Chapter 18
Molecular Markers in Date Palm

C. Cullis

Abstract Molecular markers are an increasingly important resource for all crops. DNA markers, especially those based on simple sequence repeats and single nucleotide polymorphisms, are playing an increasingly important role in plant variety identification, germplasm resource collection and breeding activities. The major types of DNA markers are described and the resources available to the date palm community are identified. In general, the molecular marker resources for date palm are somewhat limited. However, most of the available DNA marker types have been used on some material, mostly to cluster date palm varieties into related groups. The most profound effect on the development of the DNA marker resources for date palm is the newly available shotgun sequence. Mining this sequence database and the steady lowering of the costs of high throughput sequencing will increase rapidly the molecular marker resources and their application to date palm over the next few years.

Keywords DNA markers • Simple sequence repeats • Single nucleotide polymorphisms • Variety identification

18.1 Introduction

The development of molecular tools has changed the way in which individual varieties can be identified and useful information concerning the genetic control of many agronomic characteristics can be analyzed. The ability to apply these molecular tools depends to some extent on the amount of other genomic information available for the specific plant species. The overall molecular toolbox for date palm is

C. Cullis (✉)
Department of Biology, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7080, USA
e-mail: cac5@case.edu

S.M. Jain et al. (eds.), *Date Palm Biotechnology*,
DOI 10.1007/978-94-007-1318-5_18,

limited, although the resources including a draft shotgun genome sequence that recently has been developed, will allow a rapid expansion of the types of questions that can be asked. The molecular marker technologies reach back to the use of isozymes followed by a series of DNA marker technologies and most recently by possibilities to compare complete genomes. The focus here will be on the use of DNA fragments as markers for varietal identification, for the elimination of off-types from *in vitro* propagated date palms and for possible marker-assisted breeding for disease resistance. An increasing number of publications are appearing on the subject, including a recent review specifically focused on Tunisian date palm germplasm (Rhouma et al. 2010).

18.2 Evolution of Marker Resources: RFLPS to Complete Genome Comparisons

18.2.1 Restriction Fragment Length Polymorphisms

The difference in DNA fragment lengths for homologous sequences is the basis for most DNA marker technologies. The first of the DNA markers to be developed was based on restriction fragment length polymorphisms (RFLPs). As the name implies, these markers are based on assaying variations in the DNA sequence identified through the use of restriction enzymes. When a restriction enzyme is used to fragment genomic DNA then characteristic lengths of DNA between sites are the result. These fragments are resolved through the use of gel electrophoresis and the DNA fragments are transferred to nylon filters and then identified by hybridization to specific probes. The polymorphisms can be the result of various alterations in the genomic DNA sequence. The most usual cause is the loss of a recognition site for that restriction enzyme, thus preventing the enzyme from cutting the DNA, although in rare cases a new restriction site can be produced.

An alternative cause of fragment size variation is by the insertion or deletion of a DNA sequence between adjacent restriction sites. The most useful RFLPs have been in regions of single (low) copy number components of the genome, as these regions can also be useful genetic markers. If the polymorphism was generated by the loss of a restriction site, then two outcomes are possible following hybridization. If the probe is confined to a region of the genomic DNA completely within one of the restriction fragments then a larger band would be observed (Fig. 18.1, probe b). However, if the probe spanned the altered restriction site (Fig. 18.1, probe a) then a new band the size of the sum of the other two bands would be observed. When an insertion or deletion is the cause of the RFLP then no matter what probe is used only a single band is identified with different sizes in various individuals (assuming they are homozygous). Another indication that the RFLP is a result of an insertion or deletion is that an RFLP is identified with a single probe following digestion with a number of different restriction enzymes.

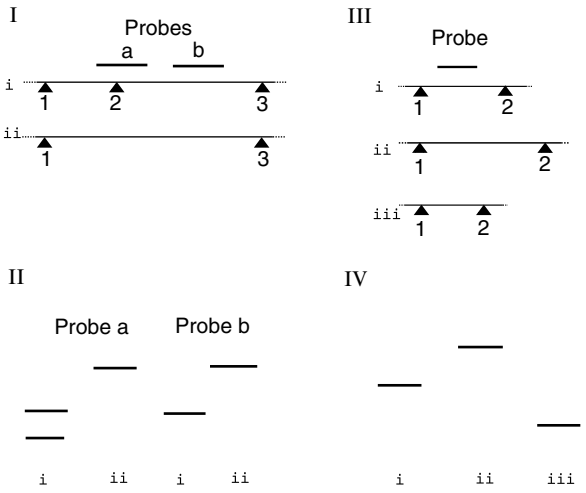

Fig. 18.1 The various outcomes of hybridization with different probes to identify the basis of various RFLPs. *I* – An RFLP caused by a loss of a restrictions site. *II* – The banding patterns after southern blotting and hybridization with probes a and b. *III* – RFLPs caused by an insertion (*ii*) or a deletion (*iii*) in the original allele (*i*). *IV* – The banding patterns after southern blotting and hybridization with the probe for these three alleles

Traditional RFLPs are identified through Southern blots and hybridization such that they are time-consuming and require large amounts of starting DNA. Therefore they are now rarely used in the standard form but have essentially been superseded by polymerase chain reaction (PCR)-based methods such as random amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs) and simple sequence repeats (SSRs). These methods, particularly AFLPs, are essentially RFLPs assayed using PCR methods.

18.2.2 Random Amplified Polymorphic DNA

The polymerase chain reaction (PCR) opened the door to many applications and has revolutionized DNA marker technology. However, the major limitation to applying PCR-based methods is that sequence information is required to design the primers to be used in the PCR. To overcome the need for sequence information it was determined from statistical calculations that a single ten base primer could be used in the PCR reaction and result in a small number of amplified bands that would be useful as molecular markers. Experimentally this theoretical analysis was confirmed to be very successful in identifying large numbers of polymorphisms (Williams et al. 1990) in plant and other genomes. Therefore, since hundreds of random decamers are commercially available, the method can be applied when no other genomic information is known. The technique does have some drawbacks.

First, the technique has been reported to suffer from irreproducibility between laboratories and sources of thermostable enzyme, although, within a laboratory, reproducible results can be achieved (Jones et al. 1997). Newer versions of the Taq polymerases have been developed which do not function as well at low temperatures, and therefore are not useful for producing RAPD patterns since the short

length of the primers requires an annealing temperature of about 40°C, rather than the more normal 60°C for most PCR reactions.

Second, the assignment of RAPD bands to a molecular map is genome specific, thus for every individual, a unique RAPD map needs to be produced since amplified fragments cannot be correlated across genomes.

Finally, RAPDs are dominant markers – a band is produced whether the fragment is homozygous or heterozygous so larger segregating populations are needed to produce accurate molecular maps. The basis for RAPDs can be a single nucleotide change in the primer sequence (since 1 base change is equivalent to a 10% mismatch in the primer sequence which will cause a 10°C change in the stability of the annealed primer) or insertion or deletion between the primer sites. Thus, overall, RAPDs are useful for DNA fingerprinting and identifying relationships between varieties, but less useful for genetic mapping.

In RFLP analysis, band size is the determining factor for identifying polymorphisms, but with RAPDs it is the presence or absence of a band. Since only one copy of an amplifiable sequence is necessary to result in an amplified band, RAPDs are dominant markers while RFLPs are generally codominant markers. For RFLPs the genomic change needs to be in the restriction site or as an insertion deletion. In the case of RAPDs single base changes anywhere in the primer will result in the loss of an amplified fragment as well as fragment size variation caused by insertions and deletions. In cases where the insertion is large, a loss of band will occur as the size of the amplifiable fragment exceeds that possible under the amplification conditions.

RAPDs have been applied to date palm variety identification (Adawy et al. 2002; Ahmed et al. 2006; Hussein et al. 2002; Sedra et al. 1998; Saker and Moursy 1998) and to the identification of somaclonal variation (Saker et al. 2006). These data were able to distinguish among varieties and place them in related groups, but no useful specific markers were found which could distinguish among somaclonal variants.

18.2.3 Amplified Fragment Length Polymorphisms

These markers are essentially RFLPs detected by PCR amplification (Vos et al. 1995). The primers used for the amplification are attached to the ends of the restriction fragments and then the complete genomic set is amplified. The method involves the digestion of the genomic DNA with two different restriction enzymes. Adaptors are then ligated to the ends of restriction fragments, with different adaptors added to each end. The two adaptors are then used as primers in a PCR reaction. The result of the amplification is a very large number of fragments. Any RFLP will result in a change in fragment size of a specific band, but this will be present against a background of all of the possible sized restriction fragments that can be amplified. The complex mixture of amplified bands needs to be separated on gels or by using automated DNA sequencers. Any polymorphic bands that are observed can be excised from the gel, cloned and sequenced to generate sequence-tagged sites (STSs).

These STS can be amplified with primers designed from the genomic sequence resulting in a single band which becomes a molecular marker. AFLPs are dominant markers since generally the presence/absence of a band is scored. However, the development of an STS will convert these regions into codomiant markers. The use of methylation-sensitive and -insensitive restriction enzyme isoschizomers allows for the analysis of potential epigenetic effects resulting from hyper- or hypo-methylated regions of the genome using AFLPs.

AFLPs have been used for date palm germplasm characterization (El-Assar et al. 2003, 2005). Forty-seven samples of date palm (*Phoenix dactylifera* L.) from Egypt were studied using four sets of amplified fragment length polymorphism (AFLP) markers (El-Khishin et al. 2003). A total of 350 bands were scored and of these 233 (66.6%) were found to be polymorphic. In another study AFLPs were used to survey genetic diversity in 40 ecotypes of date palm from oases in Tunisia (Rhouma et al. 2007). Six primer pairs resulted in a total of 428 AFLPs. The analysis of the data revealed that Tunisian date-palm germplasm has a large range of genetic diversity characteristics. The ecotypes were clustered into two groups that were independent of their geographic origin or the sex of the trees. This work has also been extended with the use of ISSRs (Zehdi-Azouzi et al. 2009). The study of regenerated plants to identify markers for somaclonal variation was carried out using AFLPs as well as RAPDS (Saker et al. 2006).

18.2.3.1 Microsatellites and SSRS

Microsatellites or SSRs are genetic markers that are derived from short (usually <6 bp) tandemly repeated sequences such as $(CT)_n$, $(AAT)_n$, $(GT)_n$. The terms microsatellite and SSR are often used interchangeably, although microsatellites are generally longer than the 2- to 3-bp unit of the SSRs. These SSR regions are widely dispersed through most animal and plant genomes and are also frequently polymorphic. The polymorphisms are due to the variability in the number of repeats that are present at a given site (Fig. 18.2). The length polymorphisms are produced by the changes in the number of the repeats such as that for a CT repeating unit as shown in Fig. 18.2. The three repeat classes shown in Fig. 18.2 are with 4 repeats, b with 2 repeats and c with 7 repeats. Since this SSR is a dinucleotide repeat, the actual size difference between the bands in a and b is 4 bp, that between a and c is 6 bp and that between b and c is 10 bp. These are small differences in length and therefore the separation of the amplified products needs to be done with high resolving power, frequently beyond the capability of agarose gels. Therefore the early ABI automatic sequencers are finding a renewed lease on life in the characterization of SSR length polymorphisms.

As with RFLPs, the primers for each SSR need to be isolated for each species. The current technology using the second-generation sequencing has made their isolation more rapid from genomic libraries or enriched genomic libraries (Panaud et al. 1996; Santana et al. 2009) or generated from an analysis of cDNA sequences. The enrichment techniques involve the use of biotinylated SSR fragments that are

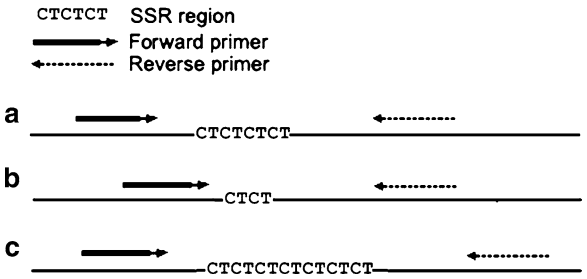

Fig. 18.2 Length polymorphisms resulting from differences in the number of SSR repeats at a locus

used to physically remove complementary sequences from a PCR amplified set of genomic fragments. The early iterations of the technique used the AFLP primers to provide the genomic representation. Initially all of the recovered fragments needed to be cloned and sequenced. However, using the 454 next generation sequencing methodology the complete set of fragments can be sequenced and assembled. The contigs are then analyzed using the program SSRIT (http://www.gramene.org/db/markers/ssrtool) to identify the contigs containing SSRs.

Where there is sufficient sequence surrounding the SSR region, primers are designed and the germplasm of interest is screened to find those which are polymorphic. Thus, although the identification of the SSRs regions is now much more rapid than before, each of them needs to be screened to find those that are polymorphic, which can be time consuming.

A second source of SSRs is from EST sequencing data (Ellis and Burke 2007). Frequently triplet repeats can occur in mRNAs since an additional amino acid will be inserted into the protein with no other changes in the structure and sequence of the protein. They can also be present in the 5′ and 3′ untranslated regions of the mRNA. The identification of these SSRs obviously needs a large EST sequence database and so is restricted to those species where a substantial sequence database is available. It is also possible that the level of polymorphisms present in EST-derived SSRs is lower than those identified from the genomic enrichment methods.

A modification of SSR markers is the inter-simple sequence repeat (ISSR) polymorphisms. ISSR is a general term for a region of the genome between microsatellite loci. They are assayed by using the complementary sequences to two neighboring microsatellites as the PCR primers. The variable region between these two regions is amplified, but by limiting the amplification parameters in the PCR reaction, the result is a mix of a variety of amplified DNA strands which can vary in length. The banding patterns frequently appear very similar to RAPD amplifications, and can therefore be used for DNA fingerprinting.

Data for date palm SSR and ISSR characterizations are available (Adawy et al. 2002; Zehdi et al. 2004; Zehdi-Azouzi et al. 2009). The genetic markers generated from seven selected ISSR primers were used to assess genetic diversity among a set of twelve Tunisian date palm varieties yielding 77 polymorphic markers which were sufficient to identify all of the varieties (Zehdi et al. 2004; Rhouana et al. 2009). The

data for all three marker types (RAPDS, AFLPs and ISSRs) gave similar phylogenies within the tested date palm varieties. Each of these methods can be used separately or together for developing relationships between date palm varieties.

18.3 Single-Nucleotide Polymorphisms

Single-nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence differs between two individual DNA samples. For example, a SNP might change the DNA sequence AGGATTCA to AGGATTTA. SNPs can occur in both coding (gene) and noncoding regions of the genome. As detailed above, SNPs are responsible for that class of RFLPs that result from the loss or gain of a restriction site. Many SNPs have no effect on cell function because they may not change protein structure (in fact, any SNP that occurs at the third position in the amino acid codon will have no effect if it does not change the amino acid sequence of the resulting protein). Their high frequency (perhaps as high as 2–3% in plant DNAs) means that they can be particularly useful in linkage mapping (Kristensen et al. 2001; Lai 2001). Again, the identification of SNPs depends on having sequence information available. Additionally, the sequence information needs to be obtained either from different varieties or from outbred individuals so that heterozygosity will be present.

Informatics tools can be used to compare the sequences and identify variations, but the raw data in the form of trace files may be important in deciding which polymorphisms may be real. When ESTs, for example, are built into unigene sets, any differences that may be present are eliminated in the formation of the consensus sequence and would need to be retrieved. Because there is no *a priori* way of differentiating between a true SNP and sequencing errors, each potential SNP must be validated. Even at a frequency of 1% these polymorphisms would generate an exceptionally large number of haplotypes if every polymorphism could be inherited independently. However, relatively few haplotypes are observed, indicating that perhaps the rate of SNP production is similar to the rate at which recombination occurs across the regions of the genome making up the haplotype blocks. Therefore, SNPs are most likely to be useful for defining haplotypes, rather than for their information individually, and so the use of SNPs is likely to involve linkage disequilibrium studies using the haplotype rather than the use of specific SNPs as individual molecular markers.

18.4 Genome Sequencing

The draft genome of date palm using second-generation sequencing has been reported for the Khalas cultivar. The draft genome, which constitutes approximately 20× coverage of the approximately 550 million base pairs genome, will contribute to a more complete understanding of date palm genetics as well as insights into

improving yield, quality and disease resistance. Availability of the draft genome sequence will also be exceptionally useful for identifying molecular markers for date palm. The complete sequence can be screened for SSR regions that can then be used to develop a full suite for both genetic mapping and varietal identification. Since date palms are outcrossing they will be heterozygous at most loci. Therefore inspection of the sequence can reveal SNPs that are apparent in the assembly of the sequence information. In the announcement of the draft genome about 850,000 new SNPs have been identified from the comparisons between parental alleles present in the cultivar Khalas. In the reporting of the data, the manual inspection of the assembled contigs appear to be consistently correct for scaffolds that are 12,000 bases or less while the longer range assembly which is developed for spanning gaps is less certain. This is likely due to the interspersion of retrotransposable elements within the date palm genome that makes the assembly of short read shotgun sequencing problematical in the absence of any additional genomic resources such as BAC sequences. The date palm draft genome is available online at: http://qatar-weill.cornell.edu/research/datepalmGenome/download.html. This resource is freely available and should therefore provide for a rapid increase in the availability and use of date palm markers.

18.5 Conclusion and Prospective

It is clear that the date palm genome is structured similarly to that of other characterized plants. Therefore all the tools that have been developed for using DNA markers are available. Preliminary studies have demonstrated that population structures and lineage relationships can be identified with the current crop of DNA markers. As noted, the availability of the complete genome sequence will facilitate the development of a suite of different marker types to be applied appropriately. The development of a series of sequenced tagged sites (probably based in SSRs) will supply resources needed for the screening of collections to reduce the number of samples kept in germplasm banks. They will also add impetus to identifying markers linked to the various disease-resistant genes. With the steady increase in the sequencing resources, SNPs will also become more useful but the relative costs of SNP and SSR analyses may well determine which of the two-marker systems becomes most widely used. Although few publications using molecular markers are currently available, it is expected that this literature will substantially increase over the next few years.

Prospects for the application of molecular markers to date palms are still very minimal. The availability of the shotgun sequence and the steady lowering of the costs of high throughput sequencing will increase the resources and their application rapidly over the next few years. As with many plant species, decisions will need to be made concerning the level of whole genome sequencing compared to targeted re-sequencing as the most efficient method for useful applications. It is undoubted that the collection of many high polymorphism information content SSR primer

pairs and validated SNPs will provide the tools for phylogenetic analyses as well as germplasm conservation. However, once genomic regions associated with important characteristics such as disease resistance, taste and post-harvest stability, the sequencing of these regions and the identification of the actual bases for these characteristics can be incorporated into the breeding and improvement programs. The identification of off-types arising in tissue culture propagation and the complete genome sequencing of normal and off-type individuals will lead to the identification of both markers for assessing off-type individuals in the regenerated plants as well as the 'mutations' responsible for these off phenotypes. Therefore these molecular markers and the tools developed through their use will facilitate the improvements in available germplasm for increasing the area under date palm cultivation as well as for the overall improvement of the plant material available to growers.

References

Adawy SS, Hussein EHA, El-Khishin D, Saker MM, El-Itriby HA (2002) Genetic variability studies and molecular fingerprinting of some Egyptian date palm (*Phoenix dactylifera* L.) cultivars II-RAPD and ISSR profiling. Arab J Biotechnol 5:225–236

Ahmed MMM, Soliman SS, Elsayed EH (2006) Molecular identification of some Egyptian date palm males by females varieties (*Phoenix dactylifera L.*) using DNA markers. J Appl Sci Res 2:270–275

El-Assar AM, Krueger R, Devanand PS, Chao CT (2003) Genetic analyses of date palms (*Phoenix dactylifera* L.) from Egypt using fluorescent – AFLP markers. HortScience 38:734

El-Assar AM, Krueger R, Devanand PS, Chao CT (2005) Genetic analysis of Egyptian date (*Phoenix dactylifera* L.) accessions using AFLP markers. Genet Resour Crop Evol 52:601–607

El-Khishin DA, Adawy SS, Hussein EHA, El Itriby HA (2003) AFLP fingerprinting of some Egyptian date palm (*Phoenix dactylifera* L.) cultivars. Arab J Biotechnol 6(2):223–234

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. Heredity 99:125–132

Hussein EHA, Adawy SS, El Khishin D, Moharam H, El-Itriby HA (2002) Genetic variability studies and molecular fingerprinting of some Egyptian date palm (*Phoenix dactylifera* L.) cultivars. 1-A preliminary study using RAPD markers. Arab J Biotechnol 5(2):217–224

Jones CJ, Edwards KJ, Castaglione S et al (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol Breed 3:381–390

Kristensen VN, Kelefiotis D, Kristensen T, Borresen-Dale AL (2001) High-throughput methods for detection of genetic variation. Biotechniques 30:318–321

Lai E (2001) Application of SNP technologies in medicine: Lessons learned and future challenges. Genome Res 11:927–929

Panaud O, Chen X, McCouch S (1996) Development of microsatellite markers and characterization of simple sequence length polymorphism (SSR) in rice (*Oryza sativa* L.). Mol Gen Genet 252:597–607

Rhouma S, Zehdi-Azouzi S, Salem AOM et al (2007) Genetic diversity in ecotypes of Tunisian date-palm (*Phoenix dactylifera* L.) assessed by AFLP markers. J Hort Sci Biotechnol 82:929–933

Rhouma S, Ould Mohamed Salem A, Zehdi-Azouzi S et al (2009) Comparative analysis of genetic diversity in Tunisian date palm (*Phoenix dactylifera* L.) as revealed by RAPDs and AFLPs. Acta Hort 814:125–130

Rhouma S, Zehdi-Azouzi S, Dakhlaoui-Dkhil S et al (2010) Genetic variation in the Tunisian date palm (*Phoenix dactylifera* L). In: Ramawat KG (ed.) Desert plants. Springer, Berlin, pp 355–370

Saker MM, Moursy HA (1998) Molecular characterization of Egyptian date palm: 11 RAPD fingerprints. Proceeding first international conference on date palms, Al-Ain, pp 173–182

Saker MM, Adawy SS, Mohamed AA, El-Itriby HA (2006) Monitoring of cultivar identity in tissue culture-derived date palms using RAPD and AFLP analysis. Biol Plant 50:198–204

Santana QC, Coetze MPA, Steenkamp ET et al (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. BioTechniques 46:217–223

Sedra MH, Lashermes P, Trouslot P et al (1998) Identification and genetic diversity analysis of date palm (*Phoenix dactylifera* L.) varieties from Morocco using RAPD markers. Euphy 103:75–82

Vos P, Hogers R, Bleeker M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Williams KJ, Kubelik A, Livak K et al (1990) DNA polymorphism amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Zehdi S, Sakka H, Rhouma S et al (2004) Analysis of Tunisian date palm germplasm using simple sequence repeat primers. Afr J Biotechnol 3:215–219

Zehdi-Azouzi S, Rhouma S, Ould Mohamed Salem A et al (2009) Comparative analysis of genetic diversity in Tunisian collections of date palm cultivars based on random amplified polymorphic DNA and inter simple sequence repeats fingerprints. Acta Hort 814:149–156