetty NP (2013) Genetic diversity of commercially grown Moringa oleifera Lam. cultivars from India by RAPD, ISSR and cytochrome P 450-based markers. Plant Syst Evol 299(7):1205–1213

Tanksley SD (1983) Molecular markers in plant breeding. Plant Mol Biol Rep 1:3–8

Wu JC, Yang J, Gu ZJ, Zhang YP (2010) Isolation and characterization of twenty polymorphic microsatellite loci for *Moringa oleifera* (Moringaceae). HortScience 45(4):690–692

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11:1–16