13
Reduced Representation Strategies and Their Application to Plant Genomes

Daniel G. Peterson

Overview

Many important crop species have large, highly repetitive genomes that make whole genome sequencing and assembly technically difficult and/or prohibitively expensive. However, there are a growing number of high-throughput "reduced representation" strategies that allow isolation and study of important and/or interesting sequence subsets from even the largest plant genomes. The following is a review of some of the major reduced representation techniques that are currently being utilized to study plants. In particular, the merits and limitations of strategies for sequencing gene space and/or repetitive elements will be discussed. Additionally, techniques for de novo discovery of DNA polymorphisms will be reviewed. Specific techniques addressed include the following:

1. Expressed sequence tag (EST) sequencing, which is currently the most economical means of elucidating the coding regions of expressed genes (Rudd 2003).
2. Methylation filtration (MF), a gene enrichment technique based on the observation that, in some plants, genes are hypomethylated compared to repeats (Rabinowicz et al. 1999).
3. Cot-based cloning and sequencing (CBCS), a technique rooted in the principles of DNA renaturation kinetics that allows enrichment for genes, repeats, or any other group of sequences based upon their relative iteration in the genome (Peterson et al. 2002a,b).
4. Reduced representation shotgun (RRS) sequencing, a means of small polymorphism discovery rooted in some of the most basic molecular biology techniques, i.e., restriction enzyme digestion of DNA and agarose gel electrophoresis (Altshuler et al. 2000).
5. Degenerate oligonucleotide primed PCR (DOP-PCR), a small polymorphism discovery tool in which partially degenerate primers are used to amplify a subset of genomic sequences which are then cloned, sequenced, and compared (Jordan et al. 2002).

The Handbook of Plant Genome Mapping.
Genetic and Physical Mapping. Edited by Khalid Meksem, Günter Kahl

ISBN 3-527-31116-5

6. Microsatellite capture techniques that use synthetic oligonucleotides composed of short tandem repeats to discover simple sequence repeats (SSRs) in genomic DNA.

13.1 Introduction

"*Everything is simpler than you think and at the same time more complex than you imagine*."

Johann Wolfgang von Goethe (1749-1832)

Seed-bearing plants (angiosperms and gymnosperms) exhibit considerable conservation with regard to relative gene order and overall gene repertoire. This conservation is evident in comparisons of even the most distantly related taxa; for instance, loblolly pine and *Arabidopsis* diverged from a common ancestor 300 million years ago, yet 90% of pine EST contigs have apparent homologues in *Arabidopsis* (Kirst et al. 2003). In contrast, the genome sizes (1C DNA contents) of seed plants exhibit tremendous plasticity, e.g., loblolly pine has a genome 162 times larger than that of *Arabidopsis*, though it is clear that the former does not have or need 162 times as many genes. Of note, species within the same family may exhibit up to 100-fold differences in 1C DNA content, and 10-fold differences have been observed for species in the same genus (Bennett and Leitch 2003). At the extreme, the angiosperms *Fritillaria assyriaca* (a lily) and *Cardamine amara* (a mustard) show a 2123-fold difference in genome size (Bennett and Leitch 2003) although both species possess comparable levels of structural sophistication. The lack of correlation between genome size and structural complexity/gene repertoire in higher eukaryotes is known as the *C-value paradox*, and its evolutionary implications have been a subject of study and debate for decades (see Hartl 2000 and Petrov 2001 for reviews).

The vast majority of genome size variation in seed plants is due to lineage-specific amplification of non-genic "repeat sequences," some of which may be found in thousands or millions of copies per 1C genome. While a few of these repeats have come to serve structural roles (e.g., centromeric and telomeric repeats), most have no known function. Many of the repetitive elements in plant genomes appear to have originated from intergenic proliferation of transposable elements (SanMiguel and Bennetzen 2000), while others have uncertain origins (Lapitan 1992). While polyploidy, gene duplication, and gene loss certainly account for some of the variation in seed plant genome sizes, their contributions to the C-value paradox appear to be rather small (Hartl 2000). For instance, polyploidy accounts for < 0.5% of the > 2000-fold variation in plant genome sizes[1], and gene duplications and losses likely account for even less.

[1] In the RBG Kew Plant C-Values Database (Bennett and Leitch 2003), seed plant species vary in ploidy from 2X to 20X (10-fold). A 10-fold variation in ploidy divided by the 2123-fold observed variation in DNA content among these same species is 0.5%.

Whole genome sequences provide the ultimate data set by which the DNA of different species can be compared. However, many important crop species have large genomes (Figure 13.1) in which repetitive sequences constitute the bulk of genomic DNA, are highly interspersed with genes, and/or have relatively recent origins (i.e., exhibit little inter-copy divergence), making gene isolation and whole-genome sequence assembly prohibitively difficult and costly (Rudd 2003). In recent years, a variety of "reduced representation" techniques have been developed to isolate and study important and/or interesting sequence subsets from large, repetitive genomes in a cost-effective manner; subsets include (but are not limited to) expressed sequences, low-copy (≈ gene-rich) genomic regions, polymorphic DNA markers, and repetitive elements. Use of some reduced representation methods may permit capture and elucidation of a species' "sequence complexity" (SqCx - Figure 13.2) and thus provide most of the benefits of whole genome sequencing at a fraction of the cost. Other reduced representation techniques allow rapid characterization of DNA polymorphisms without a priori knowledge of genomic sequence, affording inroads into the sequence diversity of under-explored genomes and providing mechanisms for efficient genotyping in those species that enjoy finished sequences.

This chapter focuses on those reduced representation strategies that will help in the sequencing of plant gene space, allow efficient characterization of the repetitive elements of genomes, and permit de novo discovery of DNA polymorphisms. The strengths and limitations of each reduced representation technique are discussed.

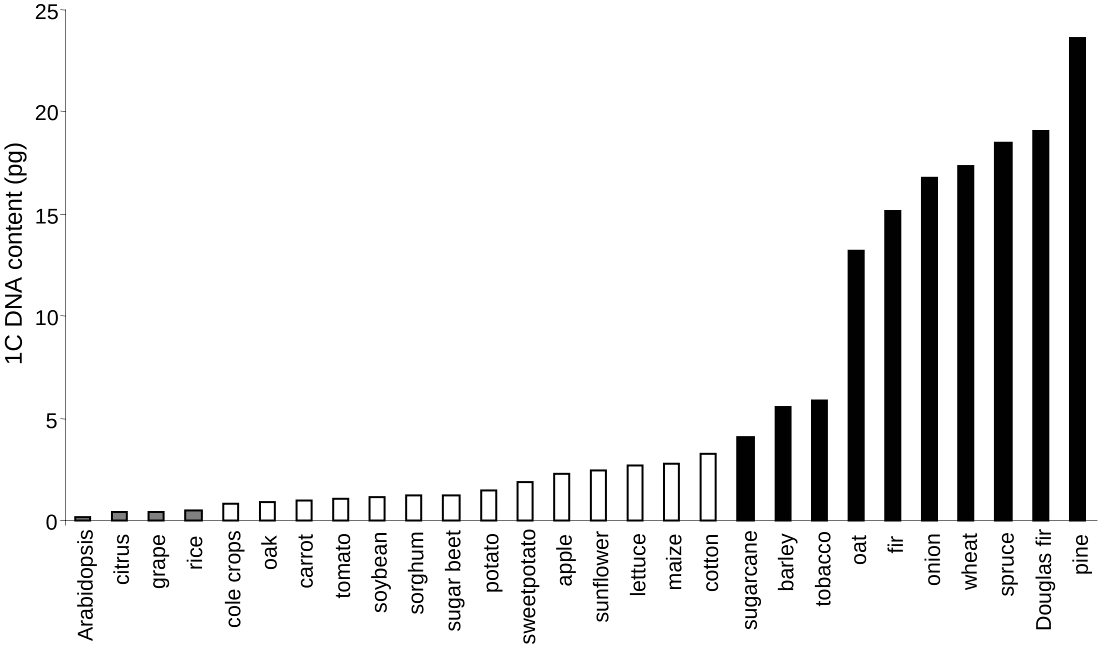

Figure 13.1. Comparative genome sizes of economically-important crop species and *Arabidopsis*. All 1C genome size data is from Bennett and Leitch (2003) except the value for *Arabidopsis* (Arabidopsis Genome Initiative 2000). Values for pine, spruce, oak, fir, and Douglas fir are mean genome sizes for their respective genera. Species with genome sizes equal to or smaller than rice are shown in gray, those with genome sizes larger than human (3.26 pg) are shown in black, and those with 1C DNA contents in between rice and human are represented in white.

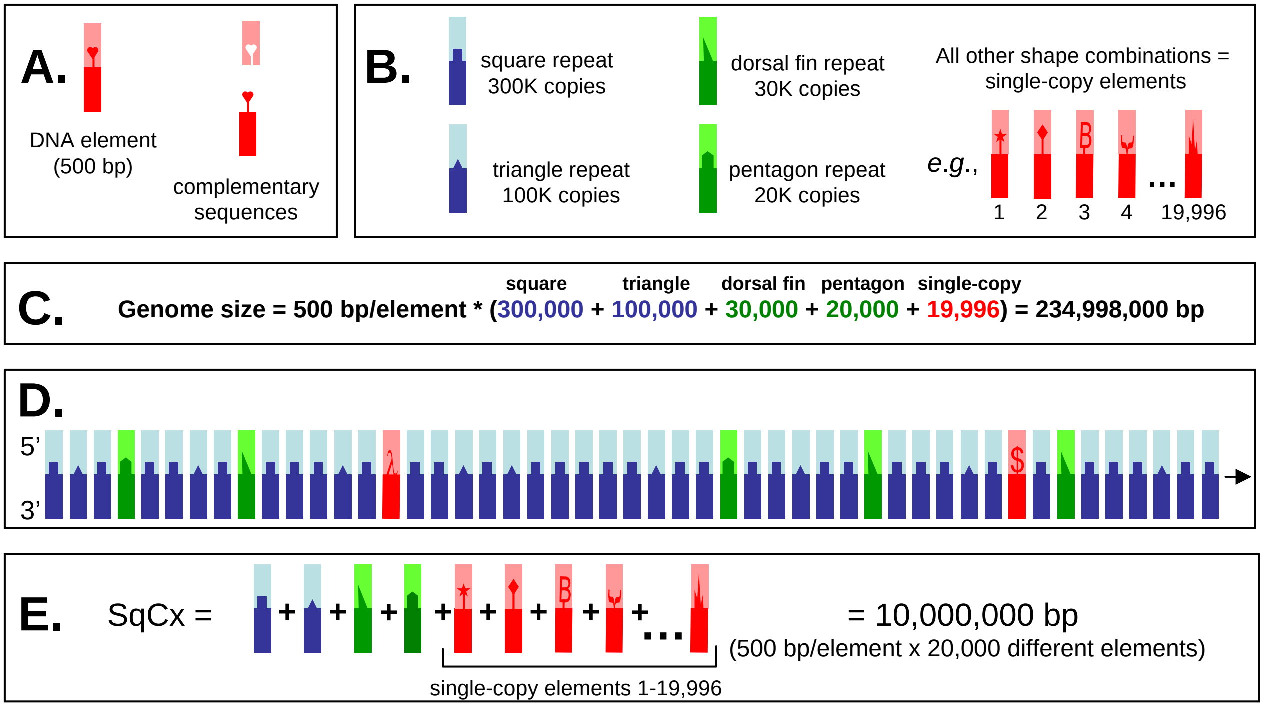
A.
DNA element
(500 bp)
complementary
sequences
B.
square repeat
300K copies
triangle repeat
100K copies
dorsal fin repeat
30K copies
pentagon repeat
20K copies
All other shape combinations =
single-copy elements
e.g.,
1
2
3
4
...
19,996
C.
square triangle dorsal fin pentagon single-copy
Genome size = 500 bp/element * (300,000 + 100,000 + 30,000 + 20,000 + 19,996) = 234,998,000 bp
D.
5'
3'
E.
SqCx =
+ + + + + + + + ...
= 10,000,000 bp
(500 bp/element x 20,000 different elements)
single-copy elements 1-19,996

Figure 13.2. Genome size vs. sequence complexity (SqCx). (A) The genome of the hypothetical plant *Planta genericus* is composed of a variety of DNA elements; each element is represented by a rectangle composed of two interlocking (i.e., complementary) pieces/sequences. For simplicity, each element is assumed to be 500 bp in length. (B) The two most highly repetitive (HR; blue) elements in the genome are "square" (300,000 copies/genome) and "triangle" (100,000 copies). Considerably fewer copies are found of the moderately repetitive (MR; green) elements "dorsal fin" (30,000 copies) and "pentagon" (20,000 copies). All other 19,996 elements in the genome are single/low-copy (SL; red) in nature. (C) 63.8% of the *P. genericus* genome is composed of copies of "square" while "triangle," "dorsal fin," and "pentagon" constitute 21.3%, 6.4%, and 4.3% of the genome, respectively. The remaining 4.2% of the genome is composed of the 19,996 different SL sequences. (D) A short stretch of *P. genericus* DNA showing the comparative frequency of HR, MR, and SL elements. (E) SqCx is the sum of all the unique sequence information in a genome. While "square" and "triangle" account for the vast majority of *P. genericus* genome size, the contribution of each of these elements to SqCx is negligible [100 x (500 bp ÷ 10,000,000 bp) = 0.0005%]. In contrast, the SL sequences, which constitute only 4.2% of genome size, account for 99.9% of SqCx. The goal of gene-enrichment reduced representation techniques is to isolate and sequence those elements which contribute most to SqCx (i.e., SL DNA) with minimum encumbrance from the repetitive elements which may make up the lion's share of the genome but contribute very little to SqCx.

13.2 Reduced Representation Techniques

13.2.1 EST Sequencing

Reverse transcriptase (Baltimore 1970; Temin and Mizutani 1970) is an RNA-dependent DNA polymerase. In nature, the enzyme is the means by which the genomes of retroviruses (i.e., viruses with RNA genomes) make *complementary DNA* (cDNA) for incorporation into a host's genome. With regard to molecular biology, the use of reverse transcriptase to generate cDNA molecules from isolated mRNA and the subsequent construction of cDNA libraries heralded in the era of gene expression research. Because each cDNA library serves as a "snapshot" of gene expression in the cells from which it was derived, comparison of cDNA libraries from different tissues, developmental stages, or from the same tissues exposed to different environmental stimuli, affords a means of correlating changes in morphology and cellular activity with changes in gene expression.

With the development of PCR and automated sequencing, it became possible to sequence large numbers of end sequences from cDNA clones. The cDNA end sequences, commonly referred to as *expressed sequence tags* (ESTs; Adams et al. 1991), can be utilized as molecular markers, a discovery that has greatly accelerated molecular mapping efforts (e.g., Komulainen et al. 2003). Because cDNA/EST sequencing results in the preferential sequencing of portions of expressed genes, it is a powerful reduced representation technique that allows an economical preliminary means of exploring plant gene space (see Rudd 2003 for review).

13.2.2 Methylation Filtration

DNA methyltransferases are enzymes that add methyl groups onto select bases of DNA. In plants, DNA methyltransferases normally catalyze the transfer of the methyl group from S-adenosyl-L-methionine to the fifth carbon in the pyrimidine ring of cytosine to form 5-methylcytosine (m5C) (Kumar et al. 1994). However, cytosine residues are not methylated indiscriminately. Rather, methylation is highly regulated, and it has been correlated with all kinds of genetic phenomena, including normal control of gene expression (e.g., Finnegan et al. 1998, 2000; Zluvova et al. 2001), imprinting (e.g., Kinoshita et al. 2004; Berger 2004), paramutation (e.g., Lisch et al. 2002), transgene silencing (e.g., Fojtova et al. 2003; Meng et al. 2003), aging (e.g., Fraga et al. 2002), repression of recombination (e.g., Fu et al. 2002), diploidization in allopolyploids (e.g., Liu and Wendel 2003), and rapid epigenetic adaptation in response to major environmental changes (e.g., Fraga et al. 2002; Kovalchuk et al. 2003; Fojtova et al. 2003).

Early attempts at cloning methylated DNA (including m5C-rich DNA from plants) were not particularly successful (see Redaschi and Bickle 1996 for review).

The principal reason for this difficulty was elucidated in the late 1980s when it was shown that some *E. coli* strains possess enzymes that preferentially restrict foreign methylated DNA sequences. Like other restriction endonucleases, the three *E. coli* methylation-specific restriction enzymes (MSREs), known as McrA, McrBC, and Mrr, apparently evolved to protect the bacterium from invading bacteriophages. However, they are quite capable of cleaving methylated plant DNA as well (Redaschi and Bickle 1996). After the discovery of MSREs, *E. coli* strains were engineered with non-functional MSRE genes (i.e., with a *mcrA*-, *mcrBC*-, and/or *mrr*-genotype) to facilitate cloning of methylated DNA.

In some plant species, the presence of a high density of m5C residues (i.e., hypermethylation) is strongly correlated with certain repeat sequences (e.g., retroelements) while low-copy sequences typically contain few or no m5Cs (i.e., they are hypomethylated). A striking example of this type of methylation pattern is seen in maize (Bennetzen et al. 1994).

Recently, Rabinowicz et al. (1999) developed a clever and simple means of preparing genomic libraries enriched in hypomethylated sequences. In short, they cloned mechanically-sheared maize DNA into *mcrA*+, *mcrBC*+, and/or *mrr*+ host strains. Most hypermethylated DNA is cleaved in these strains, and thus the resulting libraries are enriched in hypomethylated (ostensibly gene-rich) DNA. An overview of their technique, methylation filtration (MF), is shown in Figure 13.3. The efficacy of MF in producing gene-enriched genomic libraries has been shown for maize (Rabinowicz et al. 1999; Whitelaw et al. 2003) and claimed for canola and wheat. Methylation filtration is licensed exclusively to Orion Genomics (www.oriongenomics.com) and is marketed under the name GeneThresher.

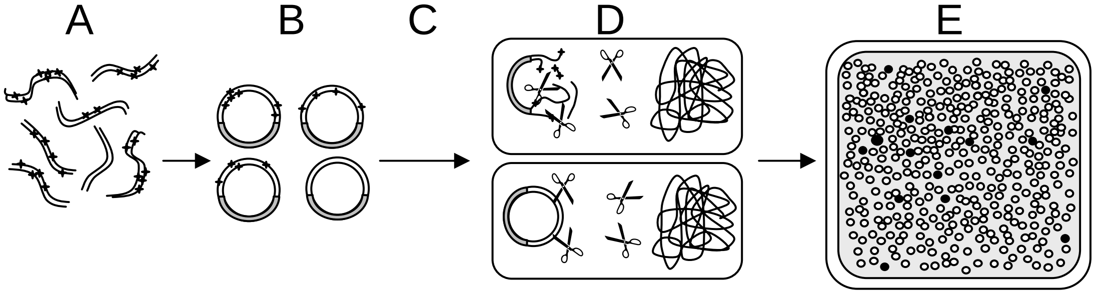

Figure 13.3. Overview of methylation filtration. (A) Plant genomic DNA is mechanically sheared into fragments. For the species shown, repeat sequences are hypermethylated compared to gene sequences. Hypermethylated DNA regions are indicated by small stars along the DNA strands. (B) The genomic DNA fragments are ligated into a vector containing an antibiotic resistance gene and a cloning site that allows alpha-complementation (blue/white selection). (C) The recombinant molecules are used to transform a strain of bacteria with active MSREs (i.e., McrAB, McrC, and/or Mrr gene products). (D) The MSREs, depicted by scissors, cleave plant hypermethylated DNA (upper cell) but do not restrict hypomethylated DNA (lower cell). (E) When challenged with an antibiotic on selective medium, only those bacteria that contain an intact circular plasmid will survive to form colonies. Colonies containing plasmids that lack an insert in their cloning site will appear blue, while those containing an insert in their cloning site will appear white. Archiving white clones results in a library enriched in hypomethylated (ostensibly gene-rich) DNA.

13.2.3 Cot-based Cloning and Sequencing

Much of what is known about eukaryote genome structure stems directly from work done by Roy Britten and colleagues during the 1960s and 1970s.[1] Britten and his collaborators at the Carnegie Institution of Washington empirically studied the re-association of genomic DNA in solution using a technique they called "Cot analysis," and it was through Cot analysis that the repetitive nature of eukaryotic genomes was discovered (Britten and Kohne 1968). In brief, when mechanically sheared DNA in solution is heated to near boiling temperature, the molecular forces holding complementary base pairs together are disrupted and the two strands of the double-helix dissociate or "denature." If the denatured DNA is then slowly returned to a cooler temperature, sequences will begin to "re-associate" (renature) with complementary strands. The temperature at which renaturation occurs can be regulated so that little or no sequence mismatch is tolerated. As predicted by the law of averages, the rate at which a sequence finds a complementary strand with which to hybridize is directly related to that sequence's iteration in the genome (Figure 13.4). In other words, those sequences that are extremely abundant (on average) find complementary strands with which to pair relatively quickly, while single-copy sequences take a much longer time to find complements. In a Cot analysis, a series of DNA samples are allowed to renature to different Cot values; a sample's Cot value is the product of its DNA concentration (C_0), re-association time (t), and, if appropriate, a buffer factor that accounts for the effect of cations on the speed of renaturation (Britten et al. 1974). The amount of re-association at each Cot value is typically determined using hydroxyapatite (HAP) chromatography to separate double- and single-stranded DNA (dsDNA and ssDNA) and spectrophotometry to quantitate the amount of DNA in ssDNA and dsDNA eluants. A graph showing re-association of genomic DNA as a function of Cot is known as a Cot curve. Through study of Cot curves of different species, Britten et al. discovered that eukaryotic genomes tend to be composed of several distinct kinetic components – specifically, highly repetitive (HR), moderately repetitive (MR), and single/low-copy (SL) DNA. Using a Cot curve as a guide, HAP chromatography can be used to isolate the different kinetic components of a genome (e.g., Britten and Kohne 1968; Goldberg 1978; Kiper and Herzfeld 1978; Peterson et al. 1998) as illustrated in Figure 13.4.

In the late 1990s, I began work as a postdoctoral associate in the lab of Andrew H. Paterson at the University of Georgia. My previous training in Cot analysis and nuclear DNA isolation (Peterson et al. 1997, 1998) coupled with Paterson's expertise in plant genetics/genomics (e.g., Paterson et al. 1995, 2000) eventually led us to develop "Cot-based cloning and sequencing" (CBCS), a synthesis of Cot methods, molecular cloning, and high-throughput DNA sequencing that permits

[1] The principles of nucleic acid re-association elucidated by Britten et al. lie at the heart of many molecular techniques utilized today, including PCR, filter hybridization (Southern/Northern blots, colony blots, macroarrays), microarrays, and chip-based re-association experiments (see Goldberg 2001 for review).

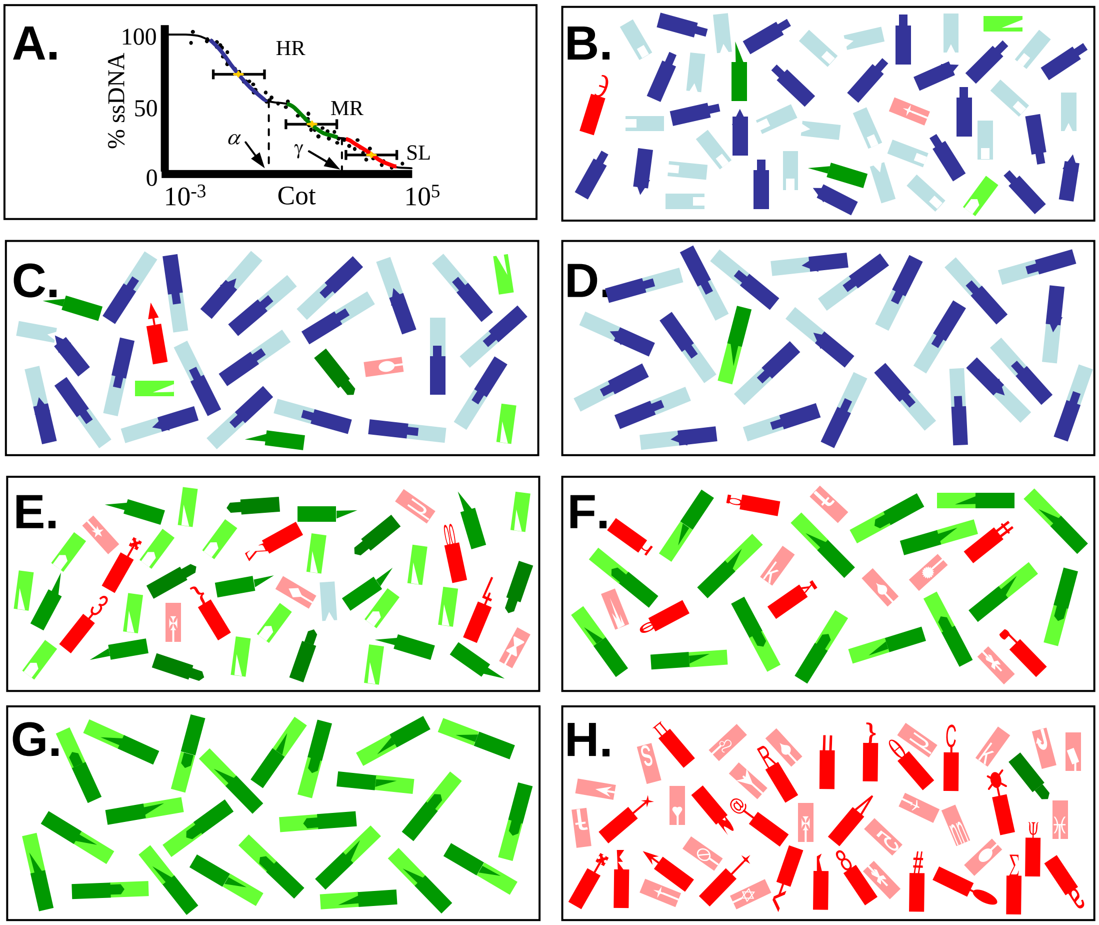

Figure 13.4. Fractionation of the hypothetical *Planta genericus* genome using CBCS. As in Figure 1, each element in the *P. genericus* genome is represented by a rectangle composed of two interlocking (i.e., complementary) pieces/sequences. Additionally, all elements are roughly 500 bp in length. (A) A Cot analysis of *P. genericus* reveals that its genome is composed of a fast (blue), a moderately fast (green), and a slow re-associating component. The fraction of the genome in each component and the kinetic complexity (estimated sequence complexity) of each component are determined from mathematical analysis of the Cot curve. Because the elements in a genome re-associate at a rate proportional to their copy number, those sequences in the fast re-associating component represent the most highly repetitive (HR; blue) sequences in the genome, while those in the moderately fast and slow re-associating components represent moderately repetitive (MR; green) and single/low-copy (SL; red) elements, respectively. (B) *P. genericus* DNA is sheared into 450-bp fragments and denatured. This illustration and each subsequent frame show a small, random part of a much larger renaturation reaction. However, each element within a frame is shown in correct proportion to elements from its own and other components. (C) The DNA is allowed to renature to the Cot value "α" (see frame A). As shown in the Cot curve (frame A), at Cot α nearly all HR elements have formed duplexes (based upon 'collisions' with complementary strands), but few MR and essentially no SL elements have found complements. Hydroxyapatite (HAP) chromatography then is used to separate (D) double-stranded (renatured) HR DNA from (E) single-stranded DNA (MR and SL DNA). The double-stranded HR DNA is cloned to produce an HRCot library. (F) The single-stranded DNA remaining after the first HAP fractionation is allowed to renature to a Cot value equivalent to γ (see frame A), at which virtually all MR elements have formed duplexes but single-copy elements are still unlikely to have found complements. HAP chromatography is used to separate the (G) double-stranded MR DNA from the (H) single-stranded SL DNA. The MR-enriched DNA fraction (frame G) is cloned to produce an MRCot library. Random primer strand synthesis is used to generate complementary strands for the single-stranded DNA molecules in the SL-enriched fraction (frame H), and the resulting duplexes are cloned to produce a SLCot library.

production and exploration of DNA libraries enriched in genes and/or repeats (Peterson et al. 2002a, 2002b). In CBCS, the results of a Cot curve (or genome parameters determined through other techniques) are used to guide HAP-based fractionation of genomic DNA into its major kinetic components (e.g., HR, MR, and SL DNA). The isolated kinetic components are cloned to produce HRCot, MRCot, and SLCot libraries, respectively, and component libraries are sequenced (Figure 13.4). For those solely interested in sequencing low-copy DNA (i.e., most genes), only an SLCot library need be prepared (e.g., Yuan et al. 2003; Lamoreux et al., submitted).[1] The efficacy of CBCS as a gene space enrichment tool has been demonstrated for sorghum (Peterson et al. 2002a), maize (Yuan et al. 2003; Whitelaw et al. 2003), wheat (Lamoreux et al., submitted), cotton (Paterson et al., in preparation), and chicken (Wicker et al., in preparation).

Of reduced representation techniques, CBCS is the only one that theoretically permits sequencing of a species' sequence complexity (SqCx; Figure 13.2). Because Cot analysis provides the kinetic complexity (i.e., the estimated SqCx) of each component, the most efficient means of capturing a species' total SqCx is to sequence each Cot library to a depth that provides a high probability that all the elements in that component are sequenced at least once (Peterson et al. 2002a, b). Since almost all of the SqCx of a plant genome will be found in its SLCot library, the vast majority of resources can be devoted to sequencing SLCot clones (Figure 13.4). For many plant genomes, CBCS should allow sequencing of SqCx at a cost of one-quarter to one-twentieth that of traditional shotgun sequencing (see Peterson et al. 2002b).

While in retrospect it seems rather obvious that DNA re-association could be used to create gene-enriched (or repeat-enriched) genomic libraries, there are several factors that likely contributed to the relatively late development of CBCS: (1) while Cot analysis was highly utilized in the 1970s, its popularity waned with the advent of molecular cloning; (2) Cot analysis (especially as it was practiced in the 1970s) was a technically demanding procedure requiring extreme standardization, and thus it was performed only in a handful of labs even in its heyday; (3) Cot analysis is considerably "less forgiving" than molecular cloning, and the tremendous appeal of the latter was such that most leaders in re-association kinetics shifted their research focus; and (4) by the time high-throughput sequencing became possible, many of the original practitioners of Cot analysis had retired.

Since publication of our original CBCS manuscript (Peterson et al. 2002a), we have been continually working to improve all aspects of CBCS. For example, in our initial experiments in sorghum we cloned SL duplexes that were the products of kinetic re-association. Although high stringency was maintained during renaturation to prevent base mismatches, any mismatch making it through to the clon-

[1] Working independently, J.L. Bennetzen and colleagues developed "high Cot" sequencing (Yuan et al. 2003), a technique that is essentially identical to enrichment for single/low-copy sequences using CBCS (i.e., sequencing of SLCot clones). Because the publications of Peterson et al. (2002a, 2002b) pre-date Yuan et al. (2003) and because, in our experience, people often confuse the term "high Cot" with "high copy" ("high Cot" DNA is actually composed of low-copy sequences), in this review we will use the terminology established by Peterson et al. (2002a).

ing process presumably would be resolved by the host cell's DNA repair mechanisms possibly resulting in generation of a sequence that does not exist in the donor genome. While this phenomenon would have minor implications in capturing SqCx, it would limit the usefulness of SLCot clones in the detection of small polymorphisms (e.g., SNPs) within gene and repeat families. By the time our first CBCS paper was published (Peterson et al. 2002a), we had already started using an enzymatic approach to synthesize complementary strands from single-stranded SL DNA templates, thus circumventing the problems associated with cloning reassociation products and increasing the utility of CBCS. We are working to successfully adapt a random priming approach to synthesize complementary strands for HR/MR molecules as well.[1] We are also conducting tests with a recently discovered nuclease that preferentially cleaves dsDNA (Shagin et al. 2002; Zhulidov et al. 2004) and consequently may allow elimination of HAP chromatography, thus speeding up and simplifying the CBCS protocol.

13.2.4 *De Novo* Polymorphism Discovery

Within a species, the vast majority of gene sequence variation is due to relatively small DNA polymorphisms. There are several types of small polymorphisms that are widely utilized to study plant gene/genome evolution.

1. Single nucleotide polymorphisms (SNPs): As their name suggests, SNPs are single base pair differences within alleles of a gene. They represent a powerful means of relating the smallest possible changes in DNA sequence to variation in phenotype. SNPs are widely utilized molecular markers in mammals and are becoming more common molecular markers in plants (Schmid et al. 2003; Törjék et al. 2003).
2. Insertion/deletion (indel) polymorphisms: Indels are typically discovered using SNP discovery approaches. Indels are useful genetic markers and are easier to score than SNPs as detection of the latter requires sequencing (or resequencing) while the former can be detected as length differences in PCR products (Bhattramakki et al. 2002).
3. Microsatellites or simple sequence repeats (SSRs): SSRs are tandemly repeated DNA sequences of 1-7 bp. They are abundant in the genomes of most eukaryotes, and because they are frequently polymorphic, codominant, and easily scored, they have been utilized as molecular markers in numerous pursuits including linkage map construction, parentage analysis, population genetics studies, and marker-aided selection (see Fisher et al. 1996 and Dekkers and Hospital 2002 for reviews). SSRs, which technically are a subclass of indels, can be discovered through analysis of large EST sets (e.g., Gupta et al. 2003).

An obvious means of studying SNPs, indels, and SSRs is to use PCR to examine polymorphisms at specific loci within a population. Likewise, genome/EST se-

[1] Single-stranded HR/MR DNA has to be imobilized to permit second-strand synthesis.

quence data from one species can be used to study parallel loci in closely related species. However, such research requires synthesis of locus-specific primers and thus a priori knowledge of the nucleotide sequence of the locus being studied. Below is a discussion of several reduced representation techniques that allow SNP, indel, and/or microsatellite discovery without existing sequence data.

13.2.4.1 Reduced-Representation Shotgun (RRS) Sequencing

Recently Altschuler et al. (2000) devised a means of studying SNPs and other small polymorphisms at a reasonable number of loci without prior knowledge of genomic sequence. Their technique, which they call reduced representation shotgun (RRS) sequencing, is quite simple and presumably applicable to most species (see Figure 13.5 for an overview). In brief, genomic DNA from a number of individuals is mixed together, digested with a single restriction enzyme, and size-fractionated by agarose gel electrophoresis. An agarose band containing DNA fragments within a relatively narrow size range (e.g., between 600 and 650 bp) is removed, and DNA fragments from the band are cloned and sequenced. Because most restriction sites will be shared by individuals in the population, the gel slice will likely contain alleles of the same loci. Polymorphisms can be detected by sequence analysis.

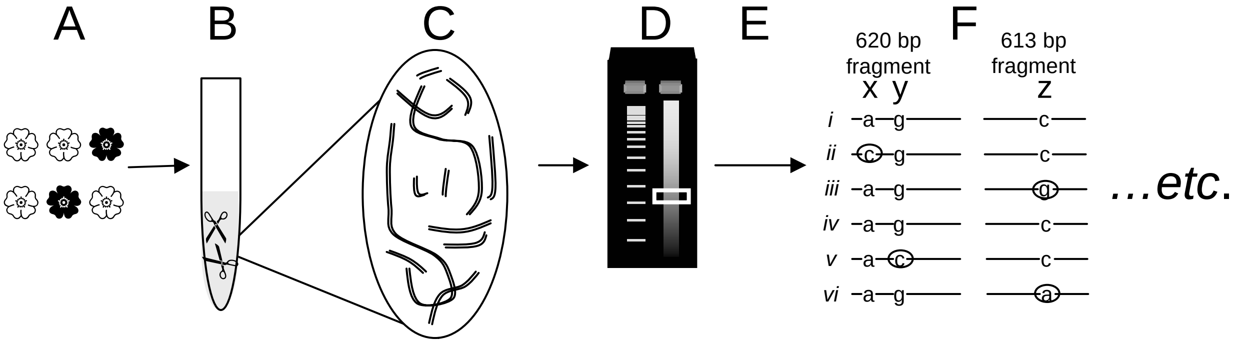

Figure 13.5. Overview of RRS sequencing. (A) DNA is extracted from multiple individuals in a population. (B) All of the DNA samples are placed in the same tube and a single restriction enzyme (depicted by scissors) is added. (C) After complete digestion, the DNA mixture consists of many fragments of different sizes. As the individuals in the population are likely to share most of the same restriction sites, most homologous DNA sequences are likely to be about the same size. (D) The digested DNA is size-fractionated via electrophoresis and a small band containing DNA fragments in a relatively narrow size range (e.g., 600-650 bp) is extricated from the gel. The DNA in the gel band is isolated, cloned, and sequenced. (E, F) Sequences are compared using sensitive alignment algorithms. SNPs and very small deletions and insertions can be discovered by comparing those sequences that appear to represent the same locus (i.e., are largely identical). For the 620-bp fragment shown, SNPs are discovered at nucleotides *x* and *y*. For the 613-bp fragment, a SNP is detected at nucleotide *z*.

13.2.4.2 **DOP-PCR**

More than a decade ago, degenerate oligonucleotide PCR (DOP-PCR) was developed as a means of amplifying (more or less) entire genomes. Using human DNA, Telenius et al. (1992) demonstrated that PCR primers with the sequence 5'-CCGACTCGAGNNNNNNATGTGG-3' (where *N* can represent any one of the four nucleotides) will bind to enough sites throughout the genome to allow amplification of most chromosomal regions. DOP-PCR has proven to be an extremely useful tool in genome studies in which DNA quantity is limited (e.g., Cheung and Nelson 1996; Dietmaier et al. 1999; Buchanan et al. 2000; Kittler et al. 2002).

Recently, Jordan et al. (2002) demonstrated the utility of DOP-PCR as a reduced representation technique for finding SNPs and other small polymorphisms in three different species (human, mouse, and *Arabidopsis*). In short, they produced a series of primers identical to the standard DOP-PCR primer (see above) except that each new primer had one to four additional nucleotides added to its 3' end. By increasing primer length, Jordan et al. effectively decreased the number of complementary sequences to which the primers were likely to bind, and thus decreased the number of PCR products generated in a particular reaction, i.e., reduced the number of loci amplified. To detect polymorphisms, the sequences of PCR products resulting from amplification using a specific primer or combination of primers were compared across multiple individuals.

13.2.4.3 **SSR Capture**

The first SSR enrichment protocols involved hybridizing "microsatellite-like"[1)] oligonucleotides to colony blots of either large- or small-insert genomic clones and identifying those clones that contain a microsatellite. While effective, these approaches are relatively cumbersome and expensive.

As an alternative to library screening, Ostrander et al. (1992) propagated a small-insert phagemid library in a *dut ung E. coli* strain. The *dut ung* genotype results in frequent substitution of dUTP for dTTP during DNA replication. Ostrander et al. then isolated circular single-stranded phagemid DNA from the bacteria using an M13 helper phage. A microsatellite-like primer and Taq polymerase then were used to generate second strands for those molecules containing a region complementary to the primer. Introduction of the DNA molecules into wild-type *E. coli* resulted in strong selection against single-stranded, dUTP-rich DNA and consequently, enrichment for double-stranded DNA containing microsatellites.

An additional means of isolating SSRs involves crosslinking microsatellite-like oligonucleotides to nylon and hybridizing the membrane with genomic DNA fragments. This effectively captures DNA sequences with regions complementary to the oligonucleotides (Edwards et al. 1996).

Currently, the most popular SSR enrichment techniques are those rooted in PCR-based primer extension (e.g., Fisher et al. 1996; Phan et al. 2000; Waldbieser

[1)] "Microsatellite-like" sequences are synthetic, single-stranded oligonucleotides that possess the characteristics of microsatellites and may be complementary in whole or part to naturally occurring SSRs.

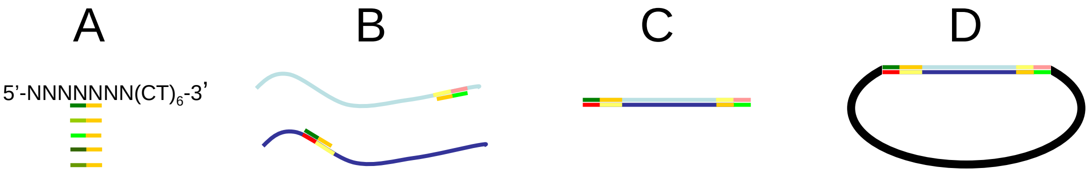

Figure 13.6. A primer extension method for SSR enrichment. The technique shown is similar to that of Fisher et al. (1996). (A) A series of primers are synthesized. Each primer possesses the same SSR-like repeat, i.e., $(CT)_6$, at its 3' end (gold line) and a partially degenerate seven-base sequence at its 5' end (green lines) that does not include any additional CT repeats. (B) The primers are used to PCR-amplify those genomic regions flanked by primer binding sites. The conditions of the PCR reaction are kept stringent to make sure that the 5' ends of the primers actually anneal with complementary sequences (red and pink lines) in the genomic DNA, thus preventing slippage of primers to the 3' ends of targeted microsatellite loci and subsequent loss of variation in repeat length. (C) Each resulting PCR product has an SSR (gold and yellow paired lines) near its ends. (D) PCR products are cloned and sequenced.

et al. 2003) and those that utilize streptavidin-coated beads to capture re-association hybrids between biotin-labeled, microsatellite-like oligonucleotides and genomic DNA fragments (e.g., Fischer and Bachmann 1998; Hamilton et al. 1999). Primer extension techniques require that (1) two SSRs be in close proximity to one another so that the region between them is amplified with SSR-based primers (e.g., Fisher et al. 1996; Figure 13.6), (2) an SSR is near a repeat sequence so that amplification can be achieved using an SSR-like primer and a repeat-based primer (e.g., Phan

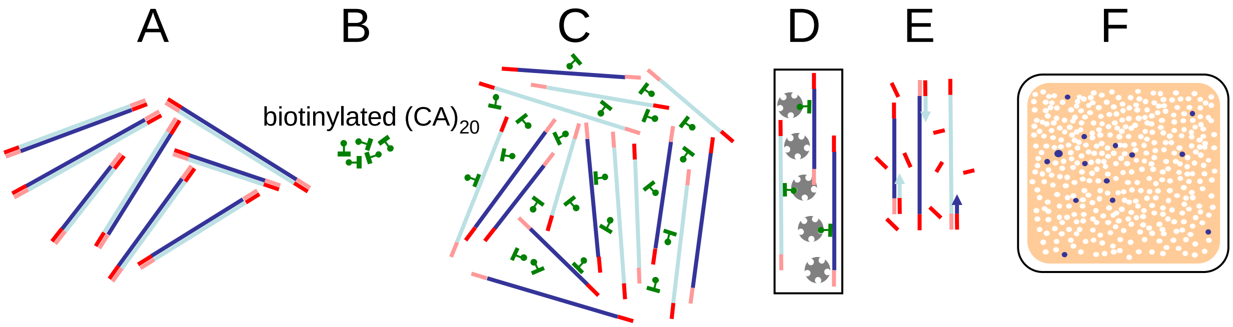

Figure 13.7. Hybrid capture method for SSR enrichment. (A) Genomic DNA is digested into fragments with restriction enzymes that produce blunt-ended cut sites, a linker is ligated to the fragments, and PCR using a primer complementary to one strand of the linker is used to amplify the DNA thus ensuring that the vast majority of molecules in the reaction have linkers on both ends. (B) One or more single-stranded biotinylated SSR-like sequences are synthesized. In the diagram a biotinylated $(CA)_{20}$ probe is shown. The ball on each $(CA)_{20}$ probe represents biotin. (C) The genomic DNA/linker molecules are denatured and allowed to re-anneal with an excess of the biotin-labeled probe(s). (D) The mixture is loaded onto a column containing streptavidin-coated beads and the column is thoroughly washed. Only those genomic DNA molecules hybridized to a biotin-labeled probe stick to the column. (In one variation, the streptavidin-labeled beads are magnetic and are removed from a slurry using a magnet). (E) The SSR-containing DNA is eluted from the column (by heating or chemical means) and amplified via PCR with the same primer used in step A. (F) The SSR-containing molecules are cloned and ultimately sequenced.

et al. 2000), or (3) genomic DNA be digested with a restriction enzyme and linkers with a primer binding site be attached to the ends, allowing PCR amplification using an SSR-like primer and a linker-based primer or two linker-based primers (i.e., Fischer and Bachmann 1998). The biotin-streptavidin "hybrid capture" approaches do not require that any particular sequence be near an SSR (Figure 13.7). Combinations of the primer extension and "hybrid capture" techniques have also been developed (e.g., Paetkau 1999).

13.3 Other Reduced Representation Techniques

In addition to the techniques listed above, there are several other reduced representation techniques that are applicable to a particular species or group of species but are likely to be of limited use to the plant genomics community as a whole. An example of this is the *RescueMu* technique which has been used to isolate genes in maize (see Raizada et al. 2001 and Raizada 2003 for reviews). In brief, *RescueMu* is a plasmid inserted into a maize *Mu1* transposon. The active *RescueMu* element preferentially inserts itself into gene regions in maize (70-90% of the time). Plasmid rescue then can be used to recover the 5-25 kb of DNA flanking the inserted element. However, *RescueMu* is currently limited to studying certain maize genotypes in which *Mu1* elements are active.

In some plants, certain class II transposons known as "miniature inverted-repeat transposable elements" (MITEs) appear to be preferentially associated with genes (see Feschotte et al. 2002 for review). Based on this observation, a modified AFLP procedure known as transposon display was used to amplify genomic regions containing a MITE known to show an insertion preference for genes (Casa et al. 2000). However, because MITE families can vary widely between species, not all MITEs are preferentially associated with genes, and some genomes have few MITEs (see Casacuberta and Santiago 2003 for review), it is unlikely that transposon display will be useful in sequencing gene space from most plant genomes.

13.4 Discussion

13.4.1 Repeat Sequence Enrichment

While repetitive sequences are often deemed "junk DNA," it is becoming increasingly clear that repeats (especially transposable elements) are one of the principal factors responsible for the evolutionary success of eukaryotes (see Britten 1996 and Wessler 1997). Repetitive DNA influences gene expression and recombination (Assaad et al. 1993; Dorer and Henikoff 1994; Sherman and Stack 1995), and some repeat sequences are involved in maintaining chromosome structure (Lee et al. 1994; Lundblad and Wright 1996). Consequently, studying repeat sequences is necessary to understanding plant genome evolution, gene function, and chromosome organization.

Of the various reduced representation techniques, CBCS is the only technique that allows efficient isolation and characterization of the repetitive sequences of genomes. Its utility as a repeat enrichment tool has been demonstrated for sorghum (Peterson et al. 2002a) and chicken (Wicker et al., submitted). Of note, CBCS can be used to generate a more accurate overview of the repeat content of a genome than sequencing a small number of large-insert clones. For example, at the time the sorghum CBCS research was conducted, there were three sorghum BAC sequences (AF061282, AF114171, and AF124045) in GenBank. Sequencing of sorghum Cot clones revealed that the most abundant repeat sequence in the sorghum genome is a previously unnamed retroelement found only once in the 424,434 bp of sequenced BAC DNA. The prevalence of this element, now known as *Retrosor-6*, was verified by BAC macroarray analysis, and it was shown to comprise roughly 6% of the sorghum genome (Peterson et al. 2002a). Since publication of the sorghum CBCS paper, the amount of sorghum sequence data in GenBank has increased dramatically. As expected, *Retrosor-6* is a common feature of many recent sorghum dbGSS entries, most notably *Sorghum propinquum* genomic clones.

While repetitive DNA is an enormous impediment to gene research in most major crops, knowledge of the sequences and distribution of repeats may circumvent many problems and, indeed, create new research opportunities. For instance, a better understanding of the physical distributions of repetitive DNA families at a resolution compatible with cloning technologies (such as over different BAC clones) may provide the means to identify "gene-rich" genomic domains that are priorities for early sequencing. Additionally, complete physical mapping of large genomes will benefit substantially from, and perhaps even require, a comprehensive knowledge of the sequences and distributions of repetitive DNA families. In this regard, HRCot and MRCot sequences from sorghum have facilitated physical mapping in this species (Peterson et al. 2002a). Identification of repetitive DNA is also valuable for masking repeats out of EST databases, significantly improving the quality of unigene sets.

The principal limitation of CBCS in the study of repeat sequences is that duplexes formed by strand reassociation are cloned, a feature that will result in occasional base mismatches and consequently make it difficult to detect small polymorphisms within repeat families. However, as mentioned above, we are working to adapt a second-strand synthesis technique for HR and MRCot library construction that may enhance the utility of CBCS.

13.4.2 Sequencing Gene space

Of the reduced representation techniques described above, EST sequencing, methylation filtration, and CBCS represent means of preferentially sequencing all or part of a plant's gene space. Toward this end, each of these techniques has its relative advantages and limitations.

13.4.2.1 EST sequencing

Compared to methylation filtration and CBCS, EST sequencing is the only reduced representation method that affords insight into both gene expression and gene space. This dual functionality makes EST sequencing a highly cost-effective genome research tool (hence the rapidly growing EST databases for many plant species; Rudd 2003). Additionally, because ESTs correspond to the exonic regions of genes, they are ideal molecular markers: each EST is not simply close to a gene but is part of a gene. However, EST sequencing has definite limitations as a gene space enrichment technique.

1. The mRNA molecules used in cDNA production have been post-transcriptionally modified; specifically, introns have been removed, a 5' m^{7}GpppN cap has been added to each transcript, and 3' poly-A tails have been appended. It is usually *via* their poly-A tails that mRNAs are isolated and from oligo(dT) primers that DNA strands are synthesized by reverse transcriptase. However, information corresponding to the 5' ends of transcripts may be lost due to the limited processivity of reverse transcriptase and/or the inhibition of enzyme movement by mRNA secondary structure (Edery et al. 1995).
2. Since intron removal precedes addition of poly-A tails, little or no information about introns is acquired by sequencing ESTs.
3. Reverse transcriptase is rather prone to mistakes, and consequently cDNA molecules are considerably more likely to contain errors than cloned genomic DNA (Menendez-Arias 2002). In fact, as many as 3% of nucleotides in reverse transcriptase-catalyzed strand synthesis reactions are likely to be incorrect (Rudd 2003).
4. Since processed mRNA is the starting material in cDNA library construction, promoter sequences, a crucial portion of gene space, are not found in EST libraries.
5. Plant tissues may be dominated by a few abundant transcripts. For example, a handful of cellular biogenesis genes account for > 40% of transcripts found in *Arabidopsis* pollen (Lee and Lee 2003). While the representation of dominant transcripts in an isolated mRNA population can be reduced using "normalization" strategies (e.g., Ko 1990; Soares et al. 1994; Neto et al. 1997; Zhulidov et al. 2004), gene copy number is not accurately reflected in cDNA libraries even if normalization techniques are employed.

6. Some transcripts are ephemeral lasting only a few minutes in cells (this is especially true of transcription factor mRNAs; e.g., Gee et al. 1991; Knauss et al. 2003; Yang et al. 2003), and others are simply transcribed at extremely low levels making their recovery difficult (Ito et al. 2003; Hsu et al. 2004).
7. The representation of genes in a given cDNA library is only indicative of gene expression in the source tissue(s) under specific environmental conditions. In order to have some confidence of acquiring the mRNA-encoding regions of every gene in a genome, one would theoretically have to obtain mRNA from every tissue exposed to every likely environmental condition/stress at every stage of development (Rudd 2003), as well as overcome the problems mentioned above.
8. Due to the limitations described (specifically points 4, 5, and 6), EST sequencing, even from libraries representing numerous developmental stages and tissues, reaches a point of diminishing returns at 50-70% transcript coverage. For example, although there are roughly 29,000 *Arabidopsis* genes as determined by genome sequencing (Arabidopsis Genome Initiative 2000), the 178,000 ESTs obtained from 61 different *Arabidopsis* tissues/stages/environments account for only 63% (i.e., 16,115) of these (Rudd 2003).

Despite its shortcomings, EST sequencing's dual ability to provide sequence information on the coding regions of probable genes and data on differential gene transcription has made it an invaluable tool in genome research. The value of ESTs as molecular markers and the incorporation of EST data into new techniques such as serial analysis of gene expression (SAGE), microarrays, transcriptome chips, etc., indicate that EST sequencing will continue to be an important tool for quite awhile.

13.4.2.2 **Methylation Filtration**

Methylation filtration is by far the simplest reduced representation technique, a fact that makes it highly appealing to those doing high-throughput genomics. It clearly enriches for gene regions in maize (Rabinowicz et al. 1999; Whitelaw et al. 2003) and several other species (www.oriongenomics.com), and it is possible that it would provide some level of gene enrichment/repeat reduction for many (most?) plant species. In this regard, a recent BLAST analysis (D. Peterson, unpublished results) indicates that of the 50,160 methylation filtered *Sorghum bicolor* sequences in GenBank (as of March 24, 2004), only 1.53% show significant (E value $< 1 \times 10^{-5}$) homology to *Retrosor-6*, a repetitive element that appears to account for about 6% of sorghum DNA (see above; Peterson et al. 2002a). Thus MF appears to reduce the level of *Retrosor-6* in sorghum DNA by roughly (768 ÷ 50,160 =) 3.9-fold.

Potential problems with using MF as a gene enrichment tool are evident by examination of the literature on DNA methylation in plants. First, it is well documented that some plant genes are normally hypermethylated and may become inactivated if hypomethylated. For example, (1) methylation of CGCG sites in several petunia genes is correlated with normal adventitious shoot bud induction, but both gene methylation and adventitious shoot budding are repressed by DNA methylase inhibitors (Prakash et al. 2003); (2) a 261-kb BAC from barley containing the powdery mildew resistance locus *Mla* has been shown to contain an extensive hypermethylated, but transcriptionally active, gene-rich island (Wei et al. 2002); (3) in at least seven ecotypes of *Arabidopsis*, the PAI1 and PAI4 genes involved in tryptophan biosynthesis are hypermethylated when active (Melquist et al. 1999; Bartee and Bender 2001); and (4) in vitro, somatic embryogenesis in carrot is blocked if cells are treated with the 5-azacytidine, an agent that causes DNA hypomethylation (LoSchiavo et al. 1989).

Second, there is considerable evidence that changes in methylation are a normal means by which plant genes, especially those involved in development or response to stress, are regulated (Finnegan et al. 2000). For example, (1) in dormant potatoes, large-scale, transient demethylation (50-70%) of 5'-CCGG-3' sequences precedes transcription of genes involved in cell division and meristem growth (Law and Suttle 2002); (2) meristematic regions in pine exhibit 35% DNA methylation in juvenile trees versus > 60% methylation in adult plants, but exposure of adult plants to reinvigoration stimuli causes a decrease in methylation to levels similar to those observed in juveniles (Fraga et al. 2002); (3) a rapid global decrease in DNA methylation occurs during seed germination and shoot apical meristem development in *Silene vulgaris* (Zluvova et al. 2001); (4) methylation appears to be involved in regulation of mRNA genes in numerous plant species (Drozdenyuk et al. 1976; Watson et al. 1987; Follman et al. 1990; Zluvova et al. 2001); (5) vernalization in *Arabidopsis* and tobacco appears to be triggered by demethylation of genes involved in the transition to flowering (Finnegan et al. 2000); (6) in maize, cold stress leads to genome-wide DNA demethylation in root tissues and subsequently to changes in transcription (Steward et al. 2002); (7) endosperm-specific demethylation and activation of specific alpha-tubulin alleles has been reported in maize (Lund et al. 1995); (8) the *P1-Blotched* and *P1-Rhoades* genes of maize show developmentally sensitive changes in methylation (Hoekenga et al. 2000); and (9) tissue-specific differences in DNA methylation have been observed in a number of plants including tomato (Messeguer et al. 1991) and rice (Xiong et al. 1999).

Finally, the hypermethylation of repeat sequences is by no means a constant or consistent characteristic of plant genomes. For example, (1) the tandem *Hae*III repeat in the grass *Pennisetum glaucum* is hypomethylated (Kamm et al. 1994); (2) when snapdragon is exposed to cold weather, methylation of the transposon Tam3 is reduced and transposition of the element increases (Hashida et al. 2003); (3) in maize, Robertson *Mutator* transposable elements undergo changes in methylation that coincide with changes in their expression (Singer et al. 2001); (4) within a genome, highly repetitive sequences can exhibit differential methylation patterns as demonstrated for the *Zingeria biebersteiniana* centromeric repeat Zbcen1 (Saunders and Houben 2001), the "*Alu*I" satellite repeats of snapdragon (Schmidt and Kudla 1996), and various high copy tobacco repeats (Kovarik et al. 2000); (5) middle repetitive DNA sequences from maize are found in both hypermethylated and hypomethylated DNA domains (Bennetzen et al. 1994); (6) the maize *suppressor-mutator* transposable element and a *Cucumis melo* satellite repeat exhibit tissue-specific differences in their methylation patterns (Banks and

Fedoroff 1989; Grisvard 1985); (7) in some instances (e.g., *Arabidopsis* centromeric regions), one strand of the DNA double helix may be hypermethylated compared to its complementary strand (i.e., there can be strand-biased methylation) (Luo and Preuss 2003).

The preceding observations indicate that methylation filtration will result in the loss of methylated genes whether currently active in the source tissue or inactive. Those genes involved in development and stress responses are particularly likely to be eliminated by MF. Additionally, certain repeat sequences will end up in methylation filtered DNA. In this regard, 33% of sequences in methylation filtered *Zea mays* libraries show significant homology to known repeats (Whitelaw et al. 2003).

Undoubtedly, some species are likely to be more amenable to methylation filtration than others. However, it is clear that before investing in MF, one should have a basic knowledge of the DNA methylation patterns found throughout the life cycle of their experimental organism.

13.4.2.3 **Cot-based Cloning And Sequencing (CBCS)**

In Cot-based cloning and sequencing (CBCS), repetitive and/or low-copy sequences are separated based upon their relative renaturation rates, which are reflective of their relative copy numbers in the genome. Such fractionation is completely independent of gene expression since the DNA used in re-association is mechanically-sheared genomic DNA. Likewise, sequence renaturation is independent of methylation patterns (Burtseva et al. 1979). Moreover, the separation of DNA sequences using Cot techniques is a well established biochemical phenomenon that can be applied to all species regardless of phylogeny (Peterson et al. 2002a, 2002b). Consequently, of the three gene-enrichment techniques, CBCS theoretically provides the most comprehensive and least biased gateway into the low-copy diversity of plant genomes.

CBCS is the most versatile of the reduced representation strategies as it can be used for enrichment of genes and/or repeats. It is the only reduced representation method that, in and of itself, could theoretically be used to sequence a genome's SqCx (see Peterson et al. 2002a, 2002b for review). In terms of gene enrichment, the parameters used in isolating the SL component can be adjusted to meet investigator goals and adapted for specific genomes (see Peterson et al. 2002a and Paterson et al. 2004 for reviews). For example, in those species in which genes are known to be grouped into "islands" (e.g., grasses), increasing the length of sequences used in constructing the SLCot library will decrease the probability that repetitive elements will elute with the SL component and result in longer clones more suitable for bidirectional sequencing. Additionally, allowing re-association to proceed to a higher Cot value will increase the stringency of SL fractionation and decrease potential repeat contamination. However, in making such changes, one runs the risk of "weeding out" short low-copy sequences near or flanked by repetitive elements. Consequently, sequencing clones from multiple SLCot libraries with different insert sizes may provide the greatest coverage of gene space per sequencing dollar (Paterson et al. 2004).

Compared to MF, CBCS appears to be a more stringent method of separating repetitive and low-copy DNA. While MF appears to result in a fourfold reduction in the level of *Retrosor-6* in sorghum DNA (see above), no SLCot clones in a 499-clone library showed homology to *Retrosor-6* (in contrast, 13.4 and 1.7% of HRCot and MRCot clones, respectively, contained *Retrosor-6* sequence). Likewise, a comparison of MF and SLCot libraries from maize indicates that more than twice as many repeat sequences are found in the former as the latter (Whitelaw et al. 2003).

As with all reduced representation techniques, CBCS has some limitations/drawbacks:

1. The DNA used in renaturation kinetics must be extremely clean as even low-level contamination from proteins, carbohydrates, and/or secondary compounds can cause serious problems (see Murray and Thompson 1976 for review). Contaminants such as polyphenols can inhibit renaturation by damaging DNA (Peterson et al. 1997). In contrast, carbohydrate contaminants may effectively decrease the area in which re-association takes place and thus artificially speed up renaturation rate. Some contaminants may absorb light at or near 260 nm and, if undetected, lead to aberrant results. Of note, many molecular biology protocols work well even if “dirty” DNA is used. However, such DNA is not suitable for Cot analysis and CBCS. Additionally, it is necessary that nuclear DNA, not total cellular DNA, is used in Cot analysis/fractionation, as significant organellar DNA contamination will complicate/confuse Cot analysis and decrease library quality. A protocol for isolating highly pure nuclear DNA suitable for Cot analysis and CBCS is available at the Mississippi Genome Exploration Laboratory (MGEL) website (www.msstate.edu/research/mgel/nucl_DNA.htm).
2. CBCS requires a fairly good understanding of re-association kinetics.
3. In a few plant species, there is a high level of non-genic low-copy DNA. For example, the tomato genome appears to have a considerable proportion of low-copy elements that are not genes (Peterson et al. 1998). For such species, CBCS may provide only marginal benefits compared to whole genome shotgun sequencing.
4. It has been speculated that SLCot sequencing may result in underrepresentation of members of large gene families (Martienssen et al. 2004), i.e., large gene families may fractionate with MR DNA. While this may be a problem if sequencing is performed using only one stringently prepared SLCot library, it is easily remedied by preparing several SLCot libraries with partially overlapping kinetic ranges.
5. Complementary regions may be found within the same single-stranded, low-copy sequence. Such molecules will “fold back” on themselves and form duplexes at Cot values approaching zero. While fold-back DNA is thought to be a minor component of genomes, low-copy fold-back sequences will be lost during SLCot library preparation. In species where fold-back DNA accounts for

> 3% of a genome, it may be advisable to clone and sequence the fold-back fraction (see Peterson et al. 2002b for further discussion).

Because a Cot analysis can be a relatively difficult, time-consuming process, the natural tendency of many researchers is to forgo actually constructing a Cot curve and skip right to re-association-based fractionation and cloning. If one's goal is solely to enrich for gene space, this may be a justifiable route, especially if it is clear that the experimental genome contains numerous repeats, the organism's genome size is well established, and the DNA used in re-association is known to be free of contamination. For example, Lamoreux et al. (submitted) recently prepared one SLCot library from bread wheat based on data from an existing wheat Cot curve and a second SLCot library based on an estimated Cot½ value for single-copy DNA as determined from genome size. The experimental Cot½ value of the SL component and the theoretical Cot½ value of single-copy DNA were fairly similar and, not surprisingly, the sequence contents of both Cot libraries were indistinguishable. To estimate the Cot½ of a species' single-copy component, the following formula can be used:

$$\mathrm{Cot½}_{\mathrm{org}} = (\mathrm{Cot½}_{\mathrm{coli}} \text{ x } G_{\mathrm{org}}) \div G_{\mathrm{coli}} \qquad (1)$$

where $\mathrm{Cot½}_{\mathrm{org}}$ is the estimated Cot½ of single-copy DNA for the organism of interest, G_{coli} is the genome size in base pairs of *E. coli*, G_{org} is the 1C DNA content of the organism of interest in base pairs, $\mathrm{Cot½}_{\mathrm{coli}}$ is the Cot½ of *E. coli* DNA. Inserting *E. coli*'s known genome size (4,639,221 bp; Blattner et al. 1997) and its Cot½ value (4.545455 M·sec; Zimmerman and Goldberg 1977) yields the following:

$$\mathrm{Cot½}_{\mathrm{org}} = (4.545455 \text{ M·sec x } G_{\mathrm{org}}) \div 4{,}639{,}221 \text{ bp} \qquad (2)$$

Placing a known genome size for an organism into Eq. (2) and solving for $\mathrm{Cot½}_{\mathrm{org}}$ provides the Cot½ for a theoretical single-copy component. From the predicted Cot½, one can decide upon a Cot value which will provide high likelihood that most low-copy elements will be isolated (e.g., Figure 13.4).

13.4.2.4 Integration of Reduced Representation Strategies

Recently, Whitelaw et al. (2003) compared the sequence content of SLCot, methyl-filtered, and unfiltered libraries from maize. Both MF and SLCot libraries showed enrichment of expressed sequences (27 and 22% of clones recognizing ESTs, respectively) compared to the unfiltered library (6% of sequences exhibited homology to ESTs). With regard to repetitive sequences, 14% of the SLCot sequences possessed significant homology to known repetitive elements, while 33% of MF sequences recognized repeats. Compared to the MF library, the SLCot library contained a larger proportion of sequences with no significant matches to any database sequences (63% vs. 39%), which partly reflects the higher repeat content of the MF library and may also reflect the ability of CBCS to capture elements (e.g., short-lived transcription factors, other low-copy sequences) that may elude EST and MF approaches. Perhaps the most significant finding of the study was that when SLCot and MF sequences were grouped into contigs, 60% of the MF clones assembled only with other MF clones, while 72% of SLCot clones assembled only with other SLCot sequences. This suggests that at least in maize, SLCot sequencing and MF may be enriching for largely different low-copy sequence subsets; alternatively, one technique may be enriching for a greater diversity of low-copy sequences than the other. Consequently, Whitelaw et al. suggest that with regard to maize and possibly other large genomes, a combination of MF and SLCot sequencing may provide better results than using one strategy alone.

13.4.3 Polymorphism Discovery

13.4.3.1 RRS Sequencing and DOP-PCR

The vast majority of polymorphism discovery techniques require a priori knowledge of sequences unique to a particular locus (or loci). Two notable exceptions are RRS sequencing and DOP-PCR. Both approaches enrich for a subset of loci more or less at random. If the DNA from many individuals is pooled, polymorphisms can be discovered by sequencing the isolated sequence subset and using computational algorithms to align probable alleles.

RRS sequencing and DOP-PCR have their relative advantages and disadvantages. Both techniques permit SNP and indel discovery. However, neither technique guarantees that polymorphic regions will be found in a sequence subset. Likewise, repetitive sequences are as likely to be sequenced as non-repetitive sequences; repeats are removed from consideration *in silico*. Of the two methods, RRS sequencing is certainly the most straightforward and probably the least expensive, as it does not require primer synthesis and is rooted in techniques common to most molecular biologists. DOP-PCR involves more experimental steps than RRS sequencing (compare methods of Jordan et al. 2002 and Altschuler et al. 2000). However, a typical DOP-PCR experiment will likely provide information about a larger subset of loci than an RRS sequencing experiment. As both techniques are known to be effective, it is likely that an investigator's preference of one over the other may reflect his/her familiarity (or lack thereof) with primer design and PCR.

13.4.3.2 Microsatellite Isolation

There are currently numerous approaches for de novo isolation of microsatellites. However, all the protocols share one thing in common: they rely upon the use of synthetic microsatellite-like oligonucleotides to "find" SSRs in genomic DNA. Of the microsatellite isolation protocols, the "nylon crosslinking" technique (i.e., crosslinking microsatellite-like oligonucleotides to a nylon membrane and incubating the membrane with genomic DNA; e.g., Edwards et al. 1996) has the greatest potential for capturing all potential SSRs, as every possible SSR oligonucleotide can easily fit on a relatively small nylon membrane. However, this approach is not widely used, reportedly because it does not provide satisfactory results for some species and/or some practitioners (Fischer and Bachmann 1998). The elegant approach of Ostrander et al. (1992) is applicable to many organisms, but it requires use of two bacterial strains and a phage intermediate and thus is less popular than simpler SSR-capture techniques. In the "primer extension" protocols, microsatellite-like sequences form part of primer sequences used in PCR (e.g., Figure 13.6). The primer extension protocols can be very fruitful and can be adjusted to meet the needs of most researchers. However, successful application of these protocols requires a fair amount of skill in primer design and some reaction optimization. Likewise, in order to amplify an SSR locus, two primer binding sites are necessary (see above). In the widely used "hybrid capture" protocols, biotinylated microsatellite-like oligonucleotides are hybridized to genomic DNA, and hybrid molecules are isolated via streptavidin-coated beads (e.g., Figure 13.7). The hybrid capture technique requires only modest PCR skills, and there is no requirement that microsatellites be near each other or near any other element. Thus the hybrid capture techniques will likely supplant primer extension techniques for those interested in straightforward SSR isolation. The primer extension techniques, however, enjoy the advantage of being easily adapted to certain genomes and/or highly-specific needs, and thus they will likely continue to be utilized as well.

13.5 Conclusions

Reduced representation strategies provide a means of exploring large, repetitive plant genomes in a cost-efficient manner. Specifically, they allow study of important sequence subsets that otherwise could be obtained only by whole-genome shotgun sequencing. The reduced representation techniques that can be employed in a given situation depend largely on the goals of the scientist and the biology of the experimental organism. EST sequencing and CBCS should be useful in capturing gene space of all seed plants, while MF will likely provide some level of gene enrichment for many plant species and high levels for others. CBCS is currently the only reduced representation technique that allows preferential study of repeat sequences. DOP-PCR, RRS sequencing, and microsatellite enrichment tools afford access to polymorphisms for mapping, molecular breeding, and characterizing genotypic/phenotypic relationships.

As the focus of genomics becomes centered less on small-genome model organisms and more on economically and socially important species, the utilization of existing reduced representation techniques and the demand for new reduced representation strategies will undoubtedly increase. Additionally, the large size of most crop genomes and the smaller pool of funding sources for plant research compared to vertebrate research make it likely that plant biologists will continue to be among the most avid users and developers of reduced representation techniques.

Acknowledgements

Thanks to Andrew H. Paterson for advice and the U.S. Department of Agriculture (USDA-NRI Award 9901330), the National Science Foundation (NSF-DBI Award 0115903), and the USDA Forest Service (Award SRS-03-CA-11330126-263) for financial support. This chapter has been approved for publication as Journal Article No. BC10511 of the Mississippi Agriculture and Forestry Experiment Station, Mississippi State University.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252:1651-1656.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513-516.

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

Assaad FF, Tucker KL, Signer ER (1993) Epigenetic repeat-induced gene silencing (RIGS) in *Arabidopsis*. *Plant Mol. Biol.* 22:1067-1085.

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

Banks JA, Federoff N (1989) Patterns of developmental and heritable change in methylation of the *suppressor-mutator* transposable element. *Dev. Genet.* 10:425-437.

Bartee L, Bender J (2001) Two *Arabidopsis* methylation-deficiency mutations confer only partial effects on a methylated endogenous gene family. *Nucleic Acids Res.* 29:2127-2134.

Bennett MD, Leitch IJ (2003) Plant DNA C-values database (release 2.0, Jan. 2003). http://www.rbgkew.org.uk/cval/homepage.html.

Bennetzen JL, Schrick K, Springer PS, Brown WE, SanMiguel P (1994) Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. *Genome* 37:565-576.

Berger F (2004) Imprinting - a green variation. *Science* 303:483-485.

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register III JC, Tingey SV, Rafalski A (2002) Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Mol. Biol.* 48:539-547.

Blattner FR, Plunkett III G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453-1474.

Britten RJ (1996) Cases of ancient mobile element DNA insertions that now affect gene regulation. *Mol. Phylogenet. Evol.* 5:13-17.

Britten RJ, Graham DE, Neufeld BR (1974) Analysis of repeating DNA sequences by reassociation. *Methods Enzymol.* 29:363-405.

Britten RJ, Kohne DE (1968) Repeated sequences in DNA. *Science* 161:529-540.

Buchanan AV, Risch GM, Robichaux M, Sherry ST, Batzer MA, Weiss KM (2000) Long DOP-PCR of rare archival anthropological samples. *Hum. Biol.* 72:911-925.

Burtseva NN, Romanov GA, Azizov YuM, Banyshin BF (1979) Intragenome distribution of 5-methylcytosine and kinetics of the reassociation of cow blood lymphocyte DNA in the normal state and in chronic lympholeukemia. *Biochemistry-Moscow* 44: 1636-1641.

Casa AM, Brouwer C, Nagel A, Wang L, Zhang Q, Kresovich S, Wessler SR (2000) The MITE family *Heartbreaker* (Hbr): Molecular markers in maize. *Proc. Natl. Acad. Sci. USA* 97: 10083-10089.

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene* 311:1-11.

Cheung VG, Nelson SF (1996) Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc. Natl. Acad. Sci. USA* 93:14676-14679.

Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews* 3: 22-32.

Dietmaier W, Hartmann A, Wallinger S, Heinmöller E, Kerner T, Endl E, Jauch K-W, Hofstädter F, Rüschoff J (1999) Multiple mutation analyses in single tumor cells with improved whole genome amplification. *Amer. J. Pathol.* 154:83-95.

Dorer DR, Henikoff S (1994) Expansions of transgene repeats cause heterochromatin formation and gene silencing in *Drosophila* . *Cell* 77:993-1002.

Drozdenyuk AP, Sulimova GE, Vanyushin BI (1976) Changes in base composition and molecular population of wheat DNA on germination. *Mol. Biol. (Moscow)* 10:1378-1386.

Edery I, Chu LL, Sonenberg N, Pelletier J (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell Biol.* 15:3363-3371.

Edwards KJ, Barkder JHA, Daly A, Jones C, Karp A (1996) Microsatellite libraries enriched for several microsatellite sequences in plants. *BioTechniques* 20:758-760.

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: Where genetics meets genomics. *Nat. Rev. Genet.* 3:329-341.

Finnegan EJ, Genger RK, Peacock WJ, Dennis ES (1998) DNA methylation in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49:223-247.

Finnegan EJ, Peacock WJ, Dennis ES (2000) DNA methylation, a key regulator of plant development and other processes. *Curr. Opin. Genet. Dev.* 10:217-223.

Fischer D, Bachmann K (1998) Microsatellite enrichment in organisms with large genomes (*Allium cepa* L.). *BioTechniques* 24:796-802.

Fisher PJ, Gardner RC, Richardson TE (1996) Single locus microsatellites isolated using 5' anchored PCR. *Nucleic Acids Res.* 24:4369-4371.

Fojtova M, Van Houdt H, Depicker A, Kovarik A (2003) Epigenetic switch from posttranscriptional to transcriptional silencing is correlated with promoter hypermethylation. *Plant Physiol.* 133:1240-1250.

Follmann H, Balzer HJ, Schleicher R (1990) Biosynthesis and distribution of methylcytosine in wheat DNA. How different are plant DNA methyltransferases? In: *Nucleic Acid Methylation*. AR Liss Inc., New York: pp 199-209.

Fraga MF, Rodriguez R, Canal MJ (2002) Genomic DNA methylation-demethylation during aging and reinvigoration of *Pinus radiata*. *Tree Physiol.* 22:813-816.

Fu H, Zheng Z, Dooner HK (2002) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl. Acad. Sci. USA* 99:1082-1087.

Gee MA, Hagen G, Guilfoyle TJ (1991) Tissue-specific and organ-specific expression of soybean auxin-responsive transcripts GH3 and SAURs. *Plant Cell* 3:419-430.

Goldberg RB (1978) DNA sequence organization in the soybean plant. *Biochemical Genetics* 16:45-68.

Goldberg RB (2001) From Cot curves to genomics. How gene cloning established new concepts in plant biology. *Plant Physiol.* 125: 4-8.

Grisvard J (1985) Different methylation pattern of melon satellite DNA sequences in hypocotyl and callus tissues. *Plant Sci.* 39:189-193.

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genomics* 270:315-323.

Hamilton MB, Pincus EL, Di Fiore A, Fleischer RC (1999) Universal linker and ligation procedures for construction of genomic DNA libraries enriched for microsatellites. *BioTechniques* 27:500-507.

Hartl DL (2000) Molecular melodies in high and low *C. Nat. Rev.* 1:145-149.

Hashida S, Kitamura K, Mikami T, Kishima Y (2003) Temperature shift coordinately changes the activity and the methylation state of transposon Tam3 in *Antirrhinum majus*. *Plant Physiol.* 132:1207-1216.

Hoekenga OA, Muszynski MG, Cone KC (2000) Developmental patterns of chromatin structure and DNA methylation responsible for epigenetic expression of a maize regulatory gene. *Genetics* 155:1889-1902.

Hsu HC, Tan LY, Au LC, Lee YM, Lieu CH, Tsai WH, You JY, Liu MD, Ho CK (2004) Detection of bcr-abl gene expression at a low level in blood cells of some patients with essential thrombocythemia. *J. Lab. Clin. Med.* 143: 125-129.

Ito T, Sakai H, Meyerowitz EM (2004) Whorl-specific expression of the *SUPERMAN* gene of *Arabidopsis* is mediated by *cis* elements in the transcribed region. *Curr. Biol.* 13: 1524-1530.

Jordan B, Charest A, Dowd JF, Blumenstiel JP, Yeh RF, Osman A, Housman DE, Landers JE (2002) Genome complexity reduction for SNP genotyping analysis. *Proc. Natl. Acad. Sci. USA* 99:2942-2947.

Kamm A, Schmidt T, Heslop-Harrison JS (1994) Molecular and physical organization of highly repetitive, undermethylated DNA from *Pennisetum glaucum*. *Mol. Gen. Genet.* 244:420-425.

Kinoshita T, Muura A, Choi Y, Kinoshita Y, Cao X, Jacobsen SE, Fischer RL, Kakutani T (2004) One-way control of *FWA* imprinting in *Arabidopsis* endosperm by DNA methylation. *Science* 303:521-522.

Kiper M, Herzfeld F (1978) DNA sequence organization in the genome of *Petroselinum sativum* (Umbelliferae). *Chromosoma* 65:335-351.

Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff RR (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 100:7383-7388.

Kittler R, Stoneking M, Kayser M (2002) A whole-genome amplification method to generate long fragments from low quantities of genomic DNA. *Anal. Biochem.* 300:237-244.

Knauss S, Rohrmeier T, Lehle L (2003) The auxin-induced maize gene *ZmSAUR2* encodes a short-lived nuclear protein expressed in elongating tissues. *J. Biol. Chem.* 278: 23936-23943.

Ko MSH (1990) An “equalized cDNA library” by reassociation of short double-stranded cDNAs. *Nucleic Acids Res.* 18:5705-5711.

Komulainen P, Brown GR, Mikkonen M, Karhu A, García-Gil R, O'Malley D, Lee B, Neale DB, Savolainen O (2003) Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda*. *Theor. Appl. Genet.* 107:667-678.

Kovalchuk O, Burke P, Arkhipov A, Kuchma N, James SJ, Kovalchuk I, Pogribny I (2003) Genome hypermethylation in *Pinus silvestris* of Chernobyl- a mechanism for radiation adaptation? *Mutation Res.* 529:13-20.

Kovarik A, Koukalova B, Lim KY, Matyasek R, Lichtenstein CP, Leitch AR, Bezdek M (2000) Comparative analysis of DNA methylation in tobacco heterochromatic sequences. *Chromosome Res.* 8:527-541.

Kumar S, Cheng X, Klimasauskas S, Mi S, Posfai J, Roberts RJ, Wilson GG (1994) The DNA (cytosine-5) methyltransferases. *Nucleic Acids Res.* 22:1-10.

Lapitan NLV (1992) Organization and evolution of higher plant nuclear genomes. *Genome* 35: 171-181.

Law DR, Suttle JC (2001) Transient decreases in methylation at 5'-CCGG-3' sequences in potato (*Solanum tuberosum* L.) meristem DNA during progression of tubers through dormancy precede the resumption of sprout growth. *Plant Mol. Biol.* 51:437-447.

Lee C, Ritchie DBC, Lin CC (1994) A tandemly repetitive, centromeric DNA sequence from the Canadian woodland caribou (*Rangifer tarandus caribou*): its conservation and evolution in several deer species. *Chromosome Res.* 2:293-306.

Lee J-Y, Lee D-H (2003) Use of serial analysis of gene expression technology to reveal changes in gene expression in Arabidopsis pollen undergoing cold stress. *Plant Physiol.* 132: 517-529.

Lisch D, Carey CC, Dorweiler JE, Chandler VL (2002) A mutation that prevents paramutation in maize also reverses *Mutator* transposon methylation and silencing. *Proc. Natl. Acad. Sci. USA* 99:6130-6135.

Liu B, Wendel JF (2003) Epigenetic phenomena and the evolution of plant allopolyploids. *Mol. Phylogenet. Evol.* 29:365-379.

LoSchiavo F, Pitto L, Giuliano G, Torti G, Nuti-Ronchi V, Marazziti D, Vergara R, Orselli S, Terzi M (1989) DNA methylation of embryonic carrot cell cultures and its variations as caused by mutation, differentiation, hormones, and hypomethylating drugs. *Theor. Appl. Genet.* 77:325-331.

Lund G, Messing J, Viotti A (1995) Endosperm-specific demethylation and activation of specific alleles of alpha-tubulin genes of *Zea mays* L. *Mol. Gen. Genet.* 246:716-722.

Lundblad V, Wright WE (1996) Telomeres and telomerase: A simple picture becomes complex. *Cell* 87:369-375.

Luo S, Preuss D (2003) Strand-biased DNA methylation associated with centromeric regions in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 100:11133-11138.

Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR (2004) Sequencing the maize genome. *Curr. Opin. Plant Biol.* 7: 102-107.

Melquist S, Luff B, Bender J (1999) Arabidopsis *PAI* gene arrangements, cytosine methylation and expression. *Genetics* 153:413.

Menendez-Arias L (2002) Molecular basis of fidelity of DNA synthesis and nucleotide specificity of retroviral reverse transcriptases. *Prog. Nucleic Acid Res. Mol. Biol.* 71: 91-147.

Meng L, Bregitzer P, Zhang S, Lemaux PG (2003) Methylation of the exon/intron region in the *Ubi1* promoter complex correlates with transgene silencing in barley. *Plant Mol. Biol.* 53:327-340.

Messeguer R, Ganal M, deVincente MC (1991) High resolution RFLP map around the root knot nematode resistance gene (Mi) in tomato. *Theor. Appl. Genet.* 82:529-536.

Murray MG, Thompson WF (1976) Contaminants affecting plant DNA reassociation. *Carnegie Inst. Wash. Yearbook* 76:255-259.

Neto ED, Harrop R, Correa-Oliveira R, Wilson RA, Pena SDJ, Simpson AJG (1997) Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the generation of EST's from nanogram quantities of mRNA. *Gene* 186:135-142.

Ostrander EA, Jong PM, Rine J, Duyk G (1992) Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. *Proc. Natl. Acad. Sci. USA* 89:3419-3423.

Paetkau D (1999) Microsatellites obtained using strand extension: an enrichment protocol. *BioTechniques* 26:690-697.

Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang C-X, Katsar CS, Lan T-H, Lin Y-R, Ming R, Wright RJ (2000) Comparative genomics of plant chromosomes. *Plant Cell* 12:1523-1539.

Paterson AH, Bowers JE, Chapman BA, Peterson DG, Rong JK, Wicker TM (2004) Comparative genome analysis of monocot and dicots, toward characterization of angiosperm diversity. *Curr. Opin. Biotech.* (in press).

Paterson AH, Schertz KF, Lin Y-R, Liu S-C, Chang Y-L (1995) The weediness of wild plants: Molecular analysis of genes influencing dispersal and persistance of johnsongrass, *Sorghum halepense* (L.) Pers. *Proc. Natl. Acad. Sci. USA* 92:6127-6131.

Peterson DG, Boehm KS, Stack S (1997) Isolation of milligram quantities of DNA from tomato (*Lycopersicon esculentum*), a plant containing high levels of polyphenolic compounds. *Plant Mol. Biol. Reptr.* 15:148-153.

Peterson DG, Pearson WR, Stack SM (1998) Characterization of the tomato (*Lycopersicon esculentum*) genome using *in vitro* and *in situ* DNA reassociation. *Genome* 41:346-356.

Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* 12:795-807.

Peterson DG, Wessler SR, Paterson AH (2002) Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.* 18:547-550.

Petrov DA (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet.* 17:23-28.

Phan J, Reue K, Peterfy M (2000) MS-IRS PCR: a simple method for the isolation of microsatellites. *BioTechniques* 28:18-20.

Prakash AP, Kush A, Lakshmanan P, Kumar PP (2003) Cytosine methylation occurs in a CDC48 homologue and a MADS-box gene during adventitious shoot induction in *Petunia* leaf explants. *J. Exp. Bot.* 54:1361-1371.

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genet.* 23:305-308.

Raizada MN (2003) *RescueMu* protocols for maize functional genomics. *Methods Mol. Biol.* 236:37-58.

Raizada MN, Nan G-L, Walbot V (2001) Somatic and germinal mobility of the *RescueMu* transposon in transgenic maize. *Plant Cell* 13:1587-1608.

Redaschi N, Bickle TA (1996) DNA restriction and modification systems. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Edited by: Neidhardt FC. *ASM Press, Washington, D.C.*: pp 773-781.

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci.* 8:321-329.

SanMiguel P, Bennetzen JL (2000) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* 82: 37-44.

Saunders VA, Houben A (2001) The pericentromeric heterochromatin of the grass *Zingeria biebersteiniana* (2n=4) is composed of Zbcen1-type tandem repeats that are intermingled with accumulated dispersedly organized sequences. *Genome* 44:955-961.

Schmid KJ, Sörensen TR, Stracke R, Törjék O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* 13:1250-1257.

Schmidt T, Kudla J (1996) The molecular structure, chromosomal organization, and interspecies distribution of a family of tandemly repeated DNA sequences of *Antirrhinum majus* L. *Genome* 39:243-248.

Shagin DA, Rebrikov DV, Kozhemyako VB, Altshuler IM, Shcheglov AS, Zhulidov PA, Bogdanova EA, Staroverov DB, Rasskazov VA, Lukyanov S (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Res.* 12:1935-1942.

Sherman JD, Stack SM (1995) Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). *Genetics* 141: 683-708.

Singer T, Yordan C, Martienssen RA (2001) Robertson's *Mutator* tranposons in *A. thaliana* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation (DDM1)*. *Genes & Dev.* 15:602.

Soares MB, Bonaldo MDF, Jelene P, Su L, Lawton L, Efstratiadis A (1994) Construction and characterization of the normalized cDNA library. *Proc. Natl. Acad. Sci. USA* 91:9228-9232.

Steward N, Ito M, Yamaguchi Y, Koizumi N, Sano H (2002) Periodic DNA methylation in maize nucleosomes and demethylation by environmental stress. *J. Biol. Chem.* 277: 37741-37746.

Telenius H, Carter NP, Bebb CE, Nordenskjold M, Ponder BA, Tunnacliffe A (1992) Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* 13:718-725.

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

Törjék O, Berger D, Meyer RC, Müssig C, Schmid KJ, Sörensen TR, Weisshaar B, Mitchell-Olds T, Altmann T (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. *Plant J.* 36:122-140.

Waldbieser GC, Quiniow SMA, Karsi A (2003) Rapid development of gene-tagged microsatellite markers from bacterial artificial chromosome clones using anchored TAA repeat primers. *BioTechniques* 35:976-979.

Watson JC, Kaufman LS, Thompson WF (1987) Developmental regulation of cytosine methylation in the nuclear ribosomal RNA genes of *Pisum sativum*. *J. Mol. Biol.* 193: 15-26.

Wei F, Wing RA, Wise RP (2002) Genome dynamics and evolution of the *Mla* (powdery mildew) resistance locus in barley. *Plant Cell* 14:1903-1917.

Wessler SR (1997) Transposable elements and the evolution of gene expression. *Exp. Biol.* 1039:115-122.

Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, van Heeringen S, Karamycheva S, Bennetzen JL, SanMiguel P, Lakey N, Bedford J, Yuan Y, Budiman MA, Resnick A, van Aken S, Utterback T, Riedmuller S, Williams SM, Feldblyum T, Schubert K, Beachy R, Fraser CM, Quackenbush J (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302: 2118-2120.

Xiong LZ, Xu CG, Saghai Maroof MA, Zhang Q (1999) Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive amplification polymorphism technique. *Mol. Gen. Genet.* 261:439-446.

Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell Jr JE (2003) Decay rates of human mRNAs: Correlation with functional characteristics and sequence attributes. *Genome Res.* 13:1863-1872.

Yuan Y, SanMiguel PJ, Bennetzen JL (2003) High-Cot sequence analysis of the maize genome. *Plant J.* 34:249-255.

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch W, Moroz LL, Lukyanov SA, Shagin DA (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32:e37.

Zimmerman JL, Goldberg RB (1977) DNA sequence organization in the genome of *Nicotiana tabacum*. *Chromosoma* 59: 227-252.

Zluvova J, Janousek B, Vyskot B (2001) Immunohistochemical study of DNA methylation dynamics during plant development. *J. .Exp. Botany* 52: 2235-2273.