Maize Genome Sequencing by Methylation Filtration

Lance E. Palmer,* Pablo D. Rabinowicz,* Andrew L. O'Shaughnessy, Vivekanand S. Balija, Lidia U. Nascimento, Sujit Dike, Melissa de la Bastide, Robert A. Martienssen,† W. Richard McCombie†

Gene enrichment strategies offer an alternative to sequencing large and repetitive genomes such as that of maize. We report the generation and analysis of nearly 100,000 undermethylated (or methylation filtration) maize sequences. Comparison with the rice genome reveals that methylation filtration results in a more comprehensive representation of maize genes than those that result from expressed sequence tags or transposon insertion sites sequences. About 7% of the repetitive DNA is unmethylated and thus selected in our libraries, but potentially active transposons and unmethylated organelle genomes can be identified. Reverse transcription polymerase chain reaction can be used to finish the maize transcriptome.

Higher plant genomes range from 100 million to 100 billion base pairs (bp) (*1*) because of amplification of repeats and changes in ploidy. The maize genome (2500 Mbp) is up to 80% repetitive, comprising mostly nested retrotransposons (*2*, *3*). This poses major challenges to complete genomic sequencing, unlike smaller genomes such as *Arabidopsis* and rice (*4*–*6*). Gene sequences can be extracted from large genomes by cDNA or expressed sequence tag (EST) sequencing (*7*). However, EST collections typically miss 40 to 50% of genes as a result of their low expression level or cell-type specificity (*8*).

Sequencing strategies that target genes can overcome underrepresentation and include introns and regulatory regions excluded in cDNAs. One strategy uses maize *Mutator* (*Mu*) transposons that preferentially insert into genes (*9*). Integration sites are recovered by plasmid rescue and sequenced (*10*). A similar approach using miniature inverted repeat transposable elements (MITEs) has been proposed (*11*). A problem with these techniques is target specificity of the transposon.

Genes are found in fewer copies and contain less DNA methylation than most transposons, and genes can be selected on this basis (*12*–*15*). High-C_0t (the product of DNA concentration and reassociation time) sequencing (*16*) depends on removal of repetitive DNA by denaturation and rapid reassociation. Methylation-selective sequencing depends on cloning in bacterial hosts that destroy methylated DNA. High-C_0t sequencing may exclude gene families, whereas methylated genes will be excluded by methylation selection. However, sampling indicates that 95% of maize exons are unmethylated, and fully methylated genes are very rare (*17*).

One in four cytosines in maize are methylated, or half of all CpG and CpXpG (where X is any base) sites (*18*). Methylation can silence transposons (*14*, *15*, *19*, *20*), but genes are rarely methylated and such methylation is restricted to the 5′ and 3′ flanking regions (*21*). Methylation filtration (MF) uses small-insert genomic libraries constructed in an *Escherichia coli* host that has the 5mC restriction system McrBC (*22*). This system prevents the propagation of clones carrying methylated inserts and results in a five- to sevenfold enrichment in genes as compared to a control library (*13*).

MF libraries were constructed from immature ear nuclear DNA, and clones were sequenced as read pairs with the use of forward and reverse primers, generating 96,576 MF reads (MFRs). Additional MFRs were downloaded from Genbank along with expressed sequence tags (EST) and Rescue Mu (RM) sequences (*23*). We used comprehensive plant organelle, repeat, and gene databases to annotate genes and repetitive elements with BLAST (*23*). Our gene database is depleted of genes annotated as hypothetical, putative, or unknown, which are often transposons misannotated as genes. Our repeat database includes most of the annotated plant repetitive DNA in GenBank nonredundant protein and nucleotide databases up to September 2002. Consistent with the results of our previous study (*13*), 8.6% of MFRs had a BLASTX match in our gene database, and 24% of the methylation-filtered reads matched repetitive elements with the use of either BLASTN or BLASTX. In contrast, among 5679 whole-genome shotgun sequences or unfiltered reads (UFRs), only 1.4% matched genes, whereas 57% matched repeats. Other reads are likely unknown repeats, intergenic regions, and a minor number of promoters and introns. Our analysis reveals that MF removed 93% of repeats (Fig. 1).

We compared the rice genome sequence to the maize MFR, EST, and RM data sets in order

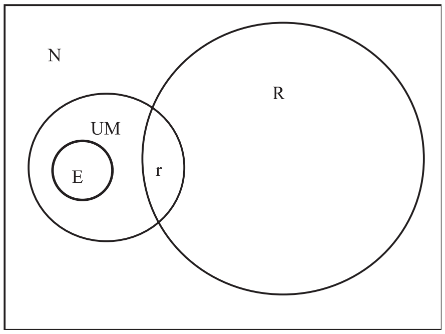

Fig. 1. Gene enrichment by MF. If the set of undermethylated sequences is *UM*, the total number of exons is *E*, and the total set of sequences is *N*, then the portion of the maize genome that is undermethylated is *UM*/*N* = (*E*/*N*)/(*E*/*UM*) = 1.44/8.57, or 17% of the genome. Because 24% of MFRs matched repeats (*r*/*UM*), the fraction of undermethylated repeats in the genome as a whole is *r*/*N* = (*r*/*UM*) × (*UM*/*N*) = 24% × 17% = 4.1%. Because 57% of the genome comprises repeats (*R*/*N*), we estimate that the remaining 52.9% (*R* – *r*) are methylated.

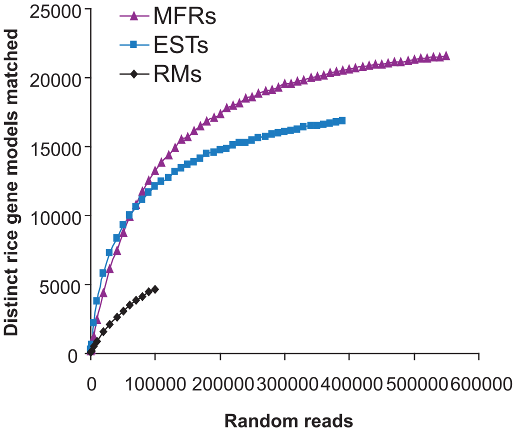

Fig. 2. Comparative gene discovery. Random maize sequences were selected, and the cumulative number of distinct rice FGENESH gene predictions matched by every 10,000 reads were plotted to simulate gene discovery. MFRs, ESTs, and RM reads were compared. Gene discovery was most efficient with ESTs up to 60,000 reads, but with MFRs after that.

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: martiens@cshl.org (R.A.M.); mccombie@cshl.edu (W.R.M.)

to estimate the ability of these approaches to capture genes. Rice and maize diverged about 55 million years ago (*24*) and have a high degree of conservation between orthologous genes but relatively low conservation of repetitive elements (*25*). Additional MFRs from Genbank (559,964 MFRs total), ESTs (383,761), and RMs (95,408) were matched to 3807 sequenced rice bacterial artificial chromosomes (BACs) and P1-derived artificial chromosomes (PACs) with the use of BLASTN (*23*). Forty percent of MFRs, 35% of RM sequences, and 68% of ESTs matched a rice BAC or PAC. We anticipated that most of these rice-maize conserved sequences were exons, and so we applied the gene prediction algorithm FGENESH (*23*) to the rice BAC sequences, generating 79,031 gene predictions. Gene predictions were filtered for repeats (*23*), resulting in 57,641 nonrepeat genes. Eighty-one percent of maize MFRs and 95 to 97% of maize RMs and ESTs overlapped a nonrepeat rice gene prediction (table S1). Thus, 33% of all MFRs, 33% of RM sequences, and 66% of ESTs matched rice gene models. However, because of redundancy, these numbers overestimate gene discovery rates. We addressed this by comparing unique FGENESH gene predictions represented in each data set (Fig. 2). With low numbers of reads, gene matches are still more frequent among ESTs, but once the number of reads exceeds 60,000 many more genes can be identified among MFRs. Thus, EST gene discovery reaches a plateau that falls short of the actual gene content, whereas genomic sequencing is more comprehensive. RM reads are even less representative, even when only optimal reads from each clone are considered. This is likely because of transposon insertion hotspots (*26*) and oversampling of individual libraries. When all rice gene predictions are included in the analysis, similar results are obtained (fig. S1). ESTs and MFRs touched a total of 16,804 (29%) and 21,629 (38%) unique nonrepeat rice gene predictions (out of 57,641 total), respectively. A total of 15,357 rice genes were touched by both MFRs and ESTs, whereas 6272 unique rice gene predictions were represented only in MFR. If hypothetical genes (*23*) are excluded from the unique rice gene predictions (26,401 genes), 14,807 (56%) genes are covered by MFRs and 11,783 (44%) are covered by ESTs.

We also examined coverage of genes on sequenced maize BACs. We found comparable representation of maize genes by ESTs and MFRs (30% and 39% of nonrepeat genes, respectively) (Fig. 3 and fig. S2). When hypothetical genes are excluded, 65% of the genes are touched by MFRs. A similar analysis of random rice BACs revealed similar levels of coverage (fig. S3). Whereas abundant transcripts were overrepresented in ESTs, repeats were overrepresented in MFRs (Fig. 3).

Coverage at the intragenic level was assessed by measuring representation of first and last exons (fig. S4). In MFRs, the first and last exons were equally represented, but ESTs had a strong bias toward the 3′ end of genes. This was not a result of directional cloning, because most of the maize cDNAs were sequenced from the 5′ end (*27*). Rather, as with most cDNA libraries, the clones were not full-length. RM matches were biased toward the 5′ end of rice gene models, consistent with the targeting of *Mutator* elements to this region of the gene (*26*). The current level of MF coverage tags about one-third of the predicted rice gene set and the limited maize sequences available, and the coverage increases to over one-half of the gene set when nonrepeat, domain-containing genes are examined (Fig. 3). However, gene coverage is rarely complete (fig. S5A). Comparative genome analysis and polymerase chain reaction (PCR) from cDNA can retrieve the remaining portion. In order to

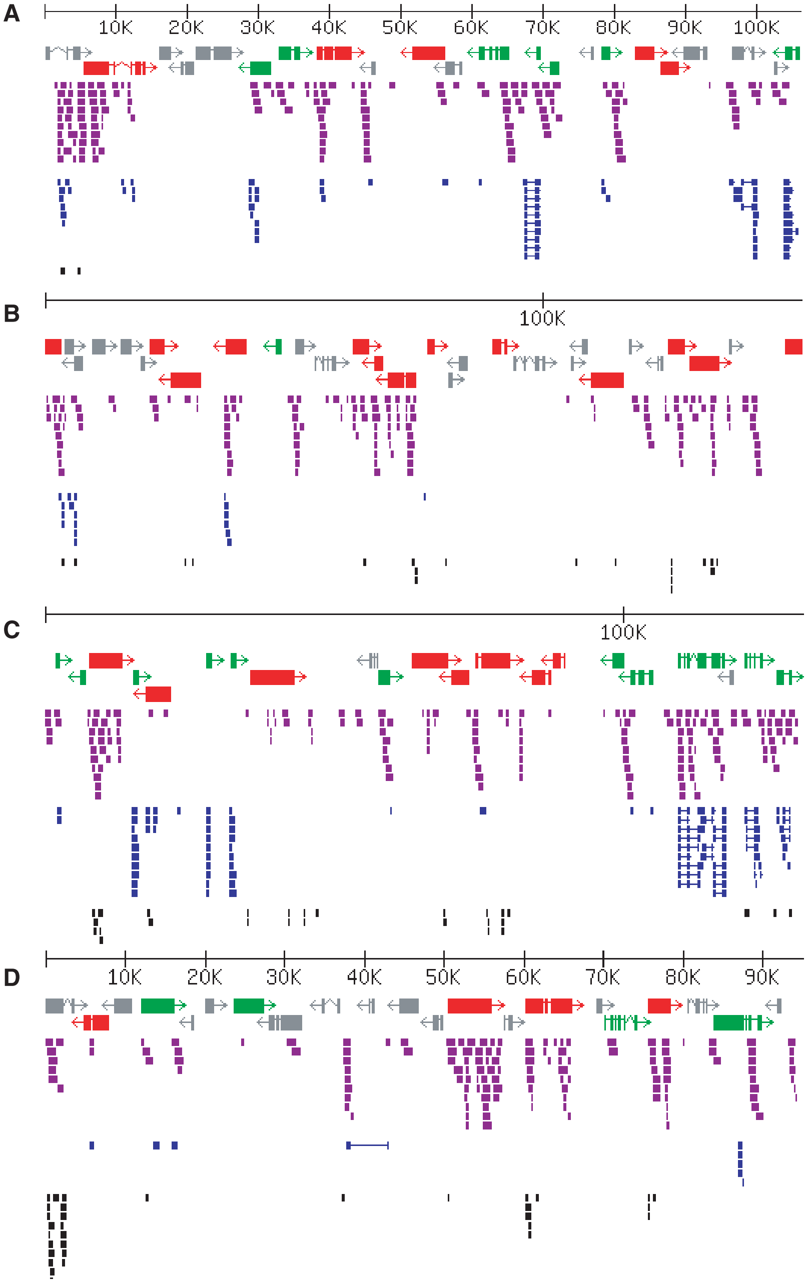

Fig. 3. Maize genome coverage. Reads were aligned to a set of 15 maize BACs with the use of BLAT (fig. S2). The MFR coverage of four representative maize BACs is shown along with partial EST coverage. The complete EST and RM coverage of these and 11 additional BACs are shown in fig. S2. FGENESH gene predictions are classified as putative if they contain a nonrepeat domain (green), hypothetical if they contain no known domain (gray), or repeat (red). MFRs, purple; ESTs, blue; RMs, black. (**A**) GI 18092333, (**B**) GI 33113959, (**C**) GI 30698556, and (**D**) GI 19908841.

identify complete transcripts, MFRs were mapped to rice, and primers [for reverse transcription PCR (RT-PCR) as well as 5′ and 3′ rapid amplification of cDNA ends PCR (RACE-PCR)] were selected from maize sequence that overlapped predicted exons (*23*). Of 92 predicted genes tested, 67 (73%) produced a spliced PCR product matching the original MFR or targeted gene. After clustering PCR, products with PHRAP (*23*), 50 targeted genes produced a sequence contig (or overlapping clone set) that linked two or more independent MF clones (fig. S5B). This suggests that RT-PCR, coupled with maize-rice comparisons, will be a powerful strategy for finishing genes.

We considered two classes of repeats recovered by MF: those that lost McrBC sites via mutation (CpG suppression) and those that lost McrBC sites via hypomethylation. McrBC requires two (A/G)5mC half-sites for restriction, and the frequency of half-sites that overlap CpG or CpXpG is shown in Fig. 4. Centromeric satellites, Ji, knobs, Opie, and Prem2 are more frequent in MFRs than UFRs, whereas Xilon, Tekay, Grande, Huck, Cinful, Zeon, and ribosomal DNA sequences show the opposite trend. In each case, the proportion of unmethylated repeats in the genome (*r*/*R*) was less than 27% (table S2). Ribosomal RNA genes, whose sequences are under selection, are efficiently removed by MF, whereas knob repeats are less efficiently selected and are depleted of McrBC sites. Ji, Opie, and Prem2 all have fewer McrBC sites in MFRs than UFRs, suggesting that older family members escape filtration through CpG suppression (*28*). In contrast, Prem1, Xilon, Grande, and Huck elements have similar McrBC frequencies in UFRs and MFRs. In this case, selection against these elements depends on methylation. We attempted to assemble entire elements (pseudocontigs) from MFRs in order to determine whether full-length elements were represented (*29*). Near-complete Opie and Xilon elements could be assembled and were not depleted of McrBC sites, suggesting that at least some full-length elements are unmethylated.

Mitochondrial DNA is unmethylated and is represented among MFRs even though nuclei are partially purified before DNA extraction. This organellar genome is currently being sequenced (*30*), but a methylation-based selective approach could be applied to sequence organelle genomes in other species [Supporting Online Material (SOM) Text]. The maize chloroplast genome, which is also unmethylated and has been fully sequenced (*31*), was represented in full in MFRs.

ESTs only match a minority of gene models, and gene discovery progresses no further after 100,000 reads. Further, only transcribed regions can be assembled from ESTs, whereas the sequence of gene islands can potentially be completed by MF. Assuming a genome size of 2500 Mbp, we estimate the undermethylated fraction of the genome or "gene space" to be 17%, or 425 Mbp in size (Fig. 1), which like the similarly sized rice genome becomes a viable sequencing project. With the current sequencing technology, less than 3 million reads would be enough to cover the maize gene space fivefold.

Maize is an ancient tetraploid and so likely contains up to 70% more genes than rice, or 50,000 to 70,000 genes (*32*). Assuming 3000 bp per gene, including 500 bp of promoter sequence (*33*–*35*), this would allow for 215 Mbp of undermethylated repeats and other intergenic regions. The elimination of more than 90% of repeats by methylation filtration reduces sequencing costs without sacrificing information, because reads within these repeats could not be assembled in any case by whole-genome shotgun analysis.

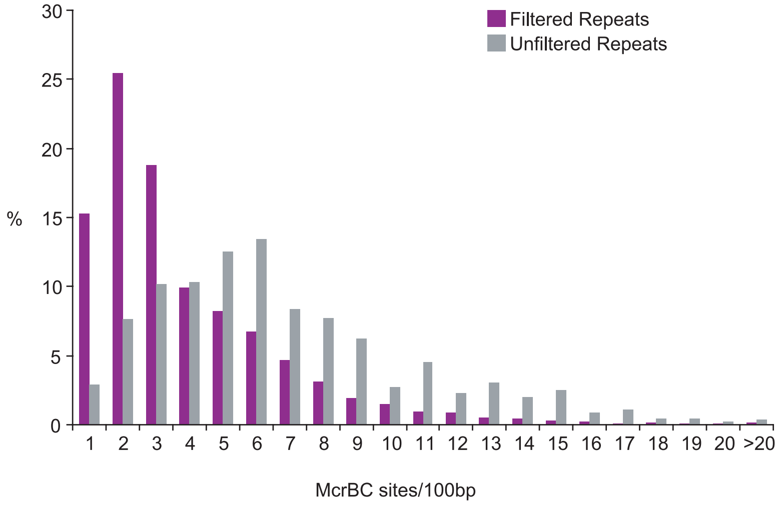

Fig. 4. Analysis of repeats. A total of 46,109 independent MFRs (average length of 545 nucleotides) and 2271 independent UFRs (average length, 549 nucleotides) were analyzed for repetitive sequence matches. The density of McrBC recognition sites per 100 bp were plotted in 10,581 MFR repeats and 1285 UFR repeats.

References and Notes

1. M. D. Bennett, I. J. Leitch, *Ann. Bot.* **76**, 113 (1995).
2. S. Hake, V. Walbot, *Chromosoma* **79**, 251 (1980).
3. P. SanMiguel *et al.*, *Science* **274**, 765 (1996).
4. J. Yu *et al.*, *Science* **296**, 79 (2002).
5. The Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).
6. S. A. Goff *et al.*, *Science* **296**, 92 (2002).
7. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
8. M. F. Bonaldo, G. Lennon, M. B. Soares, *Genome Res.* **6**, 791 (1996).
9. A. D. Cresse, S. H. Hulbert, W. E. Brown, J. R. Lucas, J. L. Bennetzen, *Genetics* **140**, 315 (1995).
10. M. N. Raizada, G. L. Nan, V. Walbot, *Plant Cell* **13**, 1587 (2001).
11. L. Mao *et al.*, *Genome Res.* **10**, 982 (2000).
12. P. D. Rabinowicz, W. R. McCombie, R. A. Martienssen, *Curr. Opin. Plant Biol.* **6**, 150 (2003).
13. P. D. Rabinowicz *et al.*, *Nature Genet.* **23**, 305 (1999).
14. J. L. Bennetzen, K. Schrick, P. S. Springer, W. E. Brown, P. SanMiguel, *Genome* **37**, 565 (1994).
15. R. Martienssen, *Trends Genet.* **14**, 263 (1998).
16. Y. Yuan, P. J. SanMiguel, J. L. Bennetzen, *Plant J.* **34**, 249 (2003).
17. P. D. Rabinowicz *et al.*, *Genome Res.* **13**, 2658 (2003).
18. C. M. Papa, N. M. Springer, M. G. Muszynski, R. Meeley, S. M. Kaeppler, *Plant Cell* **13**, 1919 (2001).
19. R. B. Flavell, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3490 (1994).
20. T. Singer, C. Yordan, R. A. Martienssen, *Genes Dev.* **15**, 591 (2001).
21. V. Walbot, C. Warren, *Plant Mol. Biol.* **15**, 121 (1990).
22. E. A. Raleigh *et al.*, *Nucleic Acids Res.* **16**, 1563 (1988).
23. Materials and methods are available as supporting material on *Science* Online.
24. E. A. Kellogg, *Plant Physiol.* **125**, 1198 (2001).
25. D. T. Morishige, K. L. Childs, L. D. Moore, J. E. Mullet, *Plant Physiol.* **130**, 1614 (2002).
26. B. P. May *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
27. C. F. Lunde, D. J. Morrow, L. M. Roy, V. Walbot, *Funct. Integr. Genomics* **3**, 25 (2003).
28. F. Antequera, A. P. Bird, *EMBO J.* **7**, 2295 (1988).
29. B. C. Meyers, S. V. Tingey, M. Morgante, *Genome Res.* **11**, 1660 (2001).
30. The project is funded by National Science Foundation award number 0110168.
31. R. M. Maier, K. Neckermann, G. L. Igloi, H. Kossel, *J. Mol. Biol.* **251**, 614 (1995).
32. B. S. Gaut, J. F. Doebley, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6809 (1997).
33. The Rice Chromosome 10 Sequencing Consortium, *Science* **300**, 1566 (2003).
34. T. Sasaki *et al.*, *Nature* **420**, 312 (2002).
35. Q. Feng *et al.*, *Nature* **420**, 316 (2002).
36. We thank the AMDeC Bioinformatics Core Facility at the Columbia Genome Center, Columbia University, for computational support. This work was supported by NSF award DBI-0110143 from the Plant Genome Research Program to W.R.M. and R.A.M. L.E.P. was supported by National Cancer Institute training grant number 2 T32-CA09311-25. W. R. M. and R. A. M. are founders, members of the board, and consultants for Orion Genomics LLC of St. Louis, MO, which has commercialized the MF technology used in this paper.

Supporting Online Material

www.sciencemag.org/cgi/content/full/302/5653/2115/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 and S2
References

8 September 2003; accepted 14 November 2003