Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes

Yan Fu*[†‡], Scott J. Emrich[‡§¶], Ling Guo[†§], Tsui-Jung Wen[‖], Daniel A. Ashlock[§**††‡‡], Srinivas Aluru[§¶**§§], and Patrick S. Schnable*[†§‖**§§¶¶]

*Interdepartmental Genetics Graduate Program, [§]Interdepartmental Bioinformatics and Computational Biology Graduate Program, **L. H. Baker Center for Bioinformatics and Biological Statistics, [§§]Center for Plant Genomics, and Departments of [†]Genetics, Development, and Cell Biology, [¶]Electrical and Computer Engineering, [‖]Agronomy, and [††]Mathematics, Iowa State University, Ames, IA 50011

Edited by Susan R. Wessler, University of Georgia, Athens, GA, and approved July 5, 2005 (received for review April 26, 2005)

Recent sequencing efforts have targeted the gene-rich regions of the maize (*Zea mays* L.) genome. We report the release of an improved assembly of maize assembled genomic islands (MAGIs). The 114,173 resulting contigs have been subjected to computational and physical quality assessments. Comparisons to the sequences of maize bacterial artificial chromosomes suggest that at least 97% (160 of 165) of MAGIs are correctly assembled. Because the rates at which junction-testing PCR primers for genomic survey sequences (90–92%) amplify genomic DNA are not significantly different from those of control primers (≈91%), we conclude that a very high percentage of genic MAGIs accurately reflect the structure of the maize genome. EST alignments, *ab initio* gene prediction, and sequence similarity searches of the MAGIs are available at the Iowa State University MAGI web site. This assembly contains 46,688 *ab initio* predicted genes. The expression of almost half (628 of 1,369) of a sample of the predicted genes that lack expression evidence was validated by RT-PCR. Our analyses suggest that the maize genome contains between ≈33,000 and ≈54,000 expressed genes. Approximately 5% (32 of 628) of the maize transcripts discovered do not have detectable paralogs among maize ESTs or detectable homologs from other species in the GenBank NR nucleotide/protein database. Analyses therefore suggest that this assembly of the maize genome contains approximately 350 previously uncharacterized expressed genes. We hypothesize that these ''orphans'' evolved quickly during maize evolution and/or domestication.

assembly validation | gene prediction | maize genome assembly | nearly identical paralog

Maize (*Zea mays* L.) is the best-studied model for cereal biology and one of the world's most important crops. Most of the maize genome consists of highly repetitive sequences; consequently, the genes in this plant comprise only 10–15% of its genomic DNA (1, 2). Because of its large repetitive fraction, the National Science Foundation funded the Maize Genomics Consortium to test two distinct filtration strategies for sequencing the "gene-rich" portion of the maize genome: methylation filtration (MF) and high C_0t (HC) selection. To date, these pilot projects have generated and deposited into GenBank 450,166 MF sequences, 445,541 HC sequences, and 50,877 random shotgun sequences as genomic survey sequences (GSSs). MF and HC strategies have proven effective in selectively recovering maize genes not captured by EST projects (3, 4).

The assembly of these GSSs into genomic contigs significantly increases their utility. Our group developed a genome assembly pipeline based on innovative parallel algorithms that can quickly assemble hundreds of thousands of nonuniformly generated genomic fragments, such as MF and HC sequence reads, in a few hours (5). A key advantage of our parallel genome assembly pipeline is that the speed with which assemblies can be generated allows experimentation on the assembly process *per se*. Specifically, this speed makes it possible to determine the effects of different assembly parameter values on the quality of the resulting assemblies.

Three research groups currently provide publicly available partial maize genome assemblies based on the GSS data [The Institute for Genomic Research (TIGR), Plant Genome Database, and our group]. To our knowledge, none of these assemblies has been subjected to systematic studies into the quality of the resulting genomic contigs, nor have attempts been made to validate the structures of potentially novel maize genes found in these assemblies that have to date eluded discovery via the extensive maize EST projects. Structure validation will provide data that can be used to design strategies to assemble the maize genome (6).

The current study reports improvements to the quality of the sequence data used for assembly and the assembly pipeline used to generate our maize assembled genomic islands (MAGIs). Computational and biological quality assessments indicate that a high percentage of the MAGIs accurately reflect the structure of the maize genome. In addition, we estimate that this assembly of the maize gene space has "tagged" >6,900 expressed genes that previously lacked evidence of transcription and that almost 350 of these genes are "orphans"; i.e., they do not exhibit similarity to genes in other species. This large-scale application of RT-PCR for the verification of the expression of predicted monocot genes is a step to developing a framework for the subsequent annotation of the entire maize transcriptome. Based on the results of these RT-PCR experiments, we estimate that the B73 genome contains between ≈33,000 and ≈54,000 expressed genes.

Materials and Methods

Maize GSS Retrieval, Trimming and Repeat Masking. Genomic Survey Sequence (GSS) sequence and quality score files generated by the Maize Genome Sequencing Consortium (Danforth Center, TIGR, Purdue University, and Orion Genomics) from the *Zea mays* inbred line B73 were downloaded from the National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov/pub/TraceDB) in late September 2003. This untrimmed, raw dataset consisted of 880,404 fragments totaling 857 MB and was subsequently trimmed with LUCY (7). The trimming parameters used for these GSSs were Bracket [20 0.003], Window [10 0.01], and Error [0.005 0.002]. Approximately 240,000 bacterial artificial chromosome (BAC) end

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: BAC, bacterial artificial chromosome; GSS, genome survey sequence; HC, high C_0t; MAGI, maize assembled genomic island; MF, methylation filtration; NIP, nearly identical paralogs; RT, reverse transcriptase; TIGR, The Institute for Genomic Research.

[‡]Y.F. and S.J.E contributed equally to this work.

[‡‡]Present address: Department of Mathematics and Statistics, University of Guelph, ON, Canada N1G 2W1.

[¶¶]To whom correspondence should be addressed at: Roy J. Carver Co-Laboratory, Room 2035B, Iowa State University, Ames, IA 50011-3650. E-mail: schnable@iastate.edu.

reads were similarly downloaded from GenBank and processed to locate additional maize statistically defined repeats (5) that were used for repeat masking.

Validation of MAGI Assemblies by Using Sequenced Maize BACs. Sixteen entire maize B73 BAC sequences (GenBank accession nos. AC144717, AF448416, AF464738, AF466202, AF466203, AF466646, AF466931, AF546189, AY325816, AY371488, AY146791, AY180107, AY180106, AF271636, AY530952, and AY530951) downloaded from GenBank on July 24, 2004, were used as benchmarks to test the structures of MAGIs. These BACs were aligned with MAGIs by using BLASTN with the low complexity filter turned off. Only MAGIs that had BLAST alignments of ≥99% identity and alignment lengths of ≥400 bp were analyzed. The overlapping region between two BACs (accession nos. AY325816 and AF464738) resulted in five pairs of identical MAGI/BAC alignments. Only one member of each pair was analyzed.

RNA Isolation and Reverse Transcription. RNA samples of maize inbred line B73 were isolated from various treatments and/or tissues (see *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site). First-strand cDNA was synthesized with SuperScript II reverse transcriptase (RT) with Oligo-dT priming (Invitrogen). The resulting product was then treated with DNaseI (Invitrogen) and purified for PCR by following a previously described protocol that prevents genomic DNA contamination (8).

Touchdown PCR Amplification and Direct Sequencing of RT-PCR Products. Primers for genomic and RT-PCRs were designed with PRIMER3 (see *Supporting Materials and Methods* for details) (9). For cDNA and genomic DNA templates, PCRs were incubated for 2 min at 92°C, followed by 10 cycles of denaturation at 94°C for 30 s, annealing for 30 s, and elongation at 72°C for 1 min and another 24 cycles of 94°C for 30 s, 61°C for 30 s, and 72°C for 1 min and a final 10-min extension at 72°C. The annealing temperature was decreased by 0.8°C per cycle during the first 10 cycles from 69°C to 61°C to increase the specificity of the amplification. PCR_{96} cleanup plates (Millipore) were used to purify PCR products for single RT bands. QIAquick spin columns (Qiagen, Valencia, CA) were used to purify individual bands for double RT bands. Each purified sample was sequenced from both directions. The sequences of the RT-PCR products are available from the authors upon request.

Gene Content Analyses. The sequences of the assembled B73 3′ Iowa State University Maize ESTs build (10) were used to assess gene coverage by querying MAGIs with low (e^{-30}) and high (e^{-100}) stringency *E*-value criteria and with the low-complexity filter turned off. FGENESH (Softberry, Mount Kisco, NY) was used for *ab initio* gene prediction with monocot parameters and the GC option that uses all potential GC donor splice sites (10). Evidence for the transcription of predicted gene models was obtained by querying another larger build of assembled maize transcripts (see *Supporting Materials and Methods*) using BLASTN (*E*-value cutoff, e^{-10}).

Annotation of Maize Genes Without Evidence of Expression. The MAGI 3.1 assembly was initially screened against the Plant Genome Database maize tentative unique genes downloaded in September, 2003, using GENESEQER (11) as described in ref. 10. A sample of MAGIs that exhibited FGENESH predictions but that did not have GENESEQER EST alignments were subjected to RT-PCR. The predicted genes tested by RT-PCR were later compared with the above-mentioned assembly of all maize ESTs by using TBLASTN with a criterion of an *E* value of $\leq e^{-10}$. RT-PCR primer pairs designed from predicted genes that exhibited significant matches to maize transcripts were used as controls for RT efficiency within the mRNA sources used in this study. The remaining candidates were then run against TIGR plant Gene Indices (see *Supporting Materials and Methods* for details). Significant matches were determined by using BLASTN and TBLASTN with a criterion of an *E* value of $\leq e^{-10}$. The cDNA sequences of predicted genes without matches to the plant transcripts were also compared with the GenBank NR protein and nucleotide database (June 2005) with BLASTX and TBLASTX, respectively, and a very conservative *E*-value cutoff (e^{-4}). Predicted genes that exhibit matches only to maize and that did not align to transposons or annotated genes were deemed novel.

Display of MAGI Annotation. GBROWSE 1.61 was downloaded from the Generic Model Organism Database web site (12) and installed on an Apple Mac OS 10.3 system. The CAP3 assembly output files (13), GENESEQER alignments using Iowa State University B73 assembled 3′ EST data (10), FGENESH predictions, BLASTX hits (*E*-value cutoff, e^{-10}), and PRIMER3 results were parsed into GFF files by using PERL and AWK scripts. All GFF files were loaded into MYSQL database for GBROWSE display.

Results and Discussion

Assembly of MAGI Version 3.1. To assemble the maize gene space, it was necessary to develop a scalable solution that used mechanisms to minimize assembly artifacts caused by the presence of repetitive elements and that also accounted for the nonuniform sampling of the genome due to gene enrichment (5, 14). In our pipeline, sequences were cleaned, repeat-masked, and clustered by using PACE (15) based on defined overlap criteria. The sequences within clusters were then unmasked and assembled with CAP3 into one or more contigs. Relative to our prior maize genome assembly (5), the assembly presented here (MAGI 3.1) incorporates further improvements in the quality of the input sequences and the repeat masking process, and it uses clone pair information during clustering.

When assembling a genome sequenced with a shotgun cloning approach, sequence errors in the input data tend to "average out" if a sufficient degree of redundancy exists. As compared with the shotgun approach, nonuniform genome sampling approaches (e.g., MF and HC enrichment) could lead to higher rates of sequence errors within poorly sampled regions. Therefore, before the assembly of version 3.1, we conducted an analysis of a sample of publicly available MF and HC sequences to determine the sources and locations of sequencing errors relative to a benchmark set of 10 genes totaling ≈79 kb of highly finished sequence (16). This study demonstrated that the average rate of errors per base in a sample of unassembled MF and HC GSSs could be reduced 6-fold (to 3.6×10^{-4}) by applying more stringent trimming parameters with minimal loss of gene content. These parameters were applied to all input sequences used in assembling MAGI Version 3.1.

Another of the improvements of MAGI 3.1 versus MAGI 2.3 was the use of an updated version of our nonredundant repeat database for repeat masking. Because repeats are overrepresented in the genome, they should also be overrepresented within a random sample of genomic fragments. Available BAC end sequences are not a random sample of the maize genome but are substantially more representative than sequences obtained by gene enrichment (e.g., MF or HC selection). Consequently, we first masked an updated collection of BAC end sequences by using known repeats to enrich for lower-copy repetitive sequences. These masked data were then subjected to single-linkage clustering to generate statistically defined repeats. This analysis resulted in the recovery of additional repetitive sequences, which were incorporated into version 2.0 of the MAGI repeat database. A larger fraction of unfiltered shotgun and BAC end data are classified as repetitive by using these new statistically defined repeats (74% versus 57.6%) relative to the previously reported repeat database (version 1.0), a value that better correlates with the estimated frequency of repetitive sequences in the maize genome (17).

The third improvement of the MAGI 3.1 pipeline over that of MAGI 2.3 relates to the use of clone pair information. Sequencing both ends of a cloned fragment of DNA generates two sequences

Table 1. Comparisons between the latest MAGI 3.1 build and the previously reported 2.3 build (5)

	MAGI 2.3	MAGI 3.1
Starting data, no. of GSSs	730,974	879,523
Input masked, %	19.6	14.7
No. of contigs	91,690	114,173
No. of clustered clones	259,920	389,799
Average GSSs per contig, *n*	4.12	5.85
Average clones per contig, *n*	2.83	3.66
Contig % GC	44.5	45.6
Average contig length, bp	1,355	1,550
Maximum contig length, bp	8,489	12,498
No. of singletons	353,558	212,127

with known physical proximity features; this information is especially useful to help the assembler resolve highly similar repeats found in complex genomes. In our pipeline, paired sequences that contain at least 100 bases of nonrepetitive DNA are grouped together and provided to PACE as initial clusters, thereby preserving all relevant clone pair information. Although a large percentage of these PACE clusters yield single contigs, proximity constraints sometimes provide evidence that clusters should be split into two or more contigs during assembly. A comparison of this build to the previously reported MAGI 2.3 is presented in Table 1. Both builds are available from the authors upon request.

Quality Assessment of MAGIs. A combination of computational and wet-laboratory approaches (illustrated in Fig. 1) was developed to assess the quality of our current partial maize genome assembly. In the following sections, we demonstrate that the contigs in the latest MAGI assembly are of high quality.

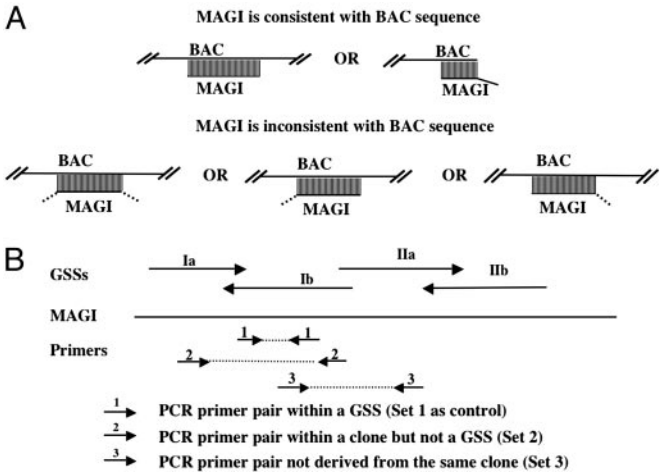

Fig. 1. Illustrations of computational and wet-laboratory strategies used for MAGI validation. (*A*) The consistency of MAGIs was assayed via alignment to maize B73 BACs. A set of potential MAGI/BAC alignments was identified by using BLAST (see *Materials and Methods*). The dashed lines mark portions of the MAGI that fail to match the BAC sequence. MAGIs were deemed to be inconsistent if they had a total overhang length (combined length of dashed lines) of >20 bp. The overhangs associated with four of the six consistent MAGI/BAC pairs that have sizes of between 6 and 20 bases can be recognized as incompletely trimmed vector sequences on a terminal GSS of a MAGI (Table 5). Four of the consistent MAGI/BAC pairs have overhangs of <6 bases, which may also be derived from incompletely trimmed vector. Terminal MAGI/BAC alignments of the type shown on the right do not provide evidence of inconsistency. Six such cases were identified. (*B*) Comparison of genomic PCR success rates: Within a MAGI, each primer pair annealed to the same GSS (set 1), two GSSs from the same clone (set 2), or two GSSs from different clones (set 3). Set 1 primer pairs served as a control to assess the success of primer design and PCR. Sets 2 and 3 primer pairs were used to validate the structure of MAGIs.

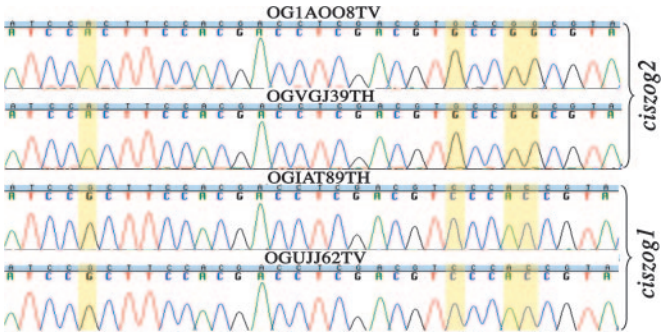

Fig. 2. Collapse of NIPs in MAGI_89783. Mismatches among GSSs are highlighted.

MAGI Validation: Comparisons to BAC Sequences. The sequences of 16 published maize B73 BACs were used as a benchmark for validating the structures of MAGIs. A BLAST search returned 173 nonredundant alignments between these 16 BACs and MAGIs. To determine whether these alignments verify the structure of a MAGI, we define the concept of consistency (Fig. 1*A*). Consistent MAGI/BAC pairs contain no more than 20 bases of a MAGI that do not align to the BAC (i.e., the sum of the two potential overhangs).

We excluded from subsequent analyses the eight inconsistent alignments that involved only a single GSS within a MAGI because these alignments do not test fragment assembly errors and are instead most likely due to misalignment of repetitive sequences. Indeed, all eight alignments of this type included repetitive sequences contained with the TIGR repeat database 4.0 (Table 5, which is published as supporting information on the PNAS web site).

Removing these eight inconsistent repetitive alignments left 165 MAGI/BAC alignments for validation (Table 5). Of these 165 MAGI/BAC alignments, 95.2% (157) are consistent. Because we observed evidence of the collapse of Near Identical Paralogs (NIPs) in the MAGI 2.3 build (5), we hypothesized that at least some of the eight inconsistent MAGIs detected in the current study could also have arisen via the collapse of NIPs into a single MAGI. Potential paramorphisms (polymorphisms between paralogs) in GSSs that comprise a MAGI have been reported previously (5, 16). Consequently, the trace files of the GSSs used to assemble each of the eight inconsistent MAGIs were examined manually. The GSSs associated with four inconsistent MAGIs are 100% identical and therefore exhibit no evidence of NIP collapse (Table 6, which is published as supporting information on the PNAS web site). The structures of three of these inconsistent MAGIs (nos. 41789, 84169, and 107229) were validated by genomic PCR (data not shown). Hence, the apparent inconsistencies associated with these three MAGIs appear to be a consequence of aligning MAGIs to highly similar but inappropriate BACs. The origin of the fourth inconsistent MAGI (no. 53496) is not known. In contrast, and consistent with the hypothesis that at least some of the inconsistent MAGIs arise because of NIP collapse, the GSSs used to assemble the remaining four inconsistent MAGIs (nos. 8097, 22812, 39419, and 89783) exhibited at least one putative paramorphism (Table 6). In the case of MAGI 89783, which encodes *cis*-zeatin *O*-glucosyltransferase, this hypothesis regarding the origin of inconsistent MAGIs is further supported by the presence in the maize inbred line B73 of two highly similar (98.3% nucleotide identity) *cis*-zeatin *O*-glucosyltransferase genes, *ciszog1* and *ciszog2* (accession nos. AF318075 and AY082660) (18). Significantly, the putative paramorphisms observed in the GSSs that comprise MAGI 89783 match those that distinguish *ciszog1* and *ciszog2* (Fig. 2). Further support for the hypothesis that at least some of the inconsistent MAGIs arise via NIP collapse is provided by the observation that

the rate of inconsistent MAGI/BAC alignments that contain putative paramorphisms (2.4%, 4 of 165) is similar to the observed rate of NIPs in the maize genome, i.e., ≈1% (S.J.E., T.-J.W., M. D. Yandeau-Nelson, Y.F., L. Li, L.G., H.-H. Chou, S.A., D.A.A., and P.S.S., unpublished data). These results suggest that the misassembly can be caused not only by highly homologous transposons but also by nearly identical nontransposon genes. The prevention of the misassembly in future assemblies of the maize genome will require access to very high-quality sequence data and the application of stringent assembly parameters.

MAGI Validation: Genomic PCR. The computational analyses described above suggest that, at minimum, ≈97% of MAGIs are correctly assembled. This observation is based on the hypothesis that if two independent assemblies (BAC and MAGI) agree, both are most likely correct. Note, however, that this is a conservative estimate; inconsistent MAGI/BAC alignments could also arise because of biological idiosyncrasies within the maize genome. To provide an estimate that incorporates such uncertainty, PCR amplification was used to independently estimate the proportion of MAGI assemblies that accurately reflect the structure of the maize genome (Fig. 1*B*). To first estimate the rate of false-negative PCR amplification, pairs of control primers that span predicted introns were designed that anneal to a single GSS (Fig. 1*B*, set 1; see also *Materials and Methods*). Each of these pairs of primers was used to conduct touchdown PCR on genomic DNA from the inbred line B73. As shown in Table 7, which is published as supporting information on the PNAS web site, ≈86% (1,165 of 1,358) of these control primers yielded a single PCR product of the size expected based on the positions at which the primers anneal to the GSS. Another 5% (68 of 1,358) of the control primers yielded a doublet PCR product, one of which was the expected size. Consistent with the structure of the maize genome (19), these doublets probably arise via the amplification of pairs of paralogous sequences. PCR failures [i.e., primer pairs that yielded either no band (6%) or multiple bands/smears (3%)] probably reflect problems in primer design, e.g., attempts to amplify multigene families.

Junction-testing primers were used to experimentally determine the quality of MAGIs. Pairs of junction-testing primers are those in which each member of a primer pair can anneal to either different GSSs within the same clone (Fig. 1*B*, set 2) or to different clones in a single MAGI (Fig. 1*B*, set 3). As such, these primer pairs can be used to test the assembly junctions of the GSSs that comprise a given MAGI. Approximately 90.9% (512 of 563; Fig. 1*B*, set 2) and 92.5% (99 of 107; Fig. 1*B*, set 3) of the junction-testing primer pairs yielded a single or doublet PCR product of the expected size (Table 7). Hence, the success rates of the junction-testing primers (90–92%) are similar to that of the control primers (≈91%). Based on a *Z* test for difference of two proportions, there is no statistical support for the hypothesis that the success rates of these classes of primers differ. We therefore conclude that a very high percentage of the GSS junctions reported in genic MAGIs are correct (i.e., they accurately reflect the structure of the maize genome).

Sequence Fidelity of MAGIs. By aligning GSSs to a benchmark set of 10 genes totaling ≈79 kb of highly finished sequence identified trimming parameters that reduced the rate of sequencing errors in a sample of GSSs from 2.3×10^{-3} to 3.6×10^{-4} (16). As mentioned above, these trimming parameters were used in the MAGI 3.1 build. We report here that the MAGIs corresponding to these 10 control genes have a sequencing error rate of 1×10^{-4}. The reduction in the rate of sequencing errors observed in MAGIs relative to GSSs is probably a consequence of the resampling of some base positions within MAGIs as compared with single-pass GSSs. About half (82 of 165) of the consistent MAGI/BAC alignments described above exhibit 100% identity, and only 213 bp of 274,689 bp (7.7×10^{-4}) within consistent alignments exhibit disagreements between the sequences of a MAGI and its respective BAC. The almost 8-fold difference between the estimated rates of sequencing errors in MAGIs obtained through alignments to BACs (7.7×10^{-4}) and alignments to the set of 10 control genes (i.e., 1×10^{-4}) may reflect higher sequencing errors in the BACs or the inappropriate alignment of a MAGI to a BAC that contains a NIP of a gene present in that MAGI.

Genic Content of MAGIs. Determining how successfully the MF and HC filtration strategies have sampled the gene space of the maize genome is complicated by the fact that a complete inventory of maize genes is not available. Even so, several computational experiments suggest that the MF and HC GSSs have captured a large fraction of the maize gene space. For example, these GSSs have been shown to tag all members of small collections of known maize genes (14, 16). In addition, ≈11% of the contigs in an assembly consisting of approximately one-fifth of the GSSs used in the MAGI 3.1 assembly exhibit similarity (BLAT settings: 95% identity and ≥20% of contig length) to the TIGR Plant Gene Index (3). Furthermore, ≈560,000 MF GSSs exhibit similarity to ≈65% of the nonrepeat, nonhypothetical maize

Table 2. FGENESH-derived gene prediction in all 114,173 MAGIs

Type of predictions	No. of predictions		
	Total (%)*	With transcription evidence in maize† (%)‡	Containing repeats§ (%)‡
Complete gene models			
With intron	13,597 (29.1)	9,096 (66.9)	1,423 (10.5)
Without intron	6,638 (14.2)	3,918 (59.0)	770 (11.6)
Subtotal	20,235 (43.3)	13,014 (64.3)	2,193 (10.8)
Incomplete gene models			
Lacking first exon	10,937 (23.4)	8,085 (73.4)	1,477 (13.5)
Lacking last exon	10,861 (23.3)	6,268 (57.7)	1,491 (13.7)
Lacking first and last exon	4,655 (10.0)	3,228 (69.3)	715 (15.4)
Subtotal	26,453 (56.7)	17,581 (66.5)	3,683 (13.9)
Total no. of predictions	46,688 (100)	30,595 (65.5)	5,876 (12.6)

*The percentage of indicated types of predicted gene models/total number of gene predictions.
†Predicted transcript matches either a maize expressed gene or maize cDNA sequence (BLASTN; *E*-value cutoff, e^{-10}).
‡The percentage of predictions that contain the indicated type of database match/number of the indicated type of gene model predictions.
§Each predicted coding sequence was screened against the nucleotide MAGI repeat database using BLASTN (*E*-value cutoff, e^{-10}). Predictions with at least one database match were deemed to be repetitive.

Table 3. RT-PCR results for all primer pairs that yielded a single genomic PCR band

RT-PCRs	Band pattern	BLAST results, *n* (%) +	BLAST results, *n* (%) −	Total, *n* (%)
RT-PCR-positive				
1 band	<	125 (34.0)	370 (27.0)	495 (28.5)
	=	32 (8.7)	165 (12.1)	197 (11.3)
2 bands	≪	14 (3.8)	35 (2.5)	49 (2.8)
	≤	18 (4.9)	58 (4.2)	76 (4.4)
Subtotal		189 (51.4)	628 (45.9)	817 (47.0)
RT-PCR-negative				
No band		134 (36.4)	582 (42.5)	716 (41.2)
1 band	>	0	10	10
2 bands	≥	5	24	29
Others*		40 (10.9)	125 (9.1)	165 (9.5)
Subtotal		179 (48.6)	741 (54.1)	920 (53.0)
Total		368 (100)	1,369 (100)	1,737 (100)

BLAST results indicate primer pairs derived from predicted genes that do (+) or do not (−) have significant BLASTN (*E*-value cutoff, e^{-10}) hits against all maize transcripts. <, The RT PCR band is smaller than genomic PCR band; =, the RT-PCR band is the same size as the genomic PCR band; ≪, both RT-PCR bands are smaller than the genomic PCR band; ≤, one RT-PCR band is smaller than genomic PCR band and the other one is the same size as the genomic PCR band; >, the RT-PCR band is larger than the genomic PCR band. Sequence analyses established that five of five RT-PCR products of this type do not exhibit similarity to the predicted genes from which the PCR primers were designed; ≥, at least one of the two RT PCR bands is larger than the genomic PCR band.

*The gel analyses of RT products yielded more than two visible bands or a smear.

Table 4. Evidence of transcription of FGENESH predicted genes

	Maize transcripts	Plant transcripts	NR databases	Maize matches only: Transposons	Maize matches only: Novel	Total
RT+	256	300	37	3	32	628
RT−	296	236	66	11	132	741
Total	552	536	103	14	164	1,369

Maize transcript values show TBLASTN hits against maize transcripts (see *Supporting Materials and Methods*) with an *E*-value cutoff of e^{-10}. Values for the plant transcripts show BLASTN and TBLASTN hits against TIGR plant gene indices (see *Supporting Materials and Methods*), with an *E*-value cutoff of e^{-10}. Values for the NR databases show BLASTX and TBLASTX hits against the GenBank NR nucleotide database and protein database, respectively, with e^{-4} as the *E*-value cutoff. The NR databases column does not include predicted genes that match only maize sequences. Most of the 103 predicted genes in this column match cereal retroelements. Entries in the maize and plant transcripts and NR databases columns did not exhibit matches to the databases shown to the left. For example, the 536 sequences with BLAST hits to plant transcripts did not exhibit matches to maize transcripts. Predicted genes that exhibited matches only to maize entries in the GenBank NR database and that did not align with transposons were deemed novel. RT+, RT-positive; RT−, RT-negative.

genes detected on published BACs (BLAT settings: 98% identity and ≥90% of read length) (3).

To estimate gene coverage within our MAGI 3.1 assembly, we used a set of assembled 3′ reads of maize ESTs from the inbred B73 that presumably corresponds to unique genes (10). Of the 19,454 unigenes in this set, 14,606 (76%) match at least one MAGI using BLAST with a stringent *E*-value cutoff of e^{-100}. Although it is not possible to directly compare these results to the previously reported estimates because of differences in algorithms and significance criteria, it is clear that the MAGIs contain a high percentage of known maize genes.

Genes can be detected not only by means of alignments to the sequences of known genes as was done above but also by *ab initio* gene prediction software. We previously used a set of >1,300 maize gene sequences to compare the performance of three *ab initio* gene prediction programs (FGENESH, GENEMARK.HMM, and GENSCAN), each of which had been trained on maize. In this analysis, FGENESH performed the best, although GENEMARK.HMM also performed well (10). These results are consistent with the observation that FGENESH was the most successful program for gene prediction in rice (20). With the 114,173 MAGIs as input, FGENESH returned 46,688 gene predictions, of which only ≈13% contained repetitive sequences (Table 2). Approximately 34% (16,093) of the predicted cDNAs had no hits against assembled maize ESTs or maize cDNAs (see *Materials and Methods*). As an additional measure of gene content, another 9,323 MAGIs did not contain a prediction but did exhibit similarity to known ESTs and/or proteins. Hence, >47% of all MAGIs in build 3.1 contain a gene or predicted gene.

Display of Annotated MAGIs. Annotated MAGIs can be viewed at the Iowa State University MAGI web site. An example is shown in Fig. 3, which is published as supporting information on the PNAS web site. Layouts of individual GSSs from parsed CAP3 output are color-coded for convenience. Sequence-based annotations against protein databases were performed with BLASTX against the Protein Information Resource International Protein Database (version 79.00). Gene structures predicted by FGENESH and GENESEQER are also displayed (*Materials and Methods*). Primers used in this study were also entered into the MAGI 3.1 GBROWSE database. The entire membership of this assembly can be downloaded along with the contigs *per se*.

RT-PCR Validation of Predicted Transcripts. As discussed above, FGENESH analysis of MAGIs resulted in the prediction of ≈16,100 genes that do not match known maize transcripts. We designed pairs of intron-spanning primers to test whether 1,590 of these *ab initio* predicted novel genes are transcribed. Another batch of 438 pairs of primers from *ab initio* predictions that do have significant BLAST hits to maize transcripts were also designed as a control. Because paralogs and nonspecific amplification can complicate the verification of putative genes by RT-PCR, we tested each pair of primers by conducting PCR on B73 genomic DNA. Approximately 86% (1,737 of 2,028) of these reactions yielded single genomic PCR bands of the size expected based on the positions at which the primers anneal to the corresponding MAGI (Table 7). The rates at which primers designed to amplify predicted genes with and without transcription evidence amplified single PCR products were similar: 84% (368 of 438) and 85% (1,369 of 1,590), respectively. The 1,737 primer pairs were also subjected to RT-PCR using a diverse cDNA pool as template (*Materials and Methods*). Reactions that yielded single bands that were smaller than or equal in size to the PCR product from genomic DNA template or that yielded double bands, one of which was smaller than or equal in size to the PCR product from genomic DNA template, were considered RT-positive. Reactions that yielded any other outcomes were deemed RT-negative. Approximately 51% (189 of 368) of the BLASTN-positive set and 46% (628 of 1,369) of the BLASTN-negative set of Table 3 were RT-positive (Table 3).

To determine the specificity of these RT reactions, we sequenced >160 PCR products from RT-positive reactions (Table 8, which is published as supporting information on the PNAS web site). These analyses demonstrated that ≈94% of these RT products were derived from the predicted genes from which the primers had been designed (data not shown). Thus, it was possible to verify the expression of 43% [94% × (628/1,369)] of predicted genes that lack evidence of transcription in maize (i.e., the BLASTN-negative set in Table 3). Consequently, the MAGIs have probably "tagged" >6,900 [43% × (46,688 − 30,595)] (Table 2) expressed genes that previously lacked evidence of transcription. Because only half of the control genes for which evidence of transcription already exists in

maize (the BLASTN-positive set in Table 3) were RT-positive in this experiment, we conclude that our RT-PCRs did not sample the entire maize transcriptome. Hence, our estimate of the number of predicted genes that are expressed is highly conservative.

Annotation of RT-Validated Genes. Of the 628 RT-positive predicted genes that previously lacked evidence for expression in the maize transcriptome, 256 (41%) exhibit significant TBLASTN hits to maize transcripts (Table 4) and are therefore probably paralogs of maize genes for which evidence of transcription exists. Another 337 (300 + 37; 54%) of the remaining genes exhibit significant similarity to plant transcripts and nonmaize genes or proteins in the GenBank NR DNA/protein databases (*Materials and Methods*). Significantly, after carefully removing sequences that exhibit similarity to transposons that are often responsible for overestimations of gene numbers in complex plant genomes (21), >5% (32 of 628) of the RT-positive predicted genes are novel based on very conservative criteria (Table 4; see also Table 9, which is published as supporting information on the PNAS web site). In all, ≈12% (164 of 1,369) of the predicted genes are novel and the expression of 20% (32 of 164) of the novel genes could be verified by RT-PCR experiments. Hence, the MAGIs are conservatively expected to contain ≈350 expressed novel genes or orphans [94% × (32/1,369) × (46,688 − 30,595)].

Estimation of the Number of Maize Genes. Based on available EST data ≈30,600 of the ≈46,700 predicted gene models in our assembly are expressed; moreover, we have shown that RT-PCR can conservatively validate the expression of ≈40% (94% × 46%) of the remaining ≈16,100 gene models (Tables 2 and 3). Taken together, these results imply that our partial maize genome assembly contains at least 37,100 genes [30,600 + (40% × 16,100)]. It is, however, possible that some of the 26,453 incomplete gene models in Table 2 do not represent unique genes. A more conservative estimate of the number of maize genes is therefore provided by considering only gene models that contain a last exon (and which could therefore be detected in our set of 3′ EST unigenes) and for which there is evidence of expression (21,099 = 13,014 + 8,085) plus the at least 40% of gene models that lack expression evidence but would be confirmed via RT-PCR experiments based on our experience (4,071 = 40% × 10,073). Dividing this sum (25,170 = 21,099 + 4,071) by the 76% of 3′ unigenes that can be identified among the MAGIs ($E = e^{-100}$) yields a lower bound of ≈33,000 genes. If we assume each nonrepetitive gene model from Table 2 is unique and expressed (40,812 = 46,688–5,876) and divide by 76%, the upper bound for the number of nonrepetitive genes in the maize genome is ≈54,000.

Conclusions

The gene enrichment strategies that have been validated by using the maize genome are likely to be applied to the genomes of other large-genome plants. Indeed, preliminary enrichment projects have already been reported for the wheat (22) and sorghum (23) genomes, and a gene enrichment project has been funded for pine. The assembly of the nonuniform genomic fragments that are generated by gene enrichment strategies poses special challenges, which we have addressed previously (5).

The current study provides two metrics (one strictly computational and the other based on large-scale PCR experiments) by which the quality of genome assemblies can be evaluated. Applying these metrics to our partial maize genome assembly demonstrates that gene-enriched sequences can be assembled into high quality contigs that facilitate biological discovery. For example, the application of large-scale RT-PCR using primers designed based on MAGIs made it possible to obtain expression data for hundreds of predicted genes.

Interestingly, these experiments also uncovered evidence for the existence of ≈350 expressed maize genes that do not have homologs in other species. We hypothesize that these orphans are quickly evolving genes that played important roles during maize evolution and/or domestication. As such, these orphans present attractive targets for reverse genetics experiments.

We thank Jia Yi for help with RNA isolation and pilot experiments and two anonymous reviewers for useful suggestions. This research was funded in part by National Science Foundation Plant Genome Program Competitive Grants DBI-9975868 and DBI-0321711 and by the Hatch Act and funds from the state of Iowa. S.J.E. was supported in part by National Science Foundation Integrative Graduate Education and Research Traineeship Fellowship DGE-9972653.

1. Bennetzen, J. L., Chandler, V. L. & Schnable, P. (2001) *Plant Physiol.* **127,** 1572–1578.
2. Martienssen, R. A., Rabinowicz, P. D., O'Shaughnessy, A. & McCombie, W. R. (2004) *Curr. Opin. Plant Biol.* **7,** 102–107.
3. Whitelaw, C. A., Barbazuk, W. B., Pertea, G., Chan, A. P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J. L., *et al.* (2003) *Science* **302,** 2118–2120.
4. Palmer, L. E., Rabinowicz, P. D., O'Shaughnessy, A. L., Balija, V. S., Nascimento, L. U., Dike, S., de la Bastide, M., Martienssen, R. A. & McCombie, W. R. (2003) *Science* **302,** 2115–2117.
5. Emrich, S. J., Aluru, S., Fu, Y., Wen, T. J., Narayanan, M., Guo, L., Ashlock, D. A. & Schnable, P. S. (2004) *Bioinformatics* **20,** 140–147.
6. National Plant Genome Initiative (2005) *Maize Genome Sequencing Project: An NSF/DOE/USDA Joint Program* (Natl. Sci. Found., Arlington, VA), available at www.nsf.gov/pubsys/ods/getpub.cfm?nsf04614.
7. Chou, H. H. & Holmes, M. H. (2001) *Bioinformatics* **17,** 1093–1104.
8. Flohr, A. M., Hackenbeck, T., Schlueter, C., Rogalla, P. & Bullerdiek, J. (2003) *Biotechniques* **35,** 920–922, 924, 926.
9. Rozen, S. & Skaletsky, H. J. (2000) in *Methods in Molecular Biology*, eds. Krawetz, S. & Misener, S. (Humana, Totowa, NJ), pp. 365–386.
10. Yao, H., Guo, L., Fu, Y., Borsuk, L. A., Wen, T. J., Skibbe, D. S., Cui, X., Scheffler, B. E., Cao, J., Emrich, S. J., *et al.* (2005) *Plant Mol. Biol.* **57,** 445–460.
11. Usuka, J., Zhu, W. & Brendel, V. (2000) *Bioinformatics* **16,** 203–211.
12. Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., *et al.* (2002) *Genome Res.* **12,** 1599–1610.
13. Huang, X. & Madan, A. (1999) *Genome Res.* **9,** 868–877.
14. Springer, N. M., Xu, X. & Barbazuk, W. B. (2004) *Plant Physiol.* **136,** 3023–3033.
15. Kalyanaraman, A., Aluru, S., Kothari, S. & Brendel, V. (2003) *Nucleic Acids Res.* **31,** 2963–2974.
16. Fu, Y., Hsia, A. P., Guo, L. & Schnable, P. S. (2004) *Plant Physiol.* **135,** 2040–2045.
17. Meyers, B. C., Tingey, S. V. & Morgante, M. (2001) *Genome Res.* **11,** 1660–1676.
18. Veach, Y. K., Martin, R. C., Mok, D. W., Malbeck, J., Vankova, R. & Mok, M. C. (2003) *Plant Physiol.* **131,** 1374–1380.
19. Blanc, G. & Wolfe, K. H. (2004) *Plant Cell* **16,** 1667–1678.
20. Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al.* (2002) *Science* **296,** 79–92.
21. Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. & Ramakrishna, W. (2004) *Curr. Opin. Plant Biol.* **7,** 732–736.
22. Li, W., Zhang, P., Fellers, J. P., Friebe, B. & Gill, B. S. (2004) *Plant J.* **40,** 500–511.
23. Bedell, J. A., Budiman, M. A., Nunberg, A., Citek, R. W., Robbins, D., Jones, J., Flick, E., Rholfing, T., Fries, J., Bradford, K., *et al.* (2005) *PLoS Biol.* **3,** e13.