Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing[1][W][OA]

Andreas P.M. Weber, Katrin L. Weber, Kevin Carr, Curtis Wilkerson, and John B. Ohlrogge*

Department of Plant Biology (A.P.M.W., K.L.W., J.B.O.), and Bioinformatic Support Core, Research Technologies Support Facility (K.C., C.W.), Michigan State University, East Lansing, Michigan 48824–1312

Massively parallel sequencing of DNA by pyrosequencing technology offers much higher throughput and lower cost than conventional Sanger sequencing. Although extensively used already for sequencing of genomes, relatively few applications of massively parallel pyrosequencing to transcriptome analysis have been reported. To test the ability of this technology to provide unbiased representation of transcripts, we analyzed mRNA from Arabidopsis (*Arabidopsis thaliana*) seedlings. Two sequencing runs yielded 541,852 expressed sequence tags (ESTs) after quality control. Mapping of the ESTs to the Arabidopsis genome and to The Arabidopsis Information Resource 7.0 cDNA models indicated: (1) massively parallel pyrosequencing detected transcription of 17,449 gene loci providing very deep coverage of the transcriptome. Performing a second sequencing run only increased the number of genes identified by 10%, but increased the overall sequence coverage by 50%. (2) Mapping of the ESTs to their predicted full-length transcripts indicated that all regions of the transcript were well represented regardless of transcript length or expression level. Furthermore, short, medium, and long transcripts were equally represented. (3) Over 16,000 of the ESTs that mapped to the genome were not represented in the existing dbEST database. In some cases, the ESTs provide the first experimental evidence for transcripts derived from predicted genes, and, for at least 60 locations in the genome, pyrosequencing identified likely protein-coding sequences that are not now annotated as genes. Together, the results indicate massively parallel pyrosequencing provides novel information helpful to improve the annotation of the Arabidopsis genome. Furthermore, the unbiased representation of transcripts will be particularly useful for gene discovery and gene expression analysis of nonmodel plants with less complete genomic information.

For approximately 30 years, sequencing of DNA by the dideoxy terminator strategy introduced by Sanger (1977) has provided the basis for almost all available information about nucleotide sequences. Pyrosequencing is an alternative technology that detects the pyrophosphate released during DNA polymerase-catalyzed incorporation of nucleotides. The pyrophosphate liberated with each nucleotide addition can generate light in a reaction coupled to ATP sulfurylase and luciferase. Although proposed as early as 1985 (for review, see Ahmadian et al., 2006), only recently have instruments become available that solve several technical details and allow large-scale use of this approach (Margulies et al., 2005). The GS20 instrument used in this study performs the sequencing reactions in a massively parallel fashion, which is referred to in this article as pyrosequencing. Double-stranded DNA is fragmented, individual molecules are attached to nanobeads, amplified, and each bead is deposited in wells of a high-density plate with picoliter reaction volumes. As sequencing reagents pass over the plate, light emitted from each well is recorded. Because typically over 300,000 wells simultaneously provide data for a collection of DNA fragments, it is possible to obtain 20 to 30 Mb or more of sequence information in a single 4.5-h run. Currently, read lengths from each DNA fragment are short (average 100–110 bp) compared to Sanger sequencing; however, this figure is expected to increase substantially with instrument, reagent, and protocol improvements.

Most applications of pyrosequencing have involved analysis of genomic DNA (e.g. Poinar et al., 2006). The goal of this study was to evaluate the ability of pyrosequencing to provide information on transcript populations from plant tissues. Sequencing of cDNA copies of transcripts has provided one of the most cost-effective approaches for gene discovery because most sequences obtained are protein coding. The absence of introns and intergenic regions greatly enhances the information content and eases interpretation of the data. Sequencing of a few thousand randomly selected cDNA clones has often been the initial step that led to the identification of key enzymes specific for biosynthesis of a wide range of natural products (Ohlrogge and Benning, 2000; Weber et al., 2004). Hundreds of EST projects that span the phylogenetic diversity of

[1] This work was supported by a Strategic Partnership Grant (Next Generation Sequencing Center) of the Michigan State University Foundation (to A.P.M.W. and J.B.O).

* Corresponding author; e-mail ohlrogge@msu.edu; fax 517–353–1926.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with journal policy described in the Instructions for Authors (www.plantphysiol.org) is: John Ohlrogge (ohlrogge@msu.edu).

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

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.107.096677

biology have also provided rich datasets for comparative genomics (Barbier et al., 2005), including variation in protein sequences that allow identification of conserved motifs, active sites, and enzyme specificity-determining residues (Mayer et al., 2005).

For this study, we chose to evaluate Arabidopsis (*Arabidopsis thaliana*) because its genome sequence is complete, more than 700,000 conventional ESTs are available, and the genome annotation is the most advanced for any higher plant. In addition, we chose to analyze 8-d-old seedlings for which the transcript population has been well characterized by microarrays (Schmid et al., 2005). Together, these factors allow advanced comparisons of pyrosequencing data to genomic and transcript data. Because of the substantially different methods used to prepare DNA for Sanger and pyrosequencing, there are a number of relative advantages and disadvantages of each approach. Pyrosequencing does not require cloning of the DNA and therefore avoids certain biases that can be introduced by enzyme steps or by instability of sequences in *Escherichia coli*. However, it is not clear whether other biases might be associated with the fragmentation, amplification, or other steps associated with massively parallel pyrosequencing. In this study, we addressed several types of potential bias and provide an analysis of the advantages and disadvantages of current massively parallel pyrosequencing data compared to conventional EST sequencing and other methods of transcript profiling.

RESULTS

To isolate transcripts, RNA was extracted from aerial tissues of 8-d-old light-grown Arabidopsis seedlings and mRNA was prepared by two rounds of oligo(dT) purification. First-strand cDNA was synthesized with oligo(dT) primer and second strand following protocols of a commercial cDNA library preparation kit. After end-repair adaptors were ligated, approximately 3 μg of the cDNA population were sheared by nebulization, and DNA sequencing was performed with the GS20 genome-sequencing system (Margulies et al., 2005).

Access to Data

Access to all EST data obtained in this study and tools for mining the data are facilitated through an Excel workbook that is available in the supplemental data (Supplemental Table S1) for download from the journal Web site. The workbook contains spreadsheets that list the number of pyrosequencing and dbEST hits to The Arabidopsis Information Resource (TAIR) 7.0 gene and cDNA models (release date March 2007) and pyrosequencing ESTs that map to the Arabidopsis genome, but that do not hit an annotated gene model. Various filters can be applied to the data to search for gene models that are hit by pyrosequencing ESTs, but not by conventional ESTs. The Generic Genome Browser, GBrowse (Stein et al., 2002), was used for visualization of gene models, Sanger ESTs, and pyrosequencing ESTs mapping to the Arabidopsis genome. Gene and EST identifiers in the Excel workbook are linked to the GBrowse display of the data, thus providing direct access to the mapping data and to EST sequences. In addition, the data can be searched and explored graphically at http://genomics.msu.edu/cgi-bin/gbrowse/A_thaliana.

Pyrosequencing Provides Very Deep Coverage of the Transcriptome

A summary of the number of ESTs and their mapping to other Arabidopsis sequences is presented in Table I. Two consecutive GS20 pyrosequencing runs generated 555,326 raw reads, totaling 60,018,332 nucleotides (nt). After quality, complexity, and primer trimming, 541,852 ESTs remained. Of these, 88.7% had at least one significant alignment to the Arabidopsis genome. The 11.3% of sequences that did not map to the genome did not produce any significant hits by BLAST to the National Center for Biotechnology Information (NCBI) nonredundant protein database. Furthermore, they disproportionately consisted of short or long reads and ESTs with extensive mononucleotide runs indicative of poly(A) tails and/or low-quality sequence.

The TAIR 7.0 Arabidopsis dataset (release date March 2007) contains 37,020 predicted cDNA models that are derived from 32,041 predicted gene loci. Most (87.1%) pyrosequencing ESTs had at least one significant alignment to a TAIR 7.0 gene model. These ESTs detected transcription of 21,877 cDNA models from 17,449 gene loci, which is 59% of the TAIR 7.0 cDNA models. Over 10,000 of the 17,449 gene loci were represented by at least three ESTs and 2,867 were represented by more than 25 ESTs (Supplemental Fig. S1). Performing a second sequencing run only increased the number of genes identified by 10%, but increased the overall sequence coverage by approximately 50% (from 7 to 10.3 Mb). Microarray data indicate 55% to 67% of Arabidopsis genes are expressed

Table I. *Summary of pyrosequencing ESTs and their mapping to the Arabidopsis genome and to conventional ESTs*

	No.	Percent
ESTs after quality control	541,852	(100)
Map to Arabidopsis genome	480,696	88.7
Map to Arabidopsis gene models	472,332	87.1
Map to predicted transcripts	470,988	86.9
Map to Arabidopsis ESTs (dbEST)	463,998	85.6
Map to genome, not to Arabidopsis ESTs	16,698	3.5
Map to genome, not to TAIR 7.0 models	9,687[a]	1.8

[a]This number is slightly different from row 2 minus row 3 due to the differences in BLAT scoring when counting a hit to the genome versus counting a hit to the gene models.

in any single organ (Schmid et al., 2005). This fact, together with the observation that a second sequencing run increased the number of genes identified by only about 10%, suggests that two pyrosequencing runs detect at least 90% of all genes expressed in this sample. Because pyrosequencing recovered the expected range of ESTs and the second run approached saturation, one or two pyrosequencing runs provide very deep and nearly complete representation of transcripts expressed by Arabidopsis seedlings, including detection of genes with very low expression.

Pyrosequencing ESTs Represent the Full Length of Transcripts

Preparation of DNA for pyrosequencing involves random shearing of the DNA by nebulization to provide short fragments suitable for sequencing. The randomness of this shearing process for cDNA has not been adequately assessed. If some cDNAs were resistant to shear forces due to their size, less complete coverage of the sequence might occur. We therefore asked whether there was bias in the regions of the transcript that were represented by pyrosequencing ESTs or in the length of the transcripts that were represented. cDNAs were analyzed based on their expression level and on their length. Mapping of the pyrosequencing ESTs to their corresponding full-length transcripts (TAIR 7.0 cDNA models) indicated that all regions of the transcripts were represented by the ESTs. There appears to be a slight strand bias with 55% to 60% of reads coming from the plus (same as mRNA) strand. We compared EST distributions representing short (<1,000 nt), medium (1,000–2,000 nt), and long (>2,000 nt) transcripts. An example that compiles 154,379 ESTs corresponding to 1,053 transcripts of 1,000 to 2,000 nt in length is shown in Figure 1. We also examined the distribution of ESTs along the length of transcripts that were highly expressed (615–1,949 ESTs per cDNA), moderately expressed (100–113 ESTs per cDNA), and minimally expressed (10 ESTs per cDNA; Supplemental Fig. S2). Although ESTs mapping to the 5′ end were in most cases more abundant than other regions, no other substantial bias of ESTs across different regions of the transcripts was observed. For short transcripts, there was a slight bias toward higher representation of the middle of the transcript (Supplemental Fig. S2). This suggests that breakage of shorter cDNA sequences near the middle is favored. Nevertheless, the bias toward the middle is not large and we conclude that other methods of cDNA preparation, such as random priming, would not substantially improve full coverage of transcripts. The observation of ESTs initiating from every percentile of the cDNAs, regardless of cDNA length or expression level, indicates that pyrosequencing is capable of reconstructing complete cDNA sequences.

Comparison to Conventional Arabidopsis ESTs

GenBank currently holds 734,275 Arabidopsis conventional ESTs (i.e. randomly picked cDNA clones sequenced by Sanger chemistry) that comprise a total of 325 million raw nucleotides of sequence. Of these ESTs, 691,589 (94.2%) had at least one significant alignment to the Arabidopsis genome. Taken together, all Arabidopsis dbEST ESTs covered 36,466,121 nt of the genome. Of the pyrosequencing ESTs that could be mapped to the Arabidopsis genome, 96.5% matched at least one Arabidopsis EST in GenBank dbEST. Over 16,000 of the ESTs that match the genome did not match sequences in the existing dbEST database and thus represent novel transcript sequences identified in this study. For these 16,698 ESTs, 13,701 matched a cDNA model and these represented 5,302 gene loci; 648 of these loci have no matching EST in dbEST and thus pyrosequencing provided new evidence that these genes are actively transcribed. For the remaining 4,654 loci, our ESTs provide coverage to portions of the models not represented in dbEST. As described below, it is likely that some of the 648 loci detected by pyrosequencing, but not in dbEST, represent difficult-to-clone sequences or DNA molecules that are toxic or otherwise unstable in *E. coli*.

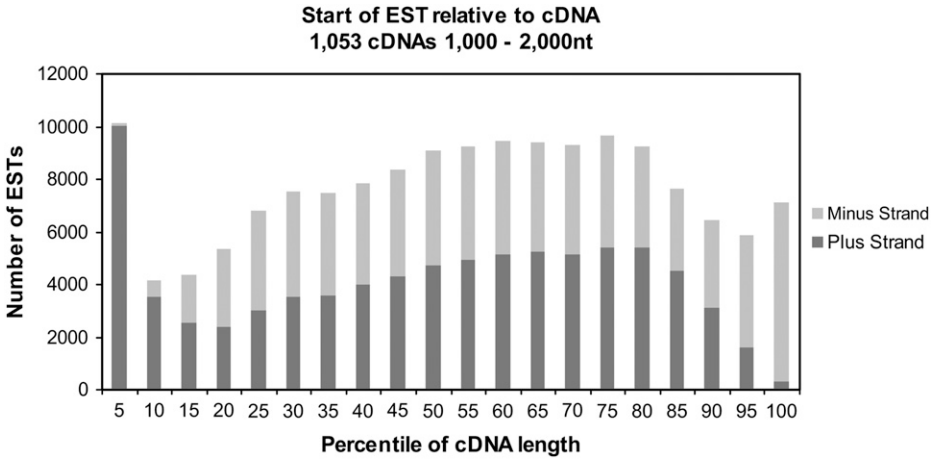

Figure 1. Pyrosequencing ESTs represent the full length of transcripts. The start (5′) position for pyrosequencing ESTs for a selection of TAIR 7.0 cDNA models is shown. The position relative to the 5′ end is expressed as a percentile of the length of the cDNA model to which it mapped; 1,053 moderately to highly expressed cDNAs (46–993 ESTs per cDNA) of 1,000–2,000 nt were selected and subdivided into 20 percentiles based on predicted cDNA length. The histogram indicates how many plus-strand (same as mRNA) and minus-strand pyrosequencing ESTs mapped to each portion of the full-length cDNA.

Two pyrosequencing runs provided sequences representing 10,280,356 nt of the Arabidopsis genome. As expected, due to their greater length and representation of multiple tissues, the 734,275 Sanger sequencing ESTs provided greater (approximately 3.5-fold) unique sequence coverage than the pyrosequencing ESTs. In addition, 23,367 Arabidopsis genes (28,301 cDNA models) were identified by all ESTs in GenBank. This compares to 17,449 genes unambiguously identified by the two pyrosequencing runs reported here. The larger number of loci represented by the dbEST dataset can largely be explained by the sampling of almost all Arabidopsis tissues.

To compare the efficiency of gene discovery by pyrosequencing to traditional EST approaches, we randomly selected five sets of 10,000 ESTs from the 734,725 ESTs in GenBank and examined how many unique loci were identified and how much genome sequence was covered by these ESTs. This number was chosen because the cost for sequencing of 10,000 ESTs is approximately equivalent to two consecutive pyrosequencing runs. On average, 10,000 randomly selected ESTs covered approximately 3,000,000 nonredundant nucleotides of genome sequence and identified 5,540 unique loci. In comparison, a single pyrosequencing run identified 3 times as many genes and covered twice as much sequence.

Representation of Chloroplast and Mitochondrial Transcripts

For 38 annotated mitochondrial open reading frames (ORFs), at least one pyrosequencing hit was detected and, with few exceptions, for most of these multiple Sanger ESTs also exist. For 71 chloroplast ORFs, we found at least one pyrosequencing EST. Similar to mitochondrial ORFs, most chloroplast transcripts detected by pyrosequencing have previously been tagged by Sanger ESTs. Only a few pyrosequencing ESTs mapped to chloroplast or mitochondrial ribosomal RNAs (209 and 48, respectively), which indicates efficient removal of ribosomal RNA during oligo(dT) purification of mRNA.

Pyrosequencing Provides Evidence for Novel Transcripts and Transcript Architecture

There were 9,687 ESTs that matched the genome, but did not match a predicted gene in TAIR 7.0. Using BLASTX, these ESTs were searched against both the RefSeq protein database and the NCBI nonredundant protein database; 278 had significant protein matches against the RefSeq and 545 had matches against the nonredundant (Supplemental Table S1B) database. After correction for those ESTs that aligned to more than one place on the genome, and multiple overlapping or adjacent ESTs, we identified approximately 60 locations in the genome that are represented by expressed sequences and are likely protein coding (based on hits to protein databases), but that were not annotated as genes in TAIR 7.0 (Supplemental Table S2). Because small peptides are underrepresented in protein databases (Lease and Walker, 2006) and because many pyrosequencing ESTs represent noncoding 5′- and 3′-untranslated regions, the tables likely underestimate the number of unannotated proteins expressed in Arabidopsis seedlings.

A specific example is shown in Figure 2. One hundred pyrosequencing ESTs and a number of Sanger ESTs map to the intergenic region between genes At1g65420 and At1g65430, a region of chromosome 1 that is not currently annotated as a gene. The transcribed sequence is 814 nt long and contains several small ORFs encoding short polypeptides of 52 and 42 amino acids in length. A BLAST search of the transcribed sequence against GenBank did not retrieve significantly similar genes in organisms other than Arabidopsis. Interestingly, this short sequence is duplicated, occurring also between genes At4g34880 and At4g34890 on chromosome 4. This transcribed region is likely not currently annotated as a gene because previous efforts in Arabidopsis genome annotation have focused on protein-coding genes with a minimum ORF length (E. Huala and D. Swarbreck, personal communication). Hence, putative genes encoding small proteins might be underrepresented in the current gene models. It is also possible that this gene encodes a long noncoding RNA of unknown function. Along the same lines, transcripts encoding a 19-kD thylakoid lumenal protein could be mapped to the extreme proximal end of chromosome 3, although no gene model is annotated in this region (http://genomics.msu.edu/cgi-bin/gbrowse/A_thaliana/?name=CHR3v01212004:23470120.23470555). This gene is also strongly supported by multiple Sanger ESTs and pyrosequencing ESTs, and by two NCBI database entries identical with the ORF derived from the ESTs (P82658, BAF019999). This putative gene thus represents a candidate for inclusion in a future version of the TAIR dataset. Further support for this gene comes from the fact that a related gene (Os08g0504500) is annotated in the genome of rice (*Oryza sativa*), encoding a protein that is 66% identical to its Arabidopsis ortholog. No paralogs were found in the Arabidopsis genome, indicating it represents a single-copy gene.

A possible concern with pyrosequencing is contamination of cDNA with genomic DNA and hence the possibility that genomic DNA fragments are wrongly identified as transcribed sequences. However, Figure 3B shows that pyrosequencing ESTs mapping to At3g54830 are clearly reflecting (and thus verifying) the predicted exon-intron structure of this gene; hence, they do represent processed transcripts, not genomic DNA. Visual examination (GBrowse) of ESTs mapping to over 100 genes supported this conclusion.

A specific example of novel transcript information is At3g11090, which is annotated as a LOB-domain family protein. To date, no ESTs mapping to this gene have been identified. However, 17 pyrosequencing ESTs unambiguously map to this locus, indicating this is

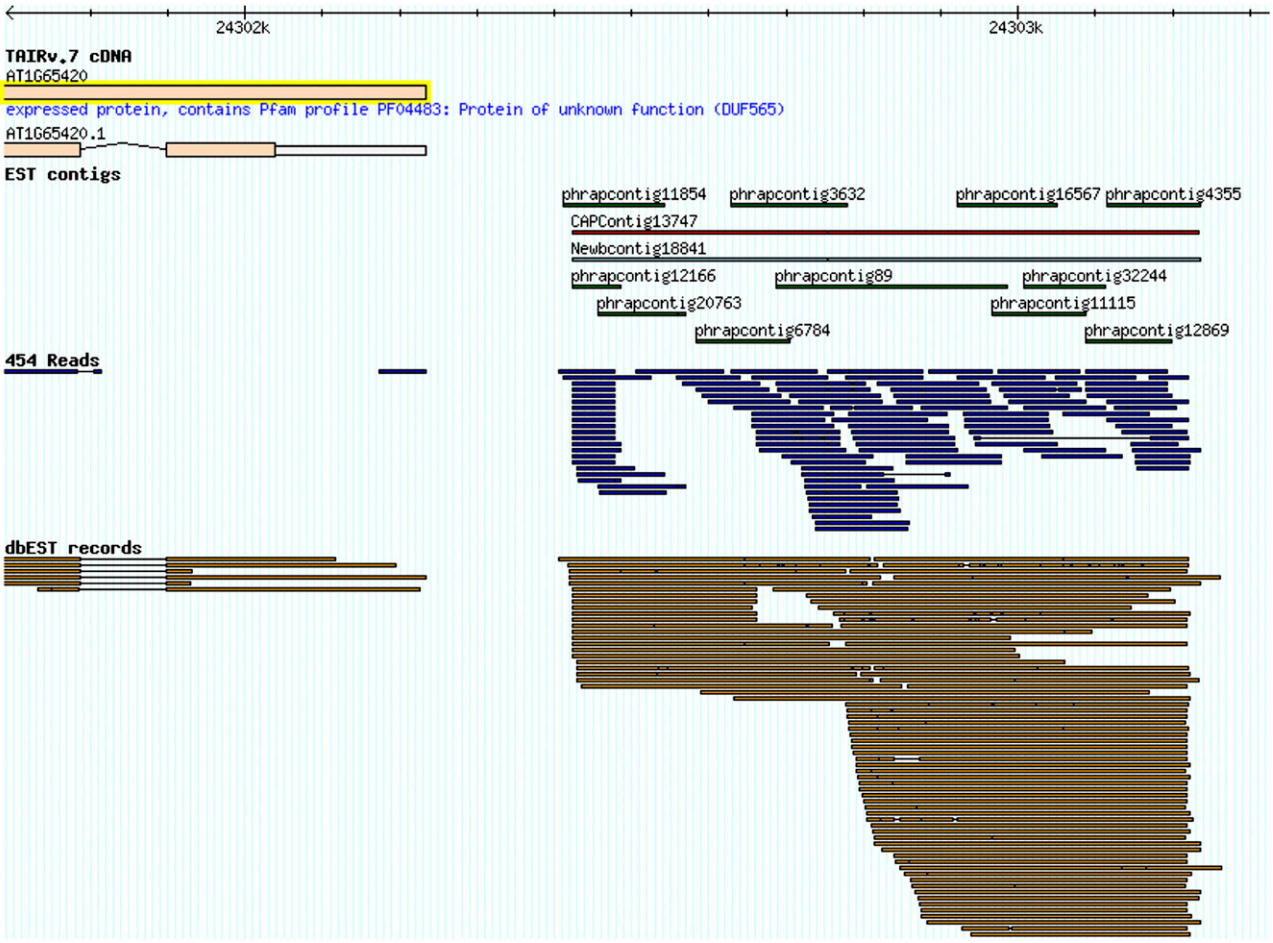

Figure 2. cDNA pyrosequencing corroborates EST evidence for an expressed gene downstream of At1g65420. Whereas multiple pyrosequencing ESTs and Sanger ESTs map to this region of chromosome 1 of the Arabidopsis genome, currently no gene model exists for this locus.

indeed an expressed gene (Fig. 3A). Interestingly, nine unique 17-bp signature sequences mapping to this gene have been previously retrieved by the Arabidopsis massively parallel signature sequencing (MPSS) plus (Meyers et al., 2004; Nakano et al., 2006) project (http://mpss.udel.edu/at/GeneAnalysis.php?featureName=AT3G11090), and a basic expression level was detected by microarray analysis (Schmid et al., 2005). Given that expression of At3g11090 was detected by three independent experimental approaches, it is surprising that no ESTs for this gene were previously found. Because all three experimental approaches do not require cloning, we posit that At3g11090 may be toxic or otherwise incompatible with cloning in *E. coli* and therefore not represented in the EST collection. Another specific example is the putative amino acid transporter At3g54830 (Fig. 3B). Also, in this case no ESTs mapping to this gene could be identified in dbEST. However, 22 pyrosequencing ESTs tagging this gene could be identified. Interestingly, 15 unique signature sequences for this gene were retrieved by MPSS (http://mpss.udel.edu/at/GeneAnalysis.php?featureName=AT3G54830), corroborating the pyrosequencing data and suggesting that this gene might be incompatible with cloning in *E. coli*.

More frequently than novel genes or genes lacking ESTs in dbEST, we detected truncated gene models that lack parts of their 5′ and/or 3′ regions. For example, pyrosequencing ESTs EB3RODY02I8QOG and EBENXNS01CGGFY map to a region on chromosome 1 upstream of gene At1g01790 that does not contain annotated gene models or sequences mapping to Sanger ESTs. At1g01790 encodes the putative potassium efflux transporter KEA1 (Maser et al., 2001). Interestingly, pyrosequencing EST EBENXNS01CGGFY, when compared to GenBank, maps to the rice gene Os04g58620, also encoding a putative potassium transporter that is closely related to Arabidopsis KEA1 (85% amino acid similarity). Thus, the Arabidopsis gene At1g01790 might require extension at the 5′ end, resulting in a significantly longer protein than currently annotated. Pyrosequencing ESTs thus can assist in identifying the correct 5′ end of the transcript.

Both Sanger ESTs and pyrosequencing ESTs map proximal and distal of the annotated gene model At5g66052, indicating that the current gene model

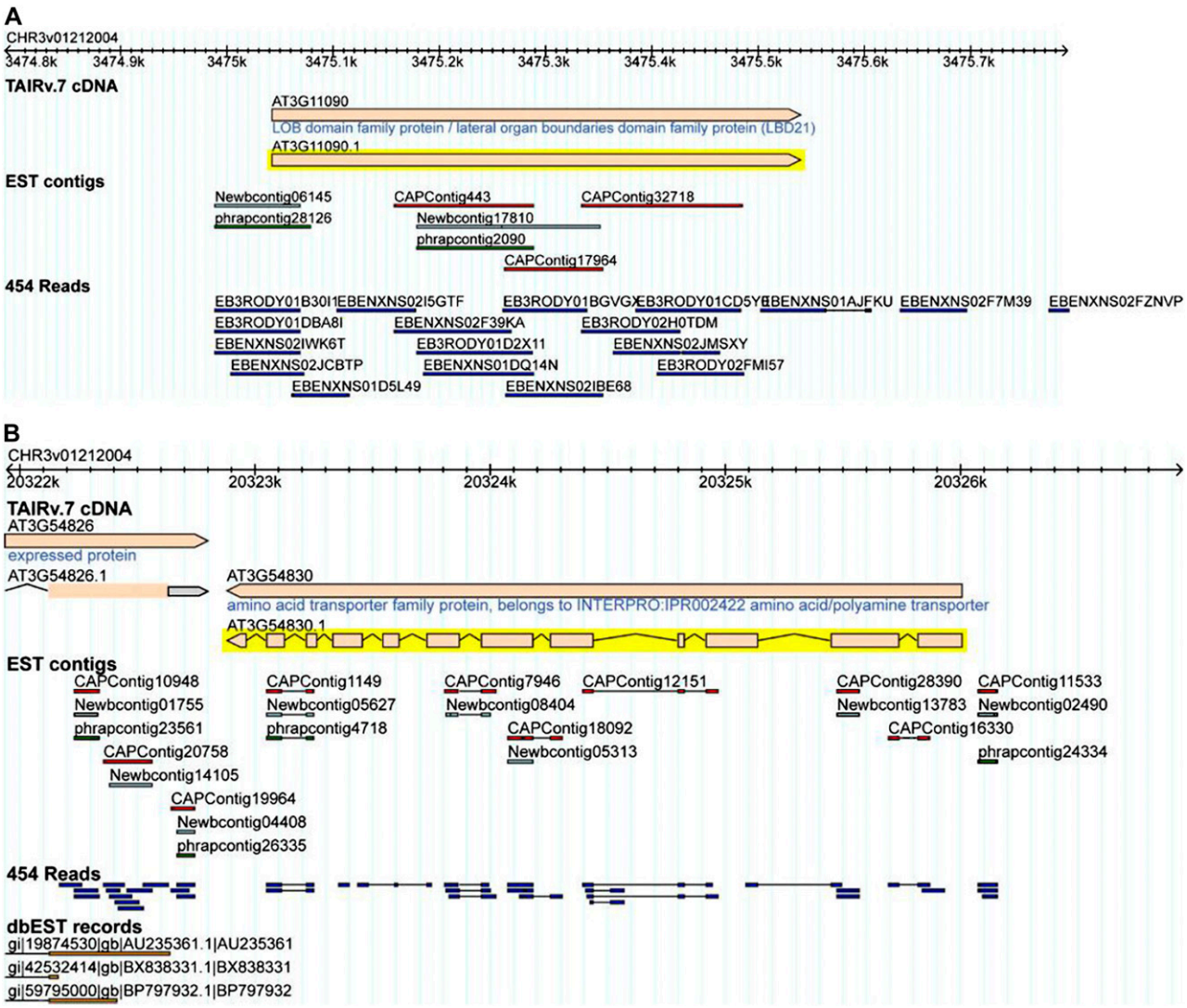

Figure 3. Pyrosequencing of Arabidopsis cDNAs reveals expression of genes for which no Sanger ESTs are available. A, Evidence for expression of the LOB domain family protein At3g11090. Seventeen pyrosequencing ESTs unambiguously mapped to this locus (pyrosequencing reads, EST contigs), whereas no Sanger ESTs mapping to this gene could be identified in dbEST. B, cDNA pyrosequencing provides evidence for expression of the amino acid family transporter protein At3g54830, whereas no dbEST records mapping to this gene could be found. The fact that pyrosequencing ESTs align with the predicted exon-intron structure of At3g54830 rules out that these ESTs represent contamination of cDNA with genomic DNA.

may not accurately reflect the transcribed region of this gene and requires extension at the 5′ and 3′ ends (Fig. 4). In another example, pyrosequencing ESTs EBENXNS01DS6VU, EBENXNS02G7LSB, and EBENXNS02IE294 all show significant homologies with phox domain-containing proteins and they map downstream of At1g15240 (http://genomics.msu.edu/cgi-bin/gbrowse/A_thaliana/?name=CHR1v01212004:5243200..5248700). In this case, it is possible that gene model At1g15240 is incomplete and should be extended to include the region tagged by pyrosequencing ESTs. However, the At1g15240 model is based on cDNA AK176485 and this cDNA appears to have a poly(A) tail indicating the poly(A) site to be where the current 3′ end of the gene model is presently annotated (D. Swarbreck, personal communication). Because genes can have more than one poly(A) site, pyrosequencing ESTs may indicate an additional, alternative downstream poly(A) site.

Application of Pyrosequencing to Analysis of Gene Expression: Digital Northerns

Comparisons of the number of ESTs for a gene between different libraries or different genes in the same library can be a reliable indicator of relative gene expression provided the ESTs map unambiguously to a single gene location (Audic and Claverie, 1997). Over 90% of the pyrosequencing ESTs from this study matched this criterion. The statistical significance of

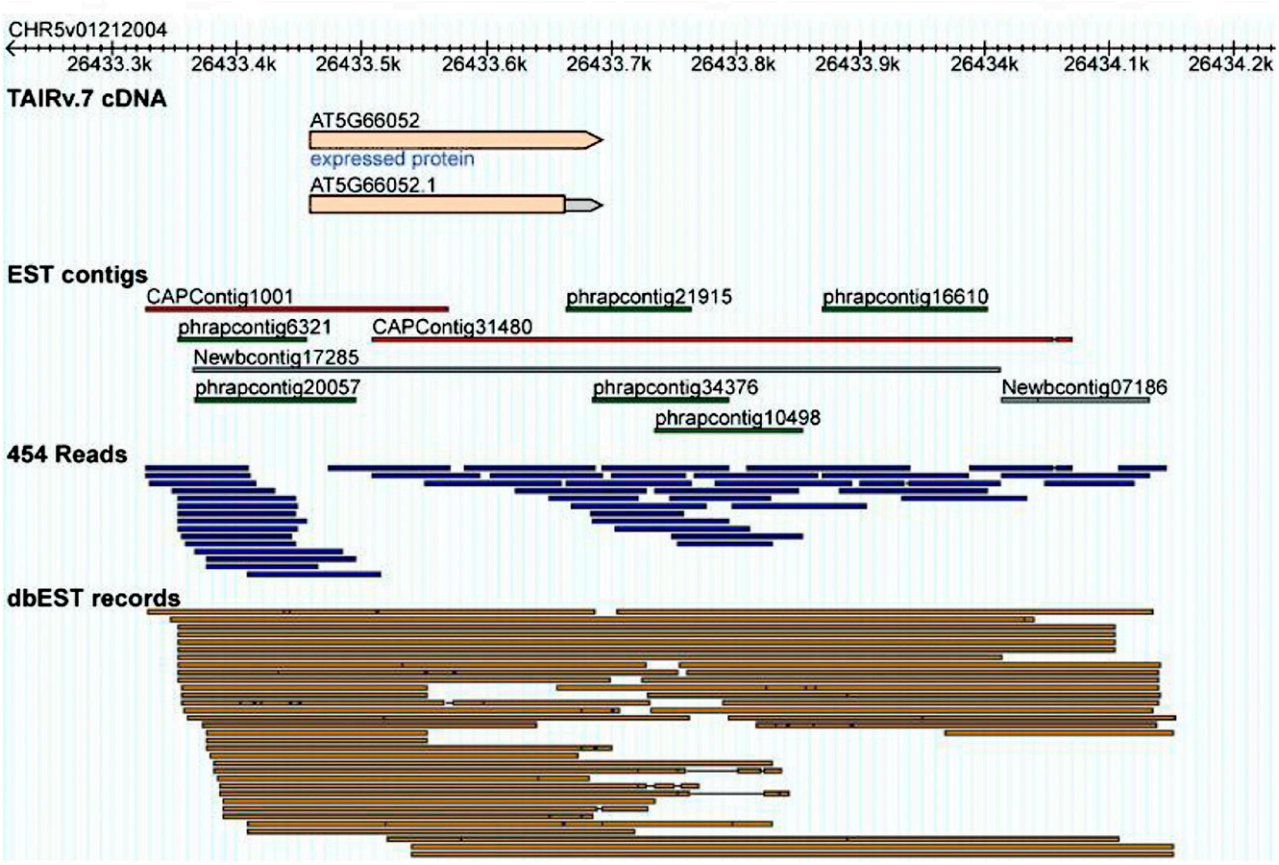

Figure 4. Evidence from both Sanger ESTs and pyrosequencing ESTs suggests revisions for model At5g66052. A large number of pyrosequencing ESTs and Sanger ESTs form contigs that map proximal and distal of gene model At5g66052, suggesting that the current gene model does not completely reflect the transcribed sequence.

a comparison of abundance of transcripts depends only on the number of ESTs (Audic and Claverie, 1997). Similar to Serial Analysis of Gene Expression (SAGE), pyrosequencing provides a very large number of individual ESTs and therefore can provide robust statistical comparisons of gene expression levels based on the number of EST reads. For the young green seedlings analyzed in this study, the most abundant transcripts were, as expected, associated with photosynthesis. Over 28,000 ESTs mapped to the Rubisco gene model At1g67090.1 and the five genes for Rubisco together contributed more than 85,000 ESTs. There are approximately 20 genes encoding chlorophyll *a/b*-binding proteins that were represented by approximately 60,000 ESTs. Thus, these two gene families together contributed 26% of the ESTs that map to the Arabidopsis transcriptome. At the low abundance end of the spectrum, 2,941 gene models were represented by only one EST. The dynamic range between highly expressed and rare transcripts is thus over four orders of magnitude and extends to transcripts that represent less than 0.001% of the mRNA population. Of the 17,449 gene loci whose expression was detected, more than 10,000 were represented by at least three ESTs.

We also compared the number of ESTs per locus to the microarray signal obtained with ATH1 arrays for aerial tissues of seedlings grown under similar conditions (Schmid et al., 2005). The correlation coefficient for loci with more than 10 ESTs was 0.45. This correlation is similar to those observed in several other studies of SAGE versus microarray data (van Ruissen et al., 2005). The correlation coefficient did not increase when only genes with higher expression were compared, suggesting the lack of strong correlation is not due to the dynamic range, but is due to characteristic differences in the methods. The absolute microarray signal for each transcript depends on a number of factors, including intensity variation between probe sets and efficiency of PCR amplification. Thus, it is likely that when very large numbers of ESTs are available, absolute gene expression levels may be better represented by EST abundance than by microarray signal. In addition, current publicly available microarrays do not provide information for many Arabidopsis genes. For example, the widely used Affymetrix Arabidopsis expression array (ATH1) contains probe sets for approximately 22,700 (71%) of the 32,041 gene loci of TAIR 7.0. We compared the loci identified by the pyrosequencing ESTs to loci represented on the ATH1

arrays and found 1,410 loci identified by pyrosequencing, which are not represented on this array. Thus, an advantage of pyrosequencing is that it provides information on gene expression for a large number of genes not currently represented on commercially available Arabidopsis microarrays.

Assembly of Pyrosequencing ESTs

In this study, pyrosequencing ESTs were mapped to a completely sequenced genome and their value for sequence annotation, gene discovery, and transcript quantification is discussed. We also addressed the use of pyrosequencing for de novo sequencing of transcripts. To this end, the pyrosequencing ESTs were assembled into contigs using three different tools, the Newbler assembler provided with the GS20 sequencer, CAP3 (Huang and Madan, 1999), and the stackPACK EST analysis pipeline (Miller et al., 1999), which uses d2_cluster (Burke et al., 1999) to partition (cluster) the ESTs and Phrap (http://www.phrap.org) to assemble each cluster. Examples of the cluster results for several transcripts are shown in Figures 2 to 4. For all three methods, relatively few full-length cDNA sequences were reconstructed, even in cases where ESTs covering the entire predicted model were available. The d2_cluster uses a transitive clustering algorithm based on similarity (96% in the default) over a large window (default 100 nt). Whereas these parameters are appropriate for traditional ESTs, they fail to adequately cluster ESTs generated by pyrosequencing because the overlapping regions of adjacent ESTs were too small to meet the threshold score for clustering. Reducing the window size while increasing the similarity did not significantly improve clustering. CAP3 placed more ESTs in contigs than the other methods and created, on average, longer contigs than stackPACK, but still failed to produce full-length contigs in the majority of instances where full coverage was possible given the available EST data. The Newbler assembler utilized the fewest ESTs of the methods tested and created the fewest contigs; however, the average length of contigs assembled by Newbler was the longest. Newbler generated significantly fewer short contigs than CAP3 or stackPACK. Additional examples of the assembly results with the three programs can be explored with GBrowse. This comparison indicated that, although pyrosequencing is able to generate sufficient sequence data to completely represent the full length of many transcripts, the assembly programs we tested are unable to efficiently create full-length contigs.

DISCUSSION

The results presented above indicate that pyrosequencing provides a very rapid, low-cost survey of a plant tissue's transcriptome and the results are robust and unbiased. Massively parallel pyrosequencing offers several additional advantages compared to previous technologies. First, no biological cloning is required. Therefore, sequences that are difficult to clone or unstable or toxic in *E. coli* are not missed. Evidence that we identified such sequences is suggested by the examples in Figure 3, where transcripts are detected in our study, by microarrays and by MPSS, but not in the previous large dataset of Arabidopsis ESTs. Second, small transcripts that are often removed during size selection in cDNA library construction are not lost. Third, data can be obtained very rapidly. The time from tissue harvesting to completion of DNA sequencing can be as little as 1 week. Fourth, the cost of pyrosequencing (each EST costs less than $0.03) is substantially less than conventional EST sequencing. Although SAGE (Velculescu et al., 1995) and MPSS (Brenner et al., 2000) have in the past provided key information on transcripts, because of the much shorter sequences and other limitations of these techniques, it is likely their use will decline for profiling of transcriptomes.

A single pyrosequencing run identified most of the genes expressed in 8-d-old Arabidopsis seedlings. Although performing a second run increased the number of transcripts detected by only 10%, the total unique sequence information increased 50%. This occurred because the additional ESTs yielded more comprehensive sequence coverage across the length of transcripts, particularly for those transcripts of genes with low expression levels. An additional benefit of multiple runs is derived from the increase in statistical accuracy available when using EST numbers to make comparisons of relative gene expression levels.

For Arabidopsis, well-characterized and widely used microarrays are available that represent a large proportion of the expressed genes. The cost of a pyrosequencing run is severalfold higher than a microarray experiment and therefore pyrosequencing, in most cases, will not be the tool of choice for routine transcript analysis of Arabidopsis. However, pyrosequencing does have the advantage of providing data for the approximately 25% of Arabidopsis genes that are not currently represented or not accurately discriminated on available microarrays.

A recent study of Bainbridge et al. (2006) detected transcripts for 10,000 loci from a human prostate cancer cell line using a single pyrosequencing cycle that resulted in 181,279 ESTs (Bainbridge et al., 2006). This study also reported a bias toward representation of 5′ and 3′ ends and to the middle of transcripts. It was speculated that these biases resulted from the accessibility of ends to sequencing and from incomplete fragmentation of the cDNA. In our analyses, we observed a higher number of ESTs from the 5′ ends of all transcripts. In addition, the 3′ ends of long (>2,000 nt) transcripts were more highly represented as would be expected from cDNA synthesis primed with oligo(dT). However, plots of the distribution of ESTs across the length of the transcript indicated that all regions were well represented. Greater representation of the middle of transcripts was primarily notable for short cDNAs. There was only a slight strand bias with 55% to 60% of ESTs coming from the plus strand. We conclude that

fragmentation of the cDNAs during nebulization did not introduce major bias in the representation of transcripts. Therefore, other methods for preparation of cDNA libraries, such as random priming, do not seem to be needed to provide complete coverage across the length of transcripts.

Approximately 3.5% of the ESTs from our study that matched the Arabidopsis genome did not match ESTs already available in GenBank. In contrast, Emrich et al. (2007) recently reported that 30% of ESTs obtained by a single pyrosequencing run for maize (*Zea mays*) shoot apical meristem did not align to any of approximately 680,000 maize ESTs available in GenBank. This much higher proportion could reflect the specialized cell type that was sampled or perhaps the greater complexity of the maize genome.

Efficient reconstruction of longer sequence contigs from pyrosequencing ESTs requires a high degree of oversampling and unbiased representation of sequence fragments. This, in contrast to genomic sequencing, is inherently problematic with transcriptome sequencing because of the large dynamic range of gene expression levels that leads to massive redundancy for coverage of some highly expressed genes, whereas transcripts of genes with baseline expression levels are underrepresented. In our study, we found that 26% of all ESTs obtained from 8-d-old Arabidopsis seedlings were derived from only 25 highly expressed genes that are members of the Rubisco and light-harvesting complex gene families, whereas over 5,000 genes were represented by less than 10 ESTs. If priority is on gene discovery and assembly of longer contigs rather than on assessing relative gene expression, it will likely be useful to normalize the cDNA population prior to sequencing to maximize coverage of less abundant transcripts present in the sample. In this regard, Cheung et al. (2006) performed a single sequencing run on a normalized cDNA population derived from mixed tissues of *Medicago truncatula*. Their sequencing yielded 23 Mb of unique sequences, which is approximately twice the amount of unique sequence information we obtained (10.3 Mb) from two runs with a non-normalized library.

Our study also revealed that currently available software tools have problems with assembly of the very large numbers of short sequences provided by pyrosequencing. This was the case even for those abundant transcripts where thousands of ESTs could be aligned to provide essentially complete coverage. The inability to assemble contigs is thus in large part related to the short overlaps. Improvement in software is currently under development and will be particularly important for the application of pyrosequencing to transcripts from species without extensive genome information. The increase in sequence length to >200 nt expected from pyrosequencing instrument upgrades will also greatly facilitate assembly of full-length cDNA sequences.

The availability of very comprehensive data for the Arabidopsis genome and a large set of conventional ESTs provided a baseline for this evaluation of pyrosequencing data. A much greater advantage of pyrosequencing will be its application to EST sequencing for those species for which little or no genomic data are available. The ability to rapidly detect sequences for almost all genes expressed in a sample will provide a more comprehensive tool for gene discovery than conventional EST sequencing. For example, genes involved in natural product biosynthesis have frequently been discovered first by EST sequencing (e.g. Bao et al., 2002). The lower cost and greater sequence coverage afforded by pyrosequencing will make it possible to more confidently identify candidate genes involved in biosynthetic pathways and will allow identification of genes with very low expression levels often missed by conventional EST projects. Finally, as more plant genome sequences become available, mapping of pyrosequencing ESTs to these genomic sequences will provide a particularly efficient means for experimental verification of predicted gene models and can also be used to train ab initio gene prediction programs.

Applications to Proteomics

Currently, proteomic analysis of organisms lacking a fully sequenced genome is difficult. This is due to the way modern proteomics data are analyzed using uninterpreted spectral assignments. This approach calculates an ideal mass spectrum for each peptide in a database and compares such spectra against observed spectra. This approach is fast enough to allow for the analysis of the thousands of spectra collected for a typical complex protein sample and thus makes the procedure amenable to high-throughput analysis (Tabb et al., 2003; Hirano et al., 2004; van Wijk, 2004). The determination of the peptide sequence from the collected spectra (i.e. de novo sequencing) is generally considered too slow and error prone to be practical for large numbers of proteins (Baginsky and Gruissem, 2006; Pevtsov et al., 2006). Unfortunately, organisms and tissues that are very amenable to biochemistry and protein isolation are frequently not model species. The potential of, for example, peas (*Pisum sativum*) for organelle proteomics is underexplored because sequence information for pea is severely limited. Pyrosequencing technology allows researches to build custom sequence libraries for their organism and tissue of interest. Because the success of a proteomics project largely depends on the size and quality of the available sequence database, the lower cost and speed of obtaining such EST data using pyrosequencing will expand the number of organisms for which this condition can be met. For proteomics approaches, however, it will be important to obtain longer EST contigs assembled from multiple reads to minimize the rate of false-positive peptide identifications. To this end, either higher sequence coverage is required or supervised methods for contig assembly based on existing genome scaffolds need to be implemented.

MATERIALS AND METHODS

Preparation of RNA and cDNA of Arabidopsis Seedlings

Arabidopsis (*Arabidopsis thaliana* ecotype Columbia) seeds were sown on soil mix, placed at 4°C for 2 d, and then germinated under continuous light (approximately 150 μmol s^{-1} m^{-2}) at 20°C. After 8 d, the aboveground green tissue was harvested and immediately frozen in liquid nitrogen. Total RNA was extracted by grinding the frozen tissue with a mortar and pestle in the single-step acid guanidinium thiocyanate-phenol-chloroform mixture as described by Chomczynski and Sacchi (1987), followed by two consecutive washes of the RNA pellet with 3 M sodium acetate (pH 6.0), as described by Logemann et al. (1987), to remove polysaccharides. Total RNA was checked for purity and degradation using the Agilent 2100 Bioanalyzer RNA chip (Agilent Technologies) and stored in 80% ethanol.

mRNA was purified using the Illustra mRNA purification kit (GE Healthcare). One milligram of total RNA was redissolved in Tris-EDTA buffer and applied to a pre-equilibrated oligo(dT) cellulose column. Poly(A)$^+$ RNA was eluted from the column and applied to a second column for another round of purification. After elution from the column, poly(A)$^+$ RNA was stored as ethanol precipitate.

cDNA was synthesized using the CLONTECH Smart PCR cDNA synthesis kit. First-strand cDNA synthesis was performed with oligo(dT) primer in a total volume of 10 μL as described in the provided protocol using 1 μg mRNA. Double-strand cDNA was prepared from 2 μL of the first-strand reaction by PCR (13 cycles) with provided primers in a 100-μL reaction. cDNA was purified using Qiagen QIAquick PCR purification spin columns and was checked for purity and degradation using the Agilent 2100 Bioanalyzer DNA chip.

DNA Sequencing and Bioinformatics

Approximately 3 μg of the final adaptor-ligated cDNA population was sheared by nebulization and DNA sequencing was preformed at the Michigan State Research Technology Support Facility following protocols for the Genome Sequencer GS20 System (Roche Diagnostic). Reads generated by the GS20 sequencer were trimmed of low quality, low complexity [e.g. poly(A)] and vector sequences using the The Institute for Genomic Research (TIGR) SeqClean software pipeline. This tool set is currently available from the Gene Index Project (http://compbio.dfci.harvarard.edu/tgi/software). After trimming, 541,852 reads remained with mean and median lengths of 89.2 and 95 nt, respectively. Alignment of these ESTs to the Arabidopsis genome (01222004 version) or predicted gene models (TAIR 7.0, courtesy of E. Huala; release date March 2007) was performed with BLAT (Kent, 2002). Stringent parameters were used for the BLAT alignments; 95% sequence identity over at least 90% of the EST length was required to assign a match. Translated BLAST searches (BLASTX) against the NCBI nonredundant and RefSeq protein databases (http://www.ncbi.nlm.nih.gov/RefSeq) were performed with the parallel BLAST implementation, mpiBLAST (Darling et al., 2003). The e-value cutoff was set at 1×10^{-10}.

Gene model and EST mapping data were displayed with GBrowse developed by Lincoln Stein (2002) and are available at: http://genomics.msu.edu/cgi-bin/gbrowse/A_thaliana.

EST sequence accession numbers in GenBank are EH795234 through EH995233 and EL000001 through EL341852.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Number of pyrosequencing ESTs versus number of gene loci.

Supplemental Figure S2. Distribution of ESTs across length of cDNAs.

Supplemental Table S1. Pyrosequencing results linked to TAIR 7.0 gene models.

Supplemental Table S2. Pyrosequencing evidence for protein-coding regions not included in TAIR 7.0 gene models.

ACKNOWLEDGMENTS

We thank Shari Tjugum-Holland and Jeff Landgraff of the Michigan State University Research Technology Support Facility for assistance with RNA and DNA analysis and DNA sequencing. We greatly appreciate Eva Huala and David Swarbreck of The Arabidopsis Information Resource for providing TAIR 7.0 datasets and for helpful discussions. We thank Andrea Brautigam and Fred Beisson for comments on the manuscript.

Received January 26, 2007; accepted February 27, 2007; published March 9, 2007.

LITERATURE CITED

Ahmadian A, Ehn M, Hober S (2006) Pyrosequencing: history, biochemistry and future. Clin Chim Acta **363:** 83–94

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res **7:** 986–995

Baginsky S, Gruissem W (2006) Arabidopsis thaliana proteomics: from proteome to genome. J Exp Bot **57:** 1485–1491

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, et al (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics **7:** 246

Bao X, Katz S, Pollard M, Ohlrogge J (2002) Carbocyclic fatty acids in plants: biochemical and molecular genetic characterization of cyclopropane fatty acid synthesis of Sterculiafoetida. Proc Natl Acad Sci USA **99:** 7172–7177

Barbier G, Oesterhelt C, Larson MD, Halgren RG, Wilkerson C, Garavito RM, Benning C, Weber APM (2005) Genome analysis. Comparative genomics of two closely related unicellular thermo-acidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic flexibility of *Galdieria* and significant differences in carbohydrate metabolism of both algae. Plant Physiol **137:** 460–474

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, 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

Burke J, Davison D, Hide W (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res **9:** 1135–1142

Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, Town CD (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. BMC Genomics **7:** 272

Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem **162:** 156–159

Darling A, Carey L, Feng W (2003) The design, implementation, and evaluation of mpiBLAST. *In* Proceedings of ClusterWorld 2003. Linux Clusters Institute. http://public.lanl.gov/radiant/pubs/bio/cwce03.pdf

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res **17:** 69–73

Hirano H, Islam N, Kawasaki H (2004) Technical aspects of functional proteomics in plants. Phytochemistry **65:** 1487–1498

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

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res **12:** 656–664

Lease KA, Walker JC (2006) The Arabidopsis unannotated secreted peptide database, a resource for plant peptidomics. Plant Physiol **142:** 831–838

Logemann J, Schell J, Willmitzer L (1987) Improved method for the isolation of RNA from plant tissues. Anal Biochem **163:** 16–20

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:** 376–380

Maser P, Thomine S, Schroeder JI, Ward JM, Hirschi K, Sze H, Talke IN, Amtmann A, Maathuis FJ, Sanders D, et al (2001) Phylogenetic relationships within cation transporter families of Arabidopsis. Plant Physiol **126:** 1646–1667

Mayer KM, McCorkle SR, Shanklin J (2005) Linking enzyme sequence to function using Conserved Property Difference Locator to identify and annotate positions likely to control specific functionality. BMC Bioinformatics **6:** 284

Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD (2004) Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol **135:** 801–813

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res **9:** 1143–1155

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res **34:** D731–735

Ohlrogge J, Benning C (2000) Unraveling plant metabolism by EST analysis. Curr Opin Plant Biol **3:** 224–228

Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X (2006) Performance evaluation of existing de novo sequencing algorithms. J Proteome Res **5:** 3018–3028

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, et al (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science **311:** 392–394

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA **74:** 5463–5467

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet **37:** 501–506

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al (2002) The generic genome browser: a building block for a model organism system database. Genome Res **12:** 1599–1610

Tabb DL, Saraf A, Yates JR (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem **75:** 6415–6421

van Ruissen F, Ruijter JM, Schaaf GJ, Asgharnegad L, Zwijnenburg DA, Kool M, Baas F (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. BMC Genomics **6:** 91

van Wijk KJ (2004) Plastid proteomics. Plant Physiol Biochem **42:** 963–977

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science **270:** 484–487

Weber APM, Oesterhelt C, Gross W, Bräutigam A, Imboden LA, Krassovskaya I, Linka N, Truchina J, Schneidereit J, Voll H, et al (2004) EST-analysis of the thermo-acidophilic red microalga *Galdieria sulphuraria* reveals potential for lipid A biosynthesis and unveils the pathway of carbon export from rhodoplasts. Plant Mol Biol **55:** 17–32