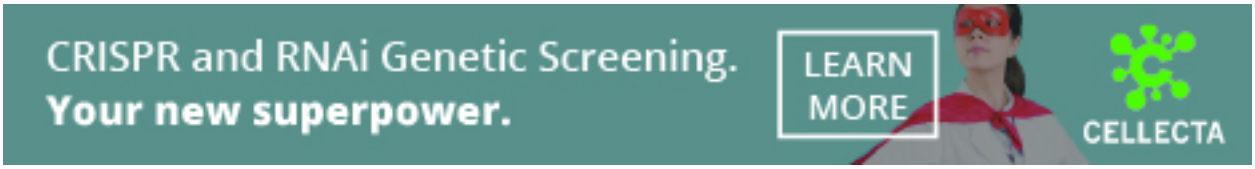
Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries

John Emberton, Jianxin Ma, Yinan Yuan, et al.

Genome Res. 2005 15: 1441-1446
Access the most recent version at doi:10.1101/gr.3362105



Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries

John Emberton,[1,4,6] Jianxin Ma,[3,6] Yinan Yuan,[1,5] Phillip SanMiguel,[2] and Jeffrey L. Bennetzen[1,3,7]

[1]*Department of Biological Sciences, and* [2]*Genomics Core Facility, Purdue University, West Lafayette, Indiana 47907, USA;* [3]*Department of Genetics, University of Georgia, Athens, Georgia 30602, USA*

A new technology was developed to assist gene-enrichment sequencing of any complex plant genome, employing maize as the test organism. Hypomethylated partial restriction (HMPR) libraries were constructed by using independent partial digestions with methylation-sensitive restriction enzymes HpaII (5′-CCGG-3′) and HpyCH4IV (5′-ACGT-3′). Fragments of 1–4 kb were purified and cloned, followed by sequence analysis of >2000 clones from 10 separate libraries. Organellar clones comprised ~10% of each library but were useful in showing that no chimeric clones were generated and that digestion efficiencies were 10%–25% in different libraries. Four separate HMPR libraries, analyzed in detail, exhibited very similar degrees of gene enrichment and repeat depletion. Known gene homologies were found in ~25% of the HMPR clones, compared with <4% in clones from a fully random set of unfiltered maize shotgun sequences. This six- to sevenfold enrichment for genes compares favorably with the best previous gene enrichment techniques in maize, High Cot analysis and methylation filtration. Compared with High Cot and methylation filtration, HMPR is exceptional in depleting retrotransposons' content to the lowest level yet observed (<5%, compared with >70% for unfiltered maize sequences) and in providing an unmatched enrichment for the "unknown" sequences that contain promoters, introns, and other gene-adjacent regions.

[The sequence data from this study have been submitted to GenBank under accession nos. CW539179–CW542054.]

Although flowering plant genomes vary enormously in nuclear DNA content, from the <150 Mb of a handful of species to the ~140,000 Mb of *Fritillaria assyriaca*, most contain between 1000 and 8000 Mb per haploid genome (with a median of ~3000 Mb) scattered across anywhere from two to >300 chromosomes (Kenton et al. 1993; Leitch et al. 2005). Gene content is much more constant, for instance varying less than twofold between the large genome of barley (*Hordeum vulgare*, ~4900 Mb) and the smaller genome of rice (*Oryza sativa*, ~440 Mb) (Van Deynze et al. 1998). Some publications (Goff et al. 2002; Yu et al. 2002; Feng et al. 2003; Rice Chromosome 10 Sequencing Consortium 2003; Sasaki et al. 2002) have concluded that rice might have about twice as many genes as does *Arabidopsis thaliana* (120–150 Mb), but other studies have indicated that most of these "extra" rice genes are artifacts of annotation, leading to the current conclusion that rice and *Arabidopsis* contain very similar gene numbers (Bennetzen 2002b; Bennetzen et al. 2004; Jabbari et al. 2004; Ma and Bennetzen 2004). Moreover, the average size of genes does not vary >50% between the small genome of *Arabidopsis* and large genome species such as barley or cotton (Dubcovsky et al. 2001; Wendel et al. 2002). Hence, the space occupied by genes within a plant nuclear genome, the "gene space," is not nearly so variable as is the overall DNA content of the genome.

In the maize (*Zea mays*) genome, ~2400 Mb of DNA is distributed across the 10 chromosomes in a haploid nucleus. A precise gene number is not known for this species, although most speculation suggests 30,000 to 50,000 protein-encoding loci. Analysis of current gene distribution and sequence variation within the maize genome has indicated that maize is derived from an ancient tetraploid (Helentjaris et al. 1988; Gaut and Doebley 1997). The most recent studies, comparing "homoeologous" segments within the duplicated genome of maize to the orthologous regions from rice and sorghum, have shown that the diploid ancestral genomes of maize diverged ~12 million years ago (Mya), approximately the same time that both diverged from a common ancestor with sorghum (Swigonová et al. 2004). Despite its recent tetraploid origin, maize now contains <50% more genes than found in its diploid ancestors (Lai et al. 2004) because of a high rate of gene loss by the accumulation of small deletions (Ilic et al. 2003).

If maize has 30,000–50,000 genes, then the minimal gene space in maize is predicted to be 150–250 Mb because the gene density in the most gene-rich portions of the maize genome is about one gene per ~5 kb (Fu et al. 2001; Song et al. 2001). The remaining ~2200 Mb of the maize genome has been shown to be comprised primarily of several classes of mobile DNAs (SanMiguel et al. 1996; Meyers et al. 2001; Whitelaw et al. 2003), especially the long terminal repeat (LTR) retrotransposons that account for >60% of the total nuclear genome (SanMiguel and Bennetzen 1998). The amplification and progressive removal of these elements both occur at prodigious rates (Ma et al. 2004), indicating that (along with polyploidy) they are the dynamic component that determines overall genome size in most flowering plants (Bennetzen 2002a). In maize and other large genome cereals such as barley and wheat, most genes are found in small islands of one to two genes surrounded by large blocks of LTR

Present addresses: [4]Department of Psychiatry, University of California, San Diego, CA 92093; [5]Plant Biotechnology Research Center, School of Forest Resources & Environmental Science, Michigan Technological University, Houghton, MI 49931.
[6]These authors contributed equally to this work.
[7]Corresponding author.
E-mail maize@uga.edu; fax (706) 583-0972.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3362105.

retrotransposons, often arranged in a nested insertion pattern (SanMiguel et al. 1996; Wicker et al. 2001; Bennetzen et al. 2005).

Several approaches have been used to try to identify all of the genes in a complex plant genome without sequencing all of the mobile and repetitive DNAs. The bulk sequencing of cDNAs, so-called expressed sequence tag (EST) analysis, is an exceptional initial technique because it yields expressed genes that are the principal target of any gene characterization program. However, EST technology rapidly becomes unproductive because of its inefficiency in finding genes with low or rare expression. Moreover, ESTs contain only the exons present within a gene, yielding no information regarding promoters or other adjacent sequences.

In all higher plants that have been investigated, most repetitive DNAs are hypermethylated at 5′-CG-3′ and 5′-CNG-3′ residues (Gruenbaum et al. 1981; Vongs et al. 1993; Bennetzen et al. 1994). For maize, *Escherichia coli* lines that do not tolerate these methylations have been used to enrich for the genes among clones generated by the insertion of small sheared fragments of total genomic DNA (Rabinowicz et al. 1999; Palmer et al. 2003). Another gene-enrichment technology employed in maize has been to use the renaturation of total genomic DNA to normalize sequence representation in the High Cot (HC) approach (Yuan et al. 2003). The methylation filtration (MF) and HC technologies each enrich for genes more than sixfold, and each diminishes LTR retrotransposon content significantly. A pilot MF and HC analysis of the maize genome involved the generation of ~895,000 sequence reads to produce ~240 Mb of assembled sequence (http://www.tigr.org/tdb/tgi/maize). Despite its very low redundancy, this project has yielded partial sequence coverage for >95% of maize protein-encoding genes (including introns, promoters, and other associated sequences), far more than discovered by many years of public and private EST projects combined and at a tiny fraction of their cost (Springer et al. 2004).

There are two intrinsic problems associated with the MF/HC approach. First, the sequences generated in any shotgun sequencing approach, including MF/HC, are not initially localized relative to physical or genetic maps. However, several technologies have been developed to resolve this limitation, including gene-enriched sequencing of bacterial artificial chromosome (BAC) ends (Yuan et al. 2002). Second, the fragments cloned in both MF and HC technologies are small so that they can yield comprehensive coverage of gene space. If MF or HC clones are large, then the presence of a small amount of methylated DNA or repetitive DNA at the edge of a clone would cause it to be lost from the libraries generated, so that MF/HC would underrepresent the sequence of boundaries between genes and the methylated/repetitive blocks.

The essential small size of MF and HC clones comes at a price. In any sequencing project, short sequence gaps are found after the shotgun stage of sequencing because certain regions are more challenging to sequence or because they were not sampled by chance in the shotgun process. This problem is often resolved by complete sequence analysis of a subclone that covers the gap. However, when subclones are all small, there is a reasonable chance that no subclone will be found that covers a particular gap. Another problem is that the absence of large subclones guarantees that most genes will not be fully covered by any single clone. Thus, subsequent investigators that wish to study a particular gene (e.g., in transgenics) will be unable to receive the full gene as a direct byproduct of the genome sequencing project.

We report here the development of hypomethylated partial restriction (HMPR) libraries as a tool to assist both the linkage of MF/HC sequence islands and to provide large gene-space clones that can be used for finishing any genome sequencing approach. Our data indicate that HMPR technologies provide exceptional gene-space enrichment.

Results

HMPR library technology relies on the observation that most genes are unmethylated much of the time in the maize genome, while most LTR retrotransposons are methylated most of the time (Bennetzen et al. 1994; Rabinowicz et al. 1999). In maize and other higher plants, the majority of this methylation is in the form of 5-methyl cytosine in the sequences 5′-CG-3′ and 5′-CNG-3′. Many bacterial restriction enzymes are unable to digest DNA that is methylated at these sites. Cytosine methylation-sensitive restriction enzymes such as PstI were initially used in maize to clone restriction fragment length polymorphism (RFLP) probes because they are enriched for low-copy-number (e.g., genic) DNAs (Burr et al. 1988). The cloning and sequencing of maize DNA completely digested with cytosine methylation-sensitive restriction enzymes is itself a powerful technique for gene discovery (Yuan et al. 2002) but is limited by a severe lack of randomness compared with that of sheared DNAs.

HMPR libraries employ a compromise in randomness between sheared DNA and complete digestion with cytosine methylation-sensitive restriction enzymes. We have used two cytosine methylation-sensitive restriction enzymes that have 4-bp sequence specificities, HpaII (5′-CCGG-3′) and HpyCH4IV (5′-ACGT-3′). Because the maize genome has an ~47% GC content (Meyers et al. 2001), HpaII and HpyCH4IV would digest maize chromosomal DNA an average of about once every 328 bp and 258 bp, respectively, if there were no cytosine methylation and bases were randomly distributed. Because of cytosine methylation, however, most maize nuclear DNA is found on 20- to 150-kb fragments when a complete HpaII digestion is employed (Springer 1992; Bennetzen et al. 1994). These large fragments primarily represent the LTR retrotransposons' blocks, while much smaller fragments represent most or all genes. Cloning and sequencing all of these tiny fragments would yield a very comprehensive process for gene discovery. However, there are two flaws to this simple approach. First, the sequencing of small DNA fragments (those <1 kb or so) wastes the ability of automated DNA sequencing apparatuses to routinely yield 800–1200 bases of high-quality sequence information. Hence, a fourfold gene enrichment would be completely negated by sequencing 200–300 bp inserts. The second problem is that a complete digestion yields fragments that cannot be located relative to each other by sequence overlap. This problem can be partly eliminated by using multiple enzymes to construct separate libraries. In HMPR technology, both of these problems are overcome by the use of partial digestion with the cytosine methylation-sensitive restriction enzyme.

Total maize genomic DNA was isolated from immature, unfertilized maize ears in order to minimize chloroplast DNA contributions to the constructed libraries. Fragments of 1–4 kb were targeted for the final product, requiring an ~10%–25% digestion by the two restriction enzymes employed. Preliminary experiments indicated that complete digestion was observed for HpaII and HpyCH4IV even when using <0.2 U/μg of maize DNA, suggesting that manufacturers were providing more enzyme than they indicated. Hence, digestions were run at 37°C for 30 min

with 0.015 to 0.1 U of either HpaII or HpyCH4IV per microgram of DNA. The products of these digestions were then sized on an 0.8% agarose gel. Under these conditions, virtually all DNA is seen as a smear that migrates with the highest-molecular-weight markers (data not shown). Regardless of the absence of any observed DNA in these size ranges, the regions of the gel that contained 1- to 2-kb, 2- to 3-kb, ~3-kb, or 2- to 4-kb fragments (relative to adjacent size markers) were excised and extracted to construct several independent HMPR libraries for HpaII and HpyCH4IV (Table 1). In order to minimize chimeric clones, fragments were dephosphorylated and then filled and "A" tailed with *Taq* polymerase. These fragments were then inserted by topoisomerase exchange into the TOPO PCR4 vector (Shuman 1994). The resultant colonies were picked and sequenced by standard automated procedures (Yuan et al. 2003).

For HpaII, six libraries were generated with different size ranges of selected inserts. With HpyCH4IV, four libraries containing different size ranges of gel-selected inserts were generated. Table 1 shows the number of clones sequenced from each library. All clones were sequenced from both ends. Detailed annotation was undertaken for the four libraries that contained the most clones that were sequenced (GenBank accession nos. CW539179–CW542054). A subset of clones were sized on 0.8% agarose gels prior to sequence analysis to see that most inserts were in the expected size range. Sequence analysis from both ends often yielded overlaps, so this could also be used to measure the size of insertions in these libraries. In most cases, the insertions in a given library exhibited a broader size range and lower average size than targeted (Table 2), but these differences were subsequently found to have little or no effect on the gene-finding and repeat-diminishing properties of these libraries (see below). Each library contained ~10% inserts from either chloroplast or mitochondrial DNA (Table 3). This is an expected result, despite the fact that we used a chloroplast-deficient tissue as the source of our total genomic DNA. Organellar DNA is mostly or completely unmethylated in most or all maize tissues and is a standard contaminant of other methylation-based gene enrichment procedures (Rabinowicz et al. 1999). Future use of this technology will benefit from the pursuit of nuclear preparations prior to purification of DNA. We have recently produced HMPR libraries from nuclear DNA preparations (data not shown). These libraries contained 7000–23,000 clones, of which a maximum of 1.2% (average, 0.6%) of the clones were of organellar origin. The organellar clones are useful, however, because they provide an indication of the quality of the library. The availability of completed sequences for maize chloroplast and mitochondrial DNAs allows determination of the frequency of chimeric clones, a precise range for insert sizes, and the degree of partial digestion.

Out of 126 organellar inserts sequenced from both ends, none were found that contained organellar DNA at one end and nuclear genomic DNA at the other. Four different Bayesian and/or "frequentist" (Samuels and Witmer, 2003) statistical analyses (data not shown) indicated a 95% certainty that the frequency of chimerics in this approach must be <2.5%.

Organellar clones also serve as an indicator of the degree of partial digestion. For nuclear DNA, any HpaII or HpyCH4IV sites found within a sequenced clone could be caused by partial digestion or by a methylated status of that site. For an organellar clone, only partial digestion is a likely origin. Of the organellar HpaII clones that were sequenced, zero to five internal HpaII sites were found per fragment, while zero to six were found for HpyCH4IV libraries (Table 2). The degree of digestion was highly variable between libraries (Table 2; data not shown), indicating that this is a difficult factor to control. Moreover, all of the digestions were more complete (25%–100% complete, averaging ~40%) than the 10% digestion that was targeted. Interestingly, the mitochondrial fragments always exhibited less digestion than did the chloroplast DNA (data not shown), suggesting that some mitochondrial fragments may be methylated at cytosine residues.

In comparison to known genes and to repeat databases, it was found that the HpaII and HpyCH4IV HMPR libraries were highly enriched for genes and depleted in LTR retrotransposons relative to an unfiltered (UF) random maize DNA set (Table 3). Despite other variables in the properties of these libraries, the relative gene enrichment and LTR retrotransposons depletion were highly similar (Fig. 1). In fact, LTR retrotransposon removal was much more efficient for this technique than for the MF or HC technologies (Fig. 2). This is not caused by a shortage of sites for these enzymes in LTR retrotransposons. A scan of the four most abundant maize LTR retrotransposons (*Huck*, *Ji*, *Opie*, and *Cinful*) indicated 4.22 HpaII sites and 1.69 HpyCH4IV sites per kilobase, while UF whole-genome shotgun sequences exhibited a respective 3.84 and 1.88 sites per kilobase for these enzymes. The HC, MF, and HMPR technologies yield quite similar enrichment of protein-encoding gene sequences, but the HMPR technology generates far more "unknown" sequences. In many cases, these "unknown" sequences are the noncoding regions of genes.

Table 1. HMPR libraries generated

Libraries	Insert size (kb)	Digestion condition[a]	No. of clones sequenced
HpaII-1	1–2	0.05	384
HpaII-2	~3	0.05	384
HpaII-3	1–2	0.05	96
HpaII-4	1–2	0.05	96
HpaII-5	2–3	0.05	96
HpaII-6	2–4	0.1	96
HpyCH4IV-1	1–2	0.015	384
HpyCH4IV-2	2–3	0.015	384
HpyCH4IV-3	~3	0.015	96
HpyCH4IV-4	2–4	0.1	96

[a]Units of enzyme used in digestion of 1 µg DNA at 37°C for 30 min.

Discussion

In a first attempt at the development of the HMPR approach, it is impressive that the technique has proven so successful in gene enrichment and LTR retrotransposon depletion. Because the distribution of restriction enzyme sites in a genome is not truly random, the use of multiple restriction enzymes and different degrees of partial digestion will be essential to cover the entire genome. On average, one expects that a 10% digestion with a four-base specificity restriction enzyme would yield the most useful fragments in a 2- to 3-kb size range. However, some regions of the genome will contain an excess of sites for any given restriction enzyme, and perhaps for multiple restriction enzymes, due to repetitive sequences, AT richness, etc. Hence, a comprehensive genome coverage and gene discovery program would use some libraries made with very partial digestion (e.g., 1%) and some with nearly complete digestion (e.g., 50%). Although 10% digestion was targeted with each enzyme employed, actual digestion

Table 2. Partial digestion of four HMPR libraries

Libraries studied[a]	HpaII-1	HpaII-2	HpyCH4IV-1	HpyCH4IV-2
Insert size (kb)				
Selected	1.0–2.0	~3	1.0–2.0	2.0–3.0
Observed	0.2–1.7	2.8–3.8	1.0–1.7	1.7–3.0
Average	1.1	3.5	1.4	2.3
HpaII or HpyCH4IV sites/insert				
Observed	0–5	0–2	0–4	1–6
Average	0.63	1.20	1.43	3.18

[a]Plasmids containing inserts from mitochondrial DNA were used to determine the insert sizes and restriction sites.

efficiencies were 30%–100% with HpaII and 25%–100% with HpyCH4IV.

The degree of gene enrichment in the HMPR technology is quite impressive compared with other techniques that have been employed. An EST project yields close to 100% nuclear gene discovery in its very early stage (although it also uncovers transcribed transposable elements and organellar genes), but it rapidly loses efficiency because of redundant sequencing of the same highly expressed genes. This redundancy problem is significantly decreased, but far from eliminated, in normalized libraries (Soares et al. 1994; Reddy et al. 2002) or by using different tissue, developmental time, and treatment sources of RNA. HMPR analysis yields all unmethylated genic regions at levels associated with their genomic representation. Other than organellar clones, ribosomal DNA, and repetitive transposons, we found zero redundancy in the first >1400 inserts that we analyzed. Hence, it appears that HMPR technology is more than a match for EST analysis in the ability to discover genes, especially as it also yields their promoters, introns, and other adjacent sequences.

The superior success of HMPR in LTR retrotransposon removal compared with HC and, especially, MF approaches was unexpected. Both MF and HMPR rely on the undermethylation of genes and hypermethylation of LTR retrotransposons as their mode of gene enrichment. Perhaps the large size of HMPR clones may require their presence in especially gene-rich regions where LTR retrotransposons are rare. Alternatively, these results could be explained by a possible sequence specificity for DNA methylation inside LTR retrotransposons such as, for instance, a bias toward greater DNA methylation of the symmetric sites recognized by most restriction enzymes. Further analysis of the relative genomic distribution of HMPR and MF clones will be needed to investigate these possibilities.

The largest category of sequence in all three gene-enrichment technologies is classified as "unknown" by our annotation system. Unknown sequences are those that do not have identified genes, organellar genome homology, or known repeats. These unknown sequences include previously unidentified genes and the portions of genes that do not encode proteins, such as introns, promoters, 5′ leaders, and 3′ trailer sequences. The unknown sequences do not include any middle or highly repetitive DNAs. Hence, we expect that the "gene space" that should be targeted in any enrichment sequence project will be the sum of what we call genes and what we call unknown sequences. In the HMPR libraries, we find that these two categories account for 90.9% of the total sequence generated, while they account for 78.2% of HC and 77.5% of MF reads. For UF sequence, 21.2% of the sequence information falls into these two categories, suggesting that the "gene space" of maize measured by this approach is ~500 Mb.

The single greatest problem that could have befallen HMPR technology was the generation of chimeric clones. These clones would yield incorrect assignments of linked contiguous blocks and could be worse than useless for finishing genomic sequence gaps. The use of A-tailing and topoisomerase exchange cloning were predicted to completely prohibit the generation of chimeric clones, and none were detected in these first libraries. However, many more HMPR clones need to be analyzed to provide an accurate estimate of rates of chimeric clone formation. In this regard, the organellar DNA clones in the libraries are quite useful. Future efforts to minimize the presence of organellar clones by nuclear DNA preparations, use of albino tissues, etc., will undoubtedly minimize their library representation but are not likely to remove organellar clones entirely.

One assessment issue is left unresolved by this preliminary analysis. It is not yet clear how random HMPR fragments can be, nor is it yet obvious how a broad range of partial digestion conditions might be optimized. A much larger scale of library generation and clone analysis will be needed to resolve this issue. Many gene-enrichment technologies, including the analysis of ESTs, HC clones, and MF clones, exhibit excellent promise in the first clones that are generated. The efficiency of these approaches in finding the last few percentages of genes is not known, although it is clear that ESTs are weak in this aspect.

We have proven that HMPR technology has significant advantages relative to other genome analysis techniques. Compared with MF and HC analyses, HMPR generates large genomic DNA inserts that can be used for linking contiguous sequence blocks and for finishing sequence gaps. HMPR clones should also span gaps in MF and HC assemblies that are caused by small blocks of repetitive and/or methylated DNA because it is not

Table 3. Sequence compositions of HMPR, HC, MF, and UF libraries

	HMPR							
Libraries studied	HpaII-1	HpaII-2	HpyCH4IV-1	HpyCH4IV-2	Total	HC	MF	UF
No. of sequences[a]	700	736	683	679	2798	1000[b]	1000[b]	1000[b]
Known genes (%)	16.6	26.0	23.7	26.0	23.1	23.4	33.7	3.6
Known retroelements (%)	3.4	7.1	4.0	2.3	4.2	17.1	20.6	72.7
Other repetitive DNA (%)	4.9	4.1	2.2	4.9	4.0	4.7	1.9	5.9
Organellar DNA (%)[c]	10.0	9.4	11.4	10.9	10.3	0.1	0.2	0.2
Unknown sequences (%)	65.1	53.5	58.7	55.9	58.3	54.7	43.6	17.6

[a]Trimmed sequences with high-quality bases >150.
[b]One thousand sequences randomly sampled from 306,557 HC clones, 281,669 MF clones, and 17,679 UF clones.
[c]Chloroplast and mitochondrial DNA.

affected by the presence of repeats or methylation inside the cloned fragment. Compared with EST analysis, HMPR technology shows less redundancy and recovers introns, promoters, and other adjacent sequences that cDNA analysis will miss. Compared with full genome shotgun sequencing, HMPR technology enriches for the genes that are the principal target of genome analysis. Hence, we feel that HMPR is a useful addition to the genome analysis toolkit.

Methods

Genomic DNA isolation

Seeds from maize inbred B73 were obtained from the U.S. Department of Agricultural Research Service, Plant Introduction Station. Genomic DNA was extracted from immature ears following the method previously described by Saghai-Maroof et al. (1984).

DNA partial digestion, size selection, and HMPR library construction

Partial digestions of 20 µg of genomic DNA were performed in 500 µL volumes with serially diluted restriction enzymes (New England Biolabs), at 37°C for 30 min, otherwise as per the manufacturer's instruction. Digestions were terminated by adding 50 µL of 0.5 M EDTA (pH 8.0). The digested fragments were dephosphorylated with shrimp alkaline phosphatase, then filled and A tailed with *Taq* polymerase. Next, the fragments were fractionated by agarose gel electrophoresis. The desired fragments were excised from the gel, recovered using the QIAEX II Gel Extraction Kit (QIAGEN Sciences), and inserted into pCR4TOPO using the Invitrogen TA cloning system. The constructed plasmids were electroporated into ElectroMax DH10B competent cells (Invitrogen/Life Technologies). HMPR libraries with as many as 28,000 clones per library have been generated by this technique, and libraries with 5000–25,000 clones are routinely produced (data not shown).

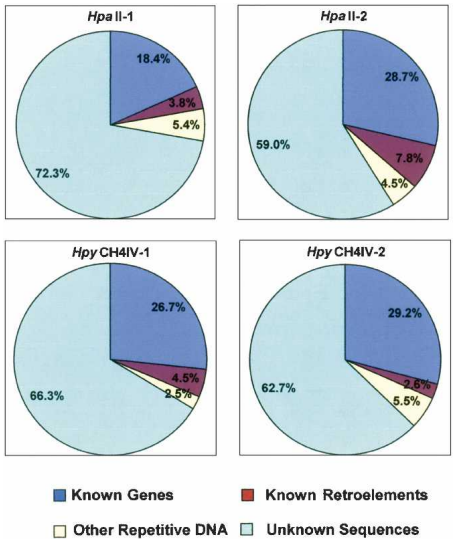


Figure 1. Nuclear sequence composition of four HMPR libraries. Organellar sequences were removed from the data set prior to the analysis.

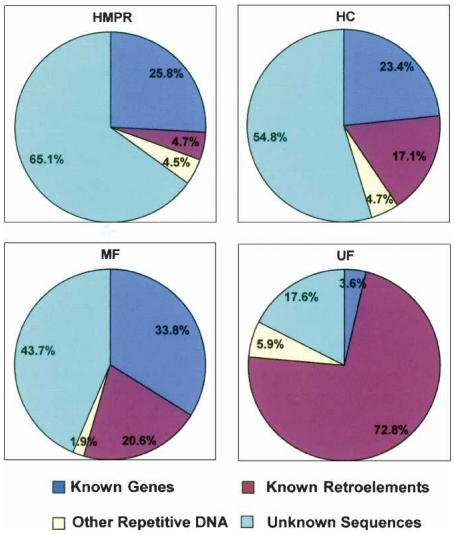


Figure 2. Nuclear sequence composition of hypomethylated partial restriction (HMPR), High Cot (HC), methylation filtration (MF) and unfiltered (UF) libraries. Organellar sequences were removed from the data set prior to the analysis.

DNA sequencing and sequence analysis

Clones were sequenced from both T3 and T7 primer sites by using big dye terminator chemistry. Sequencing reactions were loaded on an ABI3700 capillary sequencer. Resultant sequences were trimmed as previously described (Yuan et al. 2003). BLASTX was used to search the trimmed sequences against the nonredundant amino acid database deposited in the National Center for Biotechnology Information (NCBI) GenBank by August 3, 2004, to identify putative genes, retroelement polyproteins, and transposases, and a cut-off expect value of 10^{-5} was applied. The additional retrotransposons and other repetitive DNA were identified by CROSS_MATCH against a set of 715 known cereal retrotransposons (P. SanMiguel, http://data.genomics.purdue.edu/~pmiguel/projects/retros/) and the TIGR Plant Repeat Databases (http://www.tigr.org/tdb/e2k1/plant.repeats/). The organellar DNA was identified by CROSS_MATCH against complete sequences of the maize chloroplast genome (GenBank accession no. X86563) and the maize mitochondrial genome (GenBank accession no. AY566529).

A sequence was assigned to the gene category as long as it had a BLASTX hit, excluding those cases where the gene homology also matched any retrotransposon, other repeat, or organellar sequence. Sequences that did not match any gene, repetitive DNA, or organellar DNA were categorized as unknown sequences.

One thousand reads each were selected by randomly sampling clones from 306,557 HC clones, 281,669 MF clones, and 17,679 UF clones that had been previously sequenced (Whitelaw et al. 2003; http://www.tigr.org/tdb/tgi/maize/). Analysis of these clones was performed by the same approaches and criteria as described above, with both reads from each clone independently analyzed.

Acknowledgments

We thank Paul Parker for technical assistance, Dr. Ilia Leitch for information regarding genome variation in the angiosperms, and Renyi Liu for assistance in sequence analysis. This research was

supported by a Small Grant for Exploratory Research from the U.S. National Science Foundation (no. 0236505).

References

Bennetzen, J.L. 2002a. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* **115:** 29–36.

———. 2002b. The rice genome: Opening the door to comparative plant biology. *Science* **296:** 60–63.

Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E., and SanMiguel, P. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. *Genome* **37:** 565–576.

Bennetzen, J.L., Coleman, C., Liu, R., Ma, J., and Ramakrishna, W. 2004. Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7:** 732–736.

Bennetzen, J.L., Liu, R., Ma, J., and Pontaroli, A. 2005. Maize genome structure and rearrangement. *Maydica* (in press).

Burr, B., Burr, F.A., Thompson, K.H., Albertson, M.C., and Stuber, C.W. 1988. Gene mapping with recombinant inbreds in maize. *Genetics* **118:** 519–526.

Dubcovsky, J., Ramakrishna, W., SanMiguel, P.J., Busso, C.S., Yan, L., Shiloff, B.A., and Bennetzen, J.L. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol.* **125:** 1342–1353.

Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., Hu, X., et al. 2003. Sequence and analysis of rice chromosome 4. *Nature* **420:** 316–320.

Fu, H., Park, W., Yan, X., Zheng, Z., Shen, B., and Dooner, H.K. 2001. The highly recombinogenic *bz* locus lies in an unusually gene-rich region of the maize genome. *Proc. Natl. Acad. Sci.* **98:** 1082–1087.

Gaut, B.S. and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci.* **94:** 6809–6814.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296:** 92–100.

Gruenbaum, Y., Navey-Many, T., Cedar, H., and Razin, A. 1981. Sequence specificity of methylation in higher plant DNA. *Nature* **292:** 860–862.

Helentjaris, T., Weber, D., and Wright, S. 1988. Identification of the genomic locations of duplicated nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics* **118:** 353–363.

Ilic, K., SanMiguel, P.J., and Bennetzen, J.L. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes. *Proc. Natl. Acad. Sci.* **100:** 12265–12270.

Jabbari, K., Cruveiller, S., Clay, O., Le Saux, J., and Bernardi, G. 2004. The new genes of rice: A closer look. *Trends Plant Sci.* **9:** 281–285.

Kenton, A., Parokonny, A., Bennett, S.T., and Bennett, M.C. 1993. Does genome organization influence speciation? A reappraisal of karyotype studies in evolutionary biology. In: *Evolutionary patterns and processes* (eds. D.R. Lee and D. Edwards), pp. 189–206. Academic Press, London.

Lai, J., Ma, J., Swigonová, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.-J., Jeong, O.-Y., Bennetzen, J.L., et al. 2004. Gene loss and movement in the maize genome. *Genome Res.* **14:** 1924–1931.

Leitch, I.J., Soltis, D.E., Soltis, P.S., and Bennett, M.D. 2005. Evolution of DNA amounts across land plants (Embryophyta). *Annals Bot.* **95:** 207–217.

Ma, J. and Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* **101:** 12404–12410.

Ma, J., Devos, K.M., and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14:** 860–869.

Meyers, B.C., Tingey, S.V., and Morgante, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11:** 1660–1676.

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., and McCombie, W.R. 2003. Maize genome sequencing by methylation filtration. *Science* **302:** 2115–2117.

Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R., and Martienssen, R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* **23:** 305–308.

Reddy, A.R., Ramakrishna, W., Sekhar, A.C., Ithal, N., Babu, P.R., Bonaldo, M.F., Soares, B., and Bennetzen, J.L. 2002. Novel genes are enriched in normalized cDNA libraries from drought stressed seedlings of *indica* rice (*Oryza sativa* L. *cv.* Nagina22). *Genome* **45:** 204–211.

Rice Chromosome 10 Sequencing Consortium. 2003. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* **300:** 1566–1569.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A., and Allard, R.W. 1984. Ribosomal DNA spacer length polymorphism in barley. Mendalian inheritance, chromosomal location and population dynamics. *Proc. Natl. Acad. Sci.* **81:** 8014–8019.

Samuels, M.L. and Witmer, J.A. 2003. *Statistics for the life sciences*, 3rd ed. Prentice Hall, Upper Saddle River, NJ.

SanMiguel, P. and Bennetzen, J.L. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals. Bot.* **82:** 37–44.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J., Niimura, Y., Cheng, Z., Nagamura, Y., et al. 2002. The genome sequence and structure of rice chromosome 1. *Nature* **420:** 312–316.

Shuman, S. 1994. Novel approach to molecular cloning and polynucleotide synthesis using vaccinia DNA topoisomerase. *J. Biol. Chem.* **269:** 32678–32684.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Song, R., Llaca, V., Linton, E., and Messing, J. 2001. Sequence, regulation, and evolution of the maize 22-kD α zein gene family. *Genome Res.* **11:** 1817–1825.

Springer, P.A. 1992. "Genomic organization of *Zea mays* and its close relatives." Ph.D. thesis, Purdue University, West Lafayette, Indiana.

Springer, N.M., Xu, X., and Barbazuk, W.B. 2004. Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol.* **136:** 3023–3033.

Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res.* **14:** 1916–1923.

Van Deynze, A.E., Sorrells, M.E., Park, W.D., Ayres, N.M., Fu, H., Cartinhour, S.W., Paul, E., and McCouch, S.R. 1998. Anchor probes for comparative mapping of grass genera. *Theor. Appl. Genet.* **97:** 356–369.

Vongs, A., Kakutani, T., Martienssen, R.A., and Richards, E.J. 1993. *Arabidopsis thaliana* DNA methylation mutants. *Science* **260:** 1926–1928.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L., and Senchina, D.S. 2002. Intron size and genome size in plants. *Mol. Biol. Evol.* **19:** 2346–2352.

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., et al. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302:** 2118–2120.

Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.* **26:** 307–316.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79–92.

Yuan, Y., SanMiguel, P.J., and Bennetzen, J.L. 2002. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. *Genome Res.* **12:** 1345–1349.

———. 2003. High Cot sequence analysis of the maize genome. *Plant J.* **34:** 249–255.

Web site references

http://www.tigr.org/tdb/tgi/maize/; the TIGR Maize Database

http://www.tigr.org/tdb/e2k1/plant.repeats/; the TIGR Plant Repeat Databases

http://data.genomics.purdue.edu/~pmiguel/projects/retros/; Retrotransposon Database

Received October 14, 2004; accepted in revised form June 13, 2005.