SUPPLEMENTAL MATERIAL

The use of MPSS for whole-genome transcriptional analysis in Arabidopsis.

Blake C. Meyers [1, 4], Shivakundan Singh Tej [1], Tam H. Vu [1], Christian D. Haudenschild [3], Vikas Agrawal [1], Steve B. Edberg [2], Hassan Ghazal [1], and Shannon Decola [3].

[1] Department of Plant and Soil Sciences, and Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19714 USA;
[2] Department of Vegetable Crops, University of California, Davis, California 95616 USA;
[3] Lynx Therapeutics, Inc., 25861 Industrial Blvd., Hayward, California 94545 USA

This document contains additional data and analyses derived from our use and handling of MPSS data and the analysis of the genomic and expressed signatures derived from Arabidopsis. The information relates to the main published text, but some of it is not directly discussed in that text.

I. Additional analyses of MPSS genomic signatures.

a) Comparison of signature duplications in Arabidopsis versus random DNA sequence.

We compared the duplication rate of signatures from the Arabidopsis genome with that of randomly generated sequence to assess the impact of genomic duplications. First, we determined the base composition of the 858,019 17-base signatures extracted from the Arabidopsis genome. This composition was calculated using only the 13 bases of the signatures that vary; the first four bases of MPSS signatures are always GATC when the MPSS sequencing is anchored by the enzyme *Dpn*II. The composition was A = 31.03%, C = 19.43%, G = 18.73%, and T = 30.81%. Next, we randomly generated 858,019 signatures with the same overall base composition as those from the Arabidopsis genome. These random signatures were duplicated at a much lower rate (~2.5% duplicated for the 17-base signatures) than those extracted from the Arabidopsis genome, and the duplication rate never exceed three “hits” among the random signatures (Supplemental Table S1). With signatures of these lengths, duplications in the Arabidopsis genome due to length will rarely occur by chance. The difference in duplication rates between the random and Arabidopsis genomic signatures is likely due to a combination of segmental duplications within the genome (Simillion et al. 2002; Vision et al. 2000), duplications of individual genes, and duplications of intergenic sequences like transposable elements or retrotransposons. The random genome contained 21,640 duplicated signatures, 14.0% of the 154,858 signatures duplicated in the Arabidopsis genome. This indicates that the vast majority of signatures with hits >1 result from duplicated genomic regions rather than coincidental occurrences of the same sequence.

We also generated random 20-base signatures using the approach described above. These 20-base signatures demonstrated only 0.05% duplication, a rate much lower than both the 17-base random signatures and the 20-base genomic signatures from Arabidopsis (Supplemental Table S1).

b) Number of genomic signatures per gene.

Arabidopsis genes generally had multiple Class 1 signatures (mean = 7.22) and at least one Class 2 signature (mean = 2.05) (Table 3). More than 82% of the 17-base Class 1 and 2 signatures were unique sequences in the Arabidopsis genome (Table 3). For normal sense-strand transcripts, 28,911 annotated Arabidopsis genes contained Class 1, 2, 5 or 7 genomic signatures that should be detectable by MPSS (Table 3). In addition to the sense-strand

transcripts, anti-sense transcripts derived from annotated genes should be readily detected by MPSS because more than 95% of annotated genes contained Class 3 or 6 genomic signatures.

c) Secondary classifications of genomic signatures.

Signatures with secondary classifications were typically the result of our standard 500 bp of 3'UTR overlapping with the open reading frame of a gene encoded on the opposite strand. A total of 8,393 signatures were assigned a secondary class as either Class 3 or Class 6 signatures (Table 3). Only 18 signatures could have been assigned a "tertiary" class, and so we did not assign this tertiary class. For signatures with a secondary classification, we defined a precedence ranking. This precedence ranking corresponds to our estimate of the likelihood that a transcript occurs on the sense or antisense strand compared to an annotated gene. The ranking is as follows: Class 1 or 7 > Class 2 > Class 5 > Class 3 > Class 6 > Class 4, with the primary class assigned based on the higher ranking and the secondary class assigned based on the lower ranking. For example, a signature that could be considered Class 1 (compared to an exon on the same strand) or Class 3 (compared to an exon on the opposite strand) would be assigned Class 1 as the primary ranking and Class 3 as a secondary ranking. In practice, the class stored as the secondary rank was rarely used in our calculations. However, this rule was important when annotations overlapped because the primary class was used in all of our calculations.

II. Additional analyses of MPSS expressed signatures.

a) Evaluation of "reliable" and "significance" filters.

The assumption that genomic matches more likely represent "real" transcripts, and non-matched signatures are potentially erroneous can be used to evaluate the results of the other filters. For example, a total of 67,735 of 87,705 (77.2%) "significant" and "reliable" expressed signatures were matched to genomic signatures, but only 29,288 of 112,263 (26.1%) "non-significant" and "unreliable" expressed MPSS signatures matched the genome (Figure 2). The match rate that is approximately three-fold higher for significant and reliable than for non-significant and unreliable signatures suggests that these filters successfully screened signatures that may be erroneous. Most unreliable signatures were also not significant, and the unreliable signatures had a low rate of matches to the genome (Figure 2). This suggests that the reliability filter may more accurately identify spurious signatures than the significance filter. Using the rate of genomic matches as an indicator of signatures derived from real transcripts, the "reliability" filter was more effective than the "significance" filter at removing noisy signatures. The combination of the two filters and the genome comparison should provide an accurate screen to distinguish the "signal" (e.g. signatures from real transcripts) from "noise (e.g. spurious signatures). And therefore, the genomic comparison functions much as a filter like the "reliability and "significance" filters.

b) Evaluation of Class 2 "window" size using expression data.

Because the size of the Class 2 "window" was set at 500 nucleotides 3' of the stop codon for each gene, we used the MPSS expression data to determine if a smaller 3' UTR would miss many expressed signatures. Based on comparisons of the available full-length cDNA sequences from Arabidopsis, the average 3' UTR after splicing is 235 bp (Haas et al. 2003) with a standard deviation of 154 bp (M. Ayele, B. Haas, C. Town, personal communication). Therefore, 500 bp is nearly two standard deviations beyond the mean, and therefore should include the vast majority of all 3' UTRs. In our signature classification system, reducing the window size will coordinately decrease the number of Class 2 signatures and increase the number of Class 4 signatures (Supplemental Table S2). We compared the number of Class 2 and Class 4 genomic and expressed signatures with variable window sizes (Supplemental Table S2). Most 100 bp decreases in the Class 2 window size shifted ~400 to

~700 expressed signatures from Class 2 to Class 4; the decrease from a 200 bp to a 100 bp Class 2 window would dissociate 1,792 expressed signatures from annotated genes. Increasing our Class 2 window to 600 bp may more accurately map several hundred signatures currently designated as Class 4. Even larger sizes may falsely join transcripts derived from small 3'-adjacent transcripts. A planned improvement to our system would determine and adjust as necessary the 3' UTR based on the maximum distance to the poly(A) site observed among full-length cDNA sequences. This could be adjusted for each gene if the experimentally-defined 3' UTR exceeds our predefined Class 2 window size.

c) Matches between MPSS signatures and cDNA sequences or plastid genomes.

Potential transcripts from Arabidopsis other than those annotated in the five nuclear chromosomes were matched to the set of Class 0 signatures. Our initial investigation did not include the chloroplast or mitochondrial genomes, nor did it include cDNA sequences that had not been incorporated into the TIGR annotation. To determine if we might have missed matches to the Class 0 signatures by focusing on the sequence and annotation from the five nuclear chromosomes, 17-base potential signatures were extracted from the following Arabidopsis sequences: (1) the 154,478 bp chloroplast genome (GenBank AP000423); (2) the 366,923 bp mitochondrial genome (GenBank NC_001284); and (3) a set of 226,370 ESTs and full-length cDNAs present in GenBank on October 27, 2003. Genomic signatures extracted from both strands of the plastid genomes totaled 1,432 in the chloroplast and 3,318 in the mitochondrial genome. From the combined set of cDNAs and ESTs, 588,950 potential signatures were extracted, although most of this group would already have been identified among our existing Class 1 to 7 signatures. The three sets of potential signatures matched to a total of 2,341 of the 128,939 17-base Class 0 signatures (Supplemental Table S3). In each case, the rate of hits was highest among the significant and reliable Class 0 signatures (Supplemental Table S3. These data were consistent with the hypothesis that many of the higher-abundance Class 0 signatures may represent bona fide transcripts and many of the lower abundance signatures may represent sequencing errors.

d) Evaluation of Class 0 signatures.

To assess sequencing errors, the unmatched Class 0 signatures were compared to significantly-expressed genomic signatures, allowing one polymorphic base in a match. The first four bases of every signature are GATC and are invariant; the remaining 13 positions of the 17-base MPSS signatures could contain errors. For this comparison, we derived a set of signatures one base different ("OBD") from the 126,598 17-base Class 0 signatures not perfectly matched to the genome or plastid sequences in the analyses described above and in the main text. To generate the OBD signatures, each position was changed to one of three other bases, so a total of 3 x 13 = 39 OBD signatures were derived from each Class 0 signature (Supplemental Figure S1). This calculation resulted in a table of 4,937,322 Class 0 OBD signatures that were then compared with the Class 1 to 7 signatures (Supplemental Figure S1). The majority of the Class 1 to 7 signatures had no OBD matches (Supplemental Figure S2). Out of 67,735 significant and reliable signatures that were expressed and matched the genome, 28,414 (41.2%) were matched to OBD signatures corresponding to 73,815 of the Class 0 signatures (58.3%) (Supplemental Figure S2). We also performed this analysis for two-base-different ("TBD") signatures, but allowing two bases creates a set of 702 possible signatures. This set of TBD signatures was much more promiscuous than the OBD matches (Supplemental Figure S2). We had less confidence in the TBD matches because of their high level of degeneracy; relaxing two of the 13 variable bases produces 411 possible signatures and some false matches are likely to occur.

To estimate the rate of false matches that might have occurred between the OBD and 67,735 expressed genomic signatures, we also calculated the rate of OBD matches between

the 126,598 Class 0 signatures and the 614,701 genomic signatures for which no expression data were found in the 14 MPSS libraries. Allowing one polymorphic base, 79,134 of the unexpressed genomic signatures (12.8%) were matched by 54,623 Class 0 signatures (43.1%). If the comparison is restricted to only 67,735 unexpressed genomic signatures (to match the number of expressed significant and reliable signatures), only 4.8% of the Class 0 signatures matched. This false-positive rate represents less than 10% of the total OBD matches to significant and reliable expressed signatures described above. Therefore, the vast majority of the 58.3% of Class 0 signatures matched via an OBD signature are correctly matched expressed signatures.

Sequencing errors in MPSS apparently produce a large number of distinct signatures that are found at very low abundances and were usually observed in only one of the 63 runs in our database. The expression data that is lost due to sequencing errors could be re-assigned based on matching OBD Class 0 signatures to genomic signatures. More than 58% percent of the unassigned, Class 0 signatures were only one base different from expressed signatures matched to the Arabidopsis genome. We have considered the possibility of adjusting the abundance level of the expressed Class 1 to 7 signatures that match the Class 0 with one base difference. This could be performed to take into account OBD matches to multiple expressed signatures with different sequences, using a weighting system for re-distribution of the OBD abundance based on the different abundance of each matching Class 1 to 7 signature. For a one-to-one match between a Class 0 OBD signature and an OBD Class 1 to 7 signature, the adjustment would merely add the Class 0 expression value to that of the Class 1 to 7 signature. However, the sequencing error should occur at a constant rate for all bases, and the net reduction of expression levels should occur at a consistent rate across all signatures. Reassigning the Class 0 data may in fact introduce more noise, due to a one-to-many relationship between the ODB Class 0 and the Class 1 to 7 signatures. Therefore, for a systemic noise issue like sequencing error, it is best to devise filters such as we have described that can remove the erroneous signatures and leave behind signatures with a higher confidence value.

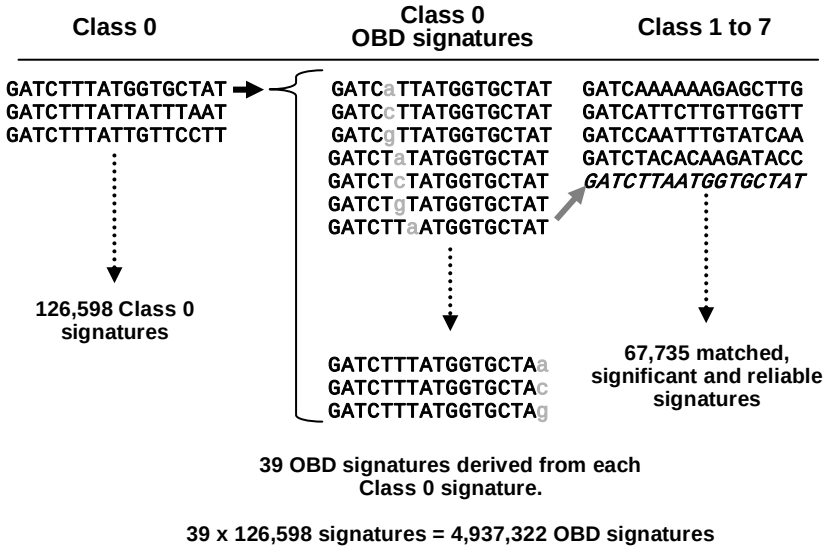
III. Additional analyses of palindromic and "bad" words.

We analyzed in more detail the effect of the 16 palindromic words in the MPSS expression data by comparing the representation of these words in genomic signatures to the representation of these words in expressed signatures (Supplemental Tables S4A, S4B and S4C). The number of palindromic four-base words was calculated for frames 1 to 8 (Figure 5A) for 17-base genomic signatures with hits = 1. The 18^{th} to 20^{th} bases are required to analyze frames 7 and 8, but do not appear in the 17-base MPSS signature; therefore, a 20-base genomic signature is required for complete assessment of palindromic words occurring in 17-base MPSS signatures. A total of 41,048 genomic signatures contained a palindrome that would affect both 2- and 4-step MPSS reactions, representing 5.2% of the distinct genomic signatures that were analyzed (Supplemental Table 4A). This observed frequency is the same as the 5.17% of signatures predicted to contain a palindrome in both steppers (see Methods). We then matched 127,982 expressed 17-base signatures to the 750,792 distinct, unique 20-base genomic signatures. The fraction of signatures that was expressed was roughly similar for signatures with one, two or three palindromes as long as one of the steppers was free of palindromic words (Supplemental Table 4A compared to 4B or 4C). However, the proportion of expressed signatures was much lower when both steppers included at least one palindromic word (Supplemental Tables 4B and 4C). Genomic signatures with palindromes in both the 2- and 4-step frames were poorly represented in the set of expressed signatures (Supplemental Table 4B and 4C). The mean abundance count for the expressed signatures was significantly lower in a stepper that contained a palindrome than in a stepper that did not contain the palindrome (Supplemental Figures 3A and 3B). The suppressing effect of palindromes was

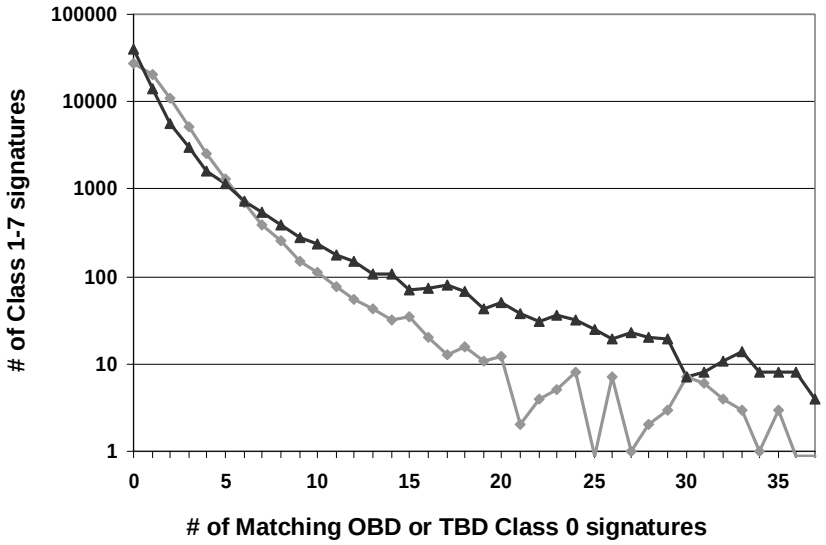
similar in the 2- and 4-step reactions. The presence of two palindromic words in the same stepper more strongly reduced the raw abundance (Supplemental Figures 3A and 3B) and had a greater overall suppressing effect than one palindromic word (Supplemental Table 4B and 4C). There are two more frames in the 20-base expressed signatures compared to the calculations performed for 17-base MPSS data, and one more frame to analyze from the 22-base genomic signatures that correspond to the 20-base expressed signatures (Figure 5B). However, similar results were obtained using the 20-base expressed signatures and 22-base genomic signatures as compared to those described for the 17-base expressed signatures and 20-base genomic signatures (data not shown).

SUPPLEMENTAL FIGURES AND TABLES

Supplemental Figure S1. One-base difference matching between Class 0 and Class 1 to 7 signatures.
An example set of Class 0 signatures is shown at left. A total of 39 one-base different ("OBD") signatures can be derived from the 13 mutable bases in this signature; examples of these are shown in the center with the one changed base indicated with gray, lowercase letters. Matching Class 1 to 7 signatures are selected by comparing the list of all possible Class 0 OBD signatures with the subset of genome-matched signatures. Dotted lines indicate the list has been abbreviated for illustrative purposes, but the complete size of the list is enumerated.



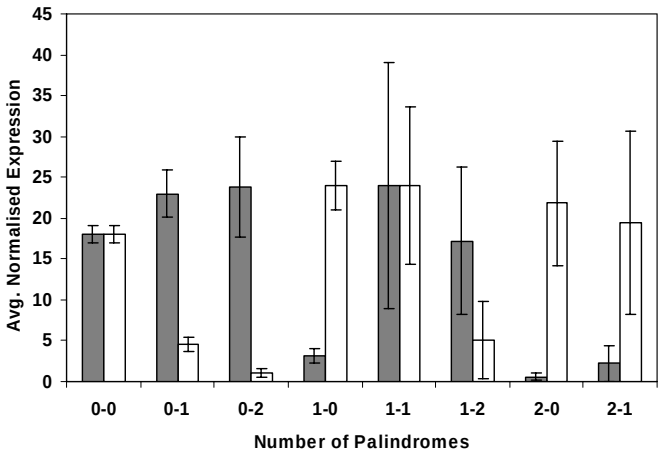
Supplemental Figure S2. One- or two-base difference matches to Class 1 to 7 signatures. The 67,735 genome-matched, significant and reliable signatures were compared to the 126,598 Class 0 signatures that remained after removing matches to the chloroplast, mitochondrial and cDNA sequences. The number of one-base different ("OBD") signatures that match between the Class 0 signatures and the Class 1 to 7 signatures is indicated in black. The number of two-base different ("TBD") signatures is indicated in gray. The X-axis corresponds to the number of OBD or TBD matches for each Class 1 to 7 signature. The Y-axis indicates the number Class 1 to 7 signatures that have OBD or TBD matches to the Class 0 signatures.



Supplemental Figure S3. Signatures containing palindromes were significantly under-represented in expression data.
A. Plot of the mean expression level for t_norm or f_norm (see Methods) for signatures with zero, one or two palindromes in either or both of the 2- and 4-step sequencing reactions. Signatures containing two or more palindromes in both steppers were expressed extremely rarely and are not shown (see Supplemental Table 4B). On the X-axis, the number of palindromes in each stepper are listed, with the number in the 2-step frames separated by a dash from the number in the 4-step frames; gray bars indicate 2-step abundances, and open bars indicate 4-step abundances. On the Y-axis, the mean expression level (in TPM) is shown for each set of signatures and steppers indicated on the X-axis. The error bars represent 95% confidence intervals. Palindromes were determined for 20-base genomic signatures that uniquely correspond to 17-base expressed signatures considered significant and reliable.

B. Statistics for values plotted in **A**. The first two columns contain the number of palindromes in each stepper. "N" indicates the number of distinct signatures in each set for which the mean expression level and the standard deviation are indicated; the mean is plotted in **A**. The P-value indicates the significance of the difference between the 2-step and 4-step mean expression levels. The P-value was calculated using a two-sample T-test.

A



B

# of palindromes			Mean ± standard deviation		P-value
2-step	4-step	N	2-step	4-step	
0	0	45,598	18.6 ± 121.5	18.4 ± 117.7	0.859
0	1	7,383	23.6 ± 133.3	4.7 ± 40.8	0.000
1	0	7,220	3.2 ± 37.7	25.1± 130.4	0.000
0	2	570	24.4 ± 77.4	1.1 ± 7.3	0.000
2	0	501	0.5 ± 4.9	21.9 ± 90.8	0.000
1	1	452	29.3± 204.2	28.7 ± 129.9	0.959
1	2	24	18.4 ± 24.1	5.6± 12.6	0.026
2	1	28	2.3 ± 7.4	23.0± 36.5	0.006

Supplemental Figures S4 – S9. Complete set of abundances for four-base words in expressed signatures.

The Excel spreadsheet contains six worksheets. For both the 17-base and 20-base expressed signatures, three worksheets were constructed. These worksheets contain the frequency of occurrence for 255 four-base words in a subset of the expressed signatures. The signatures that comprise each subsets or “Bin” are described in the Methods section. The data are the complete set from which Tables 5A and 5B are derived and the column headings are explained in more detail in those tables. Each worksheet contains a plot of the ordered ratios for 2-step versus 4-step frequency counts. The four-base word GATC was not considered because it does not appear among the expressed signatures.

Supplemental Table S1. Duplications of genomic signatures.

Modified from Table 2 to include the signatures for a random genome (see text).

	17-base signatures					20-base signatures				
"hits" [a]	Total locations [b]	% of total	# distinct [c]	% of distinct	Random	Total locations [b]	% of total	# distinct	% of distinct	Random
1	703,161	81.95	703,161	93.27	836,379	750,792	87.50	750,792	95.91	857,573
2	74,630	8.70	37,315	4.95	21,160	43,664	5.09	21,832	2.79	446
3	19,452	2.27	6,484	0.86	480	13,437	1.57	4,479	0.57	0
4	9,528	1.11	2,382	0.32	0	7,700	0.9	1,925	0.25	0
5	6,545	0.76	1,309	0.17	0	5,345	0.62	1,069	0.14	0
6	4,344	0.51	724	0.10	0	3,576	0.42	596	0.08	0
7	3,066	0.36	438	0.06	0	2,401	0.28	343	0.04	0
8	2,480	0.29	310	0.04	0	2,024	0.24	253	0.03	0
9	2,106	0.25	234	0.03	0	1,845	0.22	205	0.03	0
10	1,980	0.23	198	0.03	0	1,770	0.21	177	0.02	0
11-20	12,932	1.51	893	0.12	0	11,521	1.34	801	0.1	0
21-30	5,622	0.66	229	0.03	0	4,806	0.56	195	0.02	0
31-50	4,853	0.57	127	0.02	0	3,796	0.44	99	0.01	0
> 50	7,320	0.85	90	0.01	0	5,342	0.62	68	0.01	0
Total	858,019		753,894		858,019	858,019		782,834		858,019

[a] "Hits" refers to the total number of occurrences in the genome.

[b] Includes both spliced and unspliced versions of the 6,807 signatures that span the intron/exon boundaries (see Table 3).

[c] "Distinct" refers to the number of different sequences found within the set.

Supplemental Table S2. Effect of variable window size for Class 2 signatures.

Window size	Class 2 genomic [a]	Class 2 expressed [b]		Class 4 genomic [a]	Class 4 expressed [b]	
		Hits = 1	Hits ≥ 1		Hits = 1	Hits ≥ 1
100 bp	8,690	2,771	3,116	323,988	9,703	16,141
200 bp	17,415	4,563	5,146	315,263	7,911	14,111
250 bp	21,906	5,126	5,815	310,772	7,348	13,442
300 bp	26,327	5,538	6,309	306,351	6,936	12,948
400 bp	35,391	6,248	7,202	297,287	6,226	12,055
500 bp	44,536	6,848	7,968	288,142	5,626	11,289

[a] Calculated for genomic signatures with hits = 1.

[b] Expressed signatures were significant and reliable. Numbers for hits ≥ 1 count all genomic sites at which expressed signatures were found, and does not refer to the number of distinct signatures.

Supplemental Table S3. Class 0 matches to plastid genomes.

Filter results	Total signatures [a]	Chloroplast matches (ct)	Mitochondrial matches (mt)	cDNA matches	Union of ct, mt, cDNA matches [b]
Reliable, Significant	19,970	401	110	917	1,191
Unreliable, Significant	12,328	5	3	64	68
Reliable, Non-significant	40,125	110	78	418	550
Unreliable, Non-significant	82,975	57	47	472	532

[a] These numbers indicate only Class 0 signatures (without genomic matches), as shown in Figure 2.
[b] The union indicates the total number of distinct signatures matched in any of the three comparisons.

Supplemental Table S4. Palindrome analysis of genomic and expressed signatures.

A. Palindromes in unique genomic signatures.

# of 2-step palindromes	# of 4-step palindromes				
	0	1	2	3	4
0	470,520	111,080	10,256	450	3
1	109,148	*30,967*	*3,818*	*265*	*4*
2	9,464	*3,389*	*679*	*93*	*4*
3	376	*188*	*60*	*17*	*1*
4	5	*3*	*1*	*1*	*0*

A total of 750,792 distinct and unique 20-base genomic signatures were analyzed for palindromes. Numbers in italics indicate those signatures that have at least one palindrome in both the 2- and 4-step MPSS reactions.

B. Palindromes in expressed signatures matching uniquely in the genome.

# of 2-step palindromes	# of 4-step palindromes				
	0	1	2	3	4
0	92,932 (19.8%)	15,255 (13.7%)	1,179 (11.5%)	55 (12.2%)	0 (0%)
1	16,190 (14.8%)	*1,023 (3.3%)*	*46 (1.2%)*	*2 (0.8%)*	*0 (0%)*
2	1,188 (12.6%)	*61 (1.8%)*	*3 (0.4%)*	*0 (0%)*	*0 (0%)*
3	45 (12.0%)	*2 (1.1%)*	*1 (1.7%)*	*0 (0%)*	*0 (0%)*
4	0 (0%)	*0 (0%)*	*0 (0%)*	*0 (0%)*	*0 (0%)*

Numbers in parentheses indicate the proportion of the distinct and unique 20-base genomic signatures from part A that were uniquely matched by 17-base expressed signatures.

C. Palindromes in expressed, significant and reliable signatures.

# of 2-step palindromes	# of 4-step palindromes				
	0	1	2	3	4
0	45,598 (9.7%)	7,383 (6.6%)	570 (5.6%)	30 (6.7%)	0 (0%)
1	7,220 (6.6%)	*452 (1.5%)*	*24 (0.6%)*	*1 (0.4%)*	*0 (0%)*
2	501 (5.3%)	*28 (0.8%)*	*3 (0.4)*	*0 (0%)*	*0 (0%)*
3	20 (5.3%)	*0 (0%)*	*1 (1.7%)*	*0 (0%)*	*0 (0%)*
4	0 (0%)	*0 (0%)*	*0 (0%)*	*0 (0%)*	*0 (0%)*

From 14 libraries, we considered 61,831 distinct signatures that are unique in the genome (hits = 1) and were expressed at significant and reliable levels.