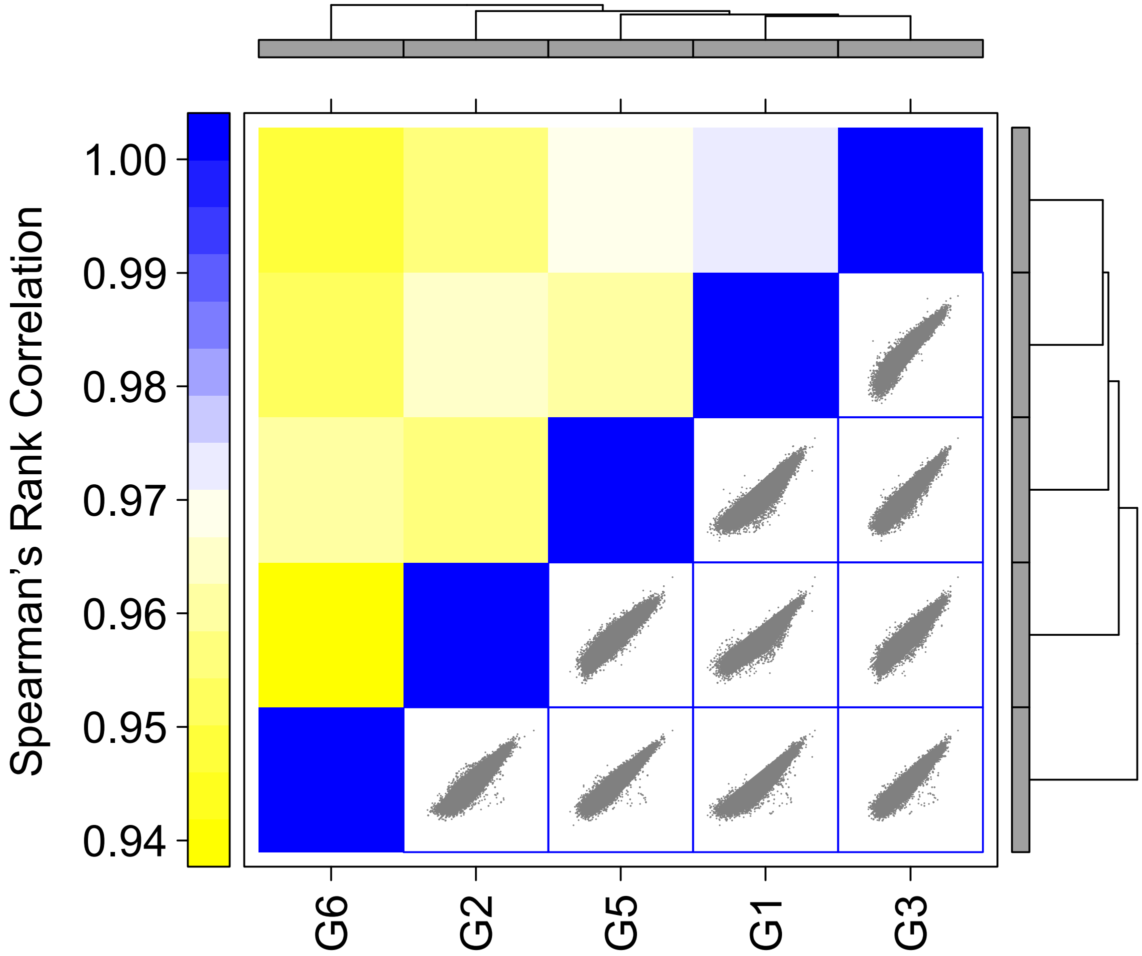
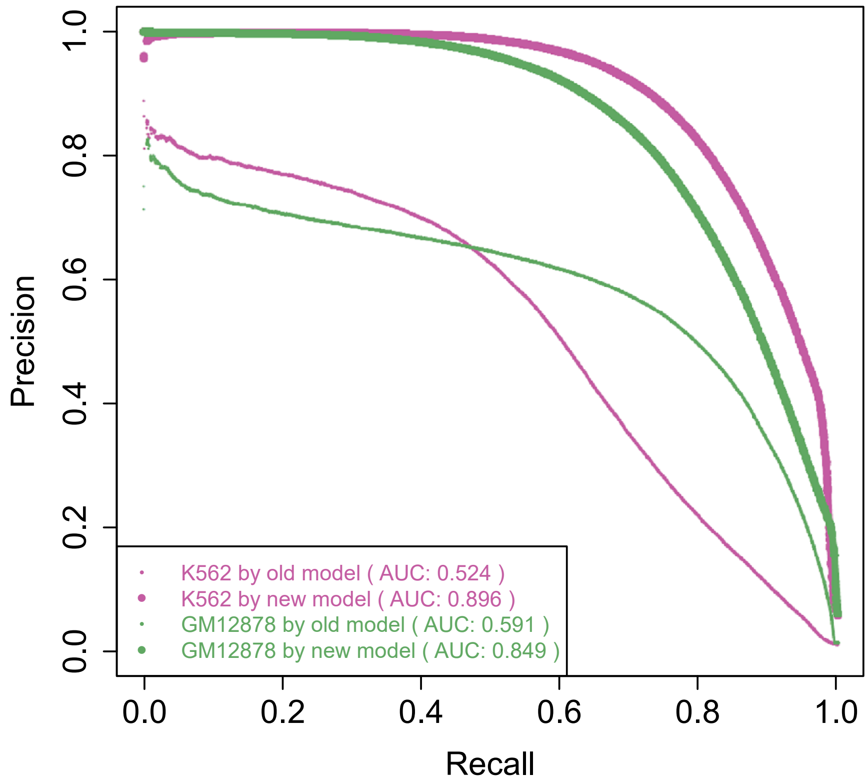
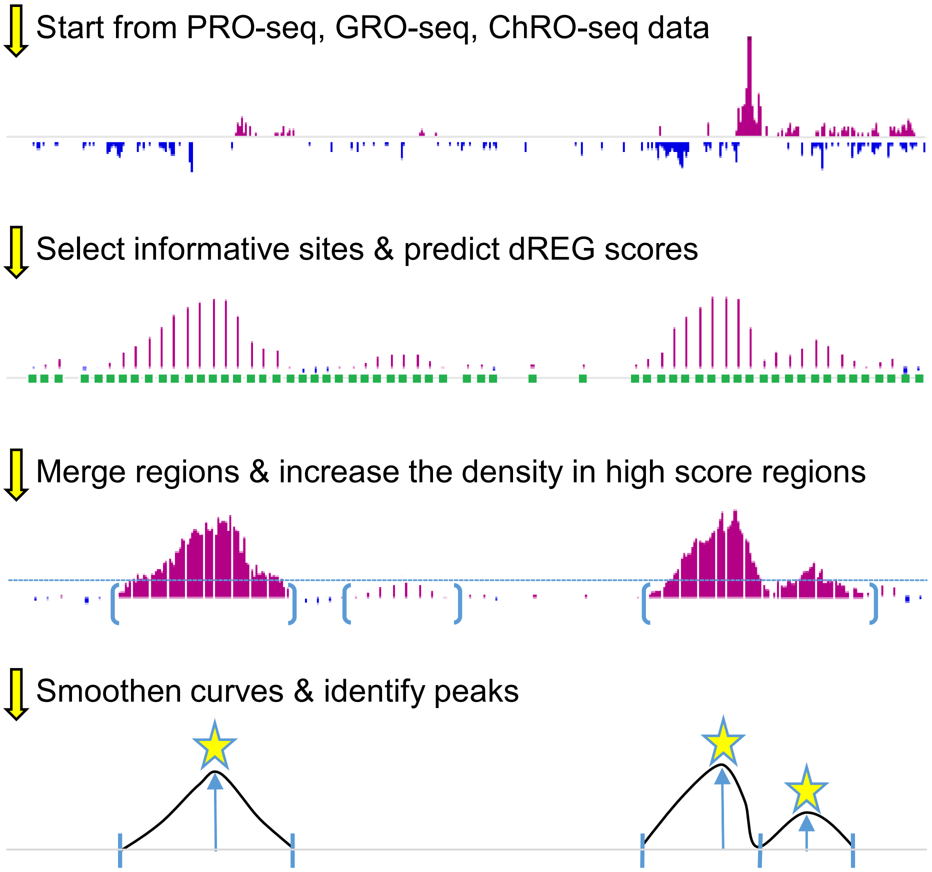


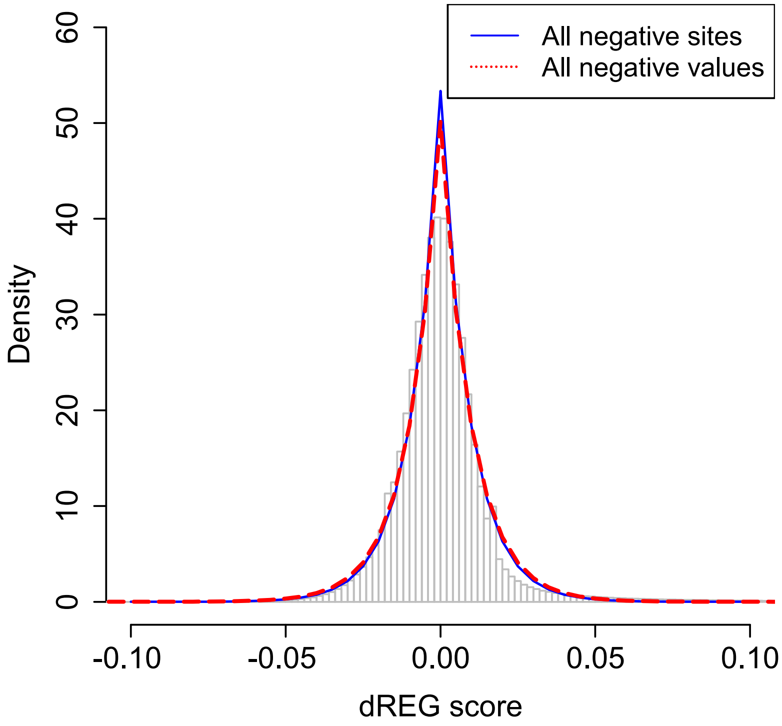
Supplemental Figure S1. PRO-seq datasets used during dREG model training. Heatmap shows Spearman's rank correlation (upper left) and raw gene-body correlations (lower right) between five PRO-seq and GRO-seq datasets used during dREG model training.



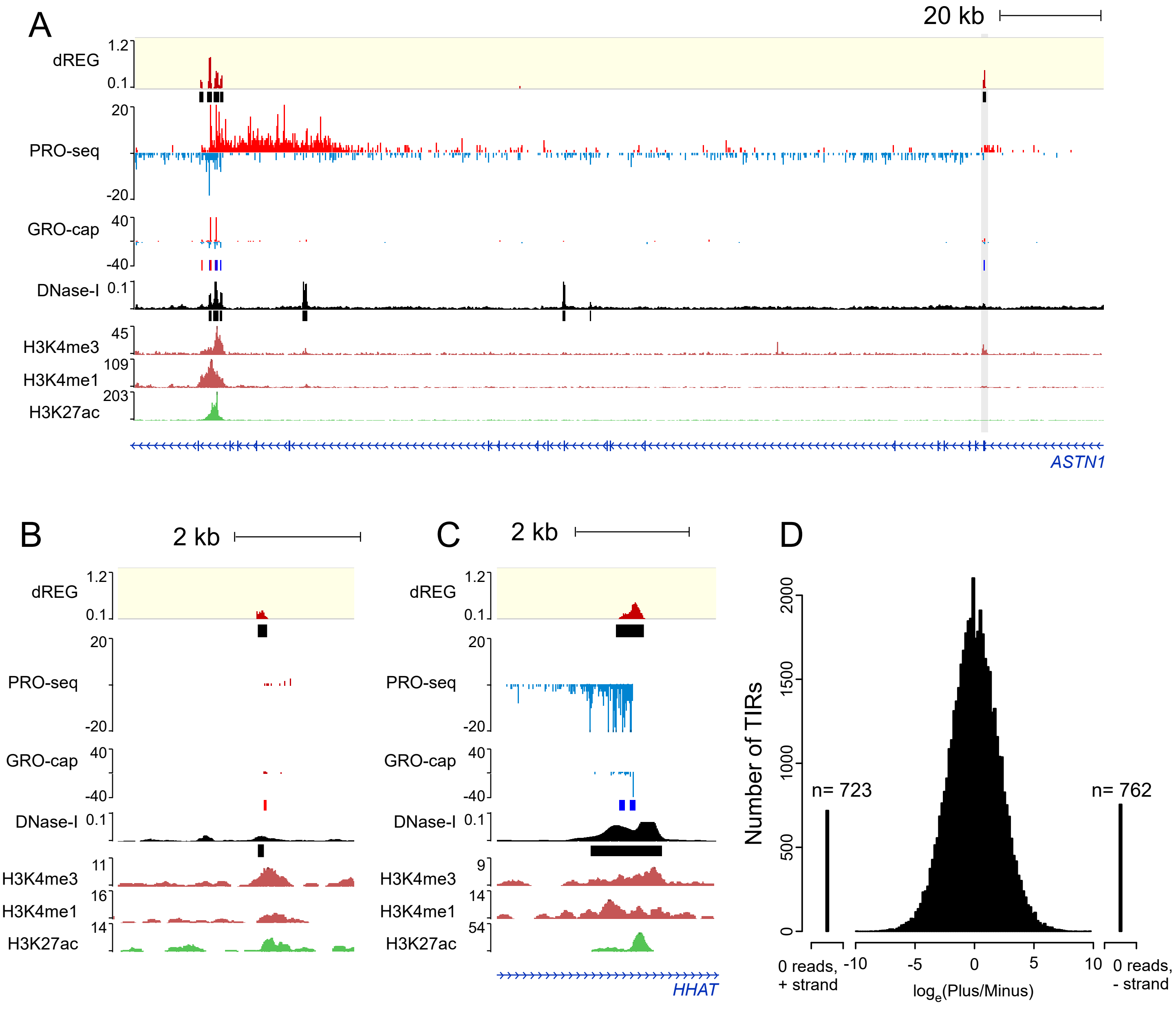
Supplemental Figure S2. dREG accurately detects the location of regulatory elements. Precision-recall curves show the precision (Y-axis; true positives/ [true positives + false positives]) and recall (X-axis; true positives / [true positives + false negatives]) of the new and previously published dREG models on the indicated dataset. The figure reflects all informative positions scored by dREG genome-wide in the indicated dataset. Both datasets were held out during model training.


Start from PRO-seq, GRO-seq, ChRO-seq data
Select informative sites & predict dREG scores
Merge regions & increase the density in high score regions
Smoothen curves & identify peaks

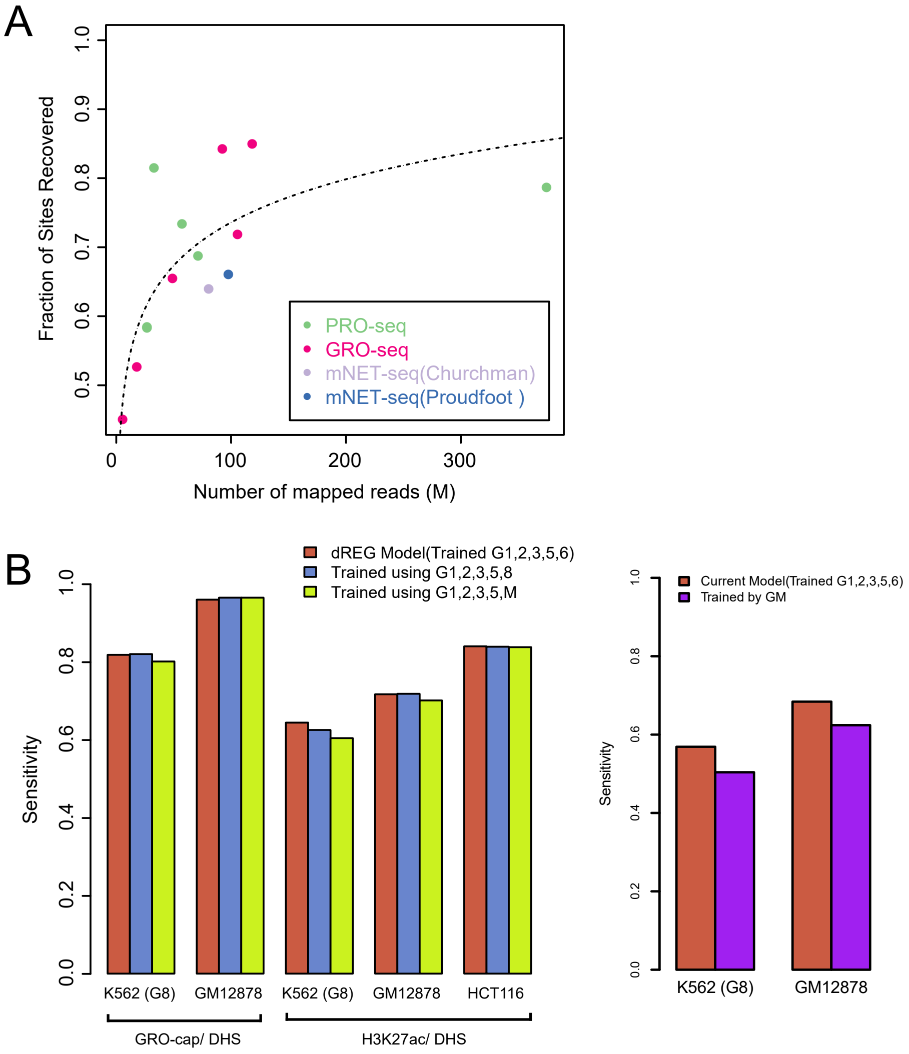
Supplemental Figure S3. Procedure for discovering transcription initiation regions (TIRs). We devised a new method for finding peaks of dREG signal, called transcription initiation regions (TIRs). dREG selects informative positions and predicts dREG signal, increases the local density in high scoring regions, and smoothes to identify local increases in signal intensity.



Supplemental Figure S4. Laplace distribution fits dREG scores in negative regions. Histogram shows the density (Y-axis) of dREG scores (X-axis) in regions that were defined as true negatives using orthogonal sources of genomic data. The lines represent the best fits to the distribution based on all true-negative sites (blue) or based on negative dREG scores (red).



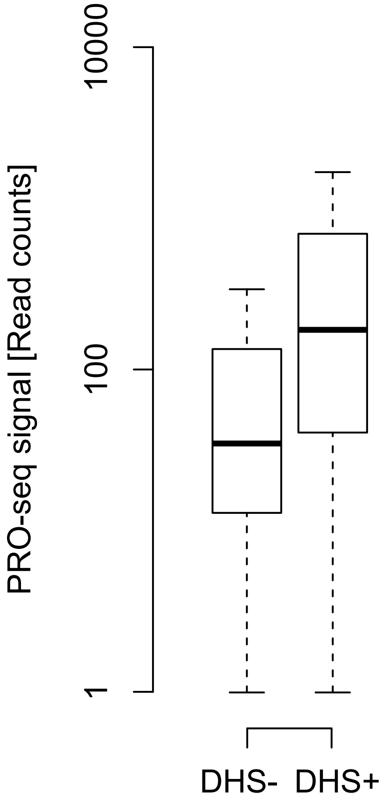
Supplemental Figure S5. dREG identifies unidirectional TREs. (A-C) WashU Epigenome Browser visualization of dREG signal, PRO-seq data, GRO-cap, DNase-seq, and H3K27ac ChIP-seq near the *ASTN1* and *HHAT* genes. TIR indicated by the gray bar (A) lacks signal in both H3K27ac and DNase-seq. Two TIRs (B-C) are supported by reads on only one strand. (D) Ratio of reads in a 500 bp window of the TIR center (TIR center to +/-250) on the plus and minus strand identified a large diversity in the directionality index on the plus and minus strand in the deeply sequenced K562 dataset.



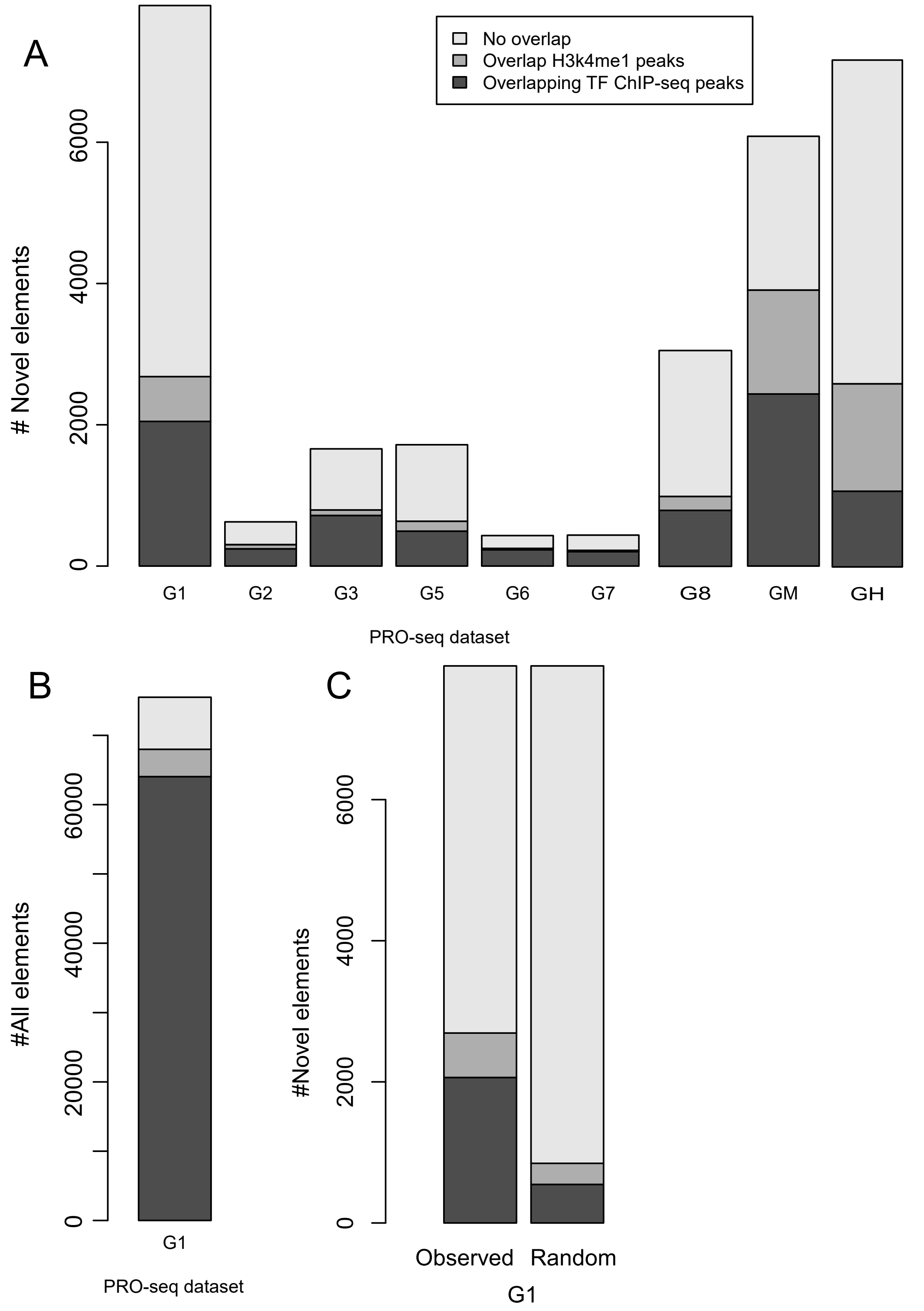
C

	TIRs	TPR	FDR
NET-seq (Proudfoot)	59,860	0.66	0.12
mNET-seq (Churchman)	62,644	0.64	0.25

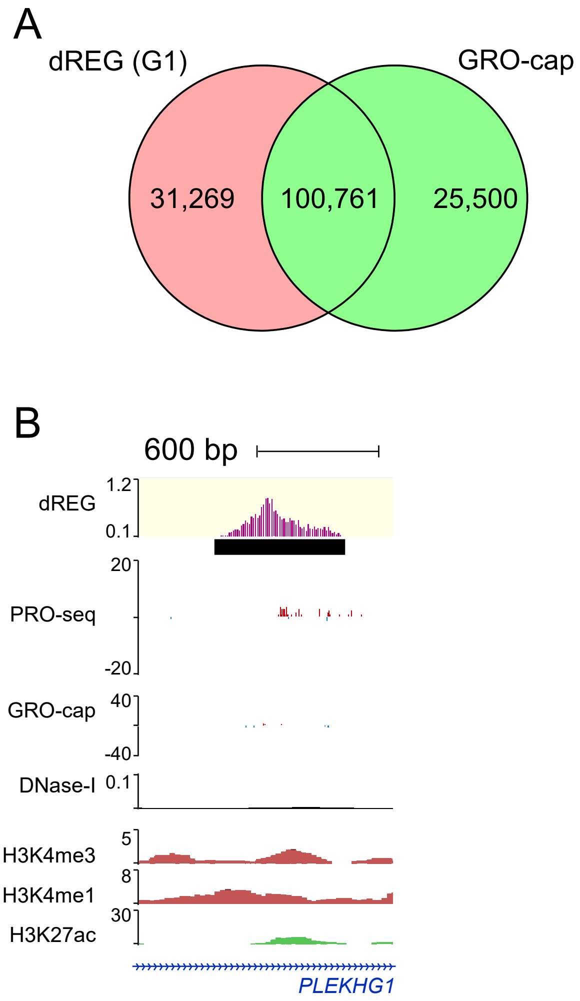
Supplemental Figure S6. The new dREG model generalizes well to other data types. (A) Scatterplot shows the fraction of sites recovered (Y-axis) at a 5% estimated false discovery rate as a function of sequencing depth (X-axis) for 13 datasets shown in Supplementary Table 1. The best fit lines are shown. The color represents the assay that was used to obtain the data, either PRO-seq (green), GRO-seq (pink), Churchman's mNET-seq (purple), or Proudfoot's mNET-seq (blue). (B) Barplots show sensitivity for different datasets at an estimated 5% FDR. Left: Sensitivity is shown for the standard dREG model (red), an alternative dREG model trained on an additional K562 GRO-seq dataset generated outside of Cornell (blue), or a dataset trained on both K562 and GM12878 (green). Right: The standard dREG model was compared to an SVR model trained using only GM12878; evaluation was performed on chr22, which was held out during training. (C) Table shows the number of TIRs recovered, the true positive rate (TPR), and the upper-bound false discovery rate (FDR) derived using the fraction of TIRs that do not intersect peak calls in H3K27ac or DNase-seq data. Proudfoot data was collected using the unphosphorylated Pol II antibody (8WG16).



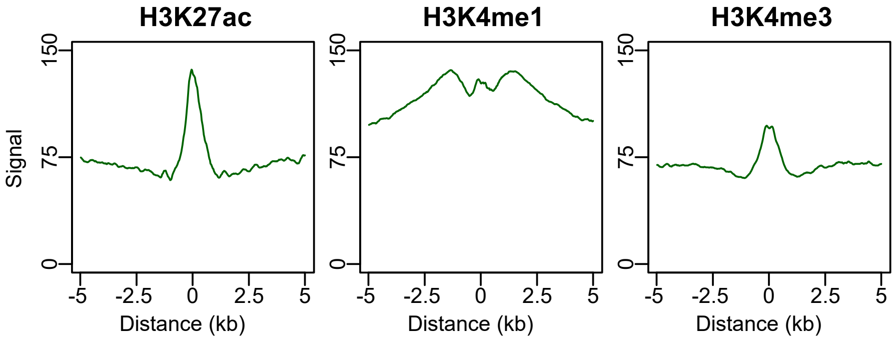
Supplemental Figure S7. Boxplots show the difference in PRO-seq signal between DHS+ and DHS- TIRs. Boxplot shows the difference in PRO-seq read counts between dREG+DHS- and dREG+DHS+ TIRs. The Y axis represents the number of reads found within 250 bp of each TIR.



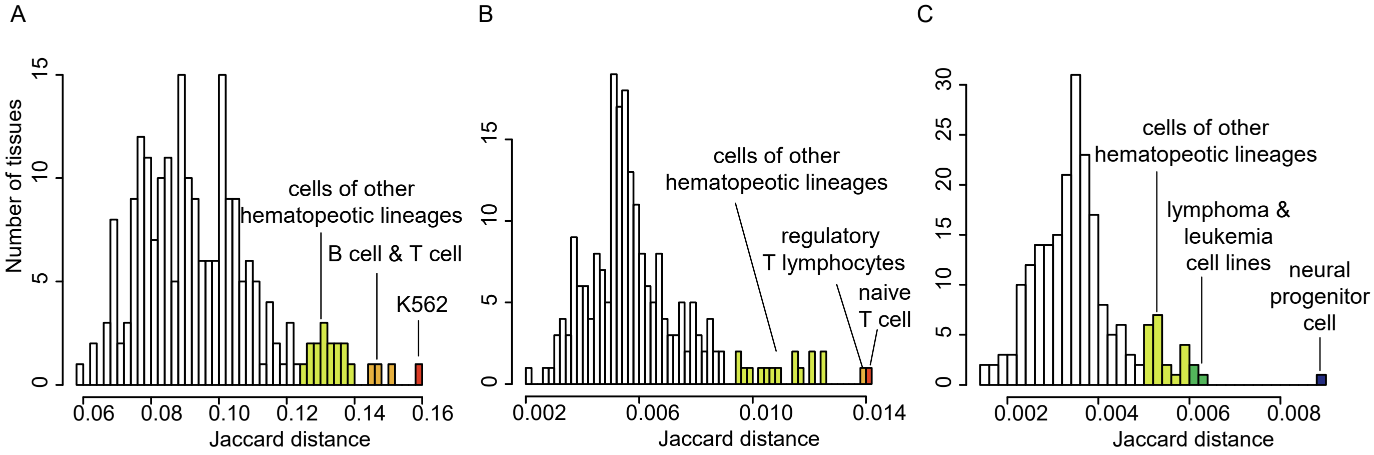
Supplemental Figure S8. Novel elements discovered using dREG frequently overlap transcription factor ChIP-seq peaks. (A-C) Plots shows the number of TIRs discovered using dREG, but not found in DNase-seq or H3K27ac ChIP-seq (Y-axis) for nine PRO-seq datasets. Seven datasets were used from K562 cells (G1-8), one dataset was used from GM12878, and one from HCT116. The number of novel dREG sites overlapping transcription factor ChIP-seq peaks (dark gray),H3K4me1 ChIP-seq peaks (gray), or not overlapping either (light gray) are shown. (A) Shows the number of novel TIRs; (B) shows the number of all TIRs; (C) Shows the number of novel TIRs for a K562 dataset (G1) compared to what is expected for randomly positioned peaks that are the same size.



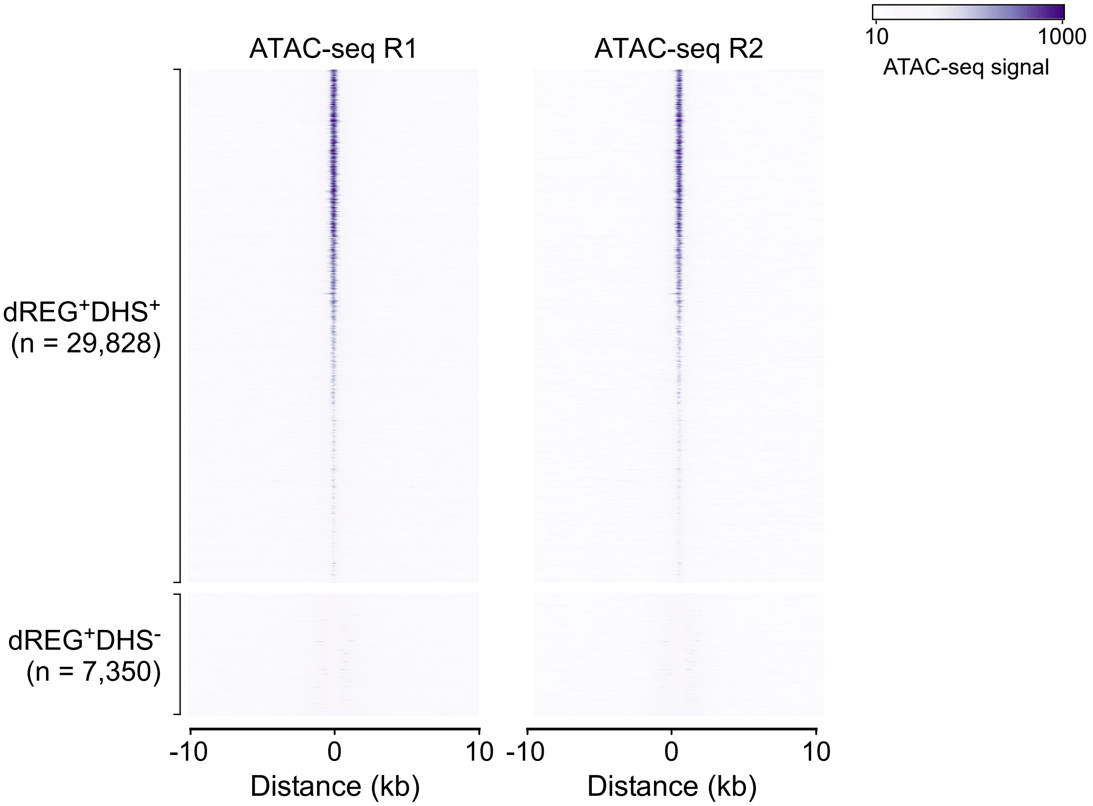
Supplemental Figure S9. Overlap between dREG and GRO-cap. (A) Venn diagram shows the overlap between GRO-cap and dREG. The intersection shows the number of GRO-cap sites. (B) Shows an example of a TIR discovered using dREG, but missed by GRO-cap.



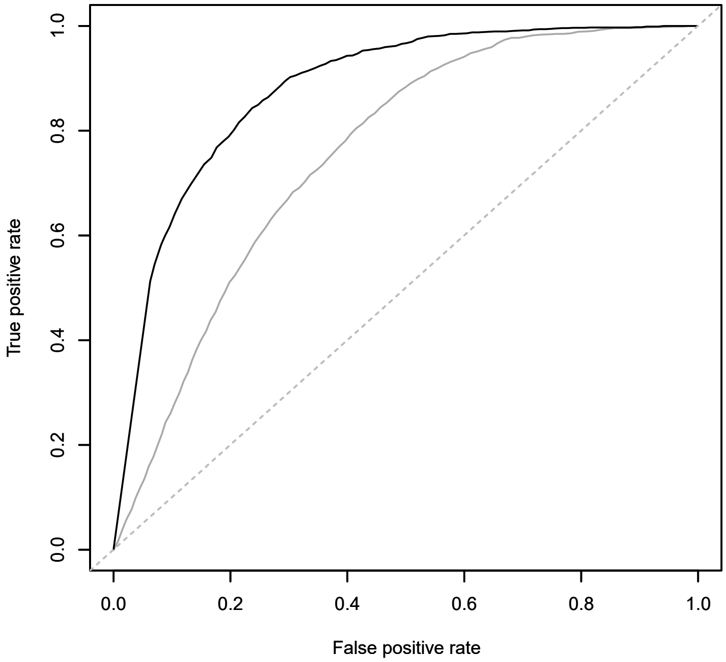
Supplemental Figure S10. Novel dREG TIRs overlap local increases in histone marks associated with enhancers. Meta plots show the raw signal for H3K27ac, H3K4me1, and H3K4me3 near TIRs identified using dREG, but were not found in peak calls for H3K27ac or DNase-seq.



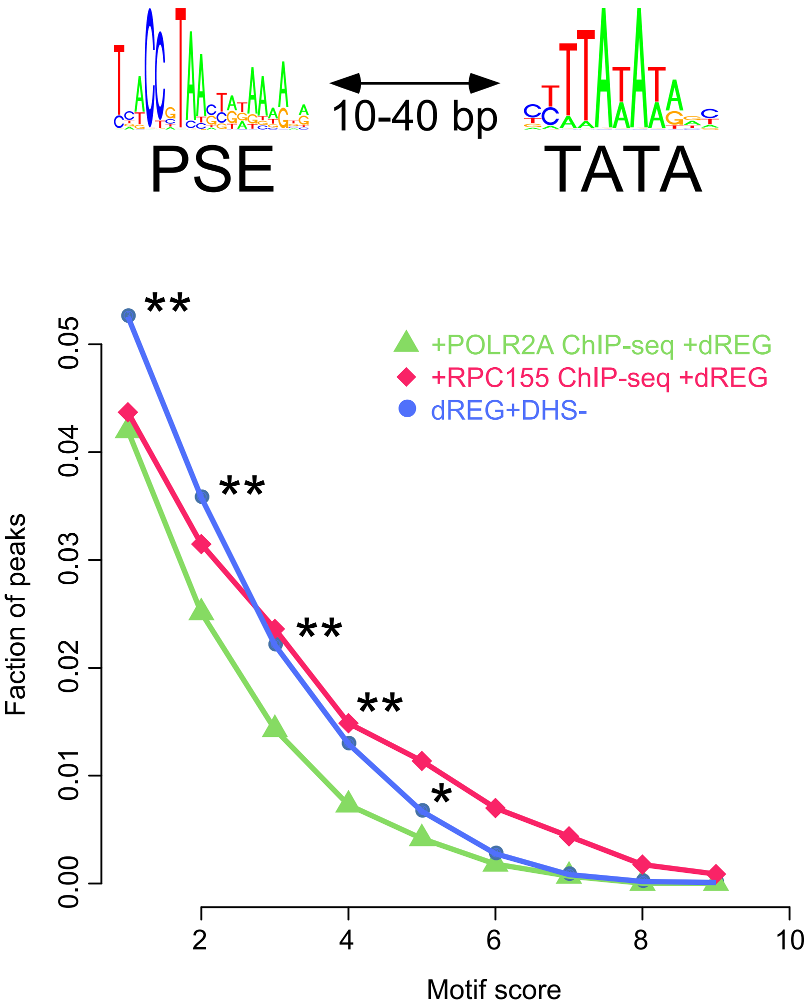
Supplemental Figure S11. Histogram shows the distribution of jaccard distance between dREG sites in K562 cell with respect to DNase-seq sites in ENCODE reference cell types. Jaccard distance was calculated for (A) all dREG sites in K562 cells, (B) dREG sites in K562 cells that do not overlap with DNase I hypersensitive sites, and (C) dREG sites in K562 cells that do not overlap with DNase I hypersensitive sites nor with H3K27ac ChIP-seq peaks. Major cell types among the outliers were colored and labeled.



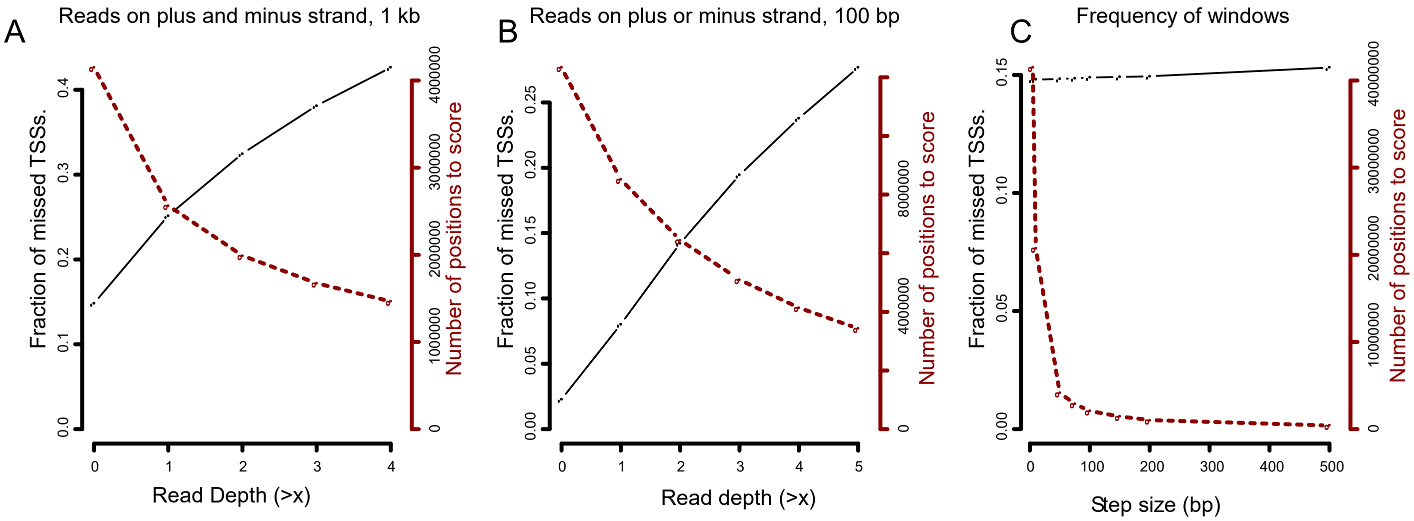
Supplemental Figure S12. dREG+DHS- sites do not reflect clonal differences in between K562 cells grown by ENCODE and by our lab. Heatmaps show raw signal for two replicates of ATAC-seq data from K562 cells grown in our lab and clonally related to K562 cells used to produce PRO-seq data. Data is shown near dREG+DHS+ (n= 29,828) and dREG+DHS- (n= 7,350). Sites were ordered by dREG score.



Supplemental Figure S13. Accuracy of classifying DHS status of dREG TIRs. Receiver operating characteristic (ROC) plot shows the accuracy of predicting whether a TIR identified using dREG was also identified as a DHS using ENCODE DNase-seq data. Classification was performed using 100 transcription factor ChIP-seq datasets in K562 cells (black, auROC= 0.88) or the dREG score alone (gray, auROC= 0.75).



Supplemental Figure S14. Pol III initiation motifs are enriched in dREG+DHS- TIRs. Scatterplot shows the fraction of TIRs that have DNA sequence motifs representing PSE and a TATA box in the right orientation with respect to each other and with a spacing of 10-40 bp between them. Motifs were obtained from (James Faresse et al. 2012) and are shown on top. P-values comparing dREG+DHS- TIRs to dREG+ POLR2A ChIP-seq+ TIRs are denoted by an asterisk (** $p < 1x10^{-5}$; * $p < 5x10^{-3}$; Fisher's exact test).



Supplemental Figure S15. Optimization of heuristics to decrease number of sites. (A-C) The fraction of missed TSSs (left, black) and the number of positions to score (right, red) as a function of the threshold read depth for reads on both the plus and minus strand in a window of 1 kb (A); the number of reads on either strand in a 100 bp window (B); or the frequency of windows examined (C).