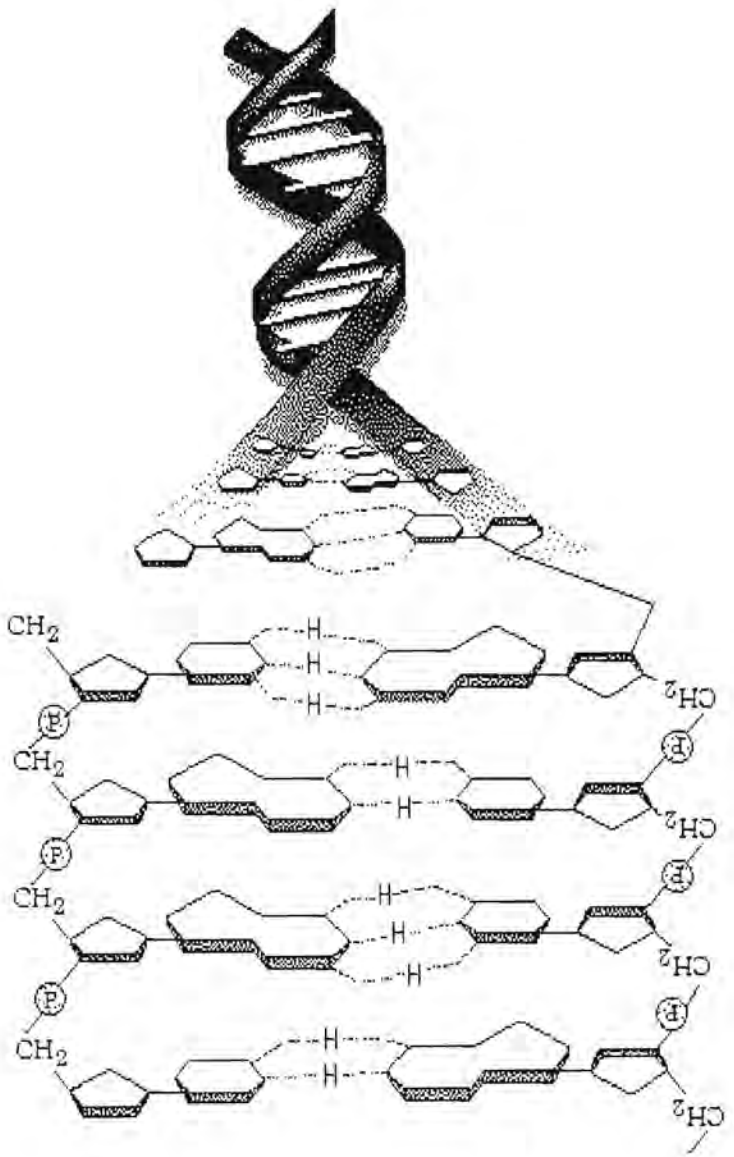
Chapter 1

The history of molecular biology and detection of genetic variation



1.1 Abstract

The past century has witnessed a breathtaking array of discoveries in the biological sciences. This is in particular in the general area of molecular biology, the scientific discipline that seeks to fully understand the molecular basis of heredity, genetic variation, and the expression patterns of individual units of heredity called genes. To fully appreciate the current status of molecular biology and where it is heading it is important to understand the early principals and the theories that gave rise to it. In this chapter a brief history of the development of the field of molecular biology and the basic philosophy on which our understanding of biological systems is based is presented. Furthermore an overview of the plant genome and its composition is given. The mechanisms that is responsible for creating variation in genomes and the most popular techniques molecular biologists use to find, study and utilize these changes are presented.

1.2 Biology: then and now

Building on observations about natural plant variation in time and space, early geneticists, such as Darwin, Mendel, and Vavilov, posed fundamental questions regarding the origin, structure, and evolution of genetic diversity. They postulated that an underlying reservoir of innate and heritable genetic possibilities delineated the options for growth, development, and reproduction of organisms at both the individual and population level. The field of plant genetics today continues to address many of the same questions while integrating new developments in molecular biology and bioinformatics. Over the last 15 to 20 years, new, highly automated tools have created unprecedented opportunities for generating and analysing large biological data sets. The systematic processing of nucleic acid and protein sequence information from many different organisms has further fundamentally changed the way that biologists approach the study of living things (McCouch, 2001).

In retrospect, the 6th decade of the 19th century was truly remarkable with respect to the development of the science of biology. By the end of those years all of the pieces were in place for the maturation of what had been a purely observational discipline into one with a theoretical basis. The result, the field of molecular biology and its attendant sub-disciplines, is grounded philosophically in a mechanistic, deterministic and reductionistic view that derives from the logical empiristic setting in which it was

born (Judson, 1996). In November 1859, Charles Darwin first published his "*On the Origin of Species*" and in 1866 Fr. Gregor Mendel his *"First and Second Laws of Heredity*" as well as his conceptualisation of the gene as the unit of inheritance. By 1869, the chemist JF Miescher isolated from human pus the substance he named nuclein, which was later called nucleic acid. At the time of these events, biology could hardly be compared to the so-called 'hard' sciences such as physics and chemistry, which rested upon strong theoretical platforms. Biology had essentially been an exercise in observation and classification. But with Darwin's theory and Mendel's laws, biology had for the first time a potential theoretical basis of it's own.

Modern biology has had great success in representing the reality of living systems in a form that yields a great deal of both theoretical and practical information. Our current understanding of all of the mechanisms by which these macromolecules act and interact as a part of the life functions of an organism derive from the reductionist paradigm and the techniques it has produced. And yet, there are clearly aspects of living systems that have not yielded to this analysis. Among these features are included the organization of the human brain and, more importantly, the origin of the mind and consciousness. It has become increasingly difficult to model these kinds of natural phenomenon in terms of the reductionist paradigm extent in much of biological thinking.

A major part of the difficulty appears to be the framework within which the natural world is viewed in modern biology. As an inheritor of the logical 'empiricists' position, the biologist believes that the only aspects of the world that are observable and open to rational investigation. In fact, relevant are those that could be, in Aristotelian terms, material cause (the set of objective possibilities) and formal cause (the shape or form of the substance). However, modern biology mostly does not recognise nor does it incorporate into theoretical considerations either the final cause (action of will) or the efficient cause (projection into reality of this act of will) (Ayala, 1970). The problem is a science that embraces Hume's assertion that the efficient and final causes of a thing can never be known. But consider for a moment the nature of chance or random events and the mistake becomes evident in this chain of reasoning. The neo-Darwinist assumes that the random nature of mutational changes in DNA eliminates the causality from the consideration. In fact, the mechanisms of mutation are well understood and proceed by quite specific steps. In a world where no system is entirely closed, that is, isolated from its surroundings, blind chance cannot exist: there are interconnections that provide a complex, if sometimes subtle, input into all

events (Laszlo, 1996). The overall result can be seen as having both an efficient cause (the agent of mutation) and a final cause (the consequence of the mutated gene on the function of the organism). Although the biologist cannot predict which gene will be mutated in a particular organism, the dependence of the functioning of that organism on that change can be observed (Ayala, 1970).

Explanation by design, or teleology, is "the use of design, purpose, or utility as an explanation of any natural phenomenon" (Webster's Third New International Dictionary, 1966). An object or behaviour is said to be teleological when it gives evidence of design or appears to be directed towards certain ends. Teleological explanations account for the existence of a certain feature in a system by demonstrating the feature's contribution to a specific property or state in the system. The idea of teleological thinking is viewed as a slur when applied to any kind of biological model or conclusion. Nonetheless, purpose and plan are obvious in living systems. Some time ago Francisco Ayala pointed out three types of teleological explanations that are appropriate to biological systems: (1) conscious anticipation of an end-state or goal, (2) self-regulating systems and (3) structures designed to perform a specific function. As Barbara McClintock (1984) expressed it: There are 'shocks' that a genome must face repeatedly, and for which it is prepared to respond in a programmed manner. An example is the 'heat shock' response in eukaryotic organisms. Each of these initiates a highly programmed sequence of events within the cell that serves to cushion the effects of the shock. Some sensing mechanisms must be present in these instances to alert the cell to imminent danger, and to set in motion the orderly sequence of events that will mitigate the danger. But there are also responses to the genome to unanticipated challenges that are not so precisely programmed. The genome is unprepared for these shocks. Nevertheless, they are sensed, and the genome responds in a discernible but initially unforeseen manner.

And as the co-holder of the 1945 Nobel prize for physiology and medicine, Sir Ernest Chain put it: "To postulate that the development and survival of the fittest is entirely a consequence of chance mutation seems to me a hypothesis based on no evidence and irreconcilable with the facts. These classical evolutionary theories are a gross over-simplification of an immensely complex and intricate mass of facts, and it amazes me that they are swallowed so uncritically and readily and for such a long time, by so many scientists without a murmur of protest." Teleology asserts that a casual non-physical agent as postulated by vitalism is purposeful and that there is purpose and design in nature.

1.3 The plant genome: size and organisation

DNA re-association kinetics studies showed that non-transcribing repeat (NTR)-DNA is an integral part of most plant genomes and its amount is proportional to the genome size (Flavell *et al.*, 1974). Most plant genomes are large and complex and NTR-DNA is primarily composed of retro-transposons (Bennetzen *et al.*, 1998). The composition of plant NTR-DNA seems to be a result of multiple invasions by different retro-transposons. Replication of retro-transposons then occurred followed by their inactivation by transposition and/or hetero-chromatization (Sandhu and Gill, 2002). The NTR-DNA is unevenly distributed in the plant genomes. Paucity of genes observed from physical maps (Sandhu *et al.*, 2001) and an abundance of hetrochromatin visualized as C-bands (Curtis and Lukaszewski, 1991; Gill *et al.*, 1991; Jiang *et al.*, 1996) allegorise that repetitive DNA is especially abundant around the centromeric regions (Copenhaver and Preuss, 1999). Regions present between two gene-rich regions are also composed of NTR-DNA as well as regions present near the tip of chromosome arms deficient in genes (Sandhu and Gill, 2002). In addition to retro-transposons, psuedo-genes also seem to be an important part of the non-transcribed regions.

Whether transposons, retro-transposons, and other repetitive elements accumulate extensively in a given evolutionary lineage may depend on several factors. Among them are the efficiency or repressive mechanisms and the rate at which the sequences undergo mutational and deletional decay. For example, methylation of C residues enhances the mutability of CG base pairs hence methylation accelerates the divergence rate of newly arising duplications. This happens at an extreme form in Neurospora, in which many methylated CGs are mutated in the span of a single generation, and at more measured rates in plants and mammals (SanMiguel *et al.*, 1998; Wang *et al.*, 1998). If multiple copies of a DNA sequence are present in a genome we can think of each sequence as a single 'species' evolving on it's own 'line of decent' because each repeat will be mutated at random.

Retro-transposons display a high degree of sequence variability (Casacuberta *et al.*, 1995; Marillonnet and Wessler, 1998) and in most cases they represent elements that have lost the ability to transpose. It is also known that both the *Ac* and *Spm* transposons of maize frequently give rise to internally deleted elements, and *Ac* ends

are much more abundant in the maize genome than are full-length elements, suggesting deletional decay of transposon sequences (Fredoroff *et al.*, 1983,1984; Masson *et al.*, 1987).

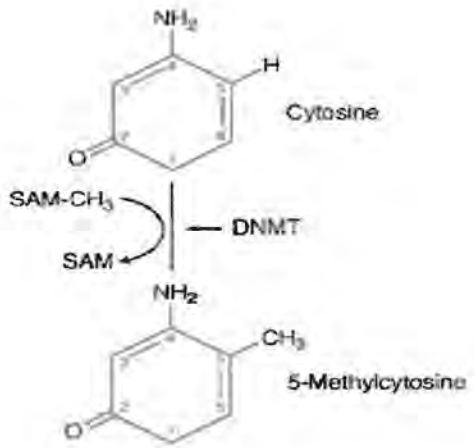


Figure 1.1. Mechanism of DNA methylation. 5-Methylcytosine is produced by the action of the DNA methyltransferases (DNMT), which catalyse the transfer of a methyl group (CH_3) from *S*-adenosylmethionine (SAM) to the carbon-5 position of cytosine (Strathdee and Brown, 2002).

1.3.1 Functional and non-functional sequences

Sequences within the genome can be classified according to a number of criteria. The most important of these is functionality and the largest class of functional DNA elements (also known as euchromatin) consists of coding sequences within transcription units. For the most part, transcription units correspond one-to-one with Mendelian genes. They usually function on behalf of the organism within which they are contained. The functional class of DNA elements also includes a number of specialized sequences that play roles in chromosome structure and transmission. The best-characterized structural elements are associated with the centromeres and telomers (Figure 1.2) (Fitzgerald-Hayes *et al.*, 1982; Bloom and Carbon, 1982; Sun *et al.*, 1997; Wright *et al.*, 1996; Pardue *et al.*, 1996; Zakian 1996).

Most of the genome appears to consist of DNA sequences that have no apparent function. This non-functional class includes psuedo-genes that derive from specific

genes but are not themselves functional with a lack of transcription or translation. For the most part, however, non-functional DNA, also known as heterocrhomatin, is present in the context of long lengths of apparently random sequence, and repetitive elements with origins that have long since become indecipherable as a consequence of constant genetic drift.

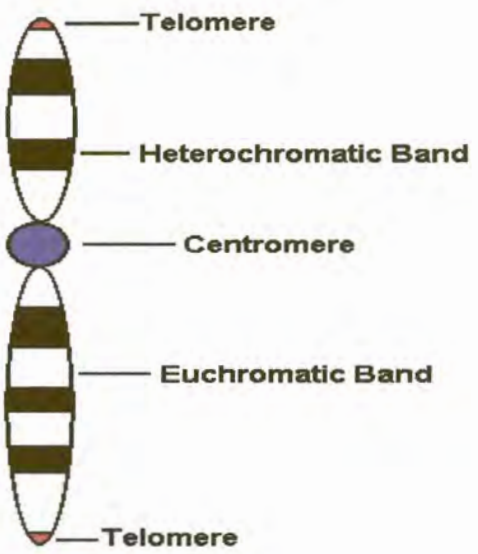


Figure 1.2. Diagram indicating different structural elements found on the chromosome (Brown, 1992).

Both functional and non-functional sequences can be distinguished by a second criterion – copy number. Sequences in a genome that do not share homology with any other sequences in the same genome are considered as single copy elements. This single copy class contains both functional and non-functional elements. Sequences that do share homology with one or more other genomic regions are considered to be repeated or multi-copy elements. The most abundant multi-copy elements found within the genome of plants are retro-transposons. Highly repetitive DNA tends to accumulate only in regions of low recombination, such as the centromeres and telomeres, were recombination is suppressed. In contrast, repeats occurring in euchromatin are much more susceptible to crossing-over and tend to be more variable in copy number relative to their array length. Much of moderately repeated DNA consists of transposable elements. The two major families, the long and short interspersed nucleotide elements (LINEs and SINEs), have significant roles in genome function and evolution.

1.4. Sources of variation: Genomic elements, genome variation and evolution

Towards the end of the sixties of the last century it became clear that the genome of eucaryotes contain, in contrast to the prokaryotic genome, a high percentage of non-coding, usually repetitive nucleotide sequences. Repetitive DNA consists of a repeated sequence of a certain size (the repeat unit) with a given copy number organized in a particular manner in space. Repeat units can be organized in three ways: (1) tandem repeats have no spaces between individual repeat units (Figure 1.2), (2) hyphenated repeats are separated by small gaps but are still grouped together and (3) dispersed repeats are spread throughout the genome. There is obviously nothing remarkable about dispersed repeats of one, or even two or three nucleotides. Hence, dispersed repeats are only significant when they involve reasonably long DNA sequences, which occur substantially more frequently than would be expected by chance. Some genome-wide dispersed repetitive DNA corresponds to members of multigene families, comprising both functional genes and pseudo-genes. Otherwise, it may represent motifs that function at the DNA level. Most dispersed repetitive DNA, however, corresponds to either functional transposable elements or their remains. Today, we know that transposons constitute a large fraction – even a majority – of the DNA in some species of plants and animals, among them mice, humans, and agriculturally important plants such as maize and wheat.

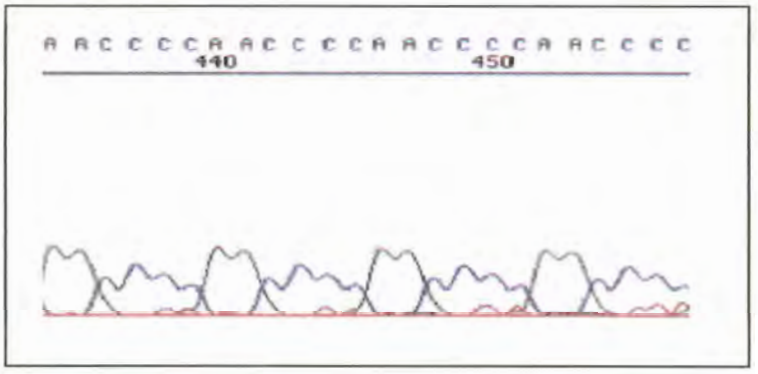


Figure 1.3. DNA sequence showing a tandem AACCC repeat

1.4.1 Retro-elements and genome variation

Retro-elements have been found in the genomes of all plant species that have been examined. But they seem to be highly abundant only in species with large genomes. This suggests that retro-elements, particularly retro-transposons, account for most of the great variation in plant genome size (SanMiguel *et al.*, 1996). Bennetzen and Kellogg (1997) raised the question if retro-transposons are the largest single component of many flowering plant genomes, because not all plant genomes have expanded with the amplification of these elements. Retro-elements transpose without excision and their mobility will always increase their copy number and thereby increase the genome size (Figure 1.4). Therefore, the question has to be asked does continuous or episodic retro-element amplification mean that all plants are on the road to larger genomes, or is there an active process for removing these interspersed repetitive DNAs from plant genomes?

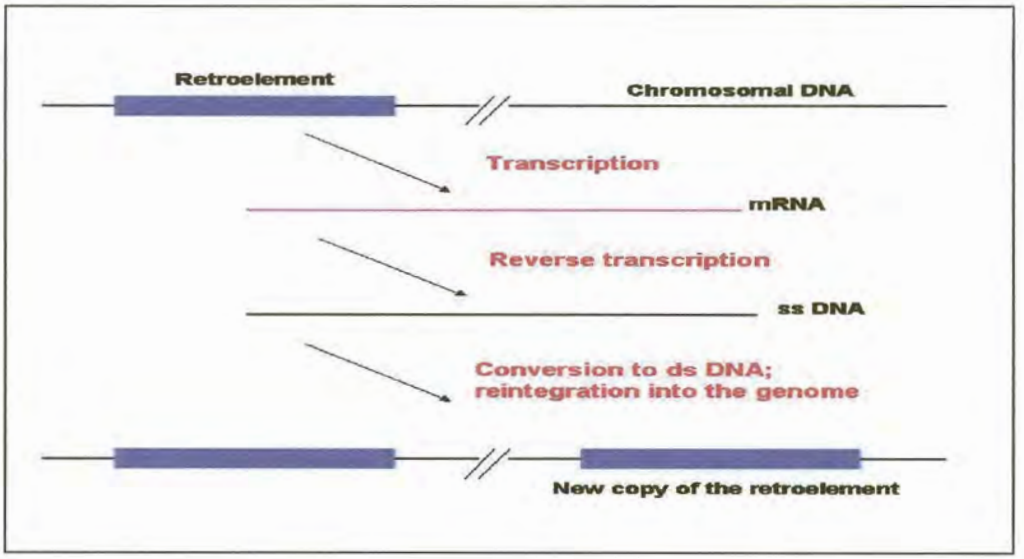


Figure 1.4. Duplication of retro-transposons in the genome *via* a RNA intermediate (Brown, 1992).

Logic dictates that there should be an upper limit to the level of transposition that genomes can endure. It is therefore quite surprising to discover the magnitude of transposable elements in genomes of plants, such as maize, which together make up more than 50% of the genome (SanMiguel *et al.*, 1996). The presence of multiple low-copy number families further indicates that many hundreds, if not thousands, of

distinct retro-element families exist in maize (Voytas, 1996). But how can a genome function with such a large burden of retro-elements? It is of course, in the element's best interest to minimize genetic damage caused by integration, because the host's survival is necessary for persistence of the element. In the yeast *Saccharomyces cerivisiae*, in which retro-transposons have been studied extensively, it appears that the five families of retro-transposons found within the *S. cerevisiae* genome, have a strong bias for sites in the genome where they integrate. Retro-elements are found particularly upstream from tRNA genes or at the telomeres. Regions targeted by yeast retro-transposons are typically devoid of open reading frames (Voytas, 1996).

The organisation of retro-elements in the interspacer regions of maize is reminiscent of the findings in yeast. Targeted integration, as opposed to amplification by recombination, is suggested by the overall structural integrity of the retro-elements and the presence of intact target-site duplications flanking most insertions. The under-representation of the most highly abundant retro-element families in the maize DNA sequence databases further suggests that these elements specifically avoid coding regions or that their presence near genes has been strongly selected against. Intergenic regions are hyper-methylated relative to gene sequences (Bennetzen *et al.*, 1994). By extrapolating from the yeast model, one might predict that some such unique chromatin feature serves as a homing device for maize retro-elements during integration. Hyper-methylated arrays of retro-transposons within retro-transposons have also been observed in the slime mold *Physarum polycephalum* (Rothnie *et al.*, 1991). This suggests that targeted integration may be a widespread strategy adopted by retro-elements to proliferate within host genomes (Voytas, 1996). According to this model, the increased number of retro-transposons in large-genome lineages may be due part to an increase in the number of possible non-deleterious insertion sites in the genome.

The fundamental issue is the likelihood of changes in genome size over evolutionary time and, in particular, the likelihood of decreases in genome size versus increases. Although increases - via amplification of retro-transposons - are clearly possible and apparently easy, decreases may be more difficult and/or may occur less frequently. The striking variation in genome size observed between closely related species has been termed the C-value paradox (Thomas, 1971), meaning that it is paradoxical that genomic complexity (i.e., size) does not correlate with biological complexity of the organism. Some genome-size variation is due to the polyploidy commonly found in the angiosperms or to tandemly repeated satellite sequences, but most is associated

with ill-defined classes of interspersed highly repetitive and middle repetitive DNAs (Flavell *et al.*, 1974). Recent studies have indicated that the majority of this reiterated DNA is composed of retro-elements (Moor *et al.*, 1991; SanMiguel *et al.*, 1996; Smyth *et al.*, 1989)

The question still stands as to whether all plant genomes are destined for genome obesity o if an active process for removing these interspersed repetitive DNAs from plant genomes exists? Studies on spontaneous mutations have shown that deletions are more frequent and longer than insertions. For example, in mammals, deletions are three to seven-times more frequent than are insertions and are, on average, somewhat longer (Graur *et al.*, 1989). In *Drosophila*, the difference is even more profound – deletions are almost ten-times more frequent and almost seven-times longer than are insertions (Petrov *et al.*, 1996). This biases in mutation frequency and size will lead to the progressive elimination of nonessential sequences. Admittedly, this process is very slow.

It is also known that both the *Ac* and *Spm* transposons of maize frequently give rise to internally deleted elements, and *Ac* ends are more abundant in the maize genome than are full-length elements, suggesting deletional decay of transposon sequences (Fedoroff *et al.*, 1983,1984; Masson *et al.*, 1987; Schwarz-Sommer *et al.*, 1985). It would not be surprising to find mechanisms that preferentially eliminate sequences, as has been found in wheat for the preferential loss of non-redundant sequences early after polyploidization (Feldman *et al.*, 1997). Whether transposons, retro-transposons, and other repetitive elements accumulate in a given evolutionary lineage may depend on several factors, among them the efficiency of repressive mechanisms and the rate at which the sequence undergo mutational and deletional decay (Fedoroff, 2000).

1.4.2 Stress and genome variation

Growth conditions are seldom optimal and when the environment changes an organism must be able to adapt in order to survive or die. Any environmental change that results in a response of an organism that is less than optimal might be considered as stressful (Levitt, 1972). Stress factors can be either biotic, imposed by other organisms, abiotic, arising from an excess or deficit in the physical or chemical

environment or of a genetic nature for example the introduction of foreign genetic material and viral infections. When plants are grown in artificial conditions additional stress might be introduced by the synthetic growth medium or additional growth regulators and antibiotics added to the plants' environment (Buchanan *et al.*, 2000).

Many of the plant retro-transposons studied to date are transcriptionally activated by various biotic and abiotic stress factors (Grandbastien, 1998). The expression of tobacco *Tnt*1 and *Tto*1 retrotransposons is greatly increased by several abiotic stresses such as cell culture, wounding, methyl jasmonate, $CuCl_2$ and salicylic acid (Hirochika, 1993; Mhiri *et al.*, 1997; Moreau-Mhiri *et al.*, 1996; Pouteau *et al.*, 1994; Takeda *et al.*, 1998, 1999). Similarly, various biotic stress factors, such as infection with various viral, bacterial or fungal pathogens (Mhiri *et al.*, 1999; Pouteau *et al.*, 1994), have been shown to activate transcription of these retrotransposons. In contrast to *Tnt*1 and *Tto*1, transcription of *Tos*17 is induced only by tissue culture.

Many factors determine how plants react to these stresses. The genetic make-up of the plant, its developmental state, the duration and severity of the stress, the number of times the plant is subjected to stress and any synergistic effects of multiple stresses influence the plants' response. In response to stress some genes are expressed more strongly, whereas the expression of others are repressed. Several mechanisms, such as quantitative modification of repetitive DNA, DNA methylation, excision and insertion of transposable elements, gene amplification or deletion and histone acetylation have been suggested as points of control for these changes (Capy, 1998; Cullis 1990; Johnson *et al.*, 1996).

1.4.3 Somaclonal variation

Somaclonal variation is a widespread phenomenon in tissue culture plants. The term, somaclonal variation, was first defined by Larkin and Scowcroft (1981) as the genetic variation in plants regenerated from the tissue culture process. In all organisms, spontaneous mutations occur from generation to generation, but somaclonal variation describes the additional mutations in plants produced via tissue culture (Bouman and De Klerk, 1997). This unexpected source of variability was once hailed as a "novel source of variation for crop improvement", but due largely to its unpredictability as a breeding tool, enthusiasm for this application has diminished and somaclonal variation has lost much popularity in recent years (Karp, 1993).

Somaclonal variation and its causes are not well understood and have not been elucidated. Plants are generated by a series of cell divisions in meristematic tissues. During the early stages of embryogenesis the apical meristem is formed from the zygote. The axillary meristem in turn originates from the apical meristem. However, new apical meristems may also originate from non-zygotic cells, in particular from somatic cells from callus or cell suspension cultures. Plants generated from these adventitious meristems are often genetically different from the mother plant (Bouman and De Klerk, 1997). Somaclonal variations are therefore often associated with callus formation (Skirvin and Janick, 1976) and the use of growth regulators, such as 2,4-dichloropenxyacetic acid (2,4-D) and benzyladenine (BA) (Figure 1.5) have been reported to play an important role in the induction of variability (Evans, 1988). Linacero *et al.* (2000) found hot spots of DNA instability in rye plants generated from immature embryos. At least 40% of the studied rye plants showed at least one variation, and the number of mutations per plant was quite high, ranging from 2 to 12. In 2001, Leroy *et al.* found in a study done on cauliflower calli using microsatellite primers, only six calli out of a total of 224 with stable original DNA patterns. Many more examples of somaclonal variation have been found in plants grown *in vitro* including beet plants (Sabir *et al.*, 1992), red clover (Nelke *et al.*, 1993), and oilseed rape (Poulsen *et al.*, 1993).

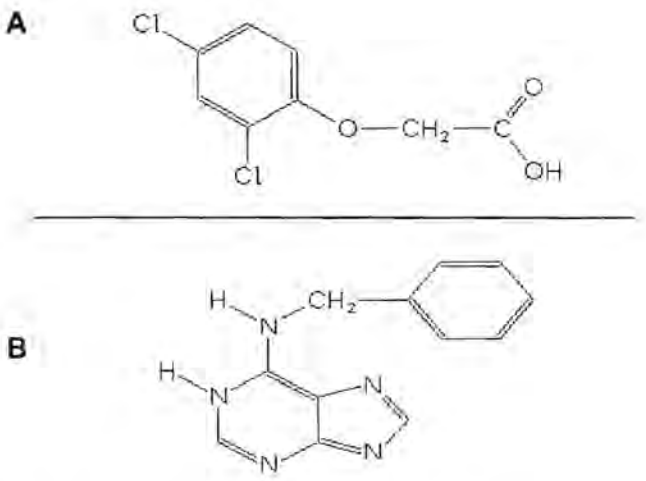


Figure 1.5. (A) The molecular structure of 2,4-dichloropenxyacetic acid (2,4-D) and (B) the structure of benzyladenine (BA).

1.4.4 Genomic re-arrangements and somaclonal variation

Many of the mechanisms by which the genome is re-organized due to stress have been observed in cells in tissue culture or in plants regenerated from such cultures. Reviews discuss a number of reasons why the tissue culture environment can be stressful to plant cells. They focus especially on the physical nature of the medium, such as high salt, water stress, mineral deficiency, an excess of metal ions and overexposure to auxins (Phillips *et al.*, 1994; Skirvin *et al.*, 1994; Cullis 1999; Arnault and Dufournel, 1994). Genomic changes in tissue culture can result in changes in ploidy level, such as aneuploidy, chromosomal rearrangements (Figure 1.6), activation of transposable elements and other genes (Figure 1.7), point mutations, genome re-arrangements, methylation changes and even altered copy number of sequences (Cullis, 1990; Peschke *et al.*, 1987; Hirochika, 1993; Phillips *et al.*, 1994). Blundy *et al.* (1987) found an almost three-fold reduction in the ribosomal RNA genes in callus cultures of flax. Lee and Phillips (1986) detected chromosomal instabilities in *in vitro* grown maize plants. The extent of these chromosome abnormalities was found to be dependent on the time the cells had been in culture (Chandler *et al.*, 1986). Translocations are a commonly observed chromosome abnormality with inversions, insertions and deletions occurring in the DNA sequence. Repetitive DNA sequences are especially sensitive to stress-related DNA changes and account for a large portion of variation in sequence copy numbers. When cultured *Cymbidium* protocorms were exposed to auxin, the amplification of AT-rich satellite DNA was observed, while exposure to gibberellic acid increased GT-rich regions (Nagl and Rucker, 1976). Zheng *et al.* (1987) found that in rice suspension cultures, highly repeated sequences were amplified up to 75-fold. Another representative of highly repetitive sequences, the ribosomal RNA DNA sequences (rDNA) are also part of the variable component and a decrease in ribosomal RNA genes has been reported in flax callus cultures (Blundy *et al.*, 1987).

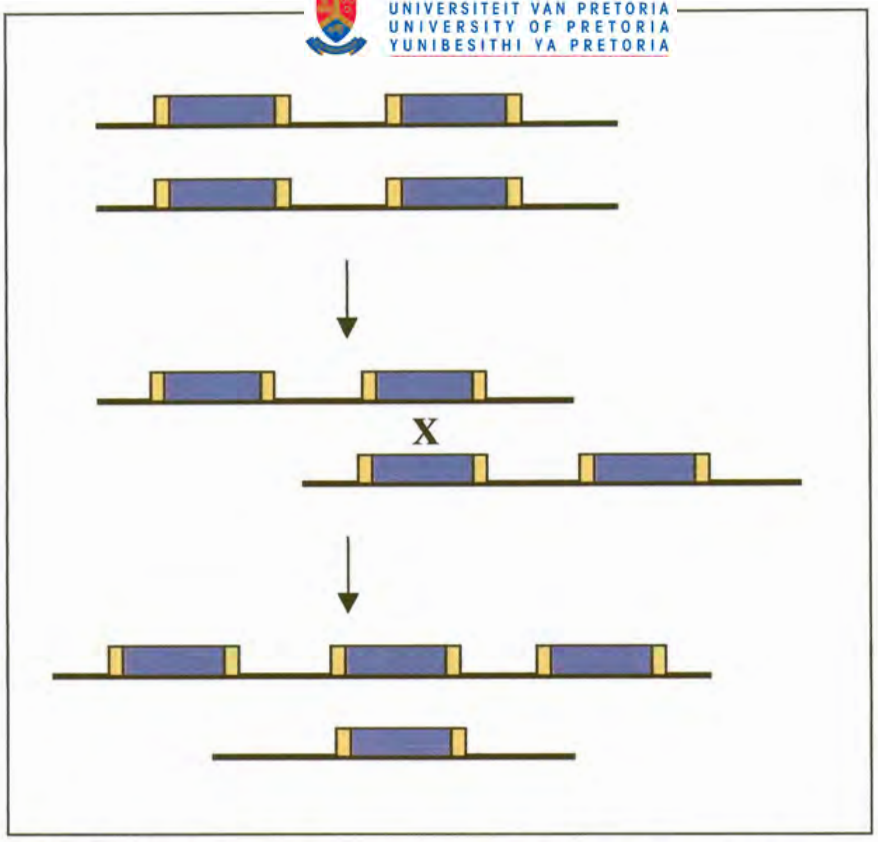


Figure 1.6. Chromosome rearrangements caused by unequal crossing over of transposable elements.

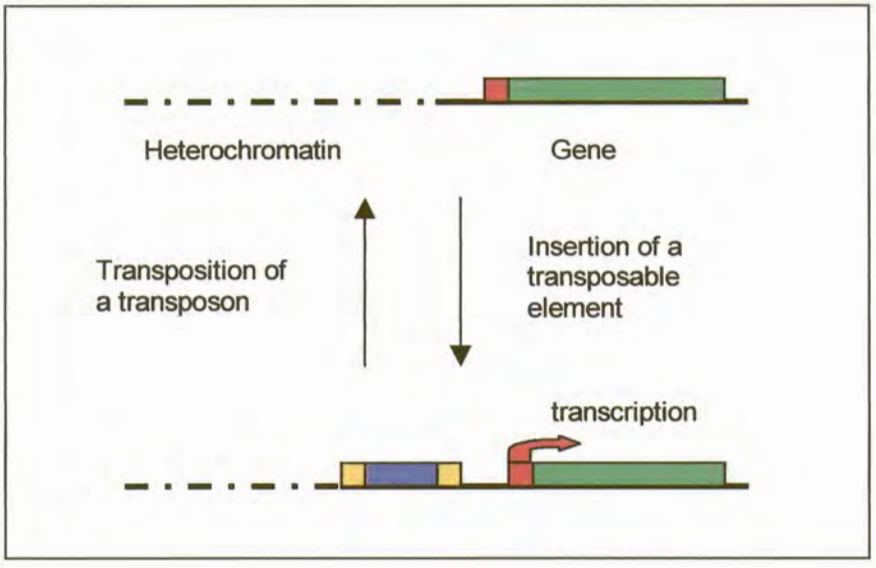


Figure 1.7. Gene activation/deactivation by transposition. Genes located near heterochromatin may be inactive due to the DNA structure. Insertion of a transposable element may move the gene further away from heterochromatin and thereby activating it. Transposition of such an element may again render the gene inactive.

1.4.5 Ribosomal RNA and their genes

Two sizes of ribosomes, 70S and 80S, are known in higher plants, with the 80S ribosomes located in the cytoplasm and the 70S located in the chloroplast and mitochondria. Each of the small subunits of the ribosomes are themselves composed of two or more smaller rRNAs and are repeated and arranged in one or more tandem arrays, termed nucleolar organizer regions (NOR) (Nierras *et al.*, 1997). With the exception of some legumes, almost all plant chloroplast genomes contain two copies of a large inverted repeat which contains the 16S, 23S and 5S rRNA, genes as well as some tRNA and ribosomal protein genes, situated in the opposite orientation and separated by a large single-copy (LSC) and small single-copy (SSC) region (Lu *et al.*, 1996). In contrast, the rRNA unit in the cytosol consist out of the 18S, 5.8S and 25S rRNA coding regions with their non-coding spacers (Haberer and Fischer, 1996), while the mitochondrion rRNA are made up by the 18S, 5S and 26S coding units and non-coding spacers (Heldt, 1997). Copy numbers of rRNA genes are high in most plants with about 570 repeats per haploid genome (Pruitt and Meyerowitz, 1986).

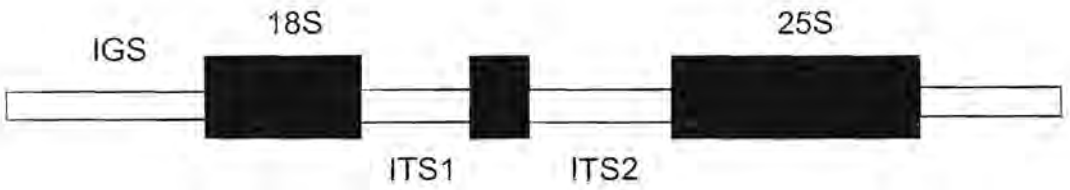


Figure 1.8. Ribosomal RNA genes in the cytosol: arrangement of the 18S-5,8S-25S RNA gene complex. IGS = intergenic spacer; ITS = internal transcribed spacer (Henry, 1997).

1.4.6 Control of retro-transposons

Because retro-transposons have the potential to dramatically alter gene function and host genome structure, it is not surprising that their transposional activities are regulated both by retro-transposon- and host-encoded factors, possibly to avoid deleterious effects on host and retro-transposon survival. The intimate relationship between retro-transposons and their plant hosts has existed for many millions of years. We are just beginning to understand how retro-transposons and their hosts'

genomes have co-evolved mechanisms to regulate transposition, insertion specificities, and mutational outcomes in order to optimise each other's survival.

Because retro-transposons cannot transpose without the presence of an RNA template available for reverse transcription, the simplest way to control their activity would be via the regulation of transcriptional initiation. Many retro-transposons show unique patterns of developmental and/or environmental regulation. A correlation between transcription and transposition of retro-transposons has been demonstrated for the tobacco *Tto*1 and rice *Tos*17 retro-transposons (Hirochika, 1993; Hirochika *et al.*, 1996). For example, transposition of *Tto*1 and *Tos*17 was associated with an increase in the levels of their RNAs, suggesting that transposition of these retro-transposons is regulated mainly at the transcriptional level. However, this is not the situation for several other retro-transposons. The *BARE*-1 of barley is highly transcribed in leaves, but its transposition has not been observed (Suoniemi *et al.*, 1996).

1.4.7 Repetitive DNA and DNA methylation

In plants and filamentous fungi, genomic methylation is restricted mostly to transposons and other repeats (Rabinowicz *et al.*, 1999; Colot and Rossignol, 1999). Most of our current knowledge concerning possible roles for methylation in eukaryotes derives from the study of organisms amenable to genome manipulation. Among these, two filamentous fungi, *Neurospora crassa* and *Ascobolus immerses*, came up with a surprise. Both fungi were found to be endowed with the ability to methylate *de novo* (and concurrently in *Neurospora*, to mutate) all gene-size duplications, and to maintain this methylation vegetatively and sexually (Colot and Rossignol, 1999).

Cytosine methylation is associated with two effects that can serve as defence mechanisms against mobile repetitive elements. First, cytosine methylation, cause a loss of RNA-polymerase-II-dependent transcription in the methylated region, either by preventing transcription initiation or by impending transcript elongation. Therefore, methylation of transposable element sequences can silence the expression of transposon-encoded genes and prevent their amplification and transposon-mediated DNA rearrangements. Secondly, cytosine methylation correlates with reduced homologous recombination between methylated regions (Eggleston *et al.*, 1995). Therefore, methylation of repetitive sequences might suppress recombination

between repeats in different genomic positions – which otherwise would lead to translocations and other chromosomal rearrangements (Bender, 1998). Striking examples of differences in methylation that correlate with differences in repeat content can be found in plants such as *Arabidiopsis thaliana* and maize (Leutwiler *et al.*, 1984; Bennetzen *et al.*, 1994).

How are repeat sequences distinguished from the rest of the genome and targeted for methylation? Evidence suggests that at least some repeated sequences are detected by a DNA-DNA pairing mechanism; unique features of the paired region mark it for methylation along the lengths of the interacting repeats (Selker *et al.*, 1987; Selker and Garrett, 1988; Vongs *et al.*, 1993). But are targeted methylation of repetitive DNA and transposons only a mechanism of silencing or is it also a mechanism to induce targeted mutations and thereby inactivate these transposons? It is known that spontaneous de-amination of 5-methylcyosine (5meC) in dividing cells causes hot spots of CG → TA mutations in *Escherichia coli,* as well as human cells (Lieb and Rehmat, 1997), hence methylation accelerates the divergence rate of newly arising duplications.

1.4.8 Retro-transposon regulation as a form of host defence

Some retro-transposons might have beneficial effects on a plant, through mutations that provide new regulatory properties to a gene or centromere function (Miller *et al.*, 1998). However, many of the expression, mutation, and insertion properties of these elements suggest that their effects are being minimized. With any highly adapted host/parasite interaction, the parasite will contribute as little as possible to the decreased host fitness. The host should also evolve such minimization/defence processes, and several seem to be acting on plant retro-transposons.

Most of the known retro-transposons appear to be defective in their ability to encode all necessary transcription functions, owing to insertions, deletions, and other mutations (Flavell *et al.*, 1992a; Flavell *et al.*, 1992b; Hu *et al.*, 1995; Jin and Bennetzen, 1989; Pearce *et al.*, 1996; Voytas *et al.*, 1992; White *et al.*, 1994). The epigenetic regulation of plant retro-transposons associated with DNA methylation and presumed hetrochromatization may also be involved in keeping retro-transposon transcription at a low level (Yoder *et al.* 1997). This DNA methylation is associated with a two- to threefold higher transition mutation rate in these elements, thus causing them to decay to a non-functioning form mare rapidly than other sequences

1.4.9 Mutations

Mutations, if they have an observable effect, are almost always harmful. Most changes however do not take place in the genes but in the great bulk of the so-called 'junk' DNA, most of which has no known function. Most mutations have very minor effects, if any, but that does not mean that they are unimportant (Crow, 1997). Spontaneous deamination of 5-methylcytosine ($5^{me}C$) causes hot spots of CG $\rightarrow$ TA mutations. Deamination of 5meC produces thymine, which is not recognised by uracil glycosylase and consequently can result in C $\rightarrow$ T mutations (Figure 1.9). It is likely that mutation hot spots at 5meC in rapidly dividing cells are attributable to insufficient time for T to G correction in the interval between deamination of 5meC and subsequent DNA replication (Lieb and Rehmat, 1997). Because cells divide rapidly during the tissue culture process as a result of hormones, such as auxins, that are added to the growth medium, and the alteration in methylation patterns observed during tissue culture, this kind of mutations can be expected to be especially higher in their cells.

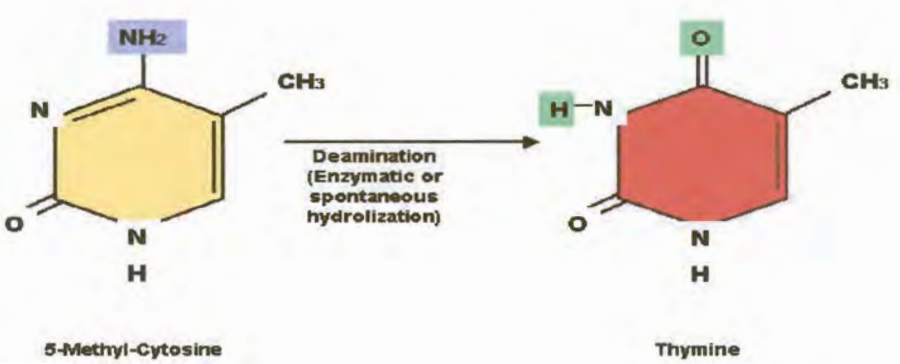


Figure 1.9. Deamination of 5-Methyl-Cytosine converts it into Thymine.

1.5 Detecting genome variation

Molecular markers and marker mapping are part of the intrusive 'new genetics' that is finding its way into all areas of modern biology, from genomics to breeding, from systematics to ecology and also into plant physiology. What are molecular markers? Molecular markers reveal neutral sites of variation at the DNA sequence level. By 'neutral' is meant that these variations do not show themselves in the phenotype, and might be nothing more than a single nucleotide difference in a gene or piece of repetitive DNA (Jones *et al.*, 1997).

1.5.1 Molecular biology and molecular markers

In April 1983, Kary Mullis chanced the course of molecular biology when he conceived the Polymerase Chain Reaction (PCR). The introduction of PCR to molecular biology has enabled the development of powerful genetic markers for the measurement of genotypic variation. By measuring genotype, rather than phenotype, genetic markers avoid complicating environmental effects and provide ideal tools for assessing genetic variation, identification and defining genetic relationships (O'Hanlon *et al.*, 2000).

PCR is an *in vitro* method whereby defined sequences of DNA are enzymatically synthesized. The reaction uses two oligonucleotide primers that hybridise to the opposite DNA strands and flank the target DNA sequence that is to be amplified. The elongation of the primers is catalysed by a heat-stable DNA polymerase. A repetitive series of cycles, involving template denaturation, primer annealing, and extension of the annealed primers by the polymerase result in exponential accumulation of a specific DNA fragment.

The PCR technique is so pervasive in molecular biology that it is difficult to think of life without it. Because of PCR, 'insufficient nucleic acid' is no longer a limitation in molecular biology. More importantly innovative research is continuously updating the definition of 'PCR applications' thereby increasing the usefulness and scope of the technique.

1.5.2 Non-PCR fingerprinting techniques

1.5.2.1 *Restriction fragment length polymorphysim (RFLP)*

If two DNA molecules are essentially the same, but nevertheless have one or more small differences in their nucleotide sequence, then the fact that they are not identical may become apparent by means of RFLPs (Buscot *et al.*, 1996). The difference between two genomes in the size of the restriction fragments at a defined genetic locus is thus termed a restriction fragment length polymorphism. RFLP analysis is predominantly used to compare closely related species by the comparison of slower evolving regions of their genomes and to assess the diversity within populations (Bruns *et al.*, 1991).

Restriction patterns are generated by the cleavage of DNA with restriction enzymes and separating the bands using electrophoresis. One of two approaches to be followed is the PCR-RFLP, which involves the restriction of PCR amplicons of a selected portion of the genome usually with four- or six-base restriction enzymes. A second approach would be the restriction of the entire genome with restriction enzymes followed by Southern-blotting using selected probes (Maclean *et al.*, 1993).

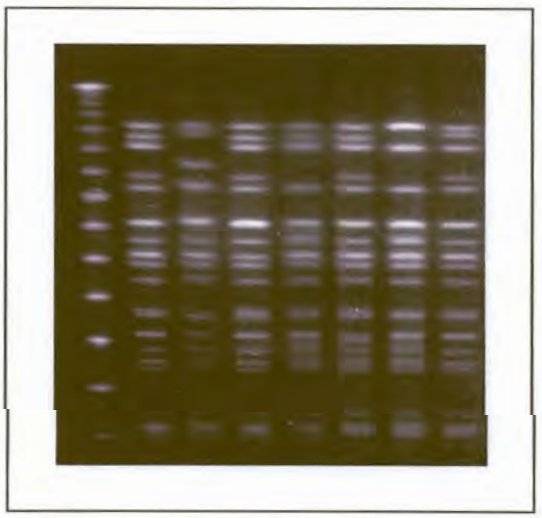


Figure1.10: A typical PCR-RFLP gel profile

There are disadvantages with RFLP analysis, as often a number of probes and enzyme combinations have to be tested in order to generate significant numbers of RFLPs. This method generally uses radioactive probes, and requires a large amount of target DNA. In fact, RFLPs detect about 1 in 10000 polymorphic nucleotides in the human genome (Soller and Beckmann, 1986). The problem with searching for direct associations between RFLP and traits of economic value is the low likelihood of finding one; at best 1 : 200, but probably 1 : 20000 (Soller and Beckmann, 1986). RFLP markers are also difficult to transfer between different linkage maps, as what is polymorphic in one population may well not be in another. This, however, is true for any randomly selected type of marker, but RFLPs do generate a reasonable number of markers, which can be easily placed onto a map.

1.5.3 PCR-based fingerprinting techniques

The genomes of closely related plants or varieties might be identical except for differences in a few coding genes or in minor genome re-organizations. A range of different approaches is used to try and detect these genetic differences. Molecular techniques based on the polymerase chain reaction (PCR) have been used very successfully as a tool in genetic mapping, molecular taxonomy, and evolutionary studies. Among the techniques that are being used in the differentiation of plants are the analyses of r-DNA intergenic regions (Scribner and Pearce 2000), simple sequence repeats (SSRs), which are also known as microsatellites, restriction fragment length polymorphic DNA (RFLP), random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphic DNA (AFLP). The two most widely used molecular techniques to detect plant variation are Random Amplified Polymorphic DNA (RAPD) analysis, which detects DNA polymorphisms amplified by arbitrary primers (Williams *et al.*, 1990) and Amplified Fragment Length Polymorphisms (AFLPs) (O'Hanlon *et al.*, 2000).

1.5.3.1 *Random amplified polymorphic DNA (RAPD)*

This technique was developed by Williams *et al.* (1990) and Welsh and McClelland (1990) to utilize the decreased specificity in binding at low annealing temperatures for short oligo-nucleotides (usually 10 nucleotides). Large numbers of fragments are amplified by this method, some of which are polymorphic. RAPD analysis can be a

very useful tool for characterizing genetic variability among different cultivars and varieties as also shown in this study. It is very simple and sensitive and provides a PCR fingerprint for related organisms based on the genome rather than individual genes (Foster *et al.*, 1993). This technique scans the DNA for short inverted repeat sequences and amplify inverting DNA segments (Hadrys *et al.*, 1992). A PCR is carried out by the use of a single primer or primer set, usually 9-10 nucleotides in length (Foster *et al.*, 1993). These primers find homology on the template DNA and generate random amplified polymorphic DNA by initiation and extension. The different band sizes can be analysed by electrophoresis to generate specific banding patterns.

The RAPD technique offers several advantages. It can produce more polymorphisms than the RFLP technique. It is simple to use as well as relatively fast, and does not require radioisotopes. A large number of bands can be produced for a single primer and a range of primers are commercially available. The major disadvantage of this technique is the inconsistency of reproducibility. Furthermore it only detects dominant markers (Williams *et al.*, 1990).

RAPD markers have been used in many species for a variety of investigations: gene cloning, medical diagnostics and trait intro-gression in breeding programs (Williams *et al.*, 1990). Levin *et al.* (1993) used RAPD markers to generate new markers on the Z chromosome of the chicken in order to identify sex-linked traits. RAPD markers have also been useful in determining phylogenetic relationships between species as demonstrated by Barral *et al.* (1993) with the *Shistosoma* genome. Recently, the RAPD technique has been applied to identify date palm varieties (Corniquel and Mercier 1994; Sedra *et al.*, 1998) which has also been a subject of this study.

Because of the unreliable nature of RAPDs certain modifications and improvements on the technique can be made to produce a more reliable differentiation system. Paran and Michelmore (1993) developed sequence-characterized amplified regions (SCARs), also done in this study. SCARS are derived from RAPD markers by developing longer primers. After the RAPD fragment is cloned and sequenced, a pair of primers are designed and synthesized. These SCAR primers are then used to amplify the specific regions of DNA. SCAR markers are advantageous over RAPD markers because they usually detect only a single locus, their PCR amplification is less sensitive to reaction conditions, and they are more likely to be co-dominant markers. SCARs have been used for mapping genes of interest, map-based cloning,

and marker-assisted selection (MAS) on several plants including lettuce (Paran and Michelmore, 1993), common bean (Gu *et al.*, 1995), oak (Bodennes *et al.*, 1997), and citrus (Deng *et al.*, 1997).

RAPDs are widely used as an easier alternative for RFLPs, because less information about the genome is required, it is less expensive and is much faster than RFLPs (Foster *et al.*, 1993). This technique provides the advantages of simplicity of minute amounts of sample DNA and a universal set of primers can be used for all living organisms, constructed without any knowledge of the organism's genome. Some disadvantages might be that this technique requires a lot of standardization, is unable to distinguish between homo- and heterozygotes and differences in band intensity. It is also difficult to obtain the same results repeatedly, which makes it sometimes difficult to draw solid conclusions. RAPD markers are the least informative of all fingerprinting techniques, but they are detected much more easily than RFLPs (Foster *et al.*, 1993).

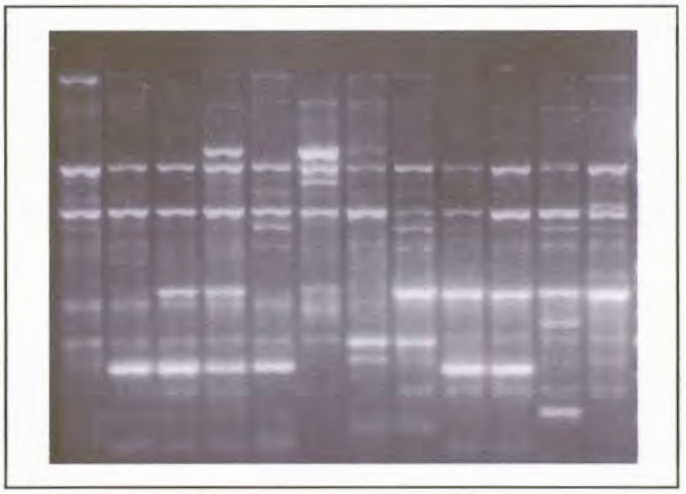


Figure 1.11: A typical RAPD gel profile.

1.5.3.2 *Amplified fragment length polymorphism (AFLP)*

AFLP is based on selective amplification of digested genomic DNA by a series of extended primers and is used to visualize hundreds of amplified DNA restriction fragments simultaneously. AFLP technology combines the power of RFLPs with the flexibility of PCR-based technology by ligating primer-recognition sequences (adaptors) to restricted DNA (Vos *et al.*, 1995). The first step involves restriction

digestion of the genomic DNA with two specific enzymes, one a rare cutter (*Mse*I) and the other a frequent cutter (*Eco*RI). Adaptors are then added to the ends of the fragments to provide a known sequence for PCR amplification. Adaptors are very important in this technique, because the restriction site sequence at the end of the fragments is insufficient for primer design (Karp *et al.*, 1997).

If these restriction fragments should be amplified, not all the fragments would be resolvable on a single gel (Karp *et al.*, 1997). Primers are thus designed to incorporate the known adaptor sequence with one to three additional base pairs. The additional base pairs are referred to as selective nucleotides. Because of the added base pair/s, PCR amplification can only occur where the primers are able to anneal to fragments that have the adaptor sequence plus the complimentary base pairs to the selective nucleotides (Karp *et al.*, 1997). This kind of amplification results in 50 - 100 fragments, which can easily be separated using poly-acrylamide gel electrophoresis. More than three additional nucleotides will result in the non-specific amplification of fragments (Vos *et al.*, 1995). Several polymorphisms are detected in a single assay.

Radio-labelled primers can be used to visualize the amplified products with exposure to X-ray film, but the cost and danger involved make non-radiolabelled and silver staining techniques preferable (Karp *et al.*, 1997). Some advantages of AFLPs are that only small amounts of DNA are needed. Unlike RAPDs that use multiple, arbitrary primers and lead to unreliable, non-reproducible results, the AFLP technique uses only two primers and gives reproducible results. Many restriction fragment subsets can be amplified by changing the nucleotide extensions on the adaptor sequences and hundreds of markers can be generated reliably. High resolution is obtained because of the stringent PCR conditions. No prior knowledge of the genomic sequence is required. The AFLP technique also works on a variety of genomic DNA samples making it very flexible (Karp *et al.*, 1997).

Figure 1.12. A typical AFLP gel profile (Russel et al., 1997)

All the evidence so far indicates that AFLPs are as reproducible as RFLPs. They need more DNA and are technically more demanding than RAPDs. Because of the speed and efficiency of the technique, compared to RFLP and RAPD, it is now being used more widely for comparative purposes.

1.5.3.3 *Representational difference analysis (RDA)*

The techniques described above rely on patterns consisting of the presence or absence of DNA fragments rather than DNA sequence variation. Understanding DNA sequence variations should allow us to understand the genetic basis of evolution, the genetic control of development, as well as the physiological abnormalities and variation. It is well known that genetic variation can result from many phenomena such as genomic rearrangements, gene duplication, viral insertions, deletions, or simple base pair changes. Detecting these sorts of variations, however, has been very difficult due to the complexity of the genomes being analysed and the small size of the differences.

In the past, subtractive hybridisation has been used with some success to identify large differences between two genomes, such as insertions or deletions. Subtractive hybridisation involves hybridising DNA containing the sequence of interest with large amounts of the DNA lacking these sequences. Hybridisation is usually followed by the physical separation of the undesired sequences from the target sequences, using methods such as chromatography. However, the larger and more complex the genomes or the smaller the difference between them, the more difficult this separation becomes. Successful subtractive hybridisation usually results in only a 10- to 100-fold enrichment of the target sequences and therefore must be followed by the

time-consuming process of sorting through large numbers of DNA sequences to find the one of interest. A new technique has recently been described that eliminates the classical problems of subtractive hybridisation and will allow direct isolation of small differences between two complex genomes without having to sort through excess products. Lisitsyn *et al.* (1993) have developed a technique called Representational Difference Analysis (RDA). RDA couples subtractive hybridisation with polymerase chain reaction (PCR) to amplify exponentially only the target DNA sequences after repeated rounds of subtractive hybridisation.

RDA is a two-step technique. In the first step a representation of the genome is created by digesting the genome with an appropriate restriction enzyme and ligating adaptors to the restriction fragments, which is then amplified by a PCR step to form amplicons. (Figure 1.13). In the second step new adapters are ligated to the amplicons of the representation of the tester genome and the amplicons of the tester and driver are subtracted from each other. This results in the amplification of unique DNA sequences found in the tester (Figure 1.14)

RDA allows one to target regions of the genome that differs in more than single nucleotide differences between two genomes. These differences may represent unique sequences, genome rearrangements differences in copy number or sequences with a high mutation rate. This allows one to find regions of genetic variation in organisms were genome variation might otherwise be low or hard to find.

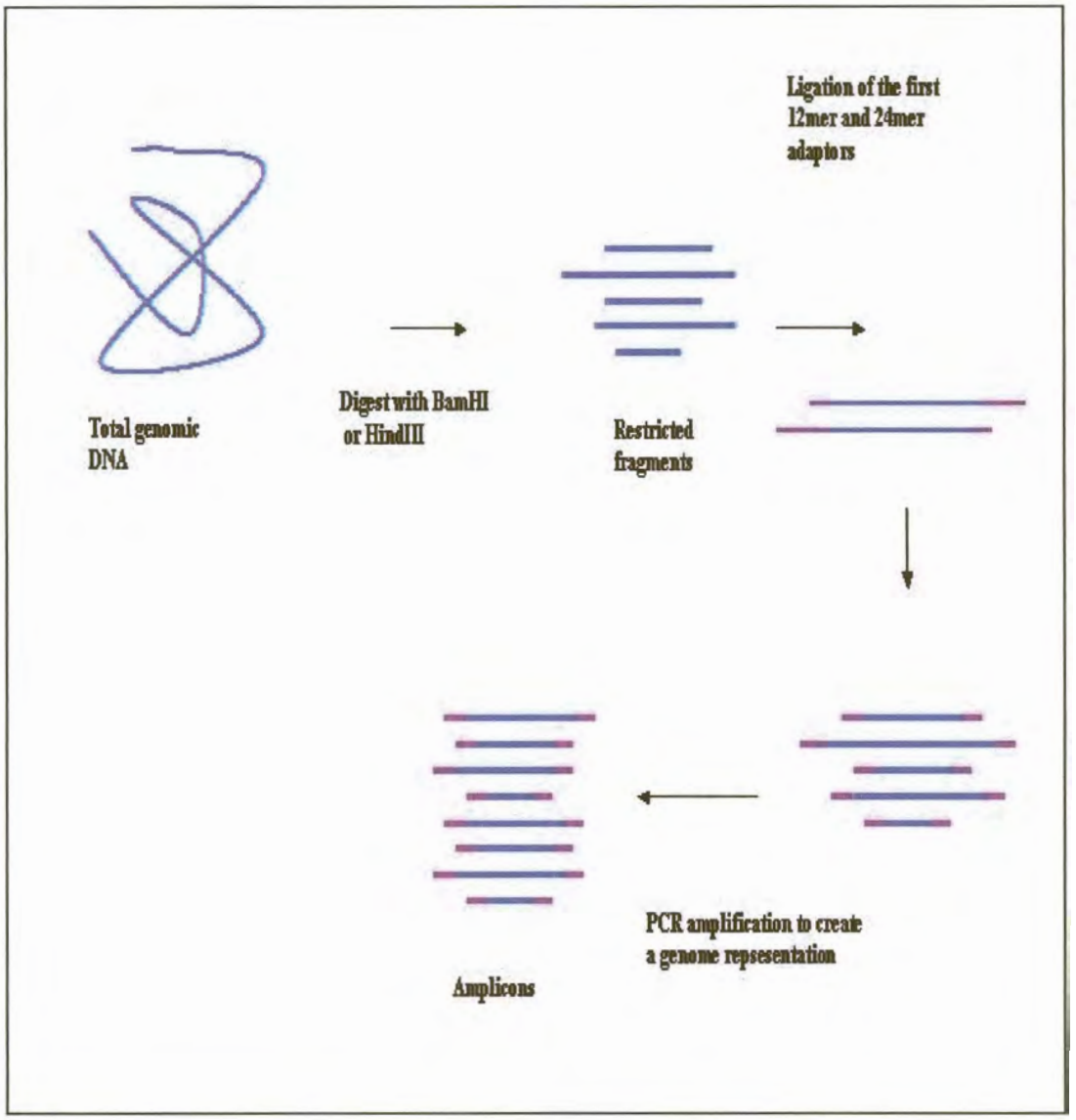


Figure 1.13. Generation of RDA amplicons from total genomic DNA

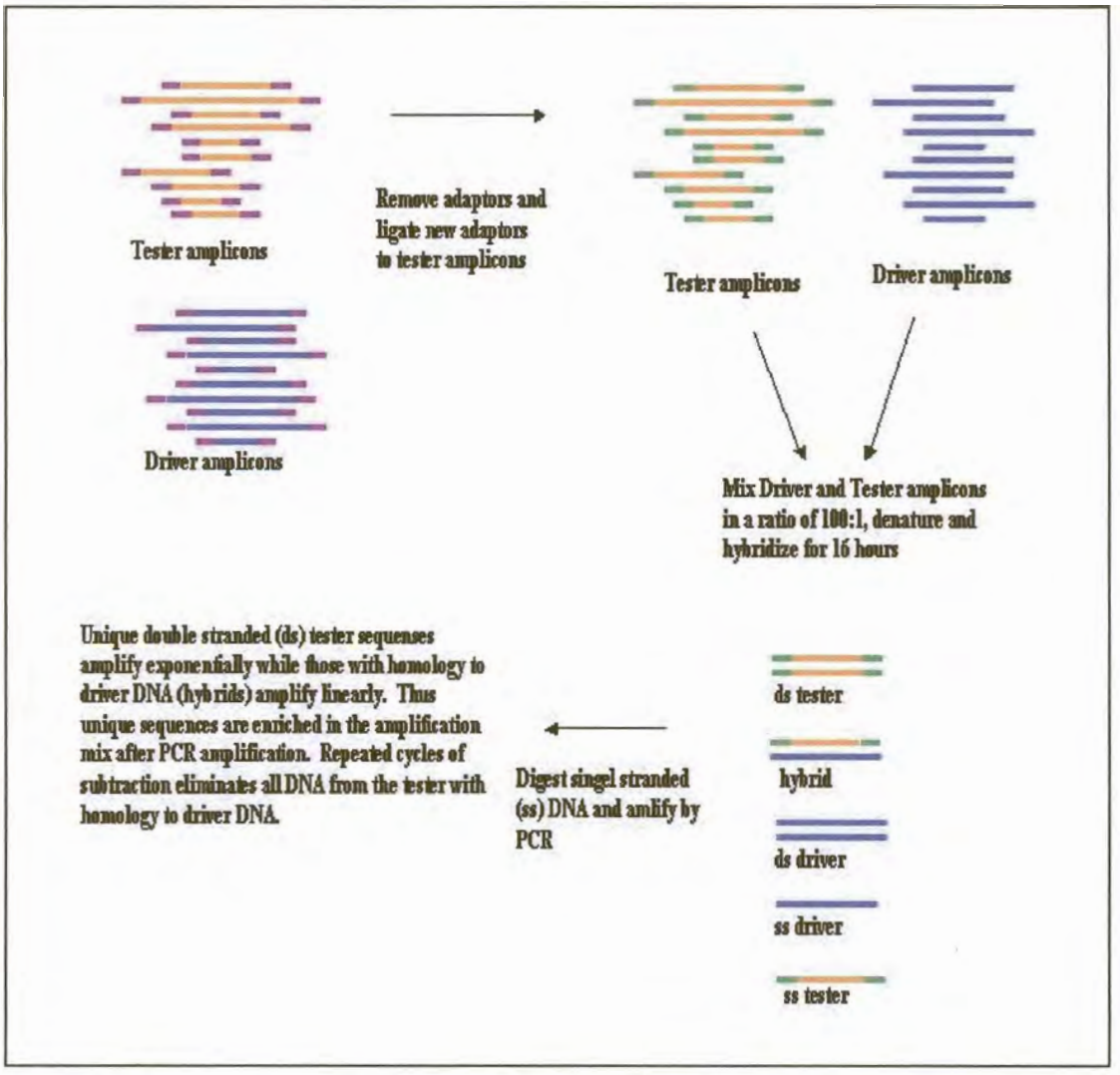


Figure 1.14. Subtractive hybridisation of tester and driver amplicons and kinetic enrichment of differences.

1.5.4 First trials of RDA

Several uses and successful trials of RDA as well as the technique were described by Lisitsyn *et al.* (1993). The originality of this approach relies on the fact that subtraction is applied to a fraction amplified by PCR (a representation) rather than the entire genome, thus making genomic subtraction applicable to complex genomes. The use of different restriction enzymes can provide several representations of the genome, if necessary. To show that RDA could detect viral DNA in the human genome, the authors used human genomic DNA, mixed with lambda phage DNA or adenovirus, as the tester. They used the same human

genomic DNA without the viral DNA as driver. These two populations, therefore, were identical except for a few additional phage or viral sequences in the tester. They found that the small restriction fragments of the adenovirus and the lambda phage were the only difference products after 3 rounds of subtractive hybridization and PCR amplification. The large restriction fragments present in the added viral DNA were not isolated by RDA, consistent with the hypothesis that only small restriction fragments would amplify during PCR.

RDA was next shown to identify RFLPs between two closely related individuals. DNA from two sisters of an Amish family with an established pedigree was analyzed using RDA. A complex, yet clear, pattern of difference products was seen on an agarose electrophoresis gel after 3 rounds of subtractive-hybridization and amplification. Five distinct difference products were isolated and all were used in Southern blot analysis to identify unique RFLPs between the sisters. The inheritance of these RFLPs was analyzed in other relatives and each identified a distinct genetic locus with two different alleles that were inherited according to Mendelian genetics. This application of RDA could be useful in identifying loci linked to inherited disorders by analyzing individuals in families segregating the disorder.

Lisitsyn *et al.* (1993) proposed several future applications of RDA (1) to detect genetic abnormalities that result in cancer, (2) to generate RFLP markers that exist between related species or individuals to be used in genetic mapping, and (3) to identify loci linked to genetic diseases that result from spontaneous mutations or rearrangements in the fertilized egg. In addition, RDA was proposed as a method of detecting RFLPs linked to a mutant gene of unknown location in any organisms that can be bred. This last possibility could greatly benefit the field of molecular biology by speeding up the process of mapping and cloning genes.

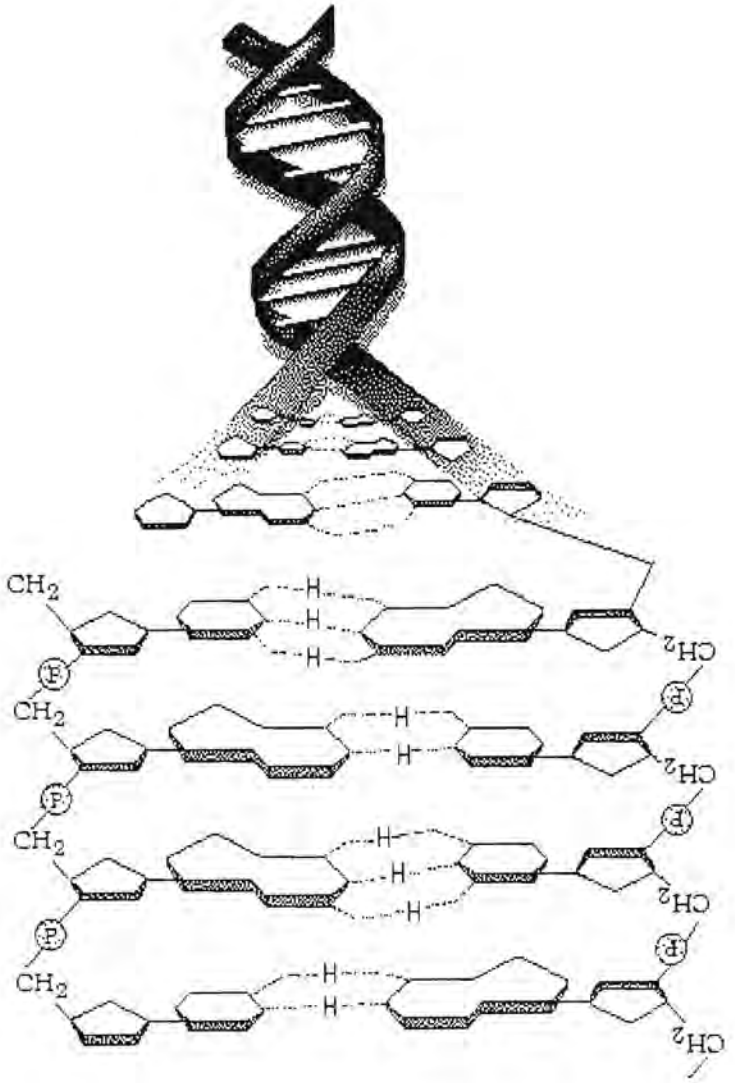
Recently, RDA was used to isolate sequences that are unique to the tumorogenic tissue of patients with AIDS-associated Kaposi's sarcoma (Chang *et al.*, 1994). RDA has also been used to isolate families of repetitive sequences present in only one of compared genomes (Navin, 1996). Further, Nekrutenko *et al.* (2000) used RDA to create a species-specific marker for voles and Toder *et al.* (2001) have applied RDA in evolutionary genomics to search for overall genome differences between humans and the great apes. RDA has also been used to determine differences between two distantly related oak species where similarities of isolated RDA fragments with known retro-transposons were found (Zoldos *et al.*, 2001). In addition, Donnison *et al.* (1996) applied RDA to identify male-specific restriction fragments in the dioecious

plant *Silene latifolia*. RDA has also been used to identify polymorphisms in banana lines that are a result of genomic rearrangements during *in vitro* propagation resulting in markers useful for the detection of early variation in the initiation of tissue culture plants (Cullis and Kunert, 2000). One of the specific advantages of RDA is that subtractions between pooled DNA samples can be performed in order to identify specific polymorphisms only present in a particular individual rather than relying on identification based on a particular pattern of polymorphic bands.

In the following chapters results are presented regarding the outcome of experiments using the RDA technique to subtract genomes from date palm varieties as well as the outcome of methylation sensitive RDA subtractions done on date palm and pine clones.

Chapter 2

Differentiation of date palm varieties with Random Amplified Polymorphic DNAs (RAPDs)



2.1 Abstract

Identification of plant cultivars where no early morphological differences are visible is crucial for many plant producers. Random Amplified Polymorphic DNA or RAPD is among the commonly used techniques for plant differentiation. To test the applicability of this technique for date palm, genomic DNA from tissue culture-derived date palm plants of the cultivars 'Barhee' and 'Medjool' were amplified for RAPD analysis using the commercially available DNA primers OPE-01 and OPE-06. Both primers could differentiate the two cultivars. In an attempt to produce a more robust amplification, sequence information of a 700 bp fragment amplified from 'Medjool' was used to create a SCAR primer pair. However, this SCAR primer pair amplified an dentical DNA fragment from both tested genomes.

2.2 Objective

The first objective of this part of the study was to confirm a reported differentiation system for date palm using the RAPD technique. The second objective was to develop a SCAR primer pair from one of the amplification products to obtain a more robust test system. In particular (1) two reported RAPD primers, OPE-01 and OPE-06, were applied in this RAPD study (Corniquel and Mercier 1994) to differentiate the two date palm cultivars 'Medjool' and 'Barhee' and (2) an attempt was made to develop a SCAR primer pair from one of the sequenced amplification products.

2.3 Results

2.3.1 DNA extraction

Two methods were used for genomic DNA extraction from date palm. The CTAB (cetyltrimethylammonium bromide) method resulted in an amount of about 320 µg genomic DNA isolated per gram fresh plant leaf material. In comparison, 5 µg genomic DNA per gram of fresh material were obtained using the Nucleon Phytopure Plant DNA extraction kit (Figures 2.1 and 2.2). DNA isolated according to the CTAB method was, however, of a much lower quality when compared to the Nucleon

Phytopure system when used for DNA amplification. After repeated attempts to optimize the PCR reaction, amplification of genomic DNA isolated with the CTAB method resulted in either smears on the agarose gel or complete failure of amplification by PCR. The shift in 'Barhee' DNA band in figure 2.2 lane 5 may be due to the larger genome size as a result of the greater number of repetitive elements found as shown in this study.

1 2 3 4 5 6

Figure 2.1. DNA extraction from date palm using the CTAB method. Lanes 1-4: 25 μg (1), 50 μg (2), 100 μg (3) and 250 μg λ-phage DNA (4) as standards to determine the amount of genomic DNA isolated; lane 5: 2 μl 'Barhee' DNA; lane 6: 2 μl 'Medjool' DNA

1 2 3 4 5 6

Figure 2.2. DNA extraction from date palm using the Nucleon Phytopure Plant DNA extraction kit. Lanes 1-4: 25 μg (1), 50 μg (2), 100 μg (3) and 250 μg λ-phage DNA (4) as standards to determine the amount of genomic DNA isolated; lane 5: 2 μl 'Barhee' DNA; lane 6: 2 μl 'Medjool' DNA

2.3.2 RAPD analysis

Application of primers OPE-01 and OPE-06 differentiated the two cultivars. OPE-01 amplified a distinct DNA fragment from 'Barhee' genomic DNA of about 1700 bp but not from 'Medjool' genomic DNA (Figure 2.3). Primer OPE-06 amplified a fragment of about 700 bp from 'Medjool' but not from 'Barhee' genomic DNA. A control test with all PCR reagents except any template DNA resulted in no amplification of any product (Figure 3.3). In addition to the two primers OPE-01 and OPE-06, the primer OPB-07 was also used randomly. However, this primer did not distinguish between the two cultivars under the conditions used in this study.

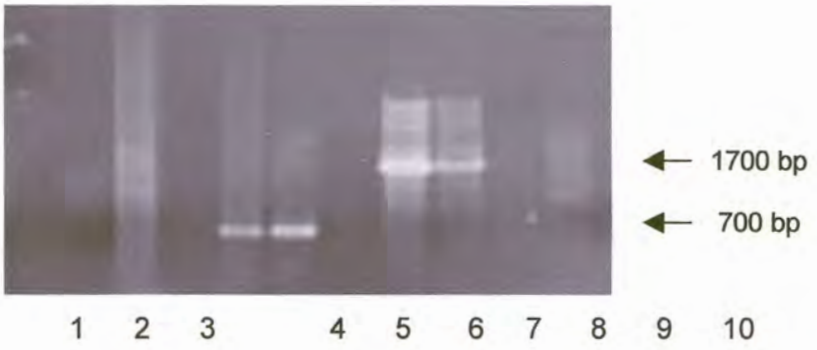


Figure 2.3. RAPD analysis of the date palm cultivars 'Medjool' and 'Barhee'. Lane 1 represents a control without genomic DNA but with primer OPE-06; lanes 2 and 3: amplification of 'Barhee' DNA with OPE-06; lanes 4 and 5 amplification of 'Medjool' DNA with OPE-01 producing a 700 bp fragment; lane 6: control without template DNA but with primer OPE-01; lanes 7 and 8: amplification of 'Barhee' DNA with primer OPE-01 producing a 1500 bp fragment; lanes 9 and 10: amplification of 'Medjool' DNA with primer OPE-01.

2.3.3 Characterization of amplification product

The 700 bp DNA fragment amplified from 'Medjool' genomic DNA was cloned into the vector p*MosBlue* and then sequenced. Sequencing of the cloned fragment was done in both directions in order to obtain a complete sequence of the DNA fragment. Analysis of the fragment including the OPE-06 primer (Figure 1.4) revealed, however,

no significant homology to any known plant sequences when a Blast search was performed.

```
  1  AAGACCCCTCCATGCTGGATTTTATTCTGATTCTACNTCAGTCCCCTCCA
 51  CTTGAATACAACTCGTCCTCGAGTTCTCATCTGGATATAAATGTTCCTCA
101  GGCTATGGTATTCATAGATATCAGCGACATTAAAAGTCTTTGAAATTCTC
151  ATATCATCTTTGTAATAACCCCAAAATATTTTTTTTATATAAAAAGGGAT
201  AGAATAGTCCTTTTCAATTAATATAAGGAGTAAAATAGGAAAGAATCAAA
251  AGATAATGGCATAGTTGTAGATGCATTGAAGTGGTAAGGGTAAAATTGGA
301  AGGAATCAAAAGTTAATGGCATAACGGTAGATATGTTGAAGTGTTAGGGG
351  CAAAATGGAAATTTAATAATATTAAACAAAGGGTGAATAGTGTTTTGATG
401  TGATTTAATTAGGGGGTTTGCTGTGGCTTTTGGATGGAAACCCGAGAGGG
451  AGTCTCAGACGGACAAAGAGGAAGAAAAAGAAGGAGAGGAAAGGAAAGGA
501  AGAAAGGAAGAGGAAAGGAAAGGAAGAAAGGAAGAGGAAGGAGATCCCGG
551  TTGCATTTCTAGCTGGAGGGAAAAACGTAGATCTCCTTCTCCTCTACACA
601  AGAACCTCGATTTCCAAAAAGAAAAAGGTAAATATCCATCCCTATCTAGT
651  CTTTTGATGACATTANGCTATACNAANGGTGGATATAGTCTANAAGGGTC
```

Figure 2.4. Sequence of the 700 bp fragment amplified with primer OPE-06 from 'Medjool' genomic DNA. OPE-06 primer site (underlined) and sites used to design the SCAR primers DpSL and DpSR are underlined.

2.3.4 SCAR primer design and testing

The SCAR primers DpSL and DpSR were constructed from the sequence obtained from the polymorphic band amplified from 'Medjool' with OPE-06 (Figure 2.4). The two 20-mer primers (Table 2.1) were designed using a standard design program (Expasy, Switzerland). The primer pair amplifying a 128 bp DNA fragment was used in a PCR reaction with 'Barhee' and 'Medjool' DNA as template at various annealing temperatures to optimize the PCR reaction. Fragments of the same size (about 128 bp) were, however, amplified from both types of genomic DNA (Figure 2.5) and therefore SCAR primers failed to differentiate between the two date palm cultivars. A further 400bp fragment was also amplified from genomic DNA of one of the 'Medjool' plants, but could not be amplified in other genomic 'Medjool' DNAs tested.

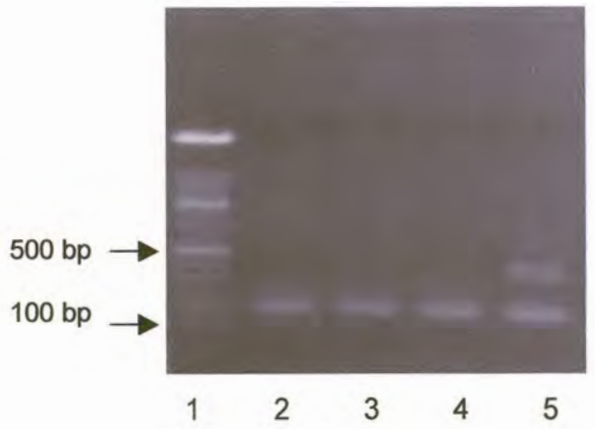


Figure 2.5. SCAR analysis of 'Barhee' and 'Medjool' genomic DNA. Lane 1 represents a 100 bp molecular marker ladder (Roche, Switzerland); lanes 2 and 3: amplification of 'Barhee' genomic DNA with SCAR primers; lanes 4 and 5: amplification of 'Medjool' genomic DNA with SCAR primers.

Table 2.1. Primer sequences used for RAPD and SCAR analysis.

Primer	Sequence
OPE-01	5'-CCCAAGGTCC-3'
OPE-06	5'-AAGACCCCTC-3'
DpSL	5'- GTGTTAGGGGCAAAATGGAA–3'
DpSR	5'- TTGTCCGTCTGAGACTCCCT–3'

2.4 Materials and methods

2.4.1 Plant material and DNA extraction

Tissue culture plants of 'Barhee' and 'Medjool' were used for genomic DNA extraction. The *in vitro* plants used were 'Medjool' derived from explant material collected in California, and 'Barhee' derived from explant material collected in the United Arabic Emirates. Total cellular DNA was isolated in a first method from the entire plant (1 g) using the Nucleon Phytopure Plant DNA extraction kit (Amersham Life Science, UK) according to the manufacturer's instructions.

In a second method for genomic DNA extraction, the protocol of Ait-Chitt *et al.* (1993) was used. This method is based on CTAB precipitation of the DNA. For DNA isolation leave tissue (1 g) from tissue culture plants was used. Leaf tissue was ground to a fine powder with a mortar and pestle in liquid nitrogen. Grinding was assisted with acid-washed sand. The powder was then homogenized in 7ml of DNA extraction buffer (1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl pH 8.0, 3% w/v CTAB, 1% v/v 2-mercaptoetanol) and incubated at 65°C for 30 min. This mixture was then extracted with an equal volume of chloroform-isoamyl alcohol (24 : 1), and the DNA in the aqueous phase precipitated with an equal volume of isopropanol. The DNA was collected by centrifugation, washed with 70% (v/v) ethanol and dissolved in TE buffer. The DNA was then treated with RNase.

2.4.2 RAPD analysis

For RAPD analysis, oligo-nucleotide primers OPE-01 and OPE-06 from Operon Technologies Inc. (Alameda, CA) were used. For detailed PCR reaction conditions see Annex.

2.4.3 DNA isolation from agarose gels and cloning

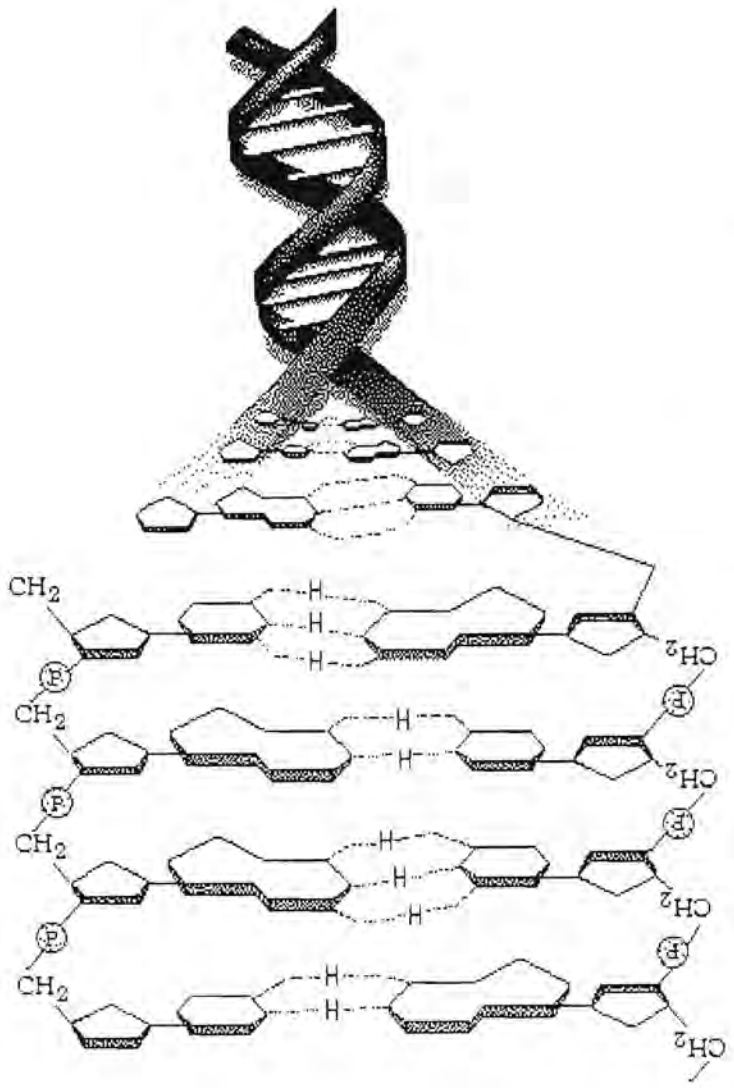
Amplified band (700 bp) from genomic 'Medjool' DNA was cut out of the gel with a scalpel and purified with a Sephaglas™ BrandPrep kit (Pharmacia Biotech Inc, USA) using the recommended protocol of the supplier for DNA purification from agarose gels. The isolated fragments were cloned into the vector pMos*Blue*. For detailed cloning procedure see Annex.

2.4.4 Sequence analysis

Sequencing reactions were carried out as described under Annex.

Chapter 3

Isolation of RDA difference products from date palm



3.1 Abstract

Several research groups have applied RDA (Representational Difference Analysis) especially in medical sciences and the technique has been predominantly used on the cDNA level to study differences in gene expression. Application of RDA on the genomic level and in plant research is still very limited, which is very likely due to its complexity in execution on large plant genomes. RDA was used in this part of the study to demonstrate its applicability in genome differentiation and to isolate differences in the genome sequence of two closely related date palm cultivars. Several RDA subtraction products with an approximately same size were isolated from 'Barhee' genomic DNA after *Bam*HI digestion of genomic date palm DNA. RDA subtraction products were cloned into an appropriate cloning vector allowing sequence analysis. Subtraction products could not be detected after digestion of genomic date palm DNA with *Hind*III followed by the subtraction of genomes from the two investigated date palm genomes.

3.2 Objective

The objective of this chapter of the study was to evaluate the applicability of the RDA for date palm genome analysis to isolate possible genomic sequence differences between two closely related date palm cultivars. In particular, the genomes of the two date palm cultivars 'Barhee' and 'Medjool' were subtracted from each other after digestion of genomic DNA of both cultivars with either the restriction enzyme *Bam*HI or *Hind*III.

3.3 Results

3.3.1 Isolation and digestion of genomic DNA

Genomic DNA of the two date palm cultivars 'Barhee' and 'Medjool' digested with either the the restriction enzyme *Bam*HI or *Hind*III revealed no obvious differences in the DNA patterns of the two cultivars after agarose gel electrophoresis and staining of DNA with ethidium bromide (Figure 3.1).

1 2 3 4

Figure 3.1. Restriction enzyme digestion of total genomic DNA from two date palm cultivars separated on a 1% agarose gel and stained with ethidium bromide. Lane 1 represents 'Barhee' genomic DNA cut with *Bam*HI cut; lane 2 'Barhee' DNA cut with *Hind*III, lane 3 'Medjool' DNA cut with *Bam*HI cut and lane 4 'Medjool' DNA cut with *Hind*III.

3.3.2 Amplification and subtraction of genomic DNA

After genomic DNA digestion, adaptor sequences were ligated to digested genomic DNA to allow amplification of digested DNA by the PCR reaction. By executing a PCR reaction using adaptor sequences as PCR primers, DNA representations (amplicons) of each date palm genome were produced using a PCR reaction. Figure 3.2 shows the amplicons produced from *Bam*HI digested genomic DNA after agarose gel electrophoresis.

After amplicon production, RDA subtractions were performed between 'Barhee' and 'Medjool' amplicons using either 'Barhee' DNA as tester and 'Medjool' as driver or with 'Medjool' DNA as tester and 'Barhee' DNA as driver. Following a single round of subtraction using a tester to driver ratio of 1 to 100 only one of the two subtractions produced subtraction products of approximately 150 bp (Figure 3.3). This was when 'Barhee' amplicon DNA was used as tester and 'Medjool' DNA as driver but not when 'Medjool' amplicon DNA was used as tester and 'Barhee' DNA as driver.

Figure 3.2. First round amplicons from *Bam*HI digested genomic DNA of two date palm cultivars after separation on a 1% agarose gel and staining with ethidium bromide. Lane 1 represents 'Barhee' amplicon and lane 2 'Medjool' amplicon.

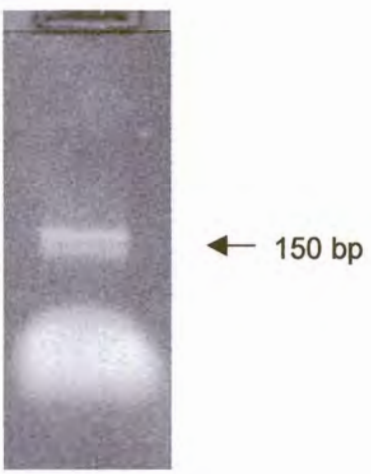


Figure 3.3. RDA subtraction product separated on a 1% agarose gel and staining with ethidium bromide after subtracting 'Barhee' amplicon DNA from 'Medjool' amplicon DNA.

3.3.3 Cloning and hybridization of subtraction products

After extraction of the region DNA from the agarose gel, the purified total extracted DNA was cloned into the unique *Bam*HI site of vector p*BlueScript*, which was used to

transform *Escherichia coli* cells of the strain JM109. After selection on an ampicillin-containing medium, fifty *Escherichia coli* (*E. coli*) colonies containing the cloned difference product were hybridized with randomly labeled 'Barhee' amplicon DNA (Figure 3.4) or 'Medjool' amplicon DNA (data not shown). Both sets of labeled amplicons hybridized to all colonies with different intensity indicating that the isolated subtraction products were not unique to 'Barhee' genomic DNA.

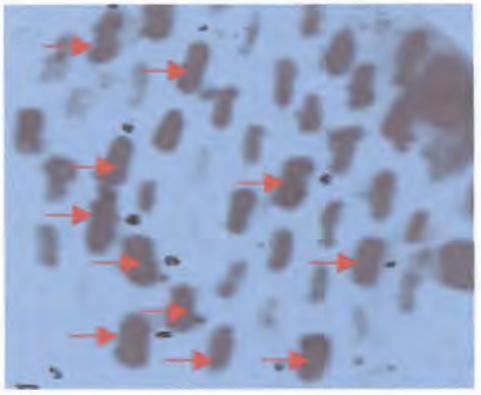


Figure 3.4. Hybridisation of *E.coli* colonies carrying plasmid DNA with subtraction product inserts to randomly labelled 'Barhee' amplicon DNA. Red arrows indicate the *E. coli* colonies that were selected.

To select for the cloned inserts with the strongest homology to 'Barhee' genomic DNA, ten *E. coli* colonies with the strongest hybridisation signal after probing with labelled 'Barhee' amplicon DNA were selected. The plasmid DNA was then isolated from colony and the cloned subtraction product inserts released after *Bam*HI digestion. Figure 3.5 shows that two different sizes of subtraction product inserts had been cloned into the plasmid.

Figure 3.5. Subtraction product inserts after separation on a 1% agarose gel and staining with ethidium bromide after release from plasmid p*BlueScript* digested with *BamH*I. (*) indicates the two inserts Dp41 and Dp50 with a different size.

3.4 Materials and methods

3.4.1 Plant material and DNA extraction

'Medjool' and 'Barhee' plant material originated from 14 plants grown in California that have been propagated solely via off-shoots from mother plant material imported in the beginning of the last century from Northern Africa (Nixon 1950). The *in vitro* plants used were 'Medjool' derived from explant material collected in California, and 'Barhee' derived from explant material collected in the United Arabic Emirates. Total cellular DNA was isolated from leaves (1 g) using the Nucleon Phytopure Plant DNA extraction kit (Amersham Life Science, UK) according to the manufacturer's instructions.

3.4.2 Representational Difference Analysis

3.4.2.1 Preparation of RDA amplicons

RDA was performed following the general outline described by Lisitsyn et al. (1993) as described in the annex. Total genomic DNA was digested with the restriction enzymes *Bam*HI and *Hind*III. Amplicons were prepared by ligating either the adaptor pair Rbam 12 & Rbam 24 or Rhind 12 & Rhind 24 (see appendix for sequences) to the digested DNA. After ligation the DNA was amplified in eight 100µl volumes using a Perkin-Elmer GeneAmp 9600 thermocycler as described in the annex.

3.4.2.2 Removal of the adaptors from amplicons

Of the amplicons that was to be used as driver 150 μg and 10 μg of the tester amplicons were digested with the appropriate enzyme. The driver and tester DNAs were both redissolved at approximately 400 μg/ml. Both the digested driver and tester amplicons were run on a 1.5% TAE agarose gel alongside an equal aliquot of undigested amplicons to check the completeness of the digestion. On the same gel, standard lambda phage DNA was run so that the final concentrations of driver and tester DNAs could be estimated and adjusted if necessary.

3.4.2.3 Change of adaptors on tester amplicons

Tester DNA was prepared by adding a second adaptor pair JBam 12 and 24 (sequences in annex) for *Bam*HI digested DNA or JHind 12 and 24 (see annex for sequences) for *Hind*III digested DNA to 1 μg of the first round amplicons. An aliquot of the ligate was amplified for 20 cycles in a reaction volume of 20 μl to check that the newly ligated adaptors would support amplification with the new primer.

3.4.2.4 Subtractive hybridisation and kinetic enrichment

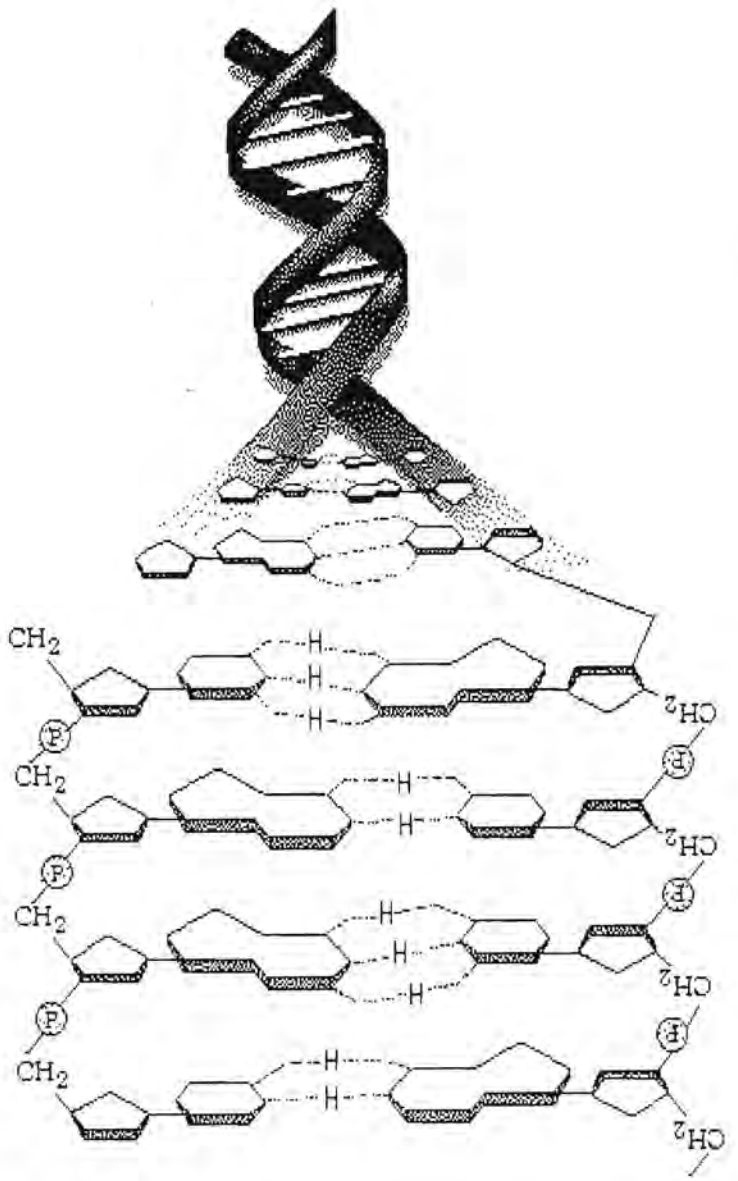
The hybridisation reaction was set up by mixing the driver and tester amplicons at a ratio of driver : tester of 100 : 1. The hybridisation mixture was then precipitated and the pellet redissolved in 4 μl 3X EE (30mM EPPS, 3 mM EDTA; pH 8) buffer. The solution was collected at the bottom of the tube and overlaid with light mineral oil so that the spherical droplet could be seen to be completely covered by oil. The DNA was then denatured at 98°C for five minutes and one μl of 5 M sodium chloride solution was added. Hybridization was done overnight at 67°C. Following hybridisation the reaction mixture was amplified by PCR for 10 cycles after which the amplicons were digested with mung bean nuclease (Amersham Life Science, UK) remove all single stranded DNA. Following the mung bean nuclease treatment the hybridization mixture was once again submitted to PCR amplification.

3.4.3 Cloning of the difference products

Two µg of these subtraction products were digested with the 50 units of the appropriate restriction enzyme and cloned into the p*Bluescript*II vector (Strategene, USA) that was used to transform competent XL1*Blue* cells. Fifty plasmid-containing colonies carrying an insert were selected and probed with either the 'Barhee' or 'Medjool' labeled driver amplicons using the *Gene Images* random prime-labelling module (Amersham Life Sciences UK). Ten colonies that showed a much stronger signal after hybridisation with the 'Barhee' amplicons than with the 'Medjool' amplicons were selected for plasmid isolation and determination of the insert sequence and size.

Chapter 4

Characterization of RDA subtraction products using bioinformatic tools



4.1 Abstract

Bioinformatics has advanced genetic research through the analysis of DNA sequences, the computerized processing of sequence data and comparison of unknown sequences with known DNA sequences through database searches. In this chapter subtraction products obtained by subtracting the two date palm genomes derived from the varieties 'Barhee' and 'Medjool' using the RDA technique were characterized with various bioinformatic tools. Results obtained showed that isolated subtraction products belong to the group of plant repetitive elements with a variety of sequence differences between the different copies of the Dp41 repetitive element, which was analysed in greater detail.

4.2 Objectives

The objective of this chapter was to characterize isolated subtraction products from the two date palm varieties 'Barhee' and 'Medjool' using bioinformatic tools to obtain detailed information about the DNA sequence of products and possible homologies to known DNA sequences available in DNA databases.

4.3 Results

4.3.1 DNA sequence analysis

Sequence analysis of the cloned subtraction products was carried out by using a commercial sequencing kit (Roche Molecular Biochemicals, Mannheim, Germany) using T7 and SP6 primers (see annex for sequence). DNA sequence analysis of 11 cloned subtraction products, revealed that these products consisted of at least three types of sequences with lengths of 141 bp (Dp41), 147 bp (Dp36), and 156 bp (Dp2) (Figure 4.1). When these three DNA sequences were aligned with all the DNA sequences determined, most of the sequenced subtraction products were homologous to Dp2, with only slight base differences occurring (Figure 4.2).

```
Dp2    CCTATCGAAC CCATTCATAC AGAGCCAGTA TTCAATGTCC CTCAACCATC  50
Dp2    GCGCGGATCT AGTAGGGTCT CCCATCCTCC CGATAGATAC TTAGGTATTC 100
Dp2    TAGAAGAGGA TACCGAGGAA ATGTTCCTAG TGGGAGATAG GGATCACATA 150
Dp2    CAGGAT                                                 156

Dp36   CCTATCGACG ACAGGCTGAC ATGGCAATTG TGCCGCACCA ATCATGCTCG  50
Dp36   GATAGGAAAG AGTCGACCTC GACGAAAGCG GCTCGGTAAA GCCCCGGTAT 150
Dp36   ACTCCAACAA GTCCGGGTCA ATCCGACGGT ATCTCCTCGC GCTGGAT    147

Dp41   CCTTCTCCCG GTAGGATCCG GCCTCACCGC AAATCCTGCA AGTATGACTG  50
Dp41   AGGGGGAAGA AGAAGGAGGG GACTCCGGAC CTGCCGTCGC GTCGTGGGGA 100
Dp41   CACCGTAGAT GGCTCGGTAG GTTGCCTTTC CTCCGTTGGA T          141
```

Figure 4.1. The three main RDA subtraction products isolated.

```
Dp2    CCTATCGAACCCATTCATACAGAGCCAGTATTCAATGTCCCTCAACCA  TCG     C
Dp3    -----A----------A----------CA-G----------CC---TAGAT-
Dp8    --------G------A----------CA-G----------CC---TAGAT-
Dp12   ----------A----------CA-G----------CC---CAGAT-
Dp18   ----------T-------AG---------CA-G----------CC---TAGAT-
Dp26   -----T----------CA-G---------G-GGGCC---TAGAT-
Dp33   ----------A----------A----------GG------CAGAT-
Dp36   -----G-CGA--GGCTGAC-TGGCA-T-GTG-CGCAC-AA---TG-    TCG  GAT
Dp41   ---TCTCCC-GT-G-A-CCGGCCT--CCGCAAATC-TG-AAGTAT-ACTGAG-GGGAAG
Dp47   ----------T----------C--CA-G----------CC---CAGAT-
Dp50   ----------A----------CA-G----------CC---TAGAT-

Dp2    GCGGATCTAGTAGGGTCTCCCATCCTCCCGATAGATACTTAGGTATTCTAGAAGAGGAT
Dp3           ----------A----------
Dp8           A----------A----------
Dp12           ----------T----------A----------
Dp18          ----------A----------
DP26          -------T----------G----------A----------
Dp33          A-CGACC--GA-GA-AG-GG-TCGGTA-AG-CCC------ACTCC-ACA --
Dp36         AGGAA--A---GA--  --GA-GAA-GC-GCT-GGTAAAGCC-CG-T-TAC  -
Dp41         AAG-AGGAGGGG--TCCG-A-GTGCCTC-CGTCGT--GGACACC-T--ATGGC
Dp47          A----------A---------
Dp50          ----------A----------

Dp2    ACCGAGGAAATGTT CCTAGTGGGAGATAGGGATCACATACAGGAT
Dp3    ------A-------  ----------A-A----------
Dp8    ----------  ----------CG------
Dp12   ------A-------T----------A----------
Dp18   ----------  ----------
DP26   ----------T----------
Dp33   ----G-TC---CC-TT----------G-G------
Dp36     --A-CA-GTCCGG GTCAA-CC--CGGTATCTC-TCGCGCT----

Dp41   TCGGTAGGTTGCC-TTCCTCC-TT                  ----
Dp47                                            ----
Dp50   ----------  ----------
```

Figure 4.2. Alignment of the different date palm RDA subtraction product sequences

4.3.2 Bioinformatic sequence analysis

Sequences were first analysed using the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990). The standard nucleotide-nucleotide BLAST (*blastn*) option was used. This allows for the unknown sequence to be compared to all nucleotide sequences in the database. Using this approach no significant homology was found between any of the isolated difference products with any known plant DNA sequences in available databases. Secondly the sequence data was subjected to the FASTA algorithm. Using FASTA version 3.4t (Pearson and Lipman, 1988), alignment algorithm provided by the European Bioinformatics Institute (http://www.ebi.ac.uk/fasta33), the Dp41 sequence showed homology against many known *Oryza sativa* (rice) genomic DNA sequences. And lastly homology searches were preformed using the Smith-Waterman Algorithm. When the Smith-Waterman algorithm (Smith and Waterman. 1981) was used to produce local alignments between the Dp41 sequence and database sequences the best homology (75%) found, was between genomic DNA from chromosome 1 of *Oryza sativa* (accession number AP002902) and Dp41 (Figure 4.3). No homology with known DNA sequences was found for Dp2 or Dp36.

4.3.3 Primer design

From the sequence information obtained by sequencing the different cloned subtraction products, three primer pairs were designed to represent the three different groups of sequences obtained. Primer pairs DP36L and DP36R; DP41L and DP41R; DP50L and DP50R (Table 4.1) were designed from the sequence information of Dp41, Dp36 and Dp50. When these primer pairs were tested on tissue culture-derived date palm plants fragments with the expected sizes were amplified from all plants regardless if 'Medjool' or 'Barhee' template was used (Figure 4.4). When genomic DNA from non-tissue culture derived plants were used as template DNA, amplification with different primer pairs varied and genomic DNA from two 'Barhee' plants did not produce any amplification product regardless which primer pair was used for amplification (Figure 4.5).

```
Dp41          6        TCCCCGTAGTAACCGGCCTCCC..CGCAAATCTT..GCAAGTA..  44
                       ::::::  :  :::  ::::::  :::   ::::  :: ::::
AP002902      89189    tccccgacgccacctccctcccggcgc...tcttcagcgagtagc
89230

Dp41          45       .....TCACTGAGGGGGA...AGAAGGAGGAGGGGCCTCCGCACG  81
                            :: : ::  ::::   ::  :::: :::::  : ::::::
AP002902      89231    tcacctcgcagatcgggatggagtgggagtagggg.ttgcgcacg
89274

Dp41          82       TGCCTCACGTC........GTGGGGACACCG.............. 104
                       : :: : :: :        :::: : : :::
AP002902      89275    t.ccacgcgccccgcgagggtggcgccgccgcggcccacgctctg
89318

Dp41          105      ..TAGATGGCTCGG.TAGGTTGCCTTTCCTCCG             134
                          ::  ::::::  :: :  ::   ::::::
AP002902      89319    gaacgacagctcggccagct..cc..gcctccg
89347


Local: Dp41 vs AP002902                    %similarity = 75.83
Score: 201.00
```

Figure 4.3 Local alignment between Dp41 and *Oryza sativa* chromosome 1 DNA, accession number AP002902, using the Smith-Waterman algorithm.

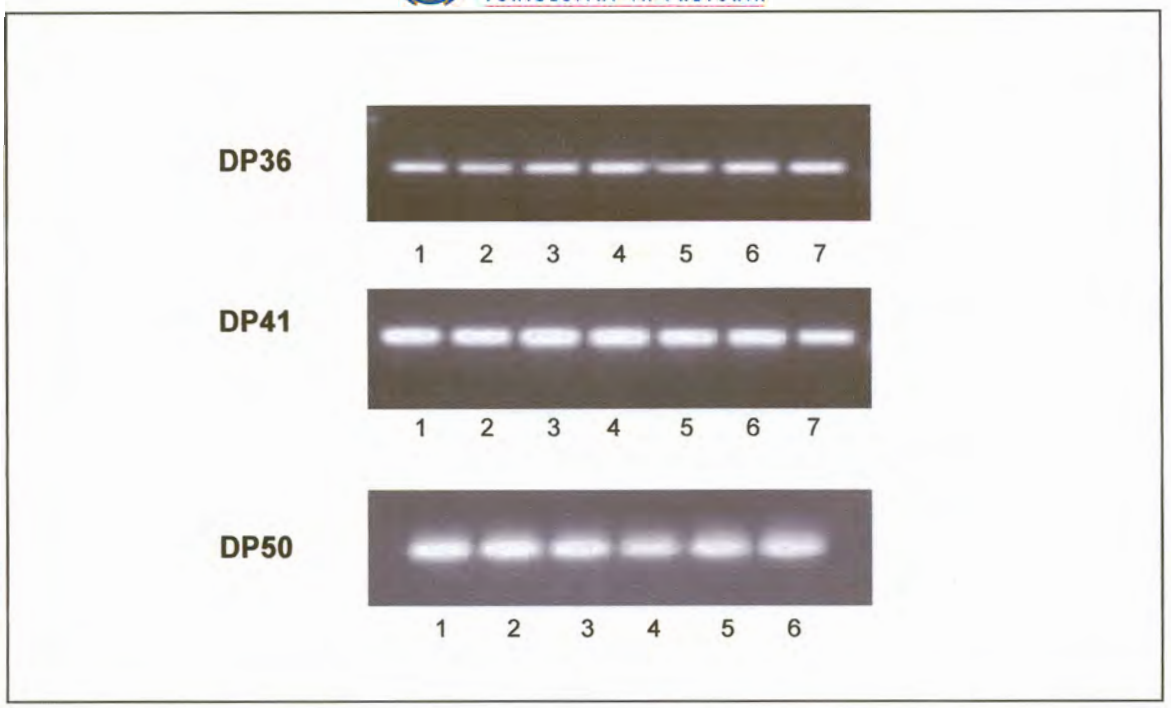


Figure 4.4 Amplification products obtained with the three different primers pairs to amplify RDA subtraction products DP36 (100 bp), DP41 (100 bp) and DP50 (120 bp). The products were amplified from seven individual tissue culture-derived 'Medjool' plants (DP36 and DP41) and six individual plants (DP50). Amplification of subtraction products using 'Barhee' DNA from tissue culture plants as template DNA resulted in identical products.

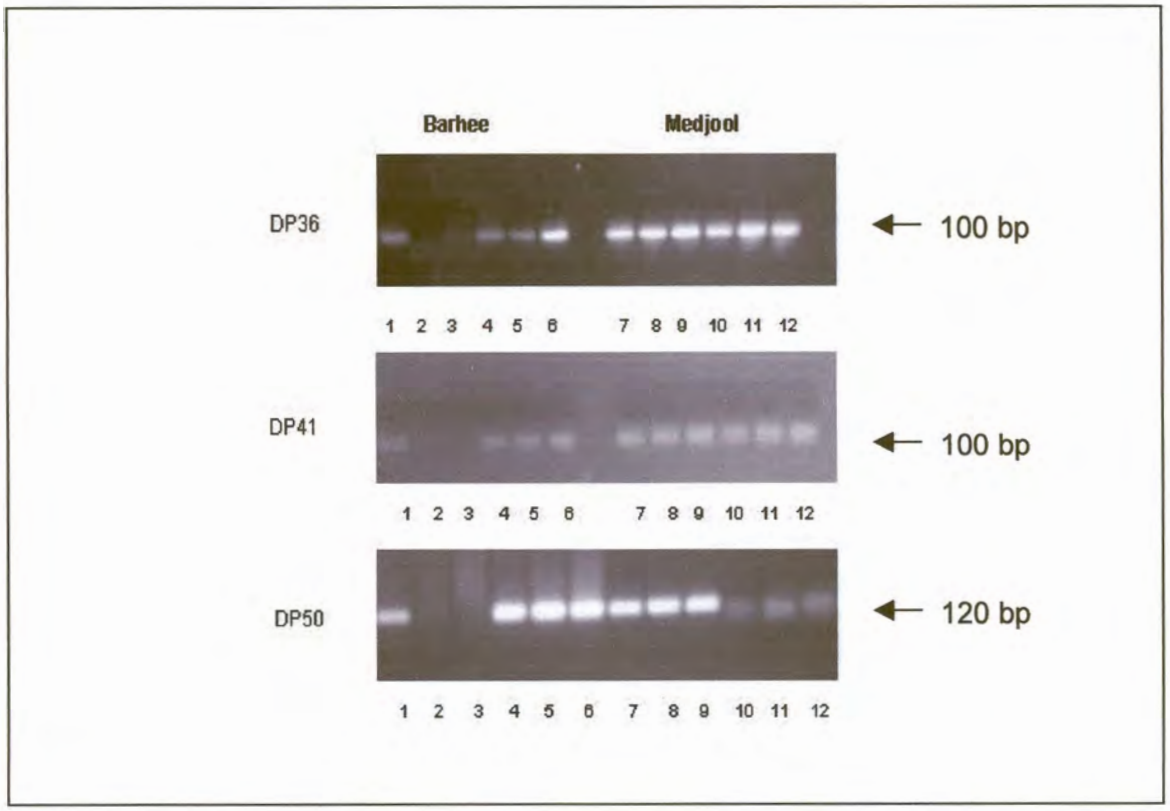


Figure 4.5 Variation in amplification of product from genomic DNA of different non-tissue culture plants. Lanes 1- 6: Amplification of DNA from 6 non-tissue culture-derived 'Barhee' plants; lanes 7- 12: Amplification of DNA from 6 non-tissue culture-derived 'Medjool' plants.

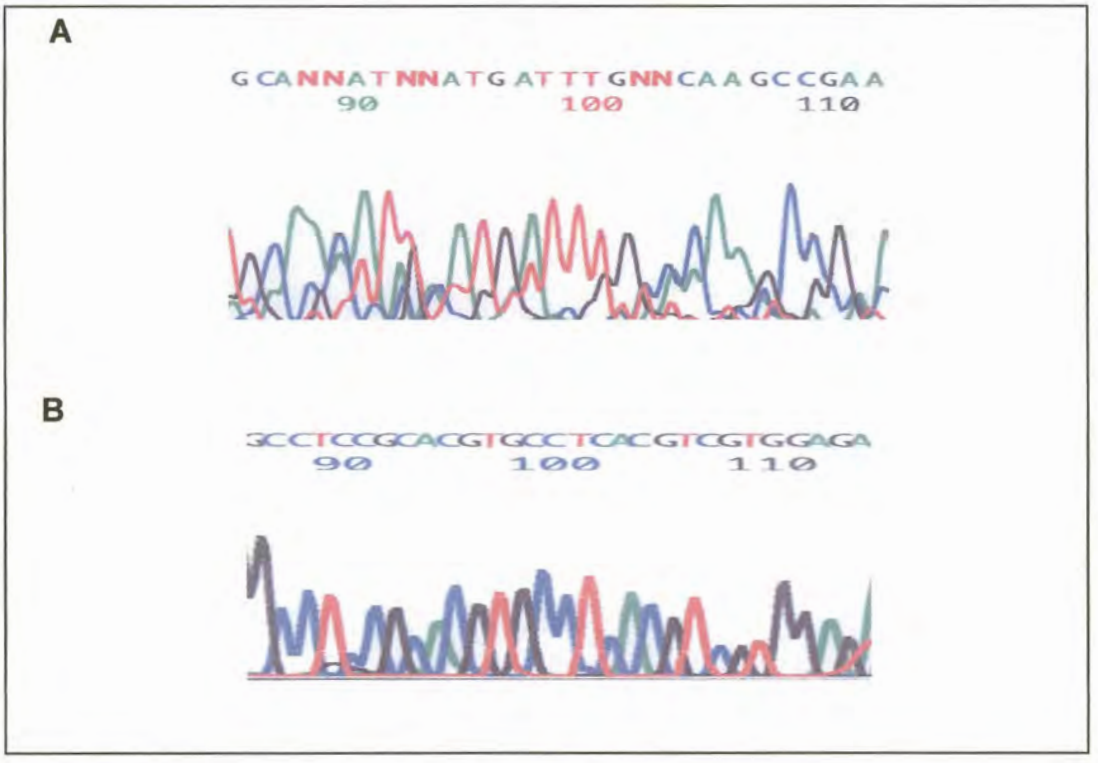


Figure 4.6 Electropherogram data from direct sequencing (A) with unacceptable high background noise, and from sequencing of cloned product (B) with acceptable background noise.

4.3.4 Detailed characterization of product Dp41

The Dp41 amplification product was characterized in more detail. From each of the six tissue culture-derived 'Barhee' and 'Medjool' plants, three independent clones of the cloned Dp41 amplification product were sequenced. As a first approach direct sequencing was applied. Using this approach, a high background noise was obtained in the sequence data and the obtained sequences were of very low quality not suitable for sequence analysis (Figure 4.6). Therefore the amplification products were first cloned and then sequence. The second approach resulted in good sequence quality.

Thirty-six sequenced clones from different 'Medjool' and 'Barhee' plants revealed a high degree of homology (above 95% sequence simalarity) with the original Dp41 subtraction product. Only those DNA sequences showing differences in comparison to Dp41 are shown in Figure 4.7. Detected differences were single base pair changes / deletions occurring mainly in a variable 45 bp region of the fragment (Figure 4.1, sequence in red). In general, more changes were observed between the individual 'Barhee' sequences than in the 'Medjool' sequences. Changes detected were specifically single base deletions in this variable region. From the 18 different sequences analyzed for each variety, 6 of the 'Medjool' sequences and 8 of the 'Barhee' sequences were identical to Dp41. This indicates that the region of the Dp41 sequence used for primer design is common to both genome types. Among the sequence variants found within the two genome- types, two, namely M6 and B9, were identical. Two sequence variants, M1 and M4, the latter with a six base deletion in its sequence, were unique to 'Medjool' and one variant (B11) was unique to 'Barhee'.

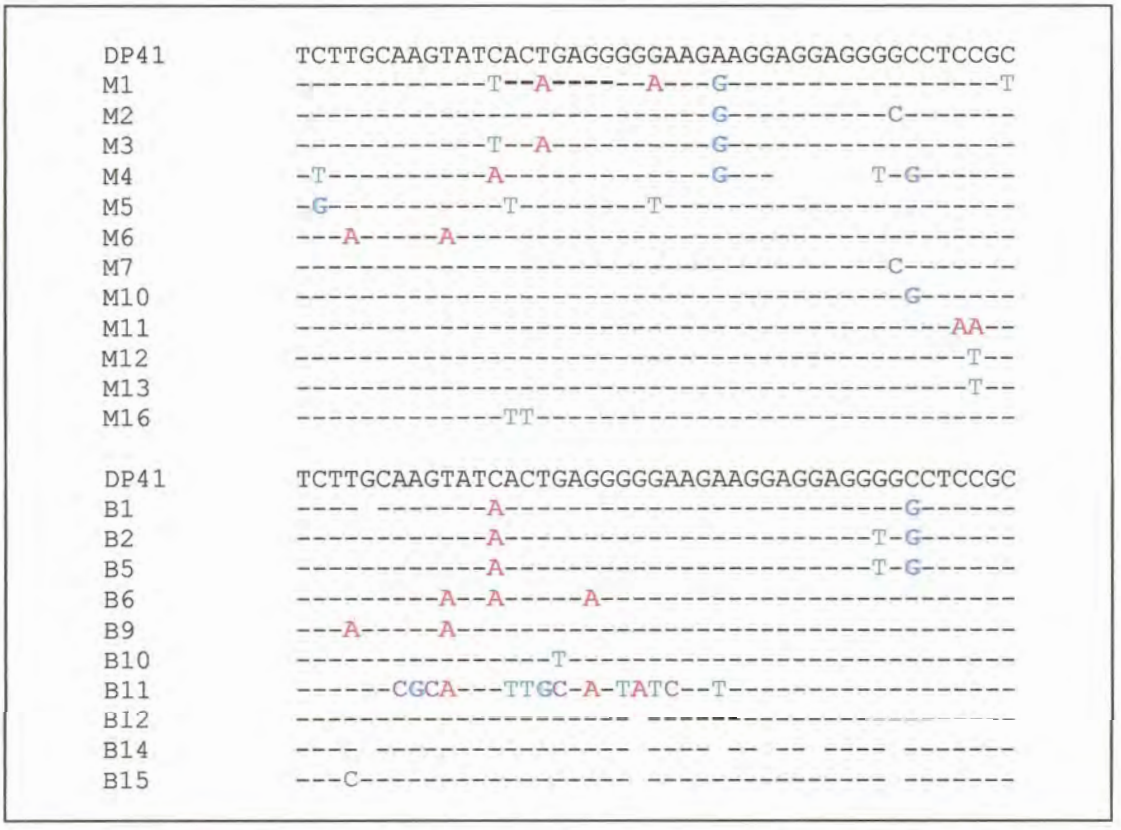


Figure 4.7. Differences found within the DNA sequence of amplified Dp41 fragment from different individual plants.

Primers PLM1, PLM4 and PLB11 (Table 4.2) were therefore designed from sequence M1, M4 and B11, to cover the variable portion of these sequences (Figure 4.6) when used in conjunction with DP41R. All 6 tissue culture-derived 'Medjool' plants and also all tissue-culture-derived 'Barhee' plants, which originated from a single mother plant, amplified a PCR product with the expected size with all three primers used. However, primer PLB11 at an optimal annealing temperature of 65°C only amplified a PCR product with the expected size of about 110 bp from 2 of 7 non-tissue culture-derived 'Barhee' and 6 of 7 'Medjool' plants. An identical result was observed with primer PLM1 at an optimal annealing temperature of 60°C. Primer PLM4 (at 65°C annealing temperature), which covered a unique 6 base pairs deletion, amplified a PCR product from DNA of all 7 non-tissue culture-derived 'Medjool' plants (Figure 4.2) but only from one 'Barhee' plant.

4.4 Materials and methods

4.4.1 Sequence analysis

Sequence analysis was done as outlined in the Annex.

4.4.2 Bioinformatic sequence analysis

For the BLAST (Altschul *et al.*, 1990) analysis the standard nucleotide-nucleotide BLAST or BlastN option was used. The 'nr' database option was used allowing the algorithm to search all GenBank, EMBL, DDBJ, PDB sequences. A low complexity filter was chosen and the 'expect' or the statistical significance threshold were set at 10 and the word size at 11. FASTA version 4.3t was used (Pearson and Lipman, 1988). From the FASTA search submission form the following options were picked. The program chosen was Fasta3. The data base option the following to parameters were chosen: (1) Nucleic acid and (2) EMLB. Gap penalties: Open -16 and Residue -4. Scores and Alignment were both set at 50. The KTUP value was set at 2 and both DNA strands were scored. The E-value threshold was set at 10. For the Smith-Waterman analysis (Smith and Waterman, 1981) the following parameters were used. Matrix: NUC4X4HB; Gap open penalty: -10 and Gap extention penalty: -5.

4.4.3 Primer design and testing

Pairs of primers were designed using a standard design program (Expasy, Switzerland) and the sequence information obtained for 4 of the different subtraction products. The primer pairs were used in a PCR reaction using 'Barhee' and 'Medjool' DNA as template at various annealing temperatures to optimize the PCR reaction. The amplification products were separated on a 1.5 % agarose gel, stained with ethidium bromide and visualized under UV light. See Annex for details on primer sequences.

Table 4.1. Primers designed for amplification of subtraction products.

Primer	Sequence
DP36L	5'-CTATCGACGACAGGCTGACA -3'
DP36R	5'-GACCCGGACTTGTTGGAGTA-3'
DP41L	5'-CCTTCTCCCCGTAGTAACCG-3'
DP41R	5'-AGGAAAGGCAACCTACCGAG-3'
DP50L	5'-TACACGATGTCCCTCAACCA-3'
DP50R	5'-GGAACATTTCCTCGGTATCC-3'

Table 4.2 Primers designed to include changes identified within the different amplified Dp41 products.

Primer	Sequence
PLM 1	5'-TTACAGAGGGGAAAGGAGGA-3'
PLM 4	5'-GGAAGGAGGTGGCTCCG-3'
PLB11	5'-CGCAATCTTGCAAGTATCAGT-3'