Plant Genomic Sequencing Using Gene-Enriched Libraries

Pablo D. Rabinowicz*

J. C. Venter Institute, 9712 Medical Center Drive, Rockville, Maryland 20850

Received January 2, 2007

Contents

1. Introduction

Plant and animal genomes share many common features, such as the presence of introns and, in many cases, a large proportion of repetitive elements. However, there are significant differences between higher plant genomes and those of higher animals, particularly mammals. For example, while the size of mammalian genomes ranges approximately between 2.5 and 3 Gigabase pairs (Gbp),[1−6] plant genomes can vary by several orders of magnitude.[7,8] For instance, the genome of *Selaginella*, an early divergent vascular plant, is estimated to be around 0.12 Gbp,[9] and the genomes of some lilies may reach over 100 Gbp. The number of genes in these genomes, however, does not vary proportionally to their size.[10] Rather, a large amount of repetitive DNA accounts for most of the genome size differences. Another factor affecting genome size in plants is polyploidy. It has been estimated that over 70% of angiosperm species have undergone one or more cycles of polyploidization.[11] Over evolutionary time polyploids may go through a process of diploidization in which duplicated genes tend to be eliminated unless they acquire new functions. In the ancient tetraploid genome of maize, for example, one of the two members of the homoeologous gene pairs, has been lost in about one-half the cases studied,[12−14] and diploidization of low-copy sequences as well as ribosomal RNA genes has also been observed in the polyploid soybean genome.[15,16]

Genome function and evolution can be best studied if a genome sequence is available, and during the last 15 years huge efforts have been devoted to elucidating the sequence of several genomes, shedding light on many fundamental biological processes. Driven by an interest in curing and diagnosing human diseases, a number of initiatives have been put forward to encourage development of faster and cheaper genomic sequencing methodologies,[17−19] which are equally applicable to plant genomes.[20] Consequently, new sequencing technologies with increased throughput and reduced costs have emerged, but there are still many hurdles to overcome before they can replace the widely used fluorescence capillary electrophoresis-based Sanger sequencing method (http://www.appliedbiosystems.com). Some of these new methodologies, such as highly parallel pyrosequencing (454 sequencing)[21] and massively parallel sequencing by synthesis (Solexa's Clonal Single Molecule Array technology; http://www.solexa.com/technology/sbs.html), can deliver large amounts of DNA sequence data in a short period of time, and the cost per base is reduced compared to Sanger sequencing. These and other emerging sequencing technologies,[22] however, have a major limitation in the short length (∼30−250 bp) of each individual sequence read, making data unsuitable for large-scale assembly. When a large and repetitive genome is broken into pieces for sequencing the 800 bp reads produced by typical capillary electrophoresis sequencing allow assembly of the original genomic sequence with much higher accuracy. For certain applications, such as sequencing genomes closely related to previously se-

Pablo Rabinowicz was born in Buenos Aires, Argentina. He obtained his B.S./M.S. degrees in Biology at the University of Buenos Aires in 1990, and in 1996 he received his Ph.D. degree in Molecular Biology of Plant Viruses from the same university. Afterward, he moved to the Cold Spring Harbor Lab (CSHL) in New York for postdoctoral training in plant genetics and genomics. At CSHL he co-developed the methylation filtration technique in Rob Martienssen's lab. Since 2004 he has had a Plant Genomics Faculty position at The Institute for Genomic Research (TIGR, now JCVI), where he is working on maize, wheat, and other grasses' genomes as well as castor bean and cassava comparative genomics, among other projects.

* To whom correspondence should be addressed. Phone: (301) 795-7787. Fax: (301) 838-0208. E-mail: pablo@jcvi.org.

10.1021/cr0682960 CCC: $65.00
Published on Web 06/27/2007

quenced ones (resequencing), shorter reads may not pose such a significant problem and the production speed and other advantages of the new technologies make them very promising.

Because 454 sequencing technology does not involve cloning DNA in *E. coli*, strategies that combine classical Sanger sequencing with 454 technology have proven to be an efficient approach to sequencing genomes that are difficult to clone due to sequence composition.[23] Nevertheless, Sanger sequencing continues to be the method of choice for sequencing large genomes such as mammalian ones.[2−6] Many animal genomes are being or will be sequenced using this technology,[24−26] and three plant genomes have also been sequenced at high levels of accuracy.[27−29] Several more plant genomes have been sequenced reaching lower quality levels or are in progress, though at a slower pace than their animal counterparts. Some examples of ongoing plant genome sequencing projects include crops such as maize, sorghum, tomato, potato, castor bean, and peach, emerging model plants such as *Brachypodium*, nonvascular plants such as the moss *Physcomitrella patens*, ancient vascular plants such as *Selaginella moellendorffii*, and flowering plants relevant for comparative and evolutionary studies such as columbine and close relatives of *Arabidopsis* (http://www.jgi.doe.gov/sequencing/why/index.html, http://www.sgn.cornell.edu/about/tomato_sequencing.pl, http://www.potatogenome.net/index.htm; http://castorbean.tigr.org, http://www.maizesequence.org). The slower rate at which new plant genomes are sequenced is due to not only a lower level of funding that plant genomic research receives relative to animal and human research but also the fact that many important plant genomes are extremely large and contain a high proportion of conserved repetitive elements.[30]

Two whole genome sequencing strategies have been commonly used for both animal and plant genomes. One is the whole genome shotgun approach (WGS),[31] where the ends of random genomic clones are sequenced at large scale. The second strategy is the bacterial artificial chromosome[32] (BAC)-based approach in which selected large-insert genomic clones are completely sequenced. Both of these strategies can yield the sequence of nearly an entire genome, which in the case of large plant genomes consists mostly of repetitive elements. Except in the cases when they affect expression of nearby genes, repetitive elements are constituted of "parasitic" DNA with the only function of self-propagation. Therefore, their repeated sequence contains little information relative to the amount of data, and sequencing approaches that capture the exonic and/or entire genic regions avoiding the repetitive DNA are fast and affordable alternatives to whole genomic sequencing. Such technologies are generally called gene-enrichment techniques. This review starts with a general introduction on plant whole genome sequencing strategies as a prelude to discuss gene-enrichment techniques for large and highly repetitive plant genomes and how these techniques compare and may synergize with the traditional whole genome sequencing methods.

2. BAC-Based Plant Genomic Sequencing

The genome of the model plant *Arabidopsis thaliana* was the first plant genome to be sequenced, and the project was carried out by an international consortium[28] using a BAC-based approach. The resulting product shows a high degree of accuracy and completeness. In a BAC-based strategy, one or more BAC libraries are constructed with an average insert size typically between 100 and 150 kilobase pairs (kbp). These libraries must consist of enough clones to represent 10−20 genome equivalents.[33] A subset of the BAC clones in the libraries that span the whole genome with minimal overlaps at the ends is selected for sequencing.

Two different methods are used to select this minimal set of BAC clones. One method identifies certain BACs as "seed clones" to be completely sequenced. In addition, the ends of all the clones in the library are sequenced. Comparison of the BAC ends and the seed clone sequences allows identification of one minimally overlapping clone at each end. These clones are then completely sequenced, and a new alignment to the BAC-end sequences is performed to identify new minimally overlapping clones at the distal ends. This process is iterated so that each chromosome can be completely sequenced.[34] Prior knowledge of BAC clones that are distributed throughout the genome accelerates the progress toward complete coverage of the genome. This information can be obtained by hybridization of all clones in the BAC library against molecular markers that are scattered around the genome as determined by their location in the genetic map. This method, known as "map as you go", requires identification of evenly separated seed clones prior to starting sequencing. Otherwise, there is a risk of leaving large regions of the genome with no seed clone, resulting in delays until such regions are sequenced. Also, serious misassemblies can be generated in the rare but possible case in which a chimeric clone is selected as seed, and complications arise when nearly identical sequences to those in BAC ends are repeated elsewhere in the genome or if the assembly of the seed BAC sequence is incorrect.

The second method makes use of a physical map of the genome in which all BAC clones are positioned relative to each other. This BAC-based physical map is constructed by determining the pattern of fragments of each clone when digested with a restriction endonuclease. The size of the fragments is determined by running the digestion products in agarose gels and imaging the pattern of bands so that the size calling can be done automatically.[35] More recently, multicolor fluorescent capillary electrophoresis (high information content fingerprinting or HICF) has been applied to increase the amount of information per BAC clone, enhancing the resolution of the map.[36,37] All patterns of restriction fragments are then compared to each other using the FPC (fingerprinted contig) software, which incorporates genetic marker information if available.[38−40] FPC counts the number of fragments of the same size that are present in any two given clones. If a significant number of fragments are "shared" by two clones they are considered to overlap. Then FPC assembles overlapping clones into physical "contigs" and the physical map is constructed. A set of minimally overlapping clones (minimal tiling path or MTP) can be extracted from the physical map, and the sequence of the genome can be efficiently determined by sequencing those clones in the MTP. Use of a physical map to select BAC clones to sequence represents a significant additional effort that the seed clone method does not require. However, when the physical map is anchored to the genetic map (for example, by hybridizing molecular markers from the genetic map to the BAC clones) it represents a powerful resource that accelerates map-based cloning of interesting genes, justifying the efforts invested.

In either BAC-based genome sequencing method once BAC clones are selected they are sequenced by a random or

"shotgun" approach in which the clone DNA is mechanically sheared and cloned for sequencing at high redundancy.[41,42] The overlapping sequences are then assembled using computational methods to reconstruct the BAC sequence[43−46] (http://www.phrap.org). A high-quality sequence is obtained by manually completing gaps and/or low-quality regions in the assembled sequence in a time-consuming process called finishing.[47] Subsequently, consensus sequences of adjacent (partially overlapping) BACs are stitched together, and the genome sequence is constructed. The genomes of rice and *Arabidopsis* were completed in this way, meeting the quality standards set for the human genome. In many other cases only a "draft" sequence is pursued and partial or no finishing is carried out.

3. Plant Whole Genome Shotgun Sequencing

After completion of the human genome sequence by two separate efforts, one using a BAC-based strategy[5] and another using a WGS strategy,[6] it became clear that both are valid approaches to sequence the large genomes of higher eukaryotes, and even combined approaches have been pursued for other mammals.[2,3] Although its product is often a more discontinuous genome sequence than that achieved by the BAC-based method, the WGS approach has the advantage of being faster and more affordable. Therefore, WGS has been frequently applied to eukaryote genomes including relatively small animal and plant genomes during the last several years.[26,27,31,48,49]

The WGS strategy is basically similar to that described above for BAC shotgun sequencing. It was first used in the early 1980s to sequence a clone containing a few kbp fragment of a mitochondrial genome using DNase I to randomly break the cloned DNA and subcloning the resulting fragments in a sequencing vector.[41] Shortly after WGS was applied to a cloned fragment of a viral genome, this time mechanically shearing the DNA using sonication.[42] Currently, genomic DNA is mechanically broken in random pieces, generally using nebulization[47] or hydrodynamic forces,[50] the fragment ends are made blunt with DNA polymerases and/or nucleases,[47] and they are then cloned into sequencing vectors. Several libraries with different insert sizes are constructed, and the bulk of the sequences are typically obtained from the small-insert libraries. Each of these libraries spans a 1−2 kbp fragment range (i.e., 2−3, 6−8, and 10−12 kbp), and clones are sequenced from both ends. Use of multiple libraries of different insert sizes compensates for possible library biases. A portion of the sequence data is also obtained from end sequencing of large-insert clones, such as BACs (over 100 kbp)[32] or lambda phage-derived fosmid clones (about 40 kbp).[51,52] Such data are very useful to assemble the genome into large pieces and resolve the assembly of repeats. Genome assembly then takes place using informatic tools that align overlapping sequences and create a consensus, contiguous sequence (sequence contig).[43−46] Information on mate reads (sequences from both ends of the same clone) and average library insert size is used to aid in the assembly. Contigs can then be ordered and oriented relative to each other when each of the end sequences from a clone (usually large-insert ones) fall in different contigs. Such groups of linked contigs are called scaffolds. Scaffolds can also be anchored to the chromosomes in the genome by aligning the sequence of molecular markers that have been genetically mapped. In this way a high-quality draft sequence of a genome can be achieved as in the case of the 500 Mbp poplar genome.[27]

However, the WGS strategy is not an efficient way to approach large plant genomes, where repetitive elements, mainly retrotransposons,[30] can account for up to 90% of the genomic DNA.[53,54] Although mammalian genomes are also vastly repetitive, plant repetitive elements differ from those of mammals in that they often belong to very conserved families.[53] This is probably due to the fact that evolutionarily recent induction of retrotransposon activity resulted in sudden expansions of retrotransposon families.[55] Another characteristic of plant retrotransposons is that they tend to insert into each other forming large stretches of nested repetitive elements in intergenic regions.[56−61] This is different than the observed distribution of repetitive DNA in mammals, where transposable elements are often inserted in introns inside genes. This abundance of nearly identical repetitive sequences in large plant genomes is a major problem for the assembly programs. Repetitive elements from multiple locations in the genome tend to be assembled together, preventing building of long intergenic sequences. Thus, application of a WGS strategy to a large plant genome may efficiently assemble the low-copy fraction of the genome, which includes most genes, but it is likely that misassemblies will occur in the repetitive intergenic regions, reducing the contiguity of the overall assembly of the genome.

Because plant whole genome sequencing approaches that can deliver highly accurate and contiguous sequences are very costly, sequencing strategies to quickly capture low-copy or protein-coding sequences in the genome (gene-enriched genomic sequencing) have become common in recent years.

4. Chromosome-Specific Library Construction

The biggest challenge is posed by genomes such as that of common wheat, which, in addition to its extremely large size (16 Gbp)[7] and high level of repetitiveness, has the complication of being a recent polyploid.[62] It is composed of three highly similar genomes, posing an additional difficulty for assembly, as conserved low-copy sequences from separate homoeologous chromosomes can be erroneously merged together. One alternative to prevent this problem and, at the same time, reduce the complexity of the genome is to isolate chromosomes by flow cytometry and construct chromosome-specific libraries for sequencing. WGS or BAC libraries can be built using DNA isolated from single chromosomes.[63−66] This approach has the limitation that not all chromosomes can be separated from the rest by flow cytometry in any species. In hexaploid wheat, only chromosome 3B can be isolated using wild-type plants, but the rest of the chromosomes can be isolated from a collection of aneuploid lines.[67] Each plant in this collection contains only one of the three members of each homoeologous chromosome group that can be separated by flow cytometry. These lines can be exploited to isolate each of the 21 wheat chromosomes and make libraries for sequencing and/or physical mapping. This approach can be carried out in a distributed way, having multiple sequencing centers, each one taking on the sequencing of one or more chromosomes or chromosome arms. The applicability of this approach to other large plant genomes will depend on the feasibility of isolating chromosomes by flow cytometry.[68] Because the amount of chromosomal DNA that can be isolated in a reasonable time is very limited, use of gene-enrichment techniques in isolated chromosomes is not straightforward.

5. Gene-Enriched Sequencing

5.1. Expressed Sequence Tags

Because genes are the most commonly sought elements in a genome, sequencing cDNA clones is an efficient method to obtain predicted mRNA sequences and deduce the putative proteins coded in them. At the same time, cDNA sequences provide evidence of the expression of the identified genes. The sequencing of random cDNA clones was first proposed as a rapid method for identifying genes[69] and clearly expandable to the whole set of transcripts of an organism,[70] now called the transcriptome. Later, the idea of large-scale cDNA sequencing to identify new genes and determine their intron/exon structure was put to practice, and random cDNA sequences were called expressed sequence tags (EST).[71−73] These ESTs provided invaluable information to annotate the human genome sequence and became the most common approach, not only to obtain the first glimpse at the gene content of a genome, but also as a complement for any genome sequencing project. As a result, there were over 40 million ESTs in GenBank by the end of 2006, and the number is continuously growing (http://www.ncbi.nlm.nih-.gov/dbEST/dbEST_summary.html). About one-quarter of the ESTs in GenBank are from plants, and they are extremely useful to identify genes in any plant genome. Nevertheless, EST sequences rarely sample more than 50−60% of the genes in the source organism, even after exhaustive sampling of normalized libraries.[74,75] Genes expressed in highly specific conditions, cell types, or developmental stages are often missing in cDNA libraries.

When standard cDNA libraries are randomly sequenced the level of expression of different genes can be estimated by the frequency with which a given sequence appears in the EST set. This is useful information, particularly if different cDNA libraries from different tissues, conditions, or developmental stages are sequenced. Comparing the number of times a given sequence is present in each library provides information comparable to a Northern-blot hybridization.[71,76,77] The downside of the EST approach using standard cDNA libraries is that highly expressed genes may account for a substantial portion of the ESTs, not providing any new sequence data. In order to reduce redundancy in EST sets, normalized cDNA libraries can be used for random sequencing. cDNA normalization is based on the difference in reassociation kinetics of unique DNA sequences versus repeated ones. Normalization can be achieved by denaturing a double-stranded cDNA sample and allowing it to slowly reanneal.[78−80] Abundant cDNAs renaturate more quickly than rare ones, which remain single stranded for a longer period. The single-stranded (i.e., low-abundance) cDNA molecules can be separated from the abundant, double-stranded ones using hydroxyapatite (HAP) chromatography, which differentially binds double- and single-stranded DNA.[81] The single-stranded cDNA is eluted from the HAP column, the second strand is synthesized, and the thus normalized cDNA is cloned for sequencing ESTs. Improvements of these methods to increase the representation of long cDNAs have been reported.[75] Normalization is a more efficient gene discovery EST approach than standard cDNA libraries. Normalized ESTs provide a qualitative (but not quantitative) estimation of expression patterns. A disadvantage of the normalization procedure is that very similar members of gene families can cross hybridize in the reannealing step and may be lost with the abundant transcripts. Although large-scale EST projects cannot capture the complete set of genes in a genome, their efficiency for gene discovery and the expression and splicing information that they deliver make EST sequencing a robust genomics methodology as a stand-alone approach or in combination with other genomics resources.

As ESTs are random, single-pass sequences, different sequences often correspond to partially overlapping or different regions of the same gene due to incomplete cDNA synthesis. ESTs can also be generated from both ends of each cDNA clone, often resulting in complementary sequences. In order to reduce the redundancy of EST data and make it easier to handle, ESTs that share a high level of sequence identity and are thus likely to correspond to the same gene can be clustered together into "unigene" sets.[82] The sequences in each cluster can also be assembled into a mixture of contigs and singletons that have been named in various ways such as transcript assemblies,[83] gene indices,[84] unique transcripts,[85] etc. This procedure yields contiguous consensus cDNA sequences that are longer than individual ESTs. The total number of resulting transcript assemblies should not be considered proportional to the number of expressed genes in the genome because different assemblies or singletons may correspond to separate regions of the same gene, overestimating the number of genes tagged. Furthermore, as ESTs may contain sequencing errors, a small proportion of mismatches must be allowed during assembly. Therefore, the resulting consensus sequences must be used with caution because nearly identical paralogous transcripts[86] can be erroneously assembled together and considered to belong to a single gene.

If a complete genome is available, ESTs from the same species can be assembled by aligning them to the genome using tools that allow spliced alignments.[87−89] This strategy has the advantage that separate EST contigs that belong to the same gene are usually linked by the corresponding genomic sequence, and cDNA sequences from nearly identical paralogous genes are less likely to be merged because ESTs are aligned to their best match in the genome.

5.2. Methylation Filtration (MF)

EST sequences are very efficient as a gene discovery tool because they usually contain coding sequences that can be easily identified by similarity searches against protein databases. However, nontranscribed flanking regulatory sequences as well as intron sequences that are excluded from EST data contain important information that can only be retrieved by sequencing genomic DNA. Techniques that provide this genomic information while minimizing the amount of repetitive sequences are available for plants. One of them, called methylation filtration (MF), is a genomic DNA-based cloning and sequencing technique that takes advantage of the fact that most of the repetitive DNA in plant genomes is extensively methylated in the form of 5-methylcytosine (Figure 1). Levels of DNA methylation are very variable across different eukaryotes. The yeast *Saccharomyces cerevisiae* has no detectable 5-methylcytosine, while in other fungi methylation is found limited to repetitive DNA.[90,91] Animal genomes show a wide spectrum of DNA methylation levels. In the worm *Caenorhabditis elegans* DNA methylation has not been found, and in the insect *Drosophila melanogaster* cytosine methylation is restricted to a narrow developmental window and only found in CpA or CpT sequences.[92] On the contrary, vertebrates show much higher levels of methylation.[93] DNA methylation in mammals

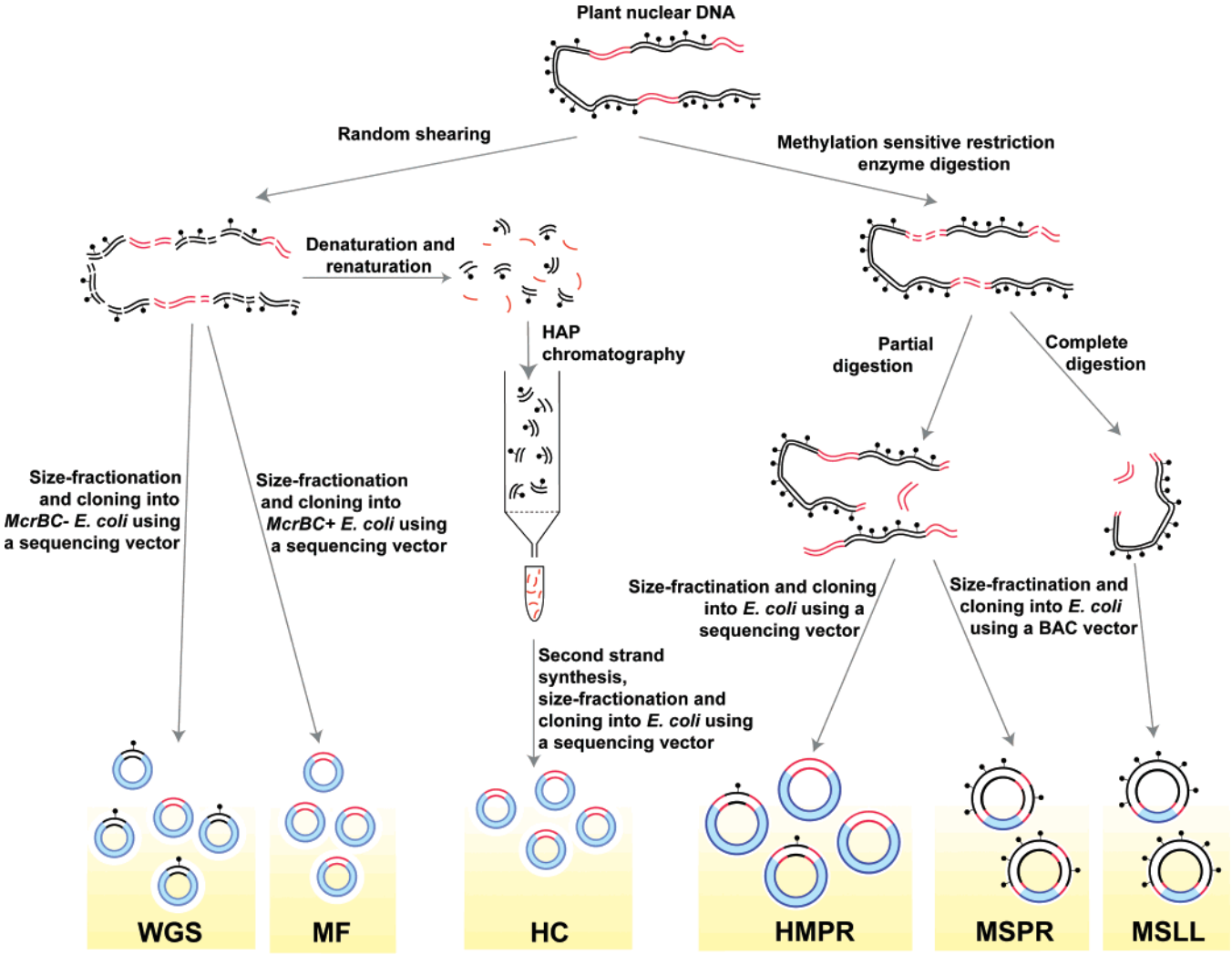

Figure 1. Summary or the different gene-enrichment sequencing strategies in cartoon style (double lines, double-stranded genomic DNA; black dots, methyl groups; circles, plasmid clones; red, genic or low-copy regions; black, repetitive regions; blue, cloning vectors).

has been found typically in CpG motifs, both in repetitive DNA as well as in genes,[94,95] except for the regulatory CpG islands.[96] DNA methylation has been found in all plants studied, frequently in CpG motifs, and also in CpNpG and asymmetric CpNpN motifs.[97,98] It has been shown that DNA methylation is associated with silent transposable elements in plants,[99−101] while genes are typically hypomethylated.[95,102] Recent genome-wide microarray analyses of DNA methylation in the small genome of *Arabidopsis* also detected DNA methylation in genes, although at lower levels than those observed in repeats, and in many cases methylation was localized toward the 3′ end of the genes.[103−105]

MF uses the *Escherichia coli* modified cytosine restriction system McrBC,[106,107] which is a restriction endonuclease that requires two recognition sites separated by 40−2000 bp, each half site consisting of a purine followed by a methylated cytosine.[108] As this is a frequent pattern in plant genomes, McrBC can digest virtually any DNA methylated at cytosines. Therefore, when plant genomic DNA is introduced in an *McrBC*$^+$ strain of *E. coli* it is frequently restricted, which is the reason why *McrBC*$^+$ strains are not routinely used for constructing eukaryotic genomic DNA libraries. MF libraries are constructed in the same way as small-insert WGS libraries but using an *McrBC*$^+$ host strain. Size fractionation of the DNA to select fragments between 1.5 and 3 kbp before cloning increases the chances of recovering low-copy DNA fragments, minimizing the presence of flanking methylated repetitive DNA. Purifying nuclear DNA is also necessary in order to reduce the amount of organelle DNA, which is non-methylated and therefore enriched in MF libraries.[109]

MF was first used in maize in a pilot study in which DNA was digested with a restriction enzyme whose recognition site does not overlap CpG or CpNpG methylation sites.[110] In this study a few hundred clones from filtered (*McrBC*$^+$ *E. coli* strains) and control (*McrBC*$^-$ strain) libraries were sequenced, and the proportion of gene-like sequences in each data set was determined using a database of protein sequences. A 6-fold increase in genes was obtained in the filtered libraries relative to the random control library (gene-enrichment ratio). On the other hand, a substantial decrease in repetitive sequences was observed among the filtered sequences. Consistently, chloroplast DNA and certain simple sequence repeats (SSRs), which are non-methylated, were also more abundant in the filtered library. This approach was later scaled up in maize, and over one-half a million MF sequences were produced by two groups, this time using randomly sheared DNA.[111,112] These studies showed that the gene discovery rate is lower than that of ESTs when less than 60 000 of each type of sequences are considered, but as more sequences are added, gene discovery by MF is more efficient and comprehensive. These studies also suggest that the non-methylated repetitive DNA found in filtered libraries

accounts for approximately 7% of the total repetitive DNA in the genome. Most of these non-methylated repetitive elements are probably ancient copies that have accumulated mutations resulting in a reduction in the proportion of cytosine and therefore cannot be methylated.[111] In some cases, they may correspond to active transposons, as it has been shown that transposons become active in mutants that reduce DNA methylation.[113−116] By comparing the frequency of gene-like sequences present in the non-methylated fraction of the genome (MF sequences) versus the frequency of gene-like sequences in the whole genome (random set of sequences) it was roughly estimated that the non-methylated space in the maize genome or the space sampled by MF was 425 Mbp.[111] These calculations were done under the assumption that all maize exons are non-methylated. Consistently, a small, random sample of maize exons was surveyed for the presence of methylation, and only 5% of them showed evidence of methylation in a methylation-sensitive PCR assay.[95] In contrast, 20−30% of the expressed genes are partially methylated in *Arabdopsis*, though at lower levels than repeats and pseudogenes.[104,105] Although comparable genome-wide analyses have not been done in maize, it is possible that gene methylation is variable among plants. Such variability could explain the differences in MF gene discovery efficiency observed in different plants as discussed below.

MF has been used on a large scale in sorghum with results consistent with those observed in maize.[117] Genes as well as SSRs and regulatory regions are enriched in sorghum MF sequences. Interestingly, MF also enriched for noncoding regulatory elements such as micro RNAs in sorghum.[118]

Gene enrichment by MF has been assessed in a range of plant genomes in pilot studies that included monocots, dicots, and non-angiosperms.[119] Assuming that most plants contain similar numbers of genes, the level of gene enrichment by MF should increase proportionally to the genome size (or subgenome size in recent polyploids), which is determined by the amount of methylated, repetitive DNA. These pilot studies suggest that the level of gene enrichment in monocot plants increases proportionally to the genome size with the exception of two wheat species. One of these is a diploid wheat, and the other is the hexaploid common wheat. In the diploid wheat, MF does not show the expected level of gene enrichment probably because of a large proportion of non-methylated repetitive elements. In the hexaploid wheat there appears to be an excessive number of gene-like sequences in the control random library, which reduces the gene-enrichment ratio. It is speculated that many of those gene-like sequences are probably pseudogenes produced by an amplification of gene sequences during the recent polyploidization event.[119] Another report of a MF analysis of diploid wheat also shows a low level of gene enrichment.[54] In this case as well a large proportion of repetitive elements is found in the MF sequence set.

In dicot plants the level of gene enrichment is somewhat lower than expected relative to the genome size. However, most of the dicot plants analyzed were ancient polyploids, and estimating the expected level of gene enrichment was difficult as loss of duplicated genes may occur without significant reduction in genome size.[12−16] Therefore, gene density is reduced although not to the level of a diploid. Regardless of this limitation, gene enrichment was observed in all dicot species tested. Other pilot studies performed in tomato showed that MF is an efficient way to discover coding sequences and regulatory regions. As expected for any gene-enrichment technique, these studies in tomato concluded that assembly of MF sequences will not produce long contiguous sequences, and intergenic regions will be missed.[120]

Non-angiosperm plants also show some level of gene enrichment,[119] including the small (~120 Mbp) genome of the early vascular, seedless plant *Selaginella* (Rabinowicz, unpublished results). An interesting case among these plants is pine, whose genome is approximately 20 Gbp. The expected high degree of gene enrichment for such a large genome was not observed in MF libraries, although enrichment in SSRs has been obtained.[121] One possible explanation is the presence of a large amount of ancient, CG-depleted transposable elements. An additional factor affecting the observed gene-enrichment ratio of MF in pine may be the presence of a large number of pseudogenes, as proposed for wheat, because a very large number of gene-like sequences was found in the random control library in pine. There is no clear evidence of polyploidy in pine, but a high level of gene duplication has been observed,[122,123] which is consistent with a pseudogene amplification.

MF has been tested in mammalian genomes to estimate the levels of enrichment in genic sequences. Mammalian genomes contain DNA methylation in CpG motifs; therefore, it can be digested by McrBC in the same way as plant DNA. However, when applied to mouse somatic tissues as well as human cells in culture MF libraries did not show a significant difference in the proportion of genic and repetitive sequences relative to control libraries.[95] Two factors contribute to these results. On one hand, mammalian repeats are mostly ancient and GpC depleted.[5] Therefore, they may not be counter-selected in MF libraries. On the other hand and more importantly, mammalian exon sequences have been shown to be methylated. In a comparative study, randomly chosen exons from maize could be amplified by PCR from a genomic DNA template previously digested with McrBC in vitro. This is consistent with the known hypomethylation of plant genes. When the same assay was performed on a random set of mammalian exons, a decrease in the amount of product was observed after PCR amplification of McrBC-treated versus untreated DNA template. This study showed that mammalian exons are methylated as often as repetitive elements are in these genomes.[95]

5.2.1. Other Uses of MF

MF has other potential uses such as selectively cloning and sequencing non-methylated genomes (i.e., bacterial genomes) in samples containing mixtures of different DNA sources. Preliminary data suggests that such an approach may be useful for selectively sequencing the genomes of phytoplasmas, which are unculturable obligate parasites.[124] When a large-insert MF library is constructed from plant DNA, the cloning efficiency is very low because the chances of recovering large fragments of DNA completely depleted of methylation are very few. Thus, when DNA from aster plants infected with the intracellular bacterial parasite Aster Yellows (AY) phytoplasma was used to construct a large-insert MF library, several clones containing AY DNA were recovered (Rabinowicz, unpublished results). As the AY genome is small (700 kbp) only a few dozen clones are necessary to cover the whole genome. Those clones can then be completely sequenced to assemble the parasite's genome. Although a large proportion of the recovered MF clones contained chloroplast DNA because it is also non-methylated, generating enough clones and sequencing their ends to detect and discard chloroplast DNA may minimize the problem.

If chloroplast sequences are of interest, the observed enrichment in such sequences in MF libraries can be useful to selectively sequence chloroplast genomes without the problem of purifying these organelles before preparing the DNA. With the purpose of performing phylogenetic studies of castor bean, chloroplast genomes from different cultivars have been sequenced using MF. Application of MF to total DNA preparations from castor bean leaves yielded up to 50% chloroplast sequences, which due to the small size of the genome (160 kbp) and the current low sequencing costs resulted in an efficient way to sequence multiple chloroplast genomes (Rabinowicz and Ravel, unpublished results).

5.3. High Cot (HC) Sequencing

In the late 1960s it was shown that when genomic DNA is denatured and slowly renatured the high-copy DNA reanneals more rapidly than medium- and low-copy DNA.[125] In a renaturation reaction the product of the concentration and time required for reassociation is called Cot, and it can be used to characterize the different components of eukaryotic genomes in terms of repetitiveness. The slow reannealing or high Cot component is mostly low-copy DNA, while the fast reannealing component corresponds to highly repetitive or low Cot DNA. The moderately repetitive DNA shows intermediate Cot value. These different components of a genome can be isolated using HAP chromatography[81] in a similar way as described above for normalization of cDNA. In order to isolate the low-copy fraction of the genome the denatured DNA is allowed to reassociate so that the highly and moderately repetitive DNA is mostly in a double-stranded form while the low-copy DNA remains in a single-stranded form, which has different affinity for HAP (Figure 1). In this way the low-copy DNA can be separated from the medium- and high-copy DNA and cloned after in vitro synthesis of the second DNA strand. Because genes reside mainly in the low-copy fraction of plant genomes, use of high Cot (HC) DNA cloning to enrich in genomic sequences containing genes has been proposed for maize,[126] and subsequent pilot studies showed its applicability in sorghum[127] and maize.[128] HC sequencing (also called Cot-based cloning and sequencing or CBCS[127]) was shown to enrich for genes and other low-copy sequences in both systems. In maize, the proportion of genic HC sequences was similar to that obtained with MF. In order to minimize the problem of losing members of multigene families, several HC libraries are made using different reannealing times. At longer reassociation times the amount of highly and moderately repetitive DNA is more efficiently reduced at the expense of increasing the chances of missing members of large gene families. At shorter reassociation times gene family recovery increases but so does the recovery of moderately repetitive DNA.

The HC results in sorghum are not completely comparable to the MF ones because the HC clone inserts were small and the library was made using an *McrBC*$^+$ *E. coli* host,[127] then inadvertently combining the HC and MF methods.

HC has also been applied to hexaploid wheat,[129] but because a random set of sequences from another (diploid) wheat species was used as a control, the level of enrichment in genes could not be estimated. In order to do this the hexaploid wheat HC data can be compared to the random sequences generated by others.[119] Such an analysis showed that the level of enrichment of HC is slightly lower than that obtained by MF (Rabinowicz, unpublished results). This low level of gene enrichment could be explained by an abundance of inactive repetitive elements that accumulated mutations to the extent that they do not easily hybridize to each other and therefore behave as low-copy DNA in a Cot experiment. In addition, large gene families (including pseudogenes) may be underrepresented. These results should be taken with caution because the number of wheat sequences analyzed is relatively small, and larger-scale analysis is required to draw more reliable conclusions. When applied to another large genome such as that of pine, HC results are consistent with the idea that heavily mutated retroelements behave as low-copy DNA,[130] and therefore, the gene discovery rate is low.

HC sequencing has only been used on a large scale in maize together with the large-scale maize MF project described above.[112] This constitutes the largest data sets of both gene-enrichment techniques for a given species, and comparison of the results of each technique produced interesting discoveries. This maize HC sequence data shows a moderately lower gene discovery capacity than MF. However, sequences with no match in DNA, EST, or protein databases are much more abundant among the HC sequences than in the MF set. Such sequences may represent noncoding portions of genes, which are less conserved and more difficult to identify as such by cross-species sequence comparisons. They could also represent still uncharacterized low-copy transposable elements.

After the maize MF and HC data was made public it was reported that mutations occurred at low frequency in the HC dataset.[131] Later, it was shown that such an artifact was a consequence of a low buffering capacity of the citrate buffer used during the slow reassociation step in which DNA is kept at relatively high temperatures for extended periods. Use of phosphate buffer eliminated the problem (Bennetzen, personal communication).

In an attempt to extend the use of this technique for efficient gene discovery in vertebrates, HC has been used to analyze the 1.2 Gbp genome of chicken. Unfortunately, this experiment resulted in no significant gene enrichment.[132] This outcome was attributed to the extensive amount of mutations accumulated in vertebrate transposable element families in a similar way as discussed above for the failure of MF to enrich for genes in mammals.

5.4. Combination of HC and MF

The large-scale maize gene-enrichment project brought an opportunity to perform a comprehensive comparison of the two methods. Overall, the findings from the parallel HC and MF analysis of the maize genome demonstrate that these two techniques recover partially overlapping fractions of the genome[112,133] and therefore show that the combination of both techniques results in a very efficient and effective way to rapidly identify the gene space of large plant genomes. When this sequencing project was half way to completion, the HC and MF sequences were assembled separately as well as combining both data sets as input. The fraction of the maize genome that would be sampled by each method was estimated by applying the Lander−Waterman algorithm,[112,134] which describes the coverage and number of gaps in a genome assembly for a given genome size and amount of sequence available. With this analysis it was predicted that the genome space sampled by MF was 260 and 280 Mbp by HC, although it has been proposed that the Lander−Waterman algorithm underestimates the size of the sampled

space in gene-enriched libraries.[135] Combination of both sequence sets spans 400 Mbp, which is less than the 540 Mbp resulting from addition of both sets. These extrapolations are consistent with the idea that each technique samples different but overlapping fractions of the genome that combined represent a 6-fold reduction of the genome size. However these extrapolations may have overestimated the sampled spaces. After completion of the project, nearly one-half a million HC reads were assembled into a total of 190 Mbp and MF assemblies span 150 Mbp, while the combined assembly contains nearly 300 Mbp (Chan et al., unpublished results; http://miaze.tigr.org).

Attempts to estimate the coverage of the gene space have been reported. Using a set of full-length cDNAs, 95% were tagged by MF and/or HC reads[133] and the average coverage of exons was 2× (Barbazuk, unpublished results). Using a curated set of gene models approximately 75% of the nucleotides were covered by either MF or HC reads (Barbazuk, unpublished results). Another study aligned MF and HC sequences to annotated maize BAC clones that are part of a large physical contig, and over 90% of the annotated genes were touched by MF or HC reads. In this report 75% of the exonic nucleotides and 49% of the nucleotides in exons and introns were covered by HC or MF.[136] The observed difference in coverage of exons versus entire gene models is due to the presence of repetitive sequences in some of the introns of the gene models analyzed.

Pseudogenes can be misleading in this kind of analyses because they can be considered gene-like sequences in fragmentary sets of MF or WGS sequences. A large number of gene fragments have been found in rice and maize forming part of transposable elements,[137,138] and hence, it is likely that most of them are methylated,[104,105] although some have been shown to be expressed[139] and, therefore, potentially non-methylated.

Assembling gene-enriched sequences not only extends low-copy sequences into contigs but also allows annotating larger gene fragments and, sometimes, entire genes, including their regulatory regions.[112,140,141] Placing such gene-enriched assemblies in the genetic map can be achieved by aligning them to sequences that have been genetically mapped.[120,142] If a BAC-based physical map is available and the ends of the mapped clones have been sequenced, gene-enriched assemblies can be anchored to the physical map by aligning them to the BAC-end sequences.

5.5. Methylation-Sensitive Digestion of DNA

Researchers have used the low level of methylation in plant genes as a way to recover low-copy sequences for mapping purposes. Cloning the 1.5−2.5 kbp fraction of maize DNA after digestion with a common methylation-sensitive restriction endonuclease resulted in a genomic library rich in low-copy sequences that could be used to design restriction fragment length polymorphism (RFLP) markers for genetic mapping.[143] This would not be an efficient genome-wide gene discovery strategy because only those low-copy sequences that contain 2 of the corresponding 6 bp recognition sites at the right distance between each other will be cloned. One way to increase the representation of hypomethylated sequences using this approach is use of partial digestion of genomic DNA with multiple frequent-cutter methylation-sensitive restriction enzymes. This idea has been called hypomethylated partial restriction (HMPR; Figure 1).[144] Use of multiple restriction enzymes with different recognition sites partially compensates for the sequence biases. In a pilot study using two such restriction enzymes a high degree of gene enrichment was observed along with a substantial proportion of sequences with no match in sequence databases. These sequences are likely to be introns and regulatory regions. In addition, the proportion of retrotransposon sequences recovered in HMPR libraries was lower than those observed in HC and MF libraries. Although this technique is very efficient as a gene discovery tool, the randomness of the recovered clones has not been assessed at on a large scale and may require using many restriction enzymes and multiple levels of partial digestion to achieve a comprehensive representation of all genes, which may amount to a large library construction effort. HMPR has an additional advantage if used in combination with HC and MF because HMPR clones that contain low-copy DNA at the ends may contain repetitive methylated sequences in the middle. Such a clone would be counter-selected in HC and MF libraries. Thus, end sequences of HMPR clones can help linking separate MF and/or HC sequence assemblies that are closely linked in the genome, but an intervening repetitive sequence prevented generation of a single sequence contig[145] (Figure 2).

In large plant genomes such as that of maize intergenic sequences can be extremely long, often reaching hundreds of kbp consisting only of heavily methylated, intact, or rearranged retrotransposons. In order to link genic sequences from HC and MF assemblies that are many kbp apart end sequences of large-insert clones from genomic libraries constructed using methylation-sensitive restriction enzymes can be useful. Methylation-spanning linker libraries (MSLL) are constructed by completely digesting nuclear DNA with 4- or 6-bp recognition site methylation-sensitive restriction enzymes[146] (Figure 1). Digestion of maize genomic DNA with frequent-cutter restriction enzymes can yield a majority of fragments bigger than 50 kbp due to the high proportion of methylated recognition site sequences and/or to GpC and GpNpG depletion.[102] MSLL has been applied to maize on a small scale using relatively short-insert size BAC libraries (10−25 kbp). This report showed that the end sequences of these clones were enriched in genes and, although retrotransposon sequences were abundant, the ends of MSLL BAC clones were outside but close to repetitive elements. Therefore, MSLL should be a useful tool in combination with HC, MF, and HMPR to construct genomic scaffolds containing gene-enriched assemblies. In these scaffolds genic sequences can be oriented relative to each other and the physical distance between them can be estimated using the MSLL library insert-size and mate-read information. The putative repetitive sequences separating those gene sequences would remain unknown (Figure 2).

A modification of the MSLL and HMPR techniques to increase randomness and genome coverage has been recently published.[147] This technique, called methylation-sensitive partial restriction (MSPR; Figure 1), differs from HMPR in that the insert size is closer to that of typical BAC libraries (∼100 kbp) and from MSLL in that the digestion with a methylation-sensitive restriction enzyme is partial, therefore increasing the randomness of the clones. One problem of constructing MSLL and MSPR libraries is that BAC vectors do not typically include multiple methylation-sensitive restriction enzyme recognition sites in their multiple cloning site. Yuan and co-workers[146] completely digested genomic DNA with *Hpa* II or *Sal* I and partially filled in the cohesive

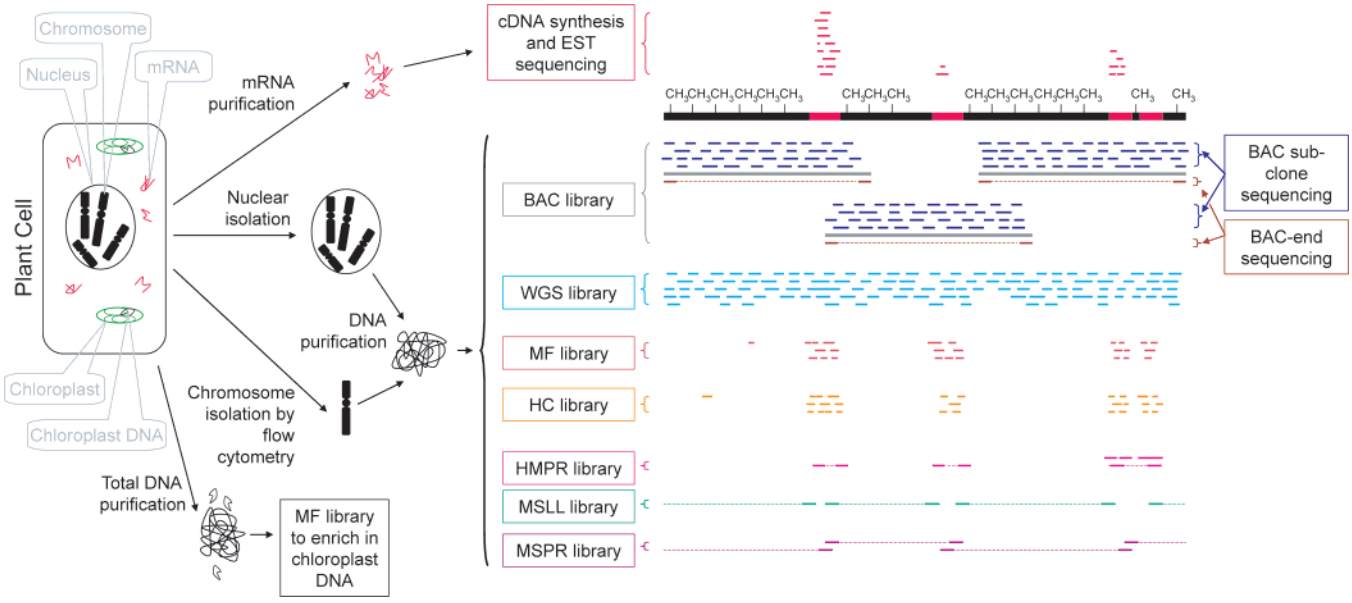

Figure 2. Schematic representation of different plant genomic sequencing approaches. DNA from whole cells, isolated nuclei, or purified chromosomes is used for construction of different libraries. A genomic region is represented as a black bar toward the top. Genes in this region are represented in red, and repetitive, methylated intergenic DNA is marked in black with "CH_3" symbols. BAC clones from a minimal tiling path (MTP) are shown in gray. Sequence reads from each library type are shown as color-coded dashes. Each strategy can be used separately or in combination with any other (see text for details). Dotted lines represent unsequenced portions of the clones.

ends to make them compatible for ligation with a partially filled-in BAC vector digested in its *Bam* HI site. In order to reduce the manipulations of the DNA prior to cloning, Yu and Li[147] constructed a BAC vector containing recognition sites for three methylation-sensitive restriction enzymes. This vector does not need a fill-in step and facilitates direct cloning of large fragments from genomic DNA partially digested with those enzymes. Analysis of end sequences from a small number of maize MSPR clones showed a low proportion of retrotransposon-related sequences,[147] suggesting that this method would be efficient in discovering and linking low-copy sequences putatively containing genes at high frequency.

5.6. Transposon Insertion-Site Sequencing

DNA transposons are typically less repetitive than retrotransposons in large plant genomes.[29] Several transposon families, particularly in maize, have been studied extensively.[148−150] One interesting feature of DNA transposons is that, opposite to what is observed in retrotransposons, they tend to insert in genic regions.[151] Taking advantage of this feature of DNA transposons has allowed their use as mutagens.[152,153] This process facilitates identification of the mutated gene by the presence of a transposon "tag". In this way a reverse genetics approach is enabled. Random mutagenesis projects have been conducted in maize using the transposons *Mutator* and *Ac/Ds* to obtain large collections of mutant lines.[154−156] Using a plant line in which the transposon is highly active it is possible to generate large populations of plants containing transposon-induced mutations. If the mutant population is large enough, it is expected that a transposon insertion can be obtained in most genes. By isolating and sequencing the regions flanking the newly inserted transposable elements a collection of gene-enriched genomic sequences can be obtained.[151] Transposon insertion site sequences can be isolated by a PCR strategy that utilizes a primer complementary to the transposon end sequence and a random primer to anneal to any flanking genomic sequence. With the goal of facilitating large-scale cloning and sequencing of transposon insertion sites, a modified *Mutator* transposon has been developed. This system called *RescueMU* consists of a transgenic plant containing a copy of the transposon engineered to include a bacterial origin of replication and an antibiotic resistance gene. In this way plasmids containing the genomic sequences flanking the transposon insertion site can be rescued from genomic DNA by restriction enzyme digestion, ligation (to circularize the plasmid), and *E. coli* transformation.[157] Sequencing the clones obtained after this procedure yields transposon insertion site sequences that are typically gene rich. However, transposon tagging resulted in an uneven representation of genes with some gene sequences being recovered at high frequency. This is due to transposon insertions that were present before the mutagenesis was induced (parental insertions) and to the bias of the transposon to insert itself in certain regions more frequently than in others.[154,157] Therefore, transposon insertion site sequencing is less effective as a gene discovery method than other gene-enrichment techniques, although the system has the advantage that if a gene is found in the collection of insertion site sequences it is likely that the plant containing it is a mutant for that gene.

5.7. Gene-Rich BAC Sequencing

As very large plant genomes can contain stretches of repetitive sequences spanning above 100 kbp, BAC libraries constructed for such genomes often contain clones that consist purely of repetitive sequences. Sequencing and assembly of such repetitive BACs often results in fragmented and misassembled consensus sequences and may be of little use for a draft sequencing project. Therefore, if a BAC-based sequencing or physical mapping approach is carried out, it is useful to identify gene-containing BAC clones so the number of clones to work with is reduced. Using gene sequences derived from gene-enriched data sets (i.e., EST, MF, HC, etc.), gene-containing BAC clones can be identified by hybridization. This approach, in some cases aided with

cytogenetic data, has been applied to the *Medicago* (http://www.medicago.org/genome/about.php), tomato (http://www.sgn.cornell.edu/about/tomato_project_overview.pl), and barley (http://phymap.ucdavis.edu:8080/barley) genomes with encouraging results. In the case of barley, the reduced set of gene-rich BACs obtained with this approach allows focusing the genomic analysis on its most fruitful portion. The limitation of the approach is that a comprehensive gene sequence set must be available to be able to identify most gene-containing BACs. Relying only in EST sequences may result in an incomplete set of gene-containing BACs because of the expression biases of EST sequencing.[75] Therefore, a combination of extensive MF/HC and EST sequence data can provide the raw material to identify most of the gene-containing BAC clones, out of which an MTP of clones can be selected for sequencing.

6. Conclusions

Gene enrichment as well as EST sequencing represent efficient approaches to extract gene information from plant genomes on both the large or moderate scale, depending on the required downstream application of the data. However, when the goal is to comprehensively sequence large plant genomes or gene repertoires, no single, easily affordable technology is currently available. In such cases, any of the approaches discussed here can yield different kinds of useful genome-wide data. One limitation of gene enrichment and EST sequencing is a lack of mapping information. Combining different approaches coupled with mapping techniques can deliver most of the genome sequence information in a cost-effective manner. In the cases of mouse and rat, a hybrid strategy using both WGS- and BAC-based approaches was shown to seize the best of both worlds. Small-insert WGS libraries have the potential of capturing genomic regions that are difficult to recover as large-insert BAC clones. Furthermore, the WGS sequences can be produced much faster than BAC sequences and made quickly available to the community. A BAC-based sequencing approach, if complemented with a physical map, allows assembling the genome in much larger contigs than those obtained in a WGS-only approach. Those long contigs can be anchored to the chromosomes using sequenced genetic markers, resulting in a comprehensive genomic resource. In plants, these approaches can be taken one step further by adding gene-enriched sequences to provide deeper coverage of genic regions (Figure 2). EST data constitutes a key complement of any genome sequencing strategy. ESTs and other cDNA sequences, particularly full-length cDNAs, not only help in gene annotation (identification and modeling of the structure of genes) but also are fundamental for ab initio gene prediction as they are useful to identify a reliable set of genes needed to "train" gene modeling programs to correctly de novo identify gene sequences in the genome under analysis.[158]

Sequencing technologies are continuously improving in throughput and quality, and at the same time their cost is decreasing. Developing technologies that have the capacity of producing larger amounts of sequence data and much faster than the currently used capillary fluorescence sequencing by the Sanger method are advancing, still with the limitation of having short read lengths. Continuous improvement of established and emerging technologies as well as creation of new ones will eventually make sequencing large genomes a routine laboratory technique. Until that becomes a reality, large-scale sequence data from the complex genomes of important crops, such as wheat and barley, can be more effectively obtained using gene-enrichment techniques combined with other genome mapping and sequencing approaches as resources allow.

7. Acknowledgments

I thank Agnes Chan for comments on the manuscript. The author is supported in part by grants from the National Science Foundation (award no. DBI-0638558) and the United States Department of Agriculture (award no. 2006-36504-17248).

8. References

(1) Lewin, B. *Genes VII*; Oxford University Press: New York, 2000.
(2) Mouse Genome Sequencing Consortium. *Nature* **2002**, *420*, 520.
(3) Gibbs, R. A.; Weinstock, G. M.; Metzker, M. L.; Muzny, D. M.; Sodergren, E. J.; Scherer, S.; Scott, G.; Steffen, D.; Worley, K. C.; Burch, P. E.; Okwuonu, G.; Hines, S.; Lewis, L.; DeRamo, C.; Delgado, O.; Dugan-Rocha, S.; Miner, G.; Morgan, M.; Hawes, A.; Gill, R.; Celera Holt, R. A.; Adams, M. D.; Amanatides, P. G.; Baden-Tillson, H.; Barnstead, M.; Chin, S.; Evans, C. A.; Ferriera, S.; Fosler, C.; Glodek, A.; Gu, Z.; Jennings, D.; Kraft, C. L.; Nguyen, T.; Pfannkoch, C. M.; Sitter, C.; Sutton, G. G.; Venter, J. C.; Woodage, T.; Smith, D.; Lee, H. M.; Gustafson, E.; Cahill, P.; Kana, A.; Doucette-Stamm, L.; Weinstock, K.; Fechtel, K.; Weiss, R. B.; Dunn, D. M.; Green, E. D.; Blakesley, R. W.; Bouffard, G. G.; De Jong, P. J.; Osoegawa, K.; Zhu, B.; Marra, M.; Schein, J.; Bosdet, I.; Fjell, C.; Jones, S.; Krzywinski, M.; Mathewson, C.; Siddiqui, A.; Wye, N.; McPherson, J.; Zhao, S.; Fraser, C. M.; Shetty, J.; Shatsman, S.; Geer, K.; Chen, Y.; Abramzon, S.; Nierman, W. C.; Havlak, P. H.; Chen, R.; Durbin, K. J.; Egan, A.; Ren, Y.; Song, X. Z.; Li, B.; Liu, Y.; Qin, X.; Cawley, S.; Cooney, A. J.; D'Souza, L. M.; Martin, K.; Wu, J. Q.; Gonzalez-Garay, M. L.; Jackson, A. R.; Kalafus, K. J.; McLeod, M. P.; Milosavljevic, A.; Virk, D.; Volkov, A.; Wheeler, D. A.; Zhang, Z.; Bailey, J. A.; Eichler, E. E.; Tuzun, E.; Birney, E.; Mongin, E.; Ureta-Vidal, A.; Woodwark, C.; Zdobnov, E.; Bork, P.; Suyama, M.; Torrents, D.; Alexandersson, M.; Trask, B. J.; Young, J. M.; Huang, H.; Wang, H.; Xing, H.; Daniels, S.; Gietzen, D.; Schmidt, J.; Stevens, K.; Vitt, U.; Wingrove, J.; Camara, F.; Mar Alba, M.; Abril, J. F.; Guigo, R.; Smit, A.; Dubchak, I.; Rubin, E. M.; Couronne, O.; Poliakov, A.; Hubner, N.; Ganten, D.; Goesele, C.; Hummel, O.; Kreitler, T.; Lee, Y. A.; Monti, J.; Schulz, H.; Zimdahl, H.; Himmelbauer, H.; Lehrach, H.; Jacob, H. J.; Bromberg, S.; Gullings-Handley, J.; Jensen-Seaman, M. I.; Kwitek, A. E.; Lazar, J.; Pasko, D.; Tonellato, P. J.; Twigger, S.; Ponting, C. P.; Duarte, J. M.; Rice, S.; Goodstadt, L.; Beatson, S. A.; Emes, R. D.; Winter, E. E.; Webber, C.; Brandt, P.; Nyakatura, G.; Adetobi, M.; Chiaromonte, F.; Elnitski, L.; Eswara, P.; Hardison, R. C.; Hou, M.; Kolbe, D.; Makova, K.; Miller, W.; Nekrutenko, A.; Riemer, C.; Schwartz, S.; Taylor, J.; Yang, S.; Zhang, Y.; Lindpaintner, K.; Andrews, T. D.; Caccamo, M.; Clamp, M.; Clarke, L.; Curwen, V.; Durbin, R.; Eyras, E.; Searle, S. M.; Cooper, G. M.; Batzoglou, S.; Brudno, M.; Sidow, A.; Stone, E. A.; Payseur, B. A.; Bourque, G.; Lopez-Otin, C.; Puente, X. S.; Chakrabarti, K.; Chatterji, S.; Dewey, C.; Pachter, L.; Bray, N.; Yap, V. B.; Caspi, A.; Tesler, G.; Pevzner, P. A.; Haussler, D.; Roskin, K. M.; Baertsch, R.; Clawson, H.; Furey, T. S.; Hinrichs, A. S.; Karolchik, D.; Kent, W. J.; Rosenbloom, K. R.; Trumbower, H.; Weirauch, M.; Cooper, D. N.; Stenson, P. D.; Ma, B.; Brent, M.; Arumugam, M.; Shteynberg, D.; Copley, R. R.; Taylor, M. S.; Riethman, H.; Mudunuri, U.; Peterson, J.; Guyer, M.; Felsenfeld, A.; Old, S.; Mockrin, S.; Collins, F. *Nature* **2004**, *428*, 493.
(4) Kirkness, E. F.; Bafna, V.; Halpern, A. L.; Levy, S.; Remington, K.; Rusch, D. B.; Delcher, A. L.; Pop, M.; Wang, W.; Fraser, C. M.; Venter, J. C. *Science* **2003**, *301*, 1898.
(5) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.;

Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; Szustakowki, J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J. *Nature* **2001**, *409*, 860.

(6) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. *Science* **2001**, *291*, 1304.

(7) Arumuganathan, K.; Earle, E. D. *Plant Mol. Biol. Rep.* **1991**, *9*, 208.

(8) Bennett, M. D.; Leitch, I. J. *Ann. Bot.* **1995**, *76*, 113.

(9) Wang, W.; Tanurdzic, M.; Luo, M.; Sisneros, N.; Kim, H. R.; Weng, J. K.; Kudrna, D.; Mueller, C.; Arumuganathan, K.; Carlson, J.; Chapple, C.; de Pamphilis, C.; Mandoli, D.; Tomkins, J.; Wing, R. A.; Banks, J. A. *BMC Plant Biol.* **2005**, *5*, 10.

(10) Flavell, R. B.; Bennett, M. D.; Smith, J. B.; Smith, D. B. *Biochem. Genet.* **1974**, *12*, 257.

(11) Wendel, J. F. *Plant Mol. Biol.* **2000**, *42*, 225.

(12) Lai, J.; Ma, J.; Swigonova, Z.; Ramakrishna, W.; Linton, E.; Llaca, V.; Tanyolac, B.; Park, Y. J.; Jeong, O. Y.; Bennetzen, J. L.; Messing, J. *Genome Res.* **2004**, *14*, 1924.

(13) Langham, R. J.; Walsh, J.; Dunn, M.; Ko, C.; Goff, S. A.; Freeling, M. *Genetics* **2004**, *166*, 935.

(14) Ilic, K.; SanMiguel, P. J.; Bennetzen, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12265.

(15) Zhu, T.; Schupp, J. M.; Oliphant, A.; Keim, P. *Mol. Gen. Genet.* **1994**, *244*, 638.

(16) Krishnan, P.; Sapra, V. T.; Soliman, K. M.; Zipf, A. *J. Hered.* **2001**, *92*, 295.

(17) Pennisi, E. *Science* **2005**, *310*, 603.

(18) Pennisi, E. *Science* **2005**, *309*, 999.

(19) Pennisi, E. *Science* **2005**, *314*, 232.

(20) Wicker, T.; Schlagenhauf, E.; Graner, A.; Close, T. J.; Keller, B.; Stein, N. *BMC Genomics* **2006**, *7*, 275.

(21) Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P.; Begley, R. F.; Rothberg, J. M. *Nature* **2005**, *437*, 376.

(22) Shendure, J.; Porreca, G. J.; Reppas, N. B.; Lin, X.; McCutcheon, J. P.; Rosenbaum, A. M.; Wang, M. D.; Zhang, K.; Mitra, R. D.; Church, G. M. *Science* **2005**, *309*, 1728.

(23) Goldberg, S. M.; Johnson, J.; Busam, D.; Feldblyum, T.; Ferriera, S.; Friedman, R.; Halpern, A.; Khouri, H.; Kravitz, S. A.; Lauro, F. M.; Li, K.; Rogers, Y. H.; Strausberg, R.; Sutton, G.; Tallon, L.; Thomas, T.; Venter, E.; Frazier, M.; Venter, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11240.

(24) Dehal, P.; Satou, Y.; Campbell, R. K.; Chapman, J.; Degnan, B.; De Tomaso, A.; Davidson, B.; Di Gregorio, A.; Gelpke, M.; Goodstein, D. M.; Harafuji, N.; Hastings, K. E.; Ho, I.; Hotta, K.; Huang, W.; Kawashima, T.; Lemaire, P.; Martinez, D.; Meinertzhagen, I. A.; Necula, S.; Nonaka, M.; Putnam, N.; Rash, S.; Saiga, H.; Satake, M.; Terry, A.; Yamada, L.; Wang, H. G.; Awazu, S.; Azumi, K.; Boore, J.; Branno, M.; Chin-Bow, S.; DeSantis, R.; Doyle, S.; Francino, P.; Keys, D. N.; Haga, S.; Hayashi, H.; Hino, K.; Imai, K. S.; Inaba, K.; Kano, S.; Kobayashi, K.; Kobayashi, M.; Lee, B. I.; Makabe, K. W.; Manohar, C.; Matassi, G.; Medina, M.; Mochizuki, Y.; Mount, S.; Morishita, T.; Miura, S.; Nakayama, A.; Nishizaka, S.; Nomoto, H.; Ohta, F.; Oishi, K.; Rigoutsos, I.; Sano, M.; Sasaki, A.; Sasakura, Y.; Shoguchi, E.; Shin-i, T.; Spagnuolo, A.; Stainier, D.; Suzuki, M. M.; Tassy, O.; Takatori, N.; Tokuoka, M.; Yagi, K.; Yoshizaki, F.; Wada, S.; Zhang, C.; Hyatt, P. D.; Larimer, F.; Detter, C.; Doggett, N.; Glavina, T.; Hawkins, T.; Richardson, P.; Lucas, S.; Kohara, Y.; Levine, M.; Satoh, N.; Rokhsar, D. S. *Science* **2002**, *298*, 2157.

(25) International Chicken Genome Sequencing Consortium. *Nature* **2004**, *432*, 695.

(26) The Honeybee Sequencing Consortium. *Nature* **2006**, *444*, 512.

(27) Tuskan, G. A.; Difazio, S.; Jansson, S.; Bohlmann, J.; Grigoriev, I.; Hellsten, U.; Putnam, N.; Ralph, S.; Rombauts, S.; Salamov, A.; Schein, J.; Sterck, L.; Aerts, A.; Bhalerao, R. R.; Bhalerao, R. P.; Blaudez, D.; Boerjan, W.; Brun, A.; Brunner, A.; Busov, V.; Campbell, M.; Carlson, J.; Chalot, M.; Chapman, J.; Chen, G. L.; Cooper, D.; Coutinho, P. M.; Couturier, J.; Covert, S.; Cronk, Q.; Cunningham, R.; Davis, J.; Degroeve, S.; Dejardin, A.; Depamphilis, C.; Detter, J.; Dirks, B.; Dubchak, I.; Duplessis, S.; Ehlting, J.; Ellis,

B.; Gendler, K.; Goodstein, D.; Gribskov, M.; Grimwood, J.; Groover, A.; Gunter, L.; Hamberger, B.; Heinze, B.; Helariutta, Y.; Henrissat, B.; Holligan, D.; Holt, R.; Huang, W.; Islam-Faridi, N.; Jones, S.; Jones-Rhoades, M.; Jorgensen, R.; Joshi, C.; Kangasjarvi, J.; Karlsson, J.; Kelleher, C.; Kirkpatrick, R.; Kirst, M.; Kohler, A.; Kalluri, U.; Larimer, F.; Leebens-Mack, J.; Leple, J. C.; Locascio, P.; Lou, Y.; Lucas, S.; Martin, F.; Montanini, B.; Napoli, C.; Nelson, D. R.; Nelson, C.; Nieminen, K.; Nilsson, O.; Pereda, V.; Peter, G.; Philippe, R.; Pilate, G.; Poliakov, A.; Razumovskaya, J.; Richardson, P.; Rinaldi, C.; Ritland, K.; Rouze, P.; Ryaboy, D.; Schmutz, J.; Schrader, J.; Segerman, B.; Shin, H.; Siddiqui, A.; Sterky, F.; Terry, A.; Tsai, C. J.; Uberbacher, E.; Unneberg, P.; Vahala, J.; Wall, K.; Wessler, S.; Yang, G.; Yin, T.; Douglas, C.; Marra, M.; Sandberg, G.; Van de Peer, Y.; Rokhsar, D. *Science* **2006**, *313*, 1596.

(28) The Arabidopsis Genome Initiative. *Nature* **2000**, *408*, 796.

(29) International Rice Genome Sequencing Project. *Nature* **2005**, *436*, 793.

(30) Kumar, A.; Bennetzen, J. L. *Annu. Rev. Genet.* **1999**, *33*, 479.

(31) Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; Hoskins, R. A.; Galle, R. F.; George, R. A.; Lewis, S. E.; Richards, S.; Ashburner, M.; Henderson, S. N.; Sutton, G. G.; Wortman, J. R.; Yandell, M. D.; Zhang, Q.; Chen, L. X.; Brandon, R. C.; Rogers, Y. H.; Blazej, R. G.; Champe, M.; Pfeiffer, B. D.; Wan, K. H.; Doyle, C.; Baxter, E. G.; Helt, G.; Nelson, C. R.; Gabor, G. L.; Abril, J. F.; Agbayani, A.; An, H. J.; Andrews-Pfannkoch, C.; Baldwin, D.; Ballew, R. M.; Basu, A.; Baxendale, J.; Bayraktaroglu, L.; Beasley, E. M.; Beeson, K. Y.; Benos, P. V.; Berman, B. P.; Bhandari, D.; Bolshakov, S.; Borkova, D.; Botchan, M. R.; Bouck, J.; Brokstein, P.; Brottier, P.; Burtis, K. C.; Busam, D. A.; Butler, H.; Cadieu, E.; Center, A.; Chandra, I.; Cherry, J. M.; Cawley, S.; Dahlke, C.; Davenport, L. B.; Davies, P.; de Pablos, B.; Delcher, A.; Deng, Z.; Mays, A. D.; Dew, I.; Dietz, S. M.; Dodson, K.; Doup, L. E.; Downes, M.; Dugan-Rocha, S.; Dunkov, B. C.; Dunn, P.; Durbin, K. J.; Evangelista, C. C.; Ferraz, C.; Ferriera, S.; Fleischmann, W.; Fosler, C.; Gabrielian, A. E.; Garg, N. S.; Gelbart, W. M.; Glasser, K.; Glodek, A.; Gong, F.; Gorrell, J. H.; Gu, Z.; Guan, P.; Harris, M.; Harris, N. L.; Harvey, D.; Heiman, T. J.; Hernandez, J. R.; Houck, J.; Hostin, D.; Houston, K. A.; Howland, T. J.; Wei, M. H.; Ibegwam, C.; Jalali, M.; Kalush, F.; Karpen, G. H.; Ke, Z.; Kennison, J. A.; Ketchum, K. A.; Kimmel, B. E.; Kodira, C. D.; Kraft, C.; Kravitz, S.; Kulp, D.; Lai, Z.; Lasko, P.; Lei, Y.; Levitsky, A. A.; Li, J.; Li, Z.; Liang, Y.; Lin, X.; Liu, X.; Mattei, B.; McIntosh, T. C.; McLeod, M. P.; McPherson, D.; Merkulov, G.; Milshina, N. V.; Mobarry, C.; Morris, J.; Moshrefi, A.; Mount, S. M.; Moy, M.; Murphy, B.; Murphy, L.; Muzny, D. M.; Nelson, D. L.; Nelson, D. R.; Nelson, K. A.; Nixon, K.; Nusskern, D. R.; Pacleb, J. M.; Palazzolo, M.; Pittman, G. S.; Pan, S.; Pollard, J.; Puri, V.; Reese, M. G.; Reinert, K.; Remington, K.; Saunders, R. D.; Scheeler, F.; Shen, H.; Shue, B. C.; Siden-Kiamos, I.; Simpson, M.; Skupski, M. P.; Smith, T.; Spier, E.; Spradling, A. C.; Stapleton, M.; Strong, R.; Sun, E.; Svirskas, R.; Tector, C.; Turner, R.; Venter, E.; Wang, A. H.; Wang, X.; Wang, Z. Y.; Wassarman, D. A.; Weinstock, G. M.; Weissenbach, J.; Williams, S. M.; Woodage, T.; Worley, K. C.; Wu, D.; Yang, S.; Yao, Q. A.; Ye, J.; Yeh, R. F.; Zaveri, J. S.; Zhan, M.; Zhang, G.; Zhao, Q.; Zheng, L.; Zheng, X. H.; Zhong, F. N.; Zhong, W.; Zhou, X.; Zhu, S.; Zhu, X.; Smith, H. O.; Gibbs, R. A.; Myers, E. W.; Rubin, G. M.; Venter, J. C. *Science* **2000**, *287*, 2185.

(32) Shizuya, H.; Birren, B.; Kim, U. J.; Mancino, V.; Slepak, T.; Tachiiri, Y.; Simon, M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8794.

(33) Nelson, W. M.; Dvorak, J.; Luo, M. C.; Messing, J.; Wing, R. A.; Soderlund, C. *Genomics* **2007**, *89*, 160.

(34) Venter, J. C.; Smith, H. O.; Hood, L. *Nature* **1996**, *381*, 364.

(35) Marra, M. A.; Kucaba, T. A.; Dietrich, N. L.; Green, E. D.; Brownstein, B.; Wilson, R. K.; McDonald, K. M.; Hillier, L. W.; McPherson, J. D.; Waterston, R. H. *Genome Res.* **1997**, *7*, 1072.

(36) Luo, M. C.; Thomas, C.; You, F. M.; Hsiao, J.; Ouyang, S.; Buell, C. R.; Malandro, M.; McGuire, P. E.; Anderson, O. D.; Dvorak, J. *Genomics* **2003**, *82*, 378.

(37) Ding, Y.; Johnson, M. D.; Chen, W. Q.; Wong, D.; Chen, Y. J.; Benson, S. C.; Lam, J. Y.; Kim, Y. M.; Shizuya, H. *Genomics* **2001**, *74*, 142.

(38) Engler, F. W.; Hatfield, J.; Nelson, W.; Soderlund, C. A. *Genome Res.* **2003**, *13*, 2152.

(39) Soderlund, C.; Humphray, S.; Dunham, A.; French, L. *Genome Res.* **2000**, *10*, 1772.

(40) Soderlund, C.; Longden, I.; Mott, R. *Comput. Appl. Biosci.* **1997**, *13*, 523.

(41) Anderson, S. *Nucleic Acids Res.* **1981**, *9*, 3015.

(42) Deininger, P. L. *Anal. Biochem.* **1983**, *129*, 216.

(43) Batzoglou, S.; Jaffe, D. B.; Stanley, K.; Butler, J.; Gnerre, S.; Mauceli, E.; Berger, B.; Mesirov, J. P.; Lander, E. S. *Genome Res.* **2002**, *12*, 177.

(44) Huang, X.; Madan, A. *Genome Res.* **1999**, *9*, 868.

(45) Sutton, G.; White, O.; Adams, M.; Kerlavage, A. R. *Genome Sci. Tech.* **1995**, *1*, 9.

(46) Myers, E. W.; Sutton, G. G.; Delcher, A. L.; Dew, I. M.; Fasulo, D. P.; Flanigan, M. J.; Kravitz, S. A.; Mobarry, C. M.; Reinert, K. H.; Remington, K. A.; Anson, E. L.; Bolanos, R. A.; Chou, H. H.; Jordan, C. M.; Halpern, A. L.; Lonardi, S.; Beasley, E. M.; Brandon, R. C.; Chen, L.; Dunn, P. J.; Lai, Z.; Liang, Y.; Nusskern, D. R.; Zhan, M.; Zhang, Q.; Zheng, X.; Rubin, G. M.; Adams, M. D.; Venter, J. C. *Science* **2000**, *287*, 2196.

(47) Birren, B.; Green, E. D.; Klapholz, S.; Myers, R. M.; Roskams, J. *Genome Analysis. A Laboratory Manual. Analyzing DNA*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1997.

(48) Goff, S. A.; Ricke, D.; Lan, T. H.; Presting, G.; Wang, R.; Dunn, M.; Glazebrook, J.; Sessions, A.; Oeller, P.; Varma, H.; Hadley, D.; Hutchison, D.; Martin, C.; Katagiri, F.; Lange, B. M.; Moughamer, T.; Xia, Y.; Budworth, P.; Zhong, J.; Miguel, T.; Paszkowski, U.; Zhang, S.; Colbert, M.; Sun, W. L.; Chen, L.; Cooper, B.; Park, S.; Wood, T. C.; Mao, L.; Quail, P.; Wing, R.; Dean, R.; Yu, Y.; Zharkikh, A.; Shen, R.; Sahasrabudhe, S.; Thomas, A.; Cannings, R.; Gutin, A.; Pruss, D.; Reid, J.; Tavtigian, S.; Mitchell, J.; Eldredge, G.; Scholl, T.; Miller, R. M.; Bhatnagar, S.; Adey, N.; Rubano, T.; Tusneem, N.; Robinson, R.; Feldhaus, J.; Macalma, T.; Oliphant, A.; Briggs, S. *Science* **2002**, *296*, 92.

(49) Yu, J.; Hu, S.; Wang, J.; Wong, G. K.; Li, S.; Liu, B.; Deng, Y.; Dai, L.; Zhou, Y.; Zhang, X.; Cao, M.; Liu, J.; Sun, J.; Tang, J.; Chen, Y.; Huang, X.; Lin, W.; Ye, C.; Tong, W.; Cong, L.; Geng, J.; Han, Y.; Li, L.; Li, W.; Hu, G.; Li, J.; Liu, Z.; Qi, Q.; Li, T.; Wang, X.; Lu, H.; Wu, T.; Zhu, M.; Ni, P.; Han, H.; Dong, W.; Ren, X.; Feng, X.; Cui, P.; Li, X.; Wang, H.; Xu, X.; Zhai, W.; Xu, Z.; Zhang, J.; He, S.; Xu, J.; Zhang, K.; Zheng, X.; Dong, J.; Zeng, W.; Tao, L.; Ye, J.; Tan, J.; Chen, X.; He, J.; Liu, D.; Tian, W.; Tian, C.; Xia, H.; Bao, Q.; Li, G.; Gao, H.; Cao, T.; Zhao, W.; Li, P.; Chen, W.; Zhang, Y.; Hu, J.; Liu, S.; Yang, J.; Zhang, G.; Xiong, Y.; Li, Z.; Mao, L.; Zhou, C.; Zhu, Z.; Chen, R.; Hao, B.; Zheng, W.; Chen, S.; Guo, W.; Tao, M.; Zhu, L.; Yuan, L.; Yang, H. *Science* **2002**, *296*, 79.

(50) Thorstenson, Y. R.; Hunicke-Smith, S. P.; Oefner, P. J.; Davis, R. W. *Genome Res.* **1998**, *8*, 848.

(51) Wild, J.; Hradecna, Z.; Szybalski, W. *Plasmid* **2001**, *45*, 142.

(52) Hradecna, Z.; Wild, J.; Szybalski, W. *Microb. Comp. Genomics* **1998**, *3*, 58.

(53) Meyers, B. C.; Tingey, S. V.; Morgante, M. *Genome Res.* **2001**, *11*, 1660.

(54) Li, W.; Zhang, P.; Fellers, J. P.; Friebe, B.; Gill, B. S. *Plant J.* **2004**, *40*, 500.

(55) SanMiguel, P.; Gaut, B. S.; Tikhonov, A.; Nakajima, Y.; Bennetzen, J. L. *Nat. Genet.* **1998**, *20*, 43.

(56) SanMiguel, P.; Tikhonov, A.; Jin, Y. K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P. S.; Edwards, K. J.; Lee, M.; Avramova, Z.; Bennetzen, J. L. *Science* **1996**, *274*, 765.

(57) Shirasu, K.; Schulman, A. H.; Lahaye, T.; Schulze-Lefert, P. *Genome Res.* **2000**, *10*, 908.

(58) Fu, H.; Park, W.; Yan, X.; Zheng, Z.; Shen, B.; Dooner, H. K. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 8903.

(59) Dubcovsky, J.; Ramakrishna, W.; SanMiguel, P. J.; Busso, C. S.; Yan, L.; Shiloff, B. A.; Bennetzen, J. L. *Plant Physiol.* **2001**, *125*, 1342.

(60) Wicker, T.; Stein, N.; Albar, L.; Feuillet, C.; Schlagenhauf, E.; Keller, B. *Plant J.* **2001**, *26*, 307.

(61) Tikhonov, A. P.; SanMiguel, P. J.; Nakajima, Y.; Gorenstein, N. M.; Bennetzen, J. L.; Avramova, Z. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 7409.

(62) Huang, S.; Sirikhachornkit, A.; Su, X.; Faris, J.; Gill, B.; Haselkorn, R.; Gornicki, P. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8133.

(63) Janda, J.; Bartos, J.; Safar, J.; Kubalakova, M.; Valarik, M.; Cihalikova, J.; Simkova, H.; Caboche, M.; Sourdille, P.; Bernard, M.; Chalhoub, B.; Dolezel, J. *Theor. Appl. Genet.* **2004**, *109*, 1337.

(64) Vrana, J.; Kubalakova, M.; Simkova, H.; Cihalikova, J.; Lysak, M. A.; Dolezel, J. *Genetics* **2000**, *156*, 2033.

(65) Kubalakova, M.; Vrana, J.; Cihalikova, J.; Simkova, H.; Dolezel, J. *Theor. Appl. Genet.* **2002**, *104*, 1362.

(66) Safar, J.; Bartos, J.; Janda, J.; Bellec, A.; Kubalakova, M.; Valarik, M.; Pateyron, S.; Weiserova, J.; Tuskova, R.; Cihalikova, J.; Vrana, J.; Simkova, H.; Faivre-Rampant, P.; Sourdille, P.; Caboche, M.; Bernard, M.; Dolezel, J.; Chalhoub, B. *Plant J.* **2004**, *39*, 960.

(67) Endo, T.; Gill, B. *J. Hered.* **1996**, *87*, 295.

(68) Suchankova, P.; Kubalakova, M.; Kovarova, P.; Bartos, J.; Cihalikova, J.; Molnar-Lang, M.; Endo, T. R.; Dolezel, J. *Theor. Appl. Genet.* **2006**, *113*, 651.

(69) Putney, S. D.; Herlihy, W. C.; Schimmel, P. *Nature* **1983**, *302*, 718.

(70) Brenner, S. *Ciba. Found. Symp.* **1990**, *149*, 6.

(71) Adams, M. D.; Kerlavage, A. R.; Fleischmann, R. D.; Fuldner, R. A.; Bult, C. J.; Lee, N. H.; Kirkness, E. F.; Weinstock, K. G.; Gocayne, J. D.; White, O.; Sutton, G.; Blake, J. A.; Brandon, R. C.; Chiu, M.; Clayton, R. A.; Cline, R. T.; Cotton, M. D.; Earle-Hughes, J.; Fine, L. D.; FitzGerald, L. M.; FitzHugh, W. M.; Fritchman, J. L.; Geoghagen, N. S. M.; Glodek, A.; Gnehm, C. L.; Hanna, M. C.; Hedblom, E.; Hinkle, P. S., Jr.; Kelley, J. M.; Klimek, K. M.; Kelley, J. C.; Liu, L.; Marmaros, S. M.; Merrick, J. M.; Moreno-Palanques, R. F.; McDonald, L. A.; Nguyen, D. T.; Pellegrino, S. M.; Phillips, C. A.; Ryder, S. E.; Scott, J. L.; Saudek, D. M.; Shirley, R.; Small, K. V.; Spriggs, T. A.; Utterback, T. R.; Weidman, J. F.; Li, Y.; Barthlow, R.; Bednarik, D. P.; Cao, L.; Cepeda, M. A.; Coleman, T. A.; Collins, E.; Dimke, D.; Feng, P.; Ferrie, A.; Fischer, C.; Hastings, G. A.; He, W.; Hu, J.; Huddleston, K. A.; Greene, J. M.; Gruber, J.; Hudson, P.; Kim, A.; Kozak, D. L.; Kunsch, C.; Ji, H.; Li, H.; Meissner, P. S.; Olsen, H.; Raymond, L.; Wei, Y.; Wing, J.; Xu, C.; Yu, G.; Ruben, S. M.; Dillon, P. J.; Fannon, M. R.; Rosen, C. A.; Haseltine, W. A.; Fields, C.; Fraser, C. M.; Venter, J. C. *Nature* **1995**, *377*, 3.

(72) Adams, M. D.; Kelley, J. M.; Gocayne, J. D.; Dubnick, M.; Polymeropoulos, M. H.; Xiao, H.; Merril, C. R.; Wu, A.; Olde, B.; Moreno, R. F.; Kerlavage, R. F.; McCombie, W. R.; Venter, J. C. *Science* **1991**, *252*, 1651.

(73) Adams, M. D.; Dubnick, M.; Kerlavage, A. R.; Moreno, R.; Kelley, J. M.; Utterback, T. R.; Nagle, J. W.; Fields, C.; Venter, J. C. *Nature* **1992**, *355*, 632.

(74) Barbazuk, W. B.; Bedell, J. A.; Rabinowicz, P. D. *Bioessays* **2005**, *27*, 839.

(75) Bonaldo, M. F.; Lennon, G.; Soares, M. B. *Genome Res.* **1996**, *6*, 791.

(76) Ewing, R. M.; Kahla, A. B.; Poirot, O.; Lopez, F.; Audic, S.; Claverie, J. M. *Genome Res.* **1999**, *9*, 950.

(77) Fernandes, J.; Brendel, V.; Gai, X.; Lal, S.; Chandler, V. L.; Elumalai, R. P.; Galbraith, D. W.; Pierson, E. A.; Walbot, V. *Plant Physiol.* **2002**, *128*, 896.

(78) Patanjali, S. R.; Parimoo, S.; Weissman, S. M. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 1943.

(79) Ko, M. S. *Nucleic Acids Res.* **1990**, *18*, 5705.

(80) Soares, M. B.; Bonaldo, M. F.; Jelene, P.; Su, L.; Lawton, L.; Efstratiadis, A. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 9228.

(81) Britten, R. J.; Graham, D. E.; Neufeld, B. R. *Methods Enzymol.* **1974**, *29*, 363.

(82) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Edgar, R.; Federhen, S.; Geer, L. Y.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Ostell, J.; Miller, V.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. *Nucleic Acids Res.* **2006**.

(83) Childs, K. L.; Hamilton, J. P.; Zhu, W.; Ly, E.; Cheung, F.; Wu, H.; Rabinowicz, P. D.; Town, C. D.; Buell, C. R.; Chan, A. P. *Nucleic Acids Res.* **2006**.

(84) Lee, Y.; Tsai, J.; Sunkara, S.; Karamycheva, S.; Pertea, G.; Sultana, R.; Antonescu, V.; Chan, A.; Cheung, F.; Quackenbush, J. *Nucleic Acids Res.* **2005**, *33*, D71.

(85) Dong, Q.; Lawrence, C. J.; Schlueter, S. D.; Wilkerson, M. D.; Kurtz, S.; Lushbough, C.; Brendel, V. *Plant Physiol.* **2005**, *139*, 610.

(86) Emrich, S. J.; Barbazuk, W. B.; Li, L.; Schnable, P. S. *Genome Res.* **2007**, *17*, 69.

(87) Haas, B. J.; Delcher, A. L.; Mount, S. M.; Wortman, J. R.; Smith, R. K., Jr.; Hannick, L. I.; Maiti, R.; Ronning, C. M.; Rusch, D. B.; Town, C. D.; Salzberg, S. L.; White, O. *Nucleic Acids Res.* **2003**, *31*, 5654.

(88) Brendel, V.; Xing, L.; Zhu, W. *Bioinformatics* **2004**, *20*, 1157.

(89) Wheelan, S. J.; Church, D. M.; Ostell, J. M. *Genome Res.* **2001**, *11*, 1952.

(90) Rossignol, J. L.; Faugeron, G. *Experientia* **1994**, *50*, 307.

(91) Selker, E. U. *Annu. Rev. Genet.* **1990**, *24*, 579.

(92) Lyko, F.; Ramsahoye, B. H.; Jaenisch, R. *Nature* **2000**, *408*, 538.

(93) Tweedie, S.; Charlton, J.; Clark, V.; Bird, A. *Mol. Cell Biol.* **1997**, *17*, 1469.

(94) Walsh, C. P.; Chaillet, J. R.; Bestor, T. H. *Nat. Genet.* **1998**, *20*, 116.

(95) Rabinowicz, P. D.; Palmer, L. E.; May, B. P.; Hemann, M. T.; Lowe, S. W.; McCombie, W. R.; Martienssen, R. A. *Genome Res.* **2003**, *13*, 2658.

(96) Cross, S. H.; Bird, A. P. *Curr. Opin. Genet. Dev.* **1995**, *5*, 309.

(97) Gruenbaum, Y.; Naveh-Many, T.; Cedar, H.; Razin, A. *Nature* **1981**, *292*, 860.

(98) Meyer, P.; Niedenhof, I.; ten Lohuis, M. *EMBO J.* **1994**, *13*, 2084.

(99) Martienssen, R. *Trends Genet.* **1998**, *14*, 263.

(100) Chandler, V. L.; Walbot, V. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 1767.

(101) Flavell, R. B. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 3490.

(102) Bennetzen, J. L.; Schrick, K.; Springer, P. S.; Brown, W. E.; SanMiguel, P. *Genome* **1994**, *37*, 565.

(103) Lippman, Z.; Gendrel, A. V.; Black, M.; Vaughn, M. W.; Dedhia, N.; McCombie, W. R.; Lavine, K.; Mittal, V.; May, B.; Kasschau, K. D.; Carrington, J. C.; Doerge, R. W.; Colot, V.; Martienssen, R. *Nature* **2004**, *430*, 471.

(104) Zhang, X.; Yazaki, J.; Sundaresan, A.; Cokus, S.; Chan, S. W.; Chen, H.; Henderson, I. R.; Shinn, P.; Pellegrini, M.; Jacobsen, S. E.; Ecker, J. R. *Cell* **2006**, *126*, 1189.

(105) Zilberman, D.; Gehring, M.; Tran, R. K.; Ballinger, T.; Henikoff, S. *Nat. Genet.* **2007**, *39*, 61.

(106) Dila, D.; Sutherland, E.; Moran, L.; Slatko, B.; Raleigh, E. A. *J. Bacteriol.* **1990**, *172*, 4888.

(107) Raleigh, E. A.; Wilson, G. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 9070.

(108) Sutherland, E.; Coe, L.; Raleigh, E. A. *J. Mol. Biol.* **1992**, *225*, 327.

(109) Rabinowicz, P. D. *Methods Mol. Biol.* **2003**, *236*, 21.

(110) Rabinowicz, P. D.; Schutz, K.; Dedhia, N.; Yordan, C.; Parnell, L. D.; Stein, L.; McCombie, W. R.; Martienssen, R. A. *Nat. Genet.* **1999**, *23*, 305.

(111) Palmer, L. E.; Rabinowicz, P. D.; O'Shaughnessy, A. L.; Balija, V. S.; Nascimento, L. U.; Dike, S.; de la Bastide, M.; Martienssen, R. A.; McCombie, W. R. *Science* **2003**, *302*, 2115.

(112) Whitelaw, C. A.; Barbazuk, W. B.; Pertea, G.; Chan, A. P.; Cheung, F.; Lee, Y.; Zheng, L.; van Heeringen, S.; Karamycheva, S.; Bennetzen, J. L.; SanMiguel, P.; Lakey, N.; Bedell, J.; Yuan, Y.; Budiman, M. A.; Resnick, A.; Van, Aken, S.; Utterback, T.; Riedmuller, S.; Williams, M.; Feldblyum, T.; Schubert, K.; Beachy, R.; Fraser, C. M.; Quackenbush, J. *Science* **2003**, *302*, 2118.

(113) Singer, T.; Yordan, C.; Martienssen, R. A. *Genes Dev.* **2001**, *15*, 591.

(114) Miura, A.; Yonebayashi, S.; Watanabe, K.; Toyama, T.; Shimada, H.; Kakutani, T. *Nature* **2001**, *411*, 212.

(115) Kato, M.; Miura, A.; Bender, J.; Jacobsen, S. E.; Kakutani, T. *Curr. Biol.* **2003**, *13*, 421.

(116) Lippman, Z.; May, B.; Yordan, C.; Singer, T.; Martienssen, R. *PLoS Biol.* **2003**, *1*, E67.

(117) Bedell, J. A.; Budiman, M. A.; Nunberg, A.; Citek, R. W.; Robbins, D.; Jones, J.; Flick, E.; Rholfing, T.; Fries, J.; Bradford, K.; McMenamy, J.; Smith, M.; Holeman, H.; Roe, B. A.; Wiley, G.; Korf, I. F.; Rabinowicz, P. D.; Lakey, N.; McCombie, W. R.; Jeddeloh, J. A.; Martienssen, R. A. *PLoS Biol.* **2005**, *3*, e13.

(118) Jones-Rhoades, M. W.; Bartel, D. P.; Bartel, B. *Annu. Rev. Plant. Biol.* **2006**, *57*, 19.

(119) Rabinowicz, P. D.; Citek, R.; Budiman, M. A.; Nunberg, A.; Bedell, J. A.; Lakey, N.; O'Shaughnessy, A. L.; Nascimento, L. U.; McCombie, W. R.; Martienssen, R. A. *Genome Res.* **2005**, *15*, 1431.

(120) Wang, Y.; van der Hoeven, R. S.; Nielsen, R.; Mueller, L. A.; Tanksley, S. D. *Theor. Appl. Genet.* **2005**, *112*, 72.

(121) Zhou, Y.; Bui, T.; Auckland, L. D.; Williams, C. G. *Genome* **2002**, *45*, 91.

(122) Kinlaw, C. S.; Neale, D. B. *Trends Plant Sci.* **1997**, *2*, 356.

(123) Krutovsky, K. V.; Troggio, M.; Brown, G. R.; Jermstad, K. D.; Neale, D. B. *Genetics* **2004**, *168*, 447.

(124) Lee, I. M.; Davis, R. E.; Gundersen-Rindal, D. E. *Annu. Rev. Microbiol.* **2000**, *54*, 221.

(125) Britten, R. J.; Kohne, D. E. *Science* **1968**, *161*, 529.

(126) Bennetzen, J. L.; Chandler, V. L.; Schnable, P. *Plant Physiol.* **2001**, *127*, 1572.

(127) Peterson, D. G.; Schulze, S. R.; Sciara, E. B.; Lee, S. A.; Bowers, J. E.; Nagel, A.; Jiang, N.; Tibbitts, D. C.; Wessler, S. R.; Paterson, A. H. *Genome Res.* **2002**, *12*, 795.

(128) Yuan, Y.; SanMiguel, P. J.; Bennetzen, J. L. *Plant J.* **2003**, *34*, 249.

(129) Lamoureux, D.; Peterson, D.; Li, W.; Fellers, J. P.; Gill, B. S. *Genome* **2005**, *48*, 1120.

(130) Elsik, C. G.; Williams, C. G. *Mol. Gen. Genet.* **2000**, *264*, 47.

(131) Fu, Y.; Hsia, A. P.; Guo, L.; Schnable, P. S. *Plant Physiol.* **2004**, *135*, 2040.

(132) Wicker, T.; Robertson, J. S.; Schulze, S. R.; Feltus, F. A.; Magrini, V.; Morrison, J. A.; Mardis, E. R.; Wilson, R. K.; Peterson, D. G.; Paterson, A. H.; Ivarie, R. *Genome Res.* **2005**, *15*, 126.

(133) Springer, N. M.; Xu, X.; Barbazuk, W. B. *Plant Physiol.* **2004**, *136*, 3023.

(134) Lander, E. S.; Waterman, M. S. *Genomics* **1988**, *2*, 231.

(135) Wendl, M. C.; Barbazuk, W. B. *BMC Bioinformatics* **2005**, *6*, 245.

(136) Bruggmann, R.; Bharti, A. K.; Gundlach, H.; Lai, J.; Young, S.; Pontaroli, A. C.; Wei, F.; Haberer, G.; Fuks, G.; Du, C.; Raymond, C.; Estep, M. C.; Liu, R.; Bennetzen, J. L.; Chan, A. P.; Rabinowicz, P. D.; Quackenbush, J.; Barbazuk, W. B.; Wing, R. A.; Birren, B.; Nusbaum, C.; Rounsley, S.; Mayer, K. F.; Messing, J. *Genome Res.* **2006**, *16*, 1241.

(137) Jiang, N.; Bao, Z.; Zhang, X.; Eddy, S. R.; Wessler, S. R. *Nature* **2004**, *431*, 569.
(138) Morgante, M.; Brunner, S.; Pea, G.; Fengler, K.; Zuccolo, A.; Rafalski, A. *Nat. Genet.* **2005**, *37*, 997.
(139) Brunner, S.; Pea, G.; Rafalski, A. *Plant J.* **2005**, *43*, 799.
(140) Fu, Y.; Emrich, S. J.; Guo, L.; Wen, T. J.; Ashlock, D. A.; Aluru, S.; Schnable, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12282.
(141) Emrich, S. J.; Aluru, S.; Fu, Y.; Wen, T. J.; Narayanan, M.; Guo, L.; Ashlock, D. A.; Schnable, P. S. *Bioinformatics* **2004**, *20*, 140.
(142) Chan, A. P.; Pertea, G.; Cheung, F.; Lee, D.; Zheng, L.; Whitelaw, C.; Pontaroli, A. C.; SanMiguel, P.; Yuan, Y.; Bennetzen, J.; Barbazuk, W. B.; Quackenbush, J.; Rabinowicz, P. D. *Nucleic Acids Res.* **2006**, *34*, D771.
(143) Burr, B.; Burr, F. A.; Thompson, K. H.; Albertson, M. C.; Stuber, C. W. *Genetics* **1988**, *118*, 519.
(144) Emberton, J.; Ma, J.; Yuan, Y.; SanMiguel, P.; Bennetzen, J. L. *Genome Res.* **2005**, *15*, 1441.
(145) Rabinowicz, P. D.; Bennetzen, J. L. *Curr. Opin. Plant Biol.* **2006**, *9*, 149.
(146) Yuan, Y.; SanMiguel, P. J.; Bennetzen, J. L. *Genome Res.* **2002**, *12*, 1345.
(147) Yu, C.; Li, Z. *Anal. Biochem.* **2006**, *359*, 141.
(148) Lisch, D. *Trends Plant Sci.* **2002**, *7*, 498.
(149) Fedoroff, N. V. *Genes Cells* **1999**, *4*, 11.
(150) Kunze, R. *Curr. Top. Microbiol. Immunol.* **1996**, *204*, 161.
(151) Hanley, S.; Edwards, D.; Stevenson, D.; Haines, S.; Hegarty, M.; Schuch, W.; Edwards, K. J. *Plant J.* **2000**, *23*, 557.
(152) Alleman, M.; Freeling, M. *Genetics* **1986**, *112*, 107.
(153) Wessler, S. R. *Science* **1988**, *242*, 399.
(154) May, B. P.; Liu, H.; Vollbrecht, E.; Senior, L.; Rabinowicz, P. D.; Roh, D.; Pan, X.; Stein, L.; Freeling, M.; Alexander, D.; Martienssen, R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11541.
(155) Bai, L.; Singh, M.; Pitt, L.; Sweeney, M.; Brutnell, T. P. *Genetics* **2007**, *175*, 981.
(156) McCarty, D. R.; Settles, A. M.; Suzuki, M.; Tan, B. C.; Latshaw, S.; Porch, T.; Robin, K.; Baier, J.; Avigne, W.; Lai, J.; Messing, J.; Koch, K. E.; Hannah, L. C. *Plant J.* **2005**, *44*, 52.
(157) Raizada, M. N.; Nan, G. L.; Walbot, V. *Plant Cell* **2001**, *13*, 1587.
(158) Guigo, R.; Flicek, P.; Abril, J. F.; Reymond, A.; Lagarde, J.; Denoeud, F.; Antonarakis, S.; Ashburner, M.; Bajic, V. B.; Birney, E.; Castelo, R.; Eyras, E.; Ucla, C.; Gingeras, T. R.; Harrow, J.; Hubbard, T.; Lewis, S. E.; Reese, M. G. *Genome Biol.* **2006**, *7* (Suppl 1), S2 1.

CR0682960