Enrichment of Gene-Coding Sequences in Maize by Genome Filtration

C. A. Whitelaw,[1] W. B. Barbazuk,[2]* G. Pertea,[1] A. P. Chan,[1] F. Cheung,[1] Y. Lee,[1] L. Zheng,[1] S. van Heeringen,[1] S. Karamycheva,[1] J. L. Bennetzen,[3] P. SanMiguel,[4] N. Lakey,[5] J. Bedell,[5] Y. Yuan,[3] M. A. Budiman,[5] A. Resnick,[1] S. Van Aken,[1] T. Utterback,[6] S. Riedmuller,[6] M. Williams,[6] T. Feldblyum,[6] K. Schubert,[2] R. Beachy,[2] C. M. Fraser,[1] J. Quackenbush[1]*

Approximately 80% of the maize genome comprises highly repetitive sequences interspersed with single-copy, gene-rich sequences, and standard genome sequencing strategies are not readily adaptable to this type of genome. Methodologies that enrich for genic sequences might more rapidly generate useful results from complex genomes. Equivalent numbers of clones from maize selected by techniques called methylation filtering and High C_0t selection were sequenced to generate ~200,000 reads (approximately 132 megabases), which were assembled into contigs. Combination of the two techniques resulted in a sixfold reduction in the effective genome size and a fourfold increase in the gene identification rate in comparison to a nonenriched library.

Maize (*Zea mays*) is an important food source and has a highly tractable genetic system. The maize genome [2300 to 2700 megabases (Mb)] (*1*) is approximately 20 and 6 times as large as those of *Arabidopsis thaliana* and rice (*Oryza sativa*), respectively. Over 60% of the maize genome consists of long terminal repeat (LTR)–retrotransposon families that vary in copy number, with up to 30,000 copies per haploid genome. Other repetitive DNA sequences account for an additional 20% of the maize genome (*2*). The genic regions of the maize genome contain low–copy number genes separated from one another by large tracts [~10 to 100 kilobases (kb)] of repetitive DNA (*3*–*5*); for the purpose of this manuscript, the term "gene" excludes any sequence contained within a transposable element. DNA sequencing of large and complex genomes is currently limited by cost considerations and difficulties in resolving repetitive sequences in assembly. Consequently, development of strategies that focus on targeted sequencing of gene-rich regions provides an alternative to whole-genome sequencing.

A technique called "methylation filtering" (MF), in which hypermethylated sequences are excluded with the use of bacterial restriction systems that cleave methylated sequences, has been used to produce libraries that are gene-enriched (*6*–*8*). Another technique, High C_0t (HC) selection, allows separation of DNA fractions into low-copy (High C_0t) or high-copy (Low C_0t) sequences, where concentration (C_0) and annealing time (t) determine the composition of the fractions (*9*, *10*). The most repetitive DNA renatures first, and the double-stranded DNA can be separated from lower copy number, unrenatured DNA. The low–copy number fraction from maize can be enriched fourfold in genes in comparison to random shotgun libraries (*11*).

DNA sequence was generated from 56,649 MF, 84,981 HC, and 17,679 unfiltered (UF) plasmid clones, as a combination of paired-end and forward-only sequence reads. A total of 95,233 MF, 100,000 HC, and 34,074 UF sequences were produced with an average edited read length of 721, 712, and 708 bases, respectively. All sequence data generated for this analysis were deposited into GenBank (*12*).

Four separate clustering and assembly analyses using MF, HC, combined MF and HC (MF+HC), and UF sequences generated 52,649 (~42 Mb); 71,492 (~54 Mb); 117,304 (~93 Mb); and 32,955 (~24 Mb) assembled *Zea mays* sequences (AZMs), respectively (*13*) (table S1). In the MF+HC assembly, 21,775 AZMs were built from two or more independent clones and were classified as MF-only, HC-only, and mixed AZMs on the basis of content. Sixty percent of the MF clones are in MF-only AZMs, and 72% of the HC clones are in HC-only AZMs. We expect the HC and MF clones to be as likely to assemble with each other as to assemble with themselves. Because each group favors self-assembly, the MF and HC libraries likely represent somewhat distinct portions of the genome.

An analysis of the repeat content of each of the MF, HC, MF+HC, and UF AZMs indicates that 35, 21, 28, and 73%, respectively, of the nucleotides align with sequences in the TIGR Cereal Repeat database (www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml) (table S1). The majority of sequence alignments are to transposable elements including retrotransposons, transposons, and miniature inverted repeat transposable elements (MITES); of the matches to transposable elements, nearly 90% are to retrotransposons for both MF and HC (Table 1).

The MF, HC, MF+HC, and UF AZMs were compared with databases of expressed transcripts with the use of BLAT (a BLAST-like alignment tool) to estimate the respective levels of gene-enrichment (*14*). These databases included a nonredundant amino acid database (NRAA) and the TIGR Plant Gene

Table 1. Repeat analysis of MF, HC, MF+HC, and UF AZMs. A sequence was defined as having a significant match to a repeat if greater than 75% of the nucleotides in the query sequence were masked by a categorized sequence in the TIGR Cereal Repeat database. The total number of repeat sequences is expressed as a percentage of the total number of unique sequences (AZMs).

Repeat category	MF	HC	MF+HC	UF
Total no. unique sequences (AZMs)	52,649	71,492	117,304	32,955
Transposable elements				
Retrotransposons	15,176	9,290	24,352	20,782
Transposons	358	373	724	223
MITES	19	80	97	14
Centromeric and telomeric sequences				
Centromere-related	94	53	147	147
Telomere-related	9	2	11	1
Ribosomal genes	118	249	367	798
Other repetitive sequences				
Known repetitive sequences	1,590	29	1,619	456
Unknown repetitive sequences	55	78	137	56
Total no. repeat sequences	17,419 (33%)	10,154 (14%)	27,454 (23%)	22,477 (68%)

[1]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132, USA. [3]Department of Genetics, University of Georgia, Athens, GA 30602, USA. [4]Purdue Genomics Core Facility, Purdue University, West Lafayette, IN 47907, USA. [5]Orion Genomics, 4041 Forest Park Avenue, St. Louis, MO 63108, USA. [6]J. Craig Venter Science Foundation Joint Technology Center (JTC), 5 Research Place, Rockville, MD 20850, USA.

*To whom correspondence should be addressed. E-mail: BBarbazuk@danforthcenter.org (W.B.B.); Johnq@tigr.org(JQ)

Indices (*15*) (Table 2; table S2). The total number of significant matches to expressed transcripts by the MF+HC AZMs was 24%, comparable to MF-only and HC-only AZMs (27 and 22%, respectively). In contrast, 6% of the UF sequences showed significant homology to expressed transcripts. This fourfold difference in the gene identification rate, achieved at a stringency of 1×10^{-10}, is similar to that observed previously (*2*, *7*, *11*). The enrichment of genes between filtered and UF AZMs is expected to increase as new repeats are discovered and as the repeat database used for masking becomes more robust.

To estimate coverage of existing maize genomic sequence by the MF, HC, MF+HC, and UF AZMs, a collection of maize sequence databases were compiled from public resources and searched (table S3). One such database consists of the "IBM2 neighbors" collection of sequence-based genetic mapping markers (www.maizemap.org/resources.htm) that are uniformly distributed across maize chromosomes 1 to 10; in some cases there are multiple markers for a single locus but most are distinct. The MF, HC, MF+HC, and UF AZMs were searched against the maize sequence markers with the use of BLAT (*14*) (Fig. 1). Approximately 27, 20, and 38% of the sequence markers show a significant match to MF, HC, and MF+HC AZMs, respectively, compared with 2% for UF AZMs. Furthermore, the MF+HC AZMs appear to be evenly distributed throughout the 10 chromosomes (Fig. 1).

Another source of sequence-based markers for mapping is the 10,643 overlapping oligonucleotide (overgo) probes (www.maizemap.org/overgos.htm) of which 96% are unique. The overgo sequences were compared with the MF, HC, MF+HC, and UF AZMs with the use of WU-BLASTN (*16*), resulting in 1142 and 1143 overgos aligning uniquely to MF and HC AZMs, whereas a further 215 were common to MF+HC. In contrast, 94 of the overgo probes mapped to UF AZMs, which reinforces the observation that both MF and HC approaches effectively enrich for unique, genic portions of the genome.

Although MF and HC AZMs match approximately equal numbers of unique sequence markers and expressed transcripts, other analyses suggest differences between the gene-enrichment technologies. For example, the HC AZMs contain almost twice as much uncharacterized DNA sequence in comparison to MF AZMs, which can be partly attributed to the fact that 29% of the MF and only 13% of the HC AZMs match retrotransposons (Table 2).

We identified retrotransposon sequences by WU-BLASTN (*16*) and determined their representation within the UF, MF, and HC sequences. Although most abundant in the maize genome (*2*), Huck and Grande elements are reduced in MF sequence (9.5 and 19% of UF levels, respectively) as are Ji and Opie, albeit less so (83 and 80% of UF levels, respectively), suggesting that Ji and Opie are less frequently methylated (Fig. 2). Prem, Ji, Gyma, and Xilon are the most commonly occurring retroelements in maize ESTs, whereas Huck and Grande elements are rare. Inefficient filtering of Ji and Opie retrotransposon sequences may reflect a lack of methylation sites (CG or CNG), transcriptional activity, or irregularities in methylation. Ji and Opie sequences derived from MF are reduced in methylation sites by approximately two-thirds in comparison to those derived from UF. However, for several other retrotransposon types there is no clear relation.

Simple sequence repeats (SSRs) are often associated with plant genes (*17*, *18*); therefore, filtration methods should enrich for SSR-containing sequences. We searched the AZMs for perfect di-, tri-, and tetra-nucleotide repeat motifs and identified 2496 and 1687 SSRs within 2236 MF (4.2%) and 1581

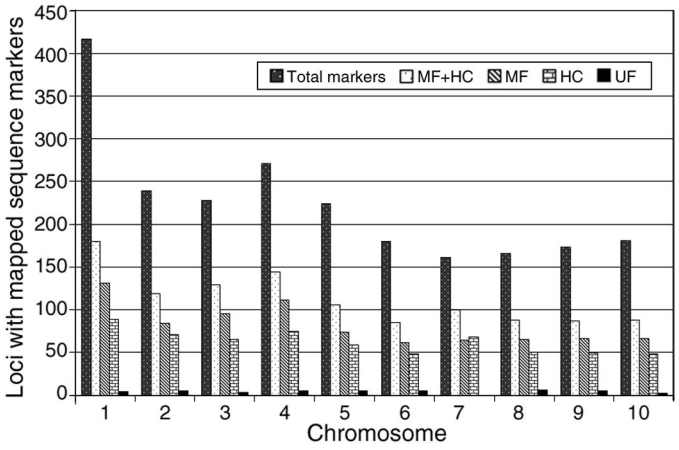

Fig. 1. The distribution of maize sequence markers and markers matching AZMs, across the 10 maize chromosomes. The MF, HC, MF+HC, and UF AZMs were searched against the IBM2 neighbors collection of maize sequence markers with the use of BLAT (*14*).

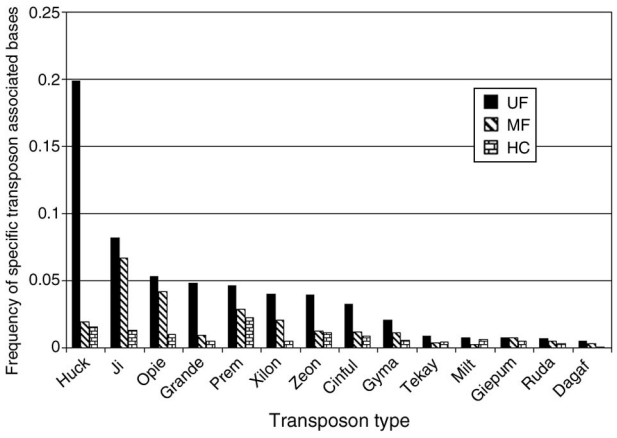

Fig. 2. The distribution of 14 classes of retrotransposon sequences within the MF, HC, and UF AZMs. Unassembled MF, HC, and UF sequences were searched against a library of known maize retrotransposon sequences extracted from the TIGR Cereal Repeat database with the use of WU-BLASTN (*16*). Frequencies are displayed as the total number of UF, MF, or HC bases that align with the transposon divided by the total number of bases in the AZMs under consideration.

Table 2. Sequence similarity search results of maize genomic assemblies. The MF, HC, MF+HC, and UF AZMs were searched against maize chloroplast and mitochondrial genomes, the TIGR Cereal Repeat database, a nonredundant amino acid database (NRAA), and the Tentative Consensus (TC) sequences in the TIGR Plant Gene Indices with the use of BLAT (*14*). The number of sequence matches in each database is expressed as a percentage of the total number of unique sequences (AZMs).

Database	MF	HC	MF+HC	UF
Total matches to chloroplast and mitochondria	93 (0.2%)	445 (0.6%)	471 (0.4%)	38 (0.1%)
Total matches to repetitive sequences*	17,419 (33%)	10,154 (14%)	27,454 (23%)	22,477 (68%)
NRAA	7,684 (15%)	7,921 (11%)	14,588 (12%)	901 (3%)
Plant TC	6,705 (13%)	7,698 (11%)	13,032 (11%)	1,218 (4%)
Total matches to expressed transcripts	14,389 (27%)	15,619 (22%)	27,620 (24%)	2,119 (6%)
No significant match	20,748 (39%)	45,274 (63%)	61,759 (53%)	8,321 (25%)
Total no. unique sequences (AZMs)	52,649	71,492	117,304	32,955

*See Table 1.

HC AZMs (2.2%), respectively. Only 371 SSRs were identified within 351 UF AZMs (1.1%). Both MF and HC enrich for SSR-containing sequences, with HC to a lesser extent. In addition, the composition of the SSRs recovered suggests that the filtration methods are targeting different genomic regions. Specifically, in comparison to MF, HC tends to bias against GC-rich SSR motifs and tri-nucleotide SSR motifs (table S4).

The Lander-Waterman model (*19*) is used to monitor the progress of genomic sequencing projects and assumes an independent sampling of the genome. We used the model and the rate of contig formation to estimate the effective genome size explored by each cloning method, which was 260 and 284 Mb for MF and HC, respectively. These represent 10- and 9-fold reductions relative to the 2500 Mb genome size (table S5). The AZMs cover approximately 16 (MF) and 19% (HC) of the effective genome size, consistent with the approximate coverage of representative unique sequences within the genome. For MF+HC, the estimated genome size (413 Mb) is larger than either of the individual libraries, but smaller than the 544 Mb sum of the separate estimates. This result is consistent with other evidence, indicating that the MF and HC strategies sample different portions of the unique sequence in maize but with an overlap of 131 Mb, or 32%, of the shared genome space. As an additional analysis, we determined that increasing the number of raw sequence reads to 500,000 (a 150% increase) results in a combined genome size of 468 Mb with an approximate 35% overlap between the genomic regions sampled by MF and HC. The 13% increase in estimated size despite a 150% increase in the number of sequence reads suggests an asymptomatic increase toward the actual genome size with increasing sequence depth. Thus MF and HC are complementary approaches to exploring the unique, gene-rich regions of the maize genome. The combined estimated size of 413 Mb is similar to the genome size of rice, which is consistent with the evolution of maize from a much smaller-genome monocot predecessor (*20*). Such a mechanism would preserve the core genome while maintaining the synteny observed in the monocots (*21*).

When the Lander-Waterman equation is applied using UF data, the estimated genome size of 2500 Mb is expected; however, the size calculated was 1285 Mb (*22*). The model was also applied to simulated genomes to determine the amount of variability in the calculation, given the current coverage versus the expected genome size. Our results show that greater discrepancies at low coverage of larger genomes are expected and that effective genome size predictions of low-coverage, reduced genomes are more accurate (table S6). The estimates of the effective genome sizes are lower bounds that will be refined as more data become available.

The overall goal of this pilot sequencing project in maize is to derive an effective strategy for the completion of the genome sequence. To meet this goal, approximately 800,000 total sequence reads will be generated from the HC and MF libraries; this manuscript is an analysis of the first 25% of the data. The HC and MF libraries provide complementary coverage of approximately 413 Mb of unique sequence. The targeted approaches described increase the likelihood of identifying the genes encoded within the genome while reducing the complexity of sequence assembly. We believe the MF and HC strategies may serve as a model for sequencing this and other large, complex genomes at reduced cost relative to conventional approaches.

References and Notes

1. K. Arumuganathan, E. D. Earle, *Plant Mol. Biol. Rep*. **9**, 208 (1991).
2. B. C. Meyers, S. V. Tingey, M. Morgante, *Genome Res.* **11**, 1660 (2001).
3. P. S. Springer, K. J. Edwards, J. L. Bennetzen, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 863 (1994).
4. S. Hake, V. Walbot, *Chromosoma* **79**, 251 (1980).
5. M. Freeling, *Annu. Rev. Plant Physiol*. **35**, 277 (1984).
6. J. L. Bennetzen, K. Schrick, P. S. Springer, W. E. Brown, P. SanMiguel, *Genome* **37**, 565 (1994).
7. P. D. Rabinowicz *et al.*, *Nature Genet.* **23**, 305 (1999).
8. Methylation filtering is marketed as GeneThresher Technology, a registered trademark of Orion Genomics LLC.
9. R. J. Britten, D. E. Khone, *Science* **161**, 529 (1968).
10. R. J. Britten, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 415 (1976).
11. Y. Yuan, P. J. SanMiguel, J. L. Bennetzen, *Plant J.* **33**, 1 (2003).
12. The GenBank accession numbers for the sequences used in this study can be downloaded from ftp://ftp.tigr.org/pub/data/MAIZE/Release2.0_accessions.
13. Materials and Methods are available as supporting material on *Science* Online.
14. J. Kent, *Genome Res.* **12**, 656 (2002).
15. J. Quackenbush *et al.*, *Nucleic Acids Res.* **29**, 159 (2001).
16. W. R. Gish, WU BLAST v. 2.0 (2003). Available at http://blast.wustl.edu.
17. M. Morgante, M. Hanafey, W. Powell, *Nature Genet.* **30**, 194 (2002).
18. S. R. McCouch *et al.*, *DNA Res.* **9**, 199 (2002).
19. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
20. J. L. Bennetzen, *Genetica* **115**, 29 (2002).
21. K. M. Devos, M. D. Gale, *Plant Cell* **12**, 637 (2000).
22. C. A. Whitelaw *et al.*, data not shown.

23. We would like to acknowledge C. R. Buell and C. D. Town for critical reading of the manuscript as well as H. Koo, V. Antonescu, J. Tsai, R. Sultana, S. Sunkara, A. Nunberg, D. Robbins, R.W. Citek, C. Tatham, E. R. Flick, J. T. Jones, and the TIGR–JTC Sequencing Core members for their valuable contribution to the Consortium for Maize Genomics project. Supported by the National Science Foundation award DBI-0221536.

Supporting Online Material

www.sciencemag.org/cgi/content/full/302/5653/2118/DC1
Materials and Methods
Tables S1 to S6
References

4 August 2003; accepted 14 November 2003

Separase Regulates INCENP–Aurora B Anaphase Spindle Function Through Cdc14

Gislene Pereira[1,2] and Elmar Schiebel[1]*

The inner centromere-like protein (INCENP) forms a complex with the evolutionarily conserved family of Aurora B kinases. The INCENP-Aurora complex helps coordinate chromosome segregation, spindle behavior, and cytokinesis during mitosis. INCENP-Aurora associates with kinetochores in metaphase and with spindle microtubules in anaphase, yet the trigger for this abrupt transfer is unknown. Here we show that the conserved phosphatase Cdc14 regulated the yeast INCENP-Aurora complex, Sli15-Ipl1. Cdc14 dephosphorylated Sli15 and thereby directed the complex to spindles. Activation of Cdc14 by separase was sufficient for Sli15 dephosphorylation and relocalization. Cdc14 not only regulates mitotic exit but also modulates spindle midzone assembly through Sli15-Ipl1.

At the metaphase-anaphase transition, the activated separase, Esp1, promotes sister chromatid separation by cleaving the cohesin complex (*1*). The spindle elongates dramatically and separates the sister chromatids to opposite poles. Separase also plays an ill-defined yet essential role in regulating the stability of anaphase spindles (*2*, *3*). This function may be mediated by separase-dependent activation of the conserved phosphatase Cdc14 (*4*). In yeast, Cdc14 is kept inactive by entrapment in the nucleolus (*5*, *6*). In early anaphase, separase, as part of the

[1]The Paterson Institute for Cancer Research, Wilmslow Road, Manchester M20 4BX, UK. [2]School of Biological Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, UK.

*To whom correspondence should be addressed. E-mail: eschiebel@picr.man.ac.uk