Exploratory analysis of a genomic segmentation with segtools

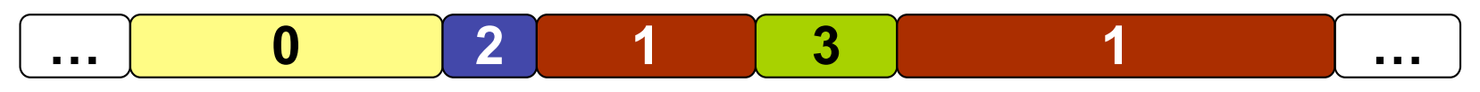
Orion Buske[1], Michael Hoffman[1], and William Noble[1,2]

Departments of [1]Genome Science and [2]Computer Science and Engineering, University of Washington, Seattle, WA 98195

Introduction

ChIP-seq, DNase-seq and other assays are generating whole-genome, single-base-resolution readouts of many genomic properties, including histone modification, open chromatin, RNA expression, and transcription factor (TF) binding. Automatic segmentation methods can label genomic regions that exhibit consistent patterns across such diverse data, but it is difficult to determine if these segmentations are biologically meaningful. We developed a software package called segtools to investigate the properties of a segmentation in a genomic context and to suggest biological interpretations of the segment labels.

segmentation: a partition of a region into segments with each segment assigned one of a small set of labels. For example, a 4-label segmentation:



Results

Segtools was provided with a 25-label whole-genome segmentation. This segmentation was produced using Segway [1] and was based on 31 ChIP-seq and DNase-seq signal tracks (histone modification, open chromatin, and transcription factor binding data) generated by the ENCODE Consortium [2]. Segway was trained on the subset of these data found in nine of the 30 ENCODE pilot regions (0.15% of the human genome). A subset of the results generated by segtools are shown and discussed here.

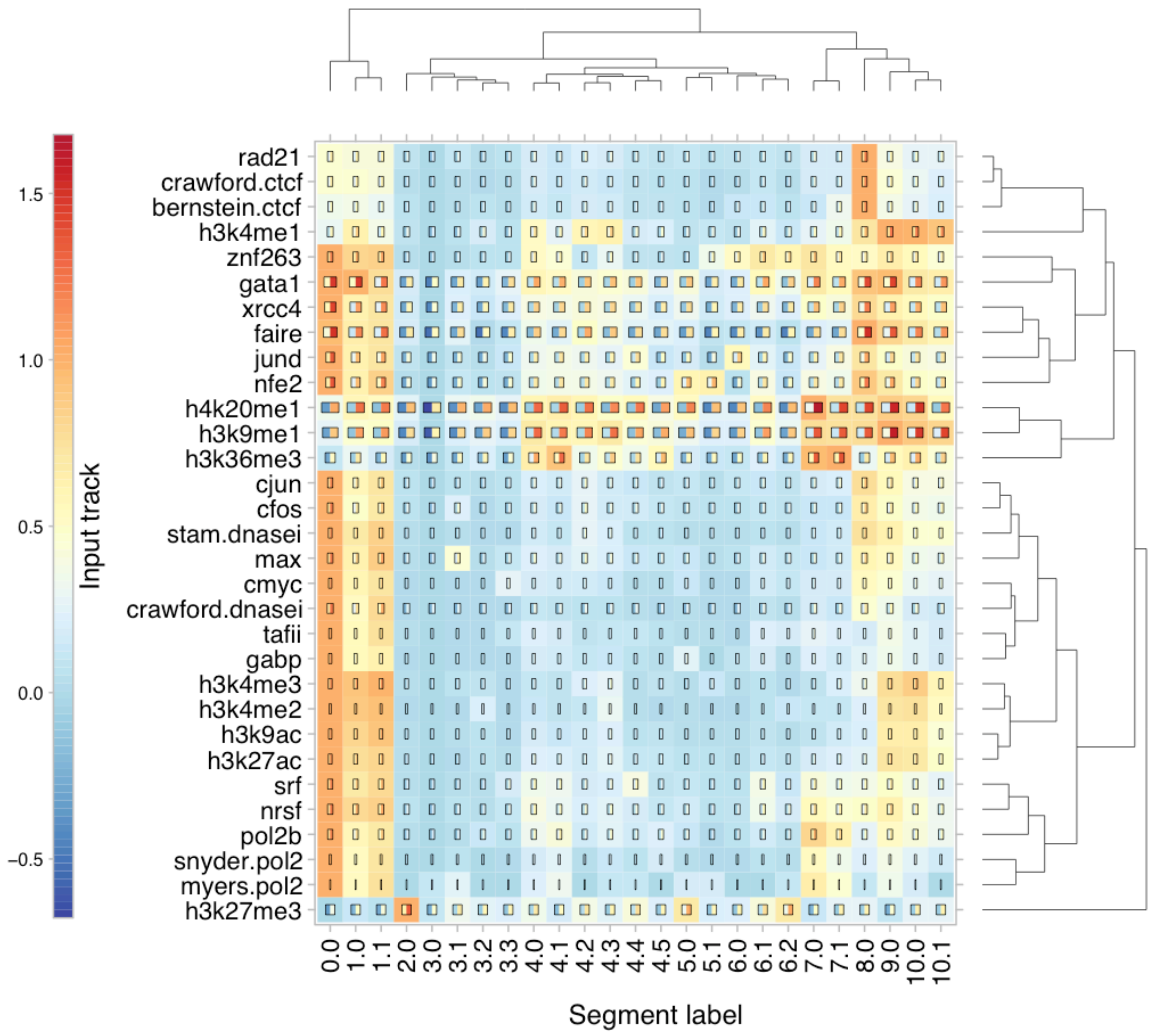


Figure 1: **A hierarchical clustering of the mean values of the 31 signal data tracks used to generate the segmentation.** The background color of each cell corresponds to the track-normalized mean signal value, μ. The horizontal bar inside each cell has a width proportional to the standard deviation, σ, and is filled with colors corresponding to μ ± σ. These data correspond to the theoretical parameters learned by Segway, but a similar plot could be made from observed data. Column-wise hierarchical clustering is used to group similar labels together, and mnemonics are assigned according to this label clustering. From these mean values, one can hypothesize that 0.0 and 1.X are associated with activated, transcribed regions, 2.0 with repression, 3.X with very little activity, and 8.0 with insulator regions. Determining the biological association of the other labels, however, is more difficult.

Mnemonic	*Association*
0.0, 1.X	Transcription initiation
2.0	Repression
3.X	Inactivation
8.0	Insulation
7.0, 9.0, 10.X	Transcription elongation

Interpretation of mnemonics, derived from information in Figure 1.

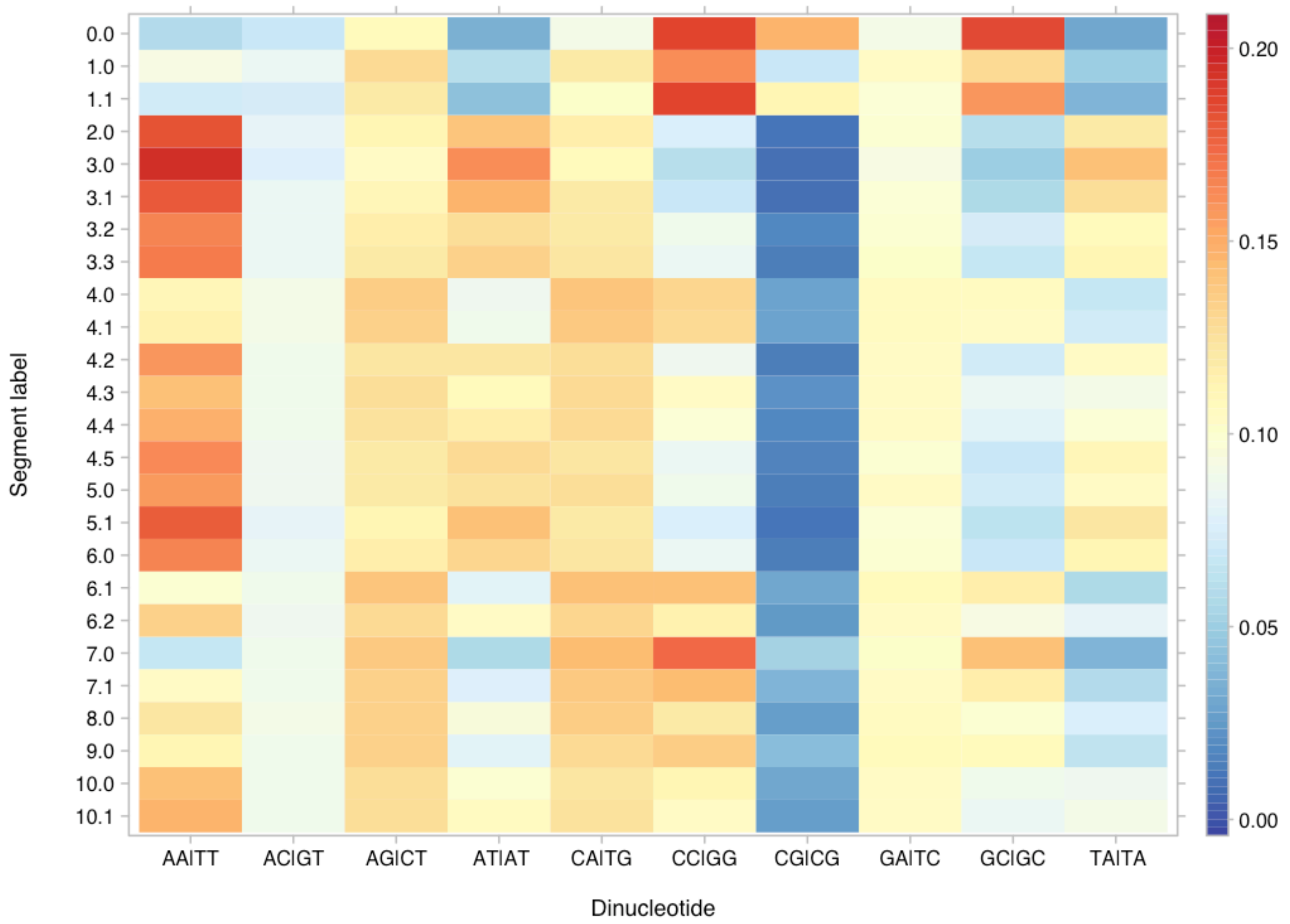
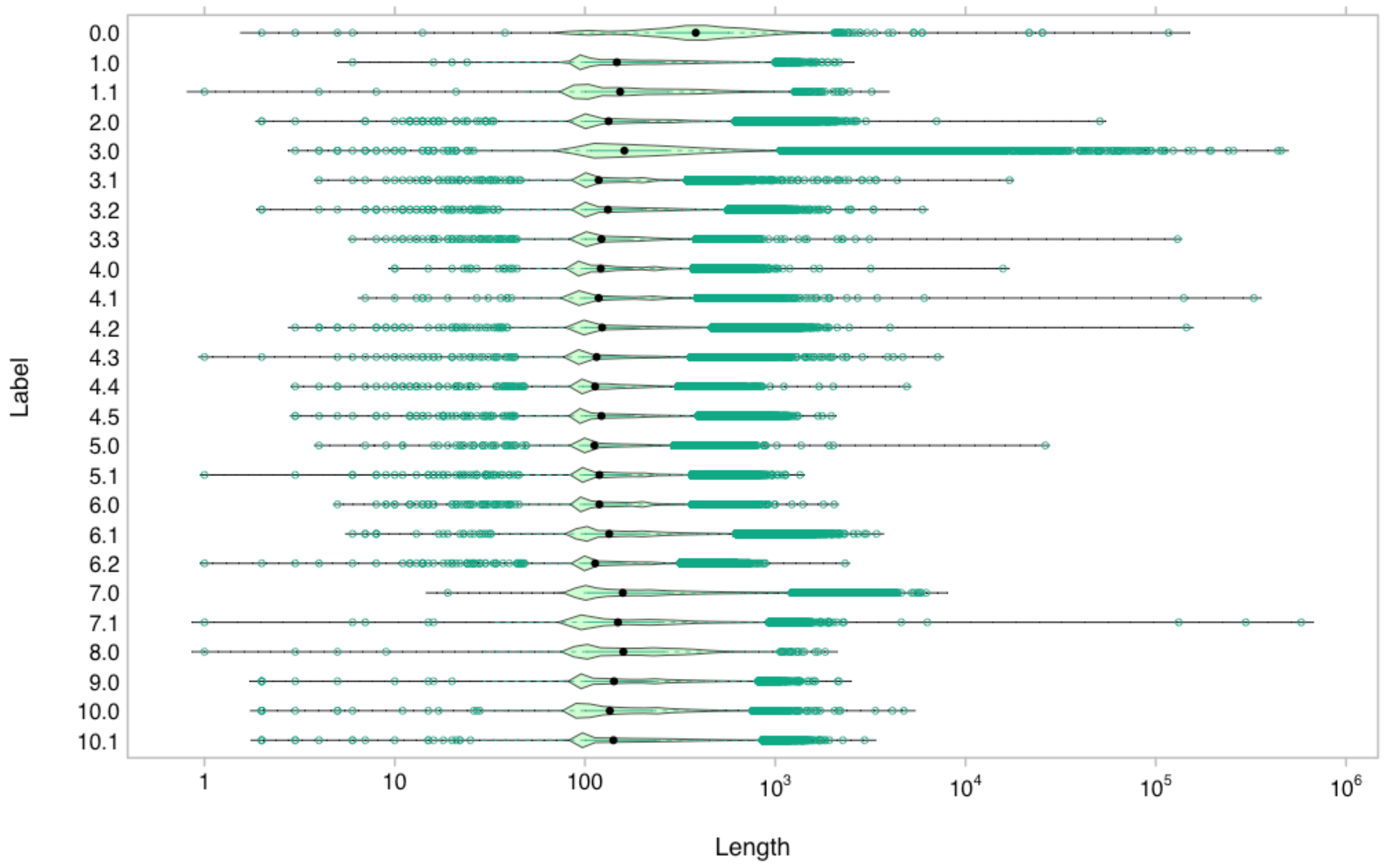


Figure 2: **Dinucleotide frequencies in the segments of each label.** The three transcription initiation labels (0.0, 1.X) all show CpG (as well as CpC and GpC) enrichment, with over 10% of the dinucleotides being CpG in 0.0 and 1.1. Such high levels are a result of the proximity of these labels with the CpG-enriched promoter regions at the start of genes (see Figure 4). ApA is enriched in labels 2.0 and 3.X (as well as 4.2-5, 5.X, and 6.0 to a lesser extend).



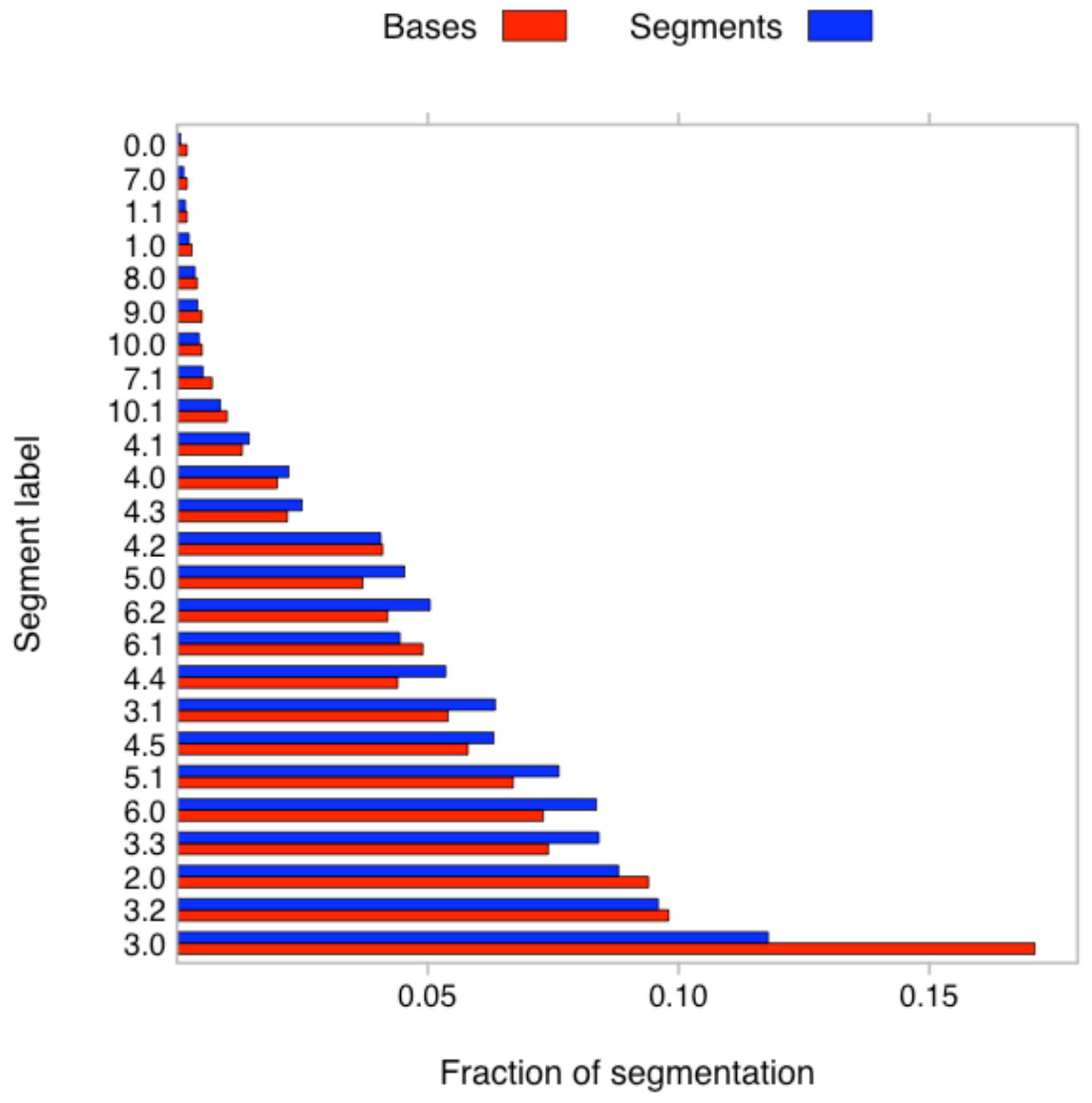


Figure 3: **The distribution of segment lengths for each label (above) and the fraction of the overall segmentation allotted to each label (at right).** In the plot above, outliers are circled and the median is labeled with a black dot. The effect of a 100-bp minimum threshold on segment length can be observed: the left edge of most distributions align with this threshold, but the short outliers revealed a minor problem in the segmentation method. The long outliers, however, were found to be a result of artifactual data on chromosome Y (for data from female cell lines). The plot at right shows the fraction of the total bases and total segments that are found in each segmentation. About a third of the segment labels are quite rare (these correspond to regions with strong signals or clear structure), with the remaining labels spanning a wide range from 2% to 17% of the segmentation.

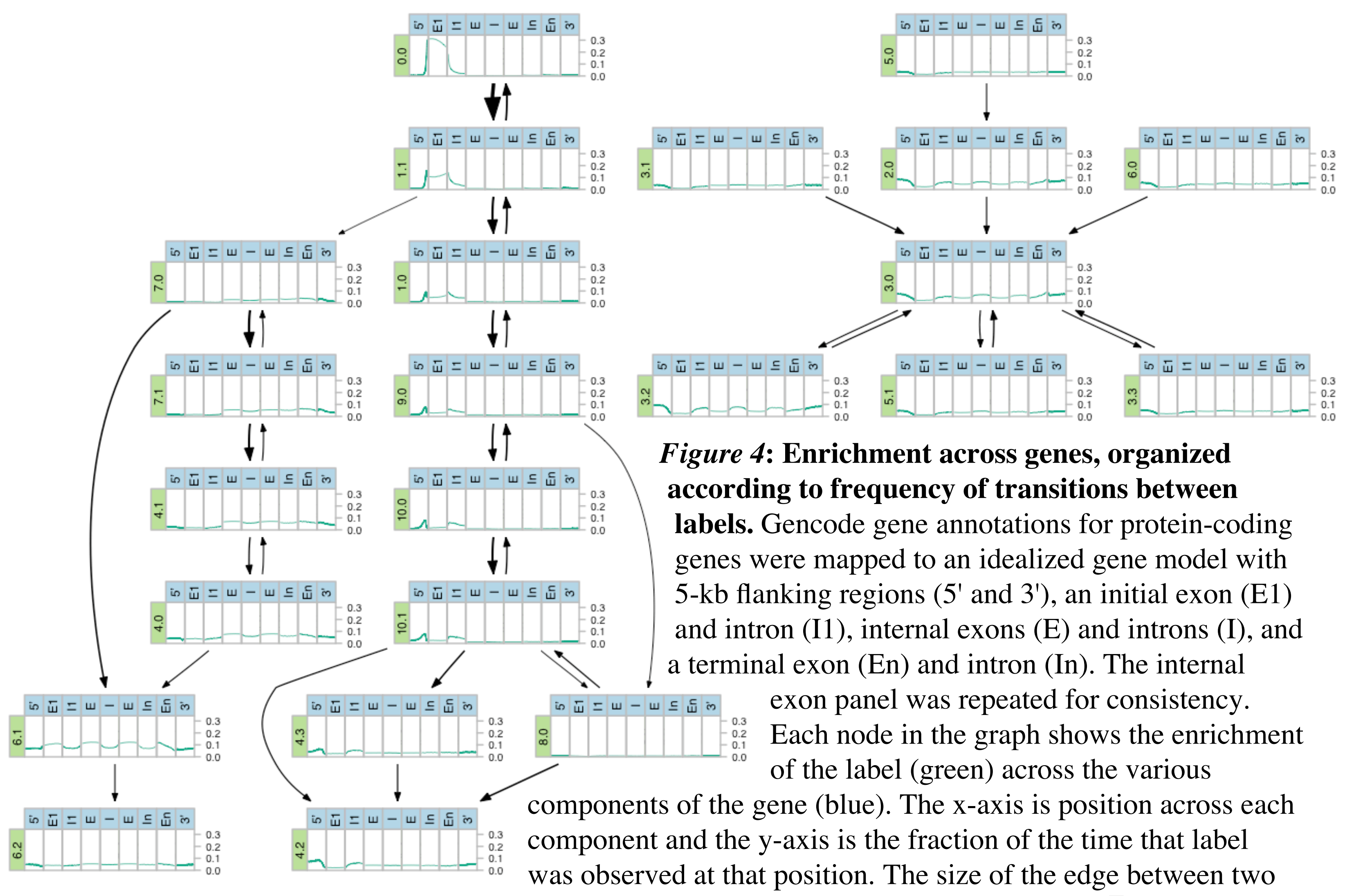


Figure 4: **Enrichment across genes, organized according to frequency of transitions between labels.** Gencode gene annotations for protein-coding genes were mapped to an idealized gene model with 5-kb flanking regions (5' and 3'), an initial exon (E1) and intron (I1), internal exons (E) and introns (I), and a terminal exon (En) and intron (In). The internal exon panel was repeated for consistency. Each node in the graph shows the enrichment of the label (green) across the various components of the gene (blue). The x-axis is position across each component and the y-axis is the fraction of the time that label was observed at that position. The size of the edge between two nodes, *a* and *b*, is proportional to the probability p_{ab} of that transition. An edge is shown if $\sum_i p_{ai} > 0.15$ *or* $\sum_j p_{jb} > 0.15$. The 15 labels in the left connected component are associated with gene structure, with six labels enriched in gene 5' ends, four labels enriched in gene 3' ends, two with little enrichment in genes, two broadly enriched upstream of genes, and one that is differentially enriched in exons.

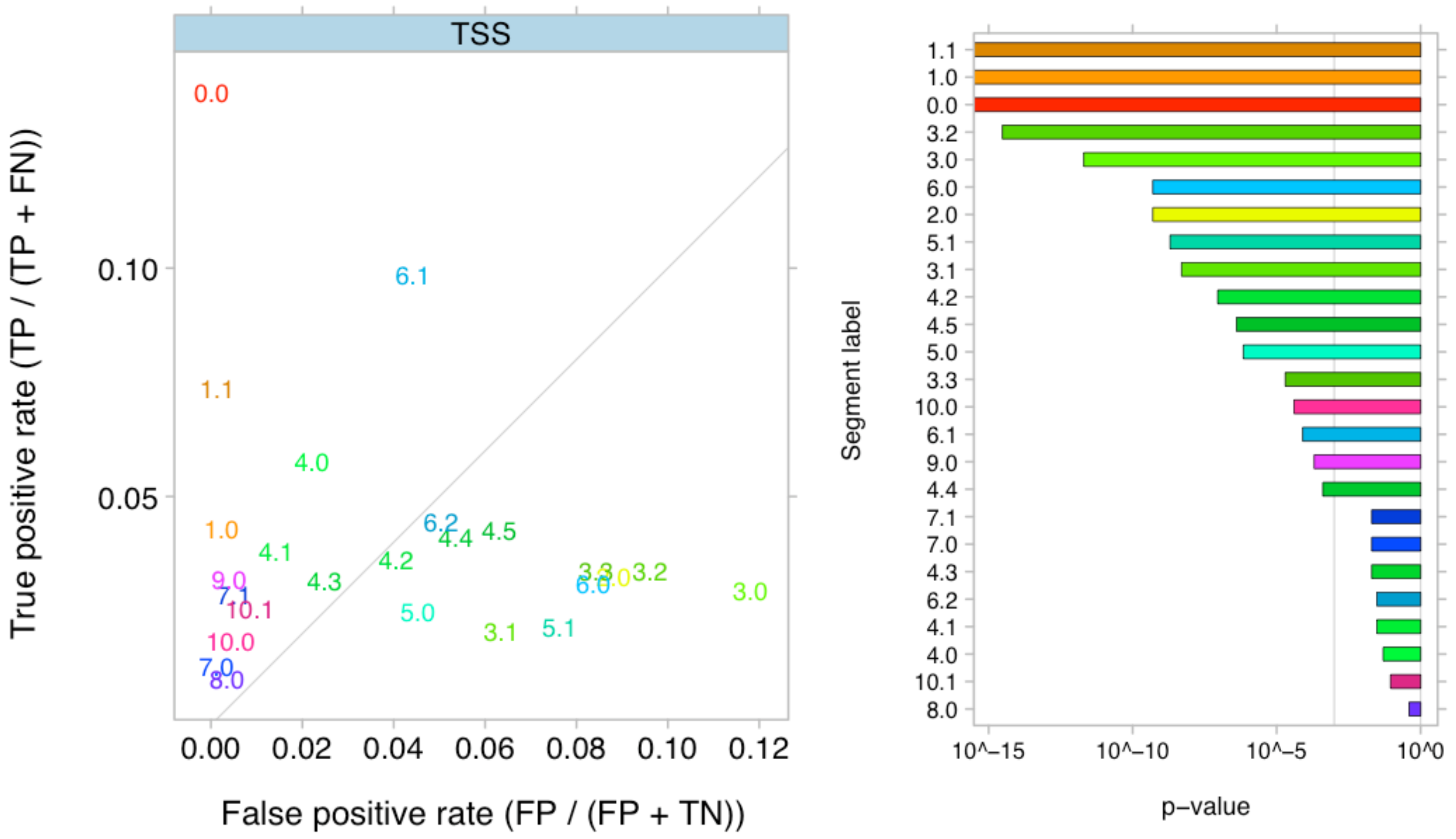


Figure 5: **Predictive value of the overlap between segment labels and Gencode transcription start site (TSS) annotations.** Every overlap of a segment with a TSS is considered a true positive (TP) for that segment's label but a false negative (FN) for every other label. Similarly, every segment that does not overlap a TSS is considered a false positive (FP) for that segment's label but a true negative (TN) for every other label. The predictive value of each segment label can thus be plotted in ROC-space (left). The line of no discrimination *(y = x)* is shown. The significance of overlap was calculated by using the Genome Structure Correction (GSC) tool [2] to compensate for non-uniform distribution of segments and bases between labels (right). The p-values of the 0.0 and 1.X are below GSC's resolution.

Conclusions

By presenting the segment labels in a variety of genomic contexts, the quality and character of a segmentation can be quickly evaluated. Segtools automates the exploratory analyses that are necessary to begin understanding a segmentation and identifying areas of additional investigation.

References

[1] Hoffman MM, Buske OJ, Bilmes JA, Noble WS. Segway: a dynamic Bayesian network for genomic segmentation. In preparation.

[2] ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447(7146):782-3.