# Supplementary Tables and Figures Referenced in the Text

## Supplementary Tables

| Training Sets | | | |
|---|---|---|---|
| | *Annotation-Supported* | *CAGE-Only-Supported* | Total |
| *CpG-Island* | 788 | 164 | 952 |
| *Non-CpG-Island* | 566 | 881 | 1447 |
| Total | 1354 | 1045 | 2399 |

| Test Sets | | | |
|---|---|---|---|
| | *Annotation-Supported* | *CAGE-Only-Supported* | Total |
| *CpG-Island* | 134 | 20 | 154 |
| *Non-CpG-Island* | 132 | 175 | 307 |
| Total | 266 | 195 | 461 |

**Supplementary Table 1**: The number of start sites within each data set (see totals), along with a breakdown of set intersections. The CpG-Island and Non-CpG-Island subdivisions of the Annotation-Supported Training Set (cross-validation performance shown in Figure 3) are shaded in blue.

## Supplementary Figures

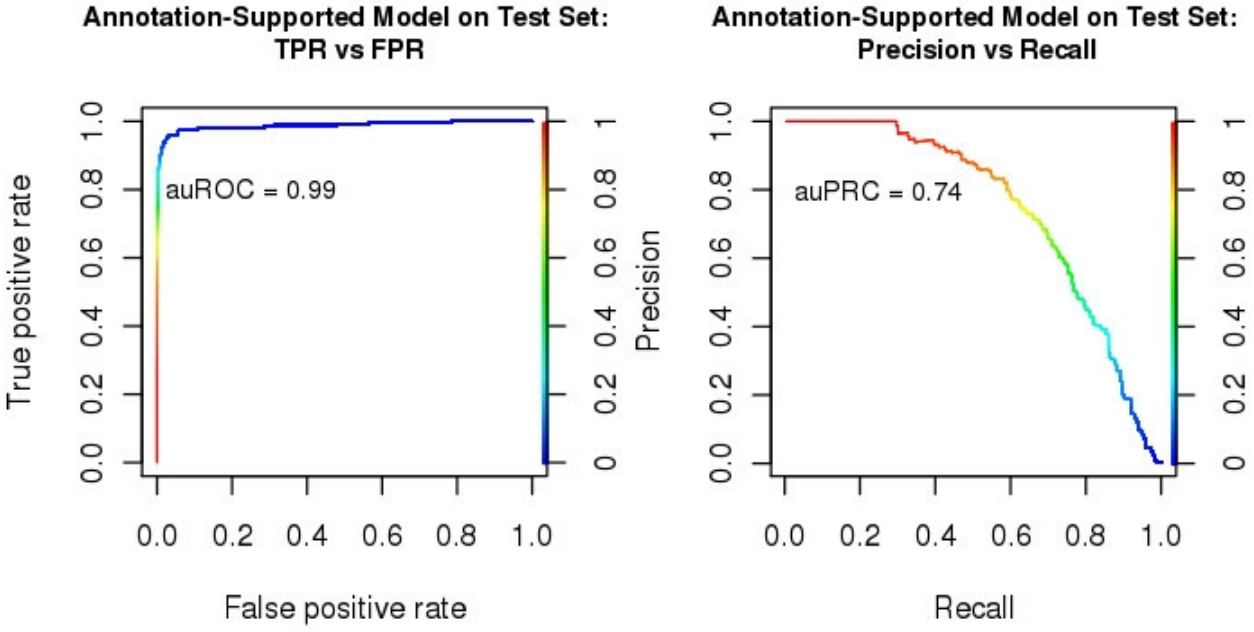
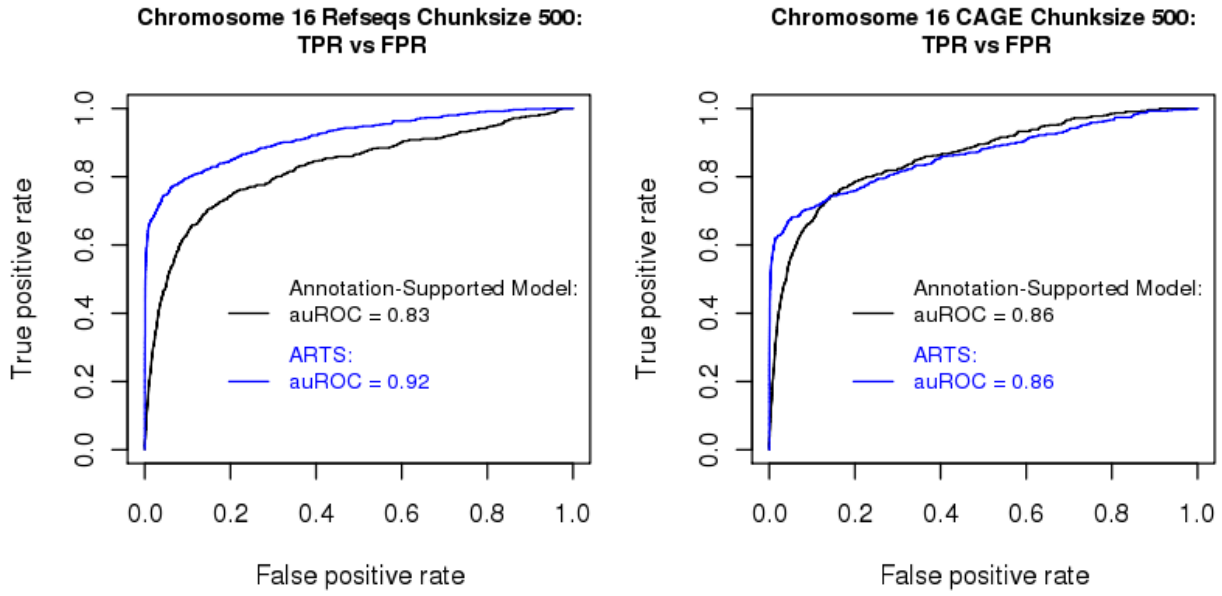
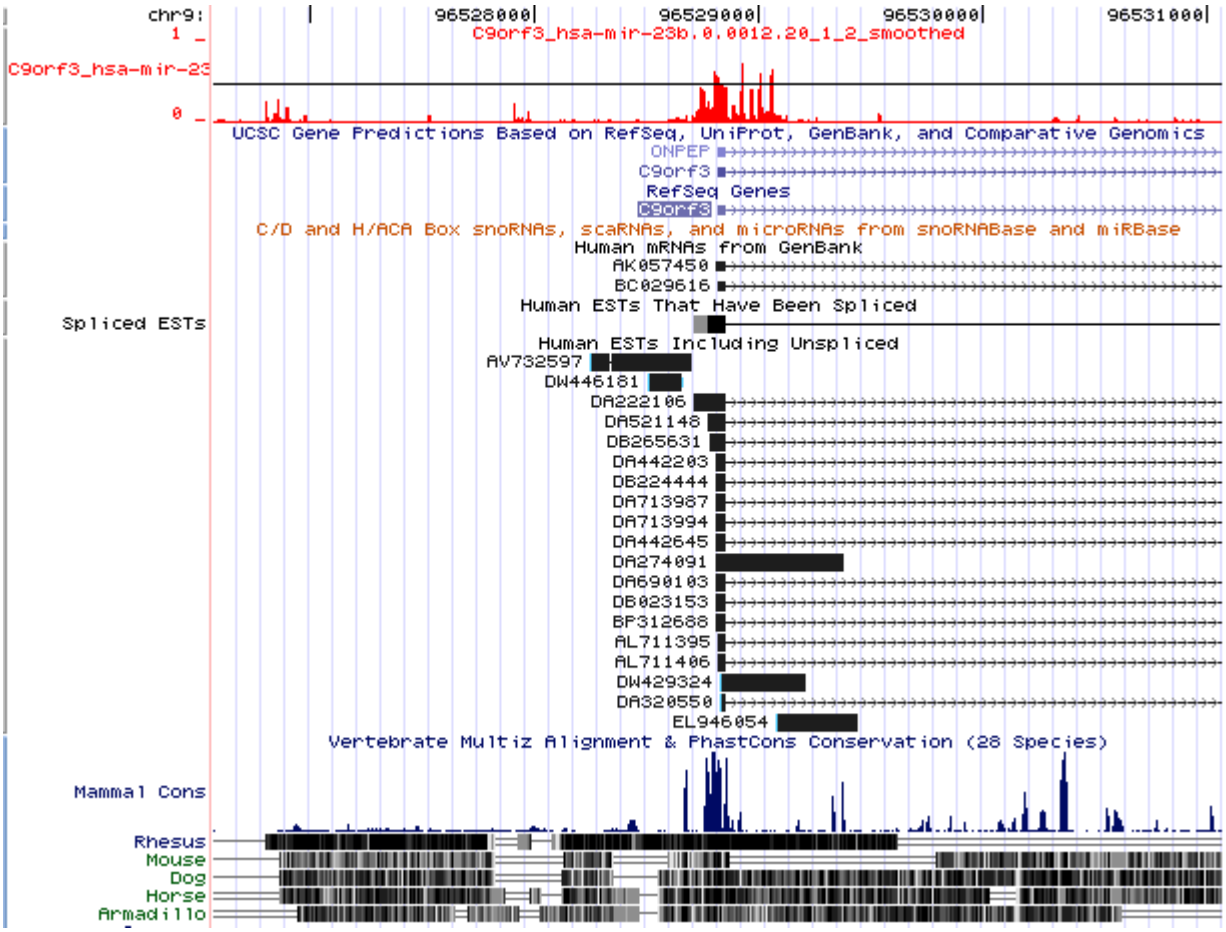



**Figure 1**: Performance of the Annotation-Supported model on a completely independent data set, the Annotation-Supported Test Set. Positive examples are the 266 CAGE tag cluster locations (location of highest CAGE tag count is selected), while negative examples are 100,000 randomly selected genomic locations from the most recent mouse genome build. The ROC curve and the Precision-Recall curve (PRC) provide two performance views. Color mapping illustrates classifier threshold value at each point on the curve.

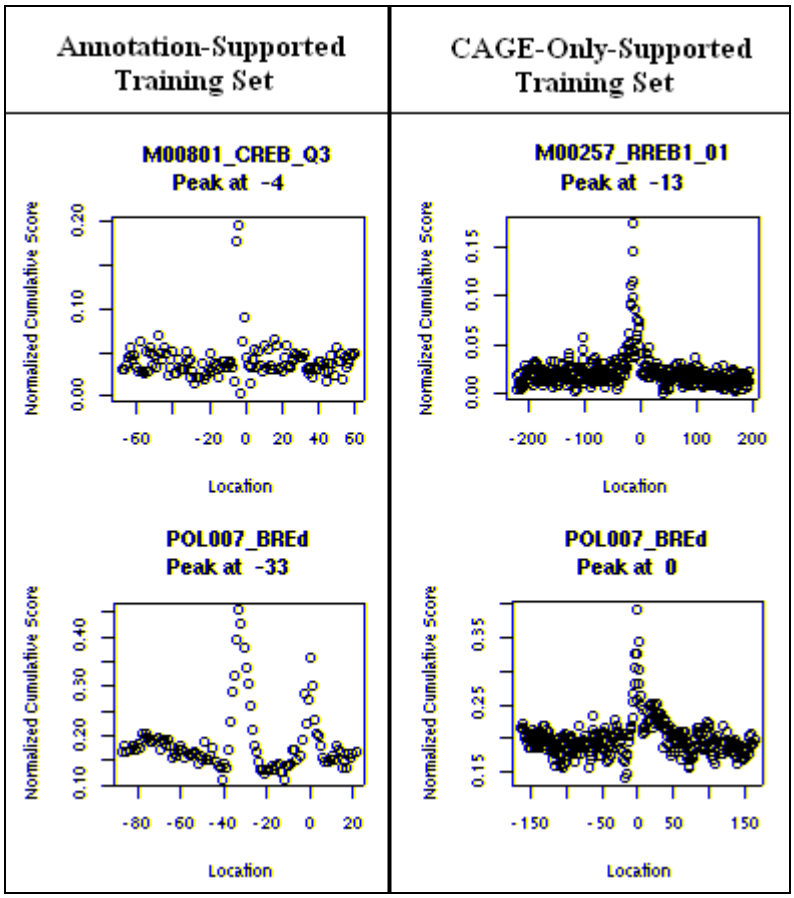



**Supplementary Figure 2**: Output comparison of the Annotation-Supported model and the ARTS TSS prediction program. Chromosome 16 is divided into 500nt chunks, and the prediction having the largest value within each chunk is computed for each program. Positive chunks contain Refseq or CAGE starts, respectively. Negative chunks comprise downstream gene portions for Refseq, and all non-positive chunks for CAGE. Area under the curve (auROC) provides a performance measure given these chunk definitions.

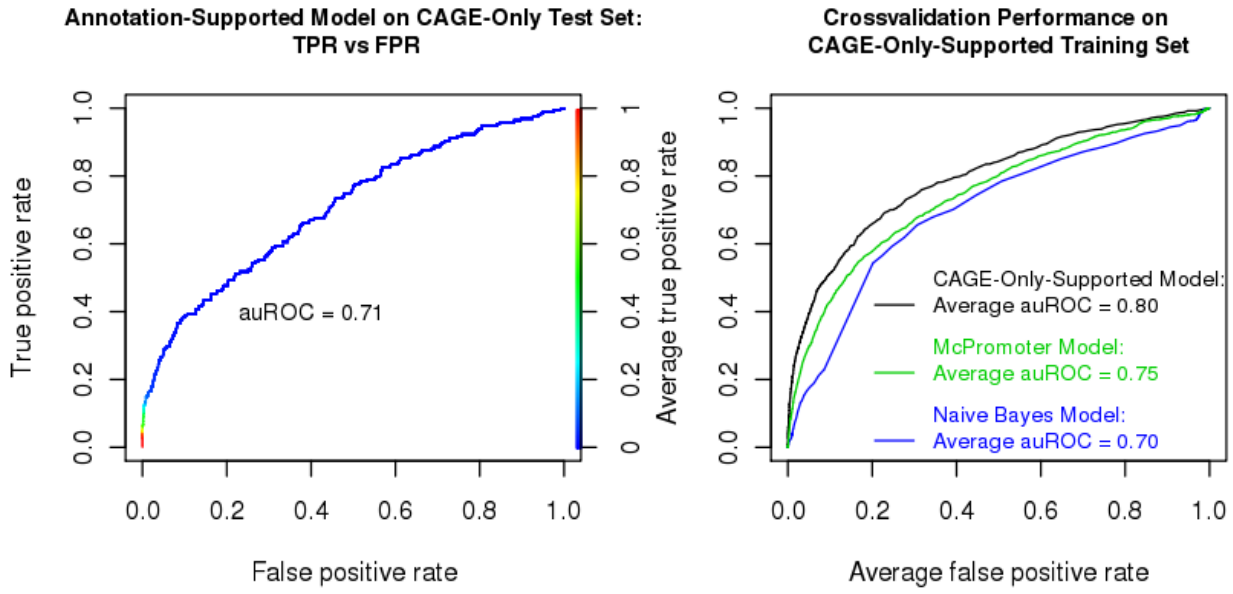



**Supplementary Figure 3**: This UCSC custom track displays a scan using the Annotation-Supported model over the region of a long UCSC transcript (C9orf3) with no annotated coding exons, containing hsa-mir-23b near the transcript end. This scan is typical of about 70% of the scans in our set of 20 putative pri-miRNAs, with a clear high probability region overlapping the annotated start of the transcript.

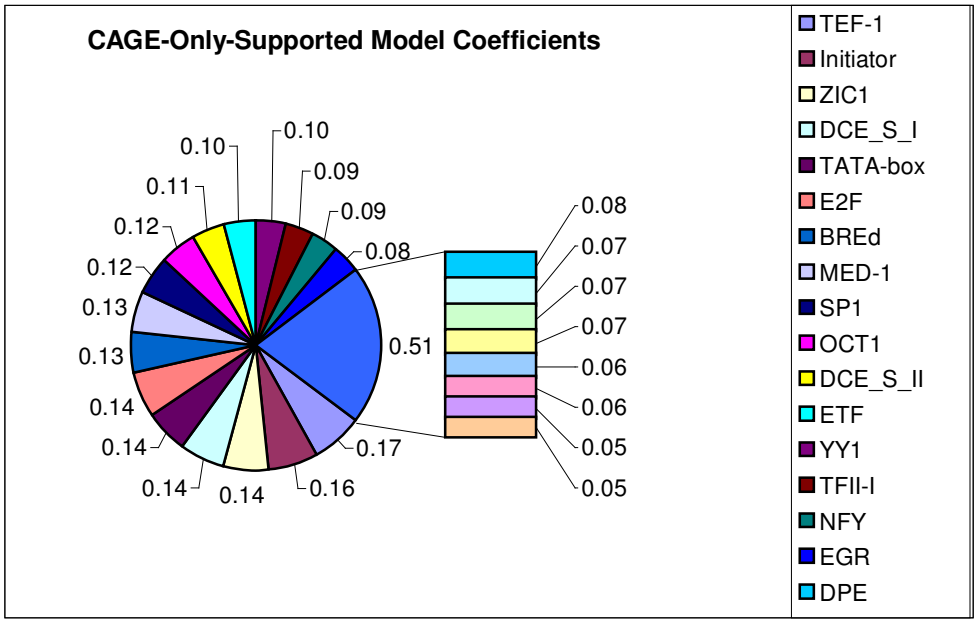

**Supplementary Figure 4**: This UCSC custom track displays a scan using the Annotation-Supported model over the region of a Genbank mRNA overlapping hsa-mir-194-2. Typical of about 30% of the scans in our noncoding set, there is no clear high-probability peak in the region of the beginning of transcript AK09202. However, there is strong evidence in this particular case for single peaks starts coinciding with several EST-supported locations downstream.




**Supplementary Figure 5**: There are numerous cases where a particular TF signal displays strong positional enrichment (on one or sometimes both strands) with respect to TSS in one data set, but little or no enrichment in another. For example, CREB (cAMP responsive element binding protein) shows a clear strong positional enrichment on the forward strand in the Annotation-Supported Single-TSS data set (upper left), whereas RREB (ras responsive element binding protein) shows strong enrichment on the reverse strand in the CAGE-Only-Supported Single-TSS set (upper right). BREd (brain and reproductive organ-expressed, downstream) is an example of a signal which is positionally enriched in both sets, but at a distinctly different strongest peak location within the Annotation-Supported Single-TSS set (lower left) than in the CAGE-only-Supported Single-TSS set (lower right).




**Supplementary Figure 6**: Performance of the Annotation-Supported model on the CAGE-Only-Supported Test Set (left) and 10-fold cross-validation performance for the retrained CAGE-Only Model (right). The plot on the right compares the performance of two additional classifiers, a Naïve Bayes classifier and McPromoter's HMM classifier, on the CAGE-Only Training Set.




**Supplementary Figure 7**: Logistic regression coefficients above 0.05 for the CAGE-Only-Supported Model. TATA-box and Initiator elements still play a role, but weights are much more evenly distributed over a variety of TFs.