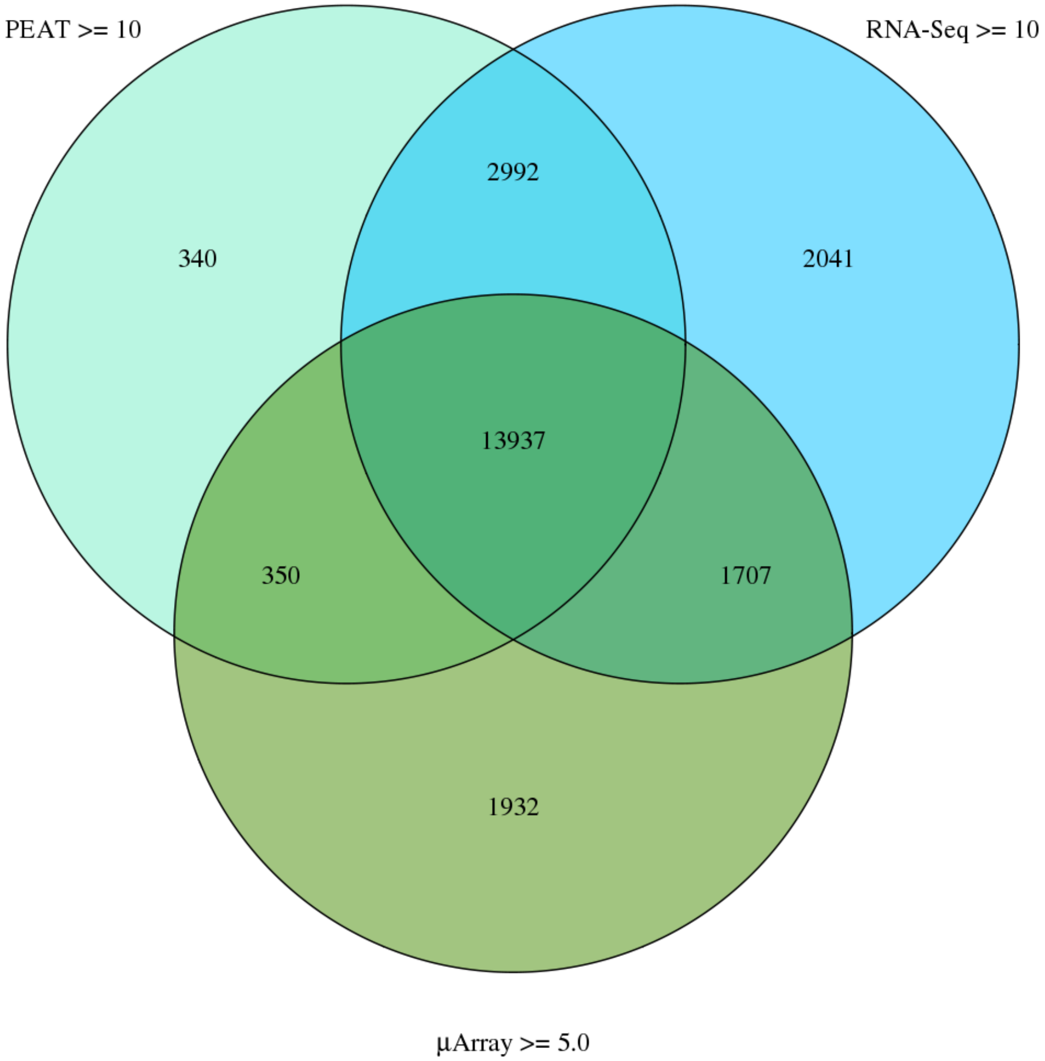
Table of Contents

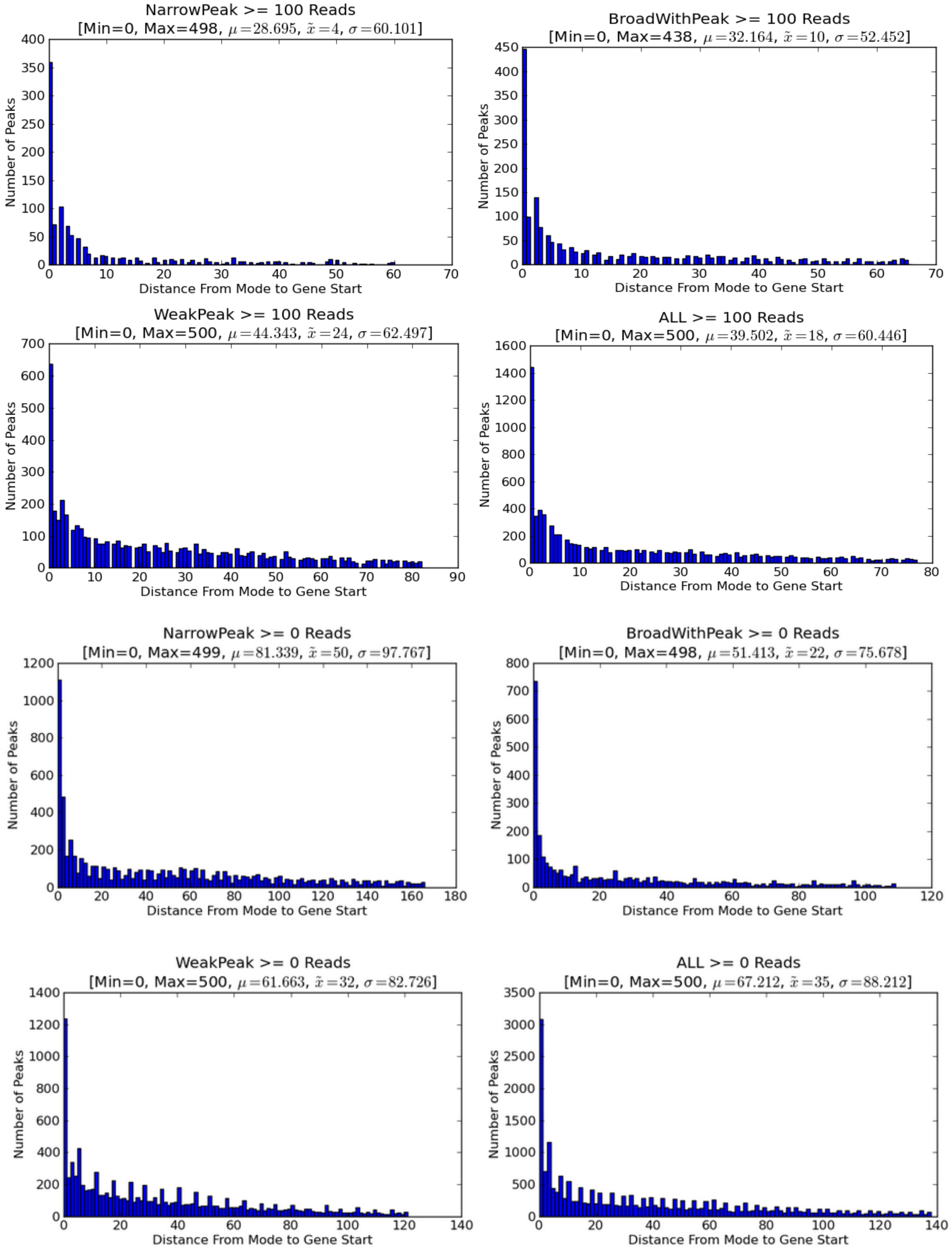
Supplemental Figures

Supplemental Figure 1: RNA-Seq gene expression comparison.



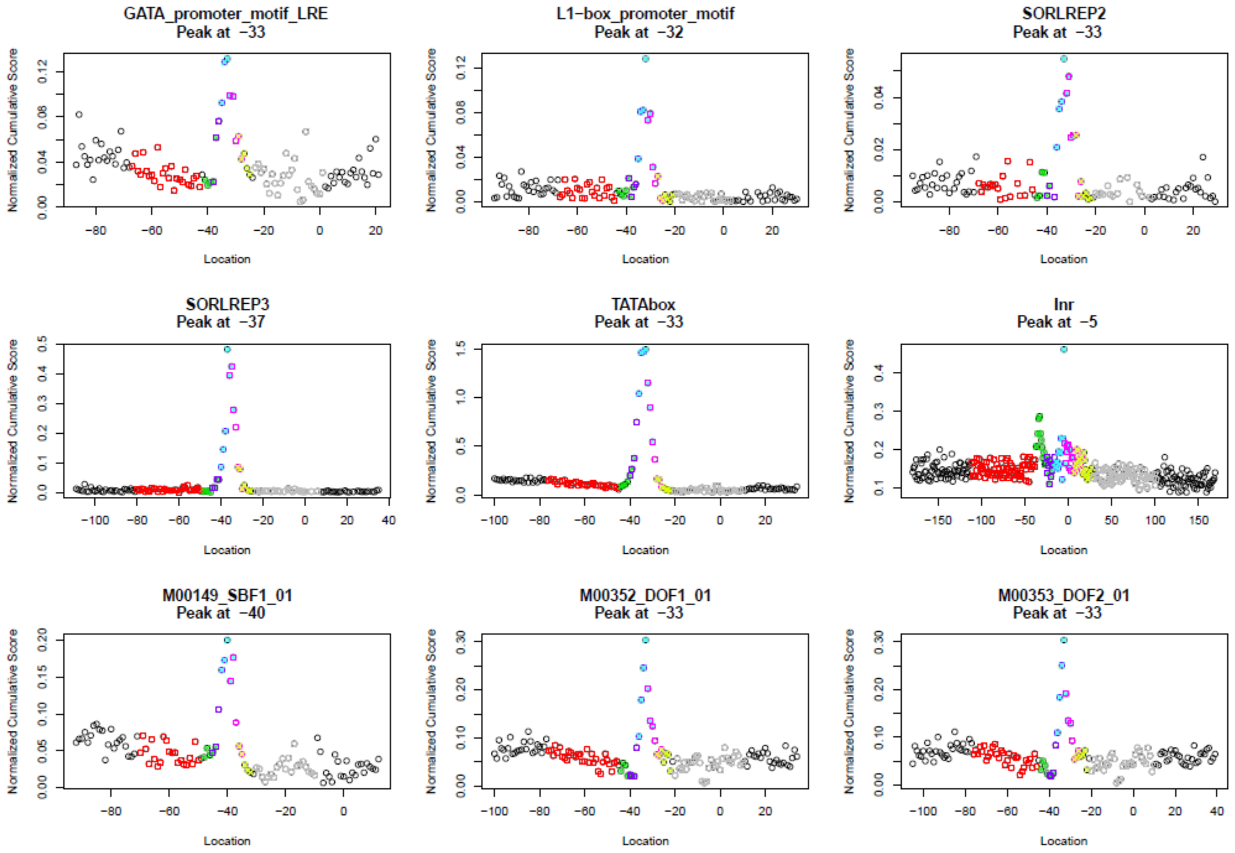
Supp. Figure 1: To assess how well the genes expressed in our PEAT pooled root sample matched those expressed in previous studies (Brady et al., 2007; Li et al., 2013), we compared the number of mutually expressed genes in our PEAT data and two published microarray and RNA-Seq datasets in Arabidopsis root. A gene was considered expressed in the RNA-Seq dataset if a total of 10 or more reads mapped to the gene across all replicates. Genes in the microarray dataset were conservatively considered expressed if the Robust Multi-array Average (RMA) normalization of any root tissue type was greater than 5.0. A gene was considered expressed in the PEAT dataset if any tag cluster containing more than 10 reads was associated with the gene. This analysis shows that overall, our data agrees quite well with previously published studies, with a majority of genes being expressed in all three datasets.

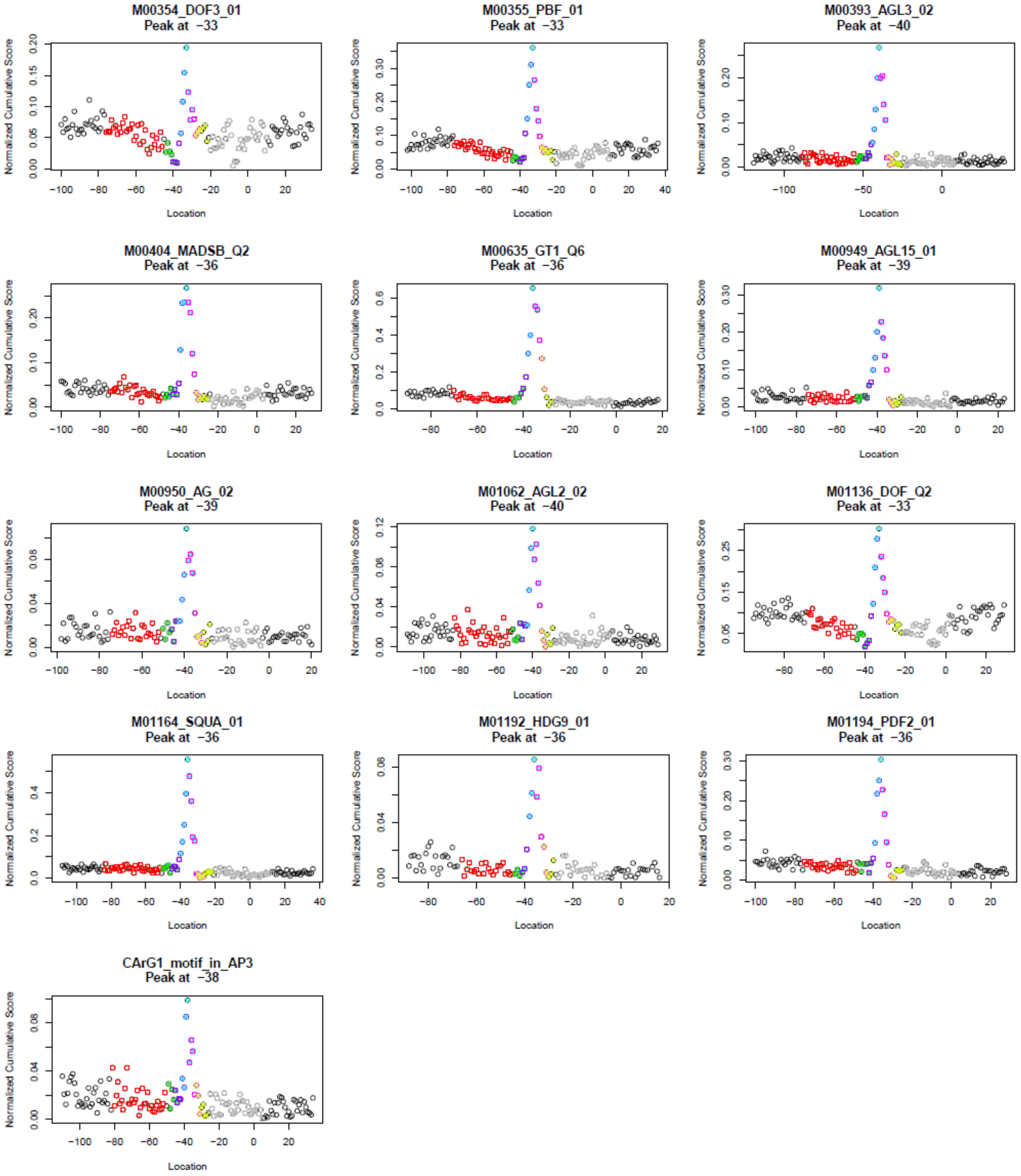
Supplemental Figure 2: Distance of PEAT peak modes to TAIR10 annotated TSSs.



Supp. Figure 2: Histograms displaying the distance of PEAT tag clusters (by shape) located no more than 500 nt upstream of an annotated TAIR10 protein coding gene. When only highly-expressed tag clusters (those with at least 100 reads) are considered, PEAT tag clusters—particularly the NP initiation pattern—are located very closely to TAIR10 annotations (50% of NP reads are within 4 nt of the gene's annotated start site).

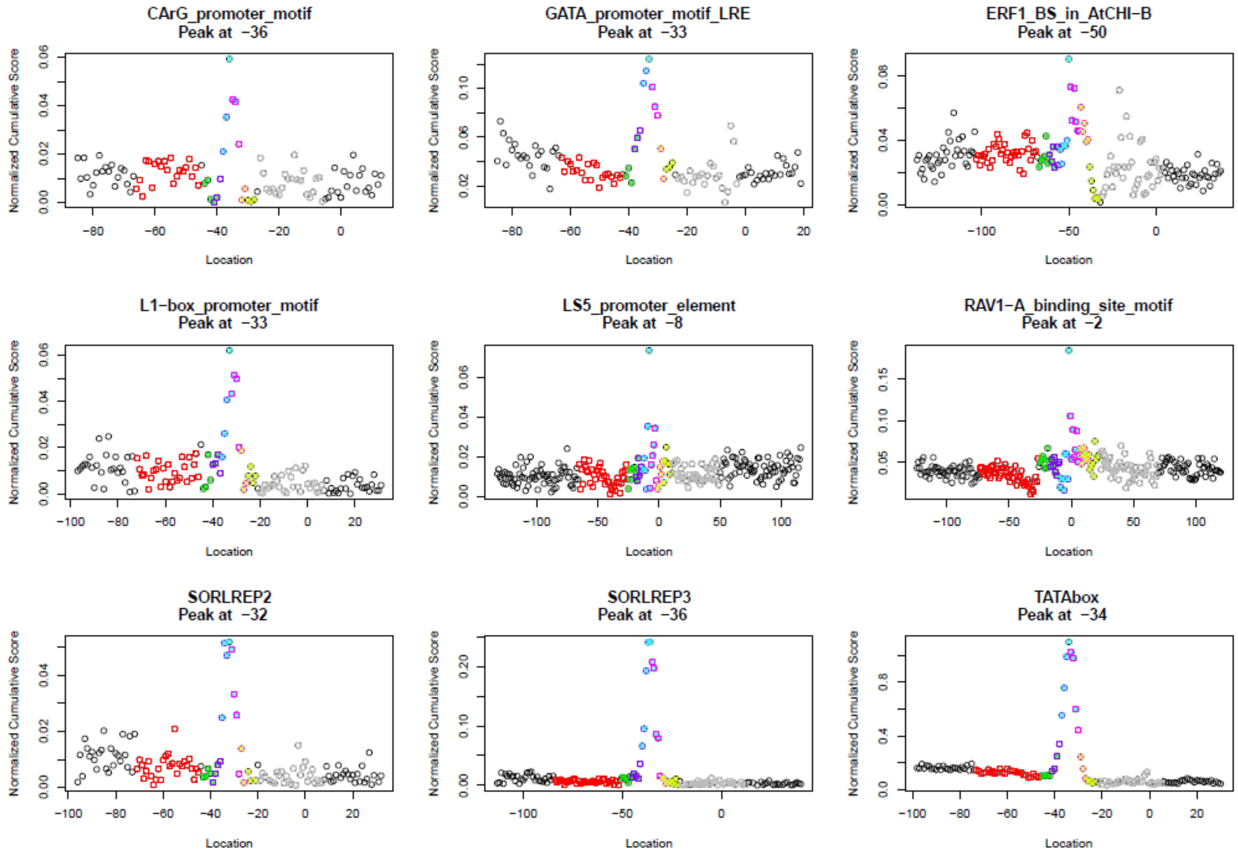
Supplemental Figure 3: Sharp, well defined Regions of Enrichment for NP promoters.



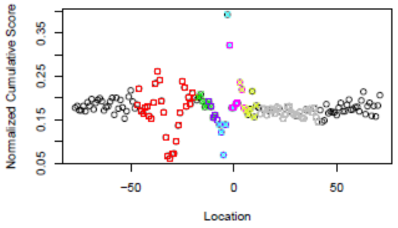


Supp. Figure 3: Sharp, well-defined Regions of Enrichment (ROEs) defined by PEAT Narrow Peak dataset. Peaks represent promoter regions which show strong, position-specific enrichment for TFBS sequences relative to strong TSSs across all training set PEAT Narrow Peak promoters and provide evidence of position-specific TFBS elements. Colored regions are those detected as portions of the enrichment and flanking area.

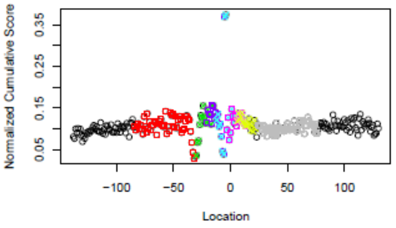
Supplemental Figure 4: Sharp, well defined Regions of Enrichment for BP promoters.



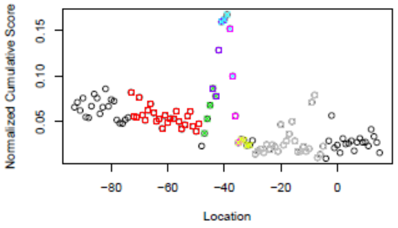
CAATbox
Peak at −3



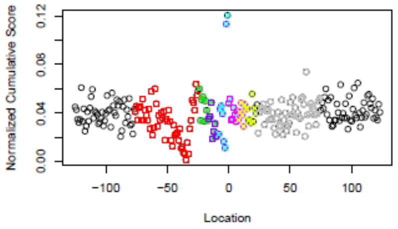
M00313_GEN_INI2_B
Peak at −4



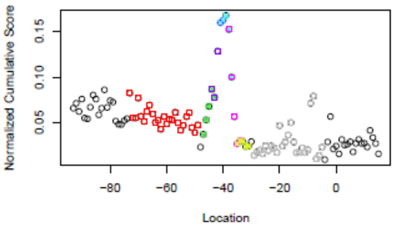
M00149_SBF1_01
Peak at −39



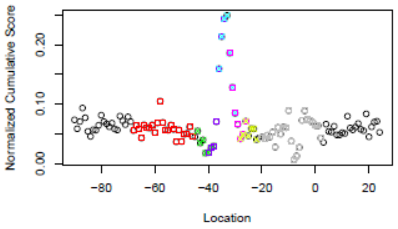
M00345_GAMYB_01
Peak at −1



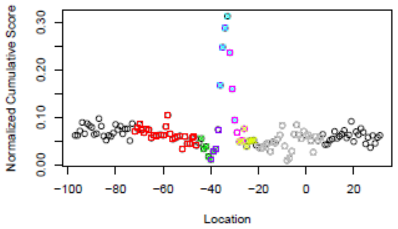
M00149_SBF1_01
Peak at −39



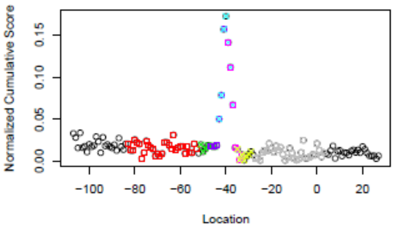
M00353_DOF2_01
Peak at −33



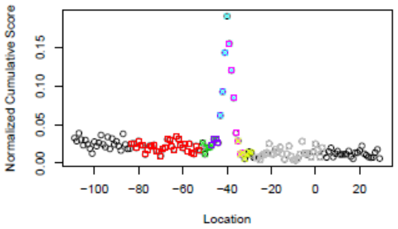
M00355_PBF_01
Peak at −33



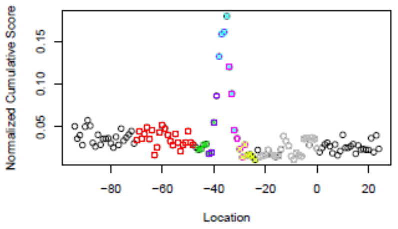
M00392_AGL3_01
Peak at −40



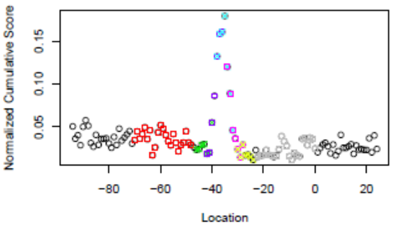
M00393_AGL3_02
Peak at −40



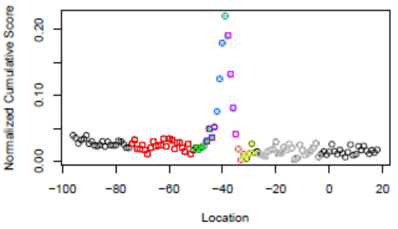
M00404_MADSB_Q2
Peak at −35



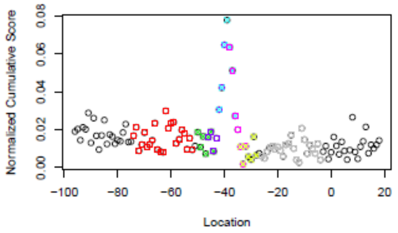
M00404_MADSB_Q2
Peak at −35



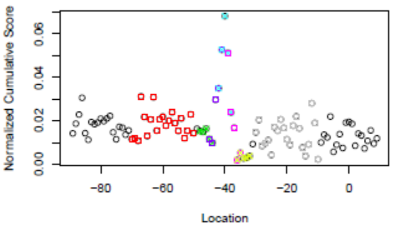
M00949_AGL15_01
Peak at −39



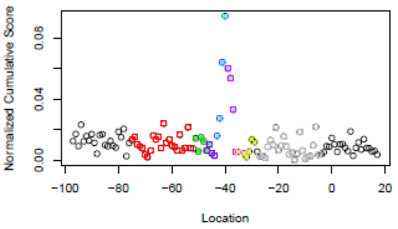
M00950_AG_02
Peak at −39



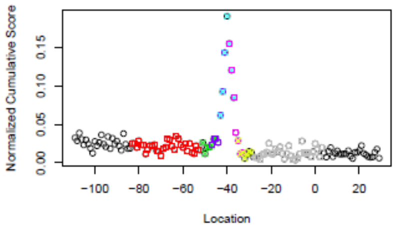
M01059_AGL1_01
Peak at −40



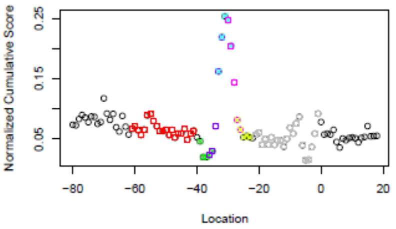
M01061_AGL2_01
Peak at −40



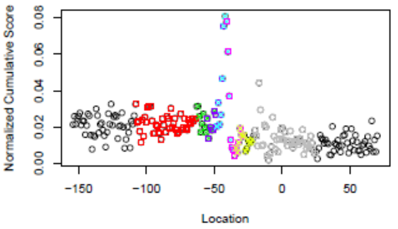
M01064_AGL3_03
Peak at −40

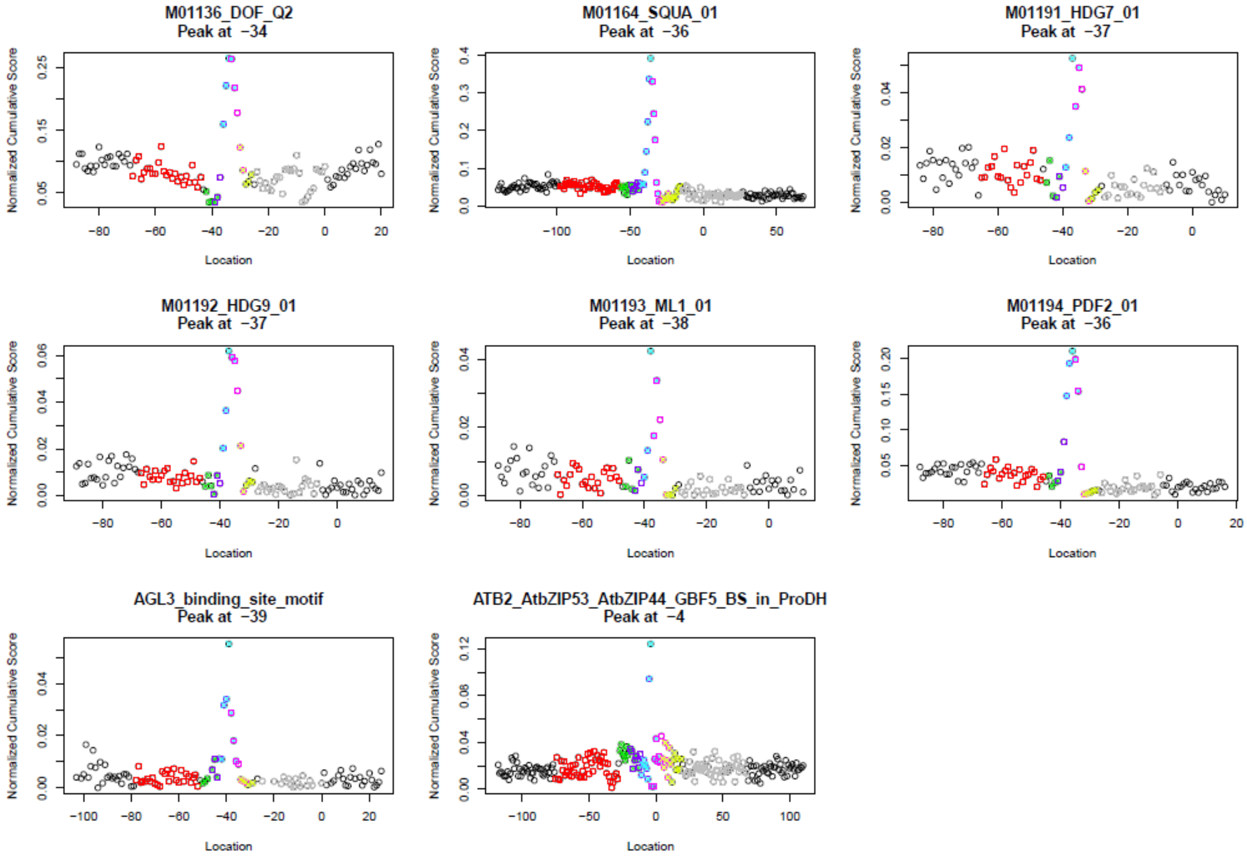


M01130_PBF_Q2
Peak at −31



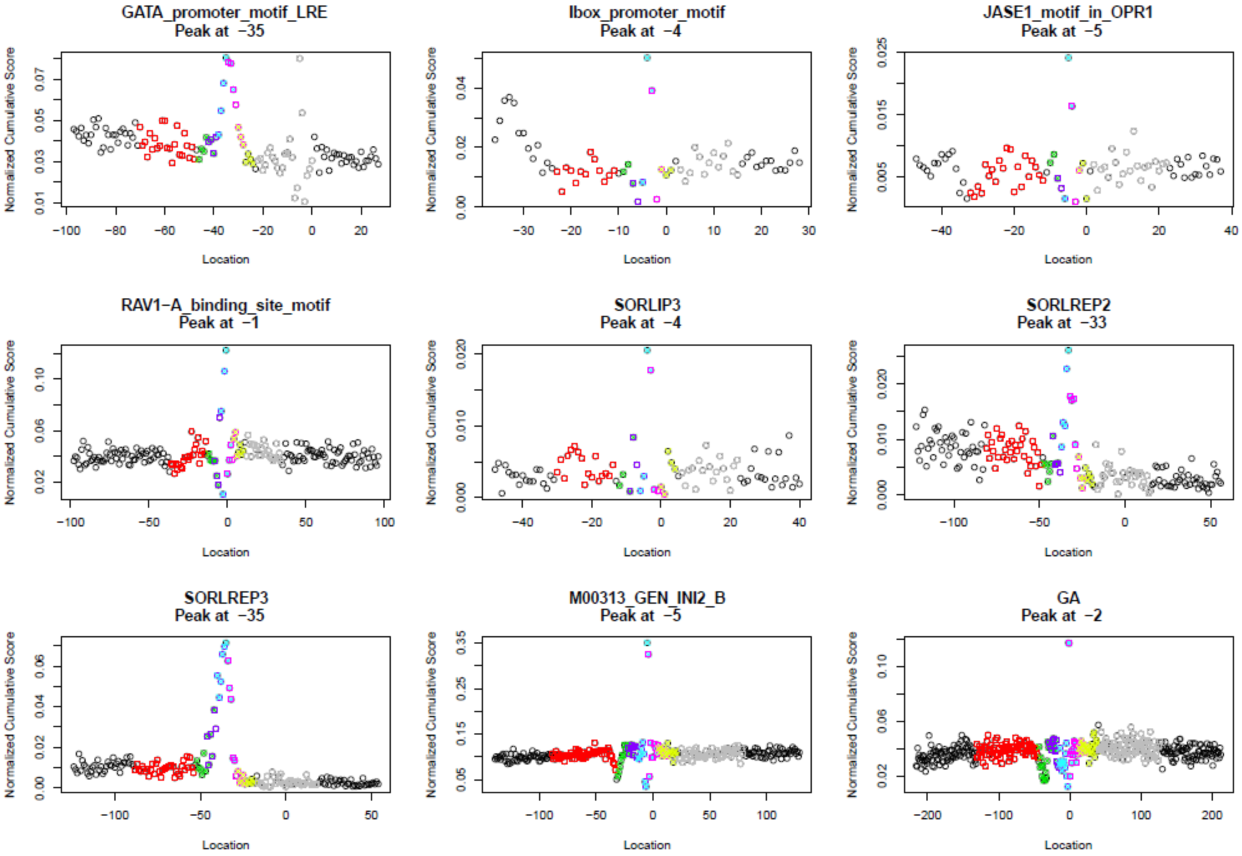
M01133_AG_Q2
Peak at −42

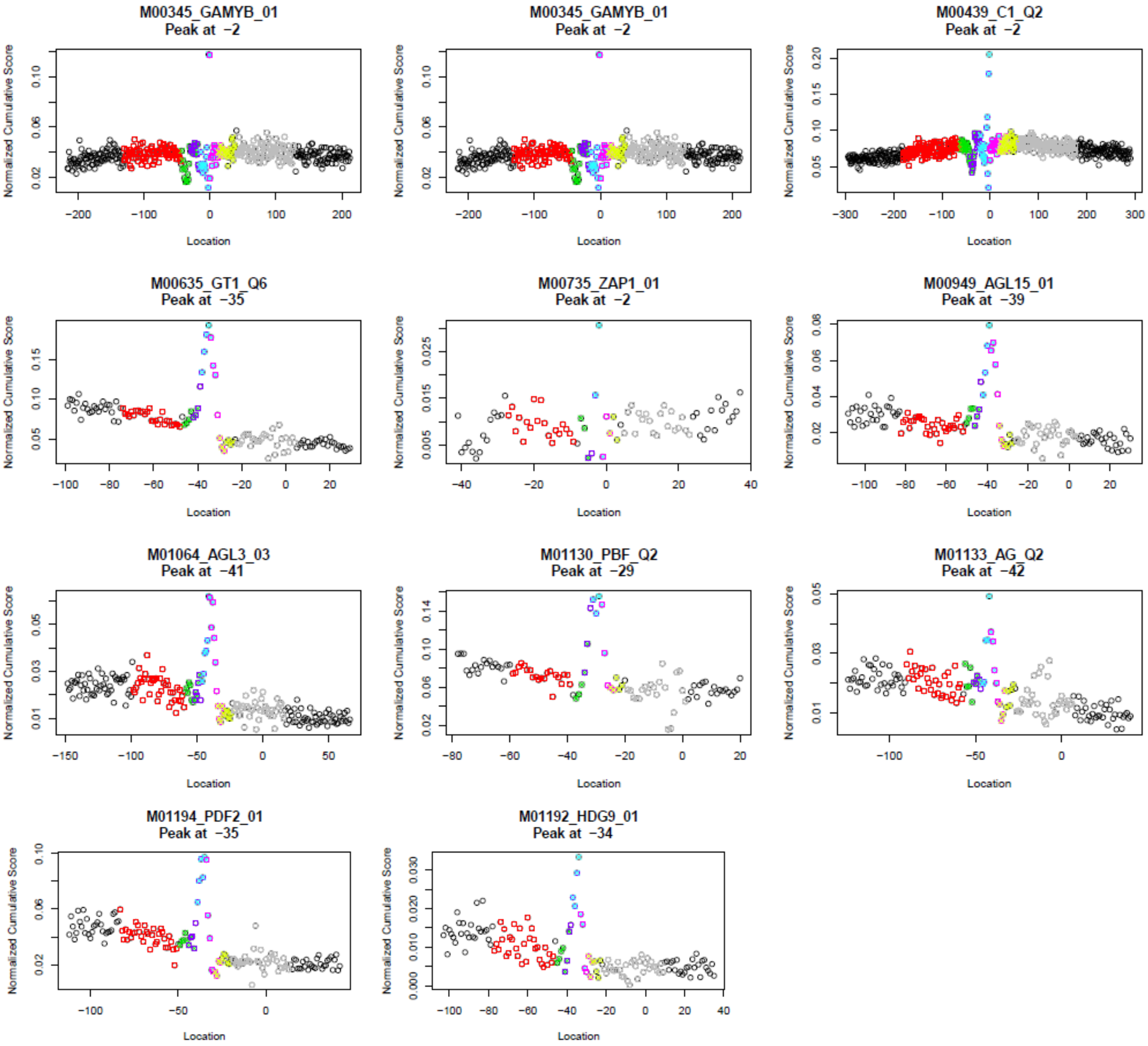




Supp. Figure 4: Sharp, well-defined Regions of Enrichment (ROEs) defined by PEAT tag clusters with Broad with Peak transcription initiation patterns.

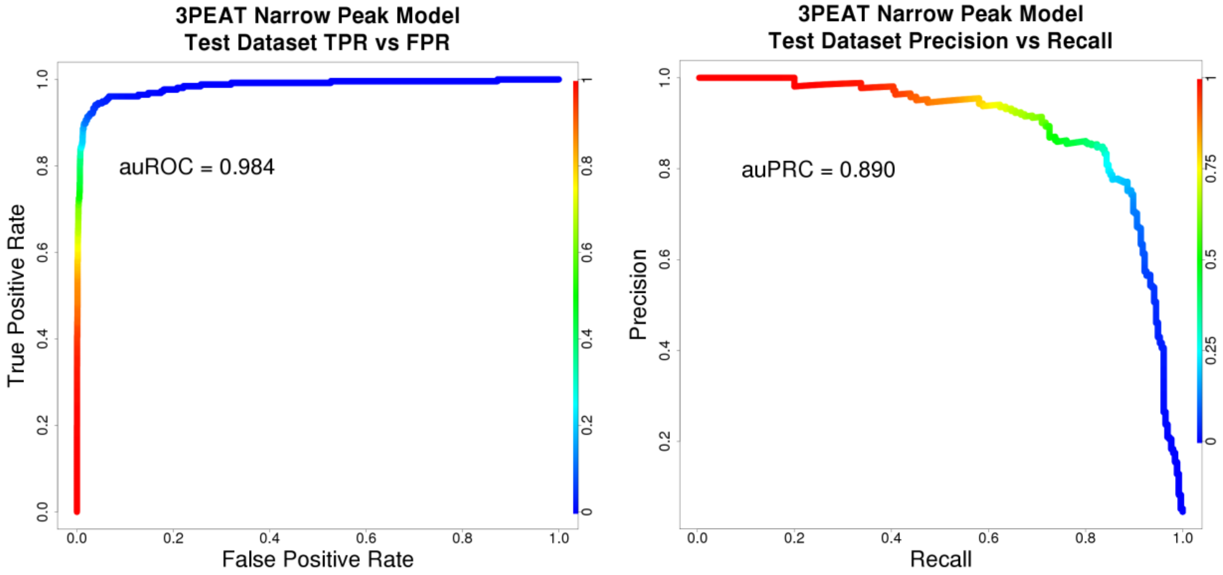
Supplemental Figure 5: Sharp, well defined Regions of Enrichment for WP promoters.





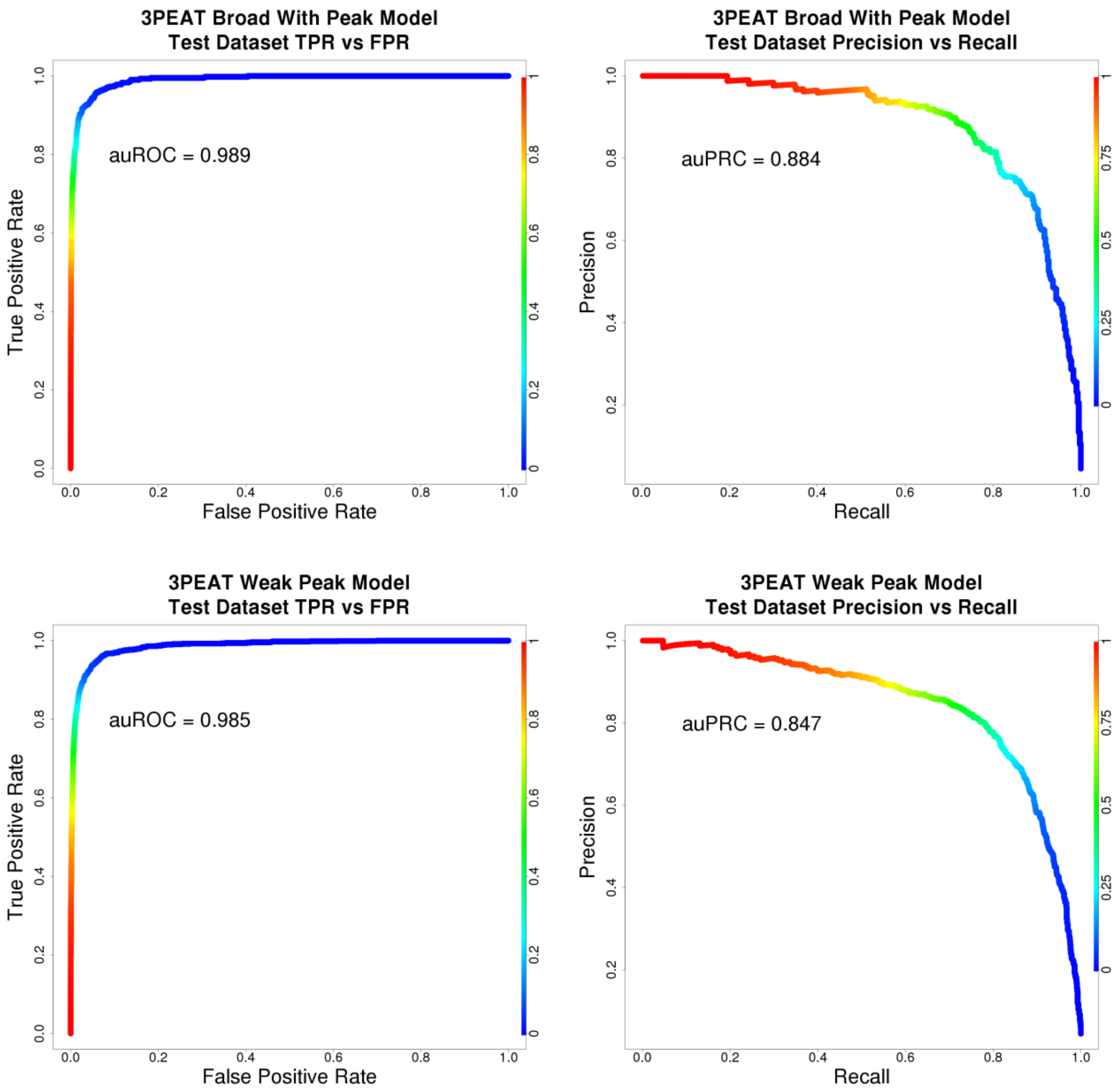
Supp. Figure 5: Sharp, well-defined Regions of Enrichment (ROEs) defined by PEAT tag clusters with Weak Peak transcription initiation patterns.

Supplemental Figure 6: Model Performance on Narrow Peak Modes.



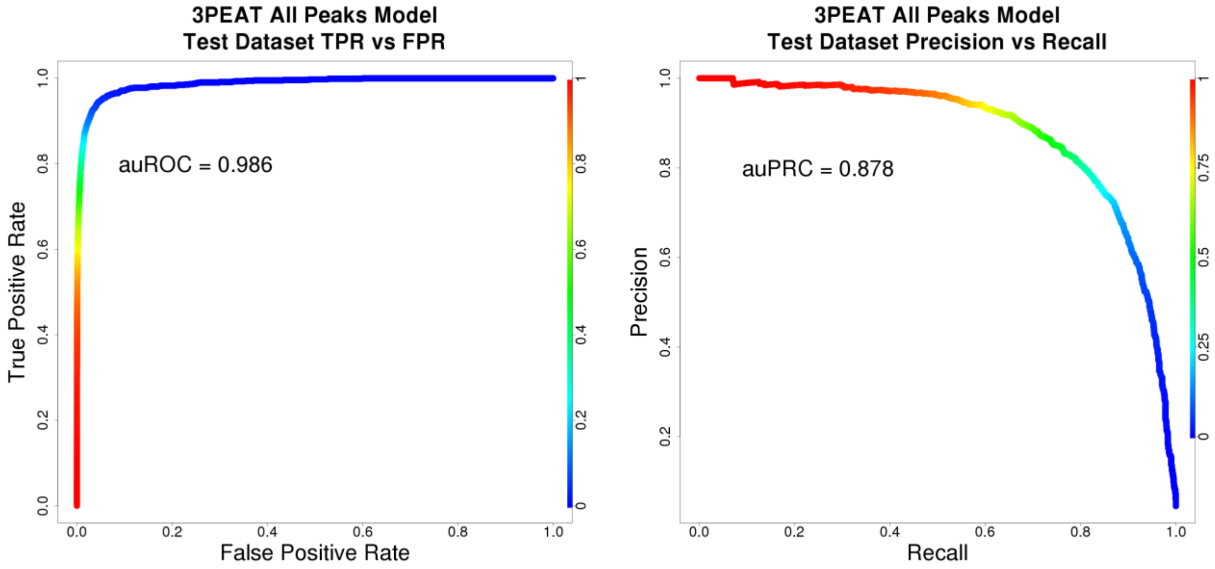
Supp. Figure 6: Plots displaying true positive rate vs false positive rate (ROC) and precision vs recall (PRC) show the 3PEAT Narrow Peak model's performance on an independent, held-out test set. The test set is comprised of NP positive examples, and negative examples drawn from nearby upstream and downstream regions. As in the training set, each NP positive example is an NP tag cluster mode—the most highly expressed location in the TSS peak distribution—for a peak that contains 100 or more TSS reads.

Supplemental Figure 7: Model Performance on Broad with Peak and Weak Peak Mode.



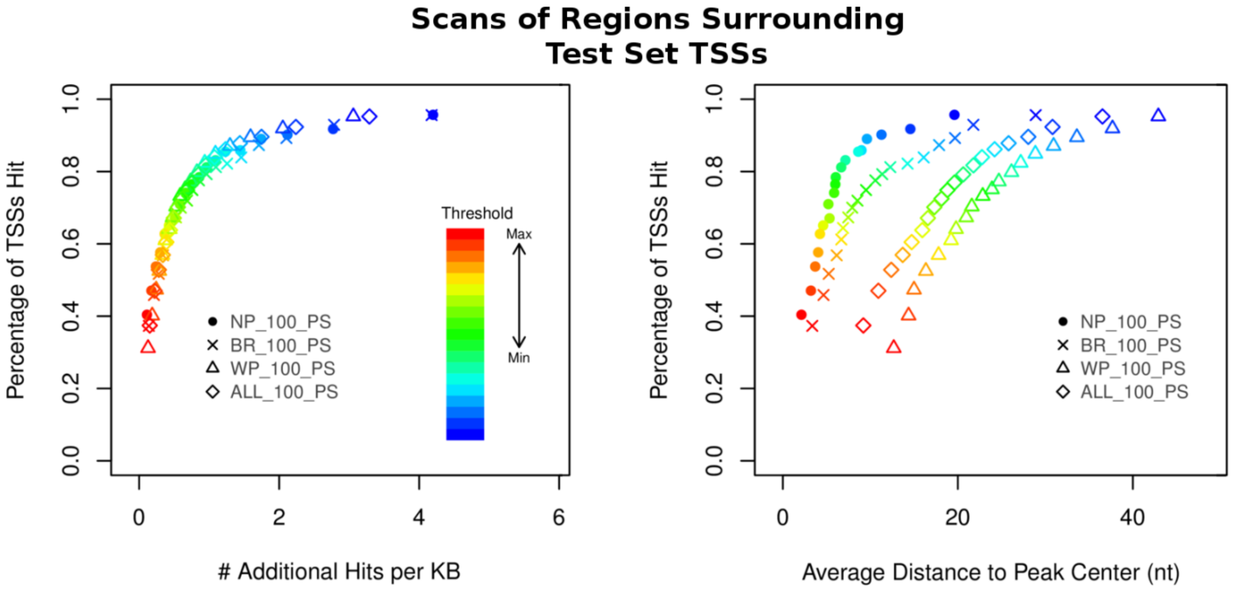
Supp. Figure 7: Plots displaying true positive rate vs false positive rates (ROC) and precision vs recall (PRC) show the performance of the 3PEAT Broad with Peak and Weak Peak initiation pattern models on independent, held-out test sets containing highly expressed TSSs and negative examples drawn from nearby upstream and downstream regions.

Supplemental Figure 8: Model performance on ALL Modes.



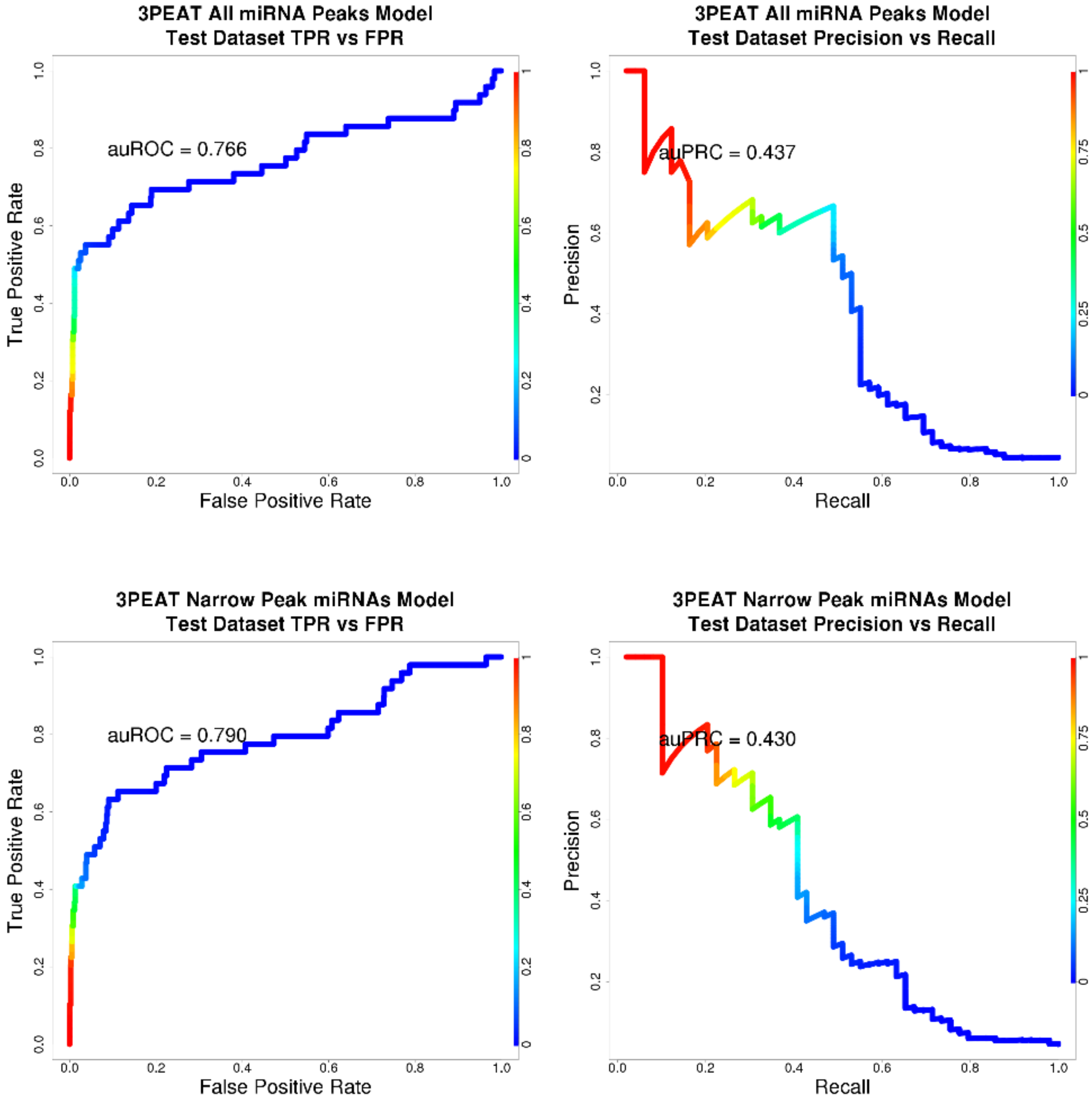
Supp. Figure 8: Plots displaying true positive rate vs false positive rates (ROC) and precision vs recall (PRC) show the performance of the 3PEAT model trained using all highly expressed PEAT tag clusters on an independent, held-out test set. All 3 PEAT initiation patterns (NP, BP, and WP) were combined together into a single dataset and used to train a general-case TSS prediction model.

Supplemental Figure 9: Model Performance on Genomic Sequence Scans.



Supp. Figure 9: Comparison of 3PEAT performance on genomic sequence scans for NP (circle), BP (cross), WP (triangle), and ALL (diamond) models. For each Test Set TSS, sequences of 4kb on each side of the TSS are considered, and a probability outcome is predicted at each nucleotide of the scanned sequence. On the left, a TSS is considered to be a ‘hit’ if a probability peak contains the TSS. The curve represented by each symbol type shows the percentage of TSSs hit as a function of number of additional hits per kilobase. On the right, each curve displays the average distance to the center of the probability peak computed over all of the peaks containing a TSS. At each threshold value (color), the plots give a comparative view of how many additional peaks are being called versus how well the TSS-containing peaks approximate actual TSS location.

Supplemental Figure 10: Model Performance on miRNA-proximal Modes.



3PEAT All miRNA Peaks Model
Test Dataset TPR vs FPR

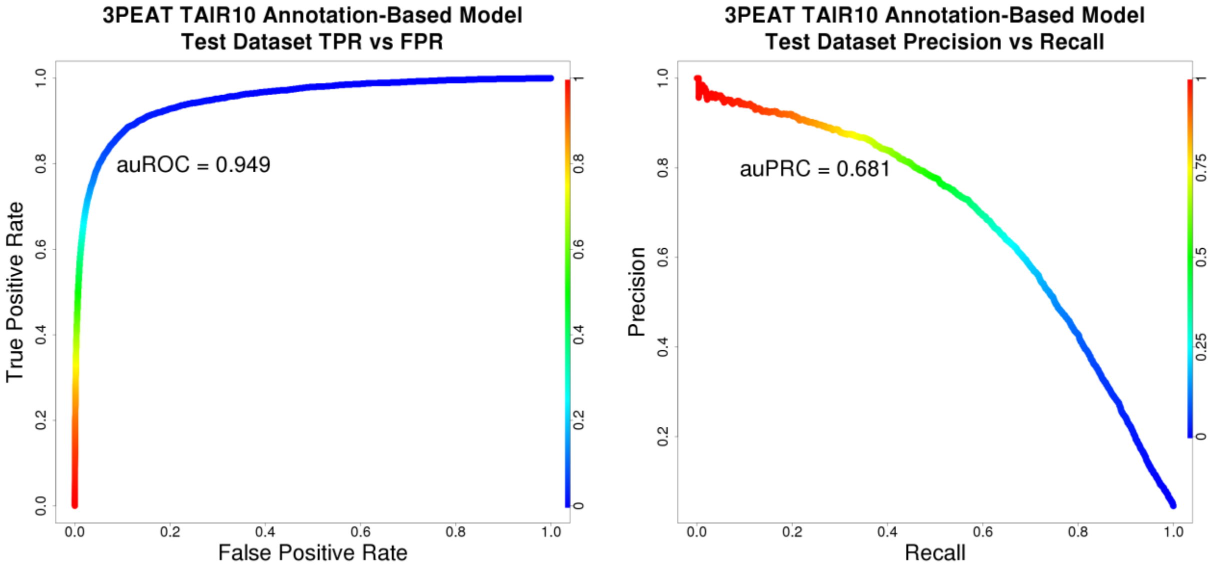
3PEAT All miRNA Peaks Model
Test Dataset Precision vs Recall

3PEAT Narrow Peak miRNAs Model
Test Dataset TPR vs FPR

3PEAT Narrow Peak miRNAs Model
Test Dataset Precision vs Recall

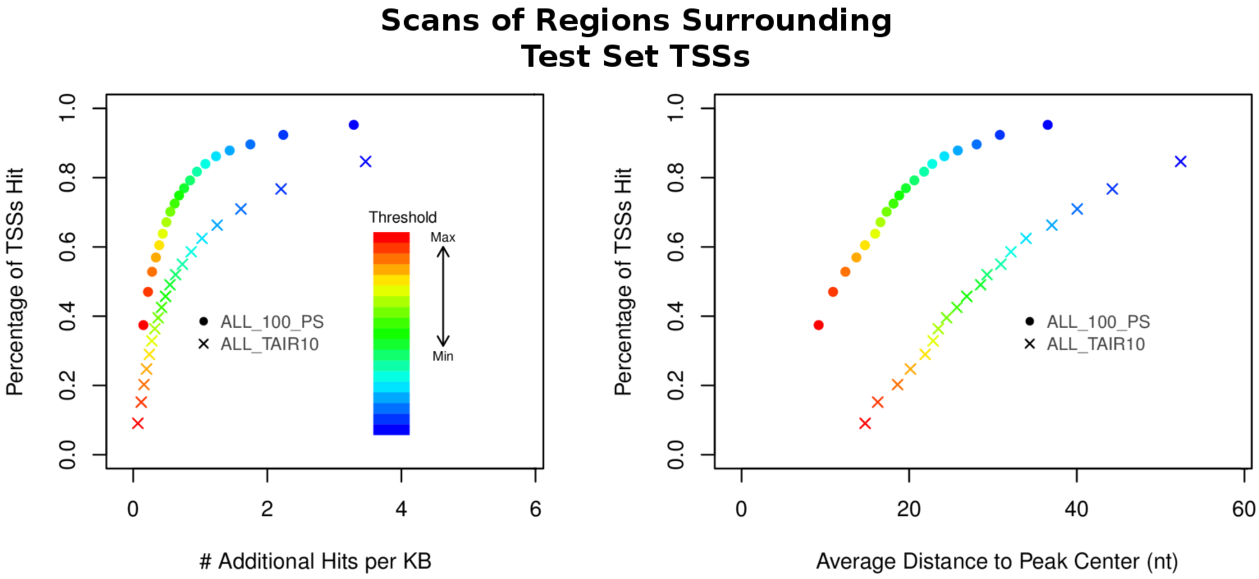
Supp. Figure 10: Plots displaying true positive rate vs false positive rates (ROC) and precision vs recall (PRC) show the performance of the general 3PEAT model used to predict All (top) and Narrow Peak (bottom) TSSs located near annotated miRNA precursors. All PEAT tag clusters located near miRNA precursors (including lowly expressed sites) were used as testing examples. These results show that the 3PEAT model is very specific but not sensitive in its prediction of miRNA precursor TSSs, suggesting that the sequence content of these promoter regions may differ from that of traditional protein coding genes.

Supplemental Figure 11: Model Performance on TAIR10 annotated TSSs.



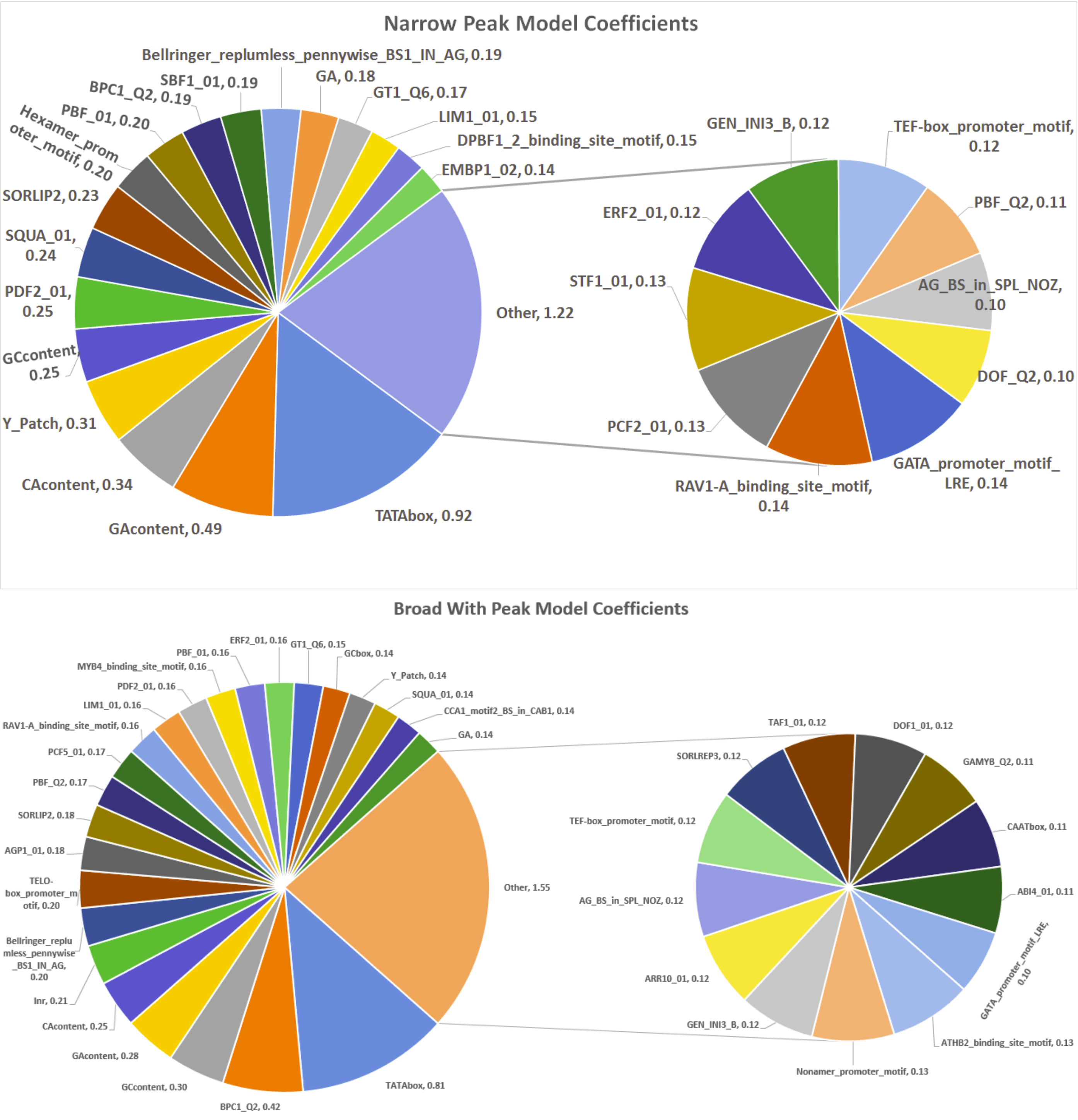
Supp. Figure 11: Plots displaying true positive rate vs false positive rates (ROC) and precision vs recall (PRC) show the testing performance of a 3PEAT model trained using annotated TAIR10 protein coding gene start sites in place of PEAT tag clusters. While auROC is changes very little when compared to PEAT data, PRC is substantially worse, resulting in a much less precise and sensitive model.

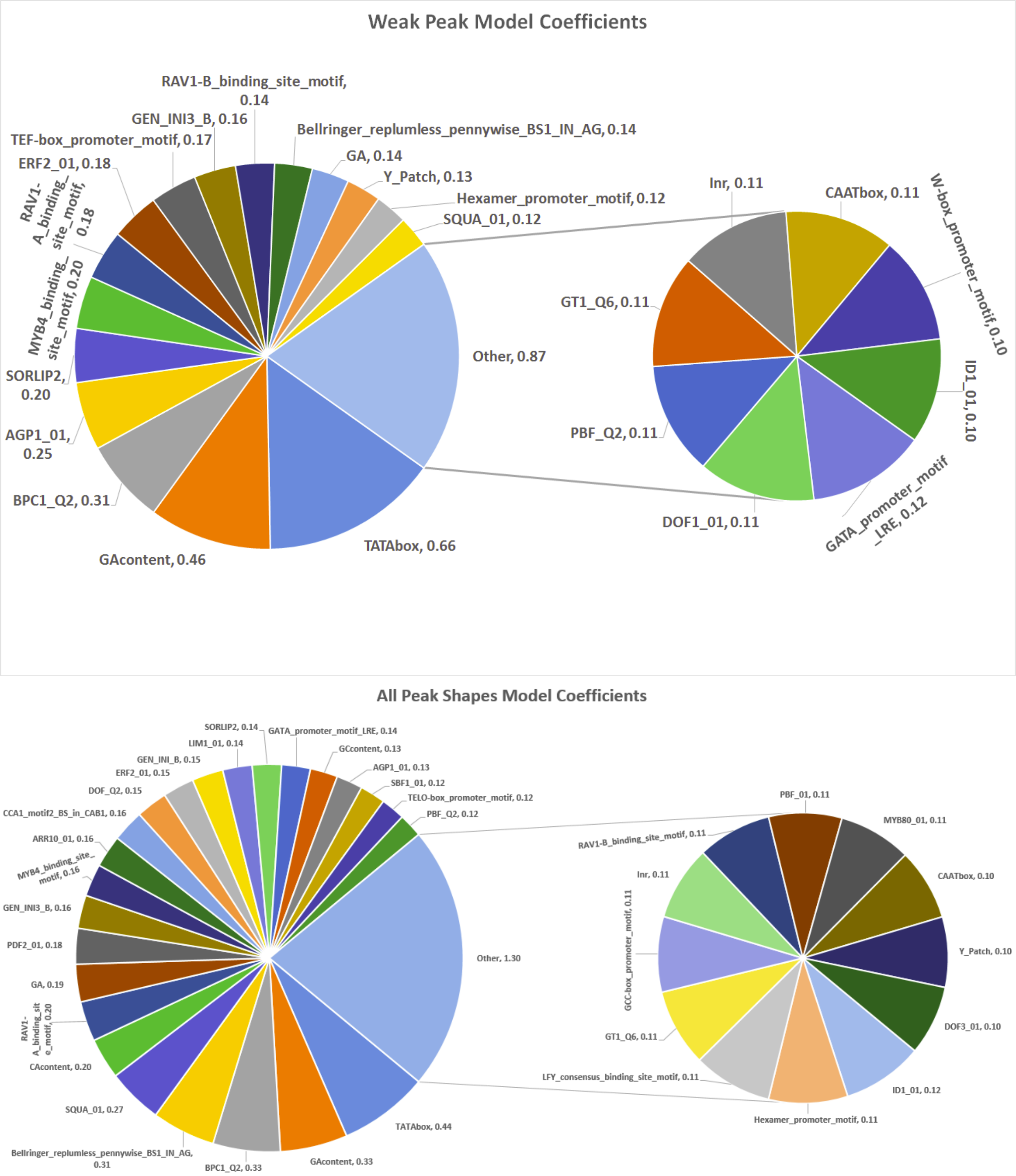
Supplemental Figure 12: Model Performance on TAIR10 Genomic Sequence Scans.



Supp. Figure 12: Comparison of 3PEAT performance on genomic sequence scans using models based on PEAT TSS (circle) and TAIR10 (cross) data. All highly expressed PEAT TSSs located near annotated protein coding genes were used in the PEAT model. The start sites of annotated protein coding genes were used to construct the TAIR10 model. For each Test Set TSS, sequences of 4kb on each side of the TSS are considered, and a probability outcome is predicted at each nucleotide of the scanned sequence. On the left, a TSS is considered to be a 'hit' if a probability peak contains the TSS. The curve represented by each symbol type shows the percentage of TSSs hit as a function of number of additional hits per kilobase. On the right, each curve displays the average distance to the center of the probability peak computed over all of the peaks containing a TSS. At each threshold value (color), the plots give a comparative view of how many additional peaks are being called vs. how well the TSS-containing peaks approximate actual TSS location. This shows the substantial increase in precision and sensitivity that the high-resolution PEAT TSS data provides.

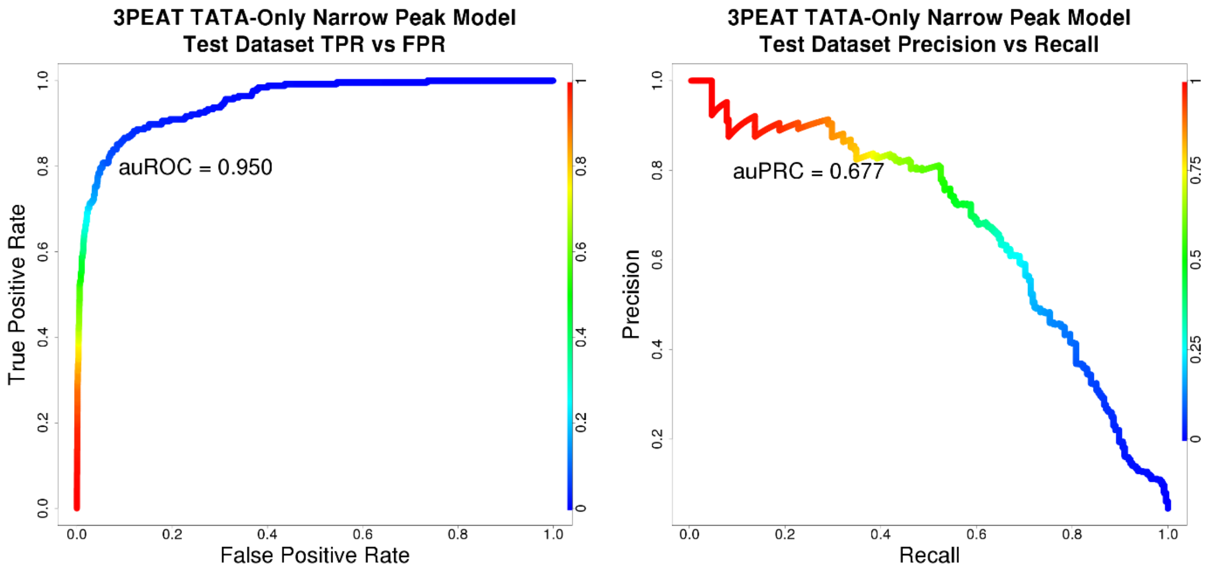
Supplemental Figure 13: 3PEAT promoter signatures, detailed for all shape models.





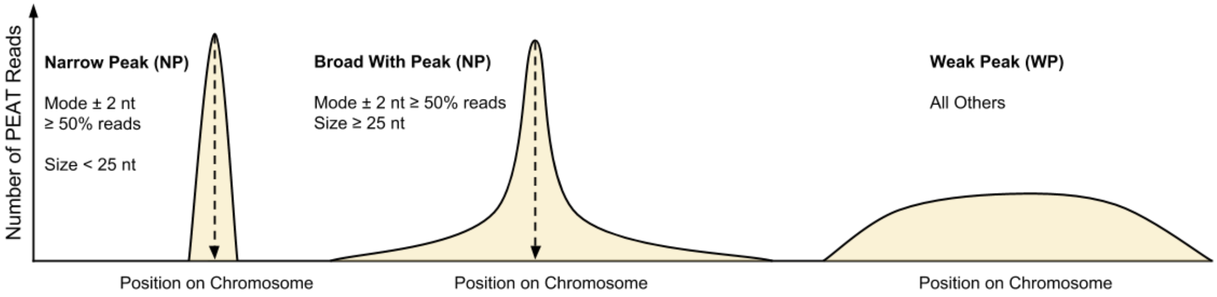
Supp. Figure 13: 3PEAT model coefficients for all models. Model coefficients can range from -1.0 to 1.0, only those exceeding 0.1 are displayed as part of the signature. Based on the model, heavily-weighted positive coefficients indicate TFs whose presence is strongly associated with the presence of a TSS peak. Numerical tables of coefficients for each model are provided in Supplemental Table 4.

Supplemental Figure 14: Model performance with TATA-box and enrichment features only.



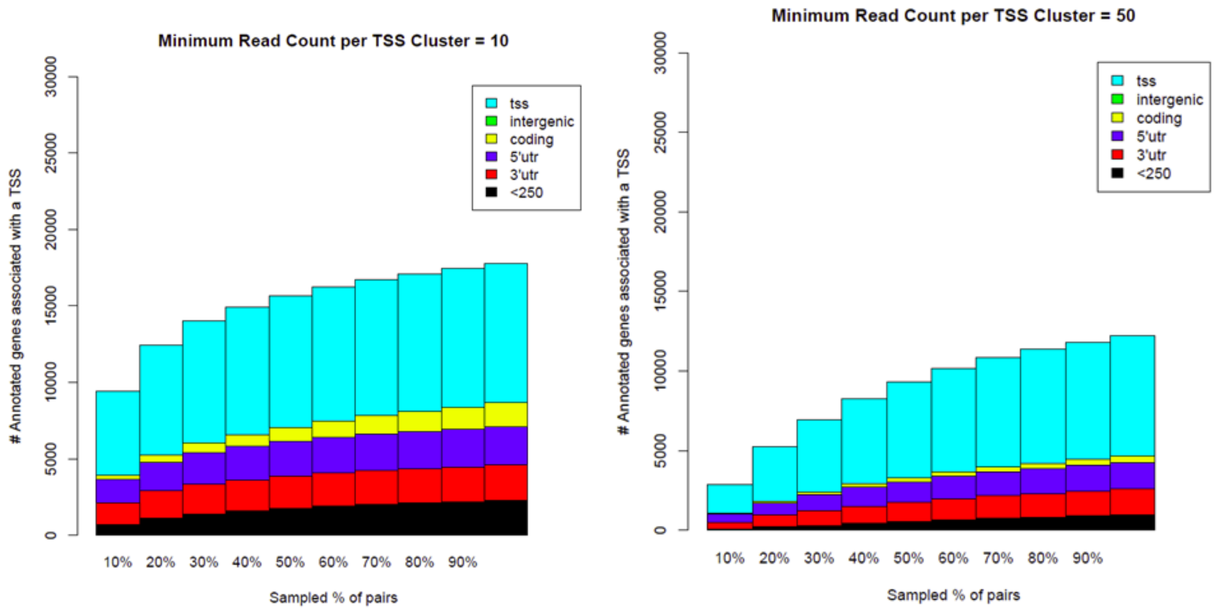
Supp. Figure 14: Plots displaying true positive rate vs false positive rate (ROC) and precision vs recall (PRC) show the 3PEAT Narrow Peak model's performance on an independent, held-out test set when trained using only TATA-box and a set of dinucleotide sequence enrichment (GC, CA, GA) features. auPRC suffered substantially compared to the NP model built using all TFBS features (Supplemental Figure 4). This result demonstrates that the presence of TATA-box alone is neither necessary nor sufficient for most transcription.

Supplemental Figure 15: Tag cluster initiation pattern definitions.



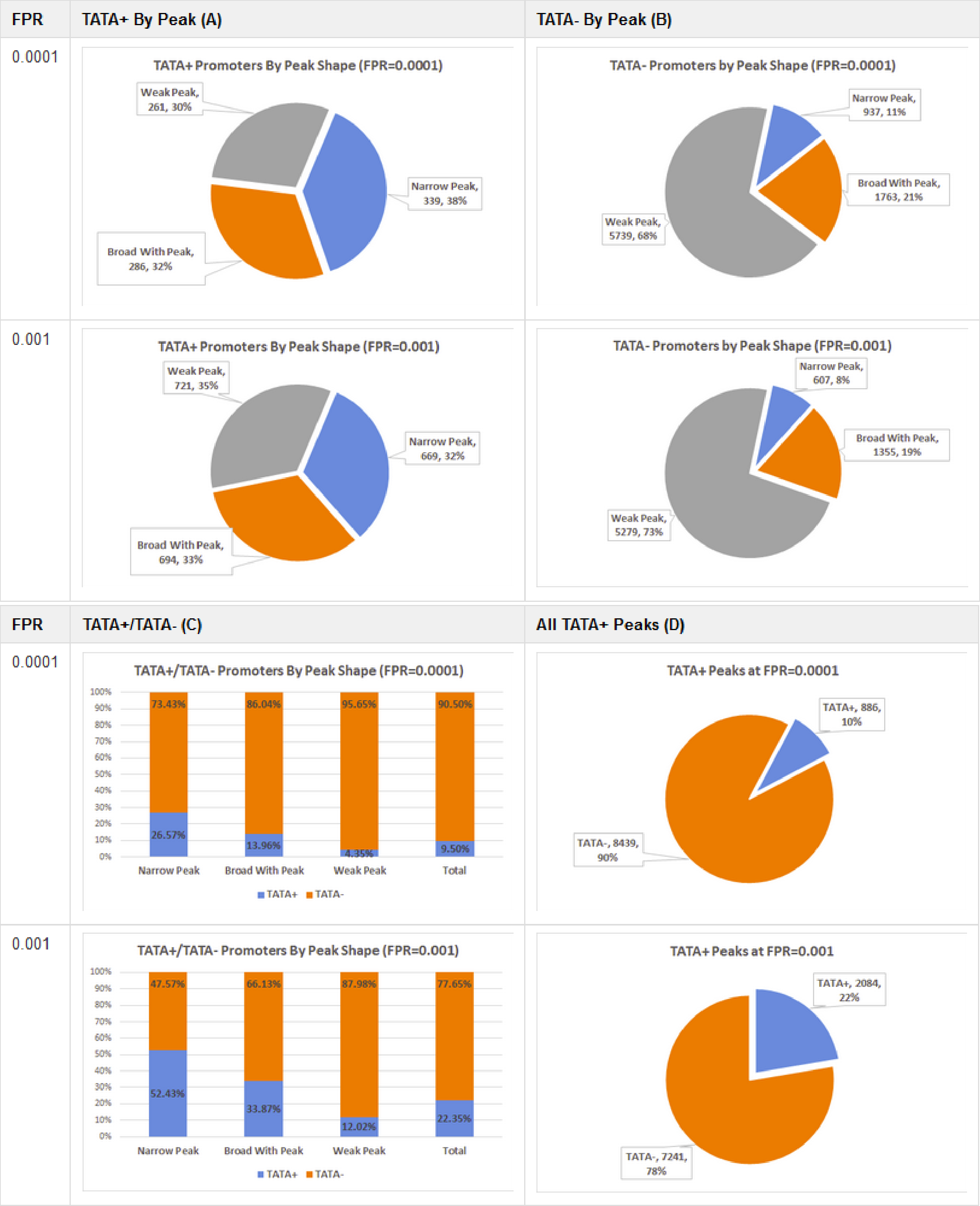
Supp. Figure 15: Definitions of PEAT tag cluster initiation patterns used in this study.

Supplemental Figure 16: Sampled read depth saturation analysis.



Supp. Figure 16: Sequencing depth analysis of PEAT data to determine if sequencing depth accurately represented gene expression in our pooled root samples. Starting with stringently mapping reads, reads were randomly sampled and reclustered according to our clustering procedure. As subsample size increases, fewer new genes are observed in the PEAT dataset. The change in number of genes has leveled off to nearly zero as the percentage of reads sampled increases from 90% to 100%. This indicates the sampling depth we achieved was within a few percent of the maximum sample depth that could be achieved.

Supplemental Figure 17: Proportions of TATA+/TATA- promoters.



Supp. Figure 17: Comparison of the percentage of TATA-containing (TATA+) and TATA-absent (TATA-) promoters by PEAT TSS initiation pattern. A promoter was considered TATA+ if it contained a TATA-box within TATA's Region of Enrichment (-45:-25 with respect to TSS mode). If the log-likelihood score of any nucleotide within this region scored above a given threshold, the promoter was considered TATA+. We evaluated 2 score thresholds corresponding to two False Positive Rates (FPRs), calculated with respect to the Arabidopsis promoter background sequence. Figures A and B show how TATA+ and TATA- promoters distribute by initiation pattern. This analysis shows that 30% to 38% of all TATA+ promoters form NP initiation patterns, despite NPs making up only 14% of the total tag clusters in the PEAT dataset. Figure C shows the proportion of TATA+ tag clusters by initiation pattern. Figure D reports the total percentage of TATA+ and TATA- tag clusters in our dataset.

Supplemental Tables

Supplemental Table 1: Counts of PEAT Tag Cluster Dataset Partitions used in 3PEAT Model.

Dataset	Abbriv.	Total Examples	Training	Testing	Training Upstream Negatives	Testing Upstream Negatives
NarrowPeak	np100_ps	1276	1021	255	460	101
BroadWithPeak	br100_ps	2050	1639[a]	410	852	213
WeakPeak	wp100_ps	6000	4800	1200	2449	608
All	all100_ps	9326	7460[a]	1865	4593	1140
TAIR10	tair10	27480	21923	5480	13530	3345
miRNA	mirna_ps	49	0	49	0	49

[a] One PEAT tag cluster (Peak_60747) was located only ~3700 nt from the end of chromosome 4, and was excluded from our dataset because the scanning procedure requires 5 KB both upstream and downstream of the TSS and could not be performed.

The number of quality-filtered PEAT TSS Tag Clusters used to build the 3PEAT models. The TAIR10 dataset is not built from PEAT TSS data, but instead from the TAIR10 annotations. The “upstream negative” columns contain the number of tag clusters from negative examples were drawn. Negative examples were drawn randomly from the upstream and downstream regions of tag clusters in regions which contained no evidence of transcription.

Supplemental Table 2: Comparison of the number of TATA+ vs TATA- promoters by PEAT TSS initiation pattern.

FPR=0.0001

Peak Shape	TATA+	TATA-	Total Examples	Proportion +	Proportion -
Narrow Peak	339	937	1276	26.57%	73.43%
Broad with Peak	286	1763	2049	13.96%	86.04%
Weak Peak	261	5739	6000	4.35%	95.65%
Total	886	8439	9325	9.50%	90.50%

FPR=0.001

Peak Shape	TATA+	TATA-	Total Examples	Proportion +	Proportion -
Narrow Peak	669	607	1276	52.43%	47.57%
Broad with Peak	694	1355	2049	33.87%	66.13%
Weak Peak	721	5279	6000	12.02%	87.98%
Total	2084	7241	9325	22.35%	77.65%

Supplemental Table 3: Cross-Validation Performance: auROC and auPRC statistics for each cross-validation fold of each 3PEAT model trained.

	Weak Peak Model		Broad with Peak Model		Weak Peak Model	
CV Fold	**auROC**	**auPRC**	**auROC**	**auPRC**	**auROC**	**auPRC**
1	0.994782521	0.933150399	0.989259293	0.904634537	0.986268808	0.874297581
2	0.989793679	0.880612174	0.992674778	0.904928444	0.9893719	0.87472384
3	0.976601368	0.877459554	0.9934135	0.932304383	0.981436425	0.875795904
4	0.993666776	0.911513	0.983283967	0.897491099	0.978160756	0.84577766
5	0.99481453	0.942739577	0.986759683	0.89015722	0.987694899	0.877039842
6	0.994732221	0.920590785	0.991606553	0.851945074	0.985043196	0.853297568
7	0.993200359	0.901232993	0.988552458	0.895095816	0.984012897	0.857648811
8	0.992893986	0.937444034	0.981429189	0.894799917	0.983710524	0.8720164
9	0.995486721	0.964125009	0.989158167	0.874990818	0.982798859	0.845746006
10	0.99717564	0.958043233	0.990525928	0.915696651	0.986113591	0.889124434

	All Peaks Model		TAIR10 Annotation Model	
CV Fold	**auROC**	**auPRC**	**auROC**	**auPRC**
1	0.985828842	0.878306217	0.948457199	0.693916637
2	0.986348144	0.861322733	0.9487788	0.681071879
3	0.989282985	0.886748964	0.94592455	0.678020212
4	0.987332157	0.897155073	0.948264939	0.68265261
5	0.988612487	0.879007199	0.950719671	0.681914664
6	0.988934644	0.896396891	0.949842212	0.696205277
7	0.987424997	0.893426747	0.949013438	0.689141172
8	0.989154892	0.892070533	0.951800102	0.695538181
9	0.987431329	0.893834812	0.949477312	0.692339718
10	0.989654	0.912980451	0.951570521	0.685226617

Supplemental Methods

Data Set Processing and Partitioning

PEAT TSS Data Processing

Paired-end reads were mapped to the TAIR10 transcriptome (Lamesch et al., 2012) using the same procedure as in (Ni et al., 2010). At most one mismatch was allowed, and reads were required to map uniquely. 4,001,991 mate pairs (8,003,982 reads) mapped to the genome with these stringent requirements. TSS tag clusters were then defined using the peak calling procedure detailed below, and partitioned into NP, BP, and WP initiation patterns. In brief, a smoothed density estimate of 5' TSS tags for each tag cluster was computed using the F-Seq tool (Boyle et al., 2008); each tag cluster boundary was then determined as exceeding a baseline score, and TSS clusters were condensed to the shortest distance containing 95% of the reads. NP clusters contained ≥ 50% of the reads within ±2 nt of the mode and span < 25 nt; BP clusters were those that contained ≥ 50% of the reads within ±2 nt of the mode and are ≥ 25 nt in length; all other clusters were classified as WP (Supplemental Figure 15). Only tag clusters with 10 or more reads were considered in the dataset. PEAT tag clusters were then labeled according to their relationship to TAIR10 gene annotation. Tag clusters were labeled with a 'TSS' location if they covered a gene's annotated start site. Other labels used in this analysis were clusters within a gene's 5' UTR and those located up to 500 nt upstream of the gene's annotated start site. Tag clusters associated with miRNAs were labeled separately. Supplemental Table 6 provides the entire dataset of PEAT tag clusters, along with the tag cluster datasets used for 3PEAT model training as described below.

Tag Cluster Annotation

PEAT tag clusters (or "peaks") are groups of contiguously mapping PEAT TSS tags, where each TSS tag corresponds to the 5' end of a capped pol-II RNA transcript. The mode of a tag cluster is defined as the location within the cluster where the greatest number of 5' ends were mapped. Clusters containing large numbers of tags were used to form the three datasets used in the 3PEAT model. First, tag clusters were filtered to consider only those located within the 5' UTR or within 500 nt upstream of an annotated protein coding gene. We limited our analysis to tag clusters for which PEAT provided strong evidence of a TSS location, requiring that each tag cluster used in model construction and evaluation contain least 100 reads. Finally, tag clusters were partitioned by their initiation pattern into individual Narrow Peak (NP), Broad with Peak (BP), and Weak Peak (WP) sets.

Sequencing Depth Analysis

To determine whether the sequencing depth achieved was sufficient to represent the gene expression state in our pooled Arabidopsis root samples, we performed a saturation analysis on our tag cluster dataset. Starting with our full set of stringently mapped reads, we randomly sampled 10%, 20%, …, 100% of the reads; for each sample, we re-performed the same process described above to produce our annotated tag cluster dataset, starting with a requirement of 10 reads per tag cluster. For each subsampled, re-clustered set of reads, we recorded the number of annotated genes associated with the resulting tag clusters (Supplemental Figure 16A). In each re-processed subsample, tag cluster shapes, locations, and annotated values can change. For example, with a much sparser dataset and considering tag clusters that may contain as few as 10 reads, many small tag clusters that associate with fewer annotated genes can be observed. As the subsample size approaches 100%, if the process is producing tag clusters such that the number of associated genes is still dramatically increasing, it could be concluded that a greater sampling depth could be achieved that would better represent gene expression in sample. However, we observed that the amount of annotated gene increase between subsamples in Supplemental Figure 16A leveled off nearly to zero as the subsample size grew to 90% and then 100% of the reads. We repeated the experiment with a stronger requirement of 50 reads per tag cluster (Supplemental Figure 16B), to test whether consideration of fewer but larger, more certain tag clusters would still show this saturation in the number of genes covered. We observed a nearly identical plot trajectory, which indicates that the sampling depth we achieved was within a few percent of the maximum depth that could be achieved.

Model Data Sets

We annotated tag clusters by initiation pattern and TAIR10 association, and selected all highly-expressed tag clusters for our analysis (Supplemental Information: Tag Cluster Annotation). Supplemental Table 3 shows the number of tag clusters in each data set. The tag clusters in each data set were randomly partitioned into training and test sets, which were used to build and test the predictive 3PEAT model. 80% of tag clusters were used to build the model, while the remaining 20% were held out and used to evaluate model performance on previously unseen data. In addition to TSS tag clusters, the datasets also included negative examples—near regions of the genome where no TSS tag clusters were present (Supplemental Information: Negative Example Construction). For the purposes of this analysis, the tag cluster mode was considered as the putative TSS for the cluster.

Model Construction

Model Features

The input features of the 3PEAT model measure the presence of individual binding elements (TATA, Ini, CAAT-box, MADS-box, etc) within their Regions of Enrichment on each DNA strand with respect to the location under examination. For example, if 3PEAT Narrow Peak model is applied at the particular sequence location Chr1:+:34,567, it will record numerical values for a TATA-box sequence signal near 33 nt upstream (-33) on the + strand, an Initiator signal near +1 on the + strand, a MADSB signal near -36 on + strand, and similarly for all defined TF ROEs on both strands. This set of numerical values for TF presence/absence within the ROE is then multiplied by feature weights and combined to form an output probability estimate that Chr1:+:34,567 is the mode of an NP-shaped TSS peak. A successfully trained model is able to find feature weights that produce an accurate probability estimate for any input sequence location.

Each ROE was subdivided into five overlapping windows of equal length, with two additional flanking windows added on either side of the ROE (see Figure 3 of Megraw et al. (2009)). To produce the features for a training example, each PWM was scanned along every nucleotide within a window in its ROE. All positive log-likelihood scores were summed together to produce the numerical feature value for this ROE's window. Sequence content features were added for GC, CA, and GA content in the 200 nt window surrounding the peak mode. Therefore, the total number of features for an individual example is seven times the number of TF strand combinations with defined ROEs, plus 3 sequence content features.

Negative Example Construction

The training and testing partitions of each dataset used to train the 3PEAT included negative examples—locations where no TSS tag cluster is present. Negative examples were created by selecting intergenic and exonic regions from the genome. For each tag cluster in a dataset, 20 intergenic locations were drawn at random from the region 1 kb – 4 kb upstream of each PEAT tag cluster that contained no other TSS tag cluster in this upstream region. Additionally, for each tag cluster in the training dataset, a location was randomly selected from a list of Arabidopsis exons as an additional negative example. This list contained all TAIR10 annotated exons, with the exception of the first exon of each protein coding gene. In the test set, this negative example was selected at random from the entire Arabidopsis genome. This selection resulted in a ratio of 21:1 for negative and positive examples in each dataset.

Model Evaluation

3PEAT Model Training and Testing

Our model uses L1-regularized logistic regression, a method which performs automatic feature selection by removing the least significant features in the model. We use the l1_logreg package, an efficient C implementation of logistic regression (Koh et al., 2007). The L1 penalty parameter controls the number of features allowed within the final model. We select this parameter through a validation partition in a double cross-validation process, measuring classification performance with auROC. This procedure was previously described in detail in (Megraw et al., 2009). A final model is constructed by building a classifier from all training examples in the data set, using the average of the optimal L1 parameters found in each cross-validation partition. After this model is constructed, it is evaluated by classifying examples from the testing dataset, which is composed entirely of examples not used in any step of the training process.

3PEAT Model Gene Scanning Procedure

The final 3PEAT models (NP, BP, WP, ALL) were used to classify each nucleotide in the 8 kb region surrounding each PEAT tag cluster in the model's test data set. At each position, the model calculates the probability that this location is the mode of a TSS tag cluster for the initiation pattern used to train the model. This procedure results in a sequence of probability predictions along the chromosome, with high probability values in regions that are most likely to be a TSS. A genomic location is considered a "hit" when the probability value rises above a threshold and then falls below the threshold for at least 10 nt. A hit is considered a "TSS hit" (accurately predicting a TSS location) if the probability peak contains a PEAT tag cluster. Calculating the average distance between TSS hit centers and PEAT tag clusters provides an assessment of how precisely the model approximates true TSS locations.

Supplemental References

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics **24,** 2537-2538.

Brady, S.M., Orlando, D.A., Lee, J.Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science **318,** 801-806.

Koh, K., Kim, S.-J., and Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. Mach. Learn. Res. **8,** 1519-1555.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research **40,** D1202-1210.

Li, S., Liberman, L.M., Mukherjee, N., Benfey, P.N., and Ohler, U. (2013). Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. Genome research **23,** 1730-1739.

Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., and Hatzigeorgiou, A.G. (2009). A transcription factor affinity-based code for mammalian transcription initiation. Genome research **19,** 644-656.

Ni, T., Corcoran, D., Rach, E., Song, S., Spana, E., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nature methods **7,** 521-527.

Supplemental Data Sets

The following data sets have been deposited in the DRYAD repository under accession number doi:10.5061/dryad.r2342.

Supplemental Data Set 1: Regions of Enrichment for TSS Initiation Pattern Models.

Defines the locations of positional enrichments of Transcription Factor Binding Sites relative to the TSS for all Transcription Factors evaluated. Each sheet contains the locational enrichments of TFBSs for an individual initiation pattern and strand.

Supplemental Data Sets 2-4: Genome Scanning Results.

These scans of 4 KB genomic regions flanking all TSS Test Set peaks by initiation pattern are also available online at http://megraw.cgrb.oregonstate.edu/suppmats/3PEAT/Scans/.

Supplemental Data Set 5: Unique GO Terms Associated with Initiation Patterns.

Lists GO terms which were overrepresented and unique within a single initiation pattern, along with the specific genes associated with the GO term.

Supplemental Data Set 6: 3PEAT TSS Prediction Model Coefficients.

The L1-regularized model coefficients from the 3PEAT model built from each initiation pattern dataset. Coefficients range between -1.0 – 1.0 and show the average value across all 7 ROE windows. The larger a coefficient, the more informative the model considers the feature.

Supplemental Data Set 7: PEAT TSS Tag Clusters

Contains genomic locations of TSS tag cluster peaks used in the 3PEAT models, grouped by initiation pattern, in addition to number of reads, associated gene, part of the gene where peak is located, and peak ID.