Characterization of the Maize Endosperm Transcriptome and Its Comparison to the Rice Genome

Jinsheng Lai,[1] Nrisingha Dey,[2] Cheol-Soo Kim,[3,5] Arvind K. Bharti,[1] Stephen Rudd,[4,6] Klaus F.X. Mayer,[4] Brian A. Larkins,[3] Philip Becraft,[2] and Joachim Messing[1,7]

[1]Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA; [2]Department of Genetics, Development & Cell Biology, Iowa State University, Ames, Iowa 50011, USA; [3]Department of Plant Science, University of Arizona, Tucson Arizona 85721, USA; [4]Munich Information Center for Protein Sequences, Institute for Bioinformatics, GSF Research Center for Environment and Health, Neuherberg, Germany

The cereal endosperm is a major organ of the seed and an important component of the world's food supply. To understand the development and physiology of the endosperm of cereal seeds, we focused on the identification of genes expressed at various times during maize endosperm development. We constructed several cDNA libraries to identify full-length clones and subjected them to a twofold enrichment. A total of 23,348 high-quality sequence-reads from 5′- and 3′-ends of cDNAs were generated and assembled into a unigene set representing 5326 genes with paired sequence-reads. Additional sequencing yielded a total of 3160 (59%) completely sequenced, full-length cDNAs. From 5326 unigenes, 4139 (78%) can be aligned with 5367 predicted rice genes and by taking only the "best hit" be mapped to 3108 positions on the rice genome. The 22% unigenes not present in rice indicate a rapid change of gene content between rice and maize in only 50 million years. Differences in rice and maize gene numbers also suggest that maize has lost a large number of duplicated genes following tetraploidization. The larger number of gene copies in rice suggests that as many as 30% of its genes arose from gene amplification, which would extrapolate to a significant proportion of the estimated 44,027 candidate genes of its entire genome. Functional classification of the maize endosperm unigene set indicated that more than a fourth of the novel functionally assignable genes found in this study are involved in carbohydrate metabolism, consistent with its role as a storage organ.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. CA398264–CA405362 and CD43287–CD44042.]

Comparative genetic mapping has shown that the chromosomes of many grass species exhibit extensive synteny (Helentjaris et al. 1988; Ahn and Tanksley 1993; Gale and Devos 1998). Although at the DNA sequence level, collinearity is interrupted and paralogous sequences are found in other genomic locations, the percentage of orthologous sequences appears to be significant, so that one could expect to map genes across closely related species (Song et al. 2002; Lai et al. 2004; Swigoňová et al. 2004). Because of synteny and its relative small size compared with other cereal genomes, rice (*Oryza sativa*) was selected as the first monocotyledonous genome to be sequenced. Although the previously published draft sequences (Goff et al. 2002; Yu et al. 2002) are useful as surveys of rice chromosomes, they are not suitable for comparative genomics. However, in the meantime, the International Rice Genome Sequencing Project, IRGSP (http://rgp.dna.affrc.go.jp), has produced a map-based sequence that has been deposited in GenBank, and pseudomolecules for all chromosomes are available to the scientific community (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml). In addition, the sequences and the analysis of three of the 12 chromosomes have been published, Chromosome 1 (Sasaki et al. 2002), Chromosome 4 (Feng et al. 2002), and Chromosome 10 (Rice Chromosome 10 Sequencing Consortium 2003).

Although the sequences of rice chromosomes permit us to use computational prediction programs to locate all the genes, predicted genes need to be verified by the identification and characterization of expressed sequence tags (ESTs). Furthermore, the tissue-specificity of predicted genes needs to be investigated based on where genes are expressed and by their allelic variants. For instance, one of the important organs of cereal grain is the endosperm. Endosperm is a nutritive tissue that is used by the germinating embryo as an energy source. Because of its high nutritional content, it is also a major food source for humans and animals. Among cereals, the endosperm of maize has been extensively studied because of its large size, and there are many mutations affecting its development and its effect on kernel appearance. Studies based on analysis of EMS (ethyl methane sulphonate) mutagenesis suggested there are at least 300 genes in maize that can cause a visible endosperm phenotype (Neuffer and Sheridan 1980). A similar estimate of seed phenotypes was made based on mutagenesis using the Mutator transposon (Scanlon et al. 1994). Still, only a small fraction of these endosperm mutants have been molecularly characterized (Scanlon and Myers 1998). Moreover, we expect that the number of genes expressed during endosperm development to be significantly

Present addresses: [5]Agricultural Plant Stress Research Center, Chonnam National University, Kwangju 500-757, Korea; [6]Turku Center for Biotechnology, Tykistoekatu 6, Turku, Finland.
[7]Corresponding author.
E-MAIL messing@waksman.rutgers.edu; FAX (732) 445-0072.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2780504.

14:1932–1937 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

higher than 300 because many of them do not yield a visible phenotype or the genes are required early in development before seeds are set.

In an effort to identify and relate the maize and rice endosperm transcriptomes, we constructed several full-length cDNA libraries using maize endosperm from early-to-middle mature growth stages (4–6 d after pollination [DAP] and 7–23 DAP). The libraries were subjected to both normalization and subtraction. Out of 5504 unique ESTs, 3160 (59%) represented completely sequenced cDNAs. All sequences have been placed onto rice chromosome sequences to provide positional information and to assess gene amplification in rice. Furthermore, functional assignments of total maize versus maize endosperm-specific cDNAs have shown that novel cDNAs contain motifs typical for carbohydrate metabolism.

RESULTS

Library Construction, Normalization, and Subtraction

A total of six libraries were constructed, four from mRNA of the 4–6-DAP tissue and two from the 7–23-DAP tissue (Table 1). To evaluate the quality (QC) of these libraries, initially a few plates (96-well) from each of the six libraries were sequenced. Based on insert size, only three out of the six libraries (endosperm_3, 4, and 5) were found to meet the required standards, and were, therefore, further analyzed (Supplemental Fig. A). To overcome representation of abundant cDNAs, 48 plates of the QC-passed libraries were arrayed on high-density filters and probed with labeled cDNAs made from the corresponding mRNA preparation. Autoradiographs were read to select against strong hybridizing clones, and the low-abundance clones were rearrayed in fresh 96-well plates for sequencing. This normalization step resulted in nearly 40% enrichment for two of the three libraries, endosperm_3 (31 plates) and endosperm_5 (32 plates). There was a less striking effect on the third (endosperm_4) library, which was directly processed for sequencing without rearraying. All clones were sequenced from both ends. Because the sequence analysis indicated that even after normalization, redundant cDNAs represented a substantial portion of the EST collection (see below), we introduced an additional enrichment step. Pools of sequenced cDNAs were used as probes to hybridize against new filters of arrayed clones from two libraries that represented the early and late immature endosperm RNA preparations. Therefore, 96 plates of endosperm_3 and 48 plates of endosperm_5 were hybridized with both cDNA probes and pooled plasmid probes; the low-abundance clones were rearrayed in 23 plates for endosperm_3 and 13 plates for endosperm_5. Clones from these plates were again sequenced from both ends.

Full-Length cDNAs

Out of a total of 35,520 attempts, 23,348 reads (66% success) were of high quality (>200 bp, Q20). Earlier samples were run on ABI 3700 DNA sequencers, later ones on an ABI 3730xl; as a consequence, the average read length improved from 576 bp to 830 bp. Of the total, 12,659 (54%) were derived from the 5′-end, whereas 10,689 (46%) were from the 3′-end of cDNAs. The number of 3′-end sequences was somewhat lower, because the poly(A) tract caused a slightly higher failure rate. In total, 8939 clones had sequences both from the 5′- and the 3′-ends, of which 5455 clones (61%) with overlapping 5′- and 3′-ends represented complete cDNAs, whereas 3484 (39%) had bigger insert sizes that are not completely covered yet. Clustering the 5455 full-length cDNAs resulted in 2198 (40%) unique sequences, thereby proving the robustness of the above-mentioned enrichment procedures. Assembling all 5′-end sequences resulted in 5504 nonredundant cDNAs.

To investigate whether these cDNAs encode full-length ESTs, we have compared our EST sequences to a database containing protein information as described under Methods. According to the results from the comparisons with our initial 1334 sequences, ~98% contained the ATG start codon in the appropriate place, with >50 nt before the ATG start codon (data not shown). Given that such a significant portion of our cDNA collection contained the ATG start codon, it is reasonable to assume that the isolation of mRNA and the construction protocol of the cDNA libraries were of sufficient quality to yield full-length cDNAs and maintain them as full-length clones in *Escherichia coli*. Alignment of protein sequences to cDNAs also proved useful in determining their predicted lengths, which enabled us to estimate gap sizes between 5′- and 3′-ends of the longer cDNA clones (data not shown). This analysis suggested that a large number of cDNAs could be completely sequenced with just one round of primer walking. Therefore, a set of 3357 primers was designed to carry out an additional cycle of sequencing for clones that had gaps in the center of their sequence, thereby yielding another 992 full-length cDNAs, bringing the total to 3160 completely sequenced clones. The size distribution of the maize endosperm-expressed cDNAs was compared with that of the set of full-length cDNAs from rice (The Rice Full-Length cDNA Consortium 2003). Although there is a high percentage of smaller cDNA length of maize endosperm cDNAs (0.8–0.9 kb) than that of the rice collection, there appears to be a higher percentage of maize endosperm cDNAs with a medium length of 1.1 to 1.2 kb (Supplemental Fig. B).

Functional Classification of the Unigene Set

From 5504 unigenes, 2911 (53%) could be assigned a putative function (Methods), which includes unclassified proteins and those whose classification is not yet clear-cut. Excluding these unclassified categories (Table 2), the majority of cDNAs fall into the class of metabolism (16%), followed by the class of cellular organization (14.1%) and then transcription (9.1%). Cellular organization also includes proteins that are localized to the organelles. This classification is consistent with the many functions of maize endosperm, where compartmentalization occurs. However, this number was not biased because of the abundance of storage proteins. Because of the normalization and subtraction of the cDNA libraries, storage protein sequences constitute only ~5% within this functional class, demonstrating that the normalization and subtraction must have worked well. Like many cross sections of cellular functions, cellular communication/signal transduction also ranked very high

Table 1. Maize Endosperm cDNA Libraries

Lib names	Rutgers ID	Iowa/Arizona ID	Tissue types	Forward seq no.	Reverse seq no.	Total seq no.
Endosperm_1	E049127	L02	7–23 DAP	337	445	782
Endosperm_2	EL01N02	L02	4–6 DAP	344	97	441
Endosperm_3	EL01N03	L04	7–23 DAP	3522	3051	6573
Endosperm_4	EL01N04	L06	4–6 DAP	4378	3912	8290
Endosperm_5	EL01N05	L08	4–6 DAP	3576	2768	6344
Endosperm_6	EK07D23	L07	4–6 DAP	502	416	918
				12,659	10,689	23,348

Table 2. Distribution of the Functional Classes of Genes in Maize Whole Genome as Well as in Endosperm-Specific and Tomato Genome-Wide ESTs

No.	Category	Whole maize genome ESTs[a]		Total maize endosperm ESTs[b]		Novel maize endosperm ESTs[c]		Whole tomato genome ESTs[d]
	Number of unigenes represented	15,373	30.8%	2911	52.9%	198	29.2%	9599
	Number of unigenes with no match	34,618	69.2%	2593	47.1%	479	70.8%	17,675
	Total unigenes	49,991	100.0%	5504	100.0%	667	100.0%	27,274
1	Metabolism	3173	16.8%	671	16.0%	102	27.3%	20.9%
2	Energy	1071	5.7%	295	7.1%	15	4.0%	4.0%
3	Cell growth, cell division, DNA synthesis	753	4.0%	199	4.8%	9	2.4%	2.8%
4	Transcription	1772	9.4%	381	9.1%	39	10.4%	11.3%
5	Protein synthesis	772	4.1%	286	6.8%	15	4.0%	4.1%
6	Protein destination	1268	6.7%	319	7.6%	26	7.0%	8.5%
7	Transport facilitation	1035	5.5%	185	4.4%	22	5.9%	5.9%
8	Cellular transport and transport mechanisms	851	4.5%	217	5.2%	14	3.7%	4.0%
9	Cellular biogenesis	1064	5.6%	193	4.6%	10	2.7%	2.7%
10	Cellular communication, signal transduction	2126	11.2%	359	8.6%	60	16.0%	13.6%
11	Cell rescue, defense, cell death, aging	1826	9.6%	334	8.0%	34	9.1%	8.0%
12	Ionic homeostasis	23	0.1%	2	0.0%	0	0.0%	0.0%
13	Cellular organization	2262	12.0%	588	14.1%	0	0.0%	13.7%
14	Motility	18	0.1%	2	0.0%	0	0.0%	0.0%
15	Tissue specificity	0	0.0%	0	0.0%	0	0.0%	0.0%
16	Development	639	3.4%	126	3.0%	4	1.1%	0.0%
17	Transposable elements, viral and plasmid	257	1.4%	25	0.6%	24	6.4%	0.0%
14	Organism-specific proteins	14	0.1%	0	0.0%	0	0.0%	0.0%
	Number of matches[e]	18,924	100.0%	4182	100.0%	374	100.0%	99.5%
15	Classification not yet clear-cut	3603		666		54		
16	Unclassified proteins	8112		1344		123		
	Total number of matches	30,639		6192		551		

[a]The maize EST whole genome data (397,000 ESTs). Out of 49,991 unigenes (singletons and clusters), 15,373 (31%) could be assigned a putative function at 1e−10, which includes unclassified proteins and those whose classification is not yet clear-cut.
[b]Out of 5504 unigenes, 2911 (53%) could be assigned a putative function, which includes unclassified proteins and those whose classification is not yet clear-cut.
[c]Out of 5504 unigenes, 677 (12.3%) were not present in the maize EST collection of GenBank, of which only 198 had a match to M*At* DB.
[d]Out of 27,274 unigenes, 9599 (35%) could be assigned a role category (Van der Hoeven et al. 2002).
[e]Number of matches exceeds the number of unigenes represented because the same unigene may have more than one match from different role categories.

(8.6%). Energy and protein destination were other large classes (7.1% and 7.6%, respectively), presumably because of starch synthesis and other energy storage functions. This was also clear from the metabolism class, where carbohydrate metabolism was the largest subgroup. A category that is relatively high at the tissue-specific level is cell rescue, defense, cell death, and aging (8%). Considering that endosperm is a terminally differentiated tissue, one would expect related genes to be expressed presumably late in endosperm development. The group of unclassified or sequences with no clear classification was rather large (nearly one-third, not listed in Table 2), indicating the importance of functional genomics in the identification of protein function. A more complete classification including all subcategories is provided as online information (Supplemental Table A).

Comparison of Maize Endosperm Unigenes With ESTs in GenBank/MaizeGDB

To investigate how many of the 5504 endosperm-specific unigenes were novel transcripts, they were searched against all the current maize ESTs available in GenBank/MaizeGDB. Out of 397,000 maize ESTs, a total of 49,991 unigenes were assembled as described under Methods (Table 2). Comparison of the two unigene sets provided a total of 677 (12.3%) novel cDNAs (BLASTN E-value 1e−20). Because the current ESTs are derived from endosperm tissues, these 677 novel ESTs most likely represent endosperm-specific or at least endosperm-preferred expressed unigenes. Nevertheless, it is interesting to note that more than a fourth of the novel functionally assignable ESTs belong to the category of "Metabolism," consistent with the tissue-specificity of the endosperm. A list of all 677 cDNAs is provided as online information in Supplemental Table B.

Recently a genome-wide analysis of a very large number of ESTs was done in tomato (Van der Hoeven et al. 2002). In their study, out of 27,274 unigenes, 9599 (35%) of the total could be assigned a role category (not including unclassified proteins nor those whose classification is not yet clear-cut). To have a side-by-side comparison, we recalculated the functional classification of the maize endosperm-specific unigene set, after eliminating the unclassified groups. The functional categories are comparable between tomato (Van der Hoeven et al. 2002) and the maize genome-wide sequences (Table 2).

Organization of Rice Genes Homologous to Maize Endosperm mRNAs

The 12 rice chromosomes sequenced by the IRGSP (http://rgp.dna.affrc.go.jp), available as pseudomolecules (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml), were used as reference to study the organization of endosperm-expressed genes. A total of 54,397 putative genes could be predicted for the rice genome from this data set (Table 3) using the FGENESH program with the default setting for monocotyledonous genes (http://www.softberry.com). A screen of all the predicted genes with a repeat database (TBLASTN *E*-value <e−50) excluded 10,370 genes that encode transposon-related proteins (both DNA

Table 3. Localization of Maize Endosperm cDNA on Rice Chromosomes

Chromosomes	Physical length	No. of predicted genes	No. of genes excluded repeats	No. of maize genes mapped	No. of orthologous rice genes	Genes/Mb	Percentage (%)
Ch01	42.96	6690	5719	572	446	10.38	7.80
Ch02	35.84	5437	4628	561	425	11.86	9.18
Ch03	36.90	5737	4928	725	522	14.15	10.59
Ch04	34.76	5355	4125	335	258	7.42	6.25
Ch05	29.26	4550	3572	358	264	9.02	7.39
Ch06	31.30	4805	3876	332	245	7.83	6.32
Ch07	29.44	4574	3640	320	251	8.53	6.90
Ch08	28.46	4277	3324	284	214	7.52	6.44
Ch09	22.76	3419	2695	208	164	7.21	6.09
Ch10	22.42	3477	2754	173	132	5.89	4.79
Ch11	27.62	4187	3310	181	129	4.67	3.90
Ch12	11.95	1889	1456	79	58	4.85	3.98
Total	353.67	54,397	44,027	4128	3108	8.79	7.06

transposons and retrotransposons), leaving a total number of 44,027 predicted genes in rice. Applying the same screening conditions to the 5504 maize unigenes led to the exclusion of 178 putative repeat-like sequences. When the remaining 5326 maize endosperm unigenes were compared with the predicted rice transcriptome, 4139 genes (78% of the total) were found to have at least one rice homolog (BLASTN *E*-value <e − 10). A list of all the maize cDNAs not present in rice is provided as online information in Supplemental Table C. This high conservation between maize and rice ESTs is consistent with the close evolutionary distance between them (50 million years ago, Mya), and, at the same time, with the evidence of rapid divergence within the grass family (Song et al. 2002). Furthermore, because the homoeologous genes in rice were spread over the entire length of all 12 chromosomes (Supplemental Fig. C), it appears the maize endosperm gene set is rather comprehensive and provides a good representation of the entire cereal transcriptome. Although clustering was moderate, there seemed to be a bias of genes to be closer to the ends rather than in the central portions of the chromosomes, which is consistent with the general assumption of gene distribution in chromosomes. Based on gene density, there were fewer endosperm-expressed genes on Chromosomes 10, 11, and 12. The highest density of endosperm-expressed genes is found on Chromosome 3, which has nearly three times more than Chromosome 11 (Table 3).

DISCUSSION

EST sequencing has been considered an efficient way of gene discovery in many common species (http://www.ncbi.nlm.nih.gov/dbEST). For maize alone, there have been two major large-scale EST sequencing projects (Gai et al. 2000; Qiu et al. 2003) and a survey of the mRNA complement of maize sperm cells (Engel et al. 2003). Here, we focused on the maize endosperm. To achieve a high coverage for endosperm transcripts, we used four strategies to enrich low-abundance RNA clones. First, the endosperm tissues were dissected at various stages of development, that is, 4–23 DAP. Tissues from 4–6 DAP and tissues from 7–23 DAP were pooled. This division was made to separate two major phases of endosperm development. The early mitotic phase of development is represented by 4–6 DAP, characterized by active cell proliferation throughout the endosperm. The basal transfer layer undergoes histodifferentiation, but there is little other cytological evidence of cell specialization (Becraft 2001). Between 7 and 10 DAP, mitotic divisions become localized to the periphery, differentiation of the aleurone becomes evident, and central cells enlarge, undergo endoreduplication, and begin to accumulate proteins and starch. By 16 DAP, the maturation program has initiated, preparing the seeds for desiccation and dormancy, and by 23 DAP, the process of desiccation has begun. Thus, these libraries encompass the major events in endosperm development. Separation between early and later phases of development provided the first level of normalization, because of the differential expression of many genes. A second, more powerful normalization step involved arraying all the randomly picked clones on filters and selecting for those with low abundance in the cDNA population. A third strategy of enrichment was the subtraction of clones already analyzed by hybridization with probes from a pool of sequenced DNAs. Fourth, the efficiency of hybridization techniques to screen against abundant cDNAs depends on having full-length cDNAs as well. By comparing the 5′-sequence reads from all the cDNA libraries analyzed to a known protein database, it was found that ~98% of the clones have start codons in the 5′-end. Because sequence reads from both the 5′- and 3′-ends were available, identification of the N- and C-terminal ends of known proteins became feasible.

At this time, it is still difficult to predict the size of the maize endosperm transcriptome. One would expect that many mutant genes might not have a phenotype and, therefore, belong to a class other than the 300 essential genes in maize endosperm (Neuffer and Sheridan 1980). There is every indication that our cDNA libraries were not exhaustively sequenced and could yield additional novel cDNAs. Nevertheless, it appears that a substantial portion of endosperm-expressed mRNAs from >5000 unique genes were captured and provided us an opportunity to gain insight into tissue-specific gene expression in maize and rice. Furthermore, the detection level of cDNAs expressed in just a few endosperm cells could be limited, regardless of the depth of additional cDNA sequencing. One also has to consider that many genes are members of tandemly amplified gene families. This is not only true for the storage protein genes that are expressed in the endosperm, but also for many other genes. A recent analysis of rice Chromosome 10 indicated that a total of 25% of its genes fall into this category, exceeding the 17% previously reported for the entire *Arabidopsis* genome (Rice Chromosome 10 Sequencing Consortium 2003). Such an extent of redundancy in gene copies would probably make it difficult to detect phenotypes of gene knockouts for a large percentage of genes, unless the sequences of members of gene families are very conserved and can be knocked out by RNAi, as recently shown for the 22-kDa α zein genes (Segal et al. 2003). Gene knockouts by insertional mutagenesis can be identified by comparing junction sequences by hybridization or in silico with full-length cDNAs. The latter would require additional sequences in the gaps of larger cDNAs, whereas hybridiza-

tion could be performed directly. Nevertheless, there is a large collection of maize sequences in the GSS section of GenBank. More importantly, a large proportion of them are derived from a uniform Mutator backcross collection of the W22 inbred line used in this study. Therefore, these W22 GSS should match the cDNA sequences unambiguously. However, a detailed analysis of tagged endosperm-expressed genes requires additional segregation analysis of mutant phenotypes that will be published somewhere else (http://endosperm.org/).

Comparative mapping to the complete genome sequence of a close maize relative, the rice genome, serves two purposes: one, it adds functional annotation to the rice genome; and two, it leads to the prediction of the location of these genes in the maize genome for those chromosomal regions that are syntenic in both. We expected that most if not all maize endosperm-expressed genes could be detected in rice. However, from 5326 repeat-free maize unigenes, only 4139 (78%) of them could be mapped to 3108 locations on the rice genome. It is not surprising to find 1020 fewer map positions in rice than the total number of maize genes. The reason for this is that maize originated by the hybridization of two progenitors that split from the sorghum progenitor 11.9 Mya (Swigoňová et al. 2004). Although our results indicate that maize might have kept, in many cases, both copies of the two duplicated genes, they also suggest that the second copy of the duplicated genes was lost at high frequency, consistent with our analysis of orthologous genomic sequences (Lai et al. 2004). Based on our mapping standard (*E*-value $\leq e-10$), ~22% (1187 out of 5326) of the maize unigenes genes have diverged between rice and maize, been lost in rice, or gained in maize over the last 50 million years, indicating a rapid change of gene content between the two species (Supplemental Table C). In this respect, it is interesting to note that the same orthologous intervals in two maize inbred lines exhibit an absence of genes, not only in the orthologous interval but also from the entire genome of one inbred line (Song and Messing 2003). Given such a difference between two inbred lines of maize, it would not be surprising if a maize inbred line would differ in the same way from rice.

On the other hand, the large percentage of conserved genes between maize and rice could be accounted for in two ways. The endosperm tissues of maize and rice have a very similar function and storage capacity, although rice seeds are much smaller than maize kernels. A large percentage of the endosperm transcriptome encodes many common cellular functions and may represent a large fraction of the total plant transcriptome. If we ask the question how many rice genes have homology to maize endosperm ESTs, then the number of 5367 was much higher than the 4139 maize endosperm cDNA matches. The difference of 1239 genes in the rice genome would suggest that 30% of the rice genes are part of gene families that could vary significantly in size and would exceed even the proportion of 25% previously determined for a single chromosome (Rice Chromosome 10 Sequencing Consortium 2003). The distribution of endosperm-expressed genes throughout the rice genome and their large number is consistent with the expectation that the majority of these genes are expressed throughout rice development and that only a small number of them are expressed uniquely in endosperm. Looking at the function of these genes, we can find many encoding proteins responsible for cellular functions that are required in many tissues. Therefore, the number of endosperm-expressed genes could be much higher than described here and represent a substantial portion of the entire transcriptome in maize and rice. The results of mapping 6591 transcripts from 19 different tissues also revealed a similar gene-rich and gene-poor chromosomal distribution as in our study (Supplemental Fig. C), consistent with the expectation that the expression of a basic set of genes in endosperm is required throughout plant development (Wu et al. 2002).

METHODS

Library Construction

Two sets of endosperm tissue were harvested from maize inbred line W22, one at 4–6 DAP and the second pooled at 2-d intervals, from 7–23 DAP, during the summer of 2001 and used to purify RNA using mRNA isolation system II (Promega). The intactness of RNA was checked by gel electrophoresis/blot hybridization, and the purity was determined by its A_{260}/A_{280} ratio. The cDNA libraries were constructed using a cDNA synthesis kit (Cat #200401–5) from Stratagene. Clones were plated on LB medium containing X-GAL and IPTG, and white colonies were selected by robotic transfer into 96-well microtiter plates. Random sampling of isolates was used to determine insert sizes by agarose gel electrophoresis.

Filter Hybridization

High-density filters were processed for hybridization experiments as described previously (Song and Messing 2002). The filters were hybridized with the same mRNA populations used for library construction by preparing [α-^{32}P]dATP-radiolabeled cDNA. Autoradiograms were read with a high-density filter reader (HDFR2) from Incogen, Inc. A weak or absent hybridization signal was then used as the criterion for selecting clones for sequencing. To avoid selecting samples that lacked a signal because of failed colony growth, filters were also hybridized with labeled pBluescript DNA.

EST Sequencing and Assembly

The cDNA inserts were sequenced from both ends using standard M13F and M13R primers (Vieira and Messing 1982) and an ABI PRISM BigDye Terminator v3.1 Cycle Sequencing Ready Reaction kit (Applied BioSystems) on automated capillary sequencers (ABI 3700 and 3730xl). Base-calling was done using the phred program (Ewing et al. 1998), with lower quality sequences being trimmed. A total of 23,348 successful sequence reads were obtained and deposited into GenBank.

Computational Analysis

All sequences were assembled using the CAP3 computer program (Huang and Madan 1999) with the parameters set at 95% identity over 40 bp. The maize cDNA sequence assemblies were imported into the Sputnik EST and cluster analysis application for functional annotation (Rudd et al. 2003). Functional assignments were performed using BLASTX (threshold value of $<1e-10$) against the MIPS catalog of functionally assigned proteins (funcat; Frishman et al. 2001).

For comparison to rice, maize cDNA sequences were related to the 12 pseudomolecules of the rice genome generated from the BAC/PAC sequences of the IRGSP (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml). The coding sequences of all the putative genes were predicted using the FGENESH program with the monocot trained program. The unigene set of maize ESTs was screened for potential repeat sequences using the cereal repeat database. The repeat-free EST sequences were then subjected to homology searches against the database of the predicted rice coding sequences.

ACKNOWLEDGMENTS

We thank G. Fuks, S. Kavchok, G. Keizer, A.B. Nelson, S. Young, and V. Zohovetz for technical assistance. This work was supported by NSF grant 0077676.

REFERENCES

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

Becraft, P.W. 2001. Cell fate specification in the cereal endosperm. *Semin. Cell Dev. Biol.* **12:** 387–394.
Engel, M., Chaboud, A., Dumas, C., and McCormick, S. 2003. Sperm cells of *Zea mays* have a complex complement of mRNAs. *Plant J.* **34:** 697–707.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.
Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., Hu, X., et al. 2002. Sequence and analysis of rice chromosome 4. *Nature* **420:** 316–320.
Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., and Mewes, H.W. 2001. Functional and structural genomics using PEDANT. *Bioinformatics* **17:** 44–57.
Gai, X., Lal, S., Xing, L., Brendel, V., and Walbot, V. 2000. Gene discovery using the maize genome database ZmDB. *Nucleic Acids Res.* **28:** 94–96.
Gale, M.D. and Devos, K. 1998. Plant comparative genetics after 10 years. *Science* **282:** 656–659.
Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296:** 92–100.
Helentjaris, T., Weber, D., and Wright, S. 1988. Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphism. *Genetics* **118:** 353–363.
Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9:** 868–877.
Lai, J., Ma, J., Swigoňová, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.-J., Jeong, O.-Y., Bennetzen, J.L., et al. 2004. Gene loss and movement in the maize genome. *Genome Res.* (this issue).
Neuffer, G. and Sheridan, W.F. 1980. Defective kernel mutants of maize. I. Genetic and lethality studies. *Genetics* **95:** 929–944.
Qiu, F., Guo, L., Wen, T., Liu, F., Ashlock, D., and Schnable, P. 2003. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. *Plant Physiol.* **133:** 475–481.
Rice Chromosome 10 Sequencing Consortium. 2003. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* **300:** 1566–1569.
The Rice Full-Length cDNA Consortium. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301:** 376–379.
Rudd, S., Mewes, H.-W., and Mayer, K.F.X. 2003. Sputnik: A database platform for comparative plant genomics. *Nucleic Acids Res.* **31:** 128–132.
Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J., Niimura, Y., Cheng, Z., Nagamura, Y., et al. 2002. The genome sequence and structure of rice chromosome 1. *Nature* **420:** 312–316.
Scanlon, M. and Myers, A. 1998. Phenotypic analysis and molecular cloning of *discolored-1* (*dsc1*), a maize gene required for early kernel development. *Plant Mol. Biol.* **37:** 483–493.
Scanlon, M., Stinard, P., James, M., Myers, A., and Robertson, D. 1994. Genetic analysis of 63 mutations affecting maize kernel development isolated from *Mutator* stocks. *Genetics* **136:** 281–294.
Segal, G., Song, R., and Messing, J. 2003. A new *opaque* variant of maize by a single dominant RNAi-inducing transgene. *Genetics* **165:** 387–397.
Song, R. and Messing, J. 2002. Contiguous genomic DNA sequence comprising the 19-kDa-zein gene family from *Zea mays*. *Plant Physiol.* **130:** 1626–1635.
———. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci.* **100:** 9055–9060.
Song, R., Llaca, V., and Messing, J. 2002. Mosaic organization of orthologous sequences in grass genomes. *Genome Res.* **12:** 1549–1555.
Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res.* (this issue).
Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G., and Tanksley, S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* **14:** 1441–1456.
Vieira, J. and Messing, J. 1982. The pUC plasmids, an M13mp7 derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19:** 259–268.
Wu, J., Maehara, T., Shimokawa, T., Yamamoto, S., Harada, C., Takazaki, Y., Ono, N., Mukai, Y., Koike, K., Yazaki, J., et al. 2002. A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14:** 525–535.
Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296:** 79–92.

WEB SITE REFERENCES

http://endosperm.org/; segregation analysis of mutant endosperm phenotypes.
http://rgp.dna.affrc.go.jp; International Rice Genome Sequencing Project, IRGSP.
http://www.ncbi.nlm.nih.gov/dbEST; NCBI EST sequencing.
http://www.softberry.com; FGENESH.
http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml; TIGR, pseudomolecules for all chromosomes.

Received April 10, 2004; accepted in revised form July 28, 2004.