Generating CAGE defined Transcriptional Start Sites

by Timo Lassmann (timolassmann at gmail dot com)

June 8, 2011

Introduction

Cap Analysis of Gene Expression (CAGE) captures the 5' ends of RNAs expressed in the cell. Analysis of the deeply sequenced cellular fractionations has shown that CAGE signal occurs mainly in the promoter regions of known genes but also along the body of genes and intergenically. The purpose of this document is to describe a strategy to separate CAGE signal into genuine transcriptional initiation sites and the remaining signal.

Workflow

Deriving a sequence based TSS model.

The mappings of all ENCODE CAGE libraries were pooled. The data was re-formatted into CAGE defined transcriptional start sites (CTSS) which are single nucleotide positions where the mapping of CAGE reads start. CTSS is a term used historically at RIKEN but given the recent findings may not be appropriate anymore. A raw expression count equal to the sum of reads mapping to the single nucleotide position are assigned to each CTSS. CTSS with fewer than 10 reads in all libraries were discarded. To identify larger regions exhibiting CAGE signal we used the parametric clustering algorithm "paraclu" [1] developed by Martin Frith specifically for CAGE data. The output is a hierarchal organization of overlapping clusters delineating very broad regions and sub-clusters focusing on increasingly denser regions of CAGE expression (see Figure 1 of [1] for further explanation). Given that the length of a nucleosome is approximately 150bp we selected all clusters shorter than 200bp from the set of generated clusters for downstream analysis. The number of obtained clusters at this stage is always much higher than the number of expressed genes in the cell. We applied the TSS predictor developed by me on these clusters to sub-classify these clusters (see below for details).

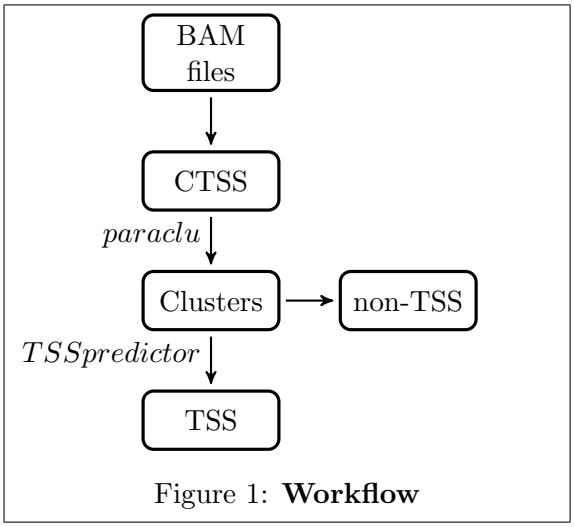


Figure 1: **Workflow**

TSS predictor

The TSS predictor is a non-supervised classifier based on modeling sequences surrounding CAGE regions via hidden markov models (HMMs). Two models are trained on all sequenced surrounding CAGE clusters. The model architecture is designed to capture sequence motifs of length 2-8 present at a certain distances from the middle of each cluster (Fig. 2). During training, the main model uses the number of raw reads observed in each cluster to proportionally weight the corresponding sequence while the background model assumes equal weight for all sequences. The posterior probability of each cluster fitting to the main model is calculated using Bayes' rule. Essentially we are asking what genomic sequence features give rise to many reads in the region.

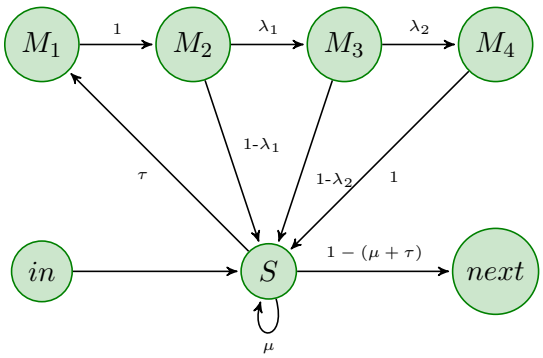


Figure 2: **Region HMM model architecture.** The HMM is composed of a series of region models where *in* and *next* are upstream and downstream S states. States M_1 - M_4 can model short sequence motifs of lengths 2,3 and 4. States M_5 - M_8 are not shown.

Predicting TSSs in individual libraries and cell lines.

To apply the predictor to cell lines we re-clustered the data as before but this time using only reads from individual cell-lines. The outcome are boundaries delineating all regions containing CAGE reads in each cell line. Boundaries containing less than 10 reads in all libraries were discarded. The TSS predictor was run on the remaining boundaries, but using the TSS model derived from all data.

Evaluation Strategy

To evaluate the predictions we overlapped CAGE clusters with windows surrounding the 5' ends of known GENCODE genes. A cluster with a high posterior probability overlapping a known promoter was counted as a true positive while intergenic clusters with a high posterior probability was counted as a false positive (Fig. 3). **NOTE**: the latter are likely novel TSSs currently not covered by GENCODE. By selecting a series of cutoffs receiver operating characteristic (ROC) and their area under ROC curve (AUC) was derived in various settings.

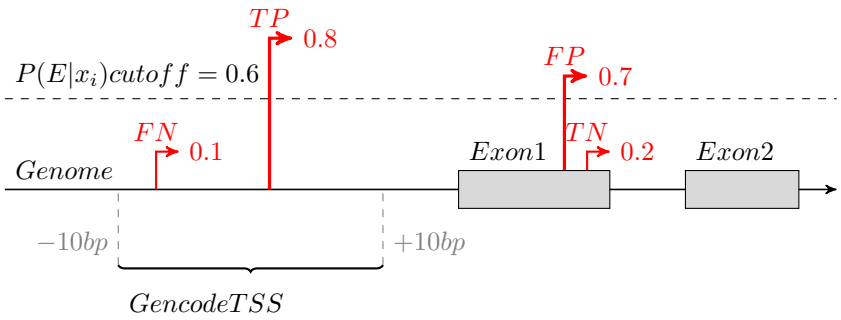


Figure 3: **Evaluation Strategy.** All CAGE clusters (shown as red arrows) are annotated by intersection with GENCODE TSSs. A series of cutoffs is selected and true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are counted as shown.

Assigning Compartment Specific Expression

For each library be intersected all CTSSs with the cell specific boundaries defined by the clustering. The cluster expression was set to sum of raw read counts making up the CTSSs. The raw expression was normalized by the number of mapped reads and multiplied by one million (tags per million - tpm).

To facilitate the downstream analysis I split the clusters in each library into three categories. All files are in ENCODE RNA Elements: BED6 + 3 Scores Format.

1. **. . . tss_high.bed** lists predicted TSS with at least 1 tpm expression. Columns correspond to:
   (a) Chromosome
   (b) Start
   (c) End
   (d) (coordinates):(paraclu cluster strength):(TSS prediction strength)
   (e) empty
   (f) Strand
   (g) level - expression level in tpm
   (h) signif - currently empty - will be IDR
   (i) score2 - raw number of reads
2. **. . . tss_low.bed** lists predicted TSS with below 1 tpm expression.
3. **. . . non-tss.bed** lists remaining non-TSS CAGE peaks.

Results:

Prediction accuracy

ROC curves demonstrate a good overall prediction performance of our predictor (Fig. 4). Differences between individual cell lines appear to be mainly due to false positives. Since these correspond to novel promoters, it is not unexpected to see such differences when using a fixed set of annotations to compare against. More pertinent is the below optimal sensitivity (the dashed line corresponds to 0.95 sensitivity) indicating that a small fraction of promoters is missed by the predictor. Manual inspection reveals that these promoters are usually on top or very close to exons confusing the predictor. I will attempt to improve the accuracy for these cases but to be practical, merging GENCODE TSSs and CAGE defined TSS will address this glitch.

Basic Statistics

The number of CAGE elements in the three categories are shown in Figure 5-7. Different sequencing depths and CAGE protocols should have effects on these statistics, especially before IDR. However, it is good to see that for polyA+ libraries approximately 10k TSS CAGE peaks are predicted independently in all samples.

GENCODE overlap

Intersection of the CAGE peaks with GENCODE annotations reveals a strong correspondence between the prediction and annotation. Highly expressed TSS predicted elements commonly overlap +/- 5 GENCODE TSSs (Figures 8 - 12) while non-TSS peaks are enriched in exons, introns and in inter-genic regions.

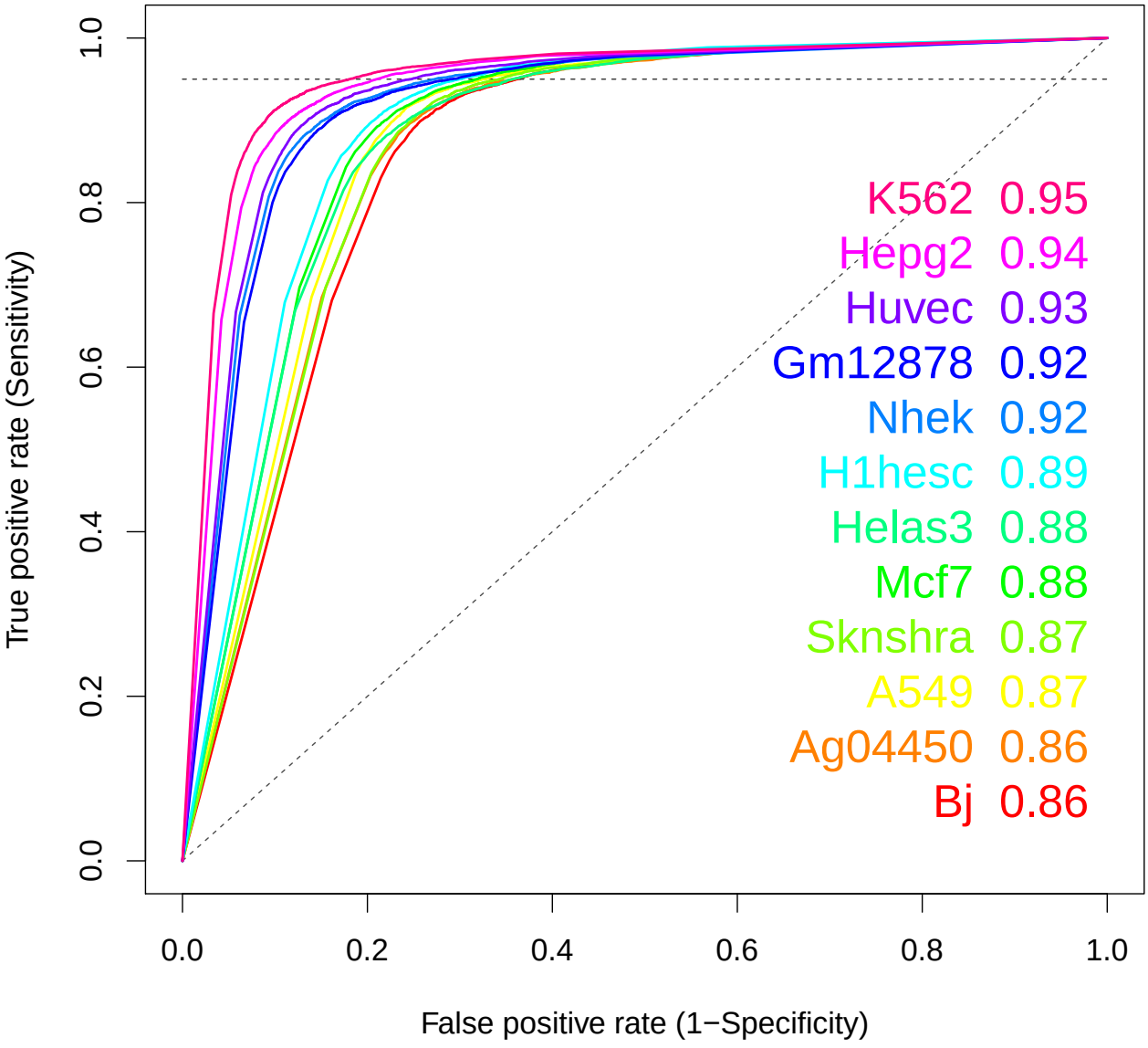


Figure 4: **ROC curve demonstrating the agreement of TSS prediction with known promoter regions.** As the standard of truth we used 10bp windows around known GENCODE gene models.

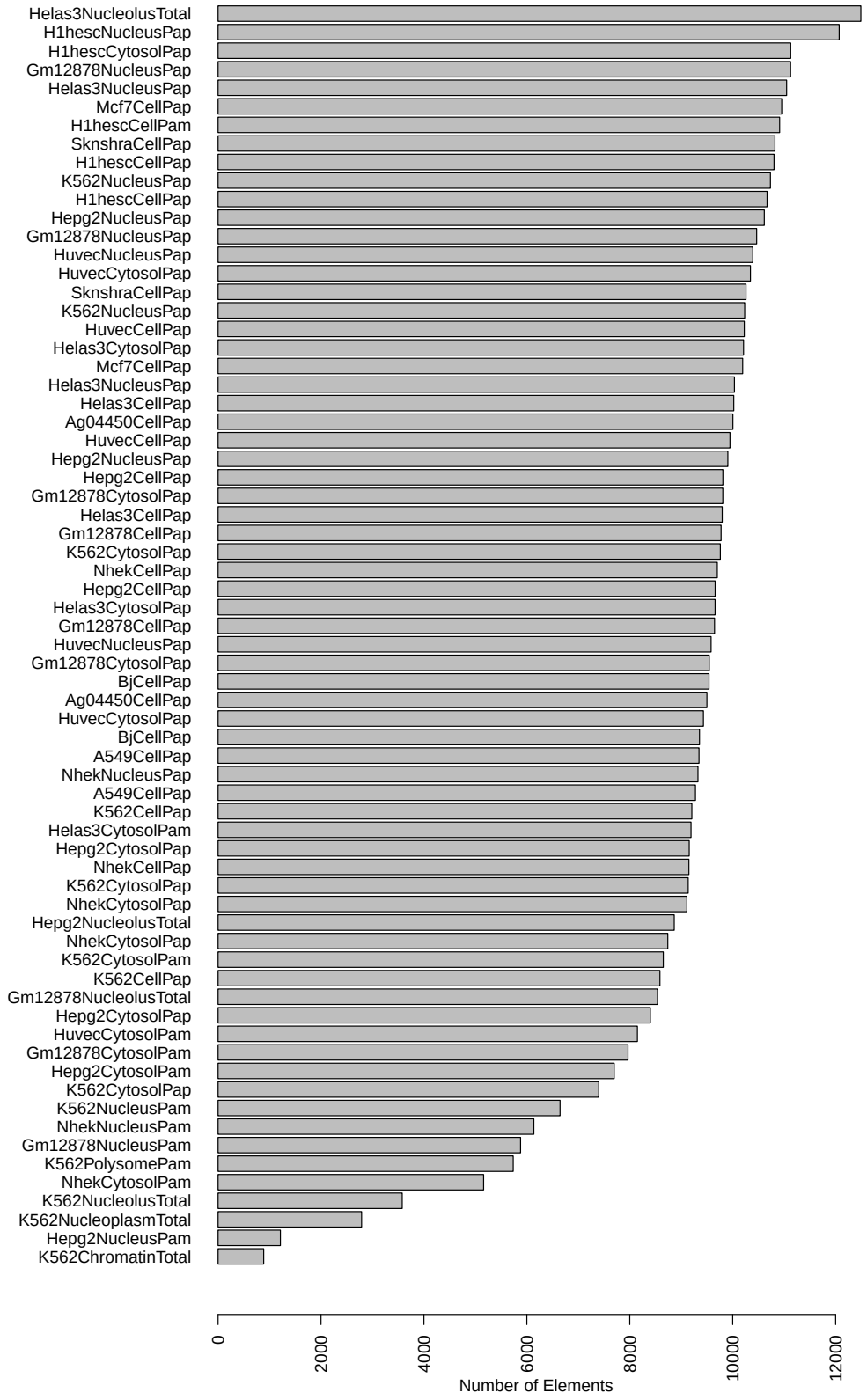


Figure 5: **Number of TSS predicted CAGE peaks with at least 1 tpm.**

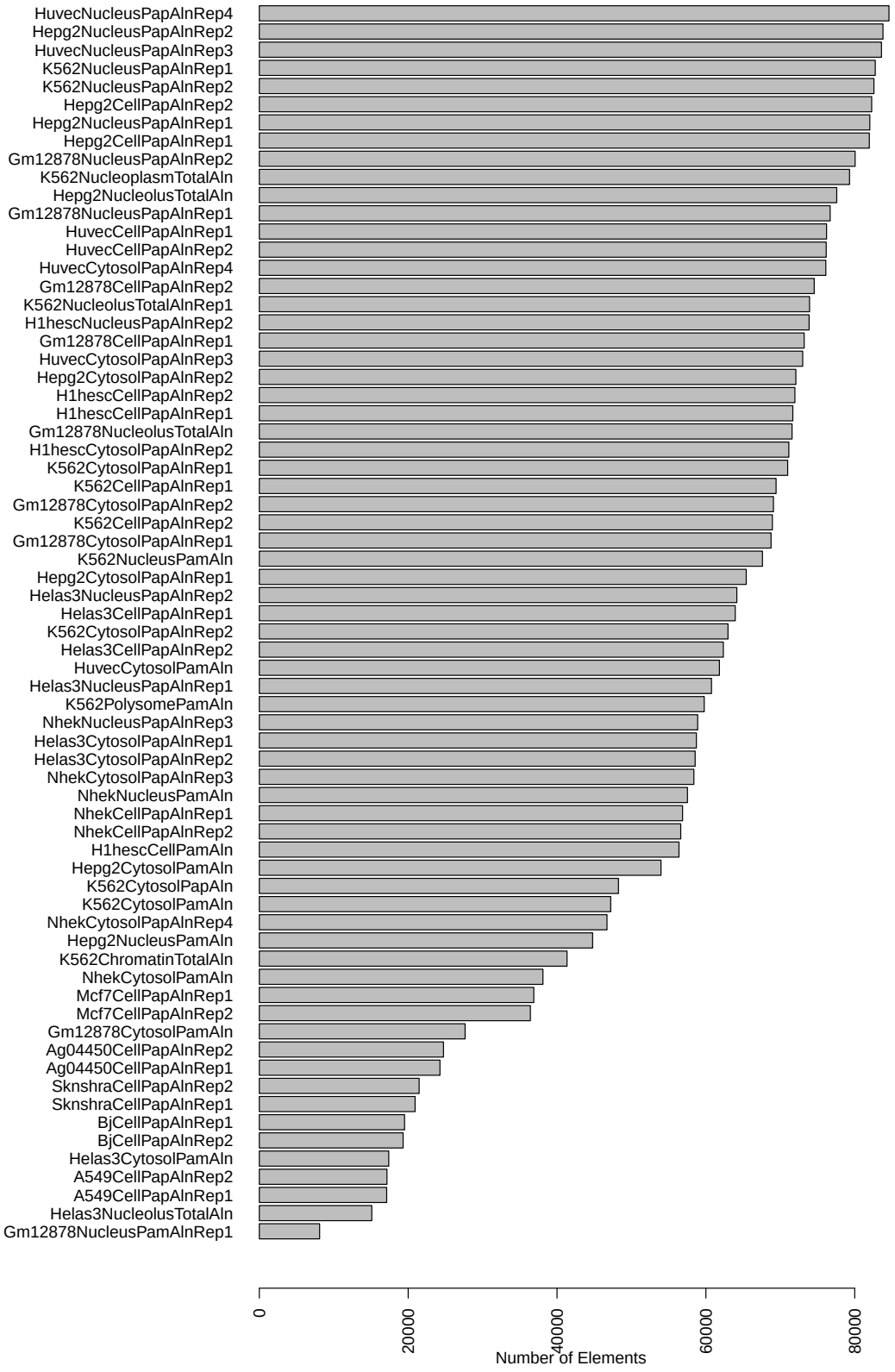


Figure 6: **Number of TSS predicted CAGE peaks with at least $<$ 1 tpm.**

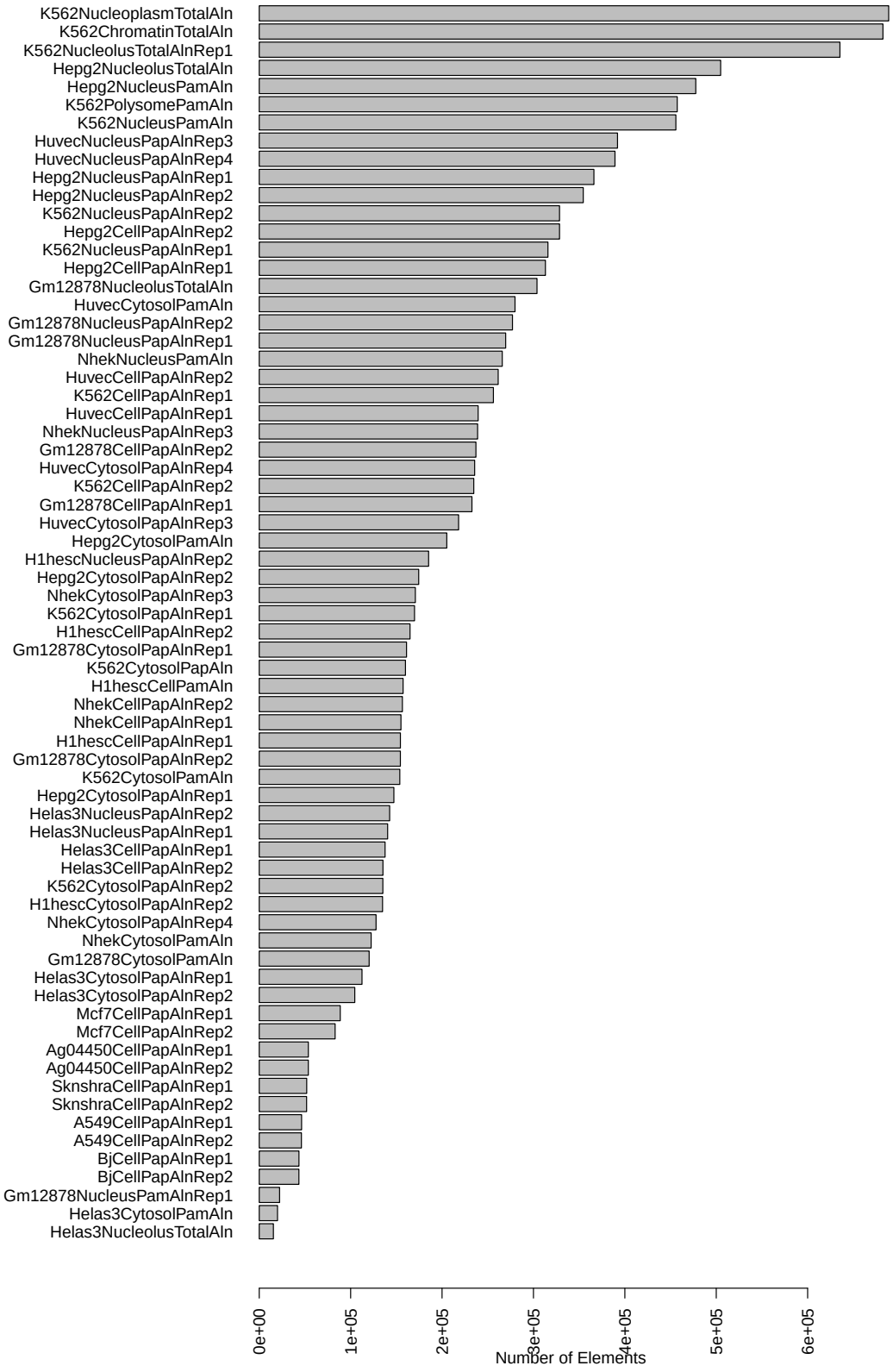
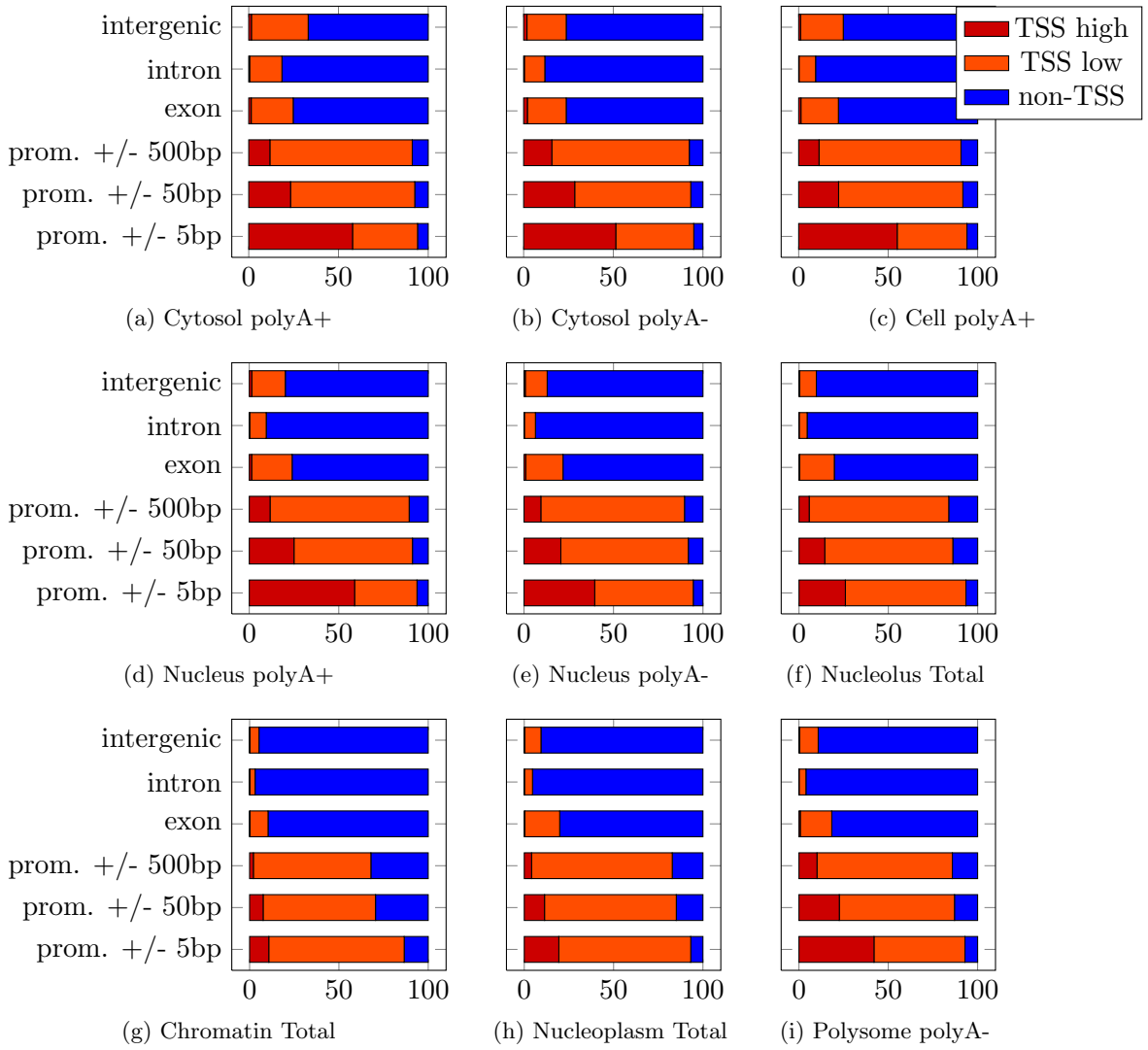


Figure 7: **Number of non-TSS CAGE peaks.**

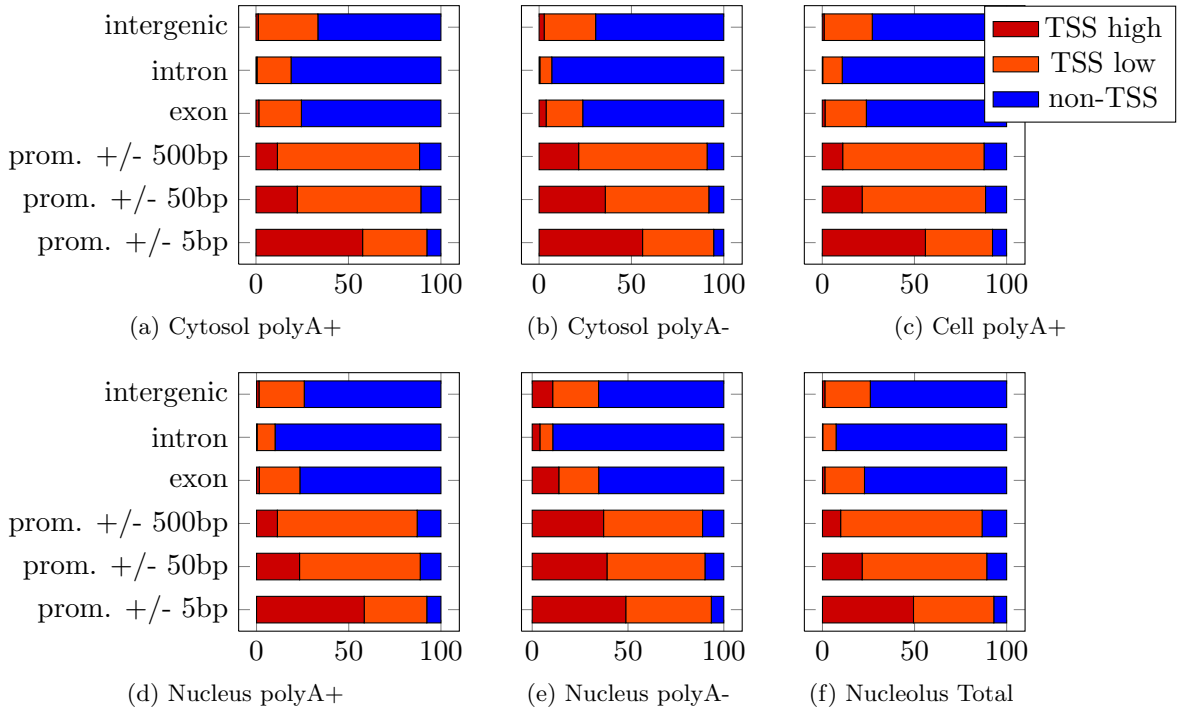


(a) Cytosol polyA+ (b) Cytosol polyA- (c) Cell polyA+

(d) Nucleus polyA+ (e) Nucleus polyA- (f) Nucleolus Total

(g) Chromatin Total (h) Nucleoplasm Total (i) Polysome polyA-

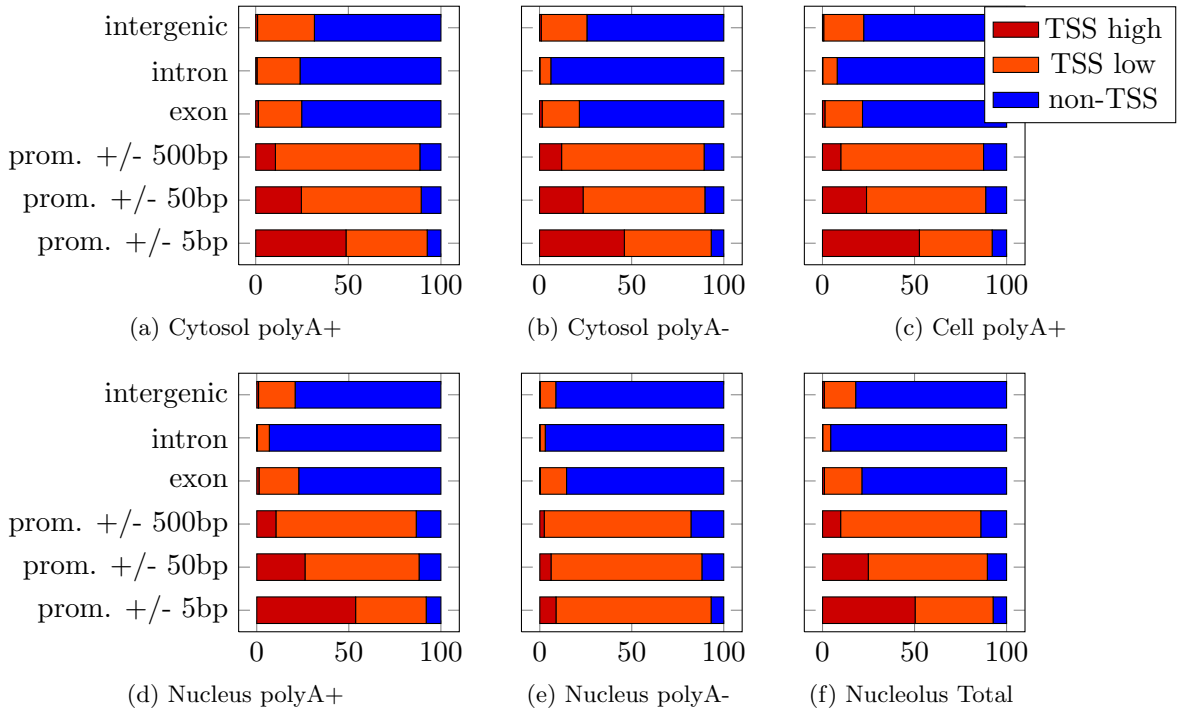
Figure 8: GENCODE annotation of predicted peaks in K562.



(a) Cytosol polyA+ (b) Cytosol polyA- (c) Cell polyA+

(d) Nucleus polyA+ (e) Nucleus polyA- (f) Nucleolus Total

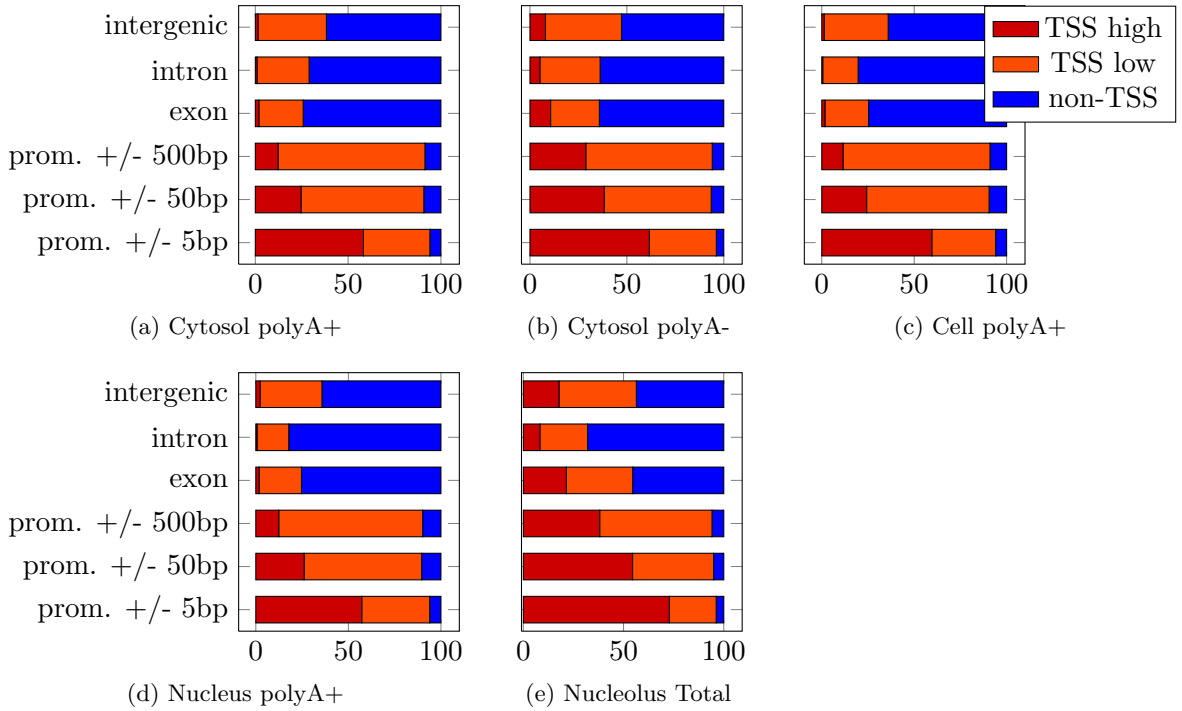
Figure 9: GENCODE annotation of predicted peaks in Gm12878



(a) Cytosol polyA+ (b) Cytosol polyA- (c) Cell polyA+

(d) Nucleus polyA+ (e) Nucleus polyA- (f) Nucleolus Total

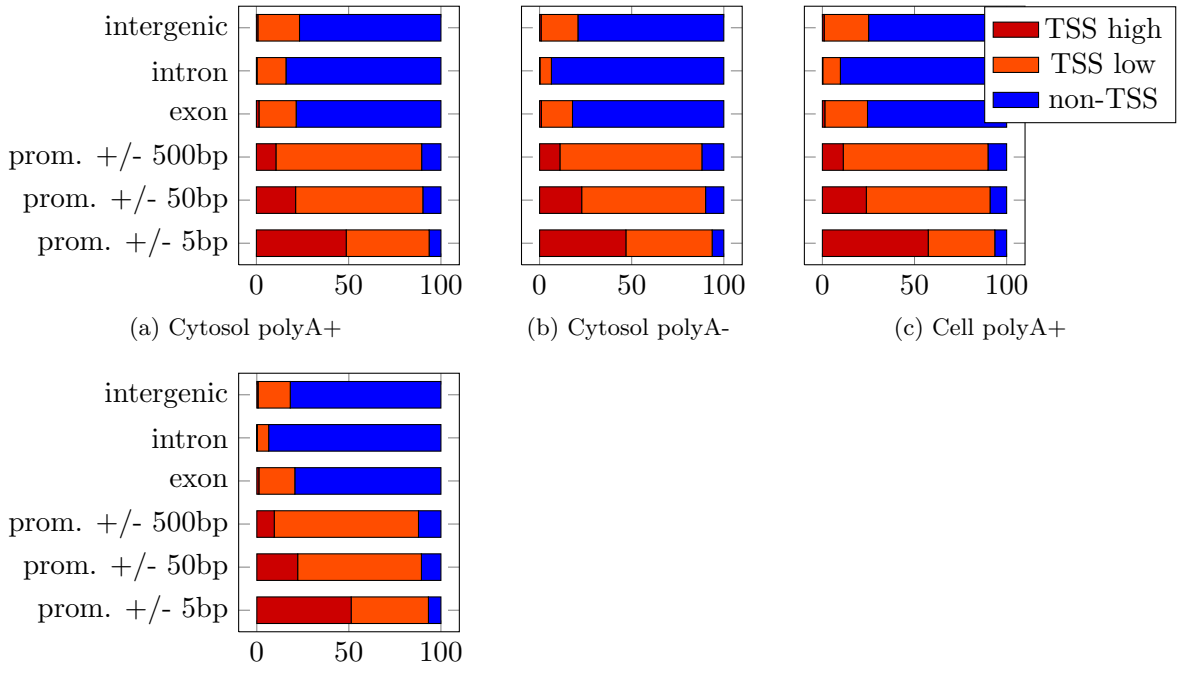
Figure 10: GENCODE annotation of predicted peaks in Hepg2



(a) Cytosol polyA+ (b) Cytosol polyA- (c) Cell polyA+

(d) Nucleus polyA+ (e) Nucleolus Total

Figure 11: GENCODE annotation of predicted peaks in HelaS3



(a) Cytosol polyA+ (b) Cytosol polyA- (c) Cell polyA+

(d) Nucleus polyA+

Figure 12: GENCODE annotation of predicted peaks in Huvec

Bibliography

[1] Frith et al. A code for transcription initiation in mammalian genomes. Genome Res (2008) vol. 18 (1) pp. 1-12