Supplementary data for “Integrative annotation of chromatin elements from ENCODE data”

Supplementary methods

Signal track generation

The ENCODE datasets span five main data types across a variety of cell-types and treatment conditions, as summarized in Supplementary Table 1. The complete list of datasets is available at the ENCODE Data Portal (http://genome.ucsc.edu/ENCODE/downloads.html). Each dataset is the result of at least two biological replicates. For a select subset of targets, multiple laboratories generated their own versions of the datasets. Laboratories generated datasets using a variety of protocols and with a diversity of sequencing parameters (such as library size, number of mapped reads, and read lengths). Hence, we developed a uniform processing pipeline to generate genome-wide signal coverage tracks by pooling data from multiple replicate experiments and combining appropriate datasets across labs when available (https://sites.google.com/site/anshulkundaje/projects/wiggler). We implemented this pipeline using the Wiggler package (http://code.google.com/p/align2rawsignal). It accounts for the individual characteristics of and differences between specific datasets and data types.

First, we downloaded from the ENCODE Data Portal all Binary Alignment/Map (BAM) files containing reads mapped to the GRCh37 human genome assembly submitted by the ENCODE production groups. To eliminate artificial differences between datasets due to different mapping stringency parameters we created custom unique-mappability tracks for the GRCh37 male and female genomes (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/referenceSequences) for read lengths ranging from 20 to 54. For a particular read length k, we labeled each coordinate on the ‘+’ strand of the genome as unique if the k-mer starting at that position and continuing 3' mapped uniquely to only that position with no mismatches. If a position x on the ‘+’ strand is labeled unique, it implies that the k-mer starting at position $x + k - 1$ on the ‘−’ strand is also unique. Hence, one can infer mappability values for any position on the ‘−’ strand directly from the ‘+’ strand mappability track. The mappability tracks are available for downloading (http://code.google.com/p/align2rawsignal). We used the mappability tracks to filter BAM files by discarding all reads that mapped to *non-unique* locations in the genome.

The sequenced reads in each dataset represent ends of target DNA fragments isolated in various ways depending on the data type. The experimental protocols result in a variety of characteristic strand-specific distribution of sequenced reads around target sites. For example, ChIP-seq datasets typically show mirror peaks of mapped reads on the ‘+’ and ‘−’ strand around binding sites of the target protein separated by a characteristic distance equal to the average lengths of the immunoprecipitated DNA fragments (44,45). Therefore, in order to faithfully represent the target of interest and generate signal tracks that show peak signal at the target binding site rather than in flanking regions, it is important to account for this strand-specific *read-shift*. We used strand cross-correlation analysis (44,46) to infer the predominant read-shifts for each replicate in each dataset. Kundaje et al. discuss read distribution characteristics of the different data types. The complete set of dataset-specific read-shift parameters are provided at

https://sites.google.com/site/anshulkundaje/projects/wiggler. The Duke and University of Washington (UW) production groups used very different protocols for performing DNase-seq, and the read-shift characteristics of their datasets diverged greatly, so we did not pool datasets across the two labs, unlike for ChIP-seq.

Multiple filtered BAM datasets, B_i, corresponding to replicates (or similar experiments from multiple labs), along with their respective estimated read-shifts, L_i, were provided as input to the signal processing engine. For each dataset B_i, we performed the following procedure:

(1) We shifted read starts in the 3' direction by $L_i / 2$.

(2) We computed shifted read-start coverage at each position on both DNA strands.

(3) For each genomic-location x and strand s, we then computed a smooth weighted sum of read counts F_{is} using a kernel of width w centered at x. We used different values of w for different data types (https://sites.google.com/site/anshulkundaje/projects/wiggler). For each data type, we selected w based on the maximum estimated fragment length for any dataset of that type (for any dataset i of data type j, we expect that $w_{jj} \geq L_{ij}$), and the general characteristics of the data type. For example, we set $w = 300$ for histone mark datasets because we expected that the sonication and size-selection protocols for the corresponding experiments would generate DNA fragments of size ≤ 300 bp. For all data types except nucleosome data, we used the Tukey window kernel, which has a central window of length c_w ($c_w \leq w$) with weights equal to 1. The weight then tapers to 0 on either end following a cosine curve. We typically set $c_w = L_i$. This procedure of shifting reads and then aggregating over a window is equivalent to a smooth read extension of length w. Hence, this aggregate signal value at each position represents the approximate number of sequenced fragments that overlap that position. The use of a common overall smoothing length for all datasets of a particular data type while allowing for dataset-specific (and replicate-specific) read-shifts provides equivalent and comparable resolution across all datasets of a data type while accounting for the dataset-specific fragment length distributions. For nucleosome data, which requires finer smoothing to distinguish individual nucleosomes, we used a sharper tri-weight kernel with a window size of 60 bp around each position after shifting reads by the dataset-specific read-shift (typically 148 bp / 2 = 74 bp) (Valouev et al. 2011).

(4) We added together the unnormalized fragment counts from both strands at each position x ($F_i(x) = F_{i+}(x) + F_{i-}(x)$).

(5) Since we used only unique mapping reads from each dataset, we needed to distinguish between mappable positions with zero signal for a particular dataset from those positions that were simply not uniquely mappable (missing data). We also needed to normalize $F_i(x)$ for the total number of mappable locations that contribute signal to any position x. Different datasets also have different library sizes. In order to compute signal on a common scale across all data sets and data types, we needed to normalize for differences in number of replicates (or pooled datasets) and the total number of mapped reads. We accomplished these normalizations thus:

> (a) Using the binary unique-mappability track corresponding to the read length for B_i, we computed the local cumulative mappability $M_i(x)$, for each position x by using the same read-shift and aggregation smoothing kernel window parameters. $M_i(x)$ represents the effective number of positions in the local kernel window around x that can potentially contribute read counts based on mappability.

(b) For each B_i, we then computed the expected fragment counts at each position x if all the mapped reads were uniformly distributed over all uniquely mappable locations on both strands in the genome $E_i(x) = (M_i(x)\ R_i) / G_i$ where R_i is the total number of mapped reads in B_i and G_i is the total mappable genome size over both strands

(c) We then add fragment counts from all B_i at each position x: $F(x) = \Sigma_i\ F_i\ (x)$.

(d) We also add expected fragment counts from all replicates/datasets at each position: $E(x) = \Sigma_i\ E_i(x)$.

(e) We then compute the normalized signal $S(x)$ at each position x as the fold-change of the observed fragment count over the expected fragment count. $S(x) = F(x) / E(x)$. This signal scale is similar to the reads per kilobasepair per million mapped reads (RPKM) measure often used to represent normalized RNA-seq read counts.

(6) Finally, we discard signal at positions that lie in regions of low dispersed mappability as these locations typically show mapping artifacts and also because the signal normalization procedure can over-correct for the small number of locations within the smoothing window that contribute signal. Specifically, we discard positions x that have $E(x) \leq 0.25\ \max_x E(x)$, including positions lying in completely unmappable windows. We represent the signal at such locations as missing data, but assign a non-negative signal value to all other positions. Therefore a position with signal value 0 is a reliable mappable location, but no reads map to it in a particular dataset. Sample screenshots of different types of data (ChIP-seq input control, CTCF and H3k4me3) are shown in Supplementary Figures 1–3.

All normalized signal tracks in bedGraph and bigWig formats are available at https://sites.google.com/site/anshulkundaje/projects/wiggler, including the tracks used as segmentation input (Supplementary Table 1).

Assignment of mnemonics to Segway labels

We assigned short human-interpretable mnemonic codes to the Segway segmentation labels through a combined process of comparing emission parameters with known biological features, and examining the enrichment of the labels for various genome annotations. One can easily identify TSS-associated labels by their enrichment of activating histone marks, such as H3K4me3 and H3K27ac, and strong Pol2, DNase, and FAIRE signal. These labels also show strong enrichment near the TSS in gene structure aggregation plots (Figure 1, Supplementary Figure 4). We segregated TSS-associated labels into two categories—Tss and TSS-flanking (TssF)—based upon the shape of the aggregation plots. The Tss labels are also associated with higher signal levels than the TssF labels. We assign labels that associate with the promoter of protein-coding genes in a less punctate fashion to the promoter (Prom) mnemonic, with modifiers for flanking (F) or weak (W) labels based upon the strength of the H3K4me3 and H3K27ac signal, and the shape and magnitude of the aggregation plots. Within the H1 hESC and HepG2 cell lines, we observe TSS-associated labels that show evidence of a bivalent (47) or poised promoter label (PromP), based on the simultaneous presence of H3K27me3 and H3K4me3.

We identify elongation labels primarily on the basis of their enrichment within genes, as shown on the gene aggregation plot, as well as the presence of histone mark H3K36me3. Among the various elongation labels, we assign the “Gen3'” label to those that show a relatively strong enrichment near the 3' end of the gene, although we note that these labels do not solely occur near the 3' end.

Strong enrichment for CTCF signal provides the rationale for the “Ctcf” label. In some cell types there is an additional CTCF label with relatively higher levels of open chromatin, which we call “CtcfO.”

We find enhancers primarily by looking for a pattern of high H3K4me1 signal and lower H3K4me3 signal. The resulting labels are separated into “weak” and “flanking” categories based upon weaker H3K4me1 and H3K27ac signals.

In the H1 hESC, HeLa-S3, and HUVEC cell lines we observe a label for which the emission parameter associated with FAIRE is relatively large, but shows no signal for DNase and no striking pattern in the gene aggregation plots. Accordingly, these labels are assigned the mnemonic "Faire." Similarly, in H1 hESC, we observe a label that shows strong signal in the DNase track from Duke, but not in the UW DNase data. Indeed, this label does not show strong signal for any other track. It is assigned the mnemonic "DnaseD."

Among the remaining labels, we assign a label to a repressed (Repr) mnemonic if the associated emission parameters indicate the presence of repressive histone mark H3K27me3. A label is quiescent (Quies) if it shows no signal in any track. In general, the emission parameters associated with the "Control" signal are difficult to interpret and are largely ignored in the assignment of mnemonics. The remaining labels exhibit some weak signal across the various tracks (except for Control and H4K20me1), and we call them "Low".

Creating the combined segmentation

First, for each segmentation, we identified states that we could group based on similar signal patterns. For the ChromHMM segmentation, the states were grouped manually based on the mean emission parameters across multiple cell lines. For the Segway segmentations run independently over multiple cell lines, multiple hierarchical clustering techniques were applied across all states in the segmentations to identify the most consistent clustering of states, both across cell lines and with respect to existing biological knowledge. Using these criteria, the Ward clustering on Euclidean distances between mean signal scores transformed to the unit interval was chosen to cluster the Segway state labels. Second, we identified pairwise relationships between the ChromHMM and Segway merged states using both overlap calculations and manual annotation (Supplementary Figure 10). Pairs of states that were viewed as concordant were assigned to one of the seven state classes. Regions of the genome occupied by concordant states between the two initial segmentations were reassigned to the new summary labels. In some cases there were combinations of states between the two segmentations that could not be reconciled, and these combinations were viewed as discordant. Regions with discordant states were not assigned a state label, and were dropped out of the summary combined segmentation. Because there are multiple concordances between segmentations, the general effect of this process was to produce longer continuous states particularly compared to Segway, but at high resolution the exact state boundaries are clipped down to the region of concordance.

Enrichment/depletion of RNA-seq in segmentation classes

We used ENCODE Project RNA contig data (ftp://genome.crg.es/pub/Encode/data_analysis/ForDeadZones/Contigs_IDR0.1_CSHL.tar.gz) to investigate how enriched segmentation labels are for different kinds of RNA. This data comprises the ENCODE CSHL RNA-seq contigs (continuous regions covered by uniquely aligned reads) that passed irreproducible discovery rate (48) threshold of 0.1 for a given

experiment performed in two biological replicates, and combines contigs from both from shotgun long RNA-seq and short RNA-seq.

We counted the number of contigs of each annotated transcript category (a "biotype") that overlaps the labels in the 12 Segway and ChromHMM segmentations in each individual cell type. The enrichment or depletion of each biotype in each label is shown in Figure 4.

We have made available the Python, R, shell script, and C code used to perform this analysis at http://www.bx.psu.edu/~rsharris/encode/manuscripts/segmentation/bundle_segmentations_vs_rnaseq_biotypes.tar.gz.

Fraction of protein-coding genes overlapping RNA-seq contigs

We combined protein-coding transcripts from GENCODE annotation levels 1–3 (validated, manually-annotated, and automatically-annotated transcripts; ftp://ftp.sanger.ac.uk/pub/gencode/release_7/gencode7_GRCh37.tgz), merging transcripts with the same gene symbol. This procedure results in nearly all genes as single intervals, but a few multi-interval genes then remain (20,677 genes, 20,713 intervals). We then tested those genes for overlap with all labels in the 12 Segway and ChromHMM segmentations in each individual cell type, creating a list of genes and intervals that overlap each label by at least 1 bp of overlap, and truncating to the overlap intersection. The number of distinct gene names identified in this step is the denominator in the plots in Supplementary Figure 20.

On a state-by-state basis, the truncated genes were tested for overlap with the same cell RNA-seq contigs. Only contigs annotated as protein coding were used, and 1 bp of overlap was considered sufficient. Of those truncated genes that have overlap, the number of distinct gene names is the numerator in the plots.

We have made available the Python, R, shell script, and C used to perform this analysis at http://www.bx.psu.edu/~rsharris/encode/manuscripts/segmentation/bundle_segmentations_vs_rnaseq_biotypes.tar.gz.

Distribution of RNA expression score for protein coding genes

We tested the truncated genes from the previous section for overlap of at least 1 bp with protein-coding RNA-seq contigs from the same cell type on a label-by-label basis. We collected the annotated expression scores from these contigs, and used only the maximum expression score for each interval that had more than one contig. We then used these scores as the data displayed in the box plot in Figure 4.

Enrichment and depletion of repeats and mappable regions in segmentation labels

We merged together all RepeatMasker (49) regions, as downloaded from the UCSC Table Browser, and then used the merge to determine which segmentation bases intersected repeat bases. We did not use repeat family information in this analysis.

We derived uniquely mappable regions from the 50-mer mappability track, on the UCSC Genome Browser, by first extracting intervals with a mapping value of 1.0 (indicating unique mappability along one strand). To allow for reads that are uniquely mappable on the opposite strand, a position *x* is considered mappable if a read beginning at *x* or an opposite strand read ending at *x* is unique. Thus the union of the intervals with a copy of the intervals shifted 3' by 49

bp yields regions which are uniquely mappable along at least one strand. The number of bases intersecting each state was then counted.

The expected number of bases for a feature and label is then derived from the total base count for that feature and the percentage of bases covered by that label. For example, if a label covers 20% of the genome, we expect 20% of the feature bases to occur in this label.

Emission parameter heat maps

We used Segtools (50) to create heat maps of emission parameters (Figure 1, Supplementary Figure 6). To increase the contrast in Segway emission parameter visualization, we truncated the color palette at the 95th percentile of mean values.

Gene structure and p300 enrichment aggregation

We used Segtools (50) to create the gene structure (Figure 1, Supplementary Figure 4) and p300 enrichment (Supplementary Figure 18) aggregation plots. We used GENCODE (51) version 7 levels 1–3 gene annotations. For the p300 plots, we used SPP (44) peak calls of HepG2 p300 peaks provided by the ENCODE Project (52,53)

Precision and recall

We calculated precision and recall between segmentations and GENCODE genes with reproducible cap analysis gene expression (CAGE) support using Segtools (50). We obtained data on the number of tags from CAGE experiments mapped to clusters overlapping a GENCODE TSS (ftp://genome.crg.es/pub/Encode/data_analysis/TSS/Gencodev7_CAGE_TSS_clusters_June2011.gff.gz ; ENCODE Project Consortium 2012; Djebali et al. 2012), and deemed genes with two or more tags mapped as having reproducible CAGE support. For each segmentation label, we defined true positives (TP) as the number of bases assigned to that label and a reproducible CAGE TSS, false positives (FP) as the number of bases assigned to that label but not a reproducible CAGE TSS, and false negatives (FN) as the number of bases assigned to some other label and a reproducible CAGE TSS. We defined the base-level precision as TP / (TP + FP), and the base-level recall as TP / (TP + FN).

Supplementary Results

Comparing the segmentations

The two segmentation methods exhibit strong concordance across most of the genome. However, there are specific methodological differences between the segmentation approaches (Table 1). ChromHMM uses a 200 bp windowed approach across the genome with a binary classification of the signal for each of the input data sets in each window. Segway utilizes the continuous range of the normalized signal for each data set considering each genomic base as input to the machine learning. These differences in approach have performance implications for each method (it is faster to run ChromHMM than Segway) but more importantly result in differences in the qualitative characteristics of the segments generated. By design, the minimum segment size for each ChromHMM state is 200 bp, and segments are generated in multiples of this unit length, while Segway expects a segment size of 100 bp but can produce both smaller and larger segments over a single base incremental range. The effect of these differences is that ChromHMM produces on the average larger segments than Segway (median GM12878

segment length: 600 bp ChromHMM, 157 bp Segway; Supplementary Figures 8 and 9). ChromHMM has fewer state changes per kilobasepair (for GM12878, 0.37 transitions per kbp) than Segway (4.8 transitions per kbp) producing more continuity across the genome. This phenomenon is particularly noticeable for instance in transcribed genes where Segway generates frequent flipping behavior between multiple related transcribed gene states. On the other hand, the greater freedom that Segway has to place segment transitions provides higher resolution that potentially has advantages for defining the boundaries of states. Manual inspection of the smaller states such as those at putative enhancers or TSSs suggested that this property of Segway facilitated a more precise representation of the underlying signal shapes in these cases. We surveyed a small number of users within the ENCODE project on the utility of the two segmentations in their own regions of interest, and received feedback that reinforced our own analyses and manual inspection. In short, they liked the continuity provided by ChromHMM over larger genomic regions such as genes, but valued the resolution provided by Segway at specific elements of interest including TSSs and enhancers.

We extended these comparisons by conducting a series of analyses side by side of the two segmentations. We analyzed each of the segmentations for overlap with selected other data from the ENCODE project that had specific interpretations in terms of genome annotation, including gene features such as TSSs and transcription termination sites, transcription factor (TF) binding regions from ChIP-seq data, as well as more dispersed signals such as RNA-seq data and methylation levels from reduced representation bisulfite sequencing. Supplementary Figure 19 shows ROC curves for representative TFs. For most TFs the performance of the two segmentations is very similar as measured by the area under the ROC curve with a handful of factors showing marginally superior overall detection of their binding regions by each segmentation (for example, NFE2, RAD21, and ELF1 for ChromHMM; REST, Z3, and SPI1 for Segway). Analysis of the distance to the nearest TSS for each segmentations (Supplementary Figure 14) shows that Segway contains states with better resolution in locating the precise TSS with a median distance from the center of a TSS segment to the nearest TSS of 189 bp for state Tss (GM12878), compared to 652 bp for ChromHMM state Tss. Analysis of the overlap with RNA-seq features (Supplementary Figure 20) and methylation levels (not shown) across the segmentations indicates that segments with similar characteristics exist in each segmentation. These results re-emphasize the contrasting properties of high continuity versus high resolution identified in the initial comparison. Overall, the analysis identifies several contrasting features of the segmentations, but does not favor one over the other. We conclude that useful features can be identified from both approaches and therefore provide both segmentations to users.

The *HBB* locus

We expect that users of the ENCODE data will find that the combined segmentations provide an informative overview of potential functionality in a locus and a guide to more detailed examination and hypothesis development. For an intensively studied locus like the *HBB* complex, which encodes the beta-like subunits of hemoglobin, it is gratifying to find that most of the known regulatory regions are captured by the appropriate states. A combination of Tss, Prom, Enh, Ctcf and DNase states from K562 cells captures 77% (by Segway) and 87% (by ChromHMM) of the known *cis*-regulatory modules (54).

Furthermore, the segmentations provide new insights for regulation within this locus (Supplementary Figure 13). It has long been known that two pairs of large deletions that remove the gene *HBB* (encoding the adult beta-globin) have similar endpoints between *HBB* and *HBG1* or *HBG2* (encoding the fetal gamma-globins), but for each pair, one deletion has a beta-thalassemia phenotype (very low or no beta-globin protein) while the other deletion maintains

expression of the *HBG* genes in adults, hence giving "hereditary persistence of fetal hemoglobin" or HPFH (55,56). The latter display mild or no symptoms, but beta-thalassemia is a serious disease requiring constant therapy. The pairs of deletions are (a) Spanish: (delta beta)0-Thal and HPFH-1; Black and (b) Chinese: Ggamma(Agamma delta beta)0-Thal and HPFH-6. In both cases, the region distal to the globin genes has been shown to have enhancer activity (57,58), indicated as red rectangles in Supplementary Figure 13. The proposed model is that the HPFH deletions bring an enhancer close to the target *HBG* genes and keep one or both expressed, whereas the thalassemia deletions remove the enhancers.

The segmentations (on the top line of Supplementary Figure 13) confirm that the known enhancers, including those at the HPFH breakpoints and the distal locus control region, are in segmentation classes predicted to be enhancers. In addition, they provide the novel insight that the chromosomal segment that is deleted is largely in the quiescent state. Genes in the quiescent state are silent or expressed at a low rate (Figure 4). Hence, the model for gene regulation in these situations can be expanded to include removal of inactive (or quiescent) chromatin as part of the process for activation of HBG genes in HPFH.

Another interesting point is that the HPFH1 enhancer is in the E state (orange), whereas the HPFH6 enhancer is in the WE (weak enhancer, yellow) state. That observation fits with the much stronger signal in HPFH1 enhancer for GATA1, p300 and CEBPB (shown by the TF binding tracks from ENCODE). Finally, the additional enhancer-related signals further away (toward the left in the figure) are intriguing. These could be previously unknown regulatory elements, and this is an important hypothesis to test experimentally.

Supplementary tables

	GM12878	H1 hESC	HeLa-S3
H3K4me1	BroadHistoneGm12878H3k4me1StdAln_2Reps	BroadHistoneH1hescH3k4me1StdAln_2Reps	BroadHistoneHelas3H3k4me1Std_1Reps
H3K4me2	BroadHistoneGm12878H3k4me2StdAln_2Reps	BroadHistoneH1hescH3k4me2StdAln_2Reps	BroadHistoneHelas3H3k4me2StdAln_2Reps
H3K4me3	AllLabsHistoneGm12878H3k4me3StdAln_4Reps	BroadHistoneH1hescH3k4me3StdAln_2Reps	AllLabsHistoneHelas3H3k4me3StdAln_4Reps
H3K9ac	BroadHistoneGm12878H3k9acStdAln_2Reps	BroadHistoneH1hescH3k9acStdAln_2Reps	BroadHistoneHelas3H3k9acStdAln_2Reps
H3K27ac	BroadHistoneGm12878H3k27acStdAln_2Reps	BroadHistoneH1hescH3k27acStdAln_2Reps	BroadHistoneHelas3H3k27acStdAln_2Reps
H3K27me3	AllLabsHistoneGm12878H3k27me3StdAln_4Reps	BroadHistoneH1hescH3k27me3StdAln_2Reps	AllLabsHistoneHelas3H3k27me3StdAln_4Reps
H3K36me3	AllLabsHistoneGm12878H3k36me3StdAln_4Reps	BroadHistoneH1hescH3k36me3StdAln_2Reps	AllLabsHistoneHelas3H3k36me3StdAln_4Reps
H4K20me1	BroadHistoneGm12878H4k20me1StdAln_2Reps	BroadHistoneH1hescH4k20me1StdAln_2Reps	BroadHistoneHelas3H4k20me1StdAln_2Reps
Pol2	AllLabsGm12878Pol2Std_9Reps	HaibTfbsH1hescPol2Pcr1xAln_2Reps	AllLabsHelas3Pol2Std_5Reps
CTCF	AllLabsGm12878CtcfStd_9Reps	AllLabsH1hescCtcfStd_5Reps	AllLabsHelas3CtcfStd_6Reps
Duke DNase	OpenChromDnaseGm12878Aln_5Reps	OpenChromDnaseH1hescAln_2Reps	OpenChromDnaseHelas3Aln_3Reps
UW DNase	UwDnaseGm12878Aln_2Reps	UwDnaseH1hescAln_1Reps	UwDnaseHelas3Aln_2Reps
FAIRE	OpenChromFaireGm12878Aln_3Reps	OpenChromFaireH1hescAln_2Reps	OpenChromFaireHelas3Aln_2Reps
Control	BroadHistoneGm12878ControlStdAln_2Reps	BroadHistoneH1hescControlStdAln_2Reps	BroadHistoneHelas3ControlStdAln_2Reps

	HepG2	HUVEC	K562
H3K4me1	BroadHistoneHepg2H3k4me1Std_1Reps	BroadHistoneHuvecH3k4me1StdAln_3Reps	AllLabsHistoneK562H3k4me1StdAln_4Reps
H3K4me2	BroadHistoneHepg2H3k4me2StdAln_2Reps	BroadHistoneHuvecH3k4me2StdAln_2Reps	BroadHistoneK562H3k4me2StdAln_2Reps
H3K4me3	AllLabsHistoneHepg2H3k4me3StdAln_4Reps	AllLabsHistoneHuvecH3k4me3StdAln_5Reps	AllLabsHistoneK562H3k4me3StdAln_6Reps
H3K9ac	BroadHistoneHepg2H3k9acStdAln_2Reps	BroadHistoneHuvecH3k9acStdAln_3Reps	BroadHistoneK562H3k9acStdAln_2Reps
H3K27ac	BroadHistoneHepg2H3k27acStdAln_2Reps	BroadHistoneHuvecH3k27acStdAln_3Reps	BroadHistoneK562H3k27acStdAln_2Reps
H3K27me3	AllLabsHistoneHepg2H3k27me3StdAln_4Reps	AllLabsHistoneHuvecH3k27me3StdAln_4Reps	AllLabsHistoneK562H3k27me3StdAln_6Reps
H3K36me3	AllLabsHistoneHepg2H3k36me3StdAln_4Reps	AllLabsHistoneHuvecH3k36me3StdAln_5Reps	AllLabsHistoneK562H3k36me3StdAln_4Reps
H4K20me1	BroadHistoneHepg2H4k20me1StdAln_2Reps	BroadHistoneHuvecH4k20me1StdAln_3Reps	BroadHistoneK562H4k20me1StdAln_2Reps
Pol2	AllLabsHepg2Pol2Std_4Reps	AllLabsHuvecPol2Std_5Reps	AllLabsK562Pol2Std_6Reps
CTCF	AllLabsHepg2CtcfStd_8Reps	AllLabsHuvecCtcfStd_7Reps	AllLabsK562CtcfStd_7Reps
Duke DNase	OpenChromDnaseHepg2Aln_3Reps	OpenChromDnaseHuvecAln_2Reps	OpenChromDnaseK562Aln_2Reps
UW DNase	UwDnaseHepg2Aln_2Reps	UwDnaseHuvecAln_2Reps	UwDnaseK562Aln_2Reps
FAIRE	OpenChromFaireHepg2Aln_3Reps	OpenChromFaireHuvecAln_2Reps	OpenChromFaireK562Aln_2Reps
Control	BroadHistoneHepg2ControlStdAln_2Reps	BroadHistoneHuvecControlStdAln_3Reps	BroadHistoneK562ControlStdAln_1Reps

Supplementary Table 1: Names of signal tracks used as input to the segmentations. Each column contains the tracks from two cell types. Each row contains a single assay target, coordinated amongst the cell types. All tracks are available at https://sites.google.com/site/anshulkundaje/projects/wiggler. Filenames are wgEncode*.norm5.rawsignal.bedgraph.gz, where * is the name used in this table.

Assay	Target	Production groups
ChIP-seq	histone modifications and variants	Broad Institute of MIT and Harvard Stanford/Yale/Davis/Harvard
	corresponding control datasets	University of Washington
ChIP-seq	transcription factor binding sites	Broad Institute of MIT and Harvard HudsonAlpha Institute for Biotechnology
	corresponding control datasets	Stanford/Yale/Davis/Harvard University of Texas at Austin University of Washington
DNase-seq	regions of open chromatin	Duke University University of Washington
FAIRE-seq	regions of open chromatin	University of North Carolina-Chapel Hill
MNase-seq	nucleosome positions	Stanford

Supplementary Table 2: Classes of ENCODE datasets we generated signal tracks for. We used a subset of these tracks, as described in Methods and specified in Supplementary Table 1, as input to the segmentations.

(A)

cell type	mnemonic	rationale
GM12878	Tss	highest TSS; high H3K4me3, H3K27ac, Pol2, DNase, FAIRE
	TssF	second-highest TSS; high H3K4me3, H3K27ac, DNase, FAIRE
	PromF	high H3K4me3, H3K27ac, DNase, FAIRE
	Enh	highest p300; high H3K4me1
	EnhF	med H3K4me1
	EnhWf	med-high K4me1, but low K27ac
	Ctcf	high CTCF
	CtcfO	highest CTCF, DNase, FAIRE
	Gen5'	high K36me3, exon
	Elon1	high K36me3, exon
	Elon2	high K36me3, exon
	ElonW	high K36me3, exon
	Gen3'	highest K36me3, med Pol2, TES after Tss/TssF
	Repr1	K27me3
	Repr2	K27me3
	Repr3	K27me3
	Repr4	K27me3
	Repr5	K27me3
	Low1	low values everywhere but control
	Low2	low values everywhere but control
	Low3	low values everywhere but control
	Low4	low values everywhere but control
	Low5	low values everywhere but control
	Low6	low values everywhere but control
	Quies	zeroes
H1 hESC	Tss	highest TSS
	TssF	near-second highest TSS
	PromF	K4me3+K27ac
	PromP	second highest TSS+K27me+K4me3+no K27ac
	EnhPr	second-highest p300+K4me1+K27ac
	EnhWf1	K4me1
	EnhWf2	K4me1
	EnhWf3	K4me1
	DnaseD	Duke DNase
	Faire	FAIRE
	CtcfO	CTCF, DNase, FAIRE
	Elon	K36me3+exon
	ElonW1	K36me3
	ElonW2	K36me3
	ElonW3	Pol2
	Gen3'	K36Me3+Pol2+exon
	Repr	K27me3
	Low1	low everything except Control, FAIRE
	Low2	low everything except Control, FAIRE
	Low3	low everything except Control, FAIRE
	Low4	low everything except Control, FAIRE
	Low5	low everything except Control, FAIRE
	Low6	low everything except Control, FAIRE
	Low7	low everything except Control, FAIRE
	Quies	zeroes
HeLa-S3	Tss	high K4me3+K27ac+TSS
	PromF	medium TSS, high K27ac, K4me3
	Enh1	highest p300+high K4me1+K27ac
	Enh2	high p300
	EnhF	some p300, K4me1
	EnhPr	some p300, K4me1, K27ac
	Faire	FAIRE but nothing else
	CtcfO	CTCF, DNase, FAIRE
	Gen5'	K36me3, high exon, medium Pol2, low TSS
	Elon	K36Me3+Pol2+exon
	ElonW1	K36me3+low pol2+exon
	ElonW2	K36Me3+Pol2+exon

	Gen3'	K36me3+TES
	Repr1	K27me3
	Repr2	K27me3
	Repr3	K27me3
	Repr4	K27me3
	Low1	low except Control, FAIRE
	Low2	low except Control, FAIRE
	Low3	low except control, H4K20
	Low4	low except Control
	Low5	low except Control
	Low6	low except Control
	Low7	low everywhere
	Quies	zeroes
HepG2	Tss	high K4me3, K27ac
	PromF	high K4me3, K27ac
	PromP1	high TSS+K27me3
	PromP2	high TSS+K27me3
	Enh	high p300
	EnhF1	med p300, K4me1, K27ac
	EnhF2	med p300, K4me1
	EnhPr	med-low p300
	EnhW	K4me1+FAIRE+p300+low K27ac
	CtcfO	CTCF, DNase, FAIRE
	Gen5'1	K36me3+med exon
	Gen5'2	K36me3+med exon
	Elon	K36me3
	ElonW	K36me3
	Repr1	high K27me3
	Repr2	high K27me3
	Repr3	high K27me3
	Repr4	K27me3
	Repr5	K27me3
	Repr6	K27me3
	Low1	low except Control
	Low2	low except Control
	Low3	low except Control
	Low4	low except Control
	Quies	zeroes
HUVEC	Tss	K4me3+exon
	PromF	K4me3+exon
	Enh	K4me1
	EnhF1	K4me1
	EnhF2	K4me1
	EnhW1	K4me1
	EnhW2	K4me1+FAIRE+low K27ac
	Faire	FAIRE
	CtcfO	CTCF, DNase, FAIRE
	Gen5'	K36me3+exon
	Elon1	K36me3+exon
	Elon2	K36me3+exon
	ElonW	K36me3+exon (weaker)
	Gen3'	K36me3+exon
	Repr1	K27me3
	Repr2	K27me3
	Repr3	K27me3
	Repr4	K27me3
	Repr5	K27me3
	Repr6	K27me3
	Repr7	K27me3
	Low1	low except Control
	Low2	low except H4K20
	Low3	low except control, H4K20
	Quies	zeroes
K562	Tss	highest TSS
	PromF	second highest TSS+K4me3
	Enh1	highest p300

Enh2	second highest p300
EnhF1	medium K4me1
EnhF2	high K4me1
EnhF3	high K4me1
EnhP	K27me3+K4me1
EnhWf	K4me1
CtcfO	CTCF
Elon	high K36me3
ElonW	high K36me3
Gen3'1	highest K36me3, TES
Gen3'2	highest K36me3, TES
Repr1	K27me3
Repr2	K27me3
Repr3	K27me3
Repr4	K27me3
Low1	low except control, FAIRE
Low2	low except control, FAIRE
Low3	low except control, FAIRE
Low4	low except control, FAIRE
Low5	low except control, FAIRE
Low6	low except control, FAIRE
Quies	zeroes

(B)

cell type	mnemonic	rationale
all	Tss	Active promoter, TSS/CpG island region
	TssF	Active promoter, flanking TSS/CpG islands
	PromF	Promoter flanking
	PromP	Inactive/Poised promoter, highly conserved
	Enh	Candidate Strong Enhancer, open chromatin
	EnhF	Candidate Strong Enhancer, flanking open chromatin
	EnhWF	Candidate poised/weak enhancer; flanking open chromatin of candidate enhancers
	EnhW	Candidate weak enhancer and open chromatin
	DNaseU	Primarily UW DNase, weaker open chromatin sites
	DNaseD	Primarily Duke DNase, candidate regulatory elements in more likely repressive locations
	FaireW	Modest Faire/Control enrichments, potential CNV
	CtcfO	Distal CTCF/Candidate Insulator with open chromatin
	Ctcf	Distal CTCF/Candidate Insulator without open chromatin
	Gen5'	Transcription transition, highly expressed genes towards 5' end
	Elon	Transcriptional Elongation, stronger H3K36me3, more exonic
	ElonW	Transcriptional Elongation, weaker H3K36me3
	Gen3'	Transcription 3' end of genes, highly expressed; more exonic
	Pol2	Pol2 specific locations, majority in genes but substantial portion in intergenic locations
	H4K20	Transcription, primarily H4K20me1, more intronic
	ReprD	Polycomb Repression with Duke DNase sites/promoter and conservation enrichment (except HELA)
	Repr	Strong Polycomb Repression
	ReprW	Weaker Polycomb Repression
	Low	Low signal proximal to active elements
	Quies	Heterochromatin/Dead Zone
	Art	Potential CNV or repetitive artifacts

Supplementary Table 3: Rationale for mnemonics. Each entry briefly describes the primary reason or reasons that the given state was assigned its mnemonic label. (A) Segway, (B) ChromHMM.

Superstate	Superstate description	Gm12878	H1 hESC	HeLa-S3	HepG2	HUVEC	K562
TSS	Predicted promoter region including TSS	Tss TssF Enh	Tss TssF	Tss Enh1	Tss PromP2 Enh	Tss Enh EnhF2	Tss PromF Enh1
E	Predicted enhancer	EnhF PromF	EnhPr PromP PromF	Enh2 EnhPr PromF	EnhPr PromF PromP1	EnhF1 PromF Repr2	EnhF3 Enh2
I	Distal CTCF	CtcfO	CtcfO	CtcfO	CtcfO	CtcfO	CtcfO
G	Transcribed Gene	Gen5' Elon1 Elon2 ElonW Gen3' Low4	Elon Gen3'	Gen5' Elon ElonW1 ElonW2 Gen3' Low4	Gen5'1 Gen5'2 Elon	Gen5' Elon1 ElonW Gen3' Low2	Gen3'1 Gen3'2 Elon ElonW

Supplementary Table 4: Clusters of states for a subset of Segway states for each cell type, used in generating the combined segmentation. All other states were clustered into the R (Predicted Repressed or Low Activity region) cluster.

State clusters	TSS	E	I	G	R
States 1-2	TSS				
State 3		PF			
State 4	TSS	E	CTCF		
States 5-6	E	E			
States 7-11		WE	CTCF		
States 12-13			CTCF	CTCF	CTCF
States 14-19			CTCF	T	T
States 20-25			CTCF	T	R

Supplementary Table 5: Combined segmentation states for combinations of ChromHMM states (y-axis) and Segway clusters of states (x-axis). If the cell is empty, the two segmentations are classed as contradictory and no combined state is assigned.

Cell line	Coverage (bp)	Coverage (%)
GM12878	2,729,345,527	96.5
H1 hESC	2,671,442,653	94.4
HeLa-S3	2,683,004,504	94.8
HepG2	2,694,700,569	95.2
HUVEC	2,693,810,504	95.2
K562	2,685,096,985	94.9

Supplementary Table 6: Coverage for the combined segmentation in each of the tier 1 and 2 cell lines compared to Segway and ChromHMM at 100%.

ChromHMM

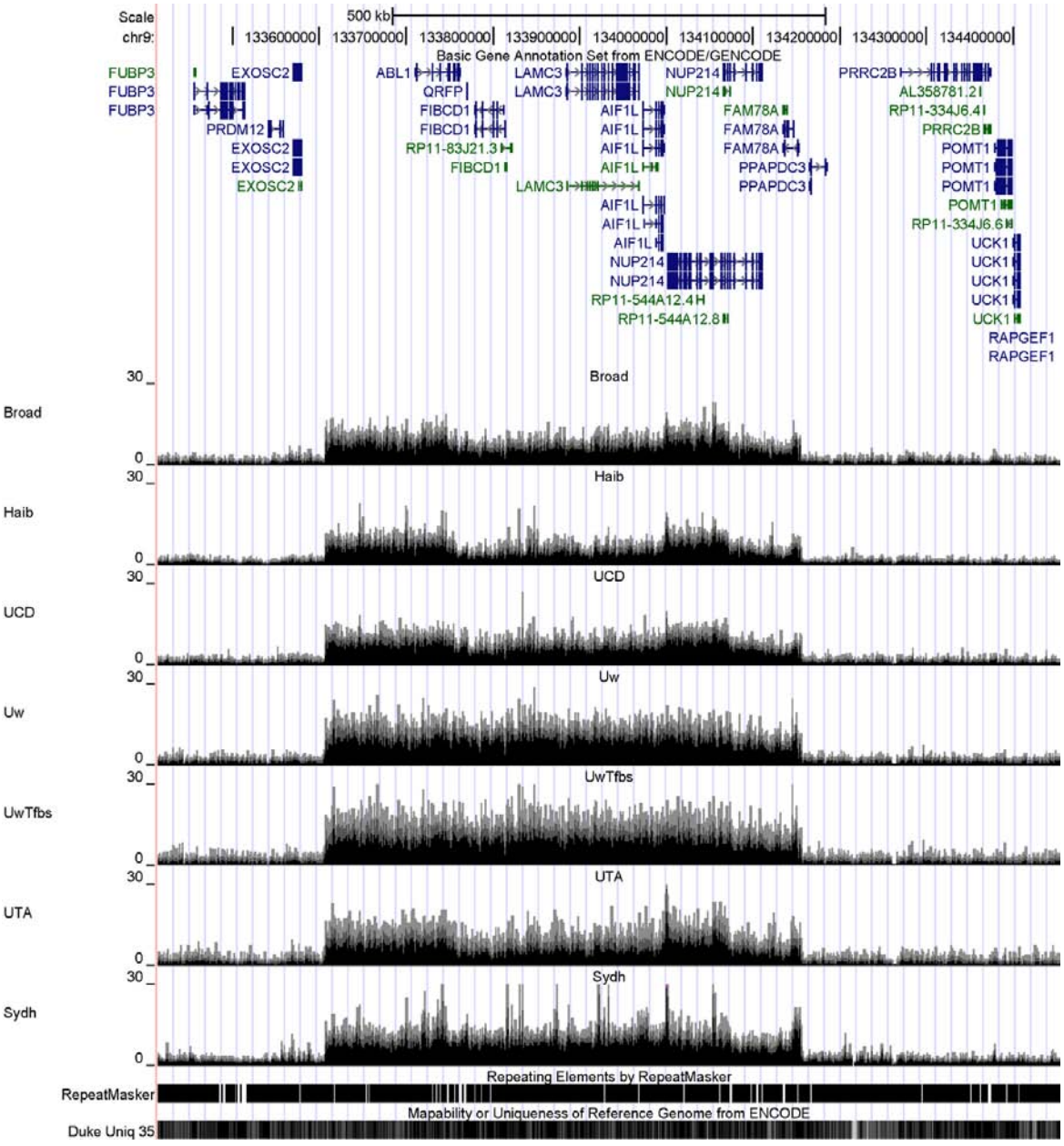
State	GM12878	H1-hESC	HeLa-s3	HepG2	K562
1_Tss	5.6	6.0	5.7	5.3	6.2
2_TssF	1.2	2.0	1.0	1.5	1.7
3_PromF	-0.5	0.5	-0.2	1.0	0.7
4_PromP	4.1	4.1	4.2	3.3	4.3
5_Enh	5.4	5.2	5.7	5.7	5.2
6_EnhF	1.5	1.4	1.6	2.1	1.2
7_EnhWF	0.1	-0.5	0.5	1.0	-0.3
8_EnhW	4.5	3.7	4.6	4.4	3.8
9_DNaseU	3.5	1.5	3.1	3.4	2.8
10_DNaseD	1.3	-0.3	-0.2	0.5	0.1
11_FaireW	-2.3	-1.4	-3.2	-0.7	-1.0
12_CtcfO	5.6	5.3	4.8	4.2	4.4
13_Ctcf	2.5	3.9	2.0	2.2	0.8
14_Gen5'	1.0	0.5	0.6	1.3	0.4
15_Elon	-3.4	-3.2	-3.1	-1.7	-2.3
16_ElonW	-4.5	-4.8	-4.1	-2.0	-3.2
17_Gen3'	-1.4	-1.4	-1.8	-1.0	-1.7
18_Pol2	-1.2	-1.3	-1.9	-1.0	-0.9
19_H4K20	-3.1	-2.1	-3.0	-1.0	-1.9
20_ReprD	1.2	1.5	-0.6	0.3	0.5
21_Repr	-3.2	-2.7	-3.9	-1.8	-2.5
22_ReprW	-3.8	-2.9	-4.1	-1.9	-2.8
23_Low	-3.5	-3.5	-4.0	-1.9	-2.5
24_Quies	-4.9	-4.3	-5.4	-2.5	-4.5
25_Art	3.7	3.1	3.1	3.9	3.2

Segway

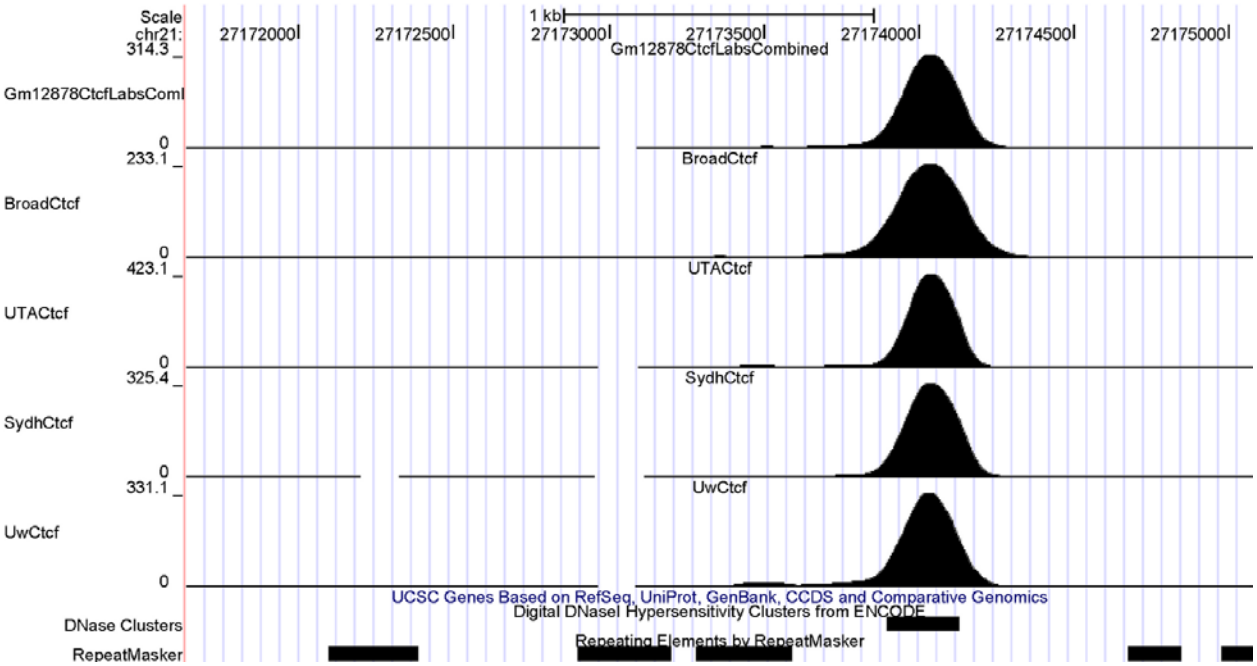
State	GM12878	H1-hESC	HeLa-s3	HepG2	K562
Tss	6.5	6.2	5.6	6.4	6.4
TssF	2.7	4.1			
PromF	0.0	1.0	2.4	1.9	3.0
PromP		3.0			
PromP1				2.9	
PromP2				4.8	
Enh	5.4			5.5	
Enh1			6.6		6.8
Enh2			5.0		4.3
EnhF	3.1		0.3		
EnhF1				2.5	-0.2
EnhF2				0.1	0.3
EnhF3					1.9
EnhPr		3.2	1.1	1.2	
EnhP					0.7
EnhW				-0.9	
EnhW1					
EnhW2					
EnhWf	0.0				-1.6
EnhWf1		0.8			
EnhWf2		-1.0			
EnhWf3		-3.6			
DnaseD		0.2			
Faire		-2.3	-5.5		
CtcfO	6.3	6.2	4.9	4.5	4.6
Ctcf	2.8				
Gen5'	0.6		-0.9		
Gen5'1				0.4	
Gen5'2				-2.0	
Elon		-2.9	-3.2	-1.6	-2.0
Elon1	-4.1				
Elon2	-3.0				
ElonW	-5.9			-2.4	-2.9
ElonW1		-4.6	-4.5		
ElonW2		-6.0	-4.9		
ElonW3		-3.4			
Gen3'	-1.4	-0.5	-1.7		
Gen3'1					-1.8
Gen3'2					0.0
Repr		0.6			
Repr1	-2.9		-4.0	-0.2	-1.9
Repr2	0.7		-3.7	-2.7	-1.8
Repr3	-4.2		-1.2	-1.3	-3.8
Repr4	-5.0		-5.8	1.8	-2.2
Repr5	-6.3			-2.6	
Repr6				-2.2	
Repr7					
Low1	-6.7	-6.2	-3.8	-1.8	-2.8
Low2	-1.9	-3.2	-2.9	-1.5	-4.8
Low3	-2.5	0.3	-5.8	-2.5	-2.6
Low4	-2.1	-3.5	-4.5	-2.4	-3.5
Low5	-5.8	-2.9	-3.7		-4.4
Low6	-4.1	-1.8	-1.5		-4.0
Low7		-5.6	-4.3		
Quies	-7.8	-7.1	-7.2	-3.1	-6.1

Supplementary Table 7: ChromHMM and Segway $\log_2$ fold enrichment for ENCODE transcription factor and general factor binding peaks calls in matched cell types to the segmentation states for the SPP set (59). Enrichments are shown for the GM12878, H1 hESC, HeLa-S3, HepG2, HUVEC, and K562 cell types which were based on combining the peak calls for 57, 31, 42, 52, 4, and 95 experiments respectively. We excluded CTCF and Pol2 because they were inputs to the segmentation.

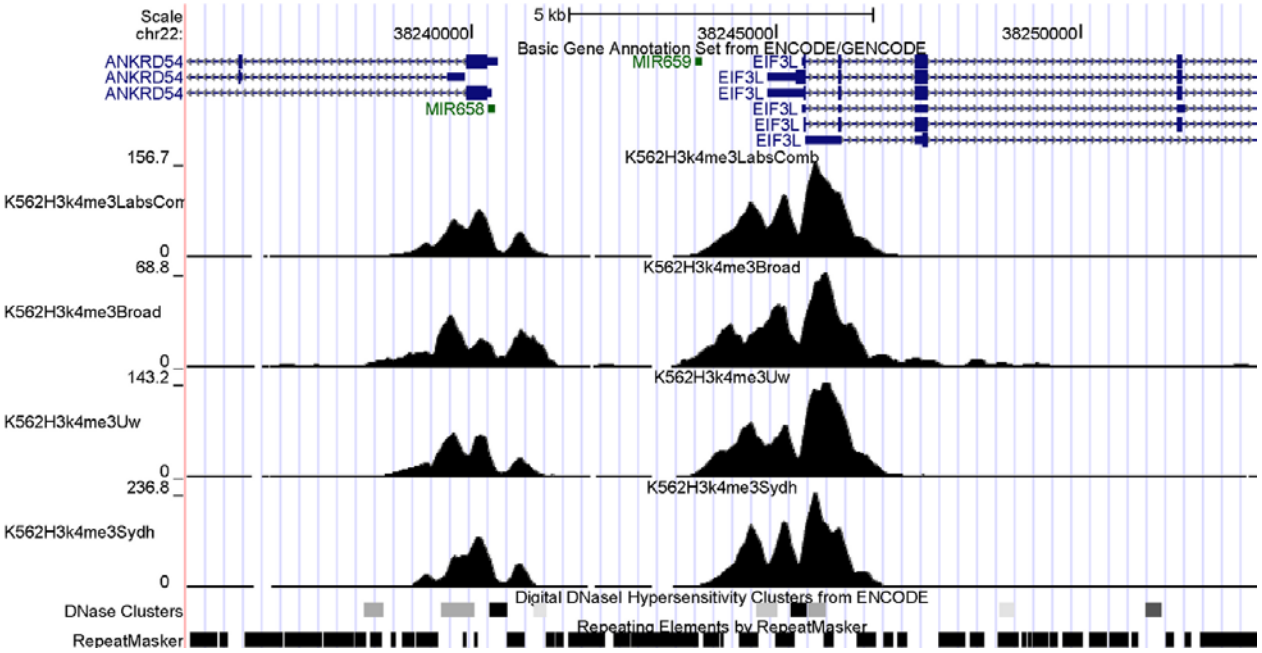
Supplementary figures



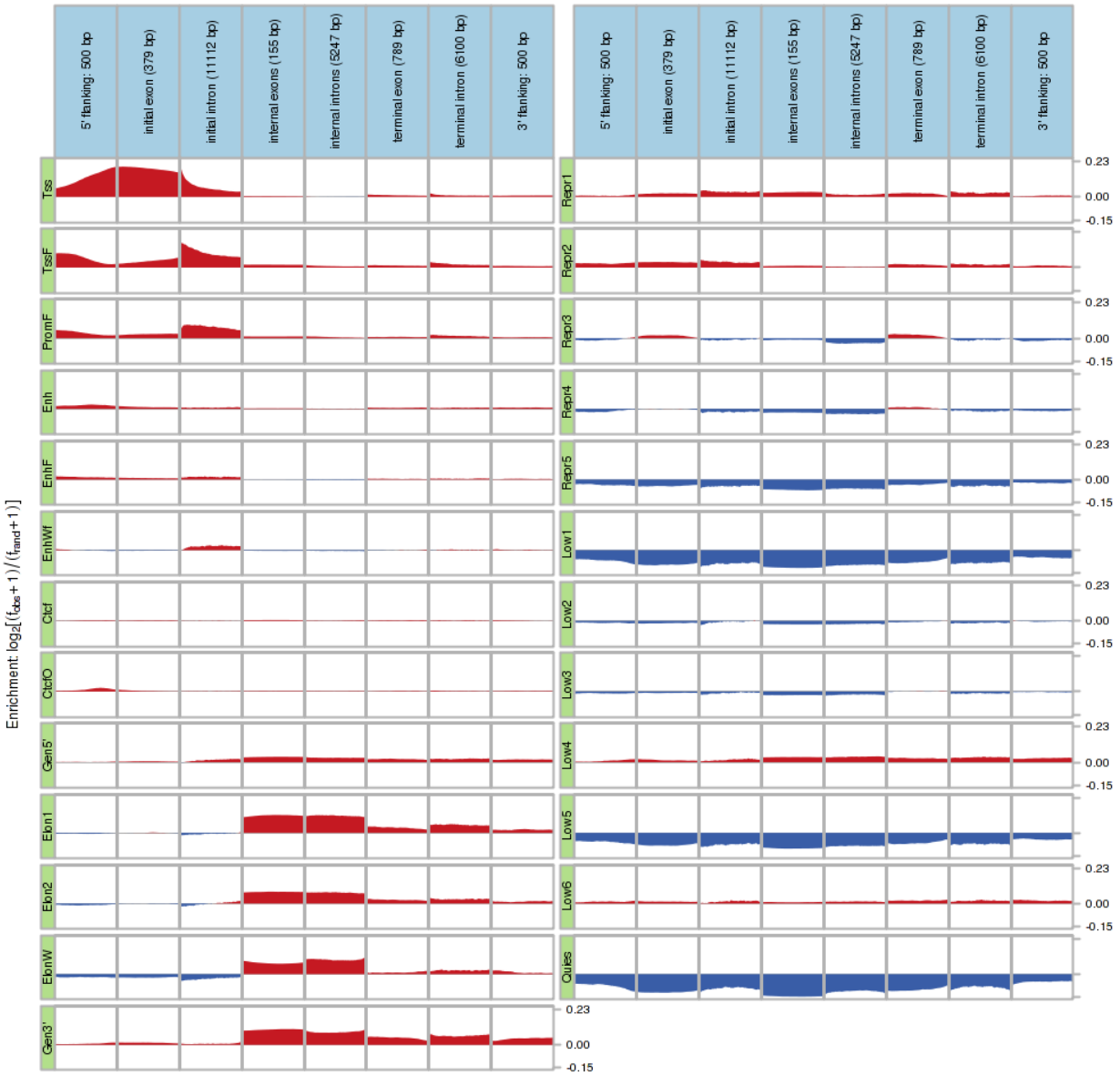
Supplementary Figure 1: Normalized signal from several input DNA tracks produced by different laboratories and production groups in the K562 chronic myeloid leukemia cell line around the *BCR-ABL* fusion locus, as shown in the UCSC Genome Browser. The tracks shown include gene models; input DNA from Broad Institute; Hudson-Alpha Institute of Biology; University of California, Davis; University of Washington DNase; University of Washington DNase ChIP; University of Texas at Austin; Stanford/Yale/Davis/Harvard group, RepeatMasker regions; and 35-bp mappability. The *BCR-ABL* locus is a known region of amplified DNA. Input DNA datasets typically show signal in the range of 0 to 3 fold, but in amplified regions they show signal > 20 fold.



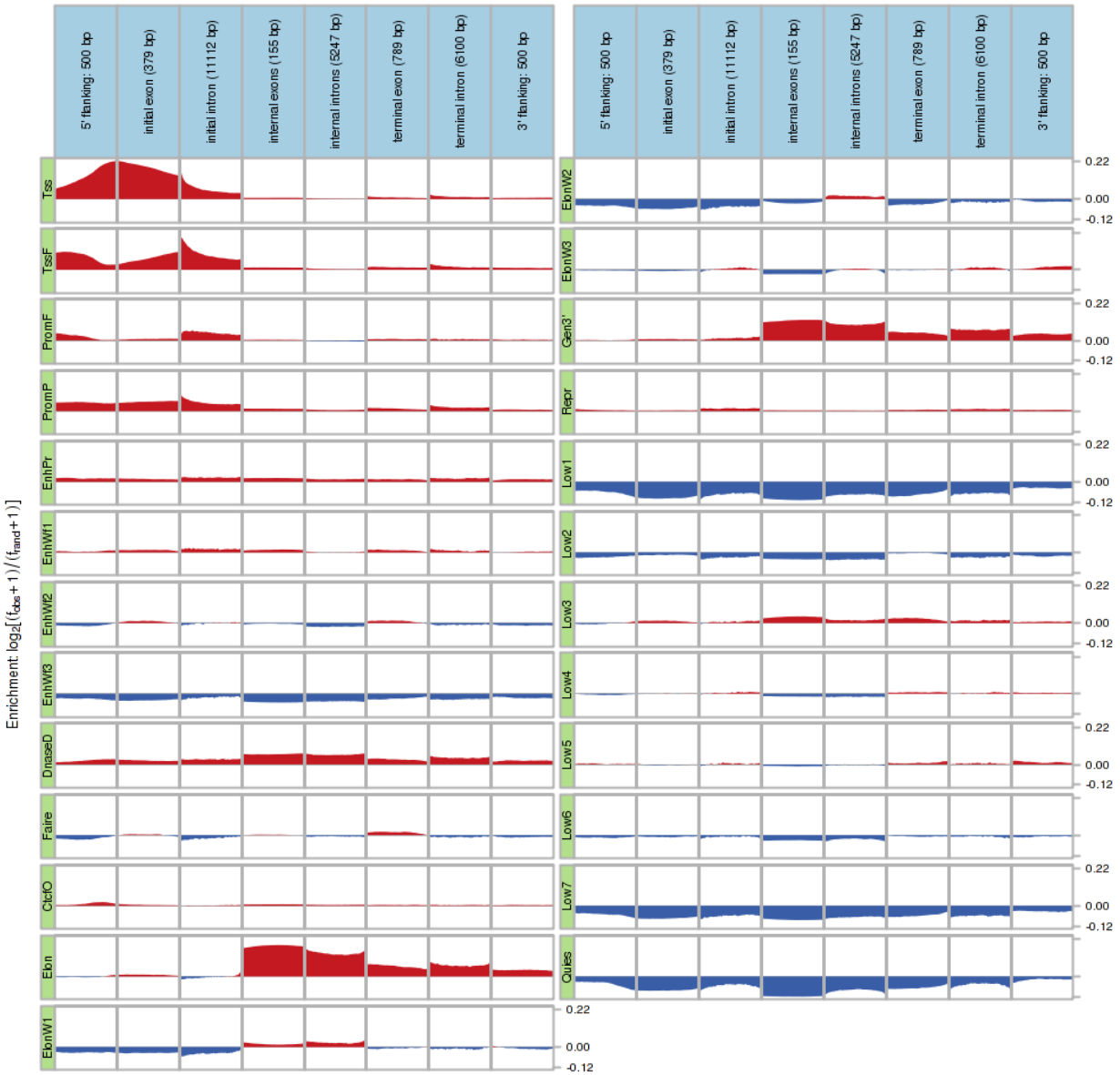
Supplementary Figure 2: Normalized signal of multiple combined-replicate CTCF datasets in the GM12878 cell line from different laboratories (Broad Institute, University of Texas at Austin, Stanford/Yale/Davis/Harvard, University of Washington), as shown in the UCSC Genome Browser. The top track shows combined signal across all of these laboratories. All tracks show similar range of signal and very similar shape of the signal profile with coincidence of the peak summits. Peak signal at typical high-confidence TF peak locations is typically > 20.



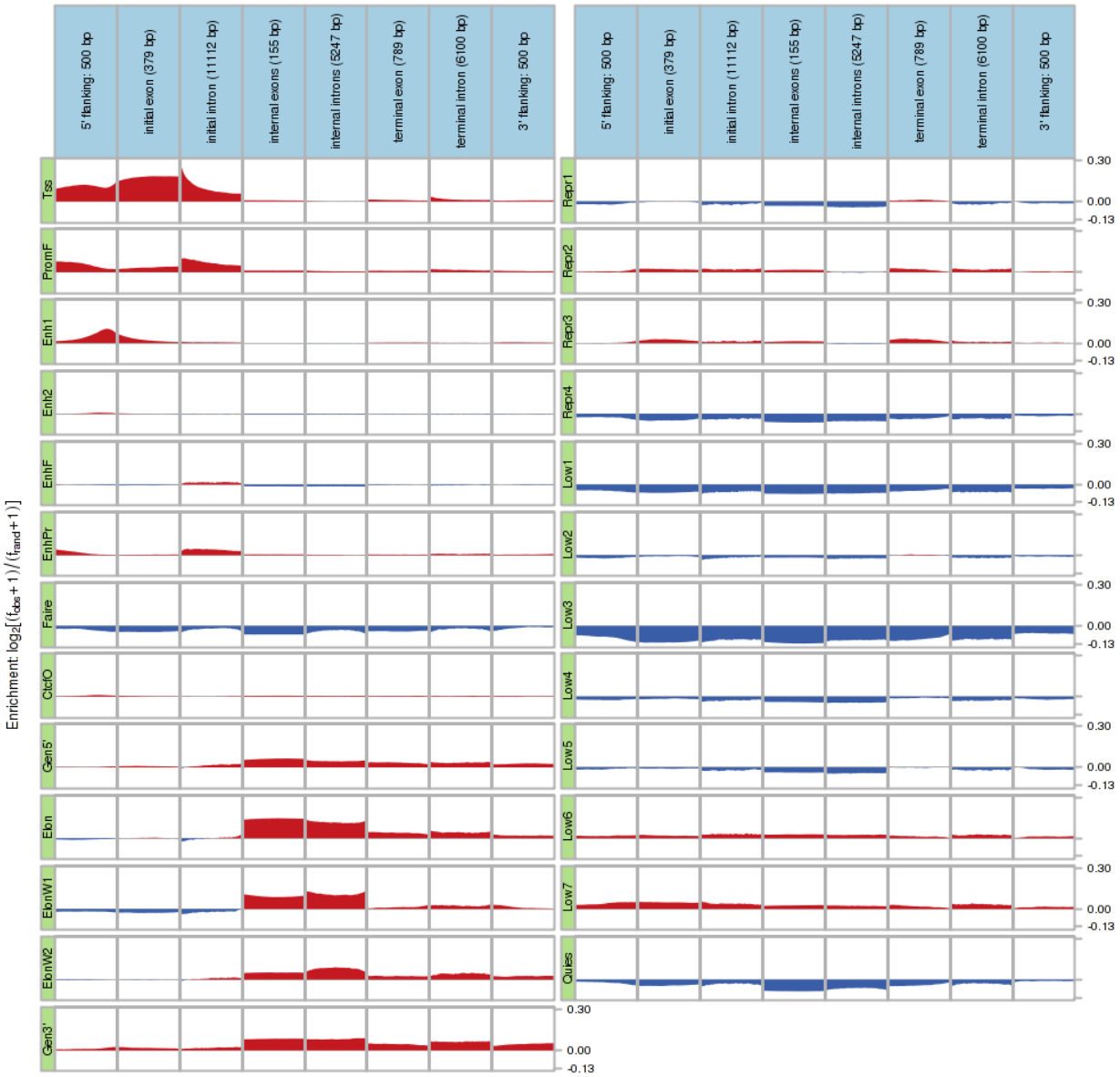
Supplementary Figure 3: Normalized signal for multiple combined-replicate H3k4me3 datasets from different laboratories and the combined signal across all labs, as shown in the UCSC Genome Browser. All tracks show similar range of signal and very similar shape of the signal profiles. High confidence regions of enrichment typically show signal > 15 fold. Note the precise dip in the H3k4me3 signal at the TSS of the genes that corresponds to the nucleosome free region.


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (3247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
Enh
EnhF
EnhWf
Ctcf
CtcfO
Gen5'
Ebn1
Ebn2
EbnW
Gen3'
Repr1
Repr2
Repr3
Repr4
Repr5
Low1
Low2
Low3
Low4
Low5
Low6
Ques
0.23
0.00
-0.15
Enrichment: log2[(fobs+1)/(frand+1)]

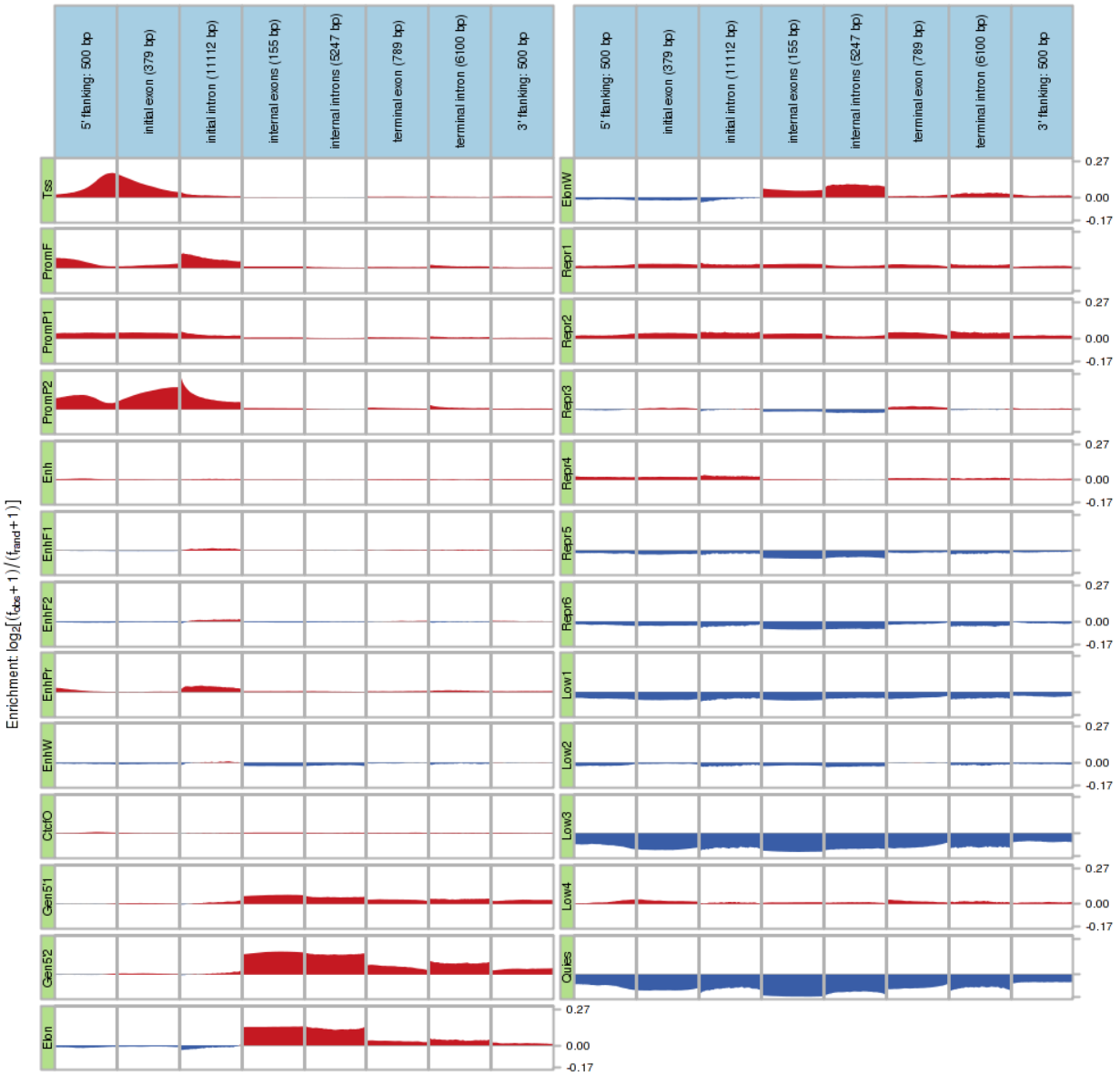
(A)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (8247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
EnhPr
EnhWf1
EnhWf2
EnhWf3
DnaseD
Faire
CtcfO
Ebn
EbnW1
EbnW2
EbnW3
Gen3'
Repr
Low1
Low2
Low3
Low4
Low5
Low6
Low7
Ques
0.22
0.00
-0.12
Enrichment: log2[(fobs + 1)/(frand + 1)]

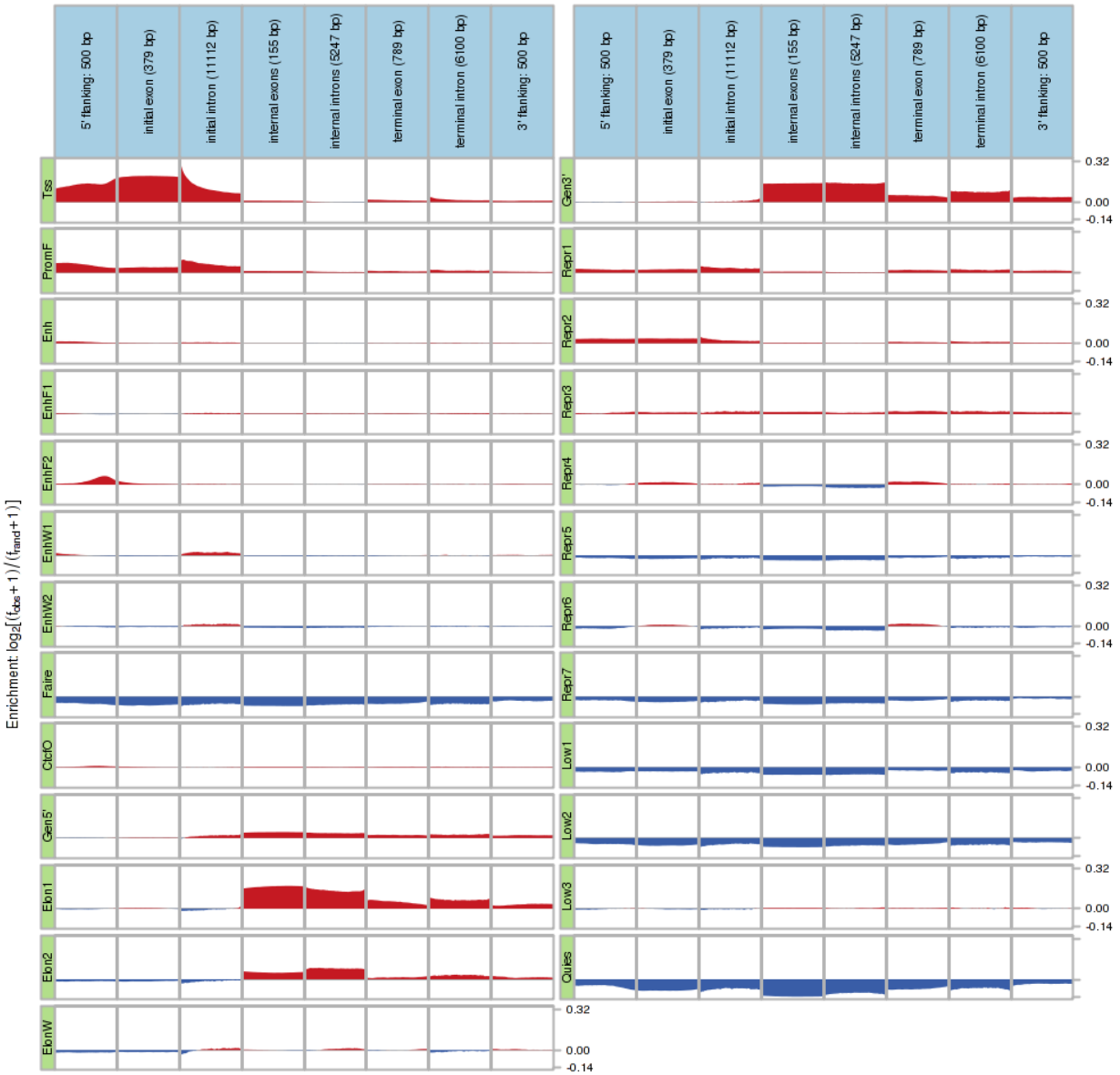
(B)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (5247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
PromF
Enh1
Enh2
EnhF
EnhPr
Faire
CtcfO
Gen5'
Ebn
EbnW1
EbnW2
Gen3'
Repr1
Repr2
Repr3
Repr4
Low1
Low2
Low3
Low4
Low5
Low6
Low7
Quies
0.30
0.00
-0.13
Enrichment: log2[(fobs + 1)/(frand + 1)]

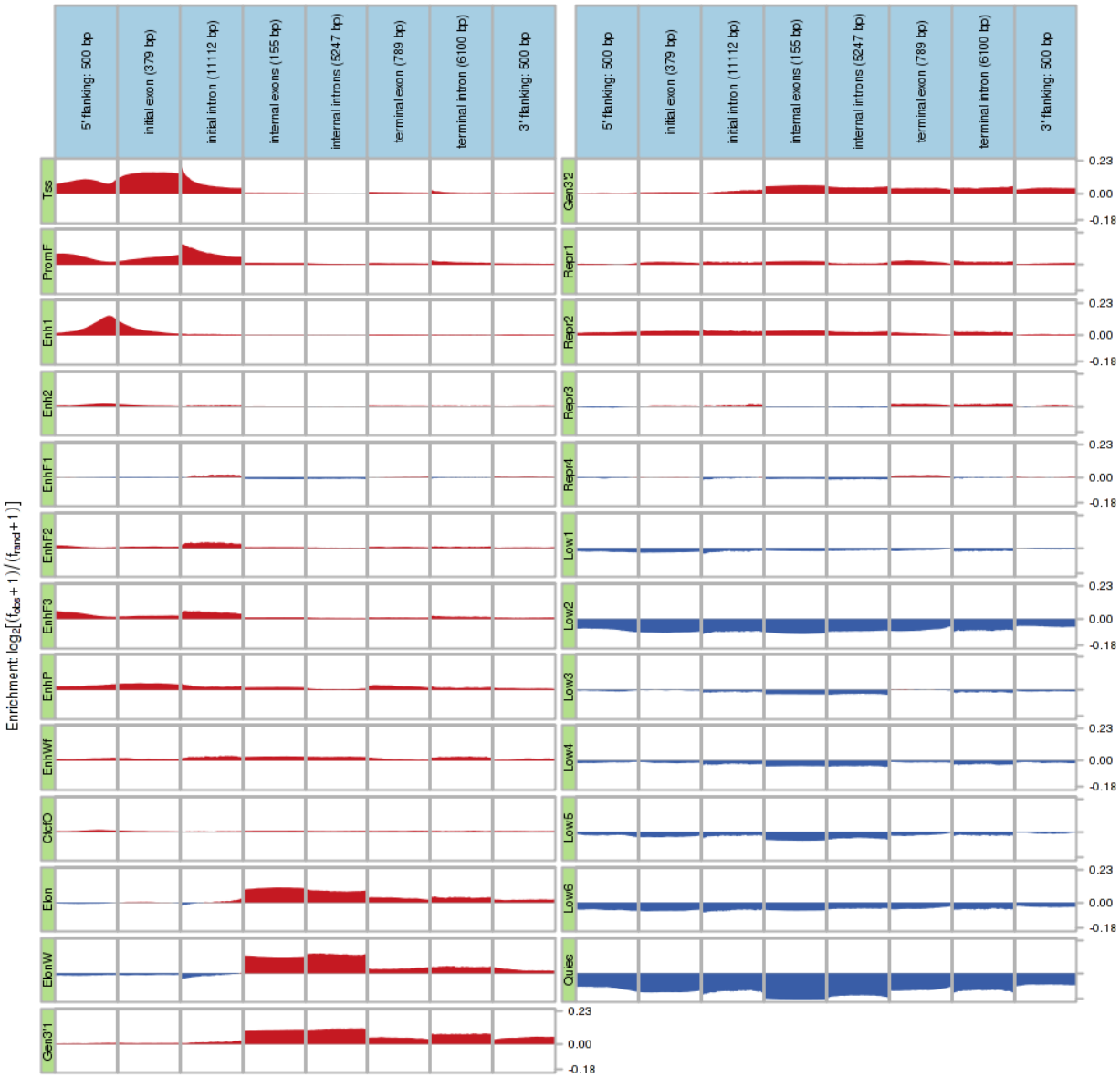
(C)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (5247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
PromF
PromP1
PromP2
Enh
EnhF1
EnhF2
EnhPr
EnhW
CtcfO
Gen5'1
Gen5'2
Ebn
EbnW
Repr1
Repr2
Repr3
Repr4
Repr5
Repr6
Low1
Low2
Low3
Low4
Quies
0.27
0.00
-0.17
Enrichment: log2[(fobs + 1)/(frand + 1)]

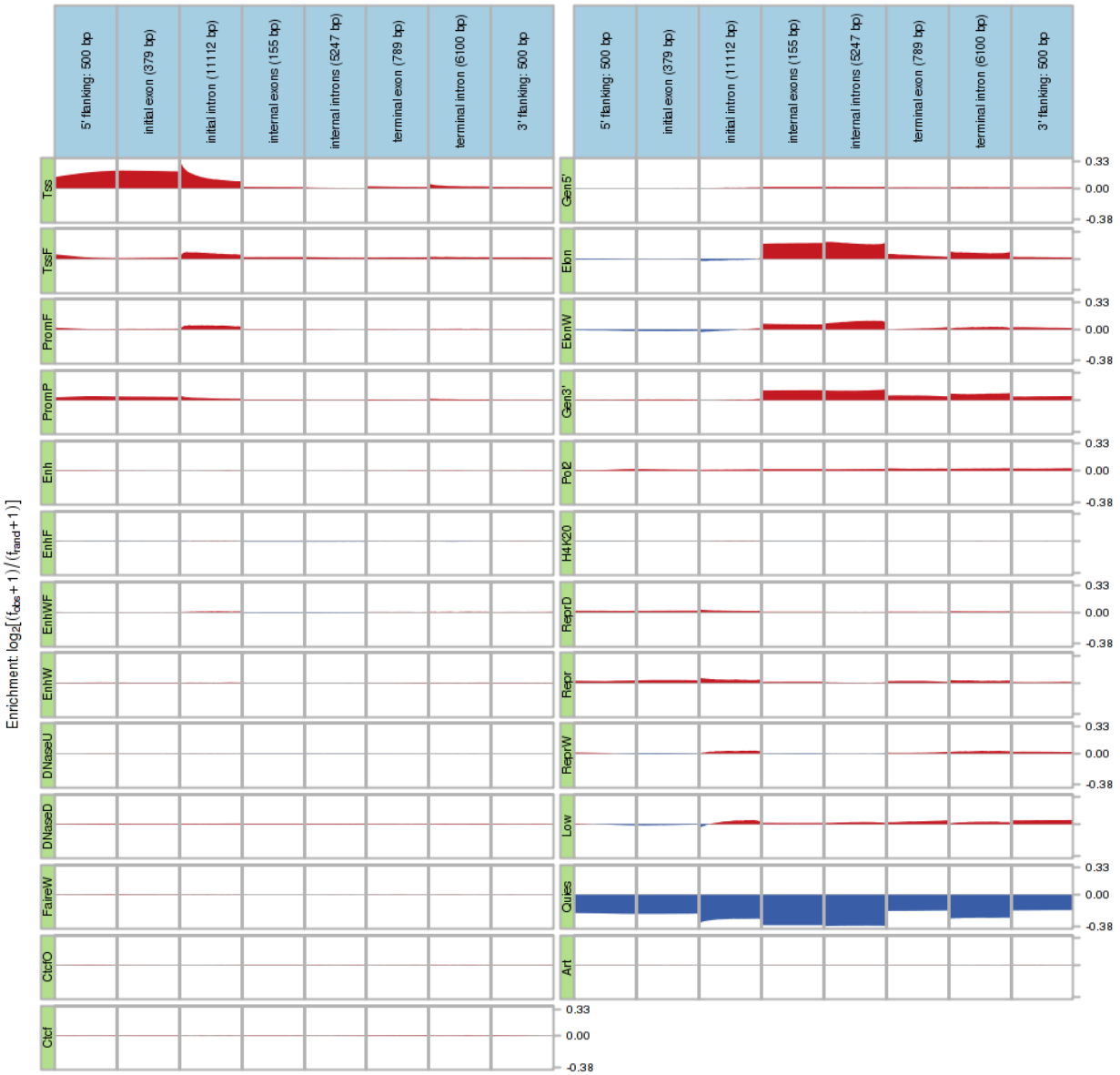
(D)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (8247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
PromF
Enh
EnhF1
EnhF2
EnhW1
EnhW2
Faire
CtcfO
Gen5'
Elon1
Elon2
ElonW
Gen3'
Repr1
Repr2
Repr3
Repr4
Repr5
Repr6
Repr7
Low1
Low2
Low3
Ques
0.32
0.00
-0.14
Enrichment: log2[(fobs + 1)/(frand + 1)]

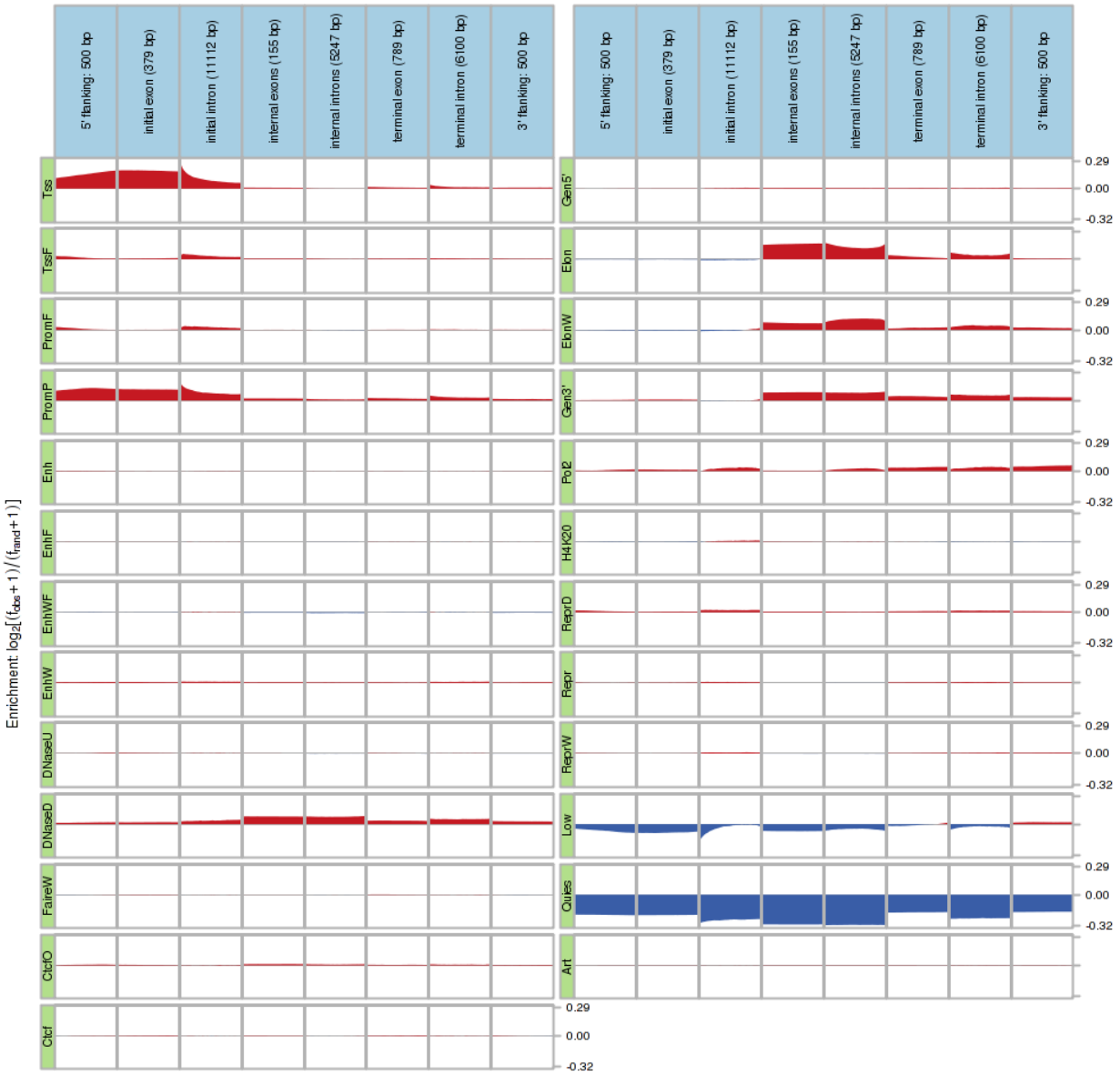
(E)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (8247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
PromF
Enh1
Enh2
EnhF1
EnhF2
EnhF3
EnhP
EnhWf
CtcfO
Ebn
EbnW
Gen3'1
Gen3'2
Repr1
Repr2
Repr3
Repr4
Low1
Low2
Low3
Low4
Low5
Low6
Quies
0.23
0.00
-0.18
Enrichment: log2[(fobs + 1)/(frand + 1)]

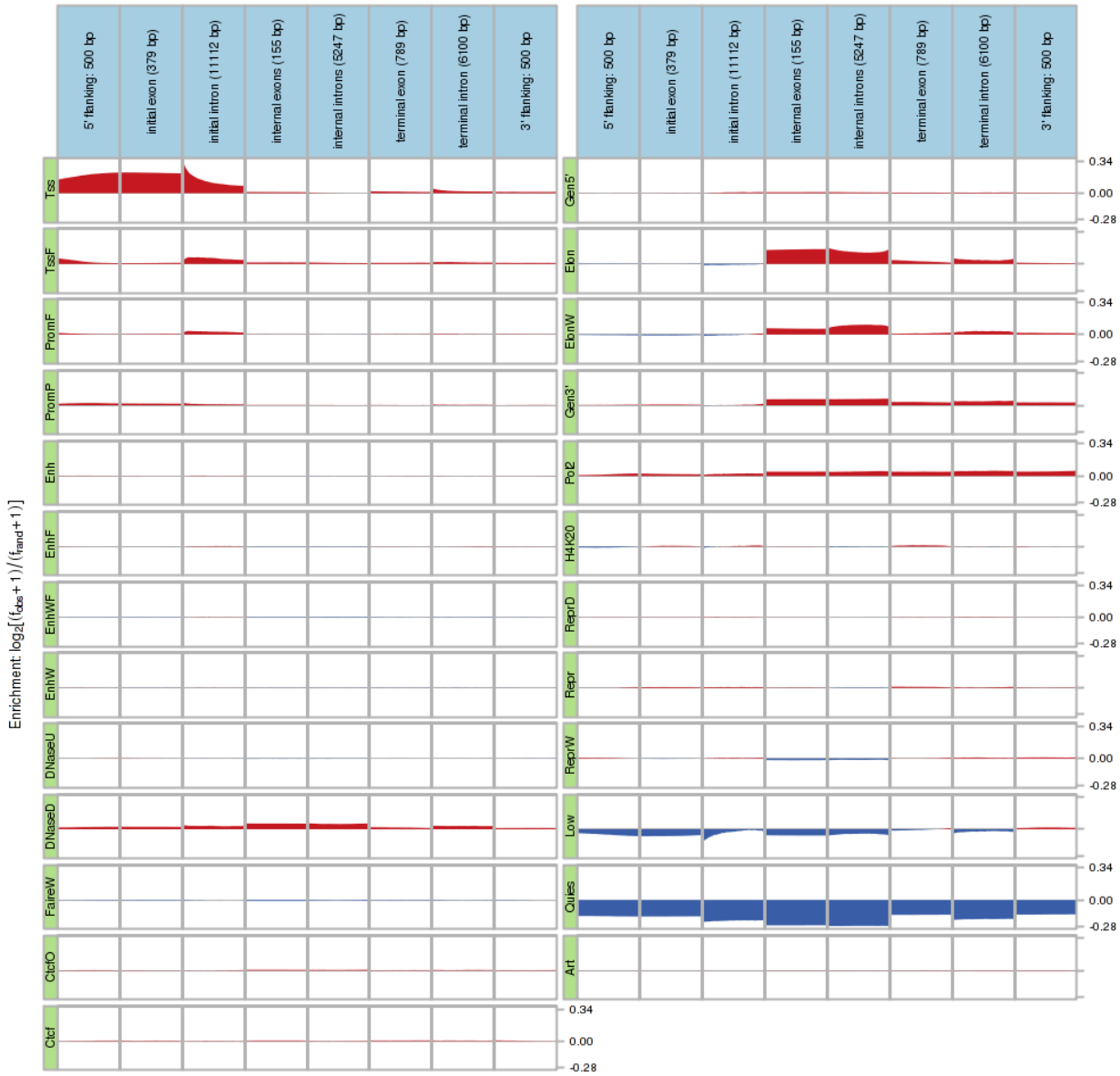
(F)


Enrichment: log2[(fobs + 1)/(frand + 1)]
5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (5247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
Enh
EnhF
EnhWF
EnhW
DNaseU
DNaseD
FaireW
CtcfO
Ctcf
Gen5'
Ebn
EbnW
Gen3'
Pol2
H4K20
ReprD
Repr
ReprW
Low
Cues
Art
0.33
0.00
-0.38

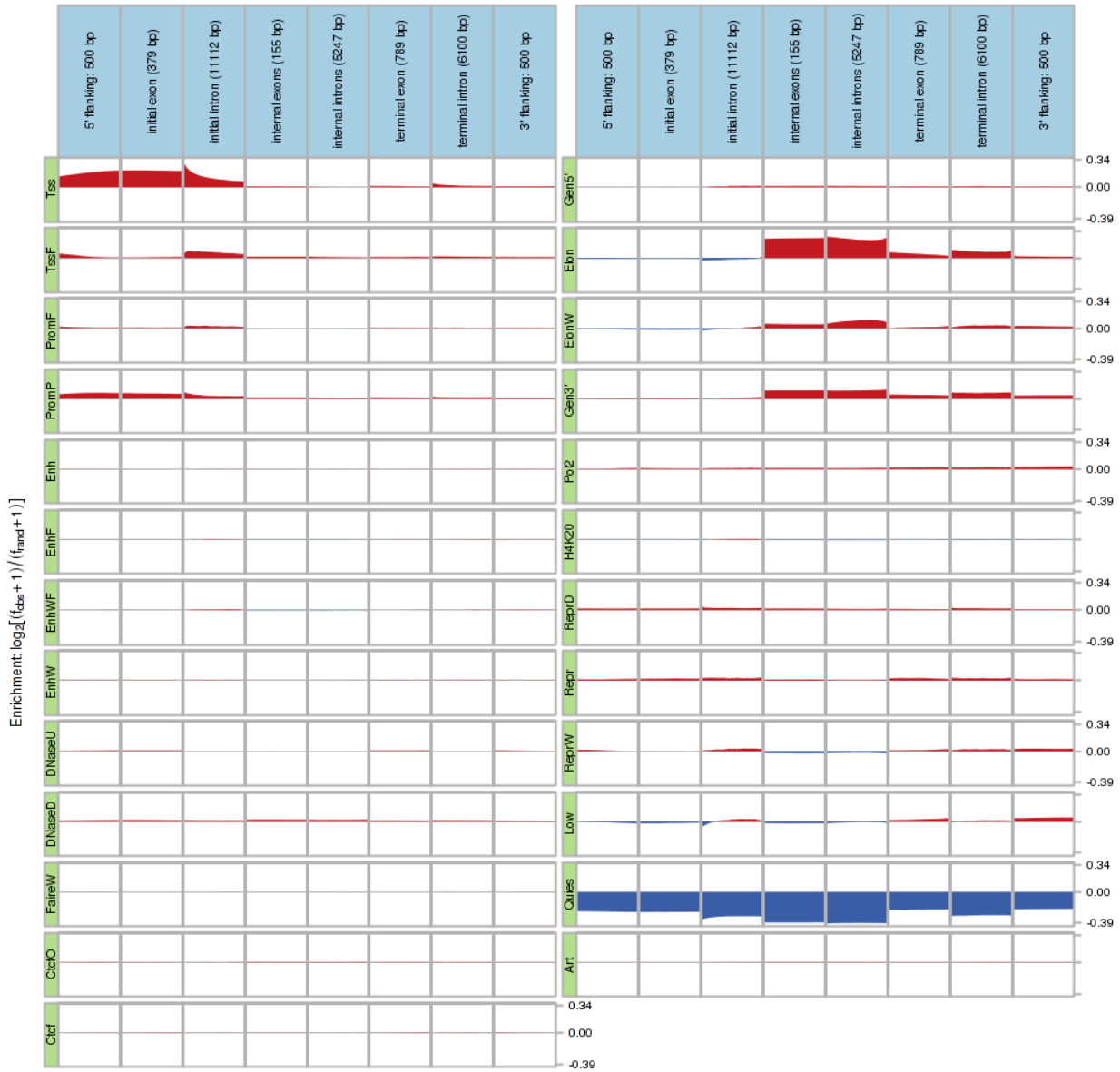
(G)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (5247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
Enh
EnhF
EnhWF
EnhW
DNaseU
DNaseD
FaireW
CtcfO
Ctcf
Gen5'
Ebn
EbnW
Gen3'
Pol2
H4K20
ReprD
Repr
ReprW
Low
Cues
Art
0.29
0.00
-0.32
Enrichment: log2[(fobs + 1)/(frand + 1)]

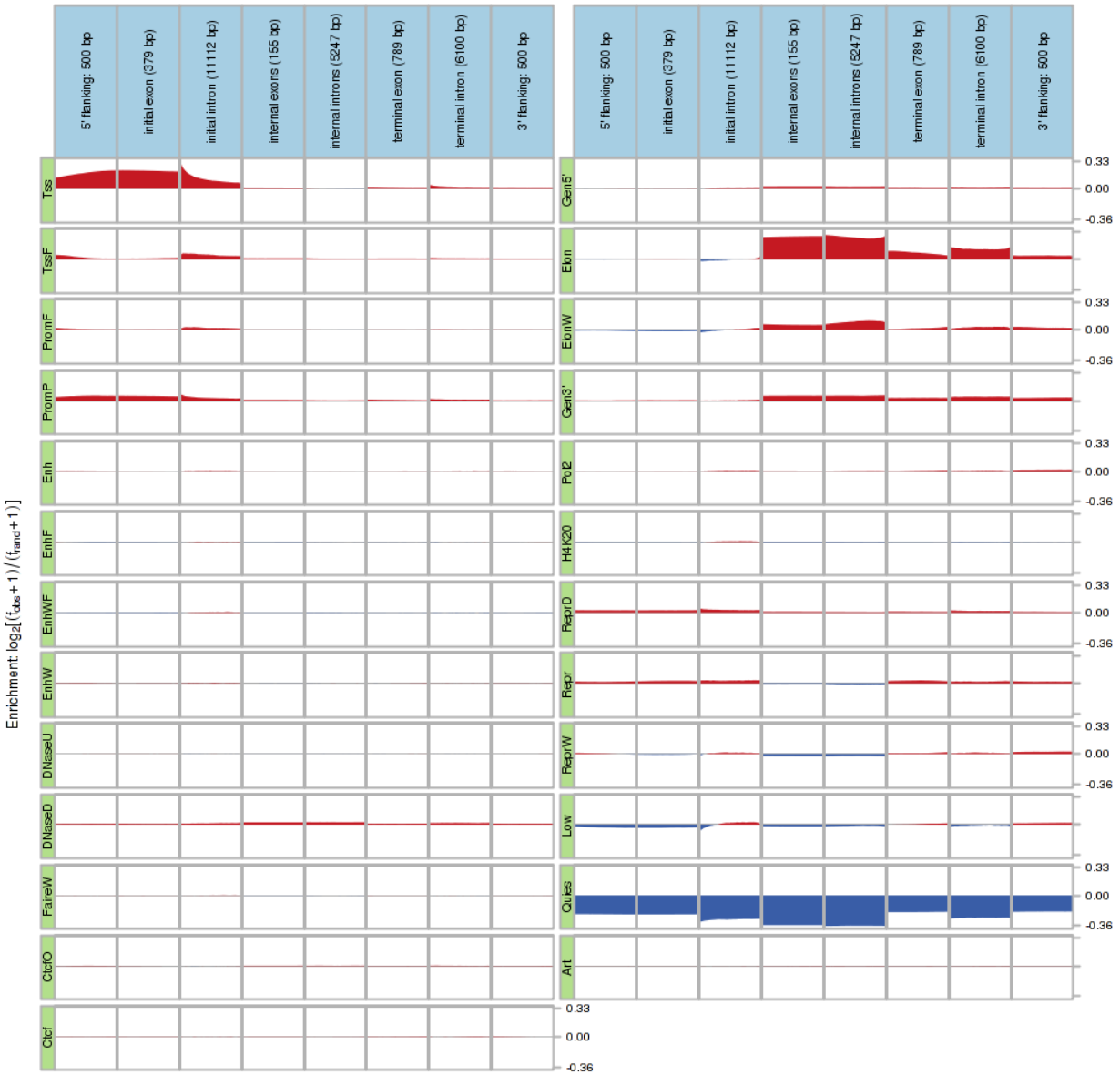
(H)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (5247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
Enh
EnhF
EnhWF
EnhW
DNaseU
DNaseD
FaireW
CtcfO
Ctcf
Gen5'
Ebn
EbnW
Gen3'
Pol2
H4K20
ReprD
Repr
ReprW
Low
Cues
Art
0.34
0.00
-0.28
Enrichment: log2[(fobs + 1)/(frand + 1)]

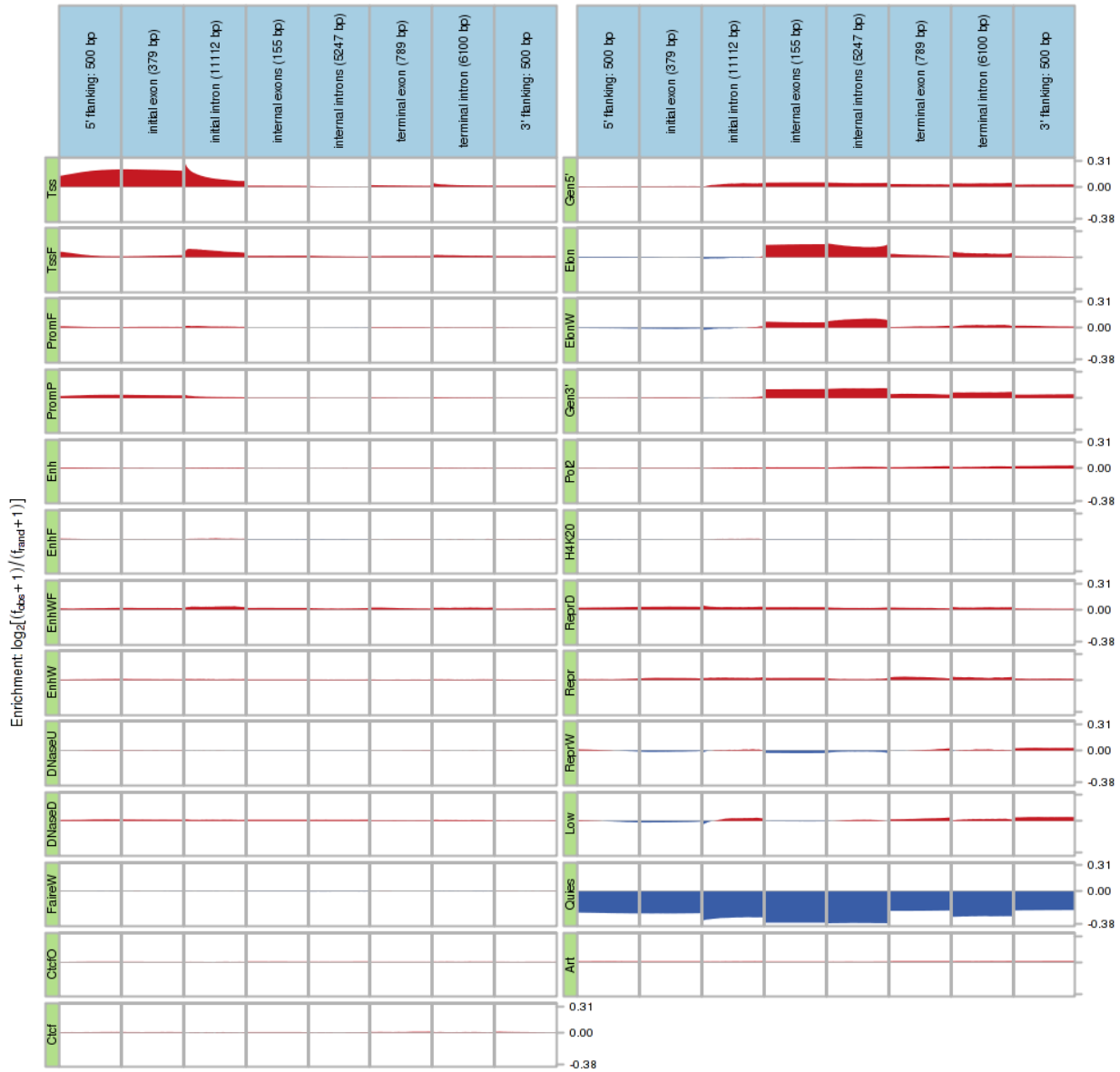
(I)


5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (3247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
Enh
EnhF
EnhWF
EnhW
DNaseU
DNaseD
FaireW
CtcfO
Ctcf
Gen5'
Ebn
EbnW
Gen3'
Pol2
H4K20
ReprD
Repr
ReprW
Low
Ques
Art
0.34
0.00
-0.39
Enrichment: log2[(fobs + 1)/(frand + 1)]

(J)

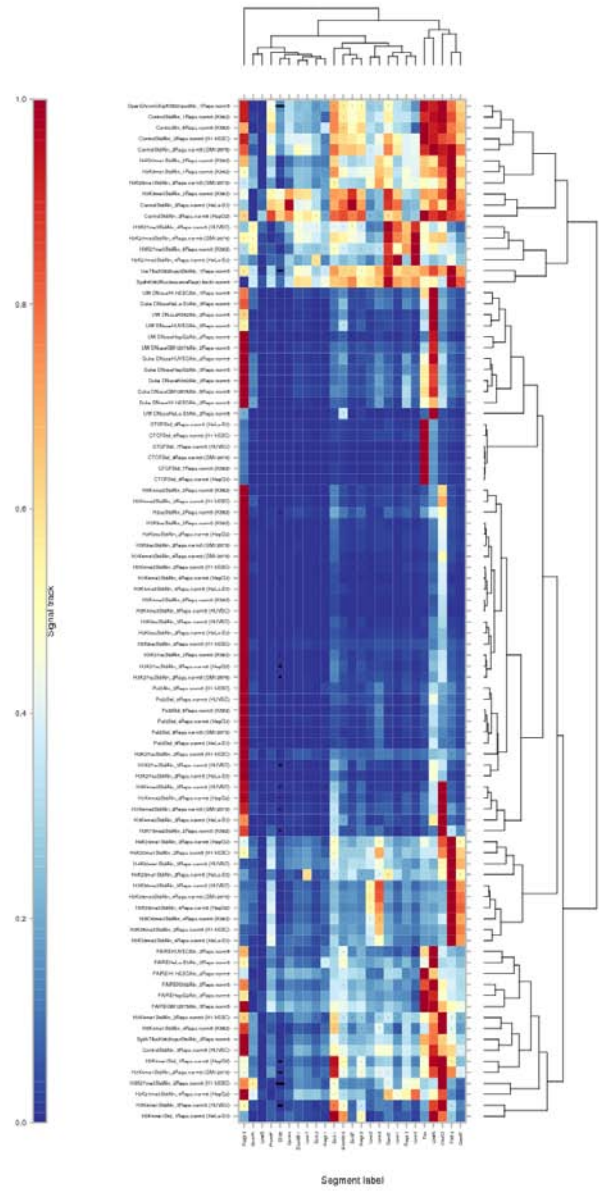

5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (8247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
Enh
EnhF
EnhWF
EnhW
DNaseU
DNaseD
FaireW
CtcfO
Ctcf
Gen5'
Ebn
EbnW
Gen3'
Pol2
H4K20
ReprD
Repr
ReprW
Low
Cues
Art
0.33
0.00
-0.36
Enrichment: log2[(fobs + 1)/(frand + 1)]

(K)

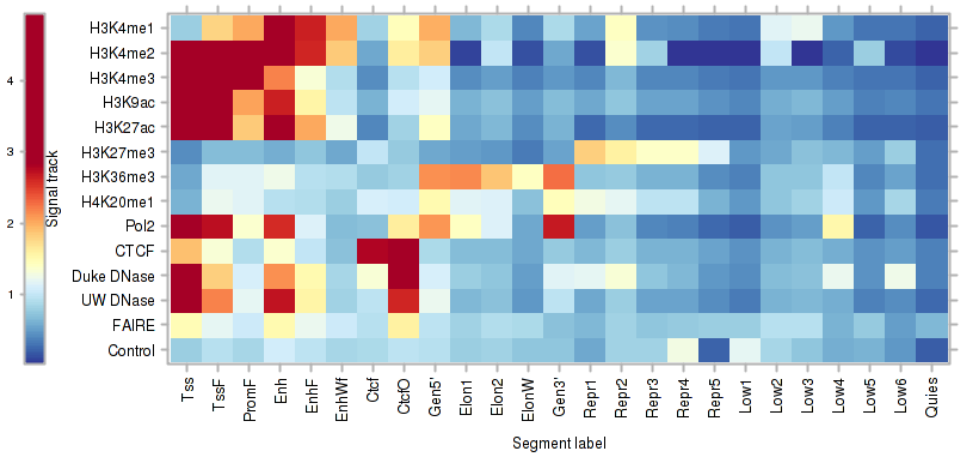

5' flanking: 500 bp
initial exon (379 bp)
initial intron (11112 bp)
internal exons (155 bp)
internal introns (3247 bp)
terminal exon (789 bp)
terminal intron (6100 bp)
3' flanking: 500 bp
Tss
TssF
PromF
PromP
Enh
EnhF
EnhWF
EnhW
DNaseU
DNaseD
FaireW
CtcfO
Ctcf
Gen5'
Ebn
EbnW
Gen3'
Pol2
H4K20
ReprD
Repr
ReprW
Low
Ques
Art
0.31
0.00
-0.38
Enrichment: log2[(fobs + 1)/(frand + 1)]

(L)

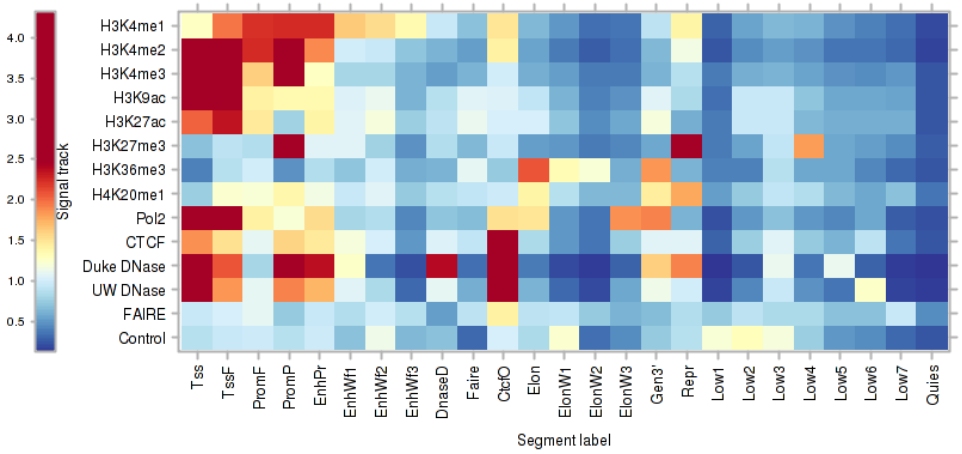
Supplementary Figure 4 (overleaf): Enrichment of various segment labels (vertically, labeled by green panels) from Segway and ChromHMM segmentations over positions on an idealized gene (horizontally, labeled by cyan panels) using the GENCODE 7 protein-coding genes. We used Segtools (50) to calculate enrichment as the base-2 logarithm of the observed frequency of a label at a particular position along a gene divided by the expected frequency of the label from its prevalence in the genome overall. Enriched positions are shown in red, and depleted positions are shown in blue. The labels for idealized gene components at the top include the mean length of that component in parentheses. Segway: (A) GM12878, (B) H1 hESC, (C) HeLa-S3, (D) HepG2, (E) HUVEC, (F) K562. ChromHMM: (G) GM12878, (H) H1 hESC, (I) HeLa-S3, (J) HepG2, (K) HUVEC, (L) K562.



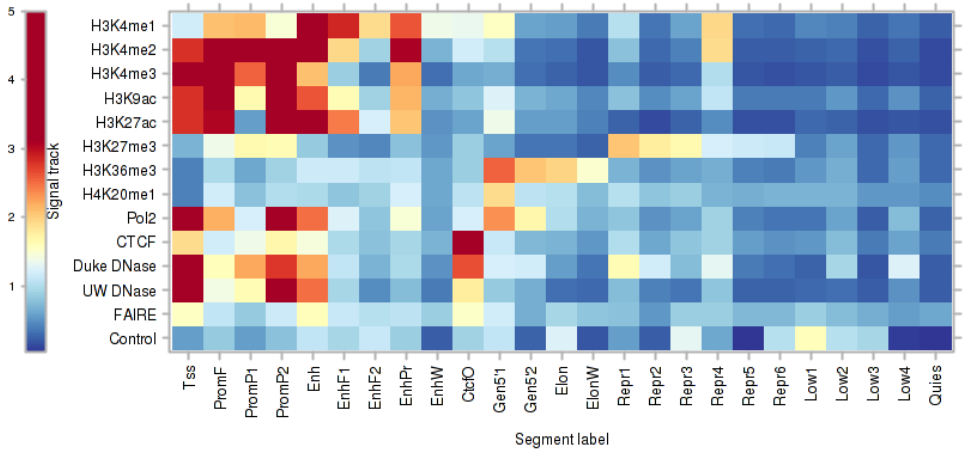
Supplementary Figure 5: Heat map of signal distribution from the Segway HeLa-S3 segmentation. The figure shows a matrix in which rows correspond to data tracks, columns correspond to Segway segment labels, and each value in the matrix is the mean data value for the given track in the given label, with values corresponding to the color bar. Rows are normalized so that the minimum and maximum values are 1 and 0, respectively.


Signal track
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K27me3
H3K36me3
H4K20me1
Pol2
CTCF
Duke DNase
UW DNase
FAIRE
Control
Tss
TssF
PromF
Enh
EnhF
EnhWf
Ctcf
CtcfO
Gen5'
Elon1
Elon2
ElonW
Gen3'
Repr1
Repr2
Repr3
Repr4
Repr5
Low1
Low2
Low3
Low4
Low5
Low6
Quies
Segment label

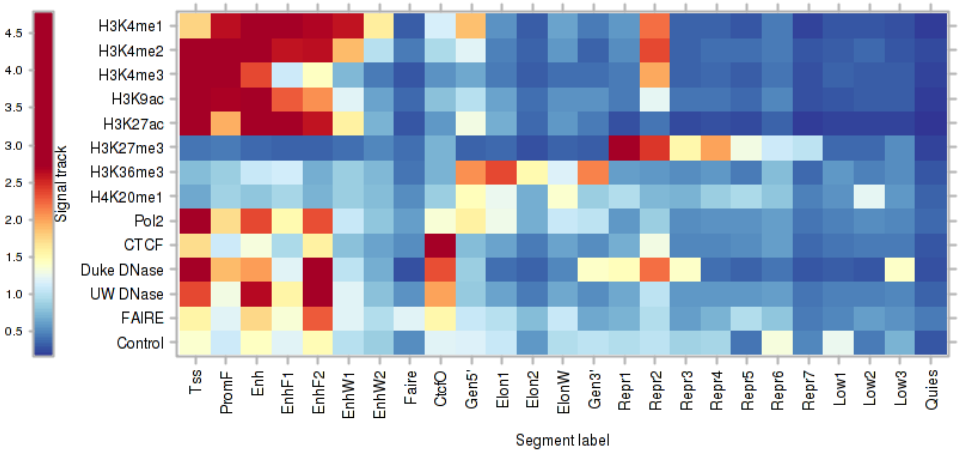
(A)


Signal track
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K27me3
H3K36me3
H4K20me1
Pol2
CTCF
Duke DNase
UW DNase
FAIRE
Control
Tss
TssF
PromF
PromP
EnhPr
EnhWf1
EnhWf2
EnhWf3
DnaseD
Faire
CtcfO
Elon
ElonW1
ElonW2
ElonW3
Gen3'
Repr
Low1
Low2
Low3
Low4
Low5
Low6
Low7
Quies
Segment label

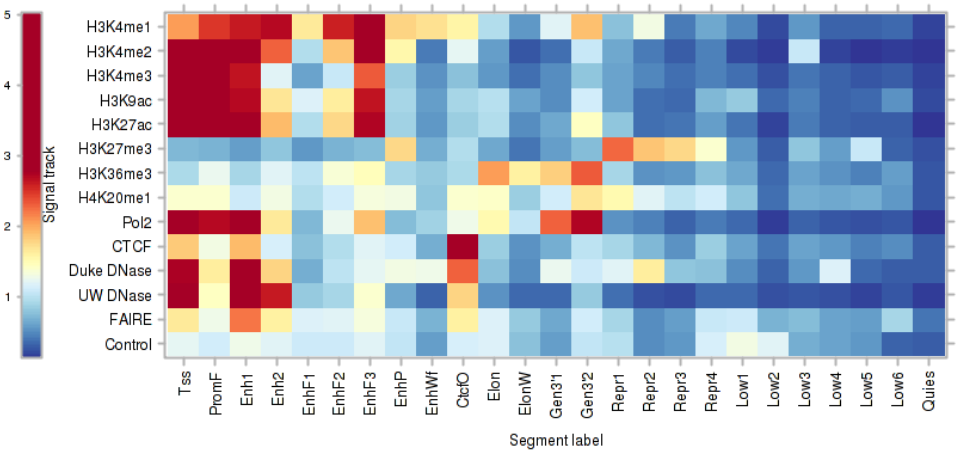
(B)


Signal track
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K27me3
H3K36me3
H4K20me1
Pol2
CTCF
Duke DNase
UW DNase
FAIRE
Control
Tss
PromF
PromP1
PromP2
Enh
EnhF1
EnhF2
EnhPr
EnhW
CtcfO
Gen5'1
Gen5'2
Elon
ElonW
Repr1
Repr2
Repr3
Repr4
Repr5
Repr6
Low1
Low2
Low3
Low4
Quies
Segment label

(C)

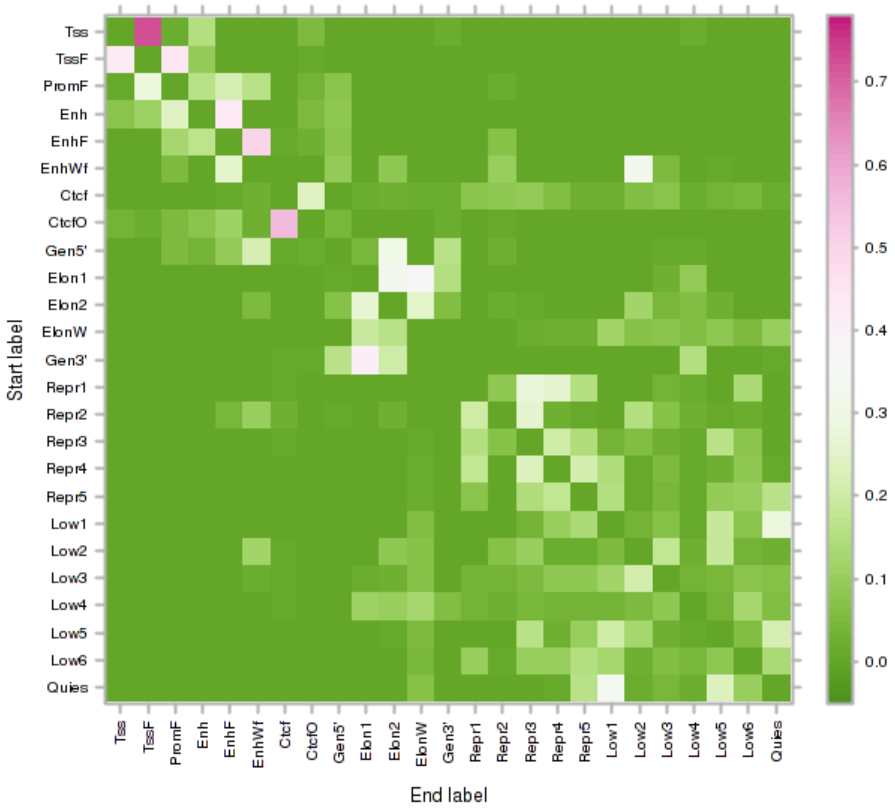


(D)

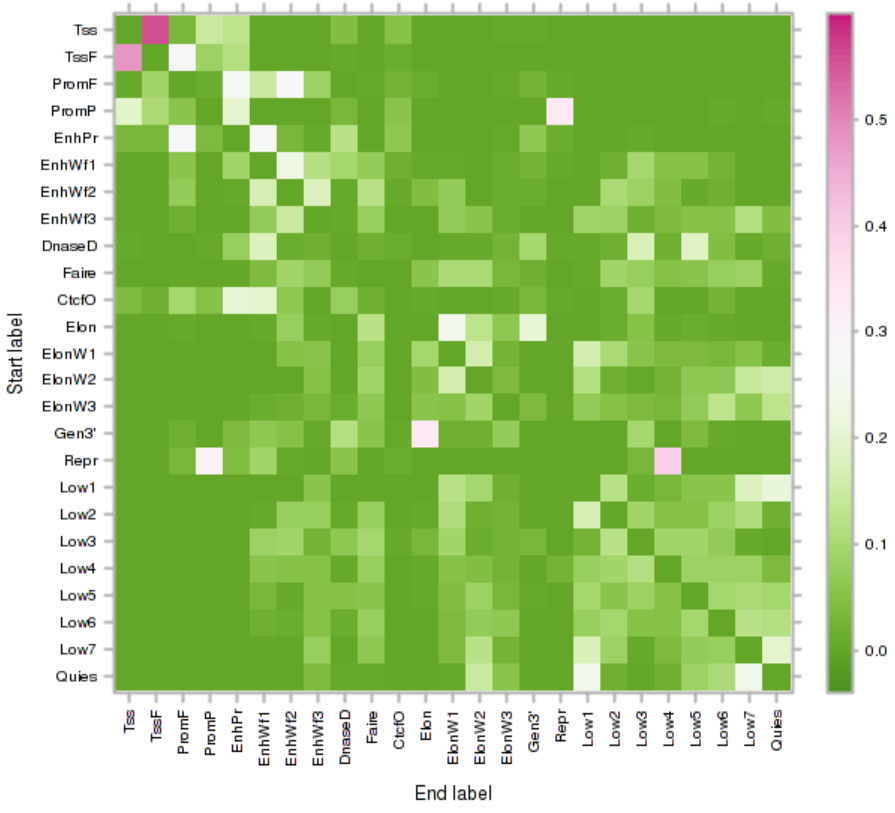


(E)

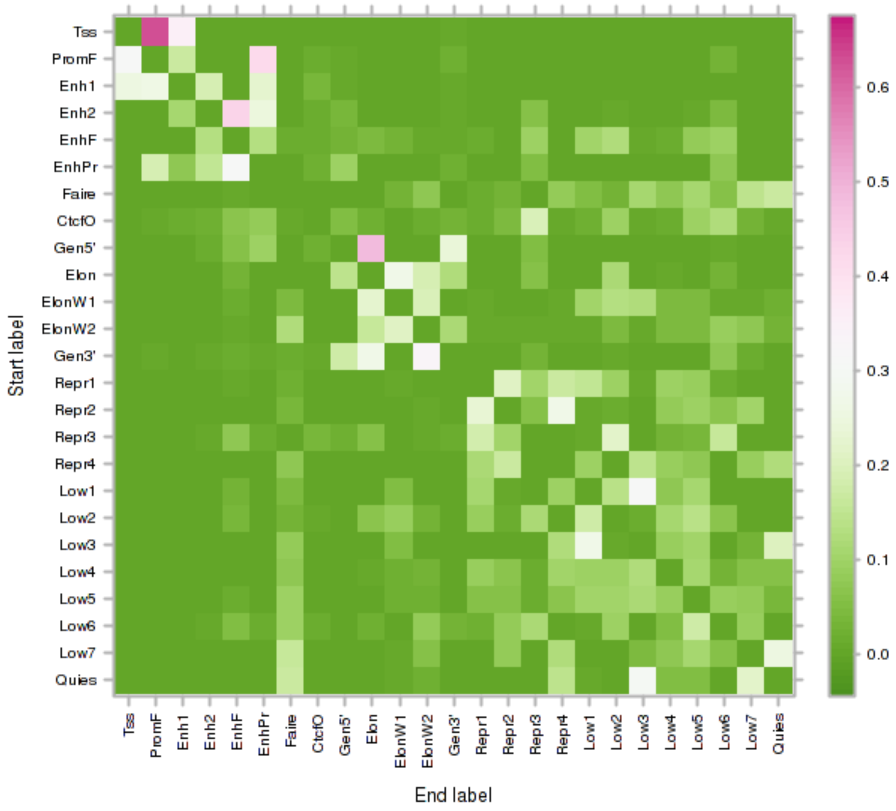
Supplementary Figure 6: Emission parameters for each segment label and signal track combination learned by Segway in various cell types. Each row corresponds to one of the input data tracks, and each column corresponds to a Segway label. Each value in the matrix corresponds to the mean of the Gaussian at the specified state. (A) GM12878, (B) H1 hESC, (C) HepG2, (D) HUVEC, (E) K562.


Start label
Tss
TssF
PromF
Enh
EnhF
EnhWf
Ctcf
CtcfO
Gen5'
Elon1
Elon2
ElonW
Gen3'
Repr1
Repr2
Repr3
Repr4
Repr5
Low1
Low2
Low3
Low4
Low5
Low6
Quies
End label
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

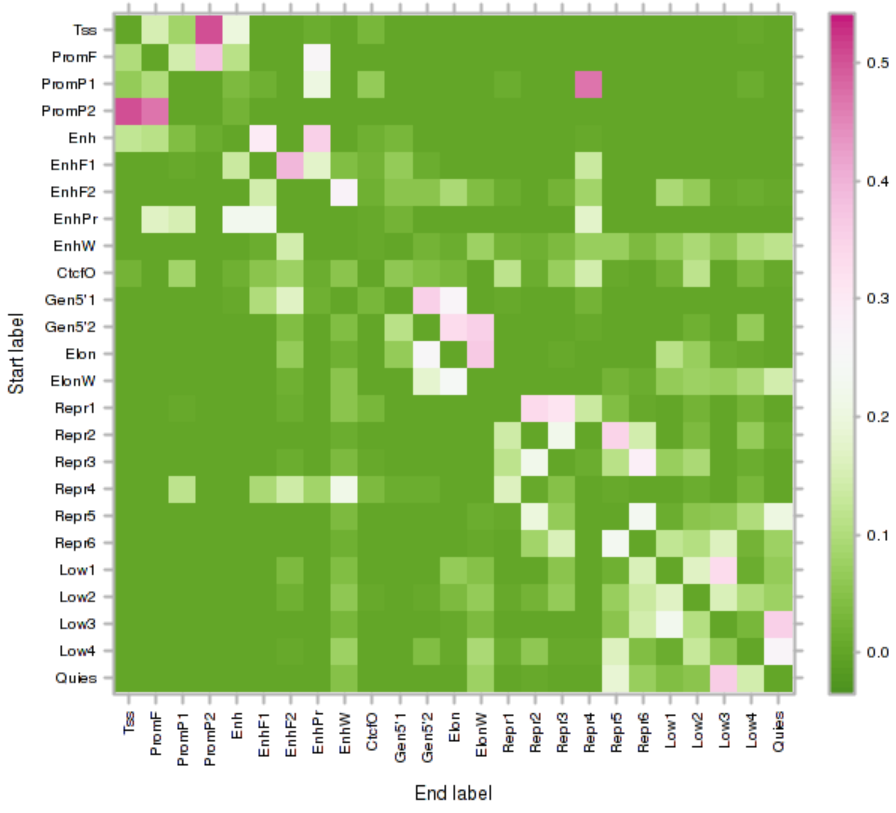
(A)


Start label
Tss
TssF
PromF
PromP
EnhPr
EnhWf1
EnhWf2
EnhWf3
DnaseD
Faire
CtcfO
Elon
ElonW1
ElonW2
ElonW3
Gen3'
Repr
Low1
Low2
Low3
Low4
Low5
Low6
Low7
Quies
End label
0.5
0.4
0.3
0.2
0.1
0.0

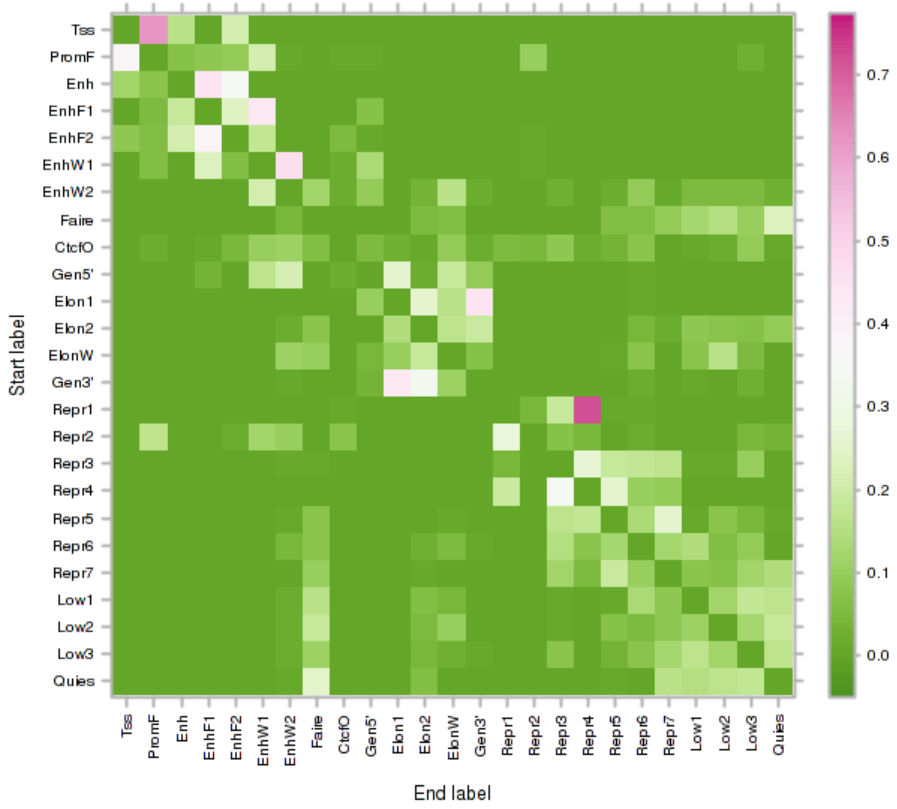
(B)


Start label
Tss
PromF
Enh1
Enh2
EnhF
EnhPr
Faire
CtcfO
Gen5'
Elon
ElonW1
ElonW2
Gen3'
Repr1
Repr2
Repr3
Repr4
Low1
Low2
Low3
Low4
Low5
Low6
Low7
Quies
End label
0.6
0.5
0.4
0.3
0.2
0.1
0.0

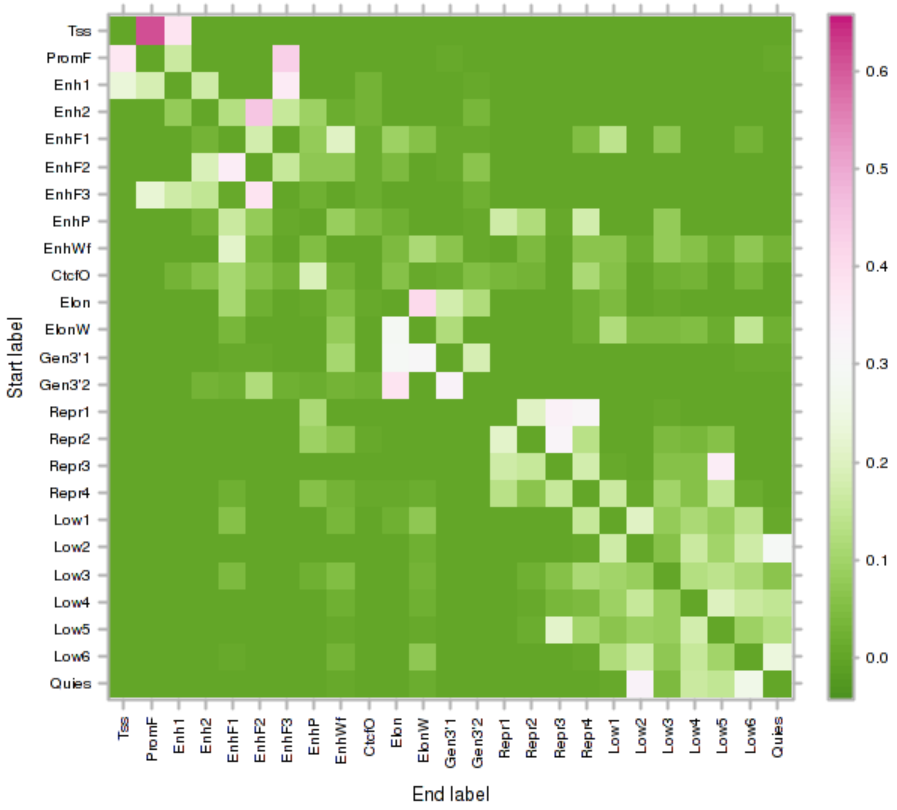
(C)


Start label
Tss
PromF
PromP1
PromP2
Enh
EnhF1
EnhF2
EnhPr
EnhW
CtcfO
Gen5'1
Gen5'2
Elon
ElonW
Repr1
Repr2
Repr3
Repr4
Repr5
Repr6
Low1
Low2
Low3
Low4
Quies
End label
0.5
0.4
0.3
0.2
0.1
0.0

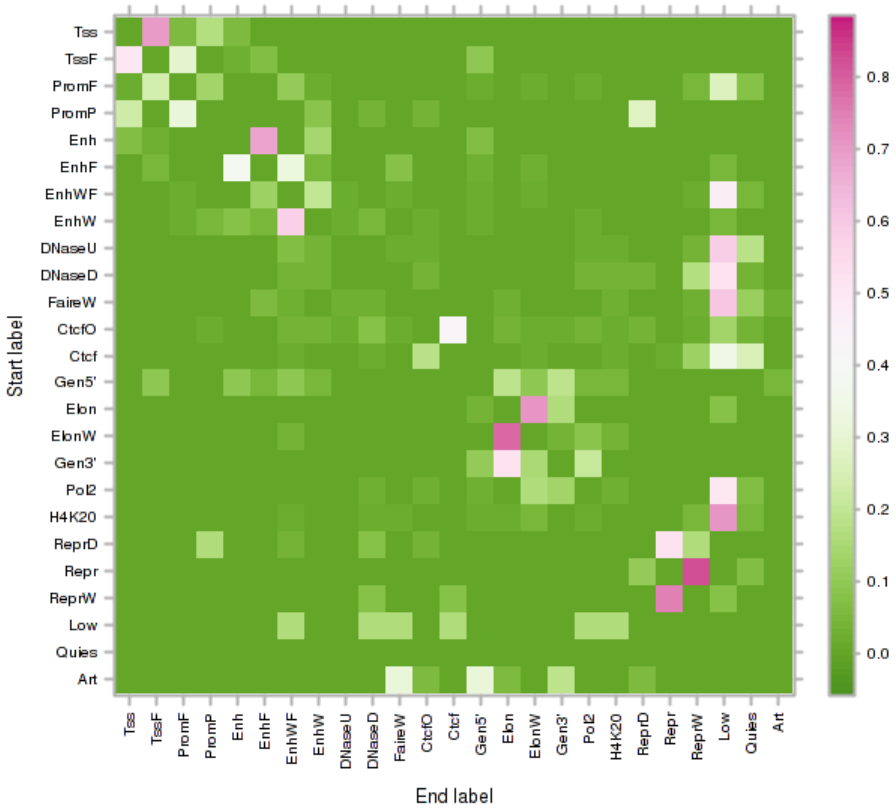
(D)


Start label
Tss
PromF
Enh
EnhF1
EnhF2
EnhW1
EnhW2
Faire
CtcfO
Gen5'
Elon1
Elon2
ElonW
Gen3'
Repr1
Repr2
Repr3
Repr4
Repr5
Repr6
Repr7
Low1
Low2
Low3
Quies
End label
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(E)

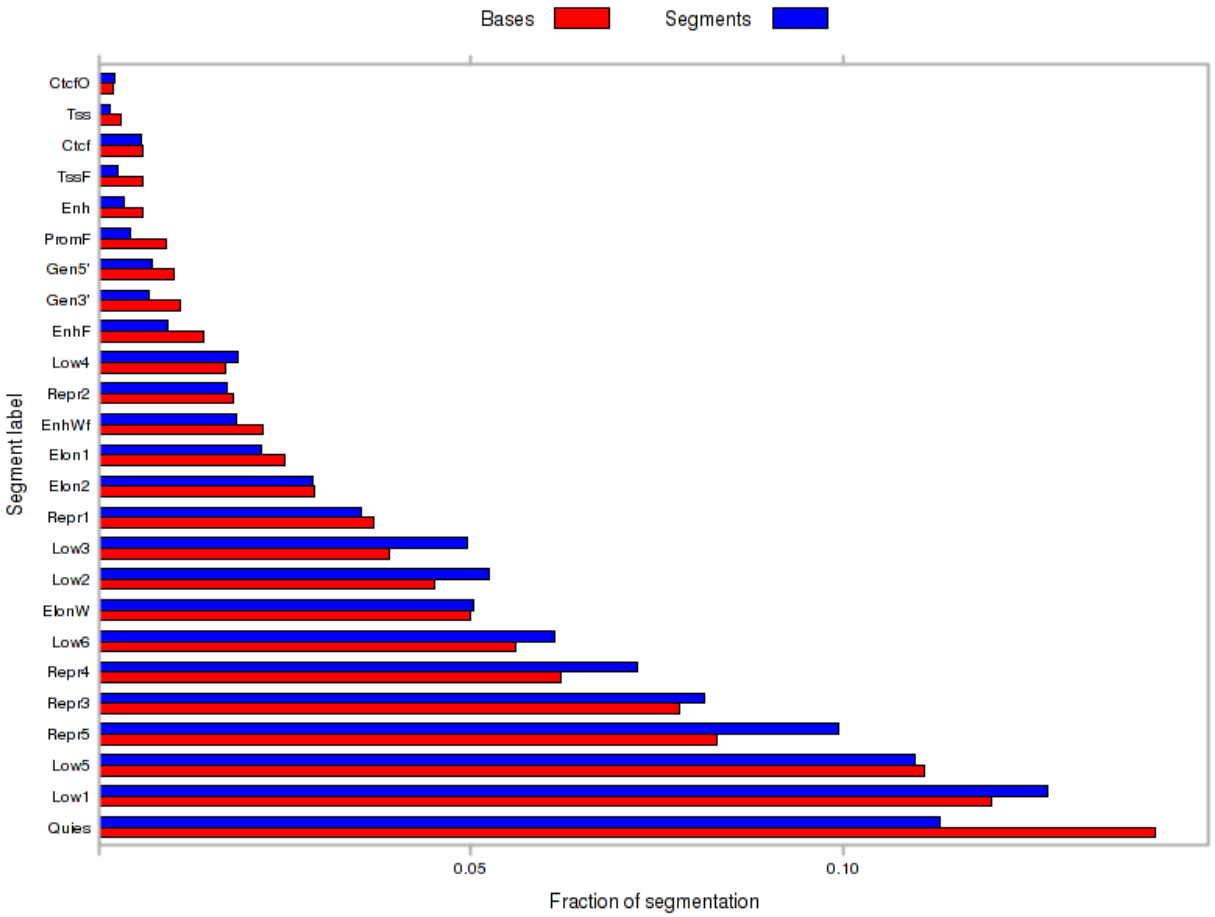

Start label
Tss
PromF
Enh1
Enh2
EnhF1
EnhF2
EnhF3
EnhP
EnhWf
CtcfO
Elon
ElonW
Gen3'1
Gen3'2
Repr1
Repr2
Repr3
Repr4
Low1
Low2
Low3
Low4
Low5
Low6
Quies
End label
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(F)

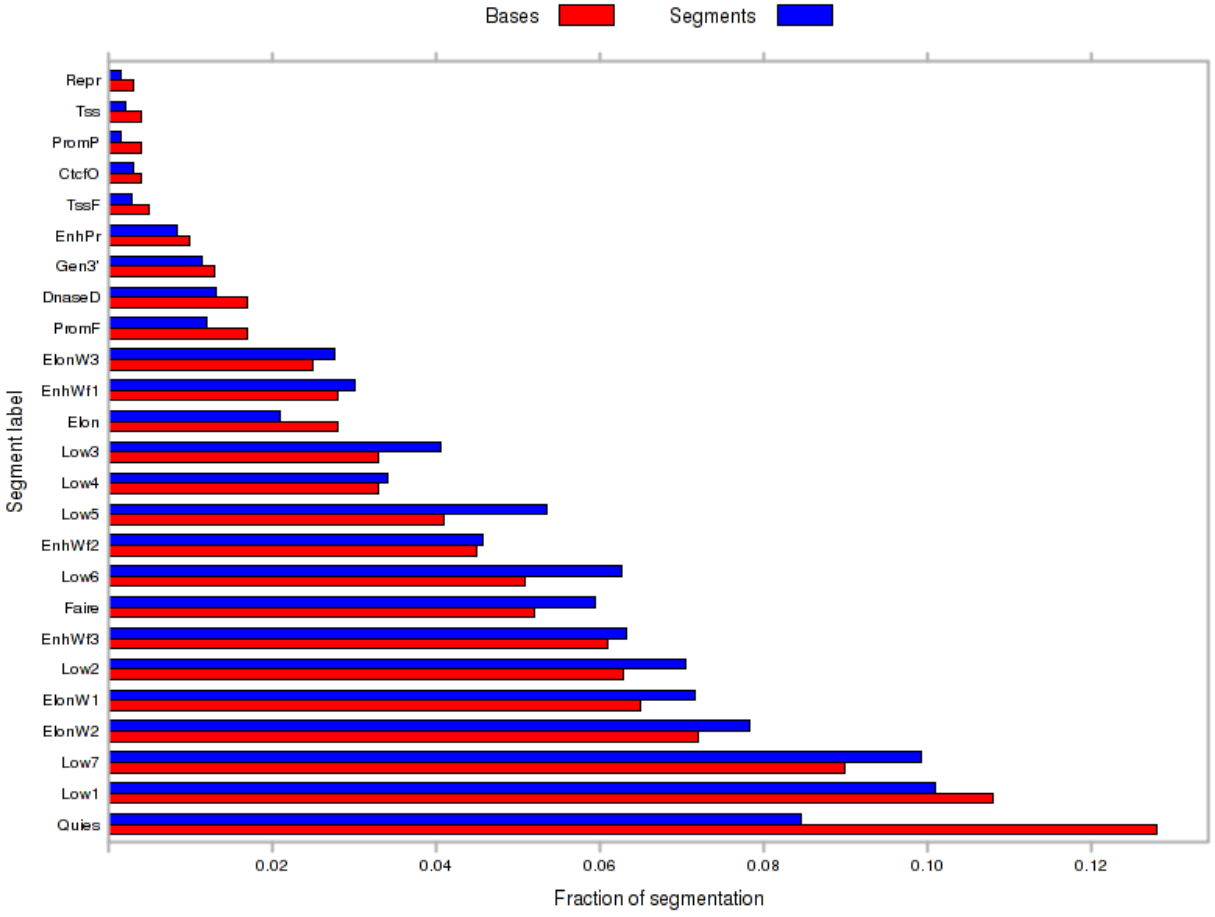


(G)

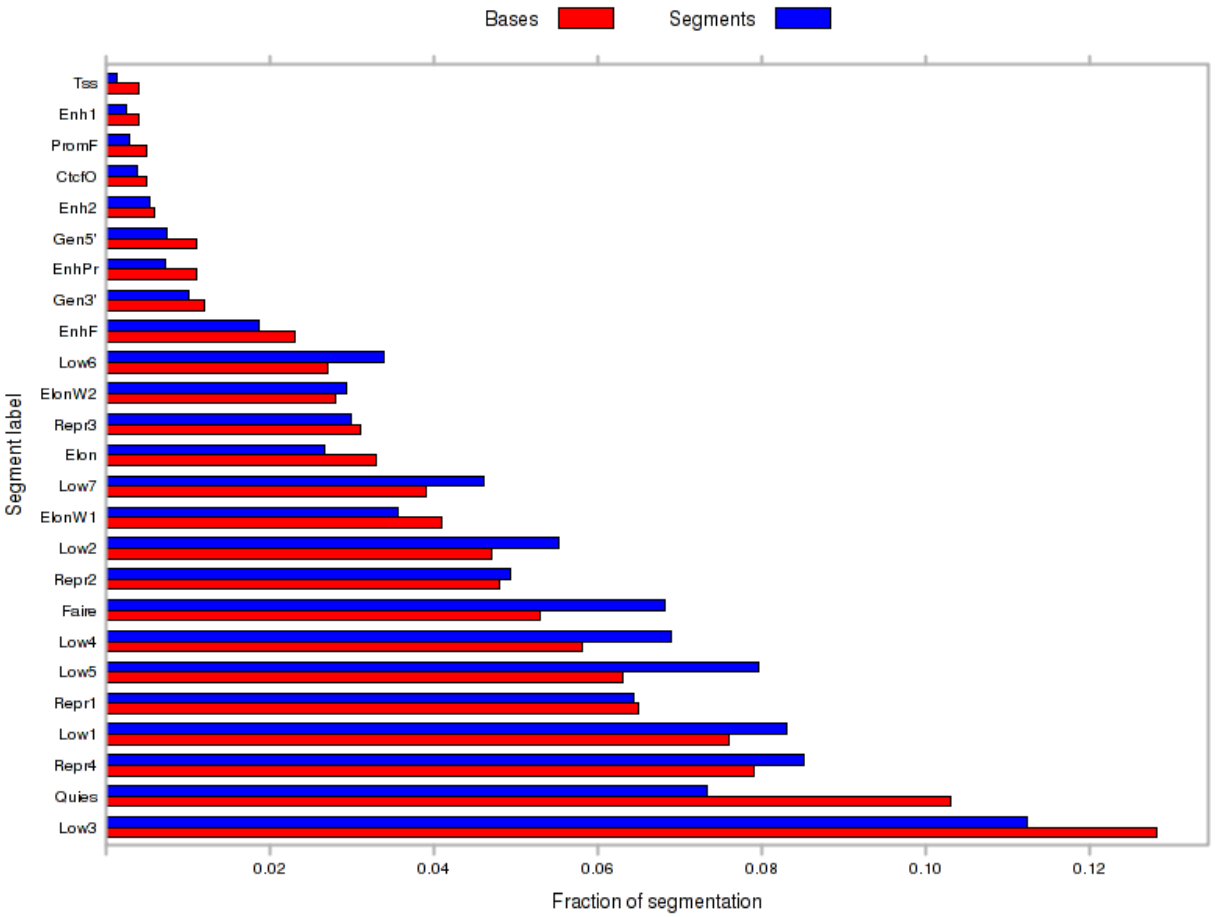
Supplementary Figure 7: Heat maps of the theoretical transition probability matrix. for the Segway (A) GM12878, (B) H1 hESC, (C) HeLa-S3, (D) HepG2, (E) HUVEC, (F) K562 models, and for the (G) ChromHMM model. Each row represents a particular start label. The cells within each row represent the probability that the model will transition to an end label given that start label, with values represented by the colors in the color bar. Diagonal cells have been zeroed to visualize non-self transitions only. Other cells that appear dark green usually represent a low nonzero value.


Bases
Segments
CtcfO
Tss
Ctcf
TssF
Enh
PromF
Gen5'
Gen3'
EnhF
Low4
Repr2
EnhWf
Elon1
Elon2
Repr1
Low3
Low2
ElonW
Low6
Repr4
Repr3
Repr5
Low5
Low1
Quies
Segment label
0.05
0.10
Fraction of segmentation

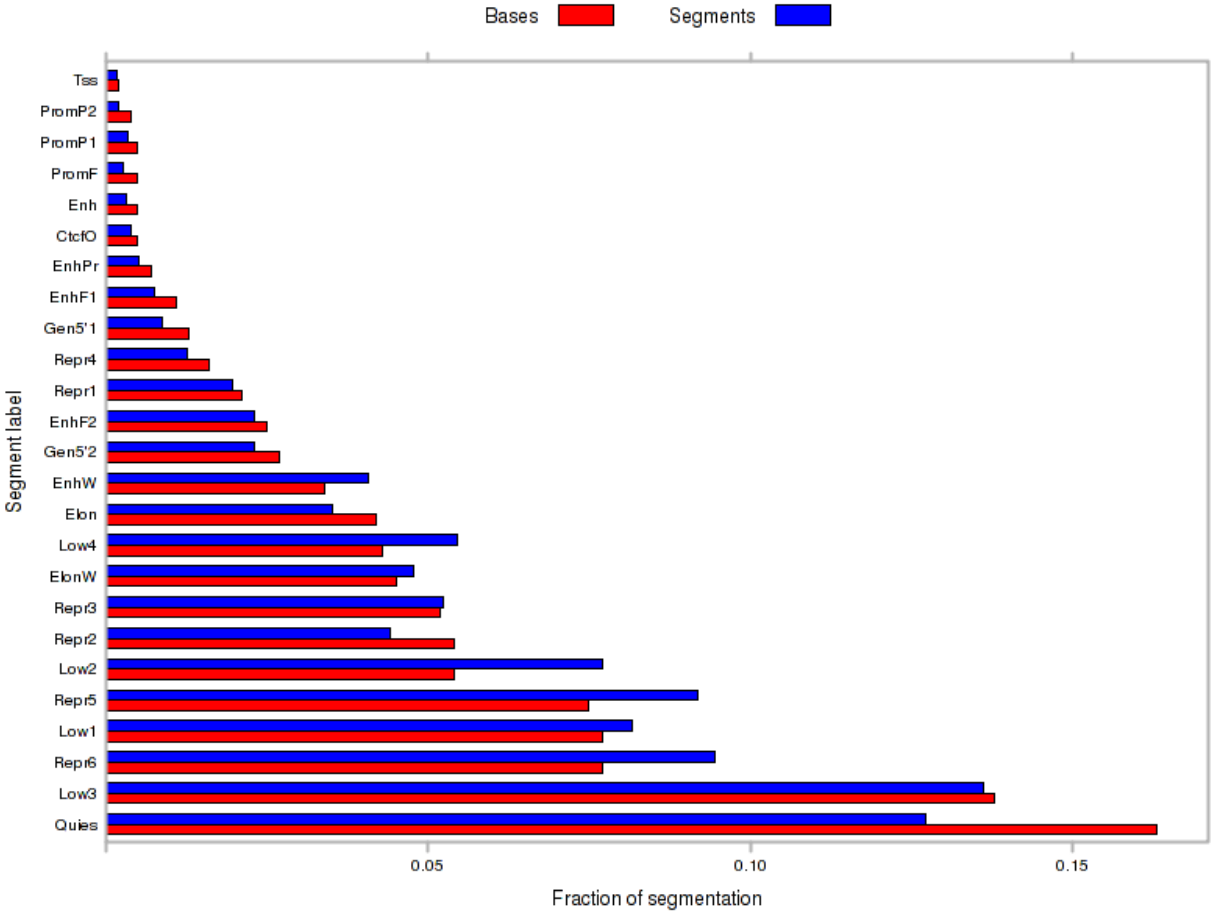
(A)


Bases
Segments
Repr
Tss
PromP
CtcfO
TssF
EnhPr
Gen3'
DnaseD
PromF
ElonW3
EnhWf1
Elon
Low3
Low4
Low5
EnhWf2
Low6
Faire
EnhWf3
Low2
ElonW1
ElonW2
Low7
Low1
Quies
Segment label
0.02
0.04
0.06
0.08
0.10
0.12
Fraction of segmentation

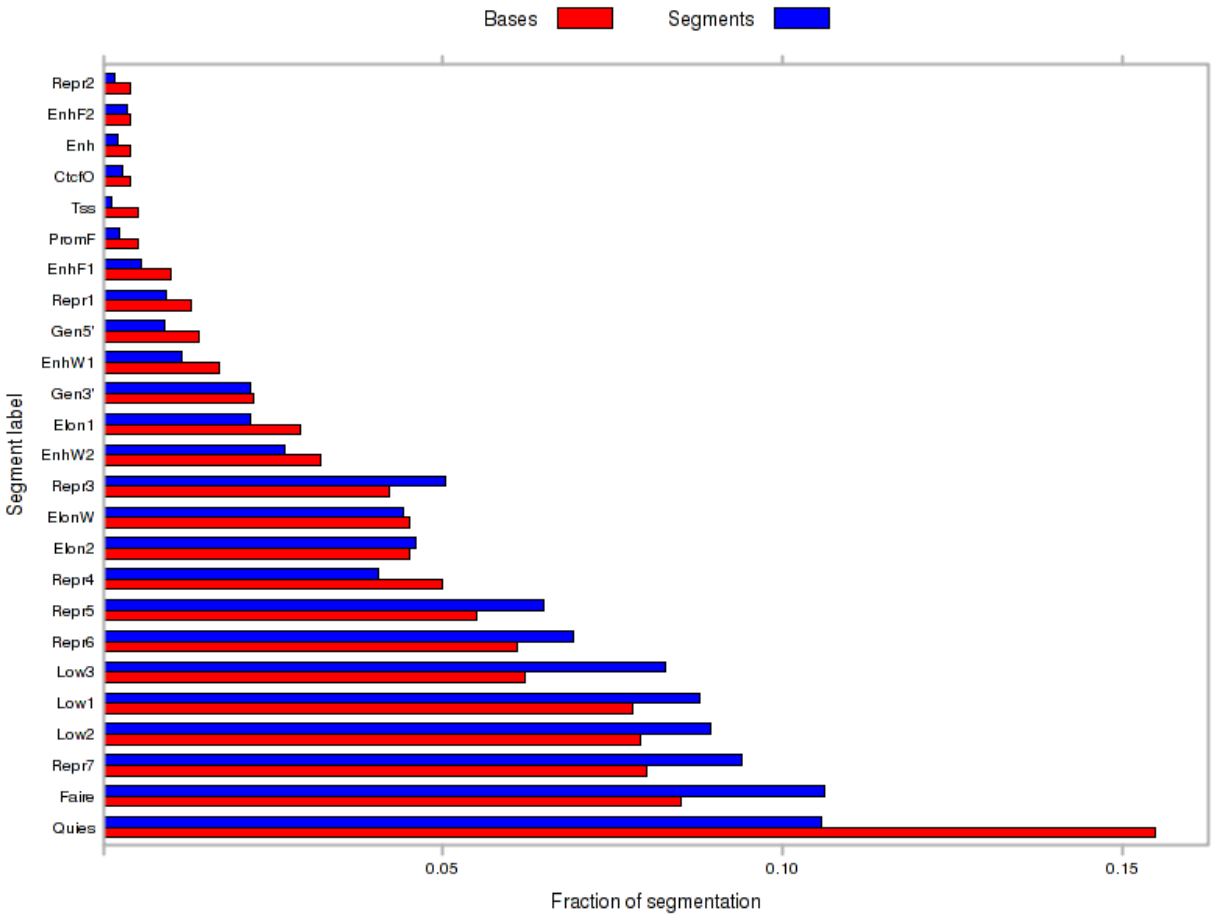
(B)


Bases
Segments
Tss
Enh1
PromF
CtcfO
Enh2
Gen5'
EnhPr
Gen3'
EnhF
Low6
ElonW2
Repr3
Elon
Low7
ElonW1
Low2
Repr2
Faire
Low4
Low5
Repr1
Low1
Repr4
Quies
Low3
Segment label
0.02
0.04
0.06
0.08
0.10
0.12
Fraction of segmentation

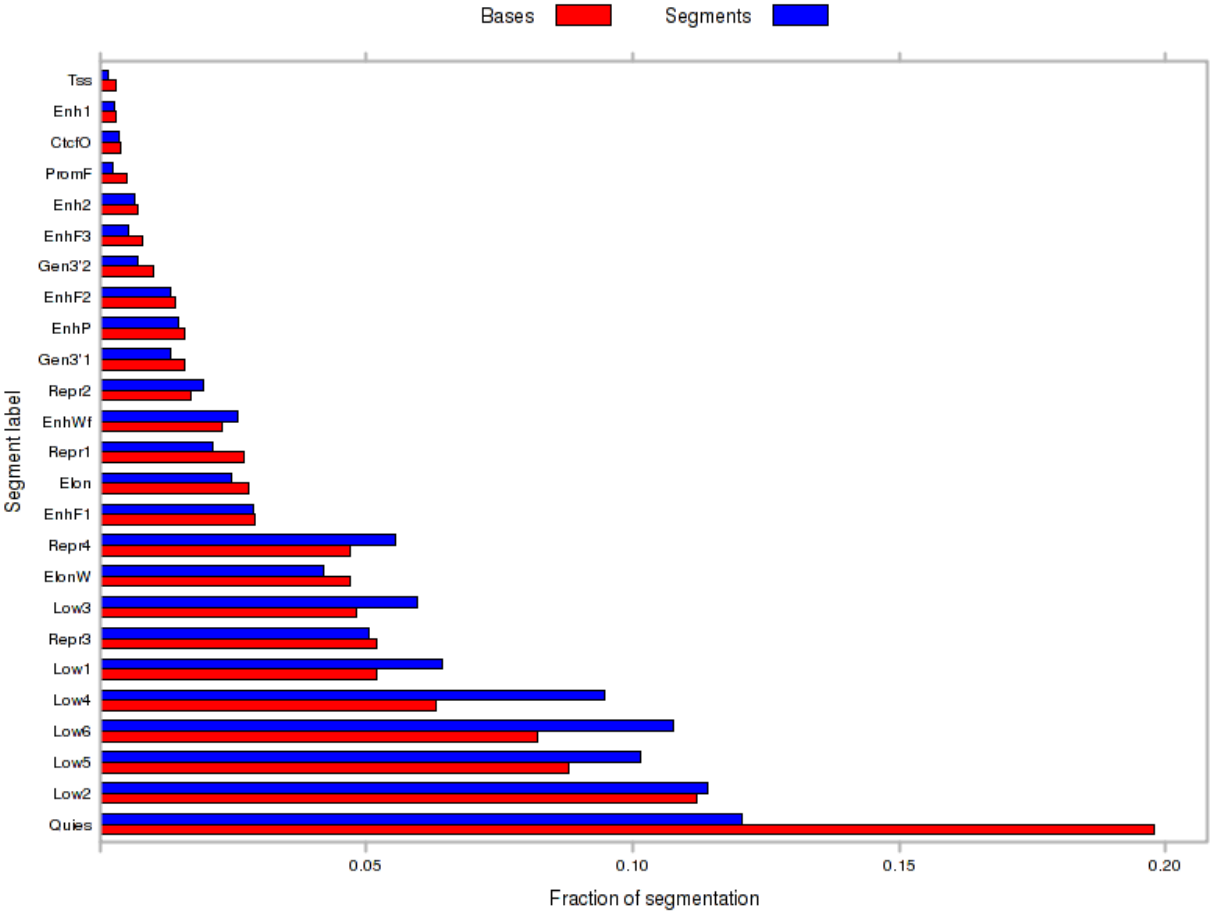
(C)


Bases
Segments
Tss
PromP2
PromP1
PromF
Enh
CtcfO
EnhPr
EnhF1
Gen5'1
Repr4
Repr1
EnhF2
Gen5'2
EnhW
Elon
Low4
ElonW
Repr3
Repr2
Low2
Repr5
Low1
Repr6
Low3
Quies
Segment label
0.05
0.10
0.15
Fraction of segmentation

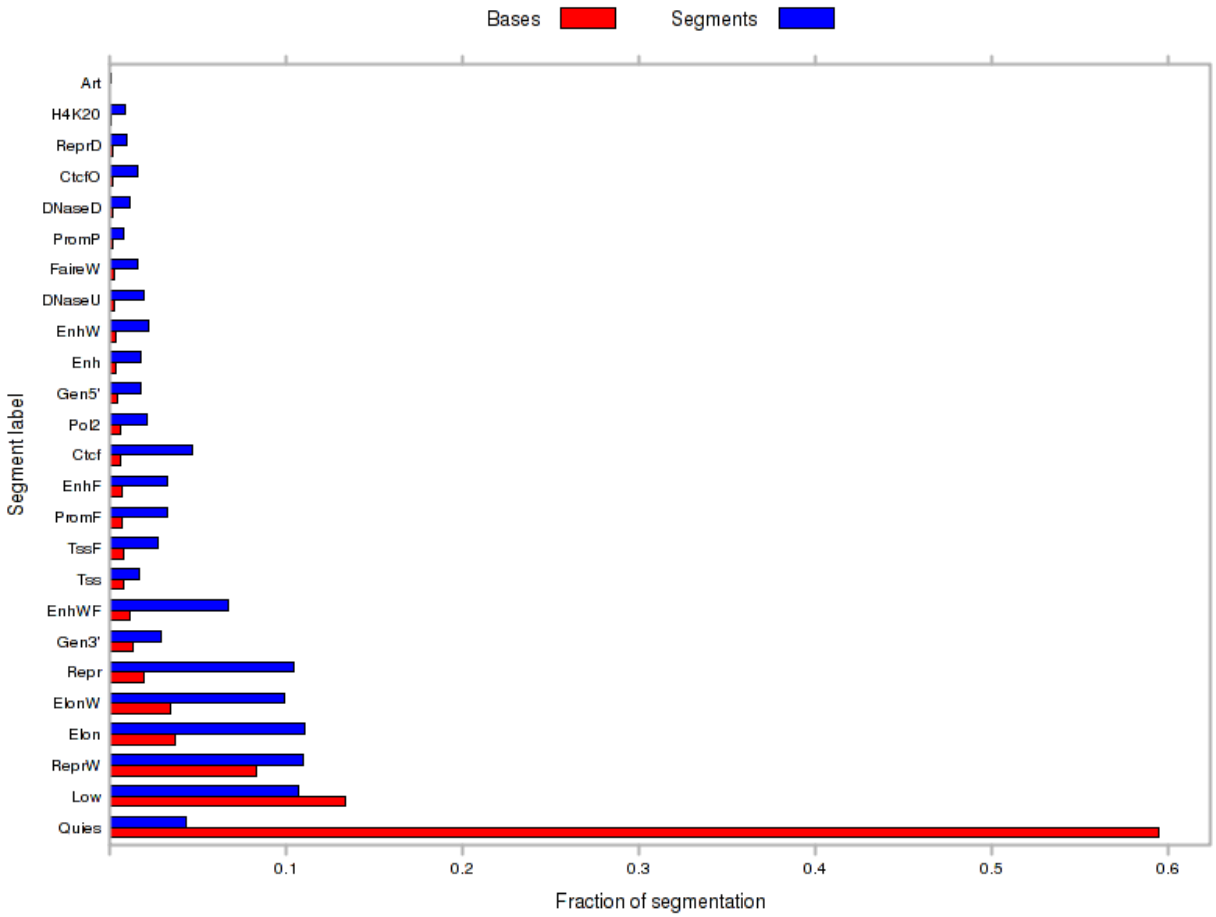
(D)


Bases
Segments
Repr2
EnhF2
Enh
CtcfO
Tss
PromF
EnhF1
Repr1
Gen5'
EnhW1
Gen3'
Elon1
EnhW2
Repr3
ElonW
Elon2
Repr4
Repr5
Repr6
Low3
Low1
Low2
Repr7
Faire
Quies
Segment label
0.05
0.10
0.15
Fraction of segmentation

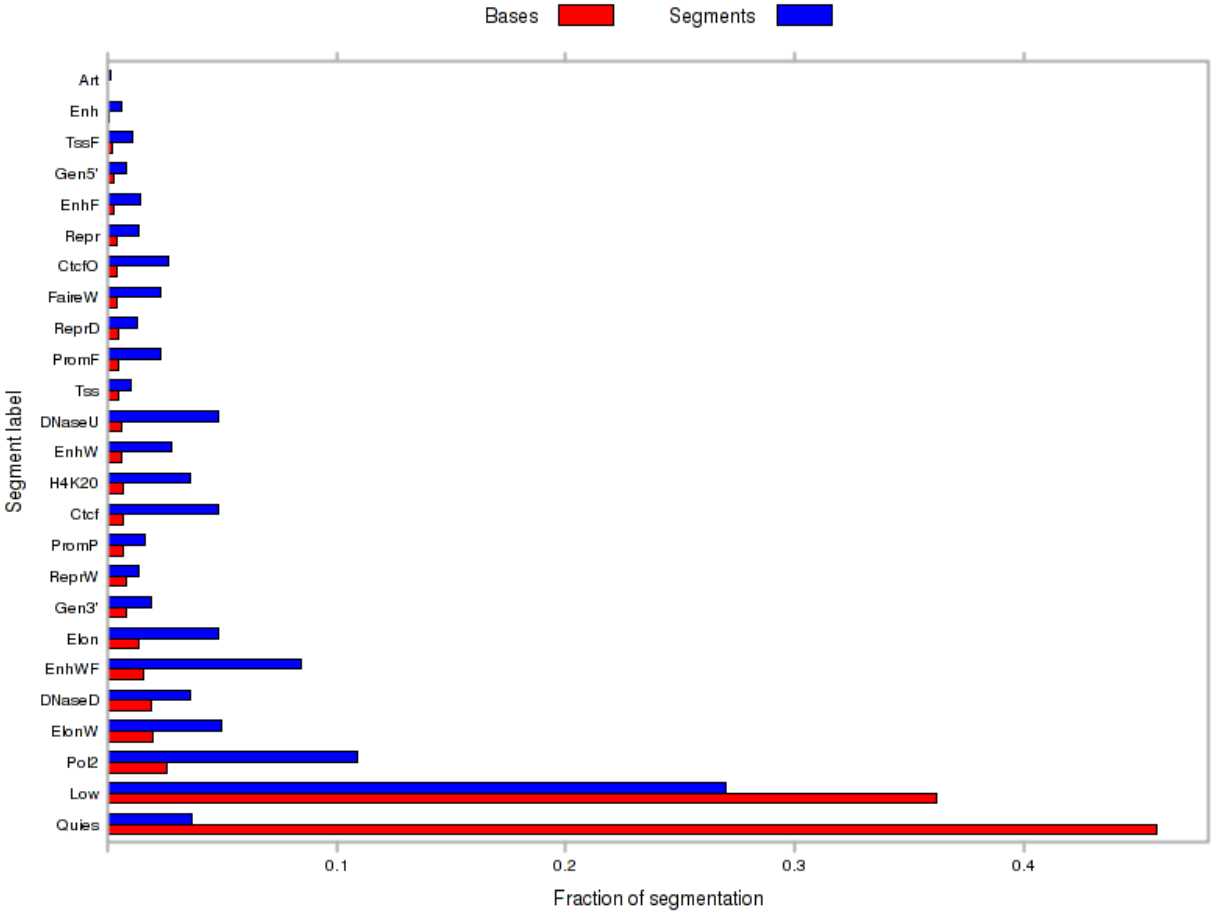
(E)


Bases
Segments
Tss
Enh1
CtcfO
PromF
Enh2
EnhF3
Gen3'2
EnhF2
EnhP
Gen3'1
Repr2
EnhWf
Repr1
Elon
EnhF1
Repr4
ElonW
Low3
Repr3
Low1
Low4
Low6
Low5
Low2
Quies
Segment label
0.05
0.10
0.15
0.20
Fraction of segmentation

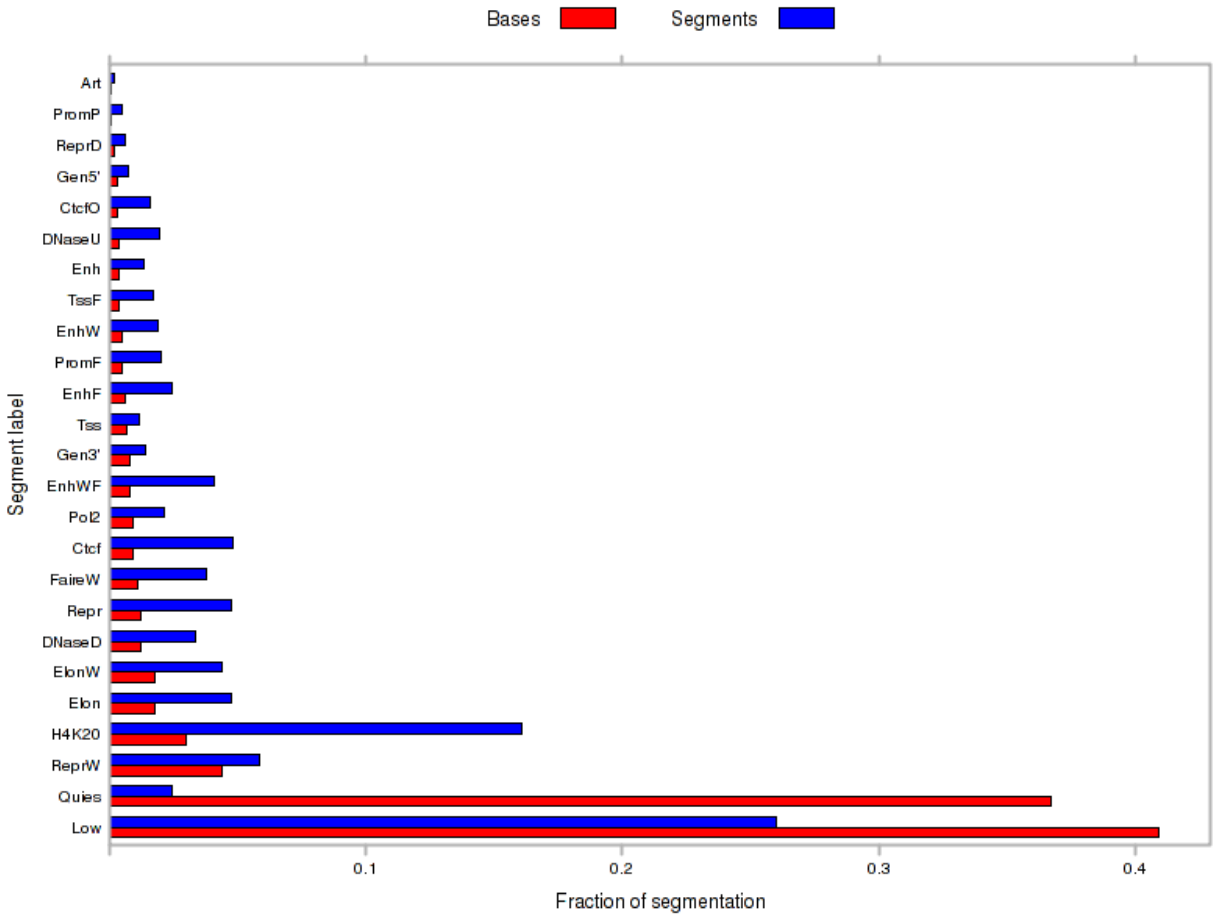
(F)


Bases
Segments
Art
H4K20
ReprD
CtcfO
DNaseD
PromP
FaireW
DNaseU
EnhW
Enh
Gen5'
Pol2
Ctcf
EnhF
PromF
TssF
Tss
EnhWF
Gen3'
Repr
ElonW
Elon
ReprW
Low
Quies
Segment label
0.1
0.2
0.3
0.4
0.5
0.6
Fraction of segmentation

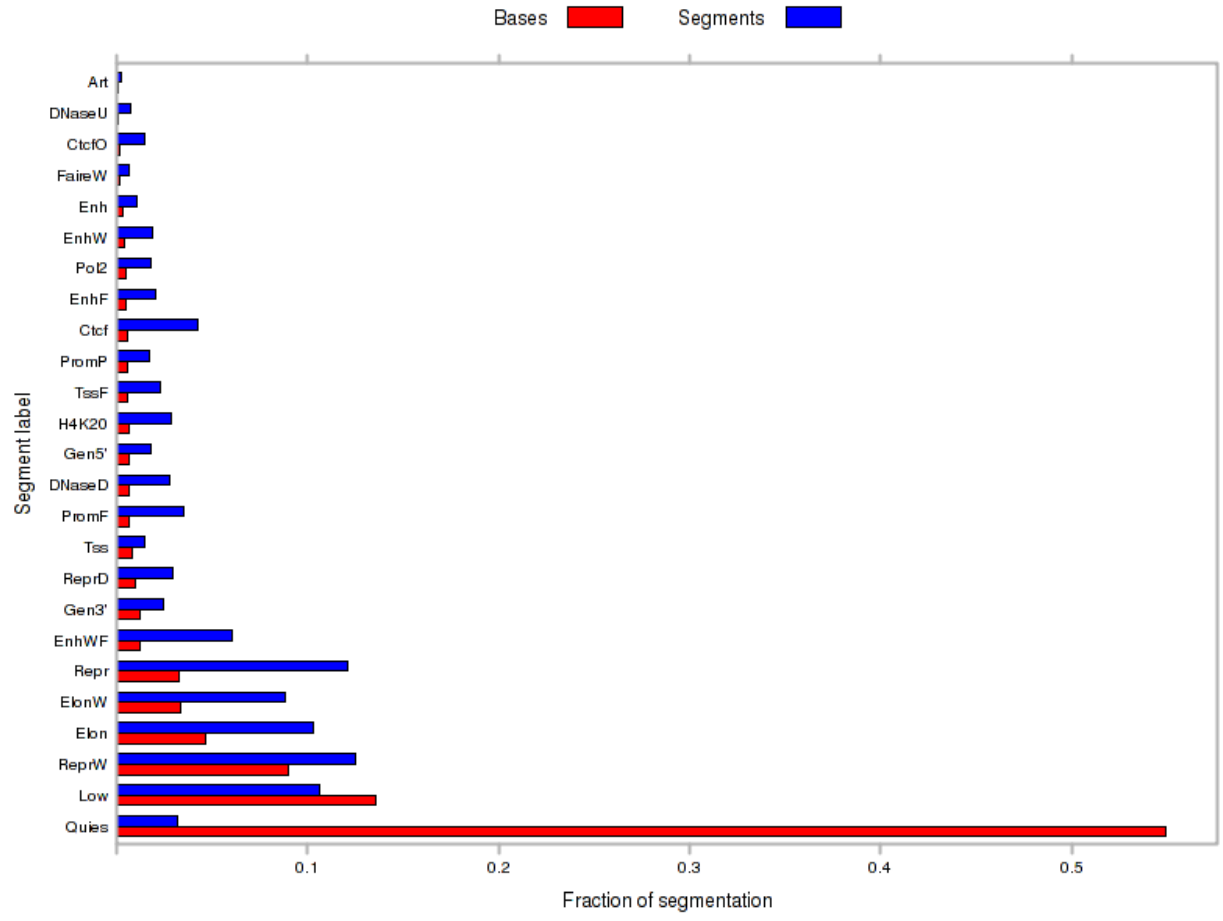
(G)


Bases
Segments
Art
Enh
TssF
Gen5'
EnhF
Repr
CtcfO
FaireW
ReprD
PromF
Tss
DNaseU
EnhW
H4K20
Ctcf
PromP
ReprW
Gen3'
Elon
EnhWF
DNaseD
ElonW
Pol2
Low
Quies
Segment label
0.1
0.2
0.3
0.4
Fraction of segmentation

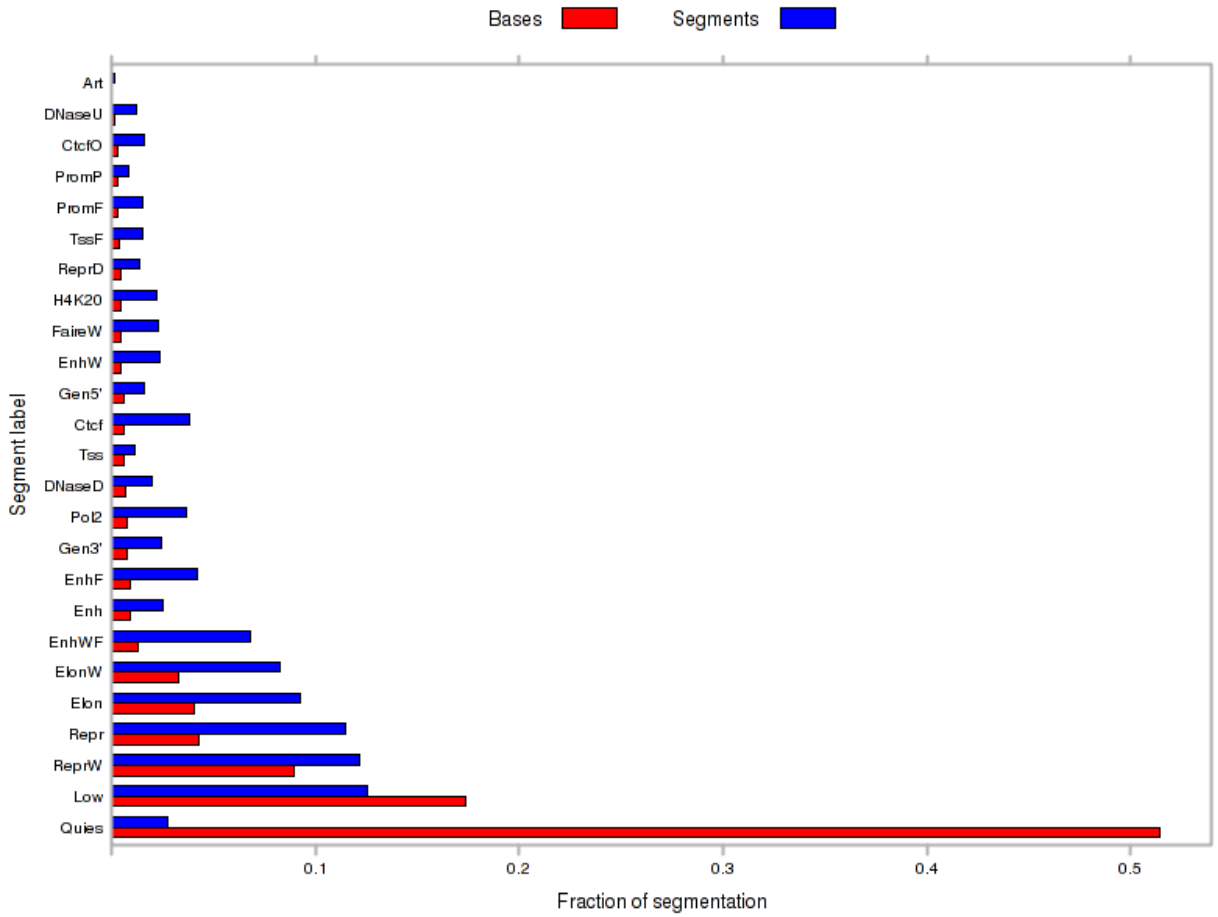
(H)


Bases
Segments
Art
PromP
ReprD
Gen5'
CtcfO
DNaseU
Enh
TssF
EnhW
PromF
EnhF
Tss
Gen3'
EnhWF
Pol2
Ctcf
FaireW
Repr
DNaseD
ElonW
Elon
H4K20
ReprW
Quies
Low
Segment label
0.1
0.2
0.3
0.4
Fraction of segmentation

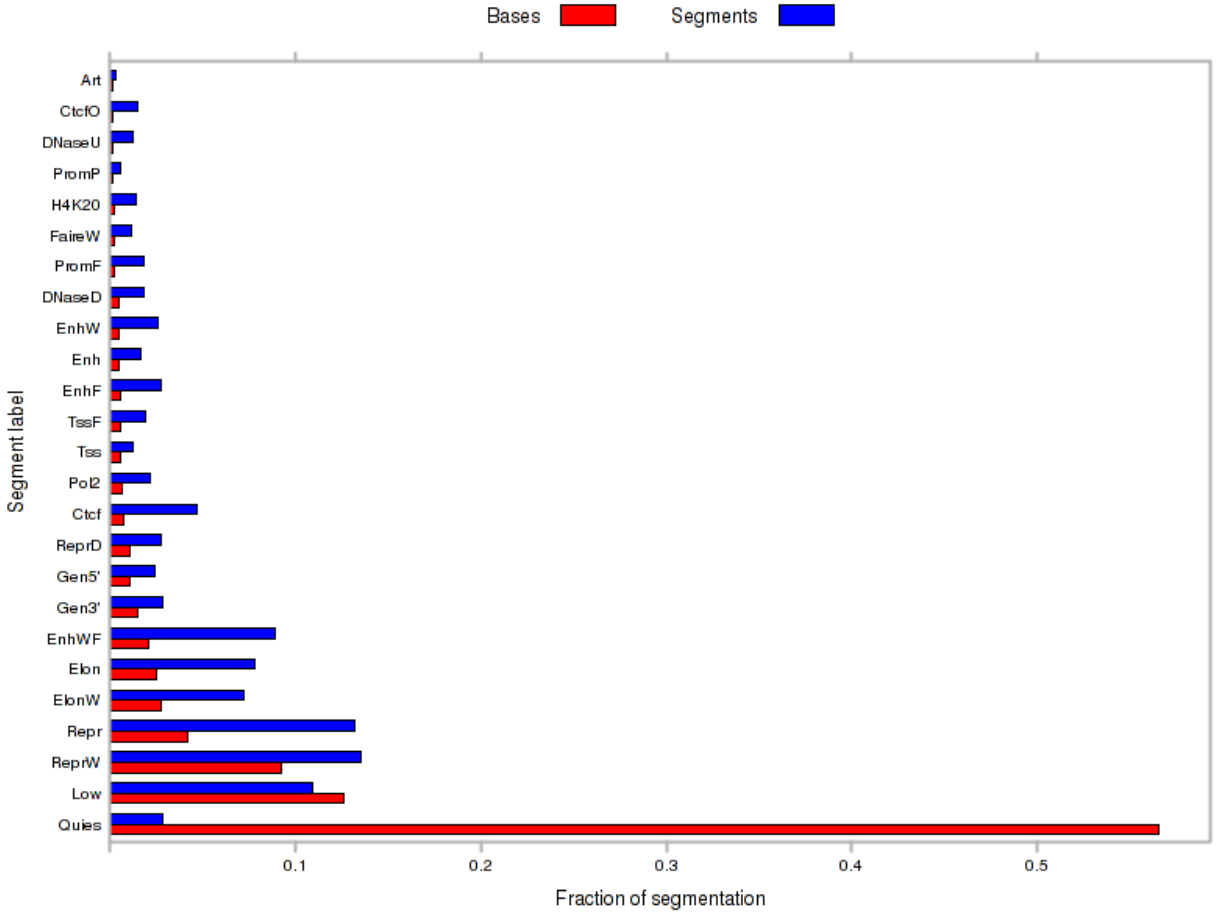
(I)


Bases
Segments
Art
DNaseU
CtcfO
FaireW
Enh
EnhW
Pol2
EnhF
Ctcf
PromP
TssF
H4K20
Gen5'
DNaseD
PromF
Tss
ReprD
Gen3'
EnhWF
Repr
ElonW
Elon
ReprW
Low
Quies
Segment label
0.1
0.2
0.3
0.4
0.5
Fraction of segmentation

(J)

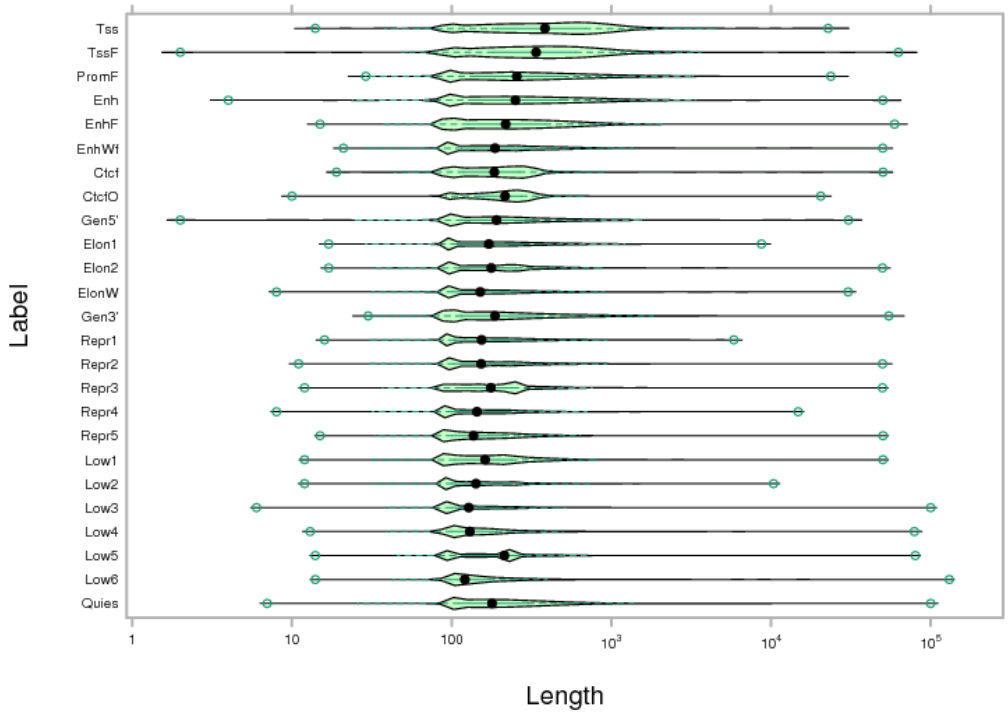

Bases
Segments
Art
DNaseU
CtcfO
PromP
PromF
TssF
ReprD
H4K20
FaireW
EnhW
Gen5'
Ctcf
Tss
DNaseD
Pol2
Gen3'
EnhF
Enh
EnhWF
ElonW
Elon
Repr
ReprW
Low
Quies
Segment label
0.1
0.2
0.3
0.4
0.5
Fraction of segmentation

(K)

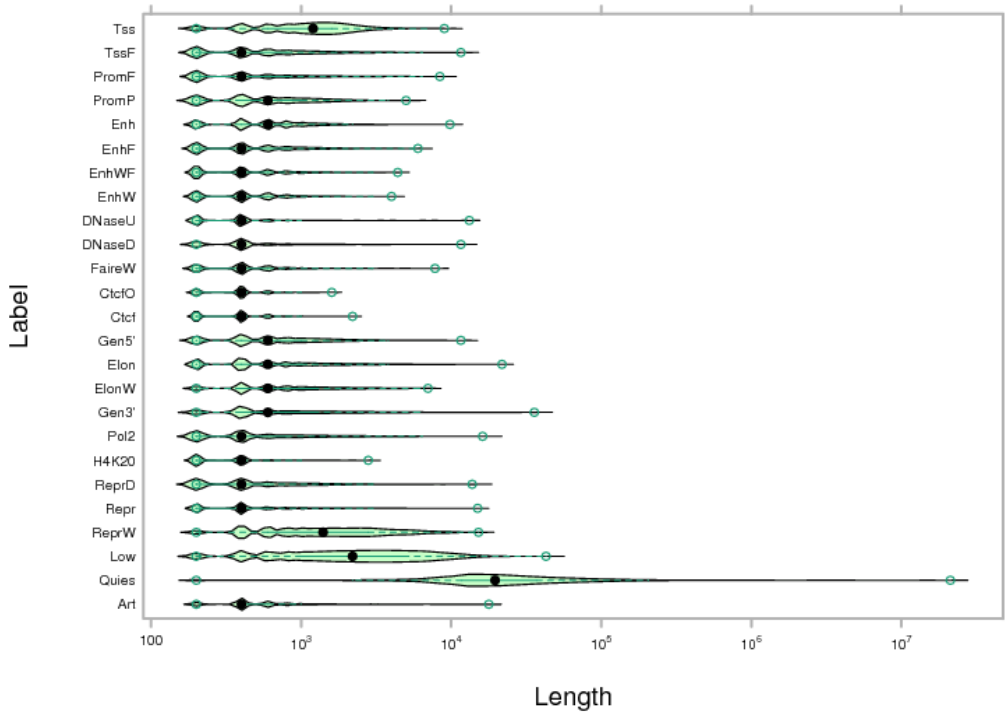

Bases
Segments
Art
CtcfO
DNaseU
PromP
H4K20
FaireW
PromF
DNaseD
EnhW
Enh
EnhF
TssF
Tss
Pol2
Ctcf
ReprD
Gen5'
Gen3'
EnhWF
Elon
ElonW
Repr
ReprW
Low
Quies
Segment label
0.1
0.2
0.3
0.4
0.5
Fraction of segmentation

(L)

Supplementary Figure 8: Fraction of segmentation covered by each label, measured in number of bases (red) and number of segments (blue) for the Segway segmentations of: (A) GM12878, (B) H1 hESC, (C) HeLa-S3, (D) HepG2, (E) HUVEC, and (F) K562, and the ChromHMM segmentations of: (G) GM12878, (H) H1 hESC, (I) HeLa-S3, (J) HepG2, (K) HUVEC, and (L) K562.

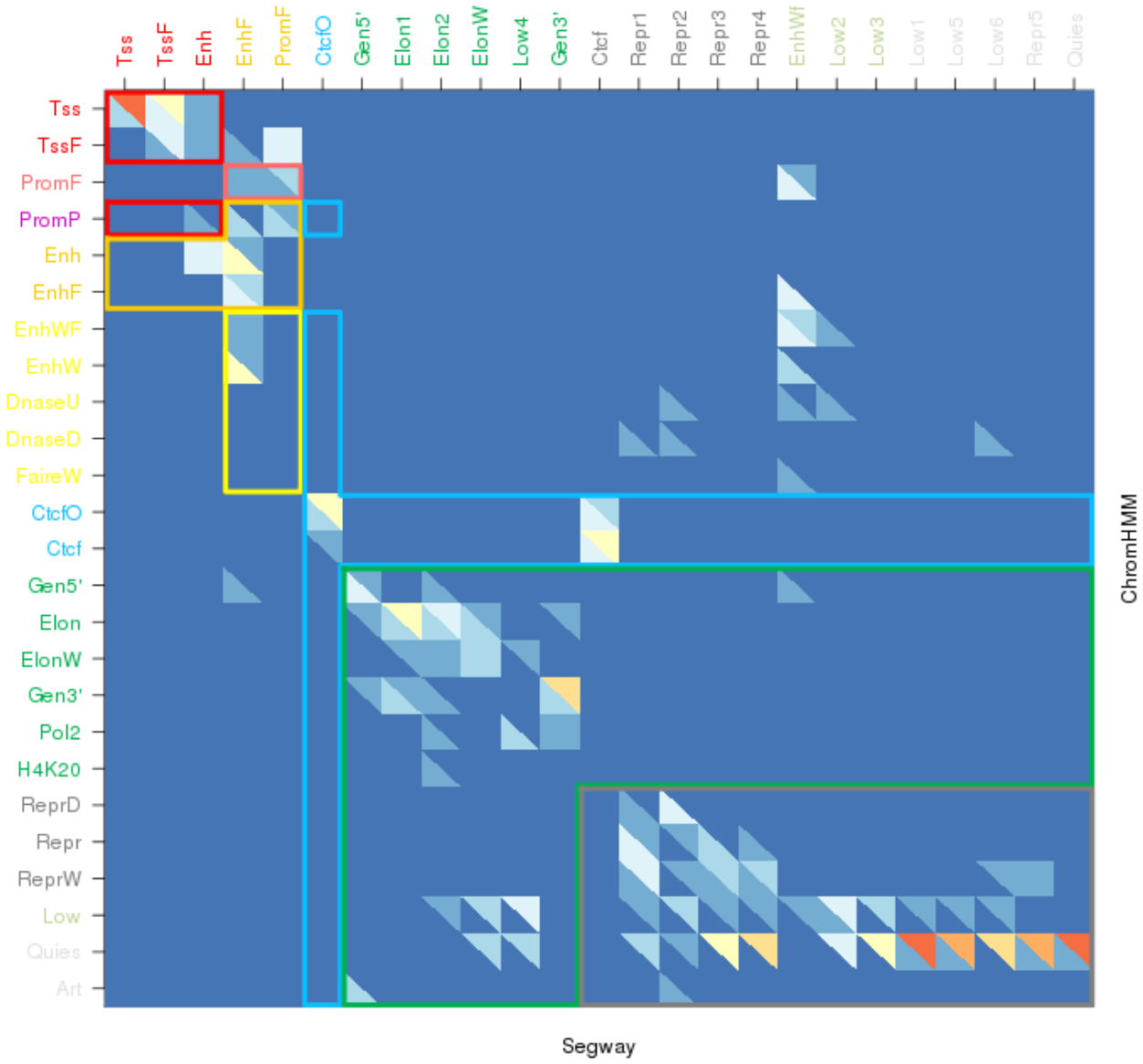


(A)

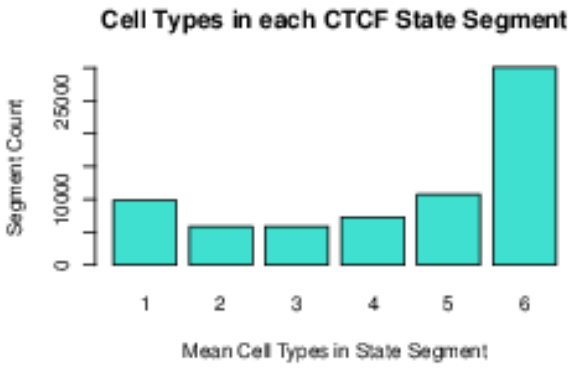


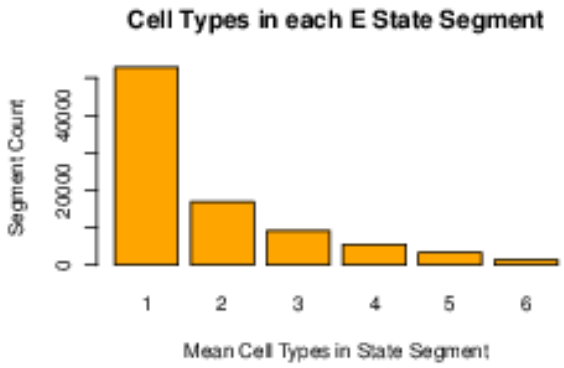
(B)

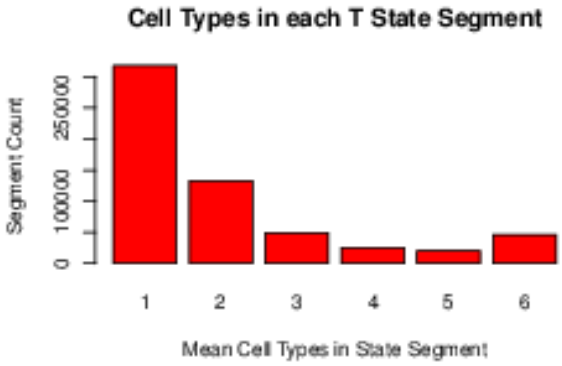
Supplementary Figure 9: Violin plots (60) showing the distribution of segment lengths on a logarithmic scale for each label, for (A) Segway and (B) ChromHMM GM12878 segmentations. Plots generated with Segtools (50). Black dots indicate the median segment length, dark green lines extend from the first to third quartiles, and dark green circles indicate outliers. The filled light green curve is a kernel density plot of each distribution.

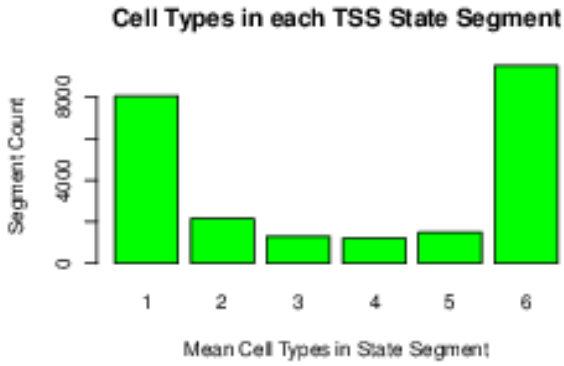


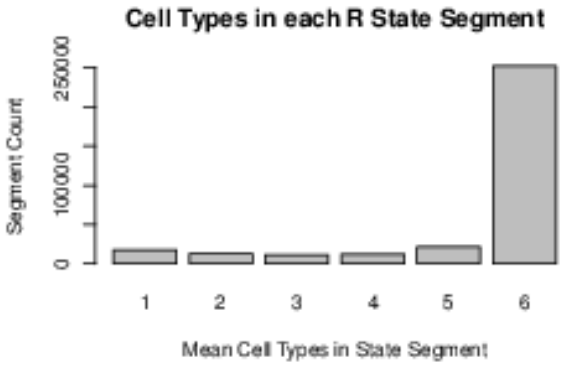
Supplementary Figure 10: Heat map of the pairwise segment overlap and combined segmentation classes for the ChromHMM and Segway GM12878 segmentations. The state mnemonics are listed along the axes, colored by their merged classes. For each pairwise combination of states, the lower-left triangle represents the proportion of segments of the ChromHMM state that overlap a segment of the Segway state by at least 20 base pairs, colored from blue (low) to red (high), and vice versa for the upper-right triangle. The colored boxes indicate the rules for deciding concordant bases in the combined segmentation: TSS (red), PF (light red), E (orange), WE (yellow), CTCF (blue), T (green) and R (gray).



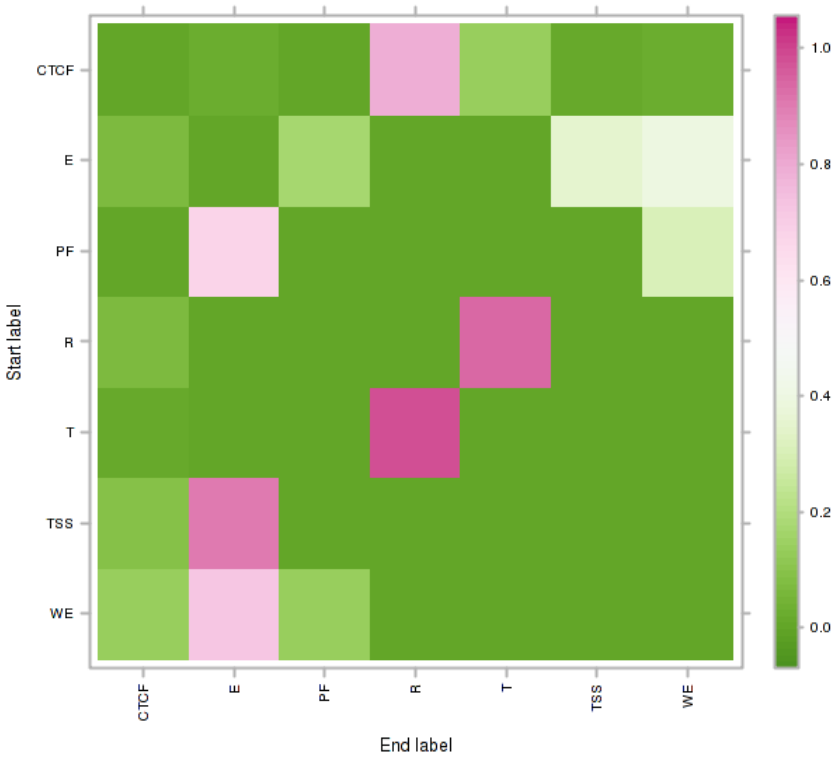




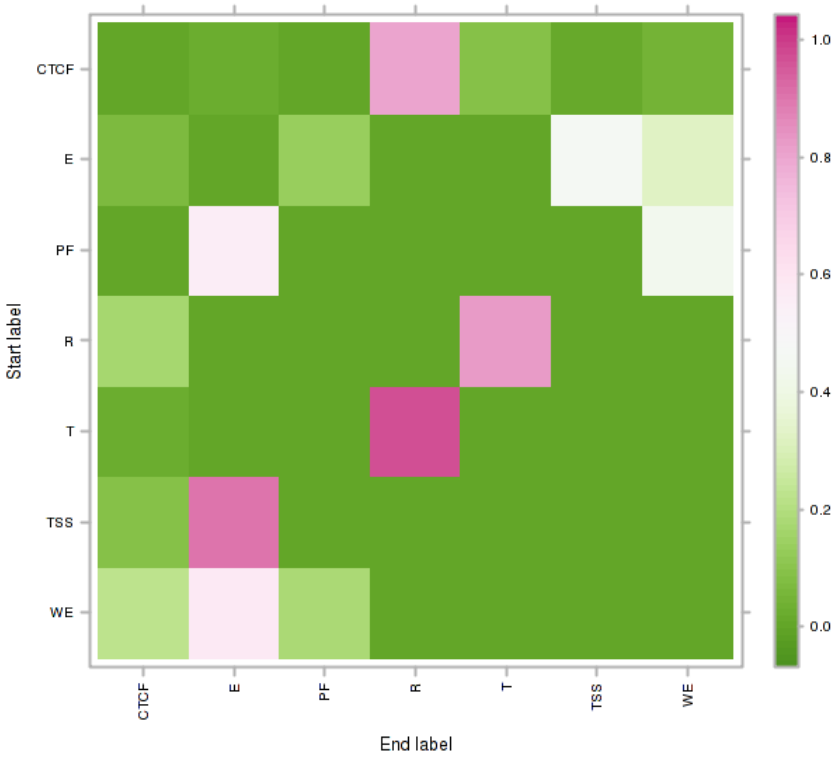




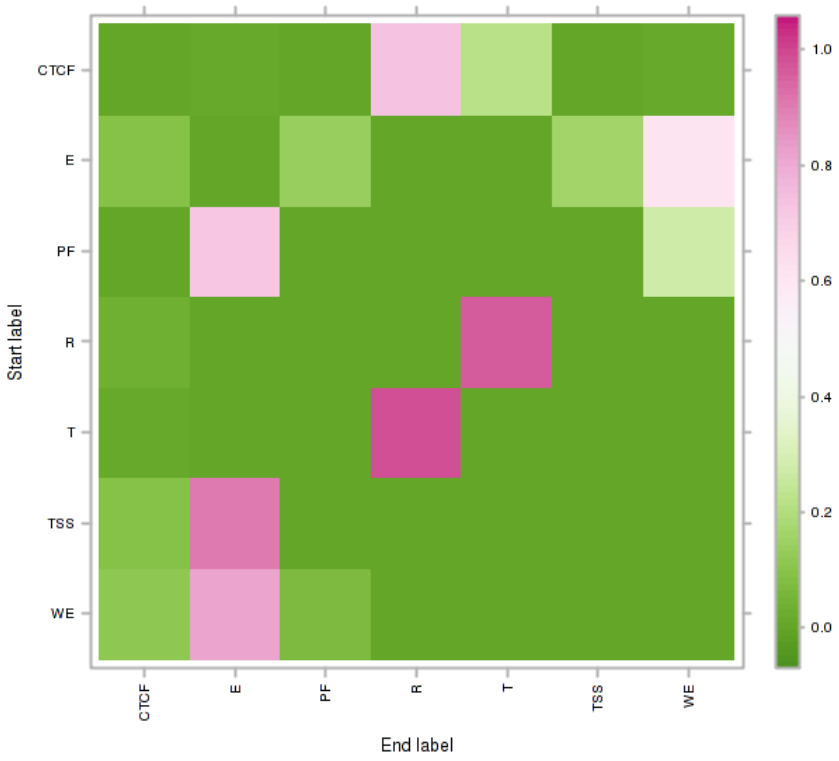
Supplementary Figure 11: State variability between cell types for the combined seven-state segmentation. The figure shows the distribution of occurrence of the state label at specific genome locations across each of the cell types from state labels that are unique to one cell type at one genome location (labeled “1”) to ubiquitous state labels that occur at the same location across all six cell lines for each of the five states (CTCF, E, T, TSS, and R; labeled “6”). For each state label, segments from each of the six cell lines were overlapped and clustered to give a bit string of state labels at each genomic region. Bit strings were then assessed by cell type. Data was then combined to give the mean occurrence of each cell count for each segment label.


CTCF
E
PF
R
T
TSS
WE
Start label
End label
1.0
0.8
0.6
0.4
0.2
0.0

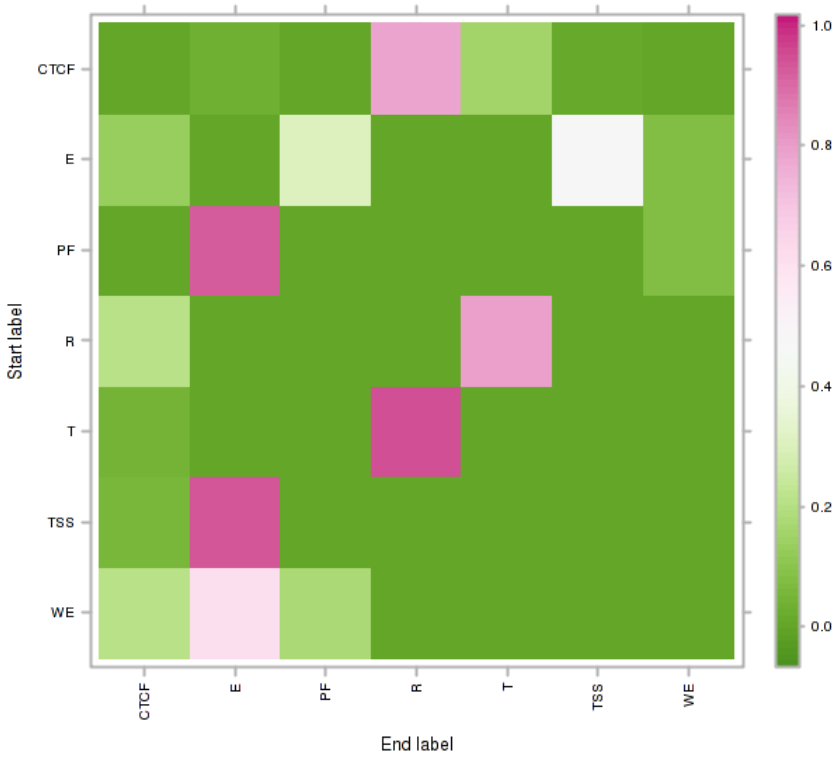
(A)


CTCF
E
PF
R
T
TSS
WE
Start label
End label
1.0
0.8
0.6
0.4
0.2
0.0

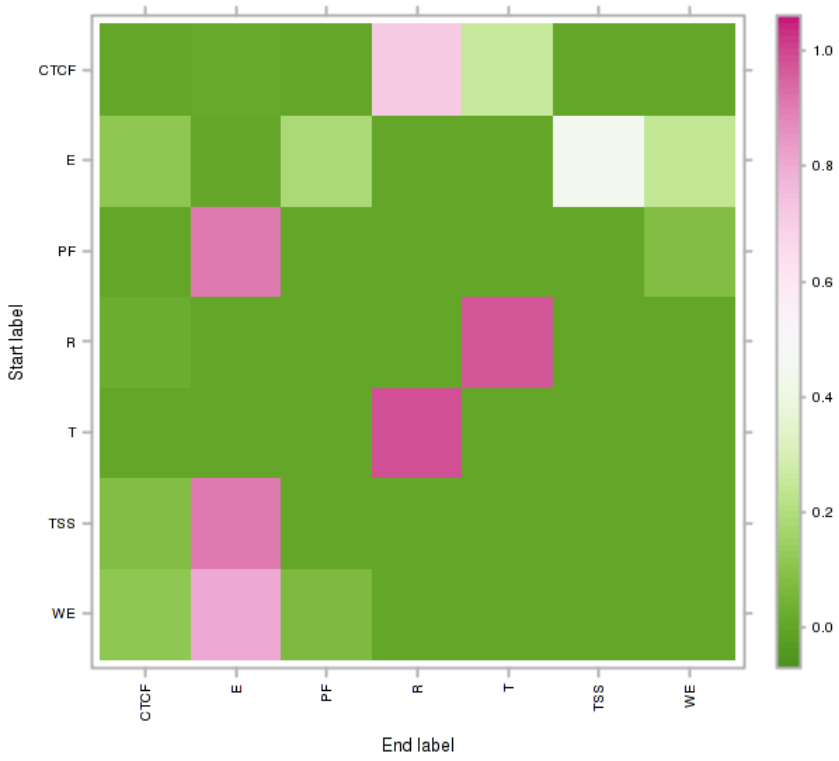
(B)


CTCF
E
PF
R
T
TSS
WE
Start label
End label
1.0
0.8
0.6
0.4
0.2
0.0

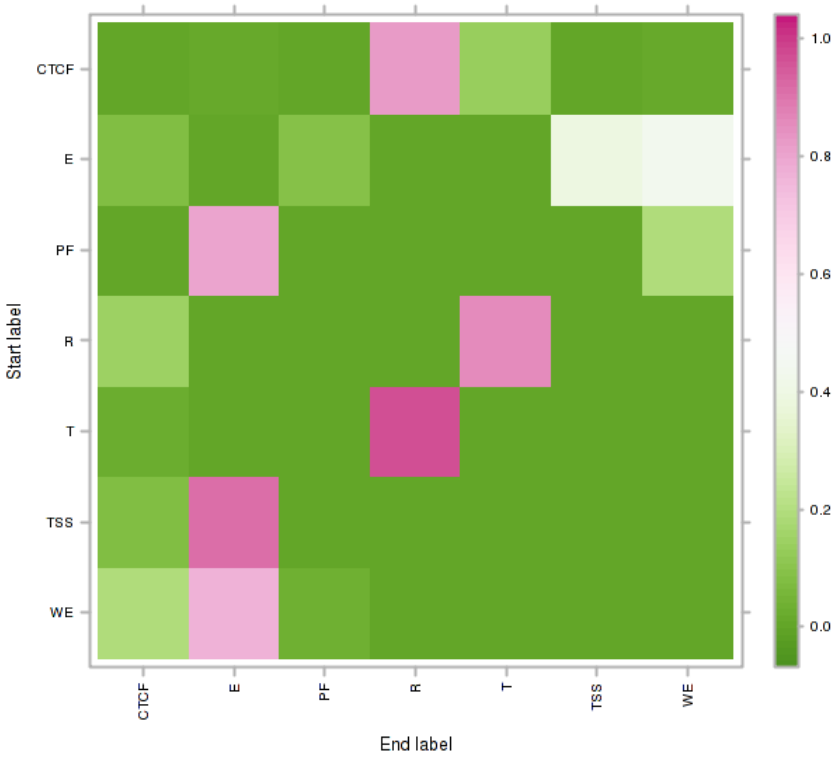
(C)


CTCF
E
PF
R
T
TSS
WE
Start label
End label
1.0
0.8
0.6
0.4
0.2
0.0

(D)

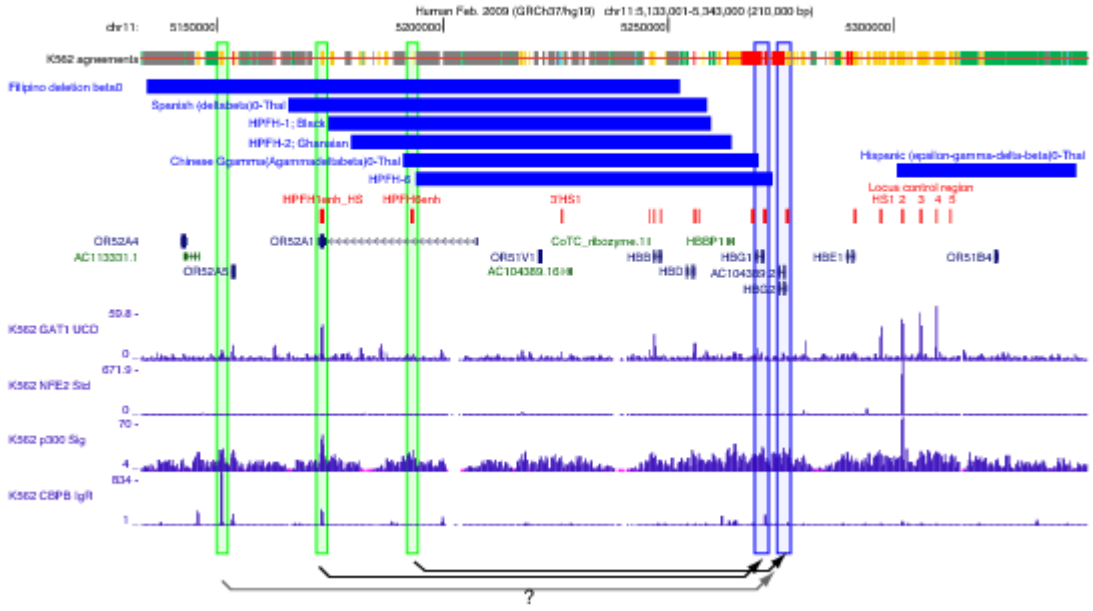


(E)

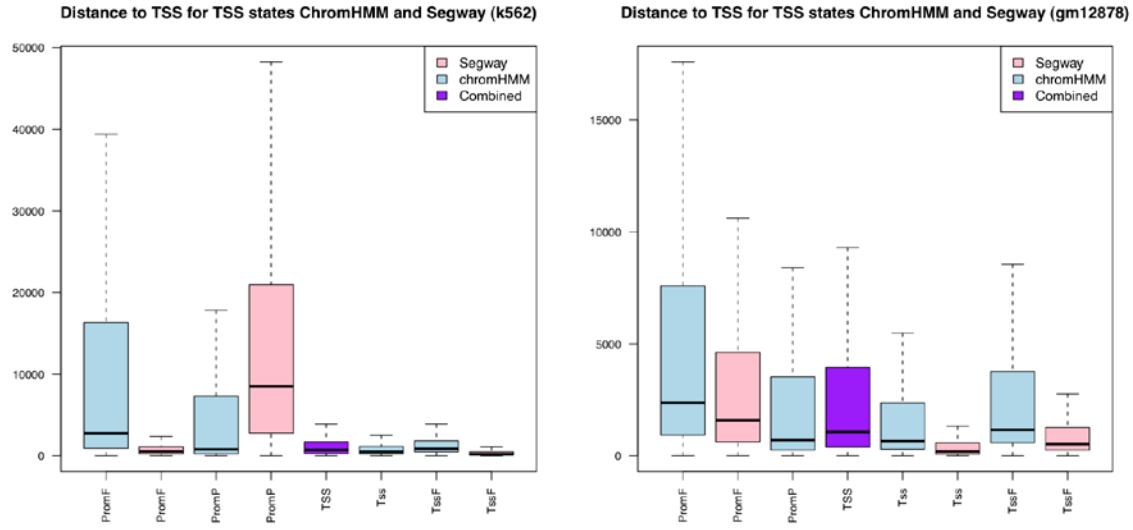


(F)

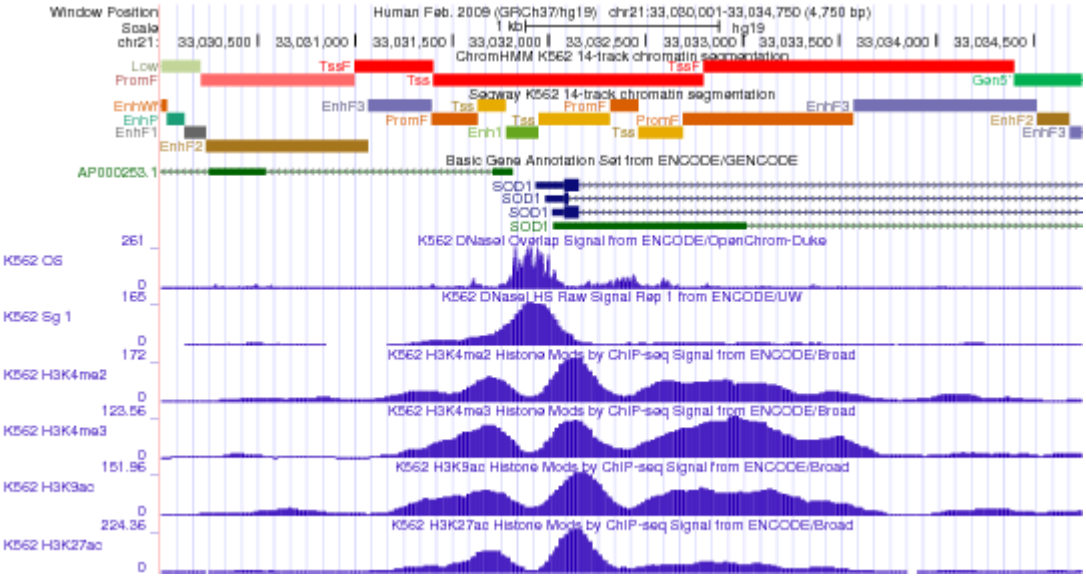
Supplementary Figure 12: Heat map of empirical transition frequencies for the combined (A) GM12878, (B) H1 hESC, (C) HeLa-S3, (D) HepG2, (E) HUVEC, (F) K562 segmentations. Each row represents a particular start label. The cells within each row represent the frequency with which segments will transition to an end label given that start label, with values represented by the colors in the color bar.



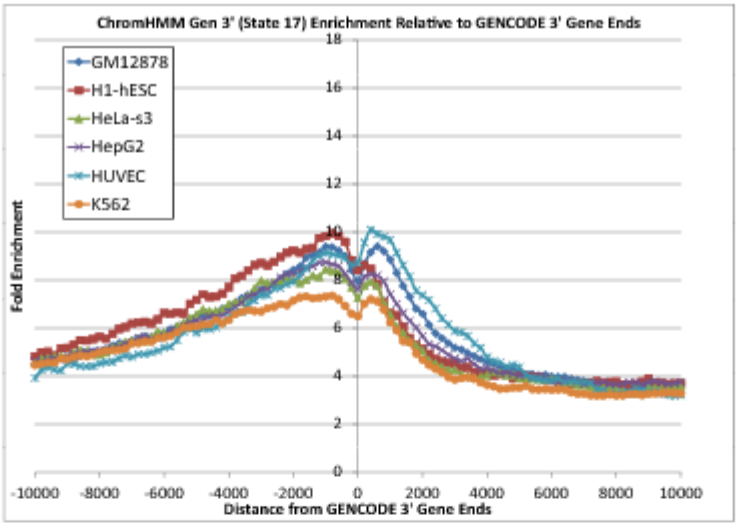
Supplementary Figure 13: Segmentations for functional interpretation in the *HBB* locus. The combined segmentation (top line) confirms the previous identification of distal regulatory regions to the left of the *HBB* genes in the diagram, and leads to the prediction of additional, more distal ones separated by regions of repressed chromatin. The long, blue horizontal rectangles show the extent of large deletions leading to phenotypes of beta-thalassemia (Thal) or hereditary persistence of fetal hemoglobin (HFPH; tracks are from the PhenCode compilation http://phencode.bx.psu.edu/ and Giardine et al. 2007). Red rectangles show the location of known *cis*-regulatory modules in the *HBB* locus (King et al. 2005). Genes are shown in the middle of the diagram, and below them are ChIP-seq signal tracks for occupancy by the transcription factors GATA1, NFE2, and CBPB along with the coactivator p300. Known and predicted enhancers to the left of *HBB* are outlined by vertical green rectangles, and the postulate that they target the *HBG1* and *HBG2* genes (outlined by blue vertical rectangles) is indicated by the arrows at the bottom.

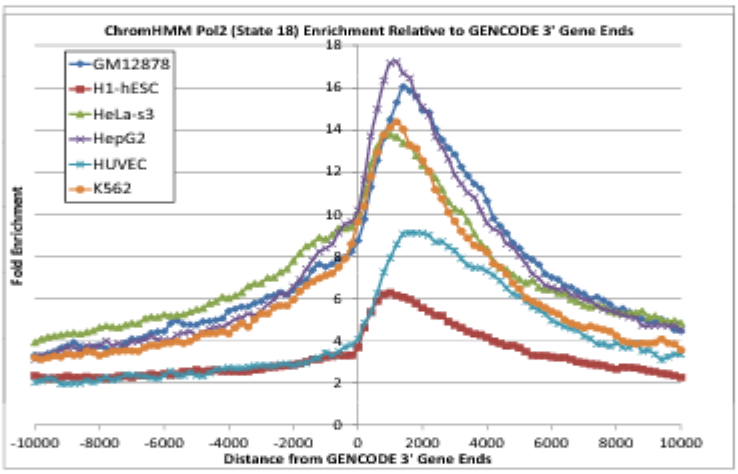


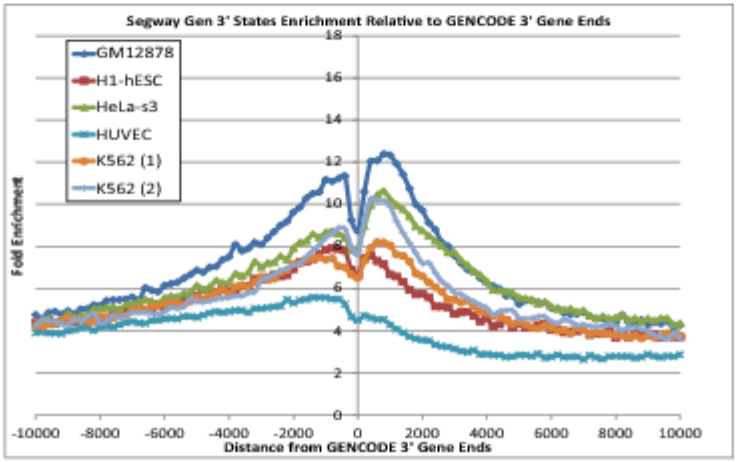
Supplementary Figure 14: Box plots of distance to the nearest TSS for selected relevant labels for ChromHMM (cyan), Segway (pink), and the combined segmentation (purple) for GM12878 and K562 cells.



Supplementary Figure 15: Promoter and flanking regions of the *SOD1* gene, displayed in the UCSC Genome Browser. The ChromHMM and Segway segmentations are shown, followed by GENCODE gene annotations, and signal tracks for DNaseI hypersensitivity, H3K4me3, H3K4me3, H3K9ac, and H3K27ac. The high-resolution Segway segmentation identifies a 171-bp Enh segment at a nucleosome-free open chromatin region, surrounded on either side by Tss and Tss flanking segments corresponding to modified well-positioned nucleosomes at the TSS. The high-continuity ChromHMM segmentation assigns this whole region to a 1400-bp Tss segment.







Supplementary Figure 16: This figure shows the fold enrichment of states relative to the 3' ends of GENCODE protein coding genes for (a) ChromHMM State 17 (Gen3') (b) ChromHMM State 18 (Pol2) (c) Segway states given the label or label prefix Gen3'. Distance relative to the 3' prime end is in base pairs.

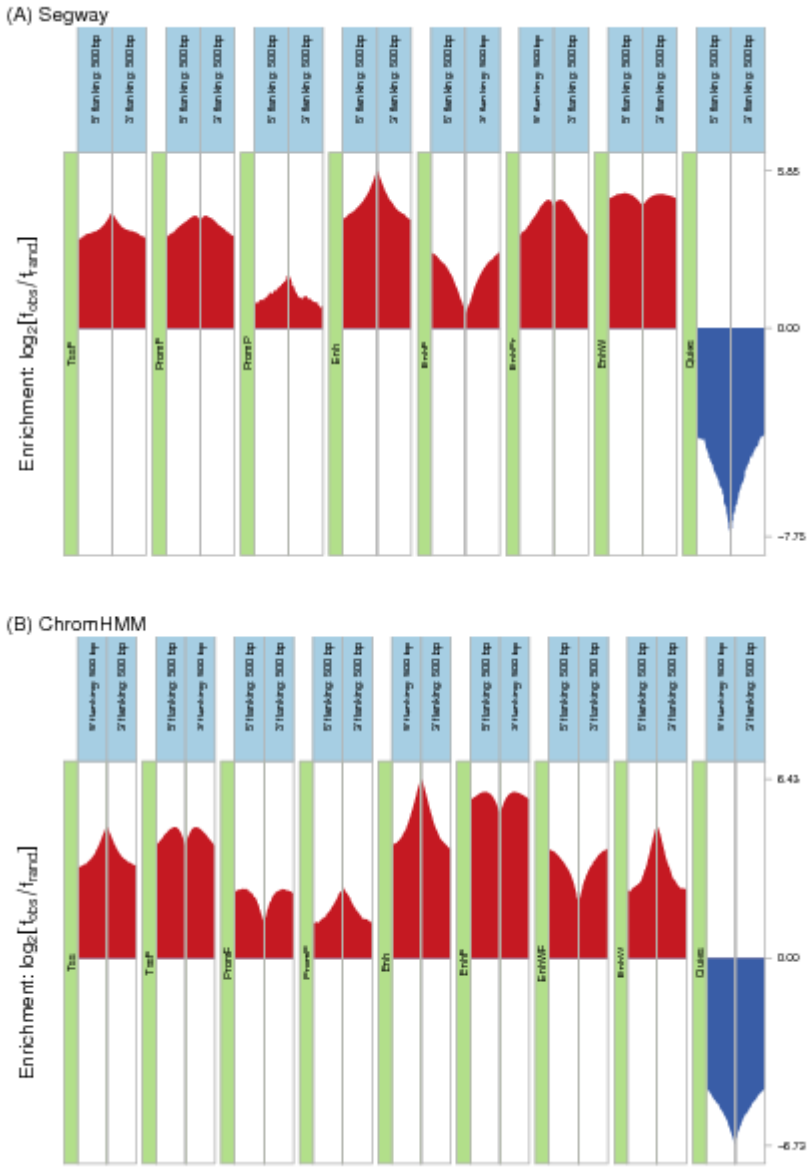
ChromHMM

State	GM12878	H1-hESC	HeLa-s3	HepG2	Huvec	K562
1_Tss	4.6	4.9	4.7	4.6	4.9	4.8
2_TssF	3.7	4.6	4.5	4.2	4.5	4.4
3_PromF	3.0	3.7	3.5	3.0	4.0	3.2
4_PromP	4.4	4.4	4.0	3.8	4.5	4.3
5_Enh	0.2	1.5	0.6	0.2	0.5	1.6
6_EnhF	0.0	0.2	0.8	0.2	0.2	1.3
7_EnhWF	0.3	-0.1	0.3	0.4	0.4	1.4
8_EnhW	0.6	1.6	0.3	0.7	0.8	1.6
9_DNaseU	0.0	-0.2	-0.6	0.9	-0.8	0.0
10_DNaseD	1.8	1.3	1.8	1.5	1.7	2.2
11_FaireW	0.6	-0.5	-0.3	-0.6	0.2	0.0
12_CtcfO	0.6	0.7	0.8	0.7	0.7	0.7
13_Ctcf	0.0	-0.6	0.2	-0.3	0.1	0.0
14_Gen5'	-0.2	1.0	1.0	0.2	0.7	1.2
15_Elon	-1.9	-1.6	-2.0	-1.9	-1.2	-1.7
16_ElonW	-1.8	-1.2	-1.8	-1.6	-1.3	-1.6
17_Gen3'	-0.4	-0.4	-0.7	-0.5	-0.4	-0.8
18_Pol2	1.5	0.5	1.3	1.6	0.9	0.8
19_H4K20	-0.6	-1.2	-0.7	-1.9	-2.3	-1.4
20_ReprD	3.7	3.2	0.4	2.1	3.5	2.0
21_Repr	1.8	2.1	-0.1	0.8	0.9	0.3
22_ReprW	0.1	1.1	-0.1	-0.1	0.0	-0.2
23_Low	0.0	-0.7	-0.4	-0.3	-0.4	-0.1
24_Quies	-1.8	-1.9	-1.7	-1.9	-2.0	-1.9
25_Art	0.4	0.9	0.4	1.2	0.9	1.0

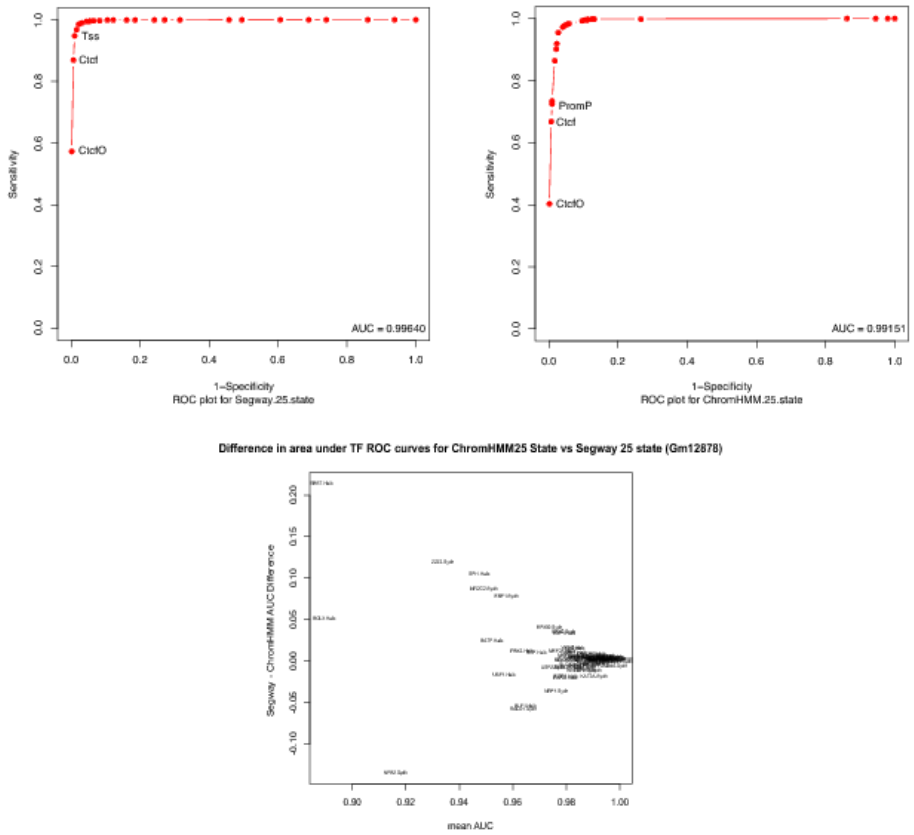
Segway

State	GM12878	H1-hESC	HeLa-s3	HepG2	Huvec	K562
Tss	4.9	4.7	4.9	4.6	5.0	5.0
TssF	4.6	4.6				
PromF	3.7	2.5	4.6	4.6	4.5	4.7
PromP		4.1				
PromP1				3.6		
PromP2				4.9		
Enh	1.8			1.5	2.0	
Enh1			2.9			3.6
Enh2			0.3			1.3
EnhF	1.5		1.1			
EnhF1				0.9	0.8	0.7
EnhF2				-0.1	2.1	2.1
EnhF3						3.5
EnhPr		2.3	2.9	3.4		
EnhP						1.6
EnhW				0.1		
EnhW1					1.7	
EnhW2					0.8	
EnhWf	1.0					1.2
EnhWF1		0.9				
EnhWF2		-0.5				
EnhWF3		-0.7				
DnaseD		1.3				
Faire		-1.2	-1.0		-1.6	
CtcfO	2.0	1.4	0.9	0.7	0.9	1.0
Ctcf	0.2					
Gen5'	0.9		0.5		0.6	
Gen5'1				-0.2		
Gen5'2				-0.5		
Elon		-1.1	-1.4	-2.0		-1.1
Elon1	-1.5				-1.5	
Elon2	-1.1				-1.6	
ElonW	-1.9			-2.0	-1.2	-1.5
ElonW1		-1.7	-2.2			
ElonW2		-1.9	-0.6			
ElonW3		-0.3				
Gen3'	0.3	0.5	0.9		-0.3	
Gen3'1						0.3
Gen3'2						0.8
Repr		3.3				
Repr1	1.0		-0.6	1.4	2.5	0.6
Repr2	2.0		0.2	0.5	3.7	1.5
Repr3	-0.3		0.5	-0.2	0.5	-0.1
Repr4	-0.7		-1.1	2.2	0.1	-0.4
Repr5	-1.2			-0.8	-0.9	
Repr6				-1.0	-0.6	
Repr7					-1.1	
Low1	-1.9	-1.7	-1.5	-1.8	-1.5	-1.1
Low2	-0.3	-1.3	-0.9	-1.2	-2.0	-2.1
Low3	-0.6	0.1	-2.0	-2.2	-0.3	-0.5
Low4	0.6	0.4	-1.1	0.1		-0.6
Low5	-1.7	-0.1	-0.5			-1.0
Low6	-0.1	-0.8	1.2			-1.8
Low7		-1.7	1.0			
Quies	-1.8	-1.1	-0.8	-1.8	-1.4	-1.7

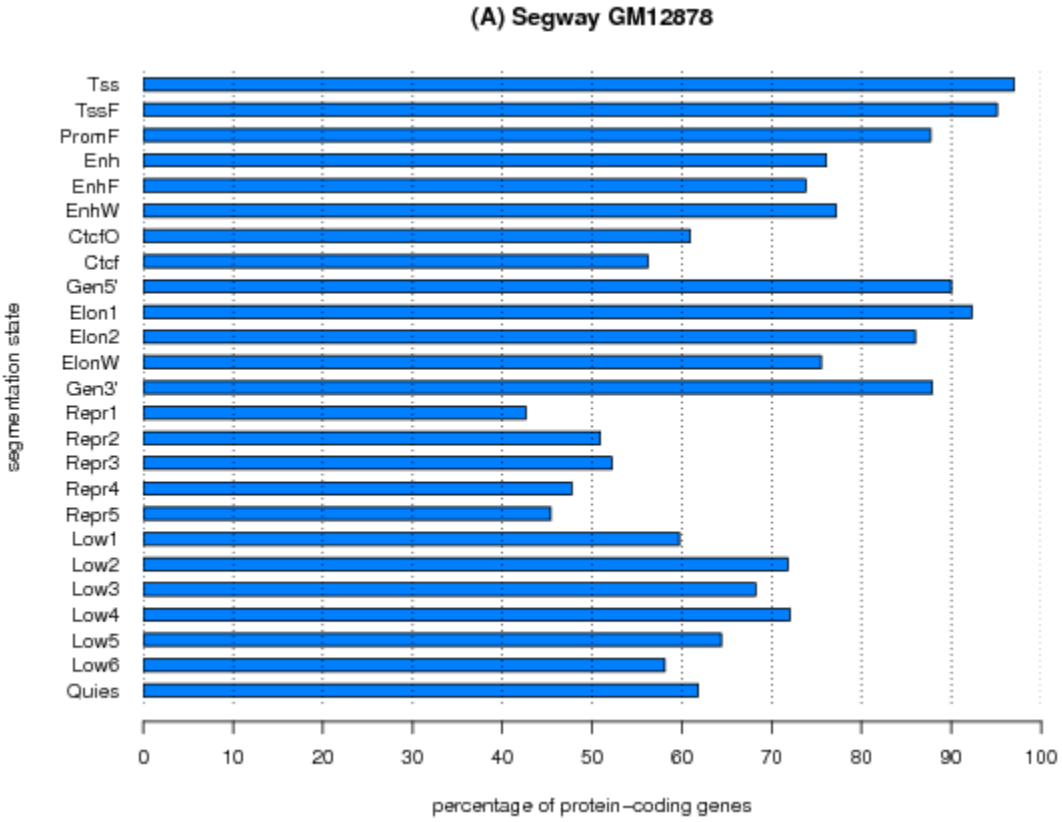
Supplementary Figure 17: ChromHMM and Segway state enrichments in each cell type for regions proximal to gene starts. These regions are those bases in the genome within 2000 bp of the 5' end of a GENCODE gene. Enrichment is defined as $\log_2$ of the ratio of the fraction of the state in the cell type that overlaps a gene start proximal region divided by the fraction of chromosomes 1-22 and X that overlap a gene start proximal region, which was 2.48%.



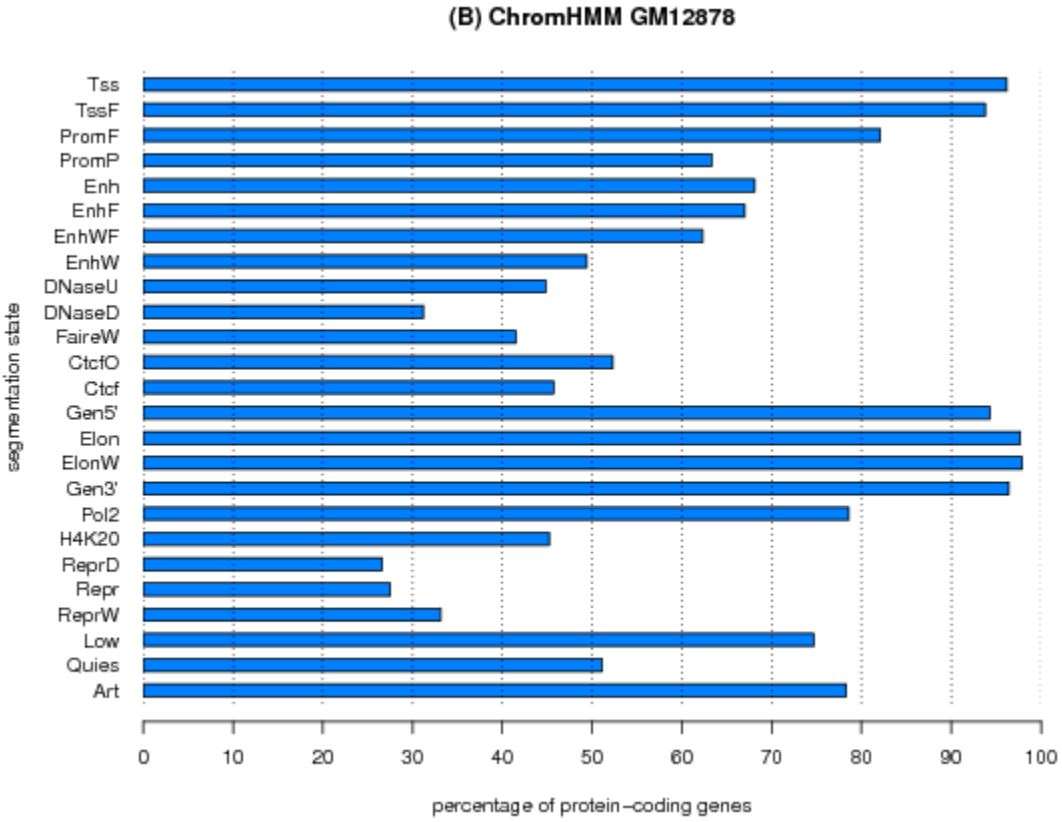
Supplementary Figure 18: Enrichment of promoter-associated, enhancer-associated, and quiescent labels in the 500 bp upstream and downstream of the center of ENCODE ChIP-seq peaks for transcriptional coactivator p300 in HepG2, for both (A) Segway and (B) ChromHMM. Each panel shows the mean $\log_2$ fold enrichment for a position relative to all p300 peaks against the position's location relative to the peak center. The largest p300 enrichment is in the Enh labels, and there is also substantial enrichment in the promoter-associated labels. The p300 peaks are highly depleted for the Quies labels.



Supplementary Figure 19: The two plots at the top are receiver operating characteristic (ROC) curves for Segway (left) and ChromHMM (right) for the SPP peak calls for the ChIP-seq data for CTCF (UW) on Gm12878 cells. The first three most specific state labels are labeled by their mnemonic. The bottom plot is the difference in area under the curve (AUC, Segway AUC minus ChromHMM AUC) for the 25 state segmentations for all Gm12878 TF ChIP-seq SPP peak datasets. The TF data is labeled by the factor detected (using the HGNC name of the gene that encodes the factor) and laboratory that generated the data.



(A)



(B)

Supplementary Figure 20: State-by-state percentage of protein coding genes supported by RNA-seq expression. Bars show, for (A) Segway or (B) ChromHMM, the percentage of protein-coding genes that also overlap a same-cell-line protein-coding RNA-seq contig for GM12878.

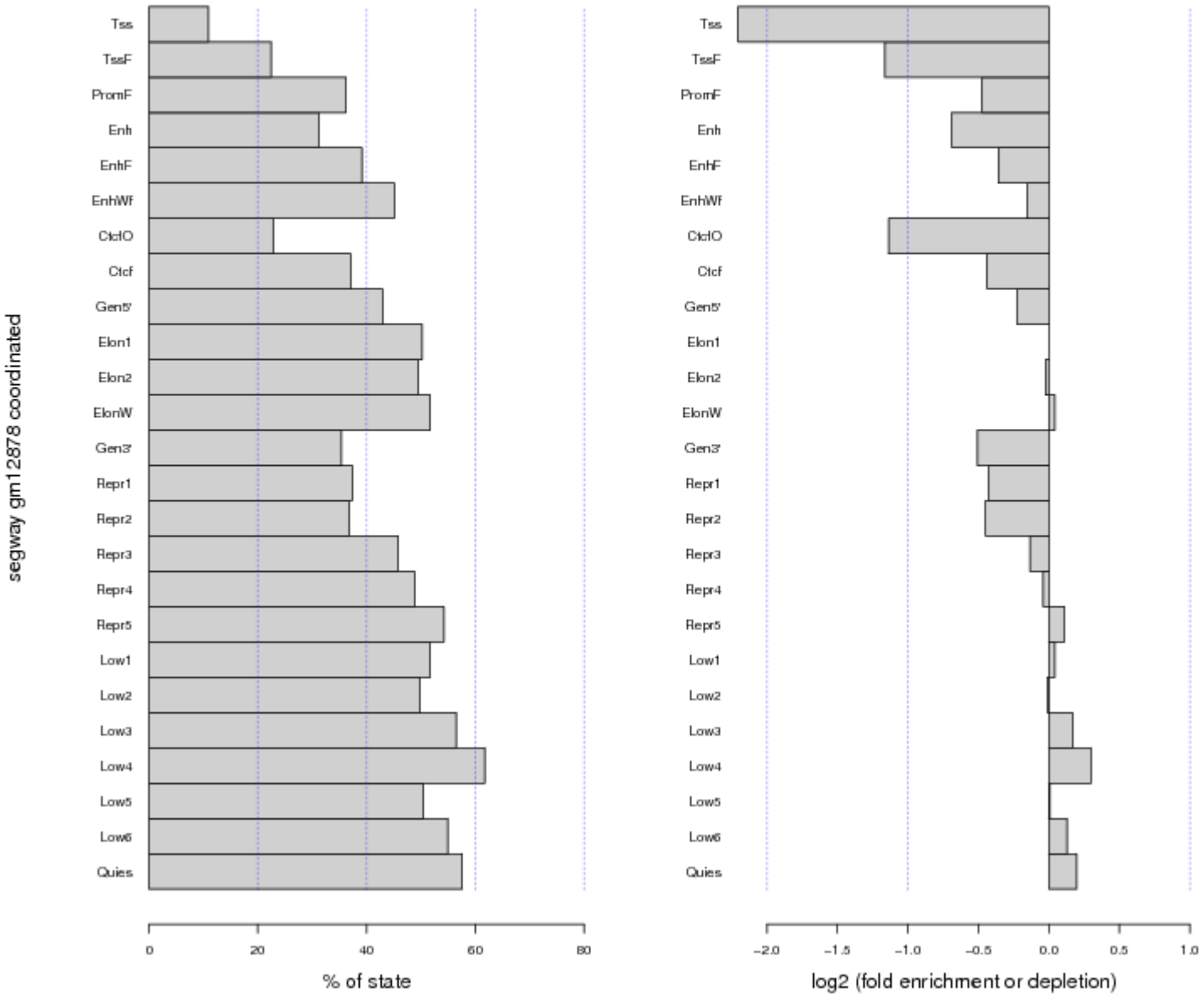
ChromHMM

State	GM12878	H1-hESC	HeLa-s3	HepG2	HUVEC	K562
1_Tss	0.2	0.2	0.2	0.2	0.2	0.2
2_TssF	0.3	0.3	0.2	0.2	0.2	0.3
3_PromF	0.4	0.7	0.6	0.5	0.4	0.8
4_PromP	0.4	0.5	0.4	0.3	0.5	0.4
5_Enh	0.4	0.5	0.3	0.4	0.3	0.4
6_EnhF	0.5	0.7	0.4	0.5	0.4	0.5
7_EnhWF	0.6	0.9	0.5	0.5	0.5	0.4
8_EnhW	0.6	0.5	0.5	0.5	0.5	0.3
9_DNaseU	0.8	0.9	0.9	0.8	0.8	0.7
10_DNaseD	0.4	0.2	0.3	0.5	0.2	0.5
11_FaireW	0.8	0.7	0.7	0.8	0.6	0.8
12_CtcfO	0.6	0.6	0.5	0.6	0.6	0.6
13_Ctcf	1.0	1.0	0.9	1.0	0.9	1.0
14_Gen5'	0.2	0.2	0.1	0.2	0.1	0.1
15_Elon	0.3	0.3	0.2	0.3	0.2	0.3
16_ElonW	0.4	0.3	0.3	0.4	0.3	0.4
17_Gen3'	0.1	0.2	0.1	0.1	0.1	0.2
18_Pol2	0.2	0.5	0.1	0.2	0.3	0.2
19_H4K20	0.4	0.3	0.9	0.4	0.2	0.3
20_ReprD	0.5	0.5	1.0	0.4	0.5	0.4
21_Repr	0.9	0.9	1.4	0.9	0.8	0.8
22_ReprW	1.0	0.8	1.5	1.0	1.0	1.0
23_Low	0.6	0.9	1.0	0.7	0.6	0.6
24_Quies	1.3	1.2	1.2	1.3	1.4	1.3
25_Art	0.2	0.2	0.1	0.2	0.2	0.5

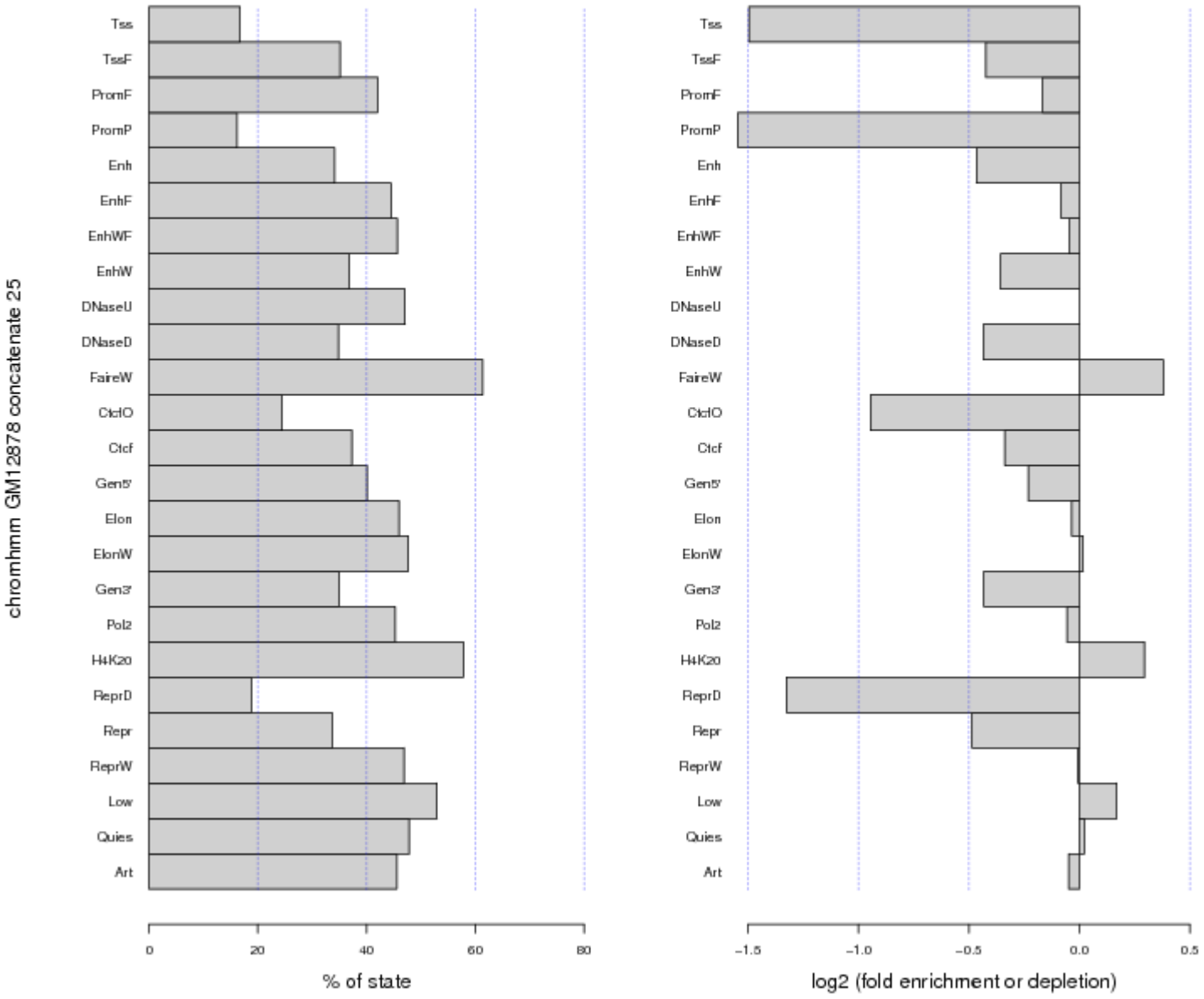
Segway

State	GM12878	H1-hESC	HeLa-s3	HepG2	HUVEC	K562
Tss	0.2	0.3	0.2	0.2	0.2	0.2
TssF	0.2	0.3				
PromF	0.3	0.7	0.2	0.2	0.3	0.3
PromP		0.6				
PromP1				0.4		
PromP2				0.2		
Enh	0.3			0.4	0.2	
Enh1			0.3			0.3
Enh2			0.6			0.4
EnhF	0.5		0.6			
EnhF1				0.4	0.3	0.6
EnhF2				0.5	0.4	0.4
EnhF3						0.3
EnhPr		0.4	0.3	0.3		
EnhP						0.6
EnhW				0.8		
EnhW1					0.4	
EnhW2					0.6	
EnhWf	0.6					0.4
EnhWf1		0.7				
EnhWf2		0.9				
EnhWf3		1.2				
DnaseD		0.3				
Faire		1.1	1.0		1.2	
CtcfO	0.6	0.7	0.6	0.7	0.7	0.6
Ctcf	0.9					
Gen5'	0.2		0.1		0.1	
Gen5'1				0.2		
Gen5'2				0.2		
Elon		0.3	0.3	0.4		0.3
Elon1	0.3				0.2	
Elon2	0.3				0.6	
ElonW	0.6			0.5	0.4	0.5
ElonW1		1.1	0.5			
ElonW2		1.2	0.4			
ElonW3		0.8				
Gen3'	0.2	0.2	0.1		0.2	
Gen3'1						0.2
Gen3'2						0.2
Repr		0.6				
Repr1	0.9		1.5	0.7	0.6	0.7
Repr2	0.7		1.2	0.8	0.7	0.6
Repr3	1.2		0.8	1.1	0.9	1.0
Repr4	1.2		1.6	0.5	1.0	1.0
Repr5	1.4			1.3	1.3	
Repr6				1.3	1.1	
Repr7					1.4	
Low1	1.4	1.4	1.3	1.3	1.4	1.0
Low2	0.9	1.2	0.8	1.2	1.4	1.4
Low3	0.9	0.8	1.4	1.5	1.2	1.2
Low4	0.5	1.1	1.3	1.0		1.3
Low5	1.4	1.0	1.1			1.3
Low6	0.9	1.1	0.5			1.3
Low7		1.3	0.9			
Quies	1.4	1.3	1.3	1.4	1.5	1.5

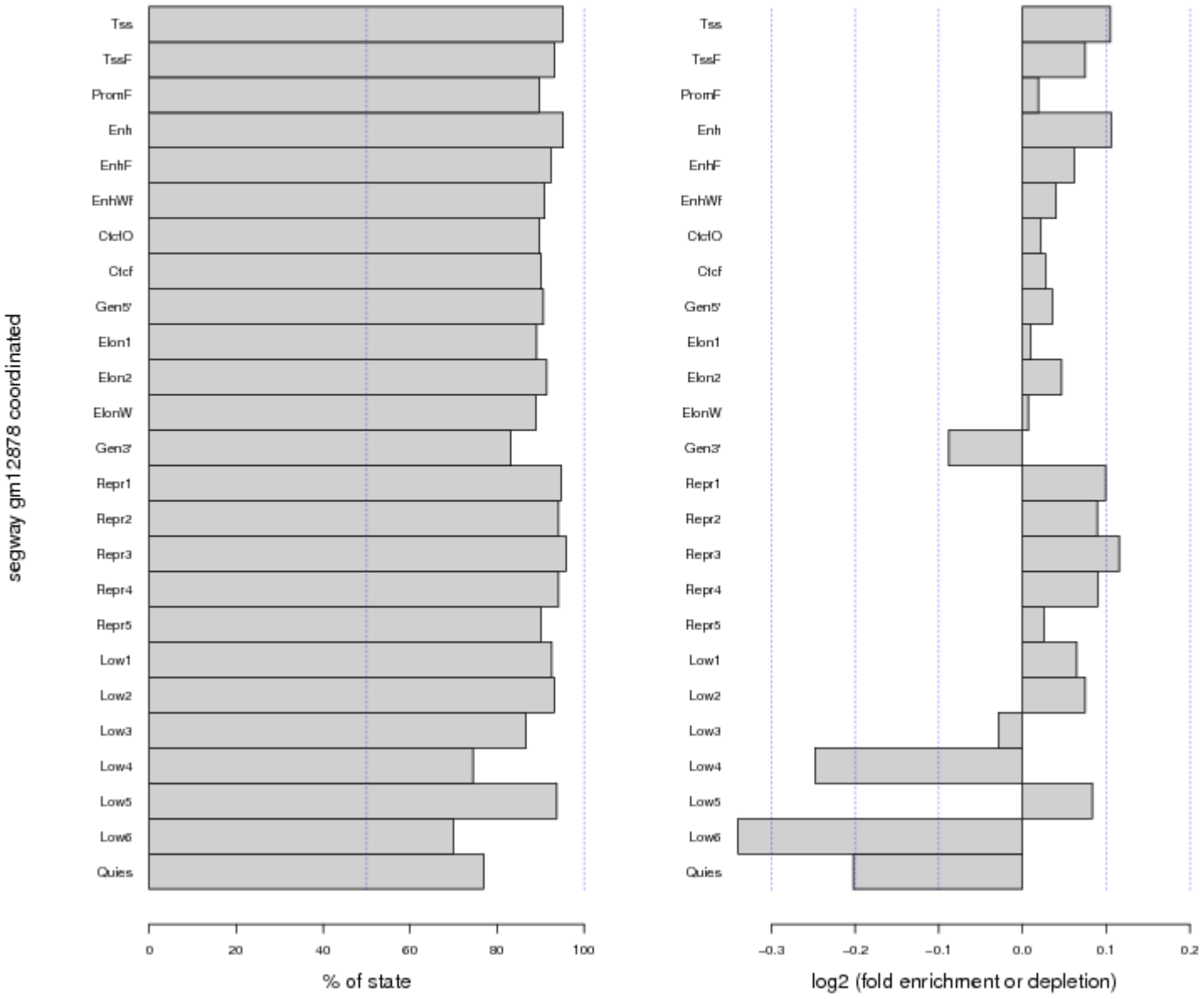
Supplementary Figure 21: Fold enrichment for nuclear lamina associated domains (Guelen et al. 2008) for all states and cell types for both ChromHMM and Segway.

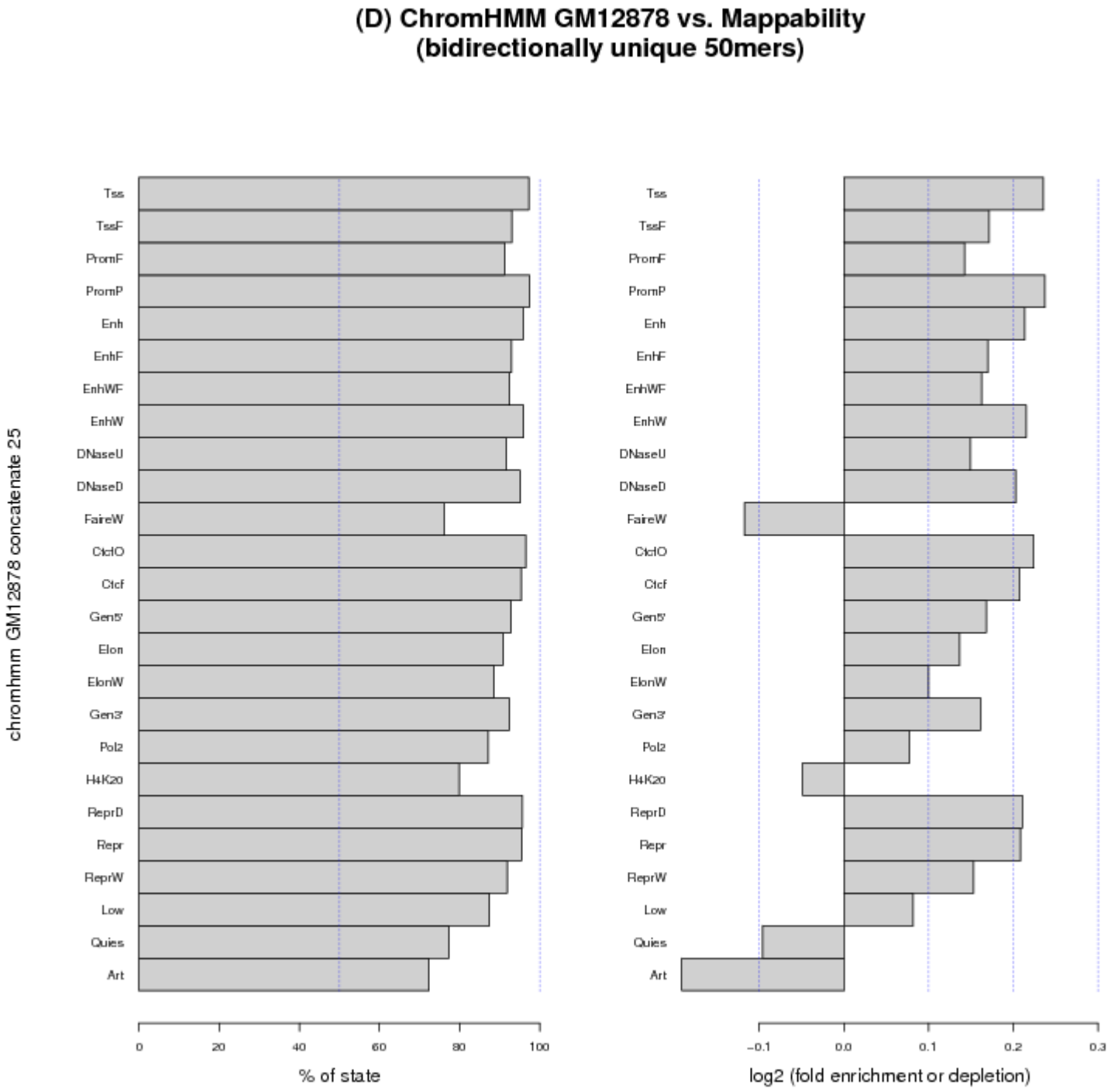

(A) Segway GM12878 vs. RepeatMasker
(all annotated repeats)
segway gm12878 coordinated
Tss
TssF
PromF
Enh
EnhF
EnhWf
CtcfO
Ctcf
Gen5'
Elon1
Elon2
ElonW
Gen3'
Repr1
Repr2
Repr3
Repr4
Repr5
Low1
Low2
Low3
Low4
Low5
Low6
Quies
0
20
40
60
80
% of state
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
log2 (fold enrichment or depletion)

(B) ChromHMM GM12878 vs. RepeatMasker (all annotated repeats)



(C) Segway GM12878 vs. Mappability (bidirectionally unique 50mers)





Supplementary Figure 22: Portion of states covered and enrichment or depletion of (A, B) RepeatMasker repeats or (C, D) mappable regions in each label for two 25-label segmentations on GM12878 data. Panels (A, C) show Segway data, and (B,D) show ChromHMM data. In each panel, the left plot shows the percentage of bases in each state occupied by the corresponding feature (repeat or mappable region). The right plot shows enrichment (positive) or depletion (negative) of the feature, measured as $\log_2(o/e)$, where o is the number of bases observed and e is the number expected (see Supplementary Methods).

ChromHMM

State	GM12878	H1-hESC	HeLa-s3	HepG2	HUVEC	K562
1_Tss	2.0	2.4	2.1	2.1	2.3	2.1
2_TssF	0.9	1.3	1.1	0.9	1.3	1.0
3_PromF	0.5	1.1	0.7	0.4	0.7	0.5
4_PromP	2.1	2.1	1.7	1.6	2.1	1.9
5_Enh	0.9	1.6	1.2	1.2	1.6	1.0
6_EnhF	0.3	0.9	0.3	0.5	0.4	0.2
7_EnhWF	0.3	1.1	0.6	0.3	0.6	0.3
8_EnhW	0.9	1.4	1.2	0.9	1.3	0.9
9_DNaseU	0.7	0.5	0.9	0.6	1.5	0.7
10_DNaseD	0.7	0.7	0.9	0.9	0.6	1.0
11_FaireW	-0.5	0.0	-2.0	-0.7	-0.9	-0.4
12_CtcfO	1.2	1.0	1.1	1.2	1.2	1.1
13_Ctcf	0.6	0.5	0.5	0.5	0.4	0.5
14_Gen5'	0.8	1.2	1.0	0.8	1.1	0.9
15_Elon	0.9	1.5	1.3	0.9	1.1	1.0
16_ElonW	0.3	0.5	0.5	0.3	0.3	0.3
17_Gen3'	1.3	1.6	1.4	1.3	1.3	1.2
18_Pol2	0.4	-0.5	0.8	0.4	-0.1	0.3
19_H4K20	-0.7	0.3	0.2	-0.2	-0.4	0.1
20_ReprD	1.8	1.4	0.1	0.8	1.6	0.9
21_Repr	0.9	0.5	0.6	0.4	0.4	0.3
22_ReprW	0.4	0.0	0.3	0.0	-0.1	0.1
23_Low	-0.1	-0.1	0.0	-0.1	-0.1	-0.1
24_Quies	-0.5	-0.6	-0.7	-0.4	-0.5	-0.4
25_Art	0.0	0.4	0.3	0.9	0.7	0.2

Segway

State	GM12878	H1-hESC	HeLa-s3	HepG2	HUVEC	K562
Tss	2.6	2.5	2.3	2.5	2.4	2.3
TssF	1.5	1.8				
PromF	0.9	1.2	1.3	1.2	1.4	1.4
PromP		2.0				
PromP1				1.5		
PromP2				2.2		
Enh	1.1			1.2	1.6	
Enh1			1.9			1.9
Enh2			1.4			0.9
EnhF	0.7		0.5			
EnhF1				0.6	0.9	0.2
EnhF2				0.3	2.1	0.4
EnhF3						0.8
EnhPr		1.2	0.6	0.7		
EnhP						0.6
EnhW				-0.1		
EnhW1					0.8	
EnhW2					0.4	
EnhWf	0.3					0.0
EnhWf1		0.6				
EnhWf2		0.8				
EnhWf3		0.6				
DnaseD		0.4				
Faire		0.2	-2.1		-0.2	
CtcfO	1.5	1.2	1.1	0.9	1.1	1.1
Ctcf	0.6					
Gen5'	0.8		1.1		0.9	
Gen5'1				1.0		
Gen5'2				1.0		
Elon		1.2	0.9	0.8		0.8
Elon1	0.8				1.2	
Elon2	0.5				0.1	
ElonW	0.2			0.2	0.1	0.4
ElonW1		0.3	0.6			
ElonW2		-0.2	-0.1			
ElonW3		-2.1				
Gen3'	1.3	1.3	1.0		1.0	
Gen3'1						0.9
Gen3'2						1.1
Repr		1.5				
Repr1	0.5		0.5	0.5	1.0	0.4
Repr2	0.9		0.3	0.1	1.6	0.6
Repr3	0.4		0.4	0.3	0.2	0.2
Repr4	0.3		0.1	0.6	0.3	0.3
Repr5	0.0			-0.1	0.1	
Repr6				0.2	0.0	
Repr7					-0.2	
Low1	-0.3	-0.5	0.1	0.0	-0.4	0.2
Low2	0.1	0.1	0.2	0.1	-0.2	-0.2
Low3	-0.1	0.2	-0.3	-0.3	-0.1	-0.2
Low4	-0.2	0.0	0.0	-0.3		0.0
Low5	-0.2	-0.3	0.0			0.0
Low6	-0.5	-0.7	0.0			-0.1
Low7		-0.2	-0.1			
Quies	-1.0	-1.6	-1.2	-0.9	-1.1	-0.7

Supplementary Figure 23: ChromHMM and Segway state enrichments in each cell type for constrained elements as defined by the SiPhy-π measure (61,62). Enrichment is defined as $\log_2$ of the ratio of the fraction of the state in the cell type that overlaps a constrained element divided by the fraction of chromosomes 1-22 and X that overlap a constrained element, which was 5.82%. The SiPhy-π elements were originally defined based on the NCBI36/hg18 assembly, but for this analysis we used a version lifted over to GRCh37/hg19.

Supplementary references

44. Kharchenko, P.V., Tolstorukov, M.Y. and Park, P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol., 26, 1351-1359.
45. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol, 9, R137.
46. Kundaje, A., Jung, Y.L., Kharchenko, P.V., Wold, B.J., Sidow, A., Batzoglou, S. and Park, P.J. (2012) Adaptive calibrated measures for rapid automated quality control of massive collections of ChIP-seq experiments. (In preparation).
47. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K. et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell, 125, 315-326.
48. Kundaje, A., Li, Q., Rozowsky, J., Brown, J.B., Harmanci, A., Wilder, S., Gerstein, M., Batzoglou, S., Sidow, A., Birney, E. et al. (2012) Reproducibility measures for automatic threshold selection and quality control in ChIP-seq datasets. (In preparation).
49. Smit, A.F.A., Hubley, R. and Green, P. (1996).
50. Buske, O.J., Hoffman, M.M., Ponts, N., Le Roch, K.G. and Noble, W.S. (2011) Exploratory analysis of genomic segmentations with Segtools. BMC Bioinformatics, 12, 415.
51. Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D. et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol, 7, S4.
52. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57-74.
53. Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R. et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature, 489, 91-100.
54. King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W. and Hardison, R.C. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res., 15, 1051-1060.
55. Tuan, D., Feingold, E., Newman, M., Weissman, S.M. and Forget, B.G. (1983) Different 3' end points of deletions causing delta beta-thalassemia and hereditary persistence of fetal hemoglobin: implications for the control of gamma-globin gene expression in man. Proc. Natl. Acad. Sci. U. S. A., 80, 6937-6941.
56. Feingold, E.A. and Forget, B.G. (1989) The breakpoint of a large deletion causing hereditary persistence of fetal hemoglobin occurs within an erythroid DNA domain remote from the beta-globin gene cluster. Blood, 74, 2178-2186.
57. Anagnou, N.P., Perez-Stable, C., Gelinas, R., Costantini, F., Liapaki, K., Constantopoulou, M., Kosteas, T., Moschonas, N.K. and Stamatoyannopoulos, G. (1995) Sequences located 3' to the breakpoint of the hereditary persistence of fetal hemoglobin-3 deletion exhibit enhancer activity and can modify the developmental expression of the human fetal A gamma-globin gene in transgenic mice. J. Biol. Chem., 270, 10256-10263.
58. Feingold, E.A., Penny, L.A., Nienhuis, A.W. and Forget, B.G. (1999) An olfactory receptor gene is located in the extended human beta-globin gene cluster and is expressed in erythroid cells. Genomics, 61, 15-23.
59. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y. et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res., 22, 1798-1812.
60. Hintze, J.L. and Nelson, R.D. (1998) Violin plots: A box plot-density trace synergism. Am. Stat., 52, 181-184.
61. Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N. and Xie, X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics, 25, i54-62.
62. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E. et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature, 478, 476-482.