Gene discovery and annotation using LCM-454 transcriptome sequencing

Scott J. Emrich,[1,2,6] W. Brad Barbazuk,[3,6] Li Li,[4] and Patrick S. Schnable[1,4,5,7]

[1]*Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, Iowa 50010, USA;* [2]*Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50010, USA;* [3]*Donald Danforth Plant Science Center, St. Louis, Missouri 63132, USA;* [4]*Interdepartmental Plant Physiology Graduate Major and Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, Iowa 50010, USA;* [5]*Department of Agronomy and Center for Plant Genomics, Iowa State University, Ames, Iowa 50010, USA*

454 DNA sequencing technology achieves significant throughput relative to traditional approaches. More than 261,000 ESTs were generated by 454 Life Sciences from cDNA isolated using laser capture microdissection (LCM) from the developmentally important shoot apical meristem (SAM) of maize (*Zea mays* L.). This single sequencing run annotated >25,000 maize genomic sequences and also captured ~400 expressed transcripts for which homologous sequences have not yet been identified in other species. Approximately 70% of the ESTs generated in this study had not been captured during a previous EST project conducted using a cDNA library constructed from hand-dissected apex tissue that is highly enriched for SAMs. In addition, at least 30% of the 454-ESTs do not align to any of the ~648,000 extant maize ESTs using conservative alignment criteria. These results indicate that the combination of LCM and the deep sequencing possible with 454 technology enriches for SAM transcripts not present in current EST collections. RT-PCR was used to validate the expression of 27 genes whose expression had been detected in the SAM via LCM-454 technology, but that lacked orthologs in GenBank. Significantly, transcripts from ~74% (20/27) of these validated SAM-expressed "orphans" were not detected in meristem-rich immature ears. We conclude that the coupling of LCM and 454 sequencing technologies facilitates the discovery of rare, possibly cell-type-specific transcripts.

[The sequence data from this study have been submitted to GenBank under accession nos. DW724699–DW985434.]

Although genome sequencing technology has become progressively more efficient over the past decade, the sequencing of complex genomes remains expensive. Expressed Sequence Tag (EST) sequencing provides an attractive alternative to whole-genome sequencing because this technique produces sequences of the transcribed portions of genes at a fraction of the cost of sequencing complete chromosomes. Even so, because genes are differentially expressed, multiple tissues must be sampled, and, when using traditional (Sanger) methods, these EST projects require substantial investments in library construction and sequencing, particularly if the goal is to capture rare transcripts.

Recently, 454 Life Sciences developed a scalable, highly parallel DNA sequencing system that is 100 times faster than standard sequencing methods and is capable of sequencing >200,000 fragments per 4-h run (Margulies et al. 2005). This increase in throughput comes at the expense of read length. On average, 454 sequence reads are only ~100 bp in length, and in addition, this technology does not capture read-pair information (Margulies et al. 2005). Hence, the assembly of 454 sequences from samples that contain large amounts of repetitive DNA such as eukaryotic genomes may prove problematic for conventional fragment assembly programs.

In contrast, the read-length limitation associated with 454 technology is less of a concern for transcriptome sequencing and analysis. This is because transcriptomes are smaller than the genomes from which they are derived and typically contain less repetitive DNA. Using laser-capture microdissection (LCM) (for review, see Schnable et al. 2004) to isolate transcripts that accumulate in specific cell types has the potential to further reduce the size of a target transcriptome. Because 454 technology avoids expensive cloning-based library construction, it is feasible to sequence a wide variety of LCM-derived cDNA samples, thereby increasing the recovery of highly specialized transcripts. Moreover, 454 technology combined with LCM is particularly well suited for EST-based gene discovery because it generates hundreds of thousands of tags per run, greatly increasing the chances of capturing rare transcripts.

Here, we report the sequencing of cDNA extracted from developmentally important Shoot Apical Meristem (SAM) cells (Baurle and Laux 2003; Guyomarc'h et al. 2005) using the LCM-454 approach. A single 454 sequencing run was able to annotate >25,000 maize genomic sequences and capture transcripts from nearly 400 "orphan genes" (Fu et al. 2005). Interestingly, experimental validation suggests that not only are "orphan" transcripts discovered using the LCM-454 approach, but most of these genes are undetectable in cDNA samples from other tissues including meristem-rich immature ears. LCM-454 sequencing is, therefore, an efficient gene-discovery platform when applied to highly specialized organs such as the SAM.

[6]These authors contributed equally to this work.
[7]Corresponding author.
E-mail schnable@iastate.edu; fax (515) 294-5256.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5145806. Freely available online through the *Genome Research* Open Access option.

Results

Gene discovery and annotation using 454 sequencing

As of December 2005, >650,000 maize EST sequences obtained from diverse tissues and genotypes had been deposited in Gen-

Bank, including sequences derived from libraries prepared from specialized structures such as the vegetative shoot apex. The apex contains both newly formed leaves and SAM cells that initiate all above-ground tissue in plants. The developmentally important SAM cells, however, comprise only a very small portion of the apex. Consequently, it is difficult to capture rare SAM-specific transcripts by sequencing ESTs from an apex library.

One means to obtain rare transcripts from specific cell types (e.g., those that comprise the SAM) is to extract and clone mRNA from individual cell types using LCM (Asano et al. 2002). This approach, however, requires a significant investment in cDNA sequencing including library construction. As a potential alternative, we attempted to discover rare transcripts by directly sequencing cDNA using the high-throughput 454 sequencing platform. Maize cDNA was extracted from multiple SAMs using LCM as described by Nakazono et al. (2003), amplified (Methods), and sequenced by 454 Life Sciences. After removing poly(A/T) tails from these reads (Methods), the ~261,000 resulting SAM ESTs had an average length of 101 bp.

The 454-ESTs were BLASTN-aligned to a variety of maize sequence databases (Table 1). In total, >93% of the 454 SAM EST sequences matched maize ESTs, GSSs, repeats, or organelle genomes. We and colleagues had previously generated ~31,000 ESTs from a cDNA library prepared from hand-dissected maize apices (Methods). The 454-ESTs were aligned to the ~18,560 unigene transcripts assembled from these Apex ESTs (Methods). More than 70% of the SAM 454-ESTs did not align to the Apex ESTs from this SAM-enriched library. GenBank contains >600,000 additional maize ESTs (Methods). More than 30% of the 454-ESTs did not align to this extensive collection of ESTs. These results indicate that this 454 sequencing run captured ESTs from many maize genes without previous evidence of expression.

We previously assembled ~880,000 "gene-enriched" B73 genomic sequences into Maize Assembled Genomic Islands (MAGIs) (http://magi.plantgenomics.iastate.edu). Previous alignments between the 114,173 MAGIs and a unigene set composed of the ~419,000 maize ESTs available in GenBank prior to February 2004 provided evidence that ~20,900 MAGIs contain at least portions of expressed genes (Fu et al. 2005). Similar alignments of the 454 SAM ESTs provide evidence that ~25,800 MAGIs contain at least portions of expressed genes. Significantly, 15,521 of these ~25,800 MAGIs did not have prior expression evidence from the alignments to the ~419,000 maize ESTs, which included the Apex ESTs. These results suggest that the representation of rare and/or SAM-specific transcripts has been enriched by the deep sequencing of cDNA isolated from SAM tissue. Hence, we conclude that LCM-454 sequencing is an efficient approach for the large-scale validation of gene expression.

Table 1. Genic sequences captured by a single 454 run compared to other gene-enriched sequencing approaches

Source database	No. of matching 454-ESTs	No. of novel 454-ESTs
UGA-ISU Apex unigenes (*N* = 18,558)	73,145	187,591 (71.9%)
GenBank maize ESTs (*N* = 647,685)	179,912	80,824 (31.0%)
ESTs + ISU MAGI 3.1 (GSS) (*N* = 862,158)	239,113	21,623 (8.3%)
ESTs + ISU MAGI 3.1 + organelle genomes + repeats (*N* = 877,431)	244,328	16,408 (6.3%)
ESTs + ISU MAGI 3.1 + organelle + monocot ESTs (*N* = 1,282,226)	245,339	15,397 (5.9%)

We previously reported (Fu et al. 2005) that ~5% of expressed maize genes are "orphans" relative to known sequence databases including GenBank and dbEST. Consistent with this previous observation, we estimate that relative to current plant databases (Methods), ~15,400 (6%) of the 454 SAM ESTs were transcribed from orphan genes. Because ESTs are differentially expressed and full-length cDNAs are not available, it is difficult to determine exactly how many unique SAM-expressed genes are orphans. It is possible, however, to estimate the overall frequency of orphans by confirming the expression of a sample of genes. A total of 9944 of the predicted maize genes described by Fu et al. (2005) were deemed, based on 454-ESTs data, to be expressed in the SAM (Methods). Of these, 914 (9%) do not have homologous sequences in monocot EST databases (Methods). Of these, 390 genes do not have matches to non-EST databases, including repeat databases (Methods). Hence, a single 454 sequencing run provided EST-based support for the expression of >9000 SAM-expressed genes, of which 390 are nonrepetitive orphans.

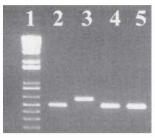
Validation of orphan expression

RT-PCR was used to confirm the expression of a sample of the orphan genes detected among the 454-ESTs. A set of 42 MAGIs that contained orphan FGENESH-predicted genes was selected for analysis that (1) aligned to 454-ESTs, (2) contained at least one intron, and (3) yielded primers that met our design criteria. Criteria 2 and 3 were used to be consistent with a prior study of maize orphans (Fu et al. 2005). As in the previous study, PCR primers were designed based on FGENESH-predicted exonic sequences in each gene. Initially, PCR amplification was performed using B73 genomic DNA as a template. A total of 38 of the 42 primer pairs yielded genomic PCR products of the expected sizes. To obtain an independent test of whether these orphan genes are indeed expressed, the 38 primer pairs were then used to conduct PCR experiments on three pools of cDNA derived from (1) SAMs, (2) meristem-rich immature ears, and (3) multiple tissues (Methods). If a single RT-PCR band was obtained, it was sequenced. Of the 38 primer pairs, 27 produced RT-PCR products that were of the correct size and whose sequence matched the MAGIs from which the primers were designed. All 27 of these orphans were expressed in the SAM (Fig. 1). Based on these results, we conclude that many of the orphans detected among the 454-ESTs are, indeed, expressed. Eleven of the 27 orphans were expressed in at least one of the other two cDNA pools. Interestingly, 20/27 (74%) of the RT-positive orphan transcripts that were detected in the SAM were not detected in the meristem-rich immature ears. This could be because of the substantial enrichment of meristems in the SAM sample and/or the existence of genes that are expressed in the SAM but not in the reproductive meristems present on the immature ears. In either case, this result provides further evidence for the value of coupling LCM and 454 sequencing for gene discovery.

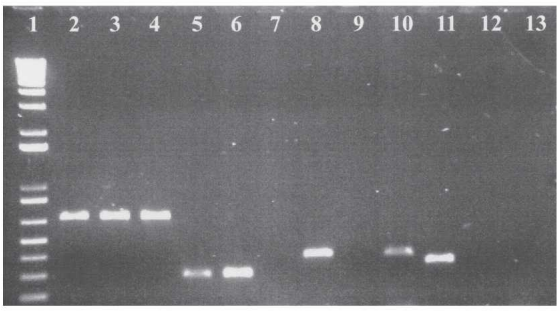
Discussion

Reductions in reagent volumes of Sanger sequencing reactions have substantially reduced costs without affecting read lengths or sequence accuracy (Smailus et al. 2005). Because of diminishing returns, replacement technologies are required to achieve additional cost savings and make possible grand challenges such as

A:



B:



C:

cDNA	RT-PCR results			
SAM	+	+	+	+
Immature ear	+	+	-	-
Complex pool	+	-	+	-
Number orphans	5	2	4	16

Figure 1. Experimental validation of the expression of orphan genes. (*A*) Test for genomic DNA contamination of cDNA. Primers that flank a 100-bp intron in the maize beta-tubulin6 (*tub6*) gene were used to amplify genomic DNA (lane *3*), SAM cDNA (lane *2*), immature ear cDNA (lane *4*), and the complex cDNA pool (lane *5*). (*B*) Examples of orphans with validated expression patterns. Primers designed to amplify (lanes *2–4*) MAGI_80343, (lanes *5–7*) MAGI_60450, (lanes *8–10*) MAGI_75030, and (lanes *11–13*) MAGI_30050 were used to amplify (lanes *2*,*5*,*8*,*11*) SAM cDNA, (lanes *3*,*6*,*8*,*12*) immature ear cDNA, and (lanes *4*,*7*,*9*,*13*) the pooled cDNA sample. (*C*) Summary of RT-PCR results for the 27 orphan genes. (+) Indicates that an RT-PCR product of the correct size was detected. Lane *1* of panels *A* and *B* contains the One KB Plus size standard (GIBCO BRL). Because primer dimers present in some lanes were cropped in both panels, the smallest size standard band shown is 200 bp.

the "$1000 genome" and the complete characterization of all expressed genes of an organism and their respective splice forms. Recently, 454 Life Sciences released a proprietary sequencing technology that quickly provides vast amounts of sequence data without the need to clone DNA prior to sequencing, further reducing the total effort required for large-scale sequencing projects. The reads obtained with 454 technology are, however, much shorter than traditional Sanger reads and are subject to a higher rate of base-calling errors, particularly in association with homopolymer runs.

This study provides experimental data that demonstrate the value of using 454 technology to sequence expressed sequences present in specific cell types isolated using laser capture microdissection (LCM). Because of its reduced size relative to the entire genome, an LCM-derived transcriptome can be more efficiently sampled, and therefore covered, by 454 sequencing. In addition, reducing the complexity of the transcriptome prior to sequencing by restricting cDNA recovery to specific tissues of interest was expected to increase the recovery of rare, tissue-specific transcripts. Approximately 261,000 454-ESTs were generated from LCM-collected SAM tissue. Only 70% of the 454 SAM ESTs align to ~648,000 maize ESTs. All potentially novel LCM-454 ESTs were aligned to the complete set of MAGIs. This corrected for LCM-454 ESTs derived from the same gene, but that did not overlap. These analyses validated the expression of >15,000 MAGIs that did not have prior evidence of expression.

As alluded to above, if a given gene is sampled by multiple nonoverlapping ESTs, the number of unique transcripts will be overestimated. Some traditional EST projects address this problem by sequencing the 3′-ends of cDNAs. It is not possible to specifically sequence the 3′-ends of cDNAs using 454 sequencing technology. Even so, our LCM-454 EST project greatly enriched for 3′-sequences and thereby minimized the overestimation of the number of unique transcripts in the SAM.

The 3′ enrichment achieved via LCM-454 sequencing is a consequence of the procedure used to amplify RNA from LCM-collected tissue (Methods), which results in relatively short cDNA fragments (~200–600 bp), all of which included the 3′-terminus of the corresponding transcripts. Prior to 454 sequencing, cDNAs are sheared. But because the target shear size is close to the size of our amplified cDNAs, most of our cDNAs were probably not sheared, or if sheared were removed via size selection prior to sequencing. Hence, we expected that a large percentage of our cDNAs were sequenced from their 3′-termini.

To test the degree to which our 454-ESTs were 3′-enriched, we identified a set of 3′-ESTs and a set of predicted maize genes that align to these 3′-ESTs (Methods) and then examined the distributions of LCM-454 EST alignments along the lengths of these genes. Using the 3′-ESTs (average length 565 bp), the beginning of the 454-EST/3′-EST alignment is within the first 20 bp upstream of the poly(A) site in 41% of the alignments, within the first 100 bp in 76% of the alignments, and within the first 300 bp in >95% of the alignments. Results for the substantially longer FGENESH-predicted genes (average length of 1039 bp) that aligned to LCM-454 ESTs were similar; the beginning of the 454-EST/MAGI alignment was within the first 20 bp upstream of the poly(A) site in 40% of the alignments, within the first 100 bp in 66% of the alignments, and within the first 300 bp in 90% of the alignments. This substantial 3′-enrichment provides confidence that the number of novel transcripts detected in this study is not substantially overestimated.

Current estimates suggest that up to 5% of expressed maize genes are "orphans" (Fu et al. 2005), that is, they match no genes isolated to date from any species. Previously, the expression of hypothetical orphan genes has been detected via large-scale efforts to specifically amplify associated transcripts from cDNA preparations (Fu et al. 2005; Xiao et al. 2005). In contrast, a single run of SAM 454-ESTs was able to detect the expression of ~400 expressed orphans; the expression of many of the tested orphans was validated via RT-PCR. Consequently, we conclude that the combination of LCM and 454 sequencing technologies is an efficient approach to discover and annotate genes.

Given the ease with which hundreds of thousands of ESTs can be generated, 454 technology makes it possible to obtain relative expression data on thousands of genes. Several high-throughput, sequencing-based quantitative expression analysis techniques are already available, most notably SAGE (Velculescu

et al. 1995) and MPSS (Brenner et al. 2000). Because both of these prior technologies produce short sequence signatures from discrete regions of transcripts, they provide a sensitive indicator of relative expression levels (Meyers et al. 2004); however, these techniques cannot provide sequence data over substantial portions of cDNAs and are therefore less well suited for applications such as SNP detection. In contrast, 454 sequencing could potentially recover virtually the entire template via "shotgun" sequencing of the transcriptome, and these tags are inherently better suited for discriminating the expression of members of highly conserved gene families because they are longer in length. Even so, under some circumstances it may be desirable to sequence SAGE libraries with 454 technology to leverage the advantages of both approaches to analyze expression digitally.

Following LCM, and prior to sequencing, we amplified RNA using a poly(T) primer. This procedure yielded fragments that are 3′-enriched relative to the entire transcriptome. The advantages of this 3′-enrichment are that it provides a better estimate of the numbers of unique transcripts within a particular transcriptome and greater depth of coverage is achieved in the 3′-ends of transcripts. The resulting data are well suited for gene discovery and in silico Northerns because transcripts are sampled at rates independent of their lengths. On the other hand, to obtain the sequence of a complete transcriptome, it would be desirable to avoid this 3′-enrichment by using random primers, rather than a poly(T) primer, to amplify the RNA following LCM. Our coverage modeling (data not shown) suggests that the ends of cDNAs will not be efficiently captured via 454 technology. Even so, 454 sequencing technology can efficiently capture the bulk of a transcriptome for use in applications such as gene discovery, annotation, and the discovery of polymorphisms. This is particularly true if transcriptome size is controlled by analyzing appropriate cell types, organs, or tissues via LCM.

Methods

Isolation of SAM mRNA

Maize (*Zea mays* inbred line B73) SAM tissue, which included Plastochron0 (P0) and P1, was extracted from ~10 14-d-old seedlings. This was achieved with modifications to the paraffin-embedding technique described by Kerk et al. (2003) and the Laser Capture Microdissection (LCM) technique described by Nakazono et al. (2003). Full details are described elsewhere (K. Ohtsu, M. Smith, S.J. Emrich, L.A. Borsuk, R.L. Zhou, T. Chen, X.L. Zhang, M.C.P. Timmermans, J. Beck, and B. Buckner, in prep.). A highly repeatable T7 RNA polymerase-based RNA amplification was performed as described by Nakazono et al. (2003) with some modifications to generate sufficient SAM cDNA for sequencing. Because a poly(T) primer was used for amplification, the resulting cDNA was enriched for the 3′-ends of transcripts.

454-EST sequencing and processing

Approximately 15 µg of LCM-derived cDNA was submitted to 454 Life Sciences, who ensured sample quality by checking the SAM cDNA on a 2% agarose gel and an Agilent bioanalyzer. The cDNA sample was then fractionated into smaller pieces (300–500 bp) that were subsequently polished (blunted). Short adaptors were then ligated on to each resulting fragment, which provide priming sequences for both amplification and sequencing, forming the basis of the single-stranded template library. Finally, one sequencing run was performed using the method of Margulies et al. (2005) and resulted in 288,992 EST sequences. 454 Life Sciences helped submit these sequences to the NCBI trace archive (http://www.ncbi.nlm.nih.gov/, accession nos. DW724699–DW985434), where they are available for independent analysis. These sequences were subsequently trimmed using Lucy (Chou and Holmes 2001) under default settings with the exception that sequences as short as 50 bp were not discarded; this returned 260,887 high-quality sequences, which we have submitted to dbEST and used for annotation. In addition, poly(A/T) tails were removed from raw 454 sequences with SeqClean (http://www.tigr.org/tgi/software) using default settings to ascertain the novelty of these sequences using longer, albeit lower quality, reads. In addition, contaminating sequences (150 sequences; 0.05% of total) were removed by SeqClean based on similarity to the *Escherichia coli* K12 (GenBank accession no. U00096) and *Lactococcus lactis* (GenBank accession no. AE005176) genomes and GenBank's Univec database.

Comparisons of 454-ESTs to public sequence databases

Maize ESTs ($N = 656{,}696$) were downloaded from GenBank in December 2005 and processed using SeqClean as described above. After eliminating 9011 contaminating or low-quality sequences, 29,615 maize ESTs (MESTs) sequenced by us from diverse cDNA libraries were extracted based on the presence of a poly(T) prefix of at least 10 bp; these were used to assess 3′-enrichment and putative sampling biases. For annotation purposes, another subset of 31,036 ESTs sequenced by us from a cDNA library generated by M. Scanlon's group (University of Georgia) from mRNA isolated from vegetative apices was extracted (K. Ohtsu, M. Smith, S.J. Emrich, L.A. Borsuk, R.L. Zhou, T. Chen, X.L. Zhang, M.C.P. Timmermans, J. Beck, and B. Buckner, in prep.). The Apex ESTs were assembled using CAP3 (Huang and Madan 1999) to generate unigenes using the following parameters: -p 98 -o 100 -y 20 -h 5.

The 454 SAM ESTs with poly(A/T) tails removed were compared to the 647,685 high-quality, unassembled maize ESTs, the maize Apex unigenes, ISU MAGIs version 3.1 (including singletons), maize chloroplast (GenBank accession no. X86563) and mitochondrial genome sequences (GenBank accession no. AY506529), and the ISU Cereal Repeat Database (http://magi.plantgenomics.iastate.edu) using BLAST. Nucleotide alignments with either an *E*-value $\leq 1e^{-8}$ or >70% identity over 50% of the EST length were deemed to have been previously discovered, providing a highly conservative estimate of novel gene discovery. The following TIGR Plant Gene Indices (http://www.tigr.org/tdb/tgi) downloaded in December 2005 were similarly searched for matches: HVGI release 9 (barley), OGI release 16 (rice), SBGI release 8 (sorghum), SOGI release 2 (sugar cane), and TAGI release 10 (wheat). Candidates were also compared to the *Arabidopsis* genome (ATH1.1con.01222004; http://www.arabidopsis.org), finished rice chromosome sequences (GenBank AP008207–AP008218), and the TIGR dicot gene indices used by Fu et al. (2005) to locate homologous sequences among plant ESTs.

Evidence of expression of SAM genes was determined by locating reciprocal best hits between predicted maize genes (Fu et al. 2005) and Lucy-trimmed 454-ESTs requiring a minimum *E*-value of $1e^{-20}$. Potential homologs were located among the monocot gene indices described above, and repeats were located against the MAGI Cereal Repeat Database v 3.1; both analyses used the novelty criteria previously described (Fu et al. 2005). Similarly, all putative orphan genes were compared to the GenBank nr database (BLASTN and BLASTX) and to the est_others database (BLASTN) on January 8, 2006 using netBLAST (blastcl3).

Annotation using 454-EST sequences

All 114,173 contigs from the partial maize inbred line B73 genome assembly MAGI 3.1 (Fu et al. 2005) were aligned to Lucy-trimmed 454 SAM ESTs using GeneSeqer and its maize-specific splice models (Usuka et al. 2000). Only alignments consisting of at least one exon of at least 50 bp in length and with identity ≥95% over at least 80% of the length of the 454-EST were used as evidence of expression. ESTs with >50 bp of repetitive sequence, as determined by a previously described masking procedure (Emrich et al. 2004), were ignored when the number of expressed MAGIs was calculated.

Validation of expression of orphan genes

RT-PCR and sequencing were conducted as described by Fu et al. (2005) on three pools of cDNA generated as described by Fu et al. (2005). The first pool was derived from amplified RNA isolated from SAMs via LCM as described above. The second pool was a complex mixture generated from multiple tissues harvested from B73 maize plants 79 d after planting in Ames, Iowa during the summer of 2005. The third pool was generated from immature, unpollinated top ears harvested from the inbred B73 59 d after planting (ears measured 1.25–2.5 cm in length). Based on RT-PCR results obtained using a pair of *tub6* primers that flank a 100-bp intron, these cDNA samples are free of detectable genomic DNA contamination.

Estimating the rate of sequencing errors in 454-ESTs

To estimate the rate of sequencing error in the ESTs generated by 454 Life Sciences, we aligned all ESTs to a collection of FGENESH-predicted maize cDNAs (Fu et al. 2005) using BLASTN and only used the best hit with an *E*-value < $1e^{-10}$. For all comparisons, at least 90% of the length of a 454 read had to match its corresponding benchmark to be considered a valid alignment. Although any disagreement is not conclusive proof of an error, we have shown that the MAGI-based maize cDNAs are of high enough quality (Fu et al. 2005) that these disagreements are likely errors in the 454 sequences.

Estimating the 3′-enrichment of 454-ESTs

A set of 8852 MAGIs was selected based on their alignment to 29,615 3′ maize ESTs with discernable poly(A/T) tails (Fu et al. 2005). Only the 5575 of these MAGIs that had an experimentally determined poly(A) site within 50 bp of the predicted termination of transcription were tested for alignment to the LCM-454 ESTs. A total of 32,075 LCM-454 ESTs aligned to these predicted genes. The LCM-454 ESTs were also directly aligned to the 29,615 3′-ESTs. A total of 36,258 LCM-454 ESTs aligned to the 3′-ESTs.

Acknowledgments

We thank Kazuhiro Ohtsu and Marianne Smith for preparing the SAM-specific cDNA used for 454 sequencing, Xiaolan Zhang of the Mike Scanlon laboratory (University of Georgia) for preparing the Apex EST library, Ruilian Zhou and other members of the Schnable Laboratory for sequencing this library, and Kazuhiro Ohtsu and Yan Fu and three anonymous reviewers for helpful comments on this manuscript. This research was supported by grants from the National Science Foundation (DBI-0321595 and DBI-0527192), ISU's Plant Science Institute, the Donald Danforth Plant Science Center, and Pioneer Hi-Bred; additional support was provided by Hatch Act and State of Iowa funds.

References

Asano, T., Masumura, T., Kusano, H., Kikuchi, S., Kurita, A., Shimada, H., and Kadowaki, K. 2002. Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: Toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J.* **32:** 401–408.

Baurle, I. and Laux, T. 2003. Apical meristems: The plant's fountain of youth. *Bioessays* **25:** 961–970.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18:** 630–634.

Chou, H.H. and Holmes, M.H. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics* **17:** 1093–1104.

Emrich, S.J., Aluru, S., Fu, Y., Wen, T.J., Narayanan, M., Guo, L., Ashlock, D.A., and Schnable, P.S. 2004. A strategy for assembling the maize (*Zea mays* L.) genome. *Bioinformatics* **20:** 140–147.

Fu, Y., Emrich, S.J., Guo, L., Wen, T.J., Ashlock, D.A., Aluru, S., and Schnable, P.S. 2005. Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. *Proc. Natl. Acad. Sci.* **102:** 12282–12287.

Guyomarc'h, S., Bertrand, C., Delarue, M., and Zhou, D.X. 2005. Regulation of meristem activity by chromatin remodelling. *Trends Plant Sci.* **10:** 332–338.

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9:** 868–877.

Kerk, N.M., Ceserani, T., Tausta, S.L., Sussex, I.M., and Nelson, T.M. 2003. Laser capture microdissection of cells from plant tissues. *Plant Physiol.* **132:** 27–35.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., Ghazal, H., and Decola, S. 2004. The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res.* **14:** 1641–1653.

Nakazono, M., Qiu, F., Borsuk, L.A., and Schnable, P.S. 2003. Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: Identification of genes expressed differentially in epidermal cells or vascular tissues of maize. *Plant Cell* **15:** 583–596.

Schnable, P.S., Hochholdinger, F., and Nakazono, M. 2004. Global expression profiling applied to plant development. *Curr. Opin. Plant Biol.* **7:** 50–56.

Smailus, D.E., Marziali, A., Dextras, P., Marra, M.A., and Holt, R.A. 2005. Simple, robust methods for high-throughput nanoliter-scale DNA sequencing. *Genome Res.* **15:** 1447–1450.

Usuka, J., Zhu, W., and Brendel, V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16:** 203–211.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.

Xiao, Y.L., Smith, S.R., Ishmael, N., Redman, J.C., Kumar, N., Monaghan, E.L., Ayele, M., Haas, B.J., Wu, H.C., and Town, C.D. 2005. Analysis of the cDNAs of hypothetical genes on *Arabidopsis* chromosome 2 reveals numerous transcript variants. *Plant Physiol.* **139:** 1323–1337.

Received January 12, 2006; accepted in revised form August 24, 2006.

Gene discovery and annotation using LCM-454 transcriptome sequencing

Scott J. Emrich, W. Brad Barbazuk, Li Li, et al.

Genome Res. 2007 17: 69-73 originally published online November 9, 2006
Access the most recent version at doi:10.1101/gr.5145806