Structure and Architecture of the Maize Genome[1][W]

Georg Haberer[2], Sarah Young[2], Arvind K. Bharti[2], Heidrun Gundlach, Christina Raymond, Galina Fuks, Ed Butler, Rod A. Wing, Steve Rounsley, Bruce Birren, Chad Nusbaum, Klaus F.X. Mayer, and Joachim Messing*

Munich Information Center for Protein Sequences, Institute for Bioinformatics, Gesellschaft für Strahlenforschung Research Center for Environment and Health, D–85764 Neuherberg, Germany (G.H., H.G., K.F.X.M.); Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02141 (S.Y., C.R., S.R., B.B., C.N.); Plant Genome Initiative at Rutgers, Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854 (A.K.B., G.F., J.M.); and Arizona Genomics Institute, University of Arizona, Tucson, Arizona 85721 (E.B., R.A.W.)

Maize (*Zea mays* or corn) plays many varied and important roles in society. It is not only an important experimental model plant, but also a major livestock feed crop and a significant source of industrial products such as sweeteners and ethanol. In this study we report the systematic analysis of contiguous sequences of the maize genome. We selected 100 random regions averaging 144 kb in size, representing about 0.6% of the genome, and generated a high-quality dataset for sequence analysis. This sampling contains 330 annotated genes, 91% of which are supported by expressed sequence tag data from maize and other cereal species. Genes averaged 4 kb in size with five exons, although the largest was over 59 kb with 31 exons. Gene density varied over a wide range from 0.5 to 10.7 genes per 100 kb and genes did not appear to cluster significantly. The total repetitive element content we observed (66%) was slightly higher than previous whole-genome estimates (58%–63%) and consisted almost exclusively of retroelements. The vast majority of genes can be aligned to at least one sequence read derived from gene-enrichment procedures, but only about 30% are fully covered. Our results indicate that much of the increase in genome size of maize relative to rice (*Oryza sativa*) and Arabidopsis (*Arabidopsis thaliana*) is attributable to an increase in number of both repetitive elements and genes.

Maize (*Zea mays* or corn) has a wide variety of uses and broad economic impact. It is a significant food source for humans, a chief ingredient in livestock feed, and is the source of a wide range of manufactured products, including sweeteners, fuel, and adhesives. It also has a long and storied history as a model organism in genetic studies. The combination of its genetic and economic importance has made maize a prime organism for genomic studies (for review, see Messing, 2005). Despite its evident value, progress toward generating a whole-genome sequence of maize has been held back by the cost and complexity of such a project. Although it is a medium-sized grass genome, at 2.4 Gb the maize genome is large compared to other sequenced plants and so will require significant funding to sequence. On top of this, its high repeat content poses computational challenges for accurately assembling a genome sequence.

In the absence of a genome sequence, studies of selected regions of the maize genome and comparisons to related species have been carried out. Comparative genetic analyses (Hulbert et al., 1990; Ahn and Tanksley, 1993; Moore et al., 1995; Gale and Devos, 1998) have suggested that significant portions of grass genomes are conserved (collinear). It has been proposed that, aside from polyploidization, large genome sizes in the grasses are caused primarily by the high content of repetitive elements (SanMiguel and Bennetzen, 1998; Meyers et al., 2001; Song et al., 2002). Several studies have investigated the local structure of orthologous regions in various grass species (Chen et al., 1997; Feuillet and Keller, 1999; Tikhonov et al., 1999; Tarchini et al., 2000; Ramakrishna et al., 2002a, 2002b; Song et al., 2002; Brunner et al., 2003; Ilic et al., 2003; Langham et al., 2004). These studies paint a picture of grass genomes that have macrocollinearity, or a general conservation of genes and gene order, but because of numerous small-scale genic rearrangements, such as insertions, deletions, amplifications, inversions, and translocations, lack perfect microcollinearity. Although the results are suggestive, the regions studied represent a tiny fraction of the genome. In addition, since all the regions were selected based

[1] This work was supported by the National Science Foundation Plant Genome (grant no. 0211851). Work at the Munich Information Center for Protein Sequences was in part supported by the Genomanalyse im biologischen System Pflanze program of the German Ministry for Education and Research.

[2] These authors contributed equally to the paper.

* Corresponding author; e-mail messing@mbcl.rutgers.edu; fax 732–445–0072.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Joachim Messing (messing@mbcl.rutgers.edu).

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.105.068718.

on the presence of mapped genes of specific interest, they are also intrinsically biased and are not likely to be representative of the general genome organization. An accurate assessment of the content and organization of the maize genome requires a more comprehensive and unbiased dataset.

Existing data suggest that plant genomes are much more dynamic than similarly related animal genomes in terms of size, gene content, organization, and repeat content (for review, see Messing, 2005). For example, grass genomes vary in size from rice (*Oryza sativa*; 0.4 Gb) to wheat (*Triticum aestivum*; 16 Gb). Because of its relatively small size and low proportion of repetitive DNA, whole-genome sequencing efforts in the grasses were initially focused on rice. Rice has about 30% more genes than Arabidopsis (*Arabidopsis thaliana*), which is largely attributed to gene family expansion (International Rice Genome Sequencing Project, 2005). Even within a single species, significant deviations from gene collinearity are observed (Fu and Dooner, 2002; Song and Messing, 2003; Brunner et al., 2005), which can involve illegitimate recombination mediated by helicases (Lai et al., 2005). Several species of grasses have undergone whole-genome duplication (WGD) events, creating large internally duplicated regions. For example, as recently as 4.8 million years ago (mya), maize underwent a WGD by the hybridization of two progenitors (Swigoňová et al., 2004). Comparison of duplicated regions from the maize genome with the orthologous regions of rice and sorghum (*Sorghum bicolor*; whose progenitor split from the progenitors of maize only 11.9 mya) indicates that the maize genome has lost many of its duplicated genes. In addition, there is increasing evidence that a significant portion of genes in all these grass species may have moved to other locations within the genome over the last 50 million years (Lai et al., 2004b).

There are a variety of strategies for sequencing whole genomes, and part of the goal of this work was to generate a reference sequence for evaluation of an appropriate sequencing strategy for the maize genome. Suitability of a sequencing strategy to a genome depends on the character of the genome, the state of the technology, and availability of funding. Published strategies include whole-genome shotgun, clone by clone, various reduced representation shotgun (RRS) methods, and various combinations of these (Lander et al., 2001; Venter et al., 2001; Waterston et al., 2002; Bedell et al., 2005). Several new RRS strategies have been developed specifically to address the challenges posed by the high repeat content of maize, with the goal of enrichment of nonrepetitive regions prior to sequencing (Rabinowicz et al., 1999; Yuan et al., 2002; Yuan et al., 2003). Two fractionation methods were used to generate about 1 million sequence reads from the genome of the maize inbred line B73 (Palmer et al., 2003; Whitelaw et al., 2003). Effective evaluation of the performance of a genome sequencing strategy will be greatly facilitated by a high-quality, randomly selected sampling of the genome in relatively large regions (containing both genic and intergenic sequences).

To this end we randomly selected 100 bacterial artificial chromosomes (BACs) from the genome of the maize inbred line B73 for sequence analysis. They were sequenced to deep coverage and manually curated to derive an accurate consensus. This provided a high-quality reference sequence representing approximately 0.6% of the genome that can serve as a basis for both an unbiased study of genome content and evaluation of potential strategies for sequencing the whole maize genome. Based on the sequence information from this large random sampling, we undertook an assessment of the organization and structure of genes, repeat sequence families, and of the coverage by RRS datasets.

RESULTS AND DISCUSSION

Sequencing and Assembly

With the goal of sampling random regions from the maize genome, we selected 100 BAC clones from inbred B73. To avoid bias, these clones were taken from three different BAC libraries made by using different restriction digests (see Nelson et al., 2005; Supplemental Table I) and were selected without regard to any known genetic markers or genes of interest. The BACs are part of a larger collection of clones for which DNA fingerprints and BAC end sequences (BES) were generated (Messing et al., 2004). Since the fingerprinted BACs were assembled into fingerprinted contigs (FPC) and anchored to the genetic map (Cone et al., 2002), the fingerprints serve as a link from the clones to the maize physical and genetic maps. The selected BAC clones fall into three categories with regard to the information known about their map location: those that have a chromosomal location by virtue of being assembled into genetically anchored FPCs (77), those that are assembled into unanchored contigs (9), and those that are singletons from the contig-building process (14; Supplemental Table II). The anchored clones are distributed across the 10 chromosomes (Supplemental Fig. 1). BLAST (Altschul et al., 1990) analysis of the BESs from singletons shows similarity to maize transposable elements, demonstrating that the singleton clones truly represented maize nuclear genomic DNA and not contaminant sequences (e.g. plastid DNA, genomic DNA from *Escherichia coli*, or other species). This was further confirmed by analysis after complete sequencing.

We sequenced, assembled, and manually curated these clones (see ''Materials and Methods'') to generate the optimal consensus sequence, producing a high-quality dataset on which all of our analyses are based. After curation, 89 BACs yielded ordered and oriented sequence assemblies, while the remaining 11 clones are not fully ordered. One of these (AC147814) represents typical tandemly repetitive regions associated with cytogenetically defined knobs. Another BAC

clone (AC150267) not included in the set of the 100 regions contains ribosomal RNA gene sequences, illustrating that the selection process yielded also clones that are recalcitrant to assembly. Since these regions were selected at random, the BACs have a wide range of sizes (22.6–227.5 kb), with an average of 143.8 kb. The singletons are smaller overall with an average size of 82.5 kb, as compared to 163.2 kb for the mapped clones (Supplemental Table III). This is not surprising, since fingerprints of smaller clones have fewer bands and thus less information content. In selecting a clone path, sequencing larger clones would typically be chosen. The combined length of the BACs is 14.38 Mb and represents roughly 0.6% of the total maize genome or, for comparison, 3.7% of the rice genome and 12.3% of the Arabidopsis genome (Table I).

Annotation

Accurate gene annotation of maize sequences poses significant challenges. The presence of transposable elements, whether whole or fragmented, in a genome often leads to overprediction of genes. To counter this, one can remove any repeated sequences from the gene set. However, as a consequence, large gene families can be mistaken for repeat sequences, leading to underprediction of genes. Thus, our annotation methods must strike a careful balance. The 100 BAC clones were annotated using a semiautomated pipeline and additional manual inspection and adjustment of gene models (see "Materials and Methods"). To address the issue of falsely predicted genes, the potential gene models were surveyed for the presence of putative repetitive sequences (Messing et al., 2004). First, the predicted coding sequences were compared with The Institute for Genomic Research (TIGR) and Munich Information Center for Protein Sequences (MIPS) plant repeat databases to identify known repeats. Second, to complement identification of known repeats, each of the remaining predicted genes was compared against the maize BES collection to identify as yet uncharacterized high-copy sequences. Genes that aligned to >10 BES with high similarity (E value equal or lower than 10^{-30}) and did not show homology to known gene families were considered repeats and excluded from the dataset. Ten BES hits would correspond to roughly 10 to 20 copies per genome (at a level of similarity detected with a 10^{-30} or lower E value). This set of removed predicted genes represents sequences that were not effectively identified using the current repeat databases and are potentially diverged or novel repeats yet to be identified in maize. Exceptions were genes homologous to expanded gene families, for example, zinc-finger proteins, ATP-binding cassette transporters, or proteins containing pre-mRNA splicing protein (PRP) domains. Using this approach, we found evidence for the presence of a total of 330 genes in the 100 random BACs. The vast majority of these genes are well supported by at least one expressed sequence tag (EST) and/or protein database entry, and a summary sheet listing experimental support for each of our annotations is provided in Supplemental Table IV.

To minimize bias in analyses that extrapolate to the entire maize genome, we defined a high-confidence gene set (HCGS) of full-length genes with end-to-end protein database matches. For example, estimation of gene size or the coverage of genic sequences by reduced representational sequencing methods requires a full-length gene set. We have used both the complete and HCGS gene sets for distinct analyses (see below). In defining the HCGS by manual inspection of protein alignments, we identified a reference subset of 172 genes that were very similar in length and sequence to previously described proteins in the nonredundant database. The average length ratio between a reference protein and its counterpart in nonredundant was 0.97 (±0.18 s), and the amino acid identity of the alignment was 0.68 (±0.16 s). In the following, we have used both datasets in our analysis.

Table I. *Comparison of maize, rice, and Arabidopsis gene statistics*

Features	Maize	Rice Genome[a]	Arabidopsis[b]
Total no. of clones	100	3,401	1,578
Average length (kb)	144 (23–227)	n/a	n/a
Total length (Mb)	14.38	n/a	132.9
Minimal tile (Mb)	n/a	370.7	117.3
Predicted genes	330	37,544	29,084
Predicted exons	1,520	175,203	142,512
Average no. of exons per gene	4.6	4.7	4.9
Average intron size (bp)	607	413	167
Average exon size (bp)	259	254	217
Average gene size (kb)[c]	4.0[d]	2.7	2.0
Average exon density/100 kb	11		122.5
Average gene density (kb per gene)	43.5	9.9	4.0
G + C content			
Overall	46.5%	43.6%	35.9%
Exons	55.4%	54.2%	43.8%
Introns	42.3%	38.3%	32.6%
Intergenic regions	46.0%	42.9%	31.7%
Protein-coding DNA	55.4%		44.0%

[a]International Rice Genome Sequencing Project, (2005). [b]Numbers derived from MAtDB, released September 2004 (Schoof et al., 2004). [c]Length from the start codon to the stop codon. [d]Average gene size for all 330 genes (including partial gene models at the end of BACs) is 3 kb.

Gene Density and Length

We set out to use our annotations to characterize the maize gene set. HCGS gene size falls into a broad distribution ranging from under 1 kb to 59.1 kb, with an average of 4 kb and a median of 2.6 kb. The average gene size for the full set of 330 genes was 3 kb. By comparison, the gene sizes of rice (2.6 kb) and Arabidopsis (2 kb) are significantly smaller (Table I; Arabidopsis Genome Initiative, 2000; International Rice

Genome Sequencing Project, 2005). The current predicted rice proteome, however, represents FGENESH models without any manual curation and therefore may have a smaller average gene size. The difference between the HCGS and the full gene set is at least partly ascribable to partial gene models at the ends of BACs, as the size distributions of exons and introns were similar (see below and Supplemental Fig. 2). HCGS exon number varies from 1 to 31 with a median value of five exons per gene (Fig. 1A). The average exon length (259 bp; median length 138 bp) is very similar to that in rice (254 bp) and slightly longer than that of experimentally confirmed exons from Arabidopsis (217 bp). Initial annotation of rice chromosome 10 (Rice Chromosome 10 Sequencing Consortium, 2003) showed a much larger average exon size (344 bp), but this was due to the inclusion of transposon-related genetic elements (which lack introns and have large

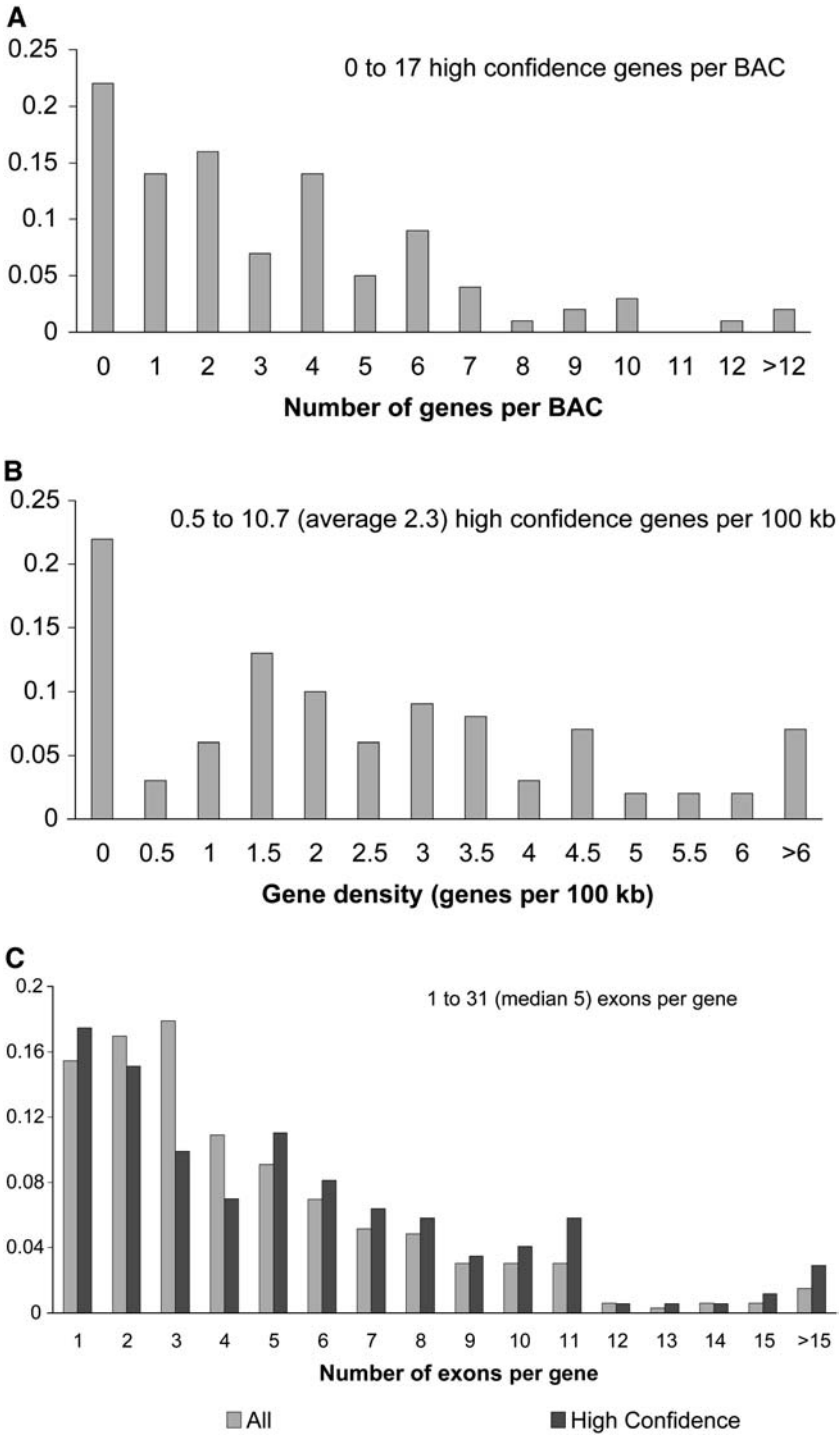

Figure 1. Gene characteristics in the 100 random regions. Graphs have been plotted to show the number of exons per gene (A), the number of genes per BAC clone (B), and the gene density expressed as number of genes per 100 kb (C).

polyproteins). The more recent value for rice exons is derived from the whole genome, and a much more rigorous annotation process (International Rice Genome Sequencing Project, 2005). Since exon size is similar, the major factor in the larger gene length in maize must be the length of introns, which have an average size of 607 bp in maize (median 166 bp) compared to 413 bp in rice. The longer average intron size in maize appears to be due to the insertion of transposable elements (see below). The size distributions of exon and intron lengths were similar between the HCGS and the full gene set (Supplemental Fig. 2).

Gene density also falls into a broad distribution. We observed between zero and 17 HCGS genes per clone (Fig. 1B). Of the 100 clones, 78 contained at least one gene, while the remaining 22 contained none. Because of the wide range in BAC sizes, it is more appropriate to use a normalized measure of gene density, such as genes per 100 kb (Fig. 1C). Using this measure, gene density varied over an 18-fold range (0.5–10.7) with an average of 1.2 genes/100 kb. Using the full set of 330 predicted genes, the average density increases to 2.3 genes/100 kb or one gene every 43.5 kb. Both of these values are markedly lower than those reported for the rice genome (one gene every 9.9 kb) and for the Arabidopsis genome (one gene every 4 kb; Table I). The broad range observed is in line with previous observations based on the sequencing of a 346-kb region containing the storage protein gene cluster on chromosome 4S in inbred BSSS53 (Song et al., 2001). Even within this single region of the genome a wide range of gene density was observed. It contains a section of 170 kb containing 25 genes and another of 70 kb containing only one.

In using the data from the random BAC clones to estimate the total density and number of genes in the maize genome, one must take into account the variability observed as well as edge effects, since the individual BACs will often contain partial genes. An alternative method would be to use the average HCGS gene length along with the predicted total gene space to calculate an approximate predicted gene number of the maize genome.

The full set of 330 genes predicted in this study cover 7% of the nucleotides in the sequenced BACs, leading to the extrapolation that the genic space for the whole 2.3-Gb genome totals approximately 167 Mb. Although this is likely to be an underestimate since partial genes are included, this number is consistent with a previous estimate of 177 Mb for the maize transcriptome (Messing et al., 2004). By using the two values for average gene size (3.0 kb from the full gene set and 4.0 kb from the HCGS), with this estimate of the total size for genic space, we estimate that the total gene number in maize is likely to be between 42,000 and 56,000 genes. Since the HCGS gene size is more likely to be representative of the whole genome, the total gene number is more likely to be at the lower end of this range. In any case, it should be noted that even the lower end of this entire range is significantly higher than the 37,544 genes in rice (International Rice Genome Sequencing Project, 2005). Although maize loses half of its duplicated genes after a WGD event (Lai et al., 2004b), one would still expect the maize genome to contain more genes than the rice genome.

GC Content

We assessed the guanine + cytosine (G + C) content of exons and introns using just the HCGS because the high level of conservation of these genes across species means the splice-site locations can be considered high confidence. Most strikingly, there were clear differences in GC content between coding and noncoding (intron and untranslated regions) sequences within genes. Exons varied from 40% to over 75% GC with a mean of 55.4% (Table I). Intronic sequences ranged from 30% to 60% with an average of 42.3%. However, there was no significant difference between the GC content of the HCGS and the full gene set.

Besides the overall difference in GC content of exons and introns, we also observed a polarity of GC content of both introns and exons decreasing in the direction of transcription, with the translational start marked by a steep increase in GC content (Supplemental Fig. 3). These observations are consistent with findings in rice (Wong et al., 2002). At this time, there is no known underlying mechanistic explanation for this observation.

Gene Expression and Codon Usage

Since expressed sequences provide the most reliable data for confirmation of gene calls, we compared our gene annotation to existing ESTs from maize and other plant genomes. The publicly available maize collection of about 397,000 ESTs has been clustered to 49,991 unigenes, although these clusters include paralogous sequences (Lai et al., 2004a). We note that these ESTs are derived from several different inbred lines (including B73) and so have some heterogeneity gene sequence, content, and expression (Song and Messing, 2003). About 85% of the 330 predicted genes from the 100 random BACs could be aligned with maize unigenes at high stringency (Fig. 2). Including other monocot-derived EST datasets increases the coverage to 91%, providing validation of our annotation set (Supplemental Table IV).

Interestingly, including ESTs from dicot and gymnosperm species (Arabidopsis, Medicago, and Pinus) yields a very different result. We compared the full set of 330 predicted genes against each EST dataset at both the DNA sequence level (using BLASTN; Altschul et al., 1990) and at the amino acid sequence level (using TBLASTN; Altschul et al., 1990). Although the fact that these collections vary in their depths has to be taken into consideration, two salient features emerged. The degree of amino acid similarity is quite similar across all plant species studied, and shows only a slight decrease proportional to the phylogenetic distance

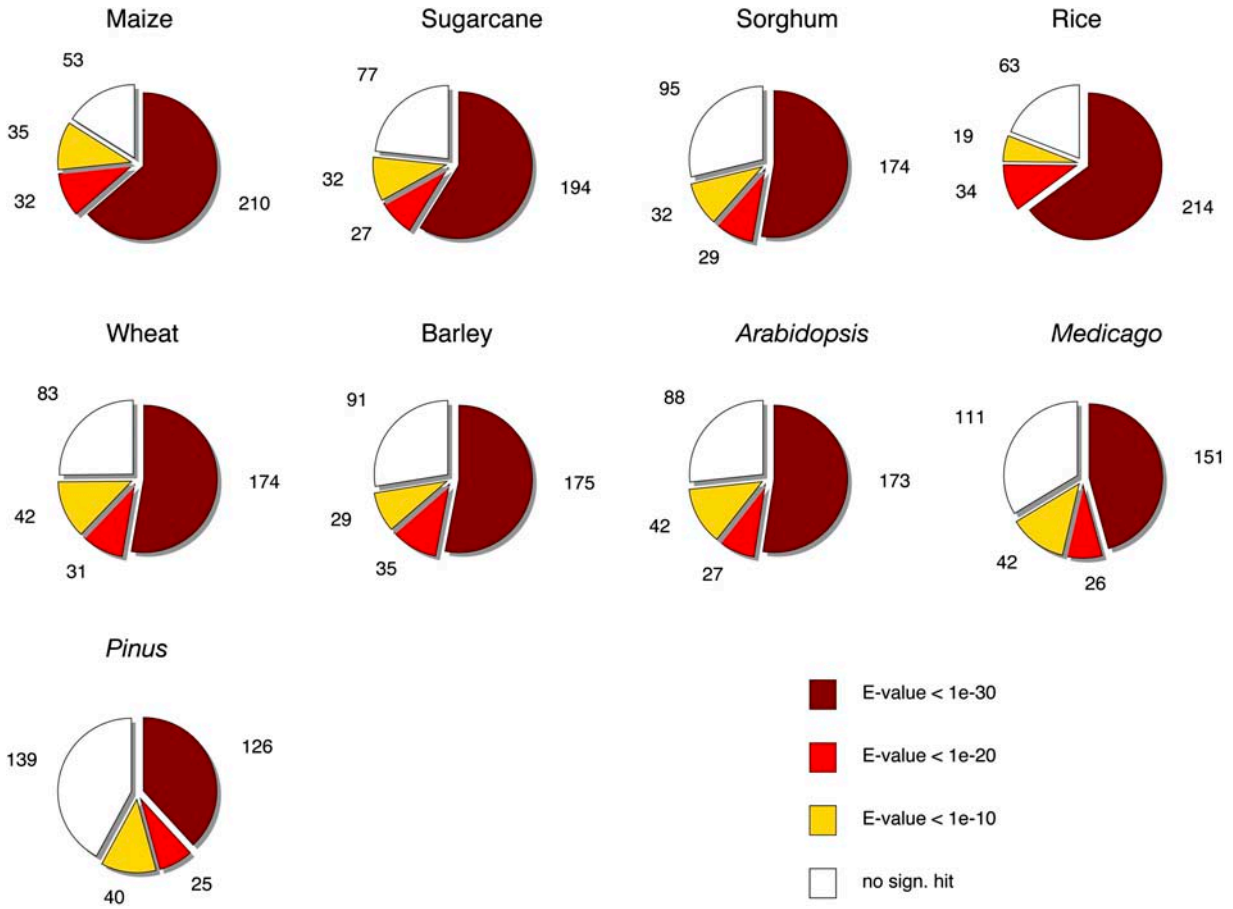

Figure 2. Comparison of genes to species-specific EST collections. Proteins derived from the gene models were compared to the EST assemblies using TBLASTN. Homologous sequences were binned into four classes: gene models with highly significant EST matches (*E* value less than 1e−30), with significant homologies (*E* value between 10^{-30} and 10^{-20}), with weak homologies (*E* values between 10^{-20} and 10^{-10}), and those exhibiting no or only very weak homologies (*E* values higher than 10^{-10}).

from maize and the depth of species-specific EST collections. By contrast, at the nucleotide level among monocots there is limited sequence divergence, but the degree of similarity drops significantly in dicots (Supplemental Fig. 4). Even the most dissimilar of the monocots has a match (<1e−10) against 62% of the 330 maize genes at the DNA level. In contrast, none of the EST sets from the other species align to more than 15% of the same set of genes.

One likely explanation for this trend is that codon usage differs greatly from monocots to dicots and gymnosperms. The marked dissimilarities in GC content between Arabidopsis and maize genes are consistent with large deviations in codon usage. For instance, maize prefers the GCC codon for Ala, while Arabidopsis prefers the GCT codon (Supplemental Table V; Supplemental Fig. 5). Knowledge of codon usage has been critical in the design of transgenes to be expressed in plants. For instance, the huge success with producing maize varieties resistant to European corn borer was largely based on synthesizing a gene for an insect-toxin protein from *Bacillus thuringensis* using codons preferred by the plant host (Perlak et al., 1991). Our results suggest that in order for a transgene to be expressed properly in nonmonocot host, it would need to have different codon usage.

Spatial Distribution of Genes

The distribution of maize genes relative to repeat sequences has been the object of much interest. Distribution of genes across a sample of 10 clones is shown in Figure 3. Our data show that in almost all cases, single or at most a few genes are separated by repeat elements, although it is possible that larger clusters of genes will be found when longer contiguous sequences become available. This raises some questions about the widely accepted theory that the maize genome consists of gene islands separated by large blocks of repeat elements (SanMiguel and Bennetzen, 1998; Yuan et al., 2002), such as defining the size of a typical gene island.

Repeat Elements

Our random sample of 0.6% of the maize genome allows us a relatively unbiased view of its repeat

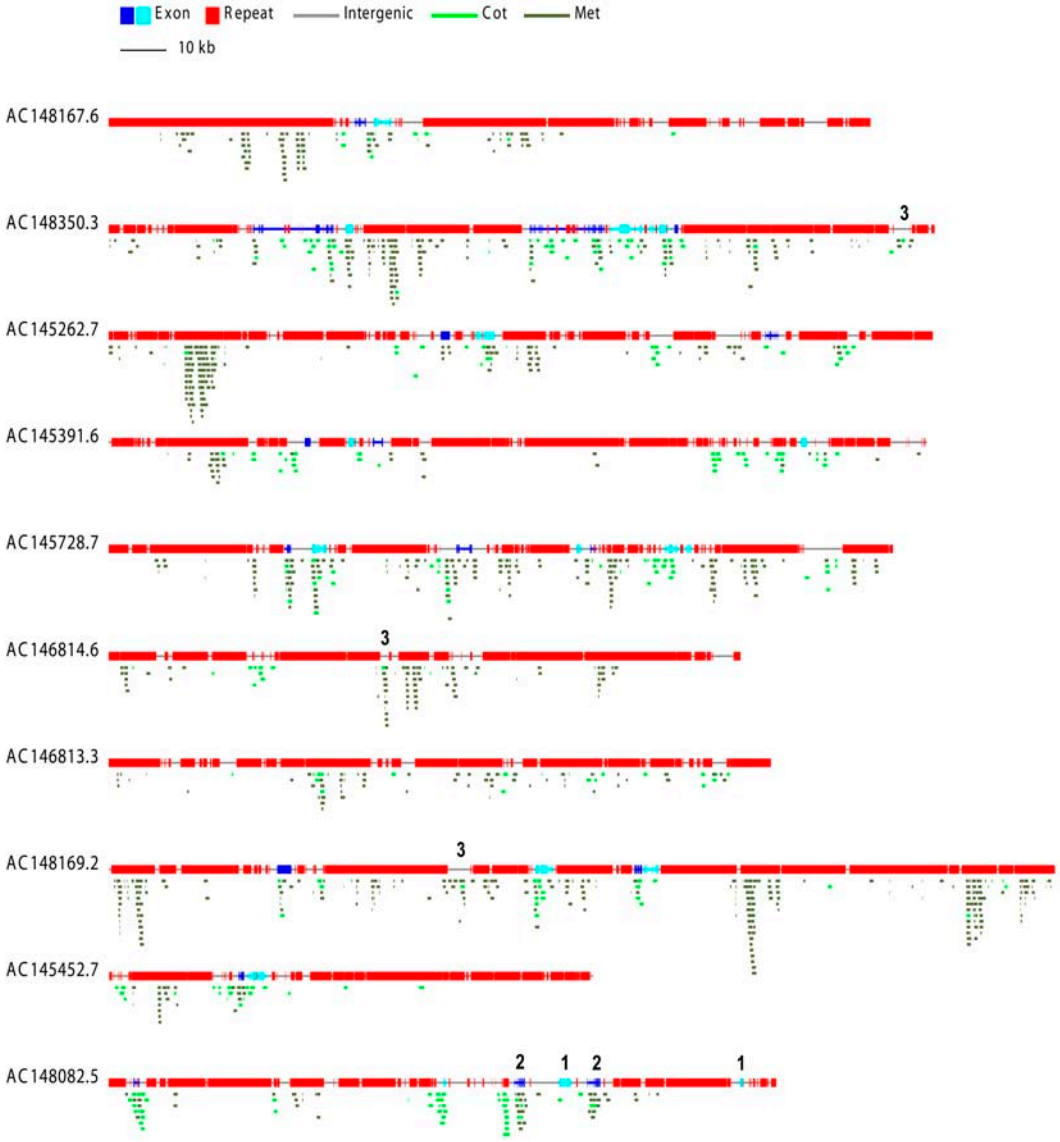

Figure 3. Graphic representation of a sample of annotated BAC clones. Ten out of 100 annotated BAC clones are arranged as bars depicting genes (blue) and regions containing repeat sequences (red). A straight gray line represents intergenic regions with no predicted gene models. To determine the coverage of our annotations by the collection of methyl- and C_0t-filtered sequence reads, we compared the BAC sequences against the respective collections obtained from TIGR (http://www.tigr.org/tdb/tgi/maize/). All filtered sequence reads were mapped to the 100 BAC sequences by BLASTN sequence comparison and subsequent quality parsing. To anchor a clone to a genomic location, a minimal sequence identity of 98% over the complete alignment length and an alignment length equal or greater than 90% of the clone length were required. Sequence matches from methyl/C_0t-filtered sequence reads are depicted in dark green and light green, respectively. Specific features are highlighted with consecutive numbers: (1) examples of low gene coverage by filtered sequences, (2) example of tandem gene copies representing highly similar hydrolases for which GSS tags could be unequivocally mapped, and (3) nonrepeat intergenic region well covered by filtered sequences.

content. Previous characterizations of the repeat content of the maize genome were based on genome survey sequences (GSS; Meyers et al., 2001; Messing et al., 2004). These approaches failed to fully describe many of the longer elements due to their reliance on single read sequences.

Analyses based on fully sequenced BACs give us the opportunity to study full-length repeats. BAC sequences were screened for repeat elements using RepeatMasker (www.repeatmasker.org; A.F.A. Smit and P. Green, RepeatMasker, version 2.1) with a customized plant repeat library (http://mips.gsf.de). The underlying repeat sequences were compiled from different sources, clustered into a nonredundant set of 5,707 sequences, and classified by a hierarchical repeat classification scheme. This repeat library was then used to mask and classify repeat sequences in BAC clones. Based on this analysis, we found the known repeat content of the 100 random BACs to be about 66% (Table II), somewhat higher than the estimates of 58% repeat elements from

Table II. *Occurrence and distribution of repetitive DNA in maize and rice BACs*

Details	Maize			Rice[a]		
	No. of Hits	No. of Bases	% of Genome[b]	No. of Hits	No. of Bases	% of Genome[c]
		bp			*bp*	
Class I retroelements	5,223	9,116,674	63.39%	11,859	6,736,074	18.82%
Ty1/*copia*-like elements	1,577	2,979,969	20.72%	4,994	1,313,472	3.67%
Ty3/*gypsy*-like elements	2,080	4,383,700	30.48%	2,866	3,187,856	8.90%
LINES	56	12,817	0.09%	160	41,210	0.12%
SINES	28	2,306	0.02%	1,031	137,363	0.38%
Other retroelements	1,482	1,737,882	12.08%	2,808	2,056,173	5.74%
Class II DNA transposons	763	184,083	1.28%	14,436	3,704,904	10.35%
hAT superfamily	66	10,612	0.07%	525	133,421	0.37%
CACTA superfamily	80	38,714	0.27%	1,185	1,012,392	2.83%
Mutator	30	2,590	0.02%	894	183,352	0.51%
Tourist-like MITEs	65	8,195	0.06%	1,565	327,362	0.91%
Other MITEs	261	35,735	0.25%	5,565	1,091,707	3.05%
Other DNA transposons	261	88,237	0.61%	4,702	956,670	2.67%
Simple repeats	326	173,354	1.21%	655	368,813	1.03%
High-copy-number genes	11	2,068	0.01%	89	20,252	0.06%
Other repeats	43	11,760	0.08%	1,401	153,272	0.43%
Total repeats	6,366	9,487,939	65.97%	28,440	10,983,315	30.68%

[a]A total of 175 pseudo BACs (i.e. 200 kb cut equally from all 12 chromosomes; Messing et al., 2004). [b]Maize genome = 2,365 Mb. [c]Rice genome = 389 Mb (International Rice Genome Sequencing Project, 2005).

BES representing one-eighth-fold coverage of the genome (Messing et al., 2004). We note that a small subset of the repeat elements (2%) is located within introns (see below).

The end sequences of a 50,000-member small insert library of sheared genomic DNA (Whitelaw et al., 2003) provided a repeat estimate for the genome of 63% (Messing et al., 2004). However, by the same method described above, we evaluated 117 nonrandom maize BAC clones present in GenBank and found they contain only 53% repeats, illustrating the bias that arises when clones are selected for containing a gene of interest. We believe our random selection of BACs is more likely to represent the genome as a whole.

As shown by the graphical distribution of repeat sequences, contiguous repetitive regions are frequently interrupted by regions that contain neither repeats nor genes. It is possible that these represent members of repeat or gene families that have degenerated beyond detection or functional sequences not yet well defined. These regions make up more than a quarter of the genome.

Of the full set of 330 genes, 34 genes (10.3%; 11.6% for the HCGS) harbor repeats within their introns. The detected repeat types within introns differed significantly from the overall repeat content in maize. About one third of these repeats belong to DNA transposons as compared to 1.28% for the entire genome, indicating a substantial enrichment of this repeat type within introns. Figure 4 shows three gene models that contain repeat elements in their introns.

To compare repeat content in maize and rice, we selected a similar number of random BACs from rice subsp. *japonica* cv Nipponbare and subjected them to the same analysis (175 pseudo BACs, i.e. 200 kb cut equally from all 12 chromosomes; see Messing et al., 2004). As expected, the repeat content found in rice is much smaller (31%; Table II), which is quite close to

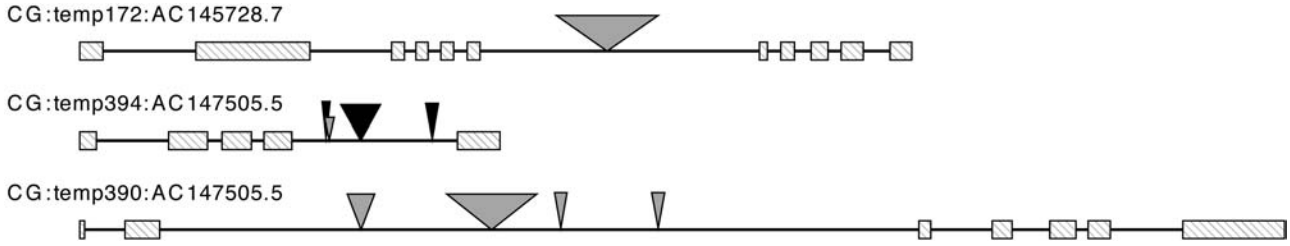

Figure 4. Three examples of genes containing repetitive sequences within their introns. CG:temp172:AC145728.7 represents an ATPase II-like protein, CG:temp394:AC147505.5 an unknown protein containing a conserved PER1 domain, and CG:temp390:AC147505.5 a protein containing two cyclin K domains. Exons are shown as striped bars, introns as black lines, and repetitive sequences as triangles. DNA transposons are represented by black and retroelements by gray triangles. CG:temp172:AC145728.7 contains a retroelement, CG:temp394:AC147505.5 three DNA transposons (tourist-, Castaway-, and MITE-*adh*-like elements) and one retroelement, and CG:temp390:AC147505.5 two copies of Ty/*copia* elements and SINEs, respectively, within their introns.

the whole rice genome statistics of 35% (International Rice Genome Sequencing Project, 2005). These findings are also consistent with results obtained by comparisons of orthologous regions between maize and rice, which exhibited insertions of retrotransposons in maize but rarely in rice (Lai et al., 2004b). Our analysis shows the predominant repetitive elements in maize are the long terminal repeat (LTR) class I retroelements. The amount of repetitive DNA is partly explained by the type of repeats present—particularly since class I elements are approximately 50 times more abundant than class II elements. The relative amounts of class I elements to class II elements differ greatly between rice (18.8%:10.4%) and maize (63.4%:1.3%; Table II; Messing et al., 2004), particularly the Ty3/*gypsy*-type and Ty1/*copia*-type elements that occupy more than half of the genome. Ty3/*gypsy* elements have been found in centromeric and heterochromatic regions, where they are mixed with short repeats (Wu et al., 2004).

Despite the significant expansion of known repeat families from rice to maize, it is not sufficient to fully explain the size difference between their genomes. In rice, 69% are repeat free, which totals 276 Mb compared to 34% in maize, totaling 804 Mb. This 3-fold increase of repeat-free sequence can in part be explained by the WGD event, which occurred as recently as 4.8 mya through the hybridization of two closely related ancestors of maize (Swigoňová et al., 2004).

Gene-Enrichment Methods

The high density of repeat sequences, low gene density, and small average gene size of the maize genome make alternative gene-enrichment sequencing strategies very attractive. To test the effectiveness of this approach, we have aligned sequence reads/contigs (GSSs) derived from two gene-enrichment protocols (Whitelaw et al., 2003), based on methyl filtering and $C_o t$ enrichment, to the 100 genomic regions. Both of the GSS datasets were derived from the same maize inbred (B73) as the 100 random BAC clones. By aligning the GSS sequences to the sequenced BAC clones, we evaluated the proportion of genes represented in the GSS collections as well as the distribution of GSS coverage of exons within the genes. It is essential for this analysis that we are able to discriminate between sequences originating from closely related paralogs. Given that many maize genes are thought to be in families with closely related paralogs (Messing et al., 2004) and that error rates of the GSS have been estimated as low as 2.3×10^{-3} (Fu et al., 2004), very stringent parameters (98% identity over at least 90% length) were used to compare sequence reads derived from these gene-enrichment methods to the sequenced BAC clones. GSS sequences, unlike ESTs, are genomic in origin and contain intronic and intergenic regions, which can be used in alignments against genomic sequences to differentiate all but the most similar of paralogs.

About 93% of the HCGS had at least one corresponding alignment within the GSS collection (Fig. 5), which is slightly higher than maize EST coverage (85%, as described above). However, only 29% of the genes have GSS alignments covering greater than 90% of their length. This result is similar to a previous study of 78 full-length cDNAs (FLCs) reporting that at least 95% aligned to at least one GSS and about 18% of the FLCs were completely covered (Springer et al., 2004). However, unlike the previous study, this analysis can distinguish between paralogs because we examined the entire gene, including intronic sequences absent in the FLCs. At the nucleotide level, the methyl-filtered reads cover 28% of the nucleotides of the genes, while the $C_o t$-enriched reads cover 34.5%. The combined enriched datasets cover 51%, illustrating the significant complementarity of the two methods.

Alignments of GSSs against annotated BAC clones (Fig. 3) revealed deep clusters of filtered sequence reads occurring both in genes and in intergenic regions—both repetitive (e.g. in BAC AC145262.7) and nonrepetitive (e.g. in BAC AC148169.2). The clusters in introns show that a significant percentage of genes contain repetitive elements (11%) such as solo LTRs or miniature inverted-repeat transposable elements (MITES) present in intronic sequences as shown above (Fig. 5). One can envision that such genes might be underrepresented by enrichment procedures. Indeed, although 94% of genes in our analysis that contain repeats in their introns were tagged by at least one GSS, their total coverage was relatively lower than the coverage for all genes (40% and 51%, respectively). The upstream and downstream sequences of genes showed decreased coverage by the GSS (Supplemental Fig. 6), although we did not observe any pronounced gradients of coverage internal to genes. As a result, UTRs and promoter sequences may be underrepresented in GSS sequences. In addition, GSS clusters that do not represent known repeats are worthy of further study, as they may either identify previously unknown repeats or a particular bias of the GSS datasets.

The two GSS datasets together have tagged the majority of the analyzed genes with at least one read, demonstrating that these methods provide a significant enrichment and enable exploration of the genic space in maize. However, upstream and downstream sequences as well as genes containing intronic repeats are underrepresented. Full-length sequences of these biologically important regions may therefore require other sequencing approaches, such as traditional shotgun sequencing of BAC clones. Alignments also show that hypomethylated DNA sequences are not restricted to gene sequences. Recently, it was shown that certain retrotransposon element families are not only hypomethylated but also transcribed (Messing et al., 2004). Since genes can be differentially methylated because of epimutations, paramutation, genomic imprinting, and tissue specificity (Lund et al., 1995; Alleman and Doctor, 2000; Lisch et al., 2002; Guo et al., 2003), all of which may affect their representation in methyl-filtered data, factors such as tissue type and developmental stage should be considered in selecting source

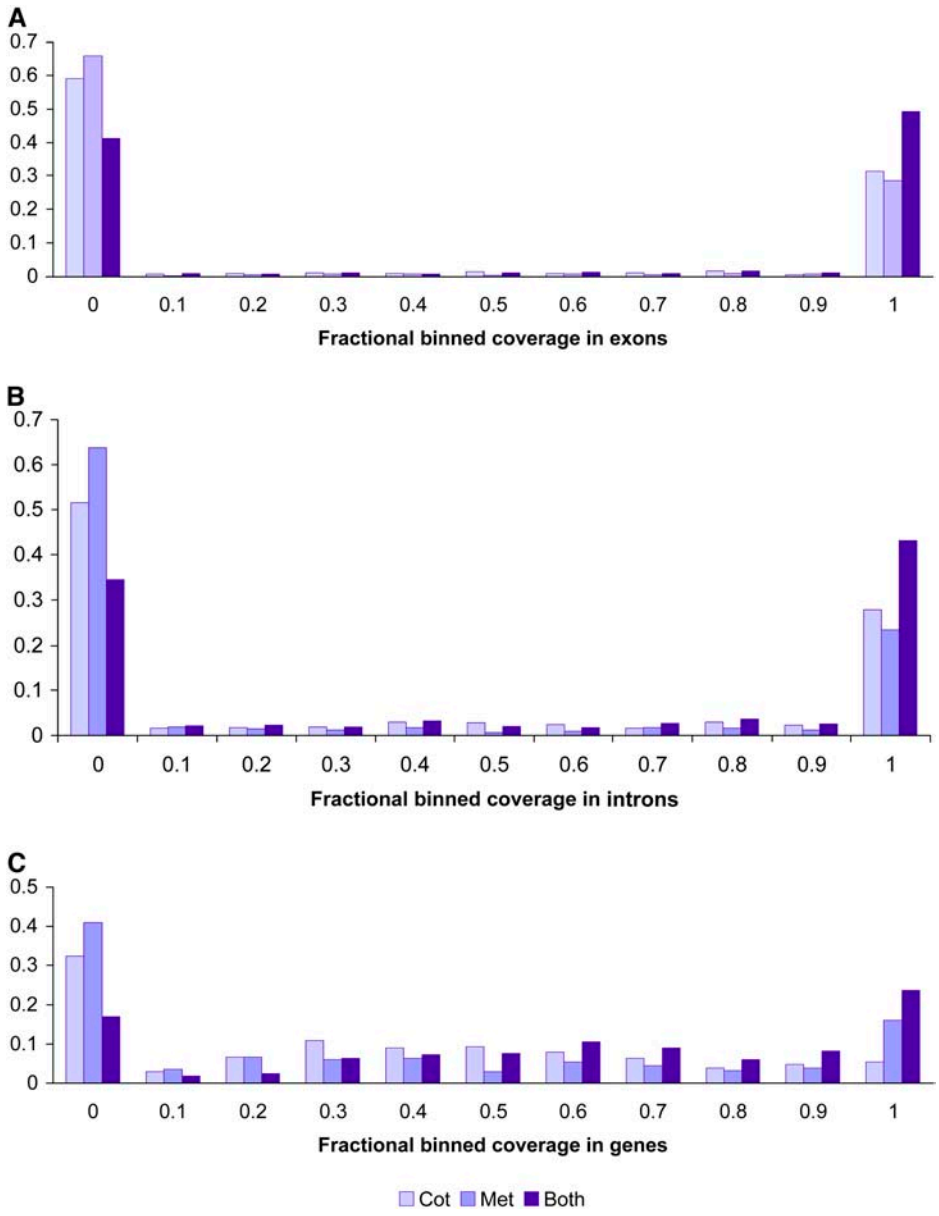

Figure 5. Coverage of exonic, intronic, and genic sequences by methyl- and C_0t-filtered sequence reads. Coverage was determined as described in Figure 3, and results were sorted into bins of size 10% of fractional coverage. Fractional coverage for exons, introns, and complete genes by methyl-filtered, C_0t-filtered, and combined sequences are shown. A to C depict the values obtained for exonic, intronic, and genic coverage. Bars in medium blue show values obtained for methyl-filtered sequence reads, bars in light blue values for C_0t-filtered clones, and dark blue bars depict cumulative values.

material. Nevertheless, where filtered sequences align at high stringency, they can provide important gap-filling functions or sequence extensions as was recently shown for the *tb1* locus (Lai et al., 2004b).

CONCLUSION

With the goal of gaining a relatively unbiased view of the maize genome, we have sampled 100 randomly selected BACs representing 0.6% of the genome, defined their content of genes and repeats, and used these data to characterize the structure and architecture of the maize genome. The maize genome is substantially larger than those of two previously sequenced plant genomes, Arabidopsis and rice. Our work shows this to be a function of the repeat, gene, and intergenic content of maize.

Our analysis shows that at least 66% of the genome consists of repetitive elements. This is a lower bound, since there are undoubtedly additional repeats in the genome including sequences that have not yet been characterized or that have diverged too far from known repeats. Retrotransposons are far more frequent than DNA transposons in the maize genome, while in rice the opposite is true. Since retrotransposons are so much larger, this partially explains the significant difference in the sizes of the maize and rice genomes. Repeats are found in the introns of 11% of genes, which explains the relative increase in size of introns compared to exons of rice and Arabidopsis.

The repeat types found within introns tend to be short, with a higher frequency of DNA transposons than the rest of the genome, and frequent occurrence of solo LTRs, indicating a possible selective pressure against large elements in these maize introns.

Of the BACs sequenced in this study, 80% were found to contain genes. Full-length genes average 4 kb in length, somewhat larger than in rice and Arabidopsis. Longer introns in maize, due in part to transposon insertions, are responsible for most of the increase in gene size. The density of genes is widely variable, ranging from 0.5 to 10.7 genes per 100 kb over a relatively even distribution, and does not suggest that a large fraction of genes are tightly clustered in islands. Based on these data, we estimate that maize has roughly 42,000 to 56,000 genes, substantially more than rice or Arabidopsis. This reflects the history of the maize genome, which includes a relatively recent WGD event, subsequent gene loss, and expansion of gene families. The WGD also appears to contribute to an increase in intergenic space void of apparent repeat sequences.

In contrast to sequencing large stretches of genomic DNA, previous samplings of the maize genome focused on methods designed to enrich unique sequences relative to repeats. Available datasets from two such methods were evaluated against our representative gene set. We found that although 93% of genes are at least partially represented in the enriched sample, less than 30% of genes are fully covered by the enriched data. Further, biases exist that indicate that not all sequences of biological interest will be obtained easily. Our results suggest that filtering methods aimed at separating genes from the rest of the genome are an efficient way to begin to sample unique sequences in the maize genome, but will probably be of limited effectiveness for generating a complete representation of the maize gene set due to inherent biases in the data.

Our data show that generating high-quality sequences from large insert clones is an effective method for sampling the repeat and gene content of the maize genome. Further, because maize BACs are linked to the physical map, they provide a resource to generate anchored sequences of the genome.

MATERIALS AND METHODS

Selection of Clones

Genomic libraries of maize (*Zea mays*) inbred B73 have been constructed in BACs with three different enzymes, *Hin*dIII, *Eco*RI, and *Mbo*I (see Nelson et al., 2005). The libraries have been quality tested with a core set of probes (Yim et al., 2002) and represent the maize genome with a 29-fold coverage based on a genome size of 2.365 Gb (Rayburn et al., 1993). To obtain clones from different random regions of the maize genome, 25 BACs from each of the three libraries were selected from the entire pool of 464,544 BACs that had been fingerprinted and end sequenced (Supplemental Table I). Based on their shared restriction fragments with overlapping BAC clones, the integrity of the first 75 clones selected from FPCs (whole-genome maize physical map; http://www.genome.arizona.edu/fpc/maize) was evident. This analysis would also exclude clones from contaminated DNAs (e.g. organellar or bacterial DNA). An additional 25 clones from the *Mbo*I library, which were singletons and did not assemble into contigs using FPC (Soderlund et al., 2002), were also selected for sequencing. To ensure that these BACs represent maize genomic DNA, their end sequences have been checked for characteristics of maize DNA by BLAST analysis (Altschul et al., 1990). Before shotgun library construction and sequencing, the *Hin*dIII agarose fingerprint profiles of all 100 BACs were cross matched with preexisting profiles.

DNA Sequencing, Sequence Assembly, and Deposit

BAC DNA was sheared into random fragments and size fractionated. Two different sizes (4 kb and 10 kb) were selected. Care was taken to hold inserts of shotgun libraries within a narrow size range. Inserts were sequenced from both ends using universal primers (Vieira and Messing, 1982), ABI 3730 capillary sequencers, and the ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction kit (Applied BioSystems). Trace files have been analyzed by the Arachne assembly program (Batzoglou et al., 2002; Jaffe et al., 2003) and deposited in GenBank. Curation was carried out by a combination of repeated reassembly with varying parameters along with manual inspection and editing. After curation, 89 BACs had known order and orientation (i.e. Phase 2), and 11 had contigs of unknown order (Phase 1). Only two BACs contain potential misassemblies, one of them (AC147814) being extremely repetitive due to knob repeat sequences while the other (AC147604) remains unresolved. For the remaining 98 correctly assembled and curated BACs, every base in the assembly is of finished quality.

Sequence Analysis

A repeat sequence library was built as described in the text and used to mask the BAC sequences that were then analyzed for their coding potential by applying extrinsic (homology based) and intrinsic (ab initio gene prediction methods) criteria and methods. As a first pass, potential gene models required either homology to known genes/ESTs or prediction by at least two gene finders. Genes were detected by applying FGeneSH++ (Salamov and Solovyev, 2000; Softberry) and GenemarkHMM (Lukashin and Borodovsky, 1998). In addition, BLAST homology searches of the respective BAC sequences against EST assemblies and protein sequences were carried out. EST collections included assemblies of Arabidopsis (*Arabidopsis thaliana*), *Medicago truncatula*, *Triticum aestivum*, sorghum (*Sorghum bicolor*), *Hordeum vulgare*, *Saccharum officinalis*, rice (*Oryza sativa*), and maize (TIGR Gene Index Database at http://www.tigr.org/tdb/tgi). Spliced alignments of the EST sequences were obtained by the GeneSeqer program (Usuka et al., 2000). Mapping of homologous proteins was carried out by BLASTX sequence comparisons of the whole BAC genomic sequence against a protein database consisting of the complete Arabidopsis genome (Schoof et al., 2004), 31,654 proteins derived from a rice full-length cDNA collection (KOME; http://cdna01.dna.affrc.go.jp/cDNA/CDNA_main_front.html) and the SWISSPROT protein database (Boeckmann et al., 2003). The annotations for each BAC can be accessed online or downloaded in the Apollo-compatible GameXML format from the MIPS maize database (http://mips.gsf.de/proj/plant/jsf/maize/index.jsp; sequence analysis and annotated gene models at MIPS).

Coverage of Filtered Clones

The methyl- and C_0t-filtered sequence reads available at TIGR (http://www.tigr.org/tdb/tgi/maize/) were used to determine the coverage of genes by the filtered sequence reads. All filtered sequence reads were compared against the 100 BACs by BLASTN sequence comparison. To anchor a clone to a genomic location, an alignment length of at least 90% of the clone length and a minimal sequence identity of 98% over the alignment length were required. Genomic/exonic/intronic coverage was determined on a nucleotide basis and was normalized to the length of the respective segment.

Web Sites Referenced

The following is a list of the Web sites referenced in this study: www.broad.mit.edu/annotation/plants/maize/randomclones.html (sequence and assembly data for the 100 random clones); pgir.rutgers.edu (the Plant Genome Initiative at Rutgers, sequencing the maize genome project); and www.maizeseq.org (the DuPont/Monsanto/Ceres maize Sequence Information Sharing program).

The list of the 100 accessions deposited into GenBank can be found in the Supplemental Table II.

Received July 21, 2005; revised September 11, 2005; accepted October 5, 2005; published December 9, 2005.

LITERATURE CITED

Ahn S, Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. Proc Natl Acad Sci USA **90:** 7980–7984

Alleman M, Doctor J (2000) Genomic imprinting in plants: observations and evolutionary implications. Plant Mol Biol **43:** 147–161

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

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

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. Genome Res **12:** 177–189

Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, Jones J, Flick E, Rholfing T, Fries J, Bradford K, et al (2005) Sorghum genome sequencing by methylation filtration. PLoS Biol **3:** e13

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31:** 365–370

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell **17:** 343–360

Brunner S, Keller B, Feuillet C (2003) A large rearrangement involving genes and low-copy DNA interrupts the microcollinearity between rice and barley at the *Rph7* locus. Genetics **164:** 673–683

Chen M, SanMiguel P, de Oliveira AC, Woo S-S, Zhang H, Wing RA, Bennetzen JL (1997) Microcolinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. Proc Natl Acad Sci USA **94:** 3431–3435

Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG, et al (2002) Genetic, physical, and informatics resources for maize: on the road to an integrated map. Plant Physiol **130:** 1598–1605

Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proc Natl Acad Sci USA **96:** 8265–8270

Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implication in maize. Proc Natl Acad Sci USA **99:** 9573–9578

Fu Y, Hsia AP, Guo L, Schnable PS (2004) Types and frequencies of sequencing errors in methyl-filtered and high $C_o t$ maize genome survey sequences. Plant Physiol **135:** 2040–2050

Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci USA **95:** 1971–1974

Guo M, Rupe MA, Danilevskaya ON, Yang X, Hu Z (2003) Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. Plant J **36:** 30–44

Hulbert SH, Richter TE, Axtell JD, Bennetzen JL (1990) Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes. Proc Natl Acad Sci USA **87:** 4251–4255

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

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature **436:** 793–800

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res **13:** 91–96

Lai J, Dey N, Kim C-S, Bharti AK, Rudd S, Mayer KFX, Larkins BA, Becraft P, Messing J (2004a) Characterization of the maize endosperm transcriptome and its comparison to the rice genome. Genome Res **14:** 1932–1937

Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA **102:** 9068–9073

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

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al (2001) Initial sequencing and analysis of the human genome. Nature **409:** 860–921

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M (2004) Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166:** 935–945

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

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res **26:** 1107–1115

Lund G, Das OP, Messing J (1995) Tissue-specific DNase I-sensitive sites of the maize *P* gene and their changes upon epimutation. Plant J **7:** 797–807

Messing J (2005) Maize genomics. *In* D Leister, ed, Plant Functional Genomics. Haworth's Food Products Press, Binghamton, NY, pp 279–303

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund C, Mayer KFX, et al (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA **101:** 14349–14354

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res **11:** 1660–1676

Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution: grasses, line up and form a circle. Curr Biol **5:** 737–739

Nelson WM, Bharti AK, Butler E, Wei F, Fuks G, Kim HR, Wing RA, Messing J, Soderlund CA (2005) Whole-genome validation of high information content fingerprinting. Plant Physiol **139:** 27–38

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science **302:** 2115–2117

Perlak FJ, Fuchs RL, Dean DA, McPherson SL, Fischhoff DA (1991) Modification of the coding sequence enhances plant expression of insect control protein genes. Proc Natl Acad Sci USA **88:** 3324–3328

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

Ramakrishna W, Dubcovsky J, Park Y-J, Busso C, Emberton J, SanMiguel P, Bennetzen JL (2002a) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics **169:** 1389–1400

Ramakrishna W, Emberton J, SanMiguel P, Ogden M, Llaca V, Messing J, Bennetzen JL (2002b) Comparative sequence analysis of the sorghum *Rph* region and the maize *Rp1* resistance gene complex. Plant Physiol **130:** 1728–1738

Rayburn AL, Biradar DP, Bullock DG, McMurphy LM (1993) Nuclear DNA content in F1 hybrids of maize. Heredity **70:** 294–300

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

Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. Genome Res **10:** 516–522

SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot (Lond) **82:** 37–44

Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KF (2004) MIPS *Arabidopsis thaliana* Database (M*At*DB): an integrated biological knowledge resource for plant genomics. Nucleic Acids Res **32:** D373–D376

Soderlund C, Humphrey S, Dunhum A, French L (2002) Contigs built with fingerprints, markers and FPC V4.7. Genome Res **10:** 1772–1787

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

Song R, Llaca V, Messing J (2002) Mosaic organization of orthologous sequences in grass genomes. Genome Res **13:** 1549–1555

Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. Proc Natl Acad Sci USA **100:** 9055–9060

Springer NM, Xu X, Barbazuk WB (2004) Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. Plant Physiol **136:** 3023–3033

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

Tarchini R, Biddle P, Wineland R, Tingy S, Rafalski A (2000) The complete sequence of 340 kb DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12:** 381–391

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc Natl Acad Sci USA **96:** 7409–7414

Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignments of homologous cDNA to a genomic DANN template. Bioinformatics **16:** 203–211

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al (2001) The sequence of the human genome. Science **291:** 1304–1351

Vieira J, Messing J (1982) The pUC plasmids, an M13mp7 derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene **19:** 259–268

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520–562

Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, et al (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science **302:** 2118–2120

Wong GKS, Wang J, Tao L, Tan J, Zhang JG, Passey DA, Yu J (2002) Compositional gradients in gramineae genes. Genome Res **12:** 851–856

Wu J, Yamagata H, Hayashi-Tsugane M, Hijishita S, Fujisawa M, Shibata M, Ito Y, Nakamura M, Sakaguchi M, Yoshihara R, et al (2004) Composition and structure of the centromeric region of rice chromosome 8. Plant Cell **16:** 967–976

Yim YS, Davis GL, Duru NA, Musket TA, Linton EW, Messing JW, McMullen MD, Soderlund CA, Polacco ML, Gardiner JM, et al (2002) Characterization of three maize bacterial artificial chromosome libraries toward anchoring of the physical map to the genetic map using high-density bacterial artificial chromosome filter hybridization. Plant Physiol **130:** 1686–1696

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

Yuan Y, SanMiguel PJ, Bennetzen JL (2003) High-$C_o t$ sequence analysis of the maize genome. Plant J **34:** 249–255; erratum **Yuan Y, SanMiguel PJ, Bennetzen JL** (2003) Plant J **36:** 430