nature biotechnology

Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing

Blake C Meyers[1], Tam H Vu[1], Shivakundan Singh Tej[1], Hassan Ghazal[1], Marta Matvienko[2,4], Vikas Agrawal[1], Jianchang Ning[1] & Christian D Haudenschild[3]

Large-scale sequencing of short mRNA-derived tags can establish the qualitative and quantitative characteristics of a complex transcriptome. We sequenced 12,304,362 tags from five diverse libraries of *Arabidopsis thaliana* using massively parallel signature sequencing (MPSS). A total of 48,572 distinct signatures, each representing a different transcript, were expressed at significant levels. These signatures were compared to the annotation of the *A. thaliana* genomic sequence; in the five libraries, this comparison yielded between 17,353 and 18,361 genes with sense expression, and between 5,487 and 8,729 genes with antisense expression. An additional 6,691 MPSS signatures mapped to unannotated regions of the genome. Expression was demonstrated for 1,168 genes for which expression data were previously unknown. Alternative polyadenylation was observed for more than 25% of *A. thaliana* genes transcribed in these libraries. The MPSS expression data suggest that the *A. thaliana* transcriptome is complex and contains many as-yet uncharacterized variants of normal coding transcripts.

The genomic sequence of *A. thaliana* has been completed in recent years[1], and that of rice is nearly complete. Experimental analyses and comprehensive descriptions of plant transcriptomes continue in parallel[2,3]. No plant transcriptome has been extensively characterized experimentally with both quantitative and qualitative expression data. Computational approaches to genome annotation can miss or incorrectly predict many genes, and validation of genome annotations with experimental data is essential[4–6]. As genomic sequencing becomes faster and more economical, it is critically important that methods are developed to detect and quantify every gene and alternatively spliced transcript within a genome.

One of the most important recent discoveries in biology is the identification of RNA molecules that do not encode proteins; these RNA molecules are called noncoding (ncRNAs)[7]. ncRNAs are difficult to predict in the absence of experimental data, although comparative approaches combined with predictions of secondary structure may identify some ncRNAs[7]. With the exception of housekeeping RNAs, like transfer RNAs or small nucleolar RNAs, the relatively few potential regulatory ncRNAs that have been characterized appear to be plant-specific[8]. Natural antisense transcripts (NATs) are another form of ncRNA. NATs overlap with transcribed coding regions and may be involved in the regulation of gene expression[9]. It is likely that ncRNAs, including NATs, are a major component of the diversity of transcripts produced in higher eukaryotes. Nearly all of the more than 29,000 predicted genes in *A. thaliana* encode proteins; very few ncRNAs are annotated[8,10]. Studies using more comprehensive transcriptional profiling approaches, such as whole-genome arrays with 'tiled' probe sets, can add important new information to the *A. thaliana* genome[2].

We have used the technology MPSS[11,12] to experimentally assess the complexity of the *A. thaliana* transcriptome. The process of MPSS starts with the cloning of a cDNA library on beads, with one transcript in the original RNA sample represented on each bead[12]. MPSS sequencing determines sets of four bases per bead by hybridization to labeled linker-probes. These bases are removed by a Type IIS restriction enzyme and the process is repeated to determine the next set of four bases. These reactions occur while the beads are immobilized, and for each bead, a sequence 'signature' of 17 or more nucleotides is obtained by successive rounds of sequencing reactions. The signatures are derived from and include the most 3′ occurrence of a specific restriction enzyme site in a transcript[11,12]. This enzyme is most often *Dpn*II, producing signatures that start with GATC. One signature is sequenced from each transcript in a library, and the technology permits the simultaneous sequencing of millions of signatures[11]. When matched to the genome to identify specific genes, the abundance of each signature represents the gene expression levels in the sampled tissue.

MPSS, like expressed sequence tags (ESTs)[13] and serial analysis of gene expression (SAGE)[14], is a tag-based method of analyzing gene expression. Combining such data with genomic sequence identifies previously unidentified genes while providing quantitative measurements of gene expression[15]. Recent applications of MPSS have identified differentially expressed genes in *A. thaliana*, but have not taken

[1]Delaware Biotechnology Institute, 15 Innovation Way, University of Delaware, Newark, Delaware 19714, USA. [2]Department of Vegetable Crops, 148 Asmundson Hall, University of California, Davis, California 95616, USA. [3]Lynx Therapeutics, Inc., 25861 Industrial Blvd., Hayward, California 94545 USA. [4]Present address: Allometra, LLC, 2604 Kalamazoo Place, Davis, California 95616, USA. Correspondence should be addressed to B.C.M. (e-mail: meyers@dbi.udel.edu).

Published online 11 July 2004; doi:10.1038/nbt992

Table 1 Libraries and signature summary statistics

		Distinct signatures						
Library	Total signatures	All TPM[a]	1–3 TPM[b,c]	4–100 TPM[b]	101–1,000 TPM[b]	>1,000 TPM[b]	Filtered total >3 TPM[b]	Filtered total[b,c]
Callus	1,959,539	40,901	4,494	17,363	2,584	134	20,081	24,575
Inflorescence	1,774,306	37,750	4,577	15,673	2,579	115	18,367	22,944
Leaves	2,884,598	53,394	7,603	18,320	2,004	141	20,465	28,068
Root	3,642,632	48,100	5,903	18,288	2,261	164	20,713	26,616
Silique	2,012,859	38,501	4,623	17,333	2,261	120	19,714	24,337
Total	12,273,934	133,375	20,801	41,357	5,609	390	42,590	68,157

[a]Includes all signatures regardless of filter results. [b]Includes only signatures which are reliable and significant. [c]Includes signatures in the range of 1 to 3 TPM that were expressed at significant levels in libraries from experiments not described here[19].

full advantage of the genomic sequence data[16–18]. In the experiments described here, we obtained 12,304,362 sequence signatures from five organs or tissues of *A. thaliana* and used these data to describe the diversity of transcripts encoded in this plant genome.

RESULTS

MPSS signatures matched to the A. thaliana genomic sequence

MPSS was performed on mRNA isolated from five tissues of *A. thaliana*, including leaf, root, silique, inflorescence and callus. The number of signatures sequenced per library varied from 1,774,306 to 3,642,632, representing between 37,750 and 53,394 distinct signatures (**Table 1**). We merged the sequencing runs and normalized the expression level of each signature in units of transcripts per million signatures (TPM)[19]. To focus on signatures in which we have the highest confidence, we disregarded 'nonsignificant' signatures never observed at levels greater than 3 TPM in any library. A second filter removed the 'unreliable' signatures observed only in one run; both filters are described elsewhere[19]. These two filters removed between 18,787 (49%) and 32,929 (61%) of the distinct signatures; however, these signatures typically represent only a small proportion of the expression data[19]. After filtering, the complexity of each library was similar (**Table 1**). More than two-thirds of the filtered signatures in each library were found in the range of 4 to 100 TPM; less than 1% of the signatures were expressed at levels above 1,000 TPM (**Table 1**). Overall, the signature abundances indicate a large number of distinct transcripts expressed at abundance levels spread over four orders of magnitude in the five libraries.

We matched the MPSS signatures with genomic sequence information to link the expression data to specific genes and genomic positions. The genomic locations of expressed signatures were determined as described[19] and compared with the 29,084 annotated genes and pseudogenes in the *A. thaliana* genomic sequence (TIGR version 3.0)[10]. A total of 30,128 signatures were matched with unique locations in the genome, 3,073 signatures mapped to duplicated locations and 9,389 remained unmatched (**Table 2** and **Supplementary Table 1**). The unmatched signatures may be derived from four sources: (i) sequencing errors (ii) spliced 3′ ends that have not yet been identified (iii) transcripts found in regions of the genome not yet sequenced or (iv) non-*A. thaliana* contaminants. These possibilities have been analyzed in greater detail elsewhere[19], but because we have filtered out the low abundance signatures and matches to full-length cDNAs, many of the remaining signatures are likely to be derived from novel transcripts or as-yet unidentified splice variants in the *A. thaliana* Col-0 transcriptome.

The set of MPSS signatures mapped uniquely in the genome ranged from 14,651 to 16,493 in each of the five libraries (**Table 2**). Because the MPSS signatures are derived from specific locations in the 3′ end of an mRNA molecule, each distinct signature corresponds to a distinct transcript, unless the sequence is duplicated in the genome. Signatures duplicated in the genome identified another 2,229 to 2,749 transcripts, although in these cases it is not clear which genomic location is transcriptionally active (**Supplementary Table 1**). The signatures in each of the five libraries were consistent with most transcripts being produced by sense-strand expression that generated normal protein-coding mRNA molecules. However, there was a substantial amount of antisense transcription, as evidenced by the more than 4,298 class 3 and 400 class 6 signatures found across the five libraries (**Table 2** and **Supplementary Table 1**).

Previously uncharacterized transcripts identified by MPSS

An examination of the number of signatures in each class revealed that numerous signatures in each of the five libraries identified sets of novel transcripts. These predicted transcripts can be grouped into three categories.

Antisense transcripts. In the five MPSS libraries, we matched 4,698 signatures with unique genomic positions (5,796 including duplicated signatures) and were antisense to annotated genes (**Table 2** and **Supplementary Table 1**). NATs are predicted to play an important role in the regulation of gene expression in higher eukaryotes[9,20]. Specific examples of antisense RNAs have been identified from plants[21] and antisense expression was recently detected for more than

Table 2 Unique genomic signatures and 17-base MPSS data from five libraries

Class	Description	TIGR Genome	Callus	Inflorescence	Leaves	Root	Silique	Total in libraries
1	Exon, sense strand	162,509	10,120	9,984	10,315	9,796	9,604	15,372
2	500 bp 3′-UTR	36,727	2,711	2,678	2,513	3,354	2,872	5,338
3	Exon, antisense strand[a]	157,645	1,105	475	1,424	1,641	749	4,298
4	Unannotated region[a]	216,717	1,655	884	935	1,089	907	3,466
5	Intron, sense strand[a]	52,218	421	338	328	350	330	908
6	Intron, antisense strand[a]	49,982	196	77	73	69	43	400
7	Span splice site, sense strand	5,682	224	215	250	194	212	346
	Total[b]	**681,480**	**16,432**	**14,651**	**15,838**	**16,493**	**14,717**	**30,128**
0	No match to genome	—	1,998	2,336	3,170	2,678	3,673	9,389

Only significant and reliable signatures expressed at greater than 3 TPM in these libraries were considered.

[a]Classes 3, 4, 5 and 6 predominantly represent previously unknown, unannotated transcripts. [b]For libraries, sum of signature classes grouped by distinct genes and distinct class 4 signatures.

Table 3 Genes and alternative transcripts detected by MPSS signatures

	All five libraries	
Description	>3 TPM	All TPM
Distinct coding transcripts, hits = 1[a]	21,964	24,641
Hits > 1	4,036	4,668
Hits ≥ 1	26,000	29,309
Distinct genes, hits = 1[b]	16,598	17,445
Hits > 1	3,535	3,972
Hits ≥ 1	18,612	19,518
Alternative transcripts, hits = 1[c]	5,366	7,196
Distinct genes with alt. transcripts, hits = 1	4,337	5,486
Distinct genes with alt. transcripts, hits > 1	422	538

[a] The sum of the class 1, 2, 5 or 7 expressed signatures indicates the number of distinct transcripts. See **Supplementary Table 2** for more details and for individual libraries. [b] The union of the set of genes identified by class 1, 2, 5 or 7 signatures. [c] Alternative transcripts calculated as the number of distinct transcripts minus the number of distinct genes.

Table 4 MPSS signatures identify both previously predicted and potentially novel transcripts

	IGRs	TIGR genes identified	Genes uniquely identified[a]
Whole-genome array[b]	1,017 (of 7,824)	21,605 (of 27,916)	3,626
cDNA transcripts[c]	1,369 (of 11,405)	18,365 (of 29,084)	809
MPSS transcripts[d]	1,256 (of 11,806)	18,162 (of 28,913)	1,168

The subset of genes detectable by each technology is indicated in parenthesis. For the MPSS data, only significant and reliable signatures with hits = 1 were used.

[a] Found in only one of the three sets of transcriptional data described here. [b] Details on the whole genome array and the definition of the IGRs are defined in ref. 2 but we recalculated the IGRs using TIGR version 3.0. Not all IGRs that met the definition were measured in the WGAs. [c] Includes ESTs and full-length cDNA sequences in GenBank on December 1, 2003. IGRs for cDNA comparisons were defined as for WGAs; a minimum length of 50 bases reduced the number slightly compared to the MPSS analysis below. These IGRs were compared by BLAST analysis requiring a match with 95% identity and longer than 50 bases. [d] The number of genes was calculated as the count of annotated genes containing class 1, 2, 5 or 7 signatures. IGRs for MPSS were defined as for WGAs; these were considered positive if they contained at least one genomic class 4 signature.

7,500 *A. thaliana* genes using whole-genome tiling arrays[2]. In our libraries, the antisense transcript complexity of the root was the highest (1,710 unique in the genome, 2,208 including duplicated signatures) and the complexity of the inflorescence was the lowest (552 unique in the genome, 914 including duplicated signatures) (**Table 2** and **Supplementary Table 1**). This may reflect a difference in the level of antisense-mediated post-transcriptional regulation in these two diverse tissues.

Unannotated transcripts in intergenic regions (IGRs). The MPSS signatures from the five libraries uniquely mapped to 3,466 distinct positions in the genome that may contain unannotated transcripts (*e.g.*, distinct class 4 signatures, **Table 2**). The most complex set of IGR-matching signatures was found in the callus library, which had 1,655 class 4 signatures mapping uniquely in the genome. The inclusion of signatures duplicated in the genome extends the matches to 4,768 IGRs (**Table 2** and **Supplementary Table 1**). In comparison, expression was detected for 2,000 *A. thaliana* IGRs using whole-genome tiling arrays[2].

Alternatively terminated or polyadenylated transcripts. We used the MPSS data and signature classifications to estimate the rate of alternative polyadenylation. Multiple class 1, 2 or 7 signatures matching to a single annotated gene may be derived from transcripts encoding identical proteins; this multiplicity of signatures could result from variation in the polyadenylation site or in the splicing of the 3′-UTRs. This type of variation is known to occur in *A. thaliana* transcripts[22]. The presence of class 5 signatures (*e.g.*, those matching annotated introns) may indicate alternative polyadenylation that truncates the open reading frame, producing a different type of transcriptional variant. To estimate the number of all types of transcriptional variants, we summed the total number of class 1, 2, 5 and 7 signatures for each library that mapped to distinct annotated genes (**Table 3** and **Supplementary Table 2**). The sum of these signatures represents the total number of distinct transcripts that were detected. We also determined the 'union,' or nonredundant merged group, of the set of genes identified by these sets of signatures. The difference between the sum and union is the number of transcriptional variants of different expressed genes; this difference per library ranged from 1,554 (inflorescence) to 2,125 (root) and totaled 5,366 transcriptional variants in the five libraries (**Supplementary Table 2**). These variants were matched with 1,416 to 1,849 distinct genes in individual libraries (2,339 to 2,902 genes, if signatures at all expression levels were included). In total, these data indicate that at least 26.1% of *A. thaliana* genes (4,337 divided by 16,598; **Table 3**) produce transcriptional variants (*e.g.*, alternatively terminated or polyadenylated transcripts). Additional variants may be detected by duplicated or low-abundance signatures (**Table 3**).

MPSS compared to other experimental approaches

The MPSS data indicate that (i) more than 60% of annotated genes are expressed in a relatively small sample of tissues and (ii) a large number of previously unannotated transcripts exist in the *A. thaliana* genome. A similar conclusion was reached by an analysis of full-length *A. thaliana* cDNA sequences and whole-genome microarrays (WGAs)[2]. We compared the set of annotated and novel transcripts with sense-strand expression in our five libraries against the set of experimentally defined *A. thaliana* cDNA sequences in GenBank and against the set of WGA data[2] (**Table 4**). Using the set of *A. thaliana* cDNA and EST sequences in GenBank, including those described[2], a total of 18,365 of the 29,084 annotated genes were matched, and only 809 genes were identified that were not present in either the MPSS or WGA data set (**Table 4**). In contrast, the five MPSS libraries identified 18,162 genes of which 1,168 were unique to that data set, and the WGAs identified 21,605 genes of which 3,626 were unique to that data set. Each of these methods has detected a set of transcripts that largely overlap with those detected by the other technologies but each is also able to detect transcripts that other technologies do not.

We directly compared the WGA and MPSS data because these technologies perform deep sampling within a single experiment, whereas the cDNAs were derived from a range of sources. Comparisons of WGA and MPSS data were performed for sense and antisense transcripts of annotated genes and for IGRs. We determined if each gene measured by WGAs contained either a class 1, 2, 5 or 7 signature (for sense-strand expression), or a class 3 or 6 signature (for antisense expression). The IGRs were defined previously[2], although we remapped these IGRs using the TIGR version 3.0 annotation and extended the IGR sequences by 500 bases at the 3′ end because MPSS signatures are derived from 3′ UTRs. Using only significantly expressed signatures, the total number of transcripts detected was greater for the WGAs than for MPSS (**Supplementary Table 3a**), particularly the number of genes with antisense transcripts. But when we included reliable signatures at all expression levels, the totals were more similar to each other, although MPSS still detected substantially less antisense transcription (**Supplementary Table 3b**). In each case, a subset of the previously undetected transcripts was detected by both MPSS and WGA, although both technologies identified transcripts not found by the other.

Patterns of transcription compared across the five libraries

The most abundant transcripts in each of the five MPSS libraries were consistent with the biology of the tissue or mapped to genes of unknown function. The top ten most abundant signatures along with the genes that they match are listed in **Supplementary Table 4**, and the data are on our web page, http://mpss.udel.edu/at. Interestingly, in callus, the top five transcripts are all from genes with poorly defined function, demonstrating how little is known about the biology of the undifferentiated plant cells. The gene list ordered from most to least abundant using normalized values for the MPSS signatures represents a nearly complete transcript inventory for each sampled tissue.

We compared the five libraries to identify signatures with tissue-specific expression patterns. These transcripts may play important and specific roles in the biology of these diverse tissues, and the regions upstream of these transcripts may have experimental utility as tissue-specific promoters. Between 176 and 456 genes in each library showed tissue specificity, corresponding to less than 0.25% of the expressed *A. thaliana* genes (**Table 5**). The root library contained the most tissue-specific expression but most genes were weakly expressed; the inflorescence library contained a larger number of tissue-specific transcripts at the higher expression levels, many of which encode pollen-related proteins. We also identified constantly expressed genes, defined as those with a summed abundance that was within a twofold range for each pair of the libraries. From the five libraries, 460 genes matched this criterion (**Table 5**). The complete set of genes with tissue-specific or constant expression is listed in **Supplementary Table 5**.

We determined the overlap in transcript abundance among the five libraries, calculating the proportion of signatures in any pair of libraries that were (i) present in both libraries, and (ii) present at levels not substantially different (that is, with a difference of 0.5 to 2.0 between the libraries) (**Supplementary Table 6**). Between 11,241 and 12,654 signatures were shared in each pair of libraries, and in each case, approximately half of the shared signatures were expressed at relatively consistent levels in both libraries. This was consistent with our finding described above that relatively few genes were expressed in a tissue-specific manner. Across the five libraries, the leaf-inflorescence-silique set of libraries shared more overlap with each other than with the callus or root libraries; this grouping indicated greater similarity among the photosynthetic versus nonphotosynthetic tissues (**Supplementary Table 6**). Another way to examine the overlap among libraries is to determine the sets of signatures found in one, two, three, four or all five libraries. In this analysis, the root library had the most in common with other libraries and the silique library had the least (**Supplementary Table 7a**). Based on signatures and genes found in only two of the five libraries, the nonphotosynthetic root and callus libraries overlapped the most, whereas silique and callus overlapped the least (**Supplementary Table 7b**). This pattern was maintained using sets of three libraries (**Supplementary Table 7c**).

DISCUSSION

The MPSS data indicate that a significant proportion of the plant genome is actively transcribed in any given tissue. We identified between 17,353 and 18,361 genes with sense expression, 5,487 and 8,729 genes with antisense expression, 1,256 and 1,400 expressed IGRs and at least 5,486 alternative transcripts. An additional 2,486 to 3,794 significantly expressed signatures from the five libraries could not be mapped to the genome, possibly because they span splice sites in 3′-UTRs that have yet to be identified. The combined data suggest that each of these plant tissues expresses a diverse set of more than 22,000 distinct transcripts. We believe these estimates are conservative because we focused on signatures mapping to unique sites in the genome, because we filtered the data, and because technical artifacts prevent detection of approximately 7.7% of genomic signatures[19].

Relatively few plant antisense RNAs have been fully characterized[21,23], but data from whole genome arrays and now from MPSS suggest that antisense transcription occurs extensively in *A. thaliana*[2]. The function of antisense RNAs is predicted to be regulatory; complementary sense and antisense transcripts would form double-stranded RNA (dsRNAs) molecules, which are processed, and then trigger post-transcriptional gene silencing[24]. Alternatively, nuclear dsRNA may be deaminated and retained in the nucleus, affecting the cytoplasmic concentration of the sense RNA[25]. Our observation of a large number of antisense transcripts indicates that this transcriptional regulatory mechanism is active in *A. thaliana*.

A second intriguing class of MPSS signatures was those that were matched with IGRs. These signatures may result from unrecognized protein or peptide-coding transcripts, or may correspond to ncRNAs. Gene prediction programs assume that genes have open reading frames, and therefore ncRNAs are underrepresented in most genome annotations[7]. ncRNAs may function as regulatory molecules[26] or may be processed to form microRNAs with regulatory functions[27]. Only a few ncRNAs have been identified from *A. thaliana*, although the functions of many of these are still unknown[8,28]. The use of tiled oligonucleotide and whole-genome microarrays has also revealed transcription from IGRs in *A. thaliana* and other organisms[2,29].

The MPSS data demonstrated that at least 25% of expressed *A. thaliana* genes show evidence of alternative polyadenylation. Variation in the polyadenylation site may influence gene function through post-transcriptional mechanisms. UTRs may contain regulatory elements affecting mRNA stability[30] or translation efficiency[31]; the use of alternative polyadenylation sites in the 3′-UTR may strongly affect RNA stability and therefore gene function. Differential polyadenylation has been shown repeatedly to occur in a tissue- or disease-specific manner[32]. Analyses of human ESTs have estimated that more than 50% of human genes use two or more polyadenylation sites[33]. Premature polyadenylation is used to regulate the activity of the human LINE retrotransposons[34], and regulates transcript levels of the *A. thaliana FCA* transcript[35]. Functional analyses of alternative transcripts may need to be done on a gene-by-gene basis.

We used the quantitative expression data to identify transcripts with tissue-specific and nonspecific expression patterns. With a limited number of libraries, these patterns are still crude measurements, but as new MPSS

Table 5 Tissue-specific or constantly expressed genes

Tissue	Strong (>250 TPM)	Moderate (25–250 TPM)	Low (10–25 TPM)	Very low (4–10 TPM)	Total (range in TPM)
Callus	16	54	78	187	325 (4–10,897 TPM)
Inflorescence	27	104	84	167	382 (4–8,428 TPM)
Leaves	6	21	40	109	176 (4–2,132 TPM)
Root	12	95	127	222	456 (4–1,094 TPM)
Silique	13	57	61	126	257 (4–6,297 TPM)
Constant	47	339	68	10	460 (4–1,591 TPM)

Class 1, 2, 5 and 7 signatures were summed for each *A. thaliana* gene identifier. Tissue-specific genes were then defined as those with 100-fold higher expression in one library than any of the other four; for signatures with abundances <100 TPM, the abundance in the other libraries was 0 TPM.

libraries are added from different tissues and treatments, specific expression patterns will be more precisely defined. Comparative approaches may lead to the identification of conserved regulatory sequences upstream of coding regions in genomic sequences. The similarities in transcriptional complexity across the five libraries belied the predicted differences in the tissue complexity. The most abundant transcripts and the overlapping transcripts were consistent with the biology of the tissues. Integration of biological knowledge with the *in silico* analyses of the overlap among the diverse tissues can explain many of the shared transcripts and may identify biological roles for transcripts of unknown function.

As with whole genome arrays[2], a relatively small set of diverse tissues can identify the expression of >60% of the total number of annotated genes. Detection of expression for the remaining 40% may be more difficult because it may require sampling of highly specialized tissues or treatments. Although slightly more genes were detected by the WGAs than MPSS for similar tissues, the observed overlap in patterns of transcriptional activity detected by different technologies is substantial. In addition, the MPSS analyses were more conservative because we excluded duplicate matches; these matches are easily detected using the signature sequence, whereas the microarray equivalent, cross-hybridization, is difficult to identify. Extensive experimentation with the different technology platforms ultimately may saturate the compendium of possible transcripts derived from this plant genome.

Does MPSS sample deeply enough to detect the most weakly expressed transcripts? Our analysis sampled more than 12 million transcripts, yet a large number of these were observed at low levels. Approximately 15% of the total distinct signatures were reliable but not significant, suggesting that even sampling at levels of more than 2.5 million signatures per library may not sufficiently probe the depths of transcriptional activity in some plant tissues. The lowest abundance levels detected by MPSS are below the linear range of detection for some microarrays[36]. Even though the estimates of complexity from MPSS are better than those obtained by many technologies, MPSS may not be sufficient to fully characterize transcription. A complex tissue like the inflorescence may contain many cell types; expression patterns in each cell type will increase in complexity if any sort of treatment is introduced. Our cut-off for significance limited the transcripts we detected to those present at >3 TPM, yet many signatures are present at levels below that, and many of these match the genome in unique locations[19]. These may represent rare transcripts. More restricted transcript sets may be identified by sampling specific cell types, using such selective technologies as laser-capture microscopy[37] or selectively isolated protoplasts[38]. The combination of approaches will provide further insight into the complexity of the *A. thaliana* transcriptome.

METHODS

Plant material. All plant material was from *A. thaliana thaliana*, ecotype Col-0. Callus was initiated from seeds grown on medium containing 1/2× Murashige and Skoog salts, 3% sucrose in presence of 2,4-dichlorophenoxyacetic acid (0.5 mg/l), indoleacetic acid (2 mg/l) and kinetin (0.1mg/l). Tissue was grown ~3 months in dark at 22 °C and transferred to fresh plates every ~10–14 d. For the floral library, immature inflorescences were harvested from plants grown in soil in a growth chamber with 16 h of light for 5 weeks. Floral tissue included the inflorescence meristem and early-stage floral buds (up to stage 11/12). Developing siliques were from plants grown under conditions identical to the floral library; siliques were harvested ~24–48 h after fertilization, when the petals begin to detach (stage 16–17) and the length of the siliques was 5–10 mm. The leaf and root libraries were taken from the same plants, grown in 16 h of light for 21 d under sterile conditions in vermiculite and perlite. For each library, total RNA was isolated using TRIzol (Invitrogen). For tissues derived from whole plants, samples were taken ~2 h after dark, in the subjective night.

Signature sequencing. MPSS was performed essentially as described[11,12]. Signatures for a given library were produced in multiple sequencing runs and in two types of sequencing reactions[11,19]; these sequencing runs and reactions were combined to calculate a single normalized abundance for each signature observed in each library[19]. All of our raw and normalized data are available on our website at http://mpss.udel.edu/at.

Analysis of MPSS data. We have implemented a classification scheme for the signatures that match the *A. thaliana* genomic sequence. More details may be found in a report describing our bioinformatics approaches in the application of MPSS[19]. Briefly, we extracted 858,019 potential MPSS signatures from the *A. thaliana* genomic sequence; potential MPSS signatures are derived from a *Dpn*II site (GATC) plus the adjacent 13 bases, with a second signature derived from the complementary strand. The position of each potential signature was compared to that of genes in the TIGR annotation version 3.0 (ref. 10) and assigned to a class depending on the position relative to exons and ORFs[19].

We applied two filters to the MPSS data to remove potentially erroneous signatures, and to isolate the subset of signatures that are expressed at significant levels. The first filter identifies signatures that are found in only one MPSS sequencing run across all libraries. Because our libraries consisted of at least four sequencing runs from the same tissue, this 'reliability' filter removes signatures that may be derived from random sequencing errors. The error rate for MPSS is estimated at ~0.25% per base. The filter for 'significant' signatures identifies those signatures expressed in any *A. thaliana* library at ≥4 TPM; because it is based on abundance, this criterion is independent of the reliability filter. This filter is called 'significant' because 4 TPM is different from 0 TPM with $P < 0.05$, whereas 1, 2 or 3 TPM is not significantly different from 0 TPM ($P > 0.05$). Approximately 15% of signatures in each library were observed at 1, 2 or 3 TPM but have been observed at significant levels (>3 TPM) in other experiments not described here[19]. The combination of the two filters removes the majority of erroneous signatures and identifies signatures most likely to be derived from real transcripts. Signatures that are reliable but not significant may represent weakly expressed transcripts. Because we have lower confidence in the nonsignificant signatures, we did not use these data for many of our calculations. The filters are described and evaluated in greater detail elsewhere[19].

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank Larry Tindell, Steve Edberg and Tanya Berardini for excellent technical assistance, and Hajime Sakai and Pam Green for critical reading of the manuscript. This work was supported by the National Science Foundation Plant Genome Research award #0110528 (B.C.M.).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Received 26 February; accepted 13 May 2004
Published online at http://www.nature.com/naturebiotechnology/

1. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
2. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
3. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
4. Andrews, J. *et al.* Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. *Genome Res.* **10**, 2030–2043 (2000).
5. Guigo, R., Agarwal, P., Abril, J.F., Burset, M. & Fickett, J.W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
6. Haas, B.J. *et al.* Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3**, published online 30 May 2002 (RESEARCH0029.21–0029.12, 2002).
7. Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2**, 919–929 (2001).
8. MacIntosh, G.C., Wilkerson, C. & Green, P.J. Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.* **127**, 765–776 (2001).
9. Vanhee-Brossollet, C. & Vaquero, C. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211**, 1–9 (1998).
10. Wortman, J.R. *et al.* Annotation of the *Arabidopsis* genome. *Plant Physiol.* **132**,

461–468 (2003).
11. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
12. Brenner, S. *et al.* In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* **97**, 1665–1670 (2000).
13. Adams, M.D. *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3–174 (1995).
14. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
15. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).
16. Hoth, S. *et al.* Monitoring genome-wide changes in gene expression in response to endogenous cytokinin reveals targets in *Arabidopsis thaliana*. *FEBS Lett.* **554**, 373–380 (2003).
17. Hoth, S. *et al.* Genome-wide gene expression profiling in *Arabidopsis thaliana* reveals new targets of abscisic acid and largely impaired gene regulation in the abi1–1 mutant. *J. Cell Sci.* **115**, 4891–4900 (2002).
18. Meyers, B.C., Morgante, M. & Michelmore, R.W. TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *Plant J.* **32**, 77–92 (2002).
19. Meyers, B.C. *et al.* The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis. Genome Res.* in the press (2004).
20. Lehner, B., Williams, G., Campbell, R.D. & Sanderson, C.M. Antisense transcripts in the human genome. *Trends Genet.* **18**, 63–65 (2002).
21. Terryn, N. & Rouze, P. The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci.* **5**, 394–396 (2000).
22. Xiao, Y.L., Malik, M., Whitelaw, C.A. & Town, C.D. Cloning and sequencing of cDNAs for hypothetical genes from chromosome 2 of *Arabidopsis*. *Plant Physiol.* **130**, 2118–2128 (2002).
23. Gibbings, J.G. *et al.* Global transcript analysis of rice leaf and seed using SAGE technology. *Plant Biotechnol. J.* **1**, 271–285 (2003).
24. Bass, B.L. Double-stranded RNA as a template for gene silencing. *Cell* **101**, 235–238 (2000).
25. Bass, B.L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
26. Mattick, J.S. & Gagen, M.J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **18**, 1611–1630 (2001).
27. Lee, Y., Jeon, K., Lee, J.T., Kim, S. & Kim, V.N. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* **21**, 4663–4670 (2002).
28. Marker, C. *et al.* Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr. Biol.* **12**, 2002–2013 (2002).
29. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
30. Touriol, C., Morillon, A., Gensac, M.C., Prats, H. & Prats, A.C. Expression of human fibroblast growth factor 2 mRNA is post-transcriptionally controlled by a unique destabilizing element present in the 3′-untranslated region between alternative polyadenylation sites. *J. Biol. Chem.* **274**, 21402–21408 (1999).
31. Knirsch, L. & Clerch, L.B. A region in the 3′ UTR of MnSOD RNA enhances translation of a heterologous RNA. *Biochem. Biophys. Res. Commun.* **272**, 164–168 (2000).
32. Edwalds-Gilbert, G., Veraldi, K.L. & Milcarek, C. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.* **25**, 2547–2561 (1997).
33. Iseli, C. *et al.* Long-range heterogeneity at the 3′ ends of human mRNAs. *Genome Res.* **12**, 1068–1074 (2002).
34. Perepelitsa-Belancio, V. & Deininger, P. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat. Genet.* **35**, 363–366 (2003).
35. Quesada, V., Macknight, R., Dean, C. & Simpson, G.G. Autoregulation of FCA pre-mRNA processing controls *Arabidopsis* flowering time. *EMBO J.* **22**, 3142–3152 (2003).
36. Chudin, E. *et al.* Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* **3**, published online 14 December 2001 (RESEARCH0005.1–0005.10, 2002).
37. Asano, T. *et al.* Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J.* **32**, 401–408 (2002).
38. Birnbaum, K. *et al.* A gene expression map of the *Arabidopsis* root. *Science* **302**, 1956–1960 (2003).