

The Maize Root Transcriptome by Serial Analysis of Gene Expression^{1[w]}

V. Poroyko, L.G. Hejlek, W.G. Spollen, G.K. Springer, H.T. Nguyen, R.E. Sharp, and H.J. Bohnert*

Department of Plant Biology (V.P., H.J.B.) and Department of Crop Sciences (H.J.B.), University of Illinois, Urbana, Illinois 61801; and Division of Plant Sciences, Plant Sciences Unit (L.G.H., W.G.S., H.T.N., R.E.S.), and Department of Computer Science (G.K.S.), University of Missouri, Columbia, Missouri 65211-7145

Serial Analysis of Gene Expression was used to define number and relative abundance of transcripts in the root tip of well-watered maize seedlings (*Zea mays* cv FR697). In total, 161,320 tags represented a minimum of 14,850 genes, based on at least two tags detected per transcript. The root transcriptome has been sampled to an estimated copy number of approximately five transcripts per cell. An extrapolation from the data and testing of single-tag identifiers by reverse transcription-PCR indicated that the maize root transcriptome should amount to at least 22,000 expressed genes. Frequency ranged from low copy number (2–5, 68.8%) to highly abundant transcripts (100→1,200; 1%). Quantitative reverse transcription-PCR for selected transcripts indicated high correlation with tag frequency. Computational analysis compared this set with known maize transcripts and other root transcriptome models. Among the 14,850 tags, 7,010 (47%) were found for which no maize cDNA or gene model existed. Comparing the maize root transcriptome with that in other plants indicated that highly expressed transcripts differed substantially; less than 5% of the most abundant transcripts were shared between maize and *Arabidopsis* (*Arabidopsis thaliana*). Transcript categories highlight functions of the maize root tip. Significant variation in abundance characterizes transcripts derived from isoforms of individual enzymes in biochemical pathways.

Serial Analysis of Gene Expression (SAGE) provides an accurate view of expressed genes in tissues or cells, the identification of transcripts, and also permits analyses that compare changes in transcript populations of organisms exposed to different conditions or between distantly related organisms (Velculescu et al., 1995; Tuteja and Tuteja, 2004). An efficient application of SAGE is the determination of transcript population structure and accurate quantity in organisms whose genomes have been sequenced, and the technique may be used to improve the annotation of genomes (Jansen and Gerstein, 2000; Saha et al., 2002; Stern et al., 2003). SAGE is based on the cloning of short segments, tags 10 to 14 bp in length, from all transcripts present at the time of RNA isolation. Further improvements by lengthening these sequences, Long-SAGE, has increased the validity of gene identification (Gowda et al., 2004; Wei et al., 2004). Segments (termed tags) of these transcripts are concatenated into inserts of clones that may include representative, identifiable portions of up to 70 genes (Velculescu et al., 1995; Tuteja and Tuteja, 2004; Wei et al., 2004). The SAGE library thus represents the true abundance profile of transcripts in

a tissue, which provides a significant advantage in time and expenses over, for example, the sequencing of random clones from cDNA libraries. The advantage of SAGE is also obvious in cases where results from normalized or subtracted cDNA libraries are available because these libraries compromise information on transcript abundance for the detection of novel transcripts. Through SAGE, the identity and abundance of tens of thousands of transcripts can be rapidly obtained and analyzed in high-throughput fashion by sequencing a few hundred to a few thousand clones. Whole genome expression profiles, relative abundance of transcripts, and dynamic transcript profiles have been obtained in animal systems (Velculescu et al., 1997; Boon et al., 2002; Liang, 2002; Ding and Cantor, 2004). In some applications, SAGE-based transcript profiling has replaced microarray- or genechip-based approaches (Evans et al., 2002).

In contrast to its application in animal systems, where millions of tags have been recorded, SAGE has only recently received attention in plant research. SAGE collections are available for rice seedlings (*Oryza sativa*; Matsumura et al., 1999), for loblolly pine (*Pinus taeda*) transcripts present in the trunk (Lorenz and Dean, 2002), for *Arabidopsis* (*Arabidopsis thaliana* ecotype Col-0) root tissue (Fizames et al., 2004) and *Arabidopsis* roots after 2,4/6-trinitrotoluene (TNT) treatment (Ekman et al., 2003), and for mature leaf and immature seed tissue of rice (subsp. japonica cv Nipponbare; Gibbings et al., 2003). Collections of SAGE tags and unigenes have also been reported for *Arabidopsis* pollen and plants that had been exposed to cold stress. Differentially expressed genes have been identified by changes in tag abundance (Jung et al.,

¹ This work was supported by the National Science Foundation (grant nos. DBI-0223905 and DBI-0211842) and by University of Illinois Urbana-Champaign and University of Missouri institutional grants.

* Corresponding author; e-mail bohnert@life.uiuc.edu; fax 217-333-5574.

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

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.104.057638.

2003; Lee and Lee, 2003). Arabidopsis root tissue has also been analyzed in depth by genechips including approximately 22,000 genes, for >10,000 of which signals above background were recorded (Birnbaum et al., 2003). For those analyses of root tissues that sampled the SAGE population to a degree that might be considered close to a complete coverage, the number of transcripts ranged from approximately 15,000 to more than 20,000 transcripts.

More recently, Massively Parallel Signature Sequencing (MPSS) identifiers have been generated that support the analysis of transcriptome complexity (Meyers et al., 2004a, 2004b). The technique has been tested by relating MPSS data with the annotated transcriptome of Arabidopsis. This resulted in the identification of approximately 20,000 root transcripts, while 89% of the expressed signature abundances matched the Arabidopsis genome. In an analysis of Arabidopsis leaf transcripts by SAGE, Robinson et al. (2004) arrived at conclusions, similar to those presented by MPSS analyses, indicating that a number of tags will identify more than one transcript in partially duplicated plant genomes. Also, complexity in polyadenylation and alternative splicing are reported by SAGE analysis, indicating that the frequency of alternative forms of transcripts may be greater than presently appreciated. Both MPSS and SAGE provide powerful tools that improve transcript annotations.

We have used SAGE to determine the complexity of the maize (*Zea mays*) root transcriptome. RNA from the primary root of well-watered seedlings was converted into a SAGE library. In total, 161,320 SAGE tags were collected, resulting in the detection of 14,850 expressed genes. In a combination of methods for tag annotation, a set of virtual tags was extracted from maize expressed sequence tag (EST) collections by the V-SAGE algorithm (Poroyko et al., 2004), by BLAST-based searches against the National Center for Biotechnology Information (NCBI) maize UniGene set, and by comparisons with the Arabidopsis genome sequence. The information is used to support the annotation of the maize root transcriptome and to begin to place transcripts into functional categories and biochemical pathways.

RESULTS

A Maize Root SAGE Library

A SAGE library has been generated with RNA from well-watered primary roots of maize (line FR697) seedlings. In total, 3,652 clones were sequenced, from which 161,320 tags could be extracted (approximately 44 tags/clone). Tags recorded only once (25,749; 15.96%) were initially eliminated because they could represent sequencing errors, contamination, or cloning artifacts (but see below). Thus, 135,571 tags that appeared at least twice were accepted as statistically significant. The distribution of tags according to fre-

quency is shown in Table I, listing 14,850 tags that appeared multiple times with the tag of highest frequency recorded 1,233 times. The majority of the accepted unambiguous unitags, 10,222 in total (68.8%), were present at low copy number (less than five copies), 4,474 tags (30%) were counted between 6 and 99 times, and 154 tags (1%) were present in copy numbers 100 or higher.

Determination of the Maize Root Transcriptome Size

The presence of 14,850 unique tags indicated a minimum number of expressed genes, which does not represent the entire maize root transcriptome. When the number of unique new tags appearing at different periods during DNA sequencing was graphed against the total number of sequenced tags, an extrapolation indicated between 17,000 and 19,000 expressed genes (Fig. 1A). Application of an alternative method, a double-reciprocal plot of all unique tags of identified transcripts versus the total number of tags sequenced after each sequencing interval (Ekman et al., 2003), resulted in a number of approximately 20,000 transcripts present in maize roots (Fig. 1B).

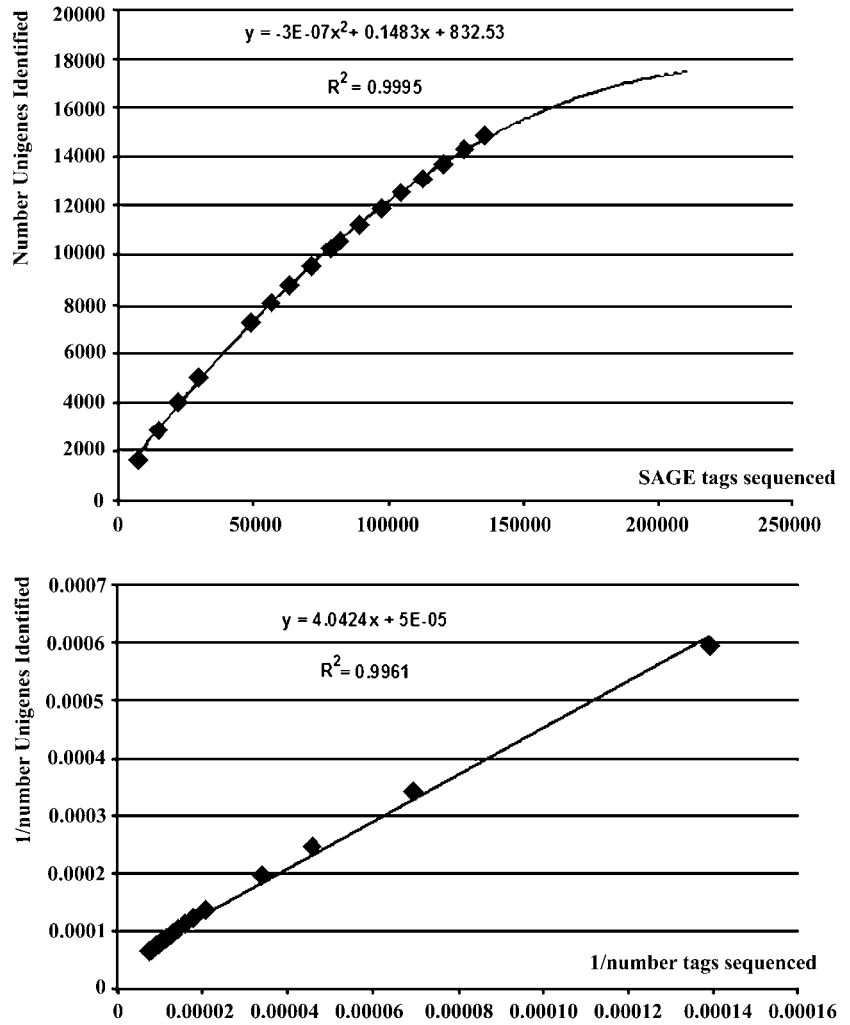
To test the possibility that single tags might represent genuine, rare transcripts, we used two strategies to analyze the 25,749 single and unidentified tags. When these orphan tags were compared to the V-SAGE collection of transcripts from the maize root, 3,072 single tags matched sequences in this collection. Alternatively, assuming that the single-tag collection might include single-nucleotide sequencing errors, we analyzed the 25,749 single tags again by a program that allowed one base to be variable. This converted 14,485, or 9% of the total number, of the single tags into tags that had been recorded before. A similar percentage, 8%, has been calculated as errors in sequencing in a SAGE analysis of the yeast transcriptome (Velculescu et al. 1997). This calculation left another 11,264 tags as orphans. These could, assuming

Table I. Distribution of SAGE tags sequenced from maize roots

Tag Copy No.	No. Different Tags
Total number of tags	161,320
Total number of different tags	40,599 ^a
Accepted tags (= transcripts)	14,850 ^b
1 ^c	25,749
2	5,317
3	2,478
4–5	2,427
6–10	2,228
11–20	1,272
21–99	974
>100	154

^aTags found only once are included. ^bOnly tags found more than once were used for BLAST searches and annotations. ^cFifty-five randomly selected one-copy tags were analyzed, 35 of which generated bands after RT-PCR amplification.

Figure 1. Extrapolating the number of transcripts present in the maize primary root. A, Plot of acquired unitags versus the total number of accepted tags recovered in successive rounds of DNA sequencing. B, Double-reciprocal plot of new tags discovery versus the total number of accepted tags.



70% as real transcripts (Chen et al., 2002a; see “Discussion”), represent additional transcripts that would increase the maize root transcriptome to at least 22,000 expressed genes.

Tag-to-Gene Assignment

To identify genes that corresponded to the 14,850 different unitags detected, virtual tags were extracted from known databases. As the reference sequence pool, 17,901 maize root cDNA sequences were available which had been sequenced from the 3'-end that included poly(A) tails. These sequences represented four maize cDNA libraries sequenced from the 3'-end that have been deposited in the NCBI database. The metadata for this collection of root cDNAs are available at <http://rootgenomics.missouri.edu>, and summarized in Supplemental Table I. All ESTs showed poly(A)⁺ structures from which the Perl-script V-SAGE (Poroyko et al., 2004) generated virtual SAGE tags. From each cDNA, all possible SAGE tags were recorded, progressing in a 3' to 5' direction. The 14,850

experimental tags were matched against this virtual tag list, which resulted in 53,559 tags from 14,391 maize ESTs that could be analyzed unequivocally. This identified 5,630 unitags, 2,547 of which matched the virtual tags from a library of well-watered maize roots, and 2,714 and 2,625 ESTs, respectively, from libraries representing maize roots water stressed for 5 or 48 h. In addition, 1,653 tags were present in a subtracted cDNA library.

A second approach used BLAST searches to identify exact matches of the SAGE tags in the NCBI database, maize UniGene Build number 40. This set consists of a set of 12,995 nonredundant maize unigenes. Searches by BLASTN of the 14-bp SAGE tags allowed matches to specific genes. Only matches with 100% identity (14 bp) to the mRNA-like strand were accepted. Strand orientation of sequences deposited in the maize UniGene set were determined by BLASTX in comparison with a database of 29,161 Arabidopsis putatively translated protein sequences and the NCBI nonredundant protein database. In total, 5,192 tags were identified for this collection of maize cDNAs.

By combining both approaches, 7,840 tags have been identified (Supplemental Table I): 2,976 tags in both reference databases, an additional 2,648 in the EST-based V-SAGE database, and 2,216 matched the UniGene set number 40 for maize deposited in NCBI. For tags that appeared in low abundance, the NCBI maize UniGene Build number 40 retrieved more hits than the V-SAGE collection.

For a number of low-complexity tags, for example CATGNAAAAAAAAA, multiple hits are inevitable. Such tags have been reported before for the Arabidopsis and loblolly pine transcriptomes (Lorenz and Dean, 2002; Fizames et al., 2004). Estimates for the probability of the appearance of an NlaIII recognition site within 3 bp of the poly(A)⁺ tail of a transcript is 1.3% for maize, and 1% for rice and Arabidopsis, while reports of rice SAGE profiles have excluded such tags (Matsumura et al., 1999; Gibbings et al., 2003). A similar condition is reflected in the maize dataset because annotations based on maize root ESTs as a reference database (<http://rootgenomics.missouri.edu>) showed 41 hits for the tag CATG(A)₁₀ in 341 copies, 42 hits CATGC(A)₉ in 255 copies, 29 hits CATGT(A)₉ in 167 copies, and eight hits for CATGG(A)₉ in 131 copies. These tags represent different transcripts, whose real abundance was determined by quantitative PCR specific for each potential target (Table II). The low-complexity tag G(A)₉ (131 copies) was used to amplify reverse transcription (RT)-PCR products, which, after DNA sequencing, revealed seven different transcripts, from a high copy number aquaporin (CF63347) to a rare transcript, a WRKY-type transcription factor (TF; CF634912).

Fifty-five SAGE tags that had appeared in only one copy were randomly selected to control for the possible presence of extremely rare transcripts. The sequences of these single tags were used in a procedure designed for the generation of long cDNA fragments from SAGE tags for gene identification (GLGI protocol; Chen et al., 2002b). Thirty-five of the 55 tags produced RT-PCR products, indicating that as much as 63% of the single-copy tags derived from genuine transcripts. As was also observed by Chen et al. (2002b), GLGI amplification occasionally generated more than one product for each primer pair, in each case with one predominant band (Fig. 2).

Validation of SAGE by Quantitative RT-PCR

Transcripts for validation by quantitative RT-PCR (qPCR) were chosen to represent tags that appeared with different frequencies to test redundancy reported by SAGE using an independent method. Primers for qPCR were chosen to contain the SAGE tag sequences anchored by a second primer to amplify a region approximately 115 bp upstream of the tag. A comparison of transcript abundance by SAGE and real-time PCR (Table II) revealed general correlation. Products from highly abundant SAGE tags appeared at the expected lower cycle numbers in the quanti-

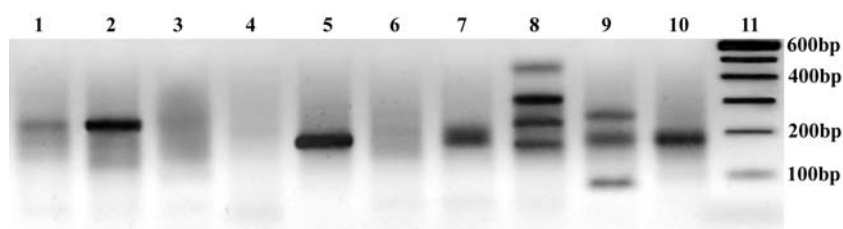
Table II. Transcript expression by SAGE and qPCR: a comparison

Low-Complexity Tags [Example: G(A) ₉] Target Gene	Accession	SAGE Tags	Ct ^a qPCR
Tonoplast membrane integral protein ZmTIP2-2 (maize)	CF633647	131	19.06 ± 0.05
Pyruvate dehydrogenase E1 α-subunit (maize)	CF627581		21.12 ± 0.11
S-Receptor kinase PK3 precursor-like protein (rice)	CF632211		22.79 ± 0.07
Putative vacuolar ATP synthase subunit G 1 (rice)	CF631447		25.14 ± 0.23
Expressed sequence Expressed protein (Arabidopsis)	CF629998 CF632710		27.45 ± 0.36 28.97 ± 0.06
WRKY11 (rice)	CF634912		30.08 ± 0.54
Copy No. Verification Target Gene	Accession	SAGE Tags	Ct qPCR
EST sequence without protein hit (maize)	CF636411	971	15.97 ± 0.09
β-Glucosidase, chloroplast precursor (maize)	CF637382	266	16.86 ± 0.08
Tubulin β-2 chain (<i>Eleusine indica</i>)	CF634356	107	17.42 ± 0.03
Putative ribosomal protein L18a, cytosolic (rice)	CK827623	75	18.64 ± 0.4
Putative disease resistance response protein (rice)	CF637046	15	18.69 ± 0.58
β-Expansin 4 (maize)	CF634719	2	21.87 ± 0.24

^aCycle threshold.

tative PCR analyses. For example, the tag for CF636411 (971 copies) appeared six cycles earlier than the tag for CF634719 (two copies), i.e. in both measurements a several hundred-fold difference was observed. In addition, quantitative real-time PCR allowed for an estimation of transcript abundance of the sampled tags. In addition, we compared the SAGE tag representation with data for 379 signals with very low intensity obtained by microarray hybridizations with RNA from the same tissue. The correlation between SAGE tag number and microarray intensity signal was 0.93. The results indicated that deviations were most common for transcripts with high %G+C (P. Li, V. Poroyko, and H.J. Bohnert, unpublished data). The data can be used to determine the depth to which the maize root transcriptome had at this stage been sampled. Assuming the total RNA amount for eukaryotic cells to be 13 pg (Okamura and Goldberg, 1989), two copies of a tag in the SAGE library represent approximately five transcripts per cell.

Figure 2. Amplification of sequences identified by single tags from root total RNA. Randomly selected tag sequences found only once generated products of different intensities (30 cycles; amplification according to Chen et al., 2002b). In total, 35 of 55 single tags showed amplification products (63.5%).



Comparison of the Arabidopsis and Maize Root Transcript Profiles

The SAGE profile of the Arabidopsis root reported 80 transcripts with tag copies exceeding 100 (up to 830 copies). These tags represented 12.9% of 144,083 tags. When sampling to the same depth of tags in maize (161,320 tags), 125 tag sequences with copy numbers ranging from 1,233 to 112 were observed. A comparison of root transcript frequencies in different species might provide an indication about similarity or divergence in root function (Table III). These abundant tags in maize and Arabidopsis showed little overlap in transcript identity or functional category. Only three transcripts in this high-abundance class were identical in both species. One transcript encoded a 40S ribosomal protein, S11 (maize tag no. 75; 153 copies) corresponding to Arabidopsis tag number 48 (158 copies). Second, a functionally unknown transcript (maize), which included the functional domain of a mitochondrial ADP/ATP carrier protein (maize tag no. 37; 231 copies), equivalent to Arabidopsis tag number 41 (165 copies; annotated as adenylate translocator). The third overlap was the 40S ribosomal protein S9 for maize tag number 19 (323 copies) and Arabidopsis tag number 70 (112 copies). When high abundance is disregarded, of the 80 most abundant tags/transcripts in Arabidopsis, 20 were also detected in maize, but most of the maize sequences were present at much lower abundance. The list (Table III) exemplifies instances where an Arabidopsis tag (GACTCTCTTA) identifies more than one gene (At3g45030 and At5g60390). Equally, cases are included where multiple tags for a particular single gene have been identified in maize, exemplifying the presence of alternatively spliced transcripts and variable 3'-end formation.

Table IV compares functional categories, according to COG (NCBI), for the most highly abundant Arabidopsis and maize transcripts detected by SAGE analyses. The juxtaposition of categories indicated general similarity, for example in transcripts in the categories ribosome biogenesis and translation, which included many of the most abundant transcripts, posttranslational modification, and ion transport. Also, transcripts for proteins in secondary metabolism were similarly numerous in both species. Differences among the groups of highly expressed transcripts existed between the two species in the categories RNA processing, chromatin structure, cytoskeleton, and energy production, which were more abundant in maize, and in the

categories defense mechanisms and cell wall biogenesis, which were more abundant in Arabidopsis. In part at least, these differences may be due to the fact that the maize SAGE profile sampled the apical 20 mm of the root while the Arabidopsis profile included the entire root. Also, the Arabidopsis roots were from 4- to 5-week-old rosette plants while in this experiment the primary root of young seedlings was harvested.

Transcripts for Biochemical Pathways, Transport Facilitators, and Gene Expression Control in Maize Roots

We present data on transcript complexity in the maize root for selected functional categories: TFs (Supplemental Table IIA), ion transporters and channels (Supplemental Table IIB), and several biochemical pathways (Fig. 3; Supplemental Table III, A–E) in a comparison with Arabidopsis models (The Arabidopsis Information Resource [TAIR], AraCyc: Arabidopsis Biochemical Pathways; <http://www.arabidopsis.org/biocyc/>). For example, multiple tags have been found representing transcripts for enzymes in all steps of the glycolysis pathway, with a significantly higher number of copies for two enzymes that are known to strongly influence the passage of metabolites through this pathway: Fru bisphosphate aldolase and glyceraldehyde phosphate dehydrogenase (Fig. 3A). Similarly, transcript abundance for enzymes in the oxidative pentose phosphate pathway (Fig. 3B), Suc metabolism (Fig. 3C), lignin biosynthesis (Fig. 3D), and sulfur assimilation (Fig. 3E) show different tag numbers for transcripts of individual enzymes in the respective pathways.

The analysis of tags revealed the expression of at least 44 TFs (Supplemental Table IIA). Many of the sequences encode proteins with a domain structure indicative of TFs, although their involvement in the regulation of gene expression, and in some cases their identity as TFs, has not been demonstrated. Some TFs that have been analyzed in other models are present, however. Unsurprisingly, components of the general transcription complex are present; e.g. subunits of the TFII complex or for the RNA-polymerase II complex are found. Others encode zinc-finger TFs. Included also are GATA-type, LIM, and several less studied factors in the bHLH and bZIP families and TFs, such as knotted or HBP-1a/b, that influence development and/or chromatin structure.

Table III. Comparison of most abundant *Arabidopsis* SAGE tags and their maize homologs

The most abundant SAGE tags identified in the *Arabidopsis* root (Ekman et al., 2003; Fizames et al., 2004) are compared with orthologous transcripts found in the SAGE population of the maize root.

Arabidopsis SAGE Tags	Copy No.	Abundance %	Arabidopsis Gene	Maize EST Homolog	Maize SAGE Tags ^a	Copy No.	Abundance %	Gene Description		
GACTCTCTTA	406	0.2818	At3g45030	CF634223	GGATCGGTTT	10	0.0061	40S ribosomal protein S20 translation elongation factor eEF-1 α -chain (gene 4)		
					At5g60390	CK347437	TGCACITTTA		221	0.1345
							GTCATCTGGA		96	0.0584
							AGGCAGACGG		38	0.0231
							AAAATCTTGC		8	0.0049
							TGCCCTCGTT		6	0.0037
							AACCACCTG		5	0.003
GTTAATGTT	367	0.2547	At5g02960	CF637692	CCAAGGGCAT	42	0.0256	40S ribosomal protein S23		
					GGGAAGACAC	2	0.0012			
AAAGTTGTAC	352	0.2443	At4g17340	CK347688	TTTTGTCCTG	3	0.0018	Membrane channel-like protein		
GGCGCGTAAG	319	0.2214	At2g33830	CF631448	TTTTGCCAGC	4	0.0024	Putative auxin-regulated protein		
GGTGGAGACA	293	0.2033	At4g14320	CK985703	AAATTGTCAA	12	0.0072	Ribosomal protein		
TGTGCCGTCA	236	0.1638	At1g26630	CF636407	AAGTTTCTTT	185	0.1126	Initiation factor 5A-4, putative		
					TCTCTCTACT	109	0.0663			
					TGAACCGTAC	55	0.0335			
GTACAGGAAG	230	0.1596	At1g33120	CF627397	TCAAGAAGAA	130	0.0791	60S ribosomal protein L9		
					At1g33140	CF627397	TGAAGAAGAA	63	0.0383	60S ribosomal protein L9
GCCCGAGGTA	226	0.1569	At3g05590	CF638015	GAAGGAAAGG	68	0.0414	60S ribosomal protein L18		
					CGCCGCCTTG	16	0.0097			
					TAAGGAAAGG	2	0.0012			
GGCATCATCT	205	0.1423	At2g39460	CF623519	TGTTATGCAC	71	0.0432	60S ribosomal protein L23A		
GGTGGTGACA	186	0.129	At3g23390	CK985703	ACCGTTTGCT	63	0.0378	ribosomal protein		
					AAATTGTCAA	12	0.0072			
TGTCTTGTT	178	0.1235	At2g27530	CF635737	TCTTTGTAAT	20	0.0122	60S ribosomal protein L10A		
					TTTGTTGTT	8	0.0049			
					GATGTTGAGG	3	0.0018			
TCTTAAGACT	171	0.1186	At1g31400	No hits						
				At5g02500	CF626370	GACGAGGACG	14	0.0084	dnaK-type molecular chaperone hsc70.1	
AGGAGGTGCC	165	0.1145	At3g08580	CF635264	TAGAGGGAGG	231	0.1431	Adenylate translocator		
GCTCAAGAGG	165	0.1145	At5g52650	CF633978	CAAAAAAAAAA	255	0.1552	Unknown protein		
					TGGCCTTTAG	90	0.0548			
CATTTTGATA	160	0.111	At1g30230	CF637384	ATGACCATCG	28	0.017	Elongation factor 1- β , putative		
AACGAGCTTG	159	0.1104	At1g66410	CF636639	GGACGATGAG	2	0.0012	Calmodulin-4		
					TGAACCTGTA	120	0.073			
					ATGGCCAAGT	26	0.0158			
					TGAAGTTGTA	10	0.0061			
CCTTTTGTTG	158	0.1096	At5g23740	CF636288	ATCAACGAGG	3	0.0018	40S ribosomal protein S11		
					TTACTACTTA	153	0.0948			
GGAGCTGGCG	137	0.0951	At1g07920	CF634765	AGGCAGACGG	38	0.0231	Elongation factor 1- α		
ATCTTCTGTT	130	0.0902	At1g07930	CF634765	AGGCAGACGG			Ubiquitin-conjugating enzyme 10		
					At1g07940	CF634765	AGGCAGACGG			
					At5g53300	CF636241	TTGGAAGAAA		17	0.0103
							TATTCGTTCT		15	0.0091
GCTGAGAAGA	128	0.0888	At1g07890	CF634835	TAGGGAGGGG	5	0.003	L-Ascorbate peroxidase 1, cytosolic		
					TGGTCTTTTG	20	0.0122			
CCGTCGTGTT	121	0.084	At3g11510	CF625147	GCATACAACA	20	0.0122	40S ribosomal protein S14		
CACACTCACA	117	0.0812	At5g17920	CF632218	GTGTTAGTGT	17	0.0103	5-Methyltetrahydropteroyltryglutamate-homo-Cys S-methyltransferase		

(Table continues on following page.)

Table III. (Continued from previous page.)

Arabidopsis SAGE Tags	Copy No.	Abundance %	Arabidopsis Gene	Maize EST Homolog	Maize SAGE Tags ^a	Copy No.	Abundance %	Gene Description
TACGATGTTG	113	0.0784	At1g14320	CF634972	AAATTTGCT	52	0.0316	60S ribosomal protein L10
					AATTCAAGT	43	0.0262	
TTTGAAGGTG	112	0.0777	At5g15200	CF626023	GATGGTCCAT	323	0.1966	40S ribosomal protein S9
					GTGAGGGTGG	3	0.0018	
GAGAGAGTCA	110	0.0763	At3g25520	CF636691	CATTTTTGTT	97	0.059	60S ribosomal protein L5
					CTGCCATCCG	35	0.0213	
GATGTTGTTT	105	0.0729	At4g33865	CF637045	TGTTAAACTC	43	0.0262	40S ribosomal protein S29
					ATGTTTGTTT	35	0.0213	
					GATGATTCTG	9	0.0055	
GCCGTACCAA	105	0.0729	At1g58983	CF624288	TCACITTTTA	61	0.0371	40S ribosomal protein S2, putative
			At1g59359	CF624288	ATAGTGTTTT	60	0.0365	40S ribosomal protein S2, putative
					TCTAAATGCC	2	0.0012	

^aMultiple tags per gene indicate alternatively spliced transcripts or transcripts with multiple 3'-ends (see Poroyko et al., 2004).

Of obvious importance for root functioning are transport facilitators, transporters, and channels. The SAGE profile showed 70 tags (for 54 different transcripts) that clearly identified functions in this category (Supplemental Table IIB). Apart from putative, uncharacterized transport proteins (31), the list includes a variety of functionally known cation and anion, carbohydrate, amino acid, and ABC-type transporters of the plasma membrane and, in addition, intracellular transport proteins. Transcripts for three K⁺-channel proteins and three voltage-dependent anion channel proteins were among the most abundant SAGE tags.

DISCUSSION

SAGE provides an economical way to sample transcript profiles under specific experimental conditions. Sequencing of a relatively small number of clones will result in identifying the majority of expressed genes, which in our example provided nearly 15,000 unigenes after sequencing of 3,652 clones. A crucial element and potential problem in SAGE studies is tag-to-gene assignment with two approaches for a resolution (Lee et al., 2002). First, BLAST searches of SAGE tags versus EST or EST-derived unigene databases are employed. As a second approach, the experimental population of tags is compared with sets of tags that are virtually generated from EST or genome sequence data. The latter extraction process considers the length of 3'-UTR regions, and the number and arrangement of possible tags adjacent to the poly(A) tails of all deposited sequence information (Unneberg et al., 2003; Poroyko et al., 2004).

The results of several studies using SAGE on plant tissue have been reported recently. SAGE has been used for a study of gene expression in rice seedlings to a depth of approximately 10,000 tags (Matsumura et al., 1999). The study used 13-bp SAGE tags as primers that resulted in the recovery of differentially

represented sequences distinguishing anaerobically grown and untreated seedlings. For loblolly pine, two SAGE libraries have been generated that sampled transcript profiles during xylem lignification from either the crown or base portions of the trunk of trees, recording 150,855 tags (including 27,279 single-copy tags; Lorenz and Dean, 2002). This number described the transcriptome of the pine trunk to a depth of 15,383 expressed transcripts. Illustrating limitations of SAGE, the study reported multiple tags found in individual ESTs and cDNAs, i.e. a single gene may be identified by more than one tag, which in fact also presents an advantageous feature because splice variants and alternative 3'-end formation of transcripts may be identified. Two SAGE-based expression profiles (50,159 tags) from mature leaf and immature seed tissue of rice (subsp. japonica cv Nipponbare) have been reported (Gibbins et al., 2003). The analysis revealed the expression of 4,546 and 711 different transcripts in leaf and seeds, respectively. A significant number of tags in this study reported transcripts encoded on the antisense strand of known mRNAs, an observation supported by another study that reported 687 bidirectional transcript pairs in rice (Osato et al., 2003).

Specifically for root tissues, SAGE (approximately 32,000 tags) has been used to profile transcript complexity in Arabidopsis and to assess responses to TNT (Ekman et al., 2003). At that level, the libraries contained 4,399 and 4,105 transcripts for control and TNT treatment, respectively. Based on these numbers, an extrapolation of the size of the Arabidopsis root transcriptome by double-reciprocal plots of the unique tags identified versus all tags sequenced resulted in a number of approximately 21,000 transcripts (Ekman et al., 2003). Quantitative PCR for several genes showed general agreement with SAGE. A large-scale study by SAGE of genes expressed in the roots of Arabidopsis has recently been reported (Fizames et al., 2004). Sequenced were 144,083 tags that represented at least 15,964 different mRNAs. For the tag-to-gene assignment, a computational approach, a cumulative

Table IV. Functional categories of 80 abundant SAGE tags in *Arabidopsis* roots compared to the most abundant transcripts in maize^a

Functional Category ^b	% Arabidopsis	% Maize
Translation, ribosomal structure	21.3	24
RNA processing and modification	0	2.4
Transcription	1.3	0.8
Chromatin structure and dynamics	0	1.6
Cell cycle control, cell division	2.5	0.8
Defense mechanisms	3.8	0
Signal transduction mechanisms	1.3	0.8
Cell wall/membrane/envelope biogenesis	3.8	0.8
Cytoskeleton	1.3	7.2
Intracellular trafficking and vesicular transport	2.5	0
Posttranslational modification, protein turnover, chaperones	5	5.6
Energy production and conversion	1.3	4
Carbohydrate transport and metabolism	2.5	4
Amino acid transport and metabolism	1.3	0
Nucleotide transport and metabolism	0	0.8
Coenzyme transport and metabolism	0	0.8
Lipid transport and metabolism	0	1.6
Inorganic ion transport and metabolism	2.5	1.6
Secondary metabolites biosynthesis	1.3	1.6
Function unknown	15	35
Dubious annotation ^c	13.8	7.5
Tags without ID	18.8	10

^aPercentage for a given category calculated based on the 80 most abundant tags in Arabidopsis. For gene products identified as having more than one function, all functions were considered. ^bFunctional categorization made by COG-delimited categories. ^cDubious annotations: tags cannot be unambiguously identified and functionally classified.

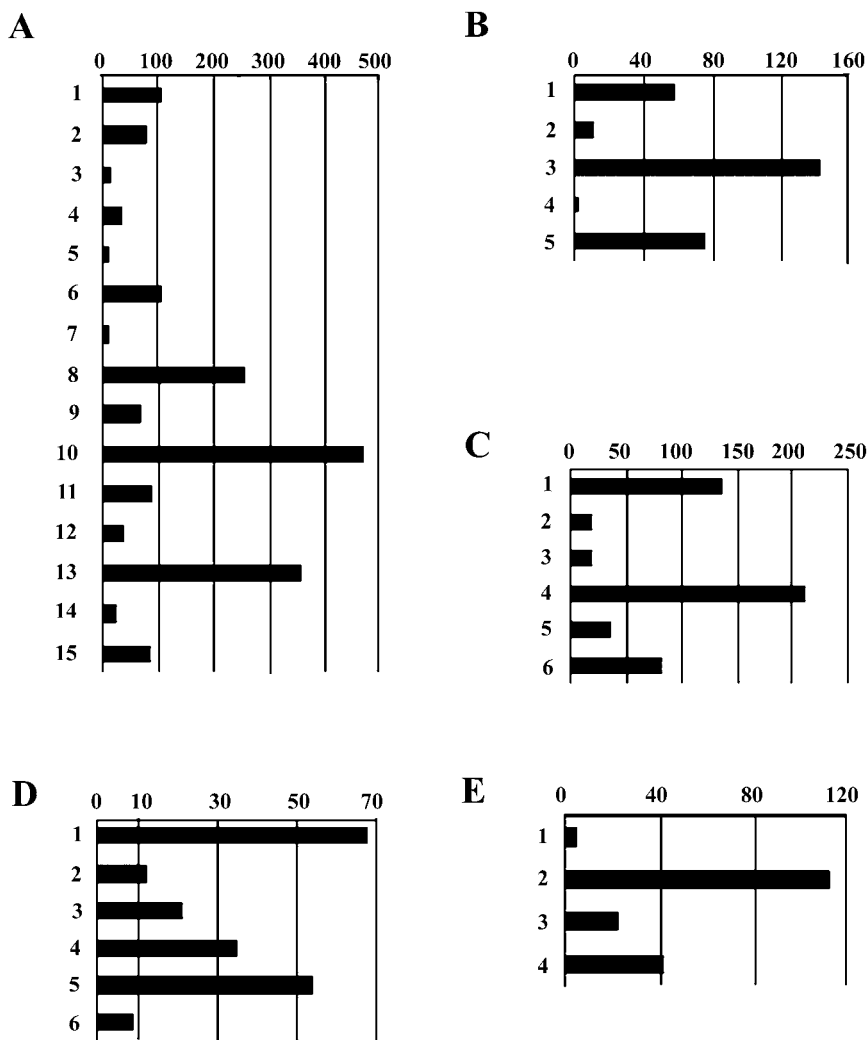
list of virtual tags, was used. It is based on 26,620 genes annotated in the sequence of the Arabidopsis genome (Arabidopsis Genome Initiative, 2000). The advantage of an entirely sequenced genome allowed for the identification of approximately 89% of the experimental tags. SAGE reported approximately 16,000 expressed genes in root RNA, and this number may be compared to a microarray-based study of gene expression in the Arabidopsis root, which identified 10,492 transcripts that produced signals in the hybridizations of root RNA to the microarray slides that contained

probes for approximately 22,000 genes (Birnbaum et al., 2003).

The complexity of the maize root transcriptome is comparable with estimates of the complexity for the Arabidopsis root transcript population. Our analyses may indicate that the number could exceed 22,000 transcripts. This is based on attempts to analyze the origin of unmatched (single) tags from human tissues, using PCR amplification of approximately 1,000 of such orphan matches in a collection of 4,285,923 SAGE tags (Chen et al., 2002a). The study discovered approximately 70% of single tags originated from transcripts previously not identified in the human genome, including alternatively spliced versions of known genes. The authors suggested that most single-copy SAGE tags were not generated from experimental error and calculated an error rate of approximately 1.7% per SAGE tag. If we followed this rationale, based on 25,749 orphan tags (Table I), an error rate of approximately 2% per tag, and 70% genuinely individual transcripts, the total number of transcripts expressed in the maize root would increase significantly by approximately 17,000 transcripts, resulting in a transcript complexity to 32,000 expressed genes in maize roots. The observation that a large number of single tags could be converted into tags that have been observed more than once by a single base change would lower this number again.

Information about transcript abundance is an important aspect of the SAGE analysis of the maize root because it can provide a control for many other analyses. In our hands, SAGE provided a more accurate or sensitive representation of the transcript population, in particular with respect to variability of 3'-end formation. Accuracy and completeness of SAGE profiling has been analyzed in a comparison of 76,790 tags for transcripts from rat hippocampal tissue with Affymetrix genechip data (Evans et al., 2002). Both techniques were comparable in detecting medium and high-abundance transcripts, and both produced inconsistent results in the detection of low-abundance transcripts. The results suggest that at the depth of approximately 77,000 SAGE sequences, SAGE sampled approximately 41% of all transcripts, whereas 30% of the transcripts in this tissue generated a signal with genechips. Statistical correlation between SAGE and quantitative PCR ($P = 0.014$), in contrast to quantitative PCR versus EST microarray data sets ($P > 0.05$), has been shown in the study of adult mouse heart tissues (Anisimov et al., 2002). A similar general agreement of SAGE data and quantitative PCR has also been reported in the study of Arabidopsis roots (Ekman et al., 2003). In this study, genes were tested that SAGE analysis had reported as either induced >14-fold (At3g28740), repressed 0.03-fold (At2g36830), or unaffected with a ratio of 1.1 (At2g39460) by TNT exposure. Quantitative PCR data reported a ratio of 46 for the first, 0.09 for the second, and 1.9 with respect to the third transcript. The apparent discrepancies between the values that had been determined by the two

Figure 3. Abundance of tags for enzymes in selected biochemical pathways in the maize primary root. Shown is tag abundance for identified transcripts in different pathways: A, glycolysis; B, pentose phosphate cycle; Further selected is a pathway, Suc degradation, a sink function of the root (C); lignin biosynthesis (D), and sulfur assimilation (E). A1, UDP-Glc pyrophosphorylase; 2, Phosphoglucomutase; 3, Hexokinase; 4, Glc-6-P isomerase; 5, Phosphofruktokinase; 6, Pyrophosphate-Fru-6-P-1-phosphotransferase; 7, Fru-6-P-2-kinase; 8, Fru bisphosphate aldolase; 9, cytosolic triosephosphate isomerase; 10, Glyceraldehyde-3-P dehydrogenase; 11, Phosphoglycerate kinase; 12, Phosphoglycerate mutase; 13, Enolase; 14, Pyruvate kinase; 15, Phosphoenolpyruvate carboxylase. B1, Glc-6-P dehydrogenase; 2, 6-Phosphogluconolactonase; 3, 6-Phosphogluconate dehydrogenase; 4, D-Ribulose-5-P 3-epimerase; 5, Transaldolase. C1, Beta-fructosidase; 2, Hexokinase; 3, Suc-UDP glucosyltransferase; 4, Fructokinase; 5, Glc-6-P isomerase; 6, Phosphoglucomutase. D1, Phenylalanine ammonia lyase; 2, Cinnamate 4-hydroxylase; 3, 4-Coumarate Co-A ligase; 4, Coumarate 3-hydroxylase; 5, Caffeoyl-CoA O-Methyltransferase; 6, Cinnamoyl Co-A reductase. E1, ATP sulfurilase; 2, Cys synthase/O-acetyl-Ser (thiol) lyase; 3, Cystathionine gamma-synthase; 4, homo-Cys methylase.



techniques were explained by the fact that values determined by quantitative PCR fit a logarithmic rather than linear function of the starting number of templates in a SAGE population (see Ekman et al., 2003). The SAGE-based value for the induction of gene At3g28740 therefore represents a significant underestimation because of the constraints originating from the depth to which SAGE libraries are sequenced. In essence, the depth of sequencing into SAGE libraries is certainly the most significant source for any discrepancies noted in data produced by the two techniques.

The value of this tag collection will increase as more maize genomic sequences become available. For example, the genechip-based analysis of the Arabidopsis root (Birnbaum et al., 2003) indicated approximately 500 transcripts for TFs (out of a total of approximately 10,000 signals, i.e. approximately 5%) expressed in the roots. Surprisingly, we identified only 44 TFs as present in the collection of 7,840 tags for which annotations were possible, or approximately 1% of all identified tags (listed in the supplemental material). The discrepancy might be explained by the different

ways the tissues were sampled, by sorting protoplasts in one and by physically collecting root segments in the second case. Also, an equal number of TFs can be expected to be among the tags, 50% of all, for which no maize gene or transcript model is yet available.

The results from this analysis will become accessible to reinterpretation once a larger segment of the maize genome is available. First, the number of alternatively terminated transcripts is high and may reveal functions in RNA turnover, silencing, or targeting, or may have a developmental and cell specificity role as documented in the Supplemental Table II where identical accession numbers identify different 3'-end tags for the same transcript. The presence of tags that match more than one gene can be resolved only when the genome has been sequenced. Importantly, exemplified in Figure 3, transcripts for proteins in primary metabolism provide information that surpasses what can be obtained by microarray analysis. Also, the diversity of transcript numbers for different enzymes/proteins in a pathway and the expression of different isoforms for pathway enzymes reveal information about regulatory circuits,

pathway networks, and possibly even protein half-life. Recording SAGE or MPSS tags in tissues and cells, at various developmental stages or under diverse experimental manipulations in the most widely used models and crop species, will eventually provide baseline values for true transcript complexity and abundance.

MATERIALS AND METHODS

Root Material, RNA Isolation, and SAGE Library Construction

Maize (*Zea mays* L. cv FR697) seeds were imbibed for 24 h in 1 mM CaSO₄ and were germinated for 28 h in vermiculite well moistened with 1 mM CaSO₄ at 29°C in the dark (Spollen et al., 2000). Seedlings with primary roots 12 to 20 mm in length were transplanted into vermiculite at high water potential (−0.03 MPa, obtained by mixing with 1 mM CaSO₄) and grown under the same conditions. At 5 and 48 h after transplanting, primary roots from 500 seedlings were harvested (green safelight; Saab et al., 1990) and the primary root apical 20 mm were sectioned into four segments (distances are from the junction of the root apex): segment (1) 0 to 3 mm plus the root cap; (2) 3 to 7 mm; (3) 7 to 12 mm; (4) 12 to 20 mm. Segments 1 to 3 constitute the elongation zone in the primary root of well-watered seedlings of this cultivar (Sharp et al., 2004). From each segment, 250 mg of material was taken and used for RNA isolation (RNeasy Maxi, Qiagen, Valencia, CA). The procedure was chosen to make the experimental material comparable to material for which a total of approximately 23,000 EST sequences have been generated (see <http://rootgenomics.missouri.edu/prgc/index.html>). Fifty μg of RNA was used to generate the SAGE library (I-SAGE; Invitrogen, Carlsbad, CA) from which 3,652 clones were sequenced.

DNA Sequencing

The library was plated on agar with zeocin 50 μg/mL and colonies were picked. Bacteria were inoculated into 96-well-deep culture plates with Luria-Bertani medium and grown overnight. Plasmid DNA was purified from bacterial cultures using the Qiagen-9600 BioRobot. Sequencing reactions were performed by BigDye terminator chemistry (Applied Biosystems, Foster City, CA) using the standard primer M13 reverse for −48. Sequencing of clones was carried out with ABI3700 and ABI3730xl capillary systems at the Keck Center, University of Illinois Urbana-Champaign.

Bioinformatics, Tag Annotation, and Data Analysis

For SAGE library analysis and SAGE tag extraction the SAGE-2000 v4.5 software package was used (<http://www.invitrogen.com/sage>). SAGE tags were identified in two ways (Lee et al., 2002). First, a comparison was made of the SAGE-2000 tag collection with the set of 3'-end NlaIII-adjacent tags extracted from a set of 3'-end sequenced ESTs obtained from three normalized cDNA libraries (zmrww00, zmrws05, zmrws48) and one subtracted library (zmsub1) with poly(A)⁺ deposited at GenBank by the Maize Root Genomics Consortium (<http://rootgenomics.missouri.edu>). EST sequences were assembled using CAP3 (Huang and Madan, 1999). The following CAP3 parameters were used: minimum 50 bp overlap; 95% similarity in the overlap; clipping range 60 bp. The 3'-end NlaIII-adjacent tags were extracted by a Perl script, termed V-SAGE (Poroyko et al., 2004). All EST contigs were annotated by BLASTX search using the NCBI nonredundant protein database. Second, we used BLASTN of the full-length 14-bp tags (CATG + 10 bp tag sequence) against the NCBI database "Zea mays UniGene Build #40." The strand orientation of maize UniGene sequences was determined by BLASTX against a database of 29,161 Arabidopsis (*Arabidopsis thaliana*; putative) translated protein sequences (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/ATH1_pep_cm_20040228) and the NCBI nonredundant protein database.

Functional categorization and annotation for parental maize datasets was done according to clustering of orthologous groups for eukaryotic complete genomes NCBI (<http://www.ncbi.nlm.nih.gov/COG/>) (Tatusov et al., 1997, 2003). The presence of identified functional domains, determined by BLASTX search, in the examined sequence was accepted for the assignment of functions. Pathway assembly was carried out in the same way and included a BLASTX search against the Arabidopsis protein database at TAIR (<ftp://>

ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/). The search results were linked to the Arabidopsis biochemical pathway map from the TAIR Web site (<http://arabidopsis.org/tools/aracyc/>).

Control of Tag Frequency by qPCR

For RT, 2 μg of total RNA were used. Reactions were done using Superscript II (Invitrogen) according to the manufacturer's instructions. The product of the RT reaction was diluted by a factor of 20 and used as a template for quantitative PCR. Reactions were performed using Smart Cycler (Cepheid, Sunnyvale, CA). Primers for qPCR were designed to produce amplicons of equal size (approximately 115 bp; Supplemental Table IVA). Primers selected for low-complexity tags are used as an additional test of SAGE veracity (Supplemental Table IVB).

The composition of reaction mixtures for quantitative PCR (final volume 25 μL) was 12.5 μL Sybr green master mix (Applied Biosystems), 2 μL of diluted cDNA template, and 1 μL of each (10 μM) of the primers. The PCR amplification program included one cycle at 95°C (15 min), 40 cycles at 95°C (each 15 s), 60°C (30 s), 72°C (30 s), and one cycle at 72°C (2 min). Melting curves for each product were made by heating of from 60°C to 95°C at 0.2°C/s. All melting curves generated a single melting point, indicating homogeneity of the products.

Absolute copy number amounts were determined by a standard curve for each selected amplicon as described (QuantiTect SYBR Green PCR; Qiagen). The transcript-per-cell ratio was calculated by the formula: $C = 2,600^N/m$, with C, transcript per cell amount; N, number of molecules determined in a QPCR reaction; m, amount of RNA taken for RT (pg). The value of 2,600 represents a constant derived from sample dilution and total RNA amount per eukaryotic cell (estimated at 13 pg; see Okamura and Goldberg, 1989).

Generation of Long DNA Fragments from SAGE Tags, the GLGI Protocol

The same total RNA sample used for SAGE analyses was used for GLGI. RT was done by using Superscript III (Invitrogen), 4 μg of total RNA and 3' adapter primer ACTATCTAGAGCGGCCGCTTTTTTTTTTTTTTTTTTTT at a final concentration of 5 μM. The reaction conditions were selected according to the manufacturer instructions. Finally, the RT reaction was diluted 10 times (ddH₂O) and used for subsequent amplification. Tag-specific amplification used the Peltier Thermal Cycler (C225, MJ Research, Reno, NV). The GLGI master mixture containing the antisense primer, cDNA template, and DNA polymerase was prepared: 21 μL per reaction, including 10 μL of Eppendorf MasterMix (2.5 ×) (Eppendorf, Westbury, NY), 4 μL of 3' universal amplification primer ACTATCTAGAGCGGCCGCTT (10 μM), 7 μL of diluted template. The tag-specific sense primer 4 μL (10 μM) was added to each well GGATCCCATG[XXXXXXXXXX]. The tag-specific 5' primers were designed for 55 randomly selected single-copy SAGE tags (Supplemental Table V). The PCR conditions used were as follows (Chen et al., 2002b): 94°C (2 min), five cycles at 94°C (30 s), 55°C (30 s), 72°C (30 s), and then 25 cycles at 94°C (30 s), 60°C (30 s), 72°C (30 s), and the reactions were kept at 72°C for 5 min after the last cycle. PCR products were examined by gel electrophoresis on 2% agarose.

ACKNOWLEDGMENTS

We thank Dr. Alvaro Hernandez and Dr. Ryan Kim (University of Illinois Urbana-Champaign) and Ruth Grene (Virginia Tech University) for discussions, and the team at the Keck Center for Comparative and Functional Genomics and Vladimir Calugaru for help. The data discussed here have been incorporated into a database, supplemental tables are part of this manuscript, and data are available from a project Web site: <http://rootgenomics.missouri.edu/prgc/index.html>.

Received January 26, 2005; revised March 15, 2005; accepted March 21, 2005; published June 17, 2005.

LITERATURE CITED

Anisimov SV, Tarasov KV, Stern MD, Lakatta EG, Boheler KR (2002) A quantitative and validated SAGE transcriptome reference for adult mouse heart. *Genomics* 80: 213–222

- Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815
- Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN** (2003) A gene expression map of the *Arabidopsis* root. *Science* **302**: 1956–1960
- Boon K, Osório EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, de Souza SJ, et al** (2002) An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci USA* **99**: 11287–11292
- Chen J, Lee S, Zhou G, Wang SM** (2002b) High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. *Genes Chromosomes Cancer* **33**: 252–261
- Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM** (2002a) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc Natl Acad Sci USA* **99**: 12257–12262
- Ding C, Cantor CR** (2004) Quantitative analysis of nucleic acids: the last few years of progress. *J Biochem Mol Biol* **37**: 1–10
- Ekman DR, Lorenz WW, Przybyla AE, Wolfe NL, Dean JFD** (2003) SAGE analysis of transcriptome responses in *Arabidopsis* roots exposed to 2,4,6-trinitrotoluene. *Plant Physiol* **133**: 1397–1406
- Evans SJ, Datson NA, Kabbaj M, Thompson RC, Vreugdenhil E, De Kloet ER, Watson SJ, Akil H** (2002) Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. *Eur J Neurosci* **16**: 409–413
- Fizames C, Muñis S, Cazettes C, Nacry P, Boucherez J, Gaymard F, Piquemal D, Delorme V, Commes T, Doumas P, et al** (2004) The *Arabidopsis* root transcriptome by serial analysis of gene expression: gene identification using the genome sequence. *Plant Physiol* **134**: 67–80
- Gibbings JG, Cook BP, Dufault MR, Madden SL, Khuri S, Turnbull CJ, Dunwell JM** (2003) Global transcript analysis of rice leaf and seed using SAGE technology. *Plant Biotechnol J* **1**: 271–285
- Gowda M, Jantauriyarat C, Dean RA, Wang GL** (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiol* **134**: 890–897
- Huang X, Madan A** (1999) CAP3: a DNA sequence assembly program. *Genome Res* **9**: 868–877
- Jansen R, Gerstein M** (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res* **28**: 1481–1488
- Jung SH, Lee JY, Lee DH** (2003) Use of SAGE technology to reveal changes in gene expression in *A. thaliana* leaves undergoing cold stress. *Plant Mol Biol* **52**: 553–567
- Lee JY, Lee DH** (2003) Use of serial analysis of gene expression technology to reveal changes in gene expression in *Arabidopsis* pollen undergoing cold stress. *Plant Physiol* **132**: 517–529
- Lee S, Clark T, Chen J, Zhou G, Scott LR, Rowley JD, Wang SM** (2002) Correct identification of genes from serial analysis of gene expression tag sequences. *Genomics* **79**: 598–602
- Liang P** (2002) SAGE Genie: a suite with panoramic view of gene expression. *Proc Natl Acad Sci USA* **99**: 11547–11548
- Lorenz WW, Dean JFD** (2002) SAGE profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (*Pinus taeda*). *Tree Physiol* **22**: 301–310
- Matsumura H, Nirasawa S, Terauchi R** (1999) Transcript profiling in rice (*O. sativa* L.) seedlings using serial analysis of gene expression (SAGE). *Plant J* **20**: 719–726
- Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD** (2004a) *Arabidopsis* MPSS: an online resource for quantitative expression analysis. *Plant Physiol* **135**: 801–813
- Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S** (2004b) The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res* **14**: 1641–1653
- Okamura JK, Goldberg RB** (1989) Regulation of plant gene expression: general principles. In: Marcus, ed, *The Biochemistry of Plants*, Vol 15. Academic Press, San Diego, pp 1–82
- Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K, et al** (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol* **5**: R5
- Poroyko V, Calugaru V, Fredricksen M, Bohnert HJ** (2004) Virtual-SAGE: a new approach to EST data analysis. *DNA Res* **11**: 145–152
- Robinson SJ, Cram DJ, Lewis CT, Parkin IA** (2004) Maximizing the efficacy of SAGE analysis identifies novel transcripts in *Arabidopsis*. *Plant Physiol* **136**: 3223–3233
- Saab IN, Sharp RE, Pritchard J, Voetberg GS** (1990) Increased endogenous abscisic acid maintains primary root growth and inhibits shoot growth of maize seedlings at low water potentials. *Plant Physiol* **93**: 1329–1336
- Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE** (2002) Using the transcriptome to annotate the genome. *Nat Biotechnol* **20**: 508–512
- Sharp RE, Poroyko V, Hejlek LG, Spollen WG, Springer GK, Bohnert HJ, Nguyen HT** (2004) Root growth maintenance during water deficits: physiology to functional genomics. *J Exp Bot* **55**: 2343–2351
- Spollen WG, LeNoble ME, Samuels TD, Bernstein N, Sharp RE** (2000) Abscisic acid accumulation maintains maize primary root elongation at low water potentials by restricting ethylene production. *Plant Physiol* **122**: 967–976
- Stern MD, Anisimov SV, Boheler KR** (2003) Can transcriptome size be estimated from SAGE catalogs? *Bioinformatics* **19**: 443–448
- Tatusov RL, Fedorova ND, Jackson JB, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al** (2003) The COG database: an updated version including eukaryotes. *BMC Bioinformatics* **4**: 41
- Tatusov RL, Koonin EV, Lipman DJ** (1997) A genomic perspective on protein families. *Science* **278**: 631–637
- Tuteja R, Tuteja N** (2004) Serial analysis of gene expression: applications in malaria parasite, yeast, plant, and animal studies. *J Biomed Biotechnol* **2**: 106–112
- Unneberg P, Wennborg A, Larsson M** (2003) Transcript identification by analysis of short sequence tags: influence of tag length, restriction site and transcript database. *Nucleic Acids Res* **31**: 2217–2226
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW** (1995) Serial analysis of gene expression. *Science* **270**: 484–487
- Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P, Vogelstein B, Kinzler KW** (1997) Characterization of the yeast transcriptome. *Cell* **88**: 243–251
- Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y** (2004) 5' Long serial analysis of gene expression (Long-SAGE) and 3' Long-SAGE for transcriptome characterization and genome annotation. *Proc Natl Acad Sci USA* **101**: 11701–11706