Supplementary Material

Enhancer RNA Profiling Predicts Transcription Factor Activity

Joseph G. Azofeifa[1,2], Mary A. Allen[2], J. R. Hendrix[1,3], Timothy Read[2,4], Jonathan D. Rubin[4] & Robin D. Dowell[1,2,3,*]

[1]*Department of Computer Science, University of Colorado, Boulder CO 80309*
[2]*BioFrontiers Institute, University of Colorado, Boulder CO 80309*
[3]*Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder CO 80309*
[4]*Department of Biochemistry, University of Colorado, Boulder CO 80309*

∗ Corresponding author: robin.dowell@colorado.edu

Contents

1 Nascent transcription data processing

SRA files were downloaded from the NCBI Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`), accession numbers provided in Supplemental Table S3. The SRA files were converted into fastq format using fastq-dump 2.3.2-5 in the SRA Toolkit. The reverse complement of data produced by a second strand synthesis kit was constructed using fastx-reverse-complement -Q33. Human and mouse fastx files were mapped to the hg19 and mm10 genomes, respectively, using bowtie2 (Langmead and Salzberg, 2012), version 2.0.2 -very-sensitive. The resulting sam files were converted to sorted bam files using samtools (Li et al., 2009), version 0.1.19. Each sorted bam was converted into two strand-separated bedgraphs (one file containing positive strand and one with negative strand reads) using bedtools (Quinlan and Hall, 2010) genomeCoverageBed version 2.22.0. We used the hg19_all.fa genome file from UCSC for human data and mm10_Bowtie2_index.fa for mouse data. The bedgraphs were sorted then converted to bigwig format using bedGraphToBigWig (Kent et al., 2010). The hg19.chrome.sizes and mm10.chrome.sizes input files were made using fetchChromSizes from UCSC and the hg19 and mm10 genome files, respectively.

2 Genomic feature data integration

Frequently, we compare two (or more) data sets for association between the genomic features. Unless otherwise stated, we say two genomic features overlap or associate if the two elements are located on the same chromosome and the center of each feature is within 1,500 base pairs (bps) of each other. For example, let some TF binding peak be located on chromosome 1 with a start coordinate of 10,000 and stop coordinate of 10,405 and an eRNA origin at chromosome 1 with a start coordinate of 10,200 and stop coordinate of 10,201. Given that the center of TF binding peak is $((10405 + 10000)/2 = 10202.5)$ and $|10202.5 - 10200.5| < 1500$ we would say these two genomic coordinates associate. The 1,500bp cutoff is justified in Supplemental Fig. 2. Furthermore, all genomic coordinates refer to hg19 or mm10 for human or mouse data sets, respectively.

3 Validation of revised Tfit algorithm

Prior work, including the earlier version of Tfit (Azofeifa and Dowell, 2017), has demonstrated a tight relationship between bidirectional transcription and enhancer associated histone marks (Danko et al., 2015; Azofeifa and Dowell, 2017). Using the modified Tfit, DNase I hypersensitivity (DHS1), histone 3 lysine 27 acetylation (H3K27ac) and histone 3 lysine 4 mono-, di-& tri-methylation significantly associate with non-promoter bidirectional transcription, as expected (Supplemental Fig. 3). Indeed, histone modifications are displaced from bidirectional centers (origins) supporting the presence of a nucleosome-free region localized precisely at the origins of bidirectional transcript initiation (Supplemental Fig. 3B).

As reported previously (Azofeifa and Dowell, 2017), we observed that super enhancers (Khan and Zhang, 2016) and regions annotated as transcribed (Azofeifa et al., 2014; Chae et al., 2015; Danko et al., 2015) often contain multiple origins. For example, as shown in Supplemental Fig. 1, the entirety of this super enhancer annotated region is transcribed and identified as a single region by most nascent transcription analysis algorithms (Allison et al., 2013; Azofeifa et al., 2014; Chae et al., 2015; Danko et al., 2015). However, since

Tfit looks for sites of initiation rather than transcribed regions, our model identifies three origins of transcription within this region, each giving rise to two transcripts proceeding in opposite directions. For clarity, we refer to the region as a "transcribed region", each inferred position of polymerase initiation as a bidirectional, regardless whether one or two transcripts are produced. If the site of polymerase initiation is not promoter associated, we refer to it as an eRNA origin, and the resulting transcripts (in Supplemental Fig. 1 we observe six) as individual eRNAs.

3.1 Depth of sequencing on eRNA inference

As the sequencing depth of each experiment varied dramatically, we sought to understand the relationship between depth and eRNA inference. To this end, we inferred eRNAs across the complete compendium of human data sets (491 experiments) and plot the number of eRNA origins predicted by Tfit against the depth of sequencing (Supplemental Fig. 11). This contour plots shows a weak but present correlation in the number of bidirectionals inferred (in any experiment) relative to the underlying sequence depth of that experiment. In other words, more depth predicts a slightly higher rate of bidirectional (or eRNA origins) inference.

3.2 Characterization of eRNA inference: comparison to ChIP

Furthermore, prior work observed a significant overlap between eRNA predictions and transcription factor binding data (Danko et al., 2015). Here, we confirm this observation using our eRNAs called by Tfit. To this end, we integrated our set of eRNA origins with the genomic binding locations of 139 proteins profiled by ChIP-seq, also in K562 cells (Supplemental Table S1). Consistent with previous results (Danko et al., 2015), 98% of eRNAs are bound by at least one regulator, where an average of 52.9 regulators localize at any one eRNA (Supplemental Fig. 5A-B). In fact, we observed three distinct patterns of TF binding (Supplemental Fig. 4A): TFs that bind all eRNAs (32 factors co-occur with over 75% of all eRNAs; clade IV); TFs that bind only a few eRNAs (39 factors associate with no more than 20% of all eRNAs ;clades I & II); and TFs that bind to many eRNAs but only with unique TF partners (58 factors occur under specific combinatorial patterning, e.g. GATA2/NR2F2/GABPA and FOSL/ATF3 strongly co-localize at eRNAs; clades III & V). TF-combinatorial control also plays a pivotal role in downstream gene regulation (Bilu and Barkai, 2005). In general, the number of TFs co-localized at sites of open chromatin is larger when an eRNA is present than not (Supplemental Fig. 6A). Furthermore, TF co-association dramatically increases when considering eRNA presence (Supplemental Fig. 6B). In summary, unique sets of TFs bind to specific eRNA origins.

Finally, we examined the co-localization of TF binding relative to eRNA origins (Supplemental Fig. 4B). We observed two classes of regulators: 84% of TFs exhibit centered, unimodal localization with eRNA origins and 16% display significantly displaced peak localization flanking eRNA origins (Supplemental Fig. 5C). For example, factors such as RBB5, PHF8 and CDH1 are significantly displaced an average of 150, 200, and 398bp from the eRNA origin, respectively (Supplemental Fig. 5D). Regulators with displaced peak localization are significantly enriched for ontological definitions such as "histone modification," "chromatin organization" and "histone deacetylation" consistent with the bimodal distribution of histone modifications observed in Supplemental Fig. 3B (p-value $< 10^{-6}$).

4 Motif curation and motif scanning

Transcription factor binding motifs are summarized as a position-specific probability distribution over the nucleotide (ATGC) alphabet, referred to commonly as a position weight matrix (PWM). These models were gathered from the HOCOMOCO (Kulakovskiy et al., 2013, 2016) database of hand-curated transcription factor binding motif models for human and mouse (downloaded from `http://hocomoco.autosome.ru/final_bundle/HUMAN/mono/` on 12/10/15). In total there exist 641 and 427 motif models for human and mouse, respectively.

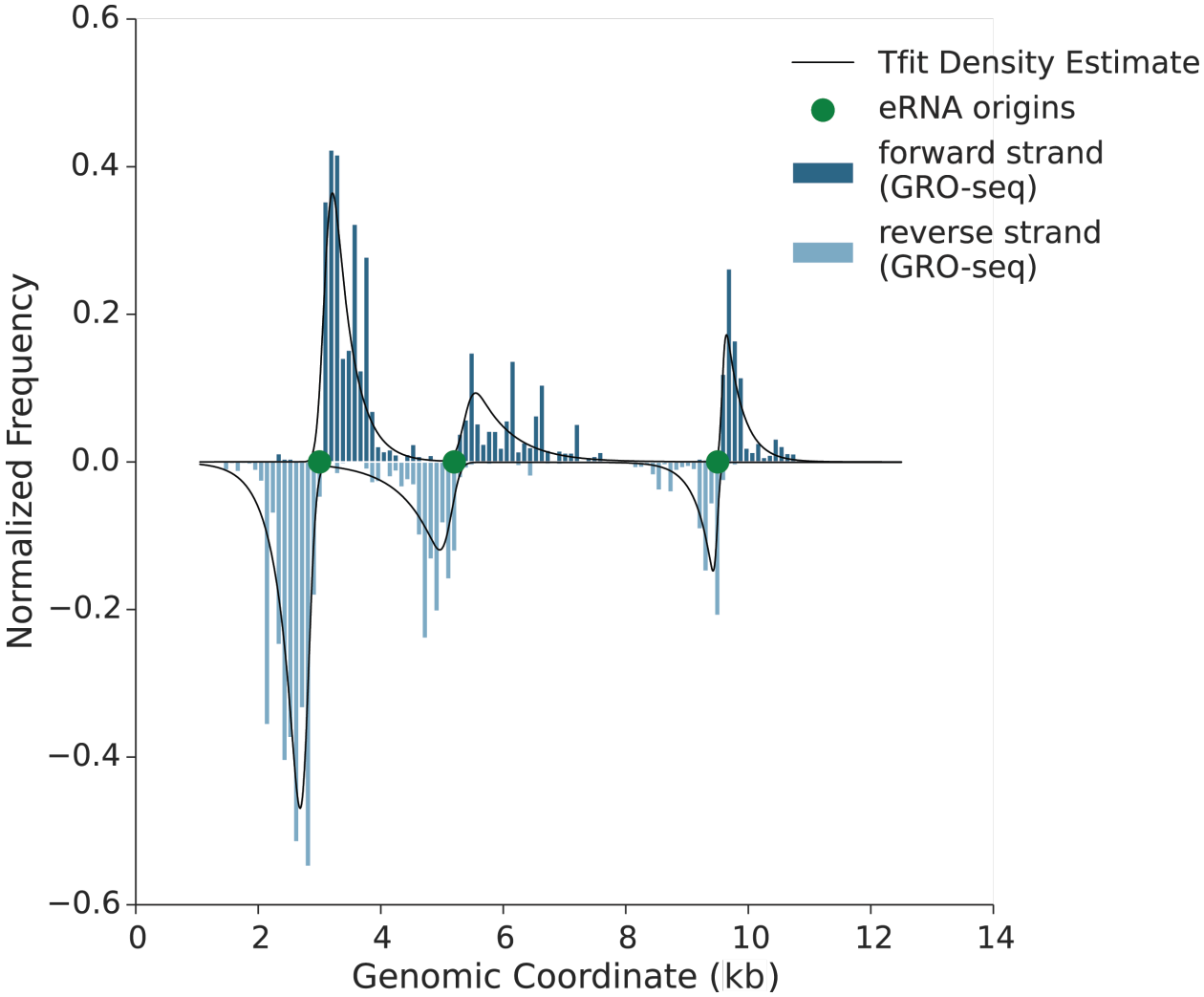
Motif scanning around bidirectionals was performed by the algorithm outlined by Staden (Staden, 1994). False discovery rate (FDR) was quantified by the approach outlined by Storey (Storey, 2002). Only sequences where the FDR did not exceed 10^{-6} were considered a significant TF sequence motif instance and the center of the matching sequence was used for all subsequent analysis. The basic stationary background model was estimated from GC content of hg19 (human, 42.3%) and mm10 (mouse, 41.2%) genome builds. Motif scanning was implemented in the C++ programming language using the popular openMPI framework to perform massive parallelization on compute clusters. This implementation, referred to as MDS, can be downloaded at `https://biof-git.colorado.edu/dowelllab/MDS`.

5 Associated File Types

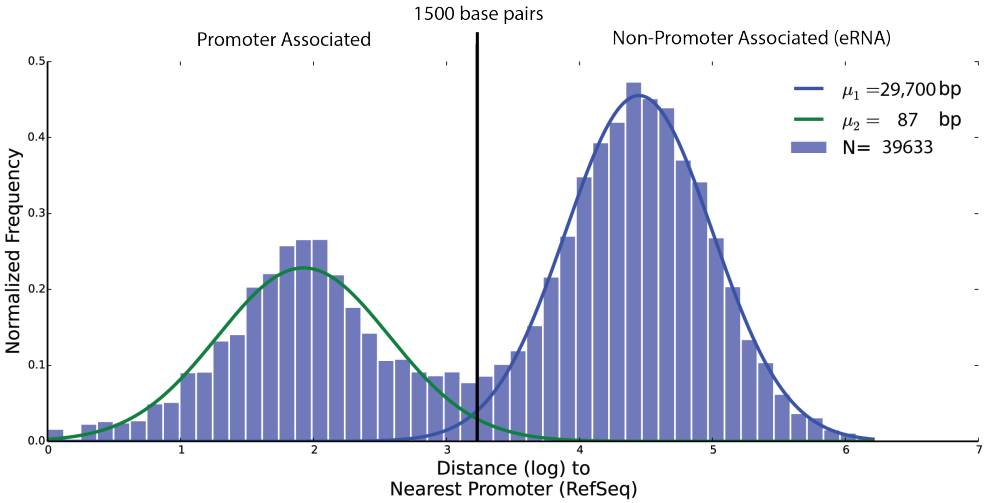
We provide here two important sets of data files: (1) a folder of Tfit predicted eRNA origins for our compendium of publicly available human and mouse nascent transcription data sets (771) (Supplemental Data S1); and (2) a histogram of motif locations surrounding eRNA origins for each of the 771 nascent transcript data sets and 641 motif models (Supplemental Data S2). These data files are available at `http://dowell.colorado.edu/pubs/MDscore/`

6 Jupyter Notebook

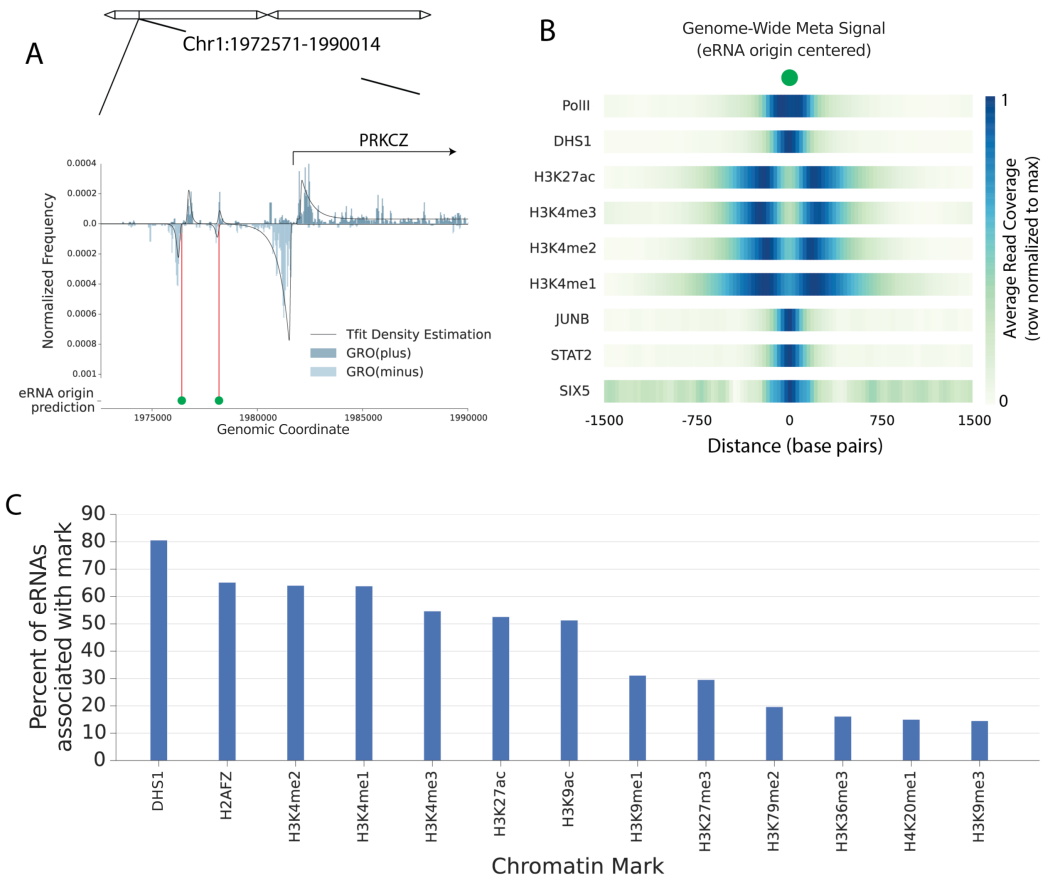
To aid in data exploration, we present a user-friendly Jupyter notebook to compute MD-score statistics, visualize changes between data set comparisons and explore Tfit output. Given the shear size of the data analyzed (771 nascent transcription data sets, 641 TF motif models), we could not explore all possible comparisons. Instead we provide here a tool, wrapped within the Jupyter Notebook environment, to explore this data resource. The python package motif_displacement_analysis along with the Jupyter Notebook environment can be downloaded at `https://biof-git.colorado.edu/dowelllab/motif_displacement_analysis`. Capabilities include drawing and displacement heatmaps related to the motif displacement distributions, quantifying different MD-scores, mean motif distances and generic KS-test statistics. As the diversity and quantity of nascent transcription data increases, eRNA profiling will likely play a significant role in defining the biological systems where individual TFs exert their regulatory influence.



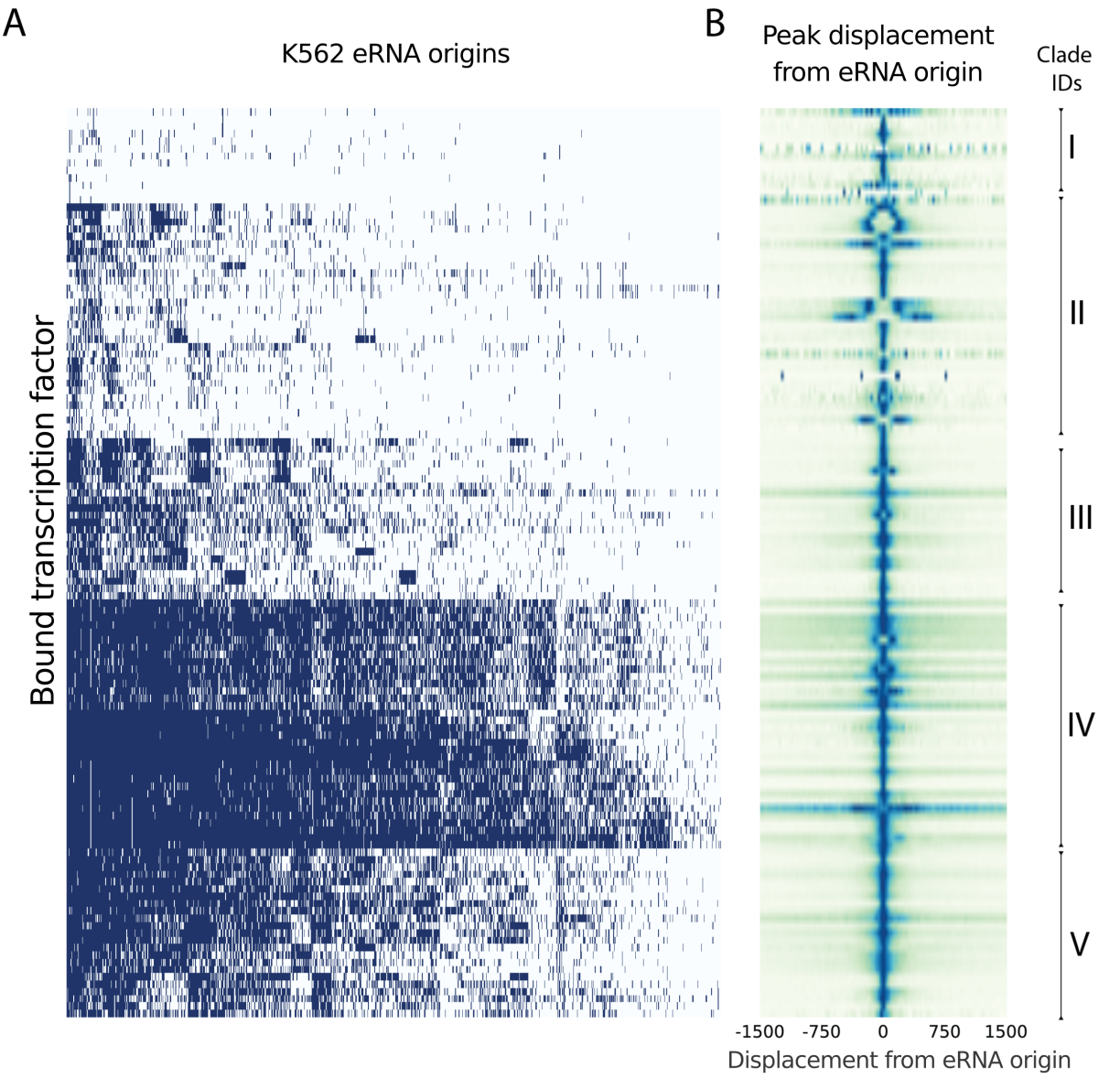
Supplemental Figure 1: **Example annotated super enhancer region showing RNAP inference.** An annotated super enhancer (starting at Chr2:10,456,371), GRO-seq read coverage from an HCT116 data set (Allen et al., 2014) (probability density normalized) and the final inferred density function obtained by Tfit (Azofeifa and Dowell, 2017). Via Bayesian model selection, three distinct eRNA origins are identified. Green dots indicate the eRNA origin estimate.



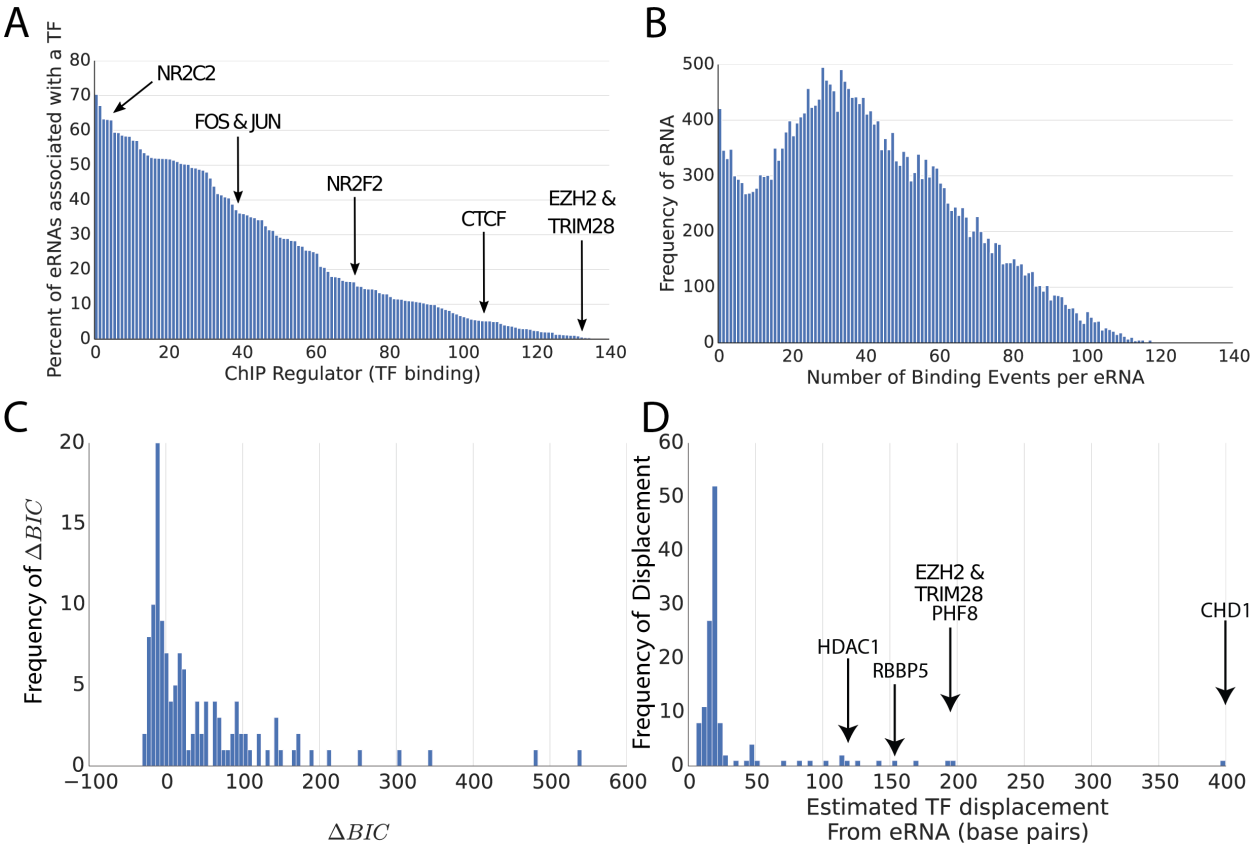
Supplemental Figure 2: **Most sites of bidirectional transcription identified by Tfit lack association with promoter regions.** Sites of bidirectional transcription were profiled in a K562 GRO-cap (Core et al., 2012) (SRR1552480) experiment using Tfit. A promoter is defined as the region associated with a RefSeq (release 76) annotated gene's start site. Bimodiality was estimated via a two component Gaussian mixture model fit with the EM algorithm. The mean of each Gaussian curve is given in the key.



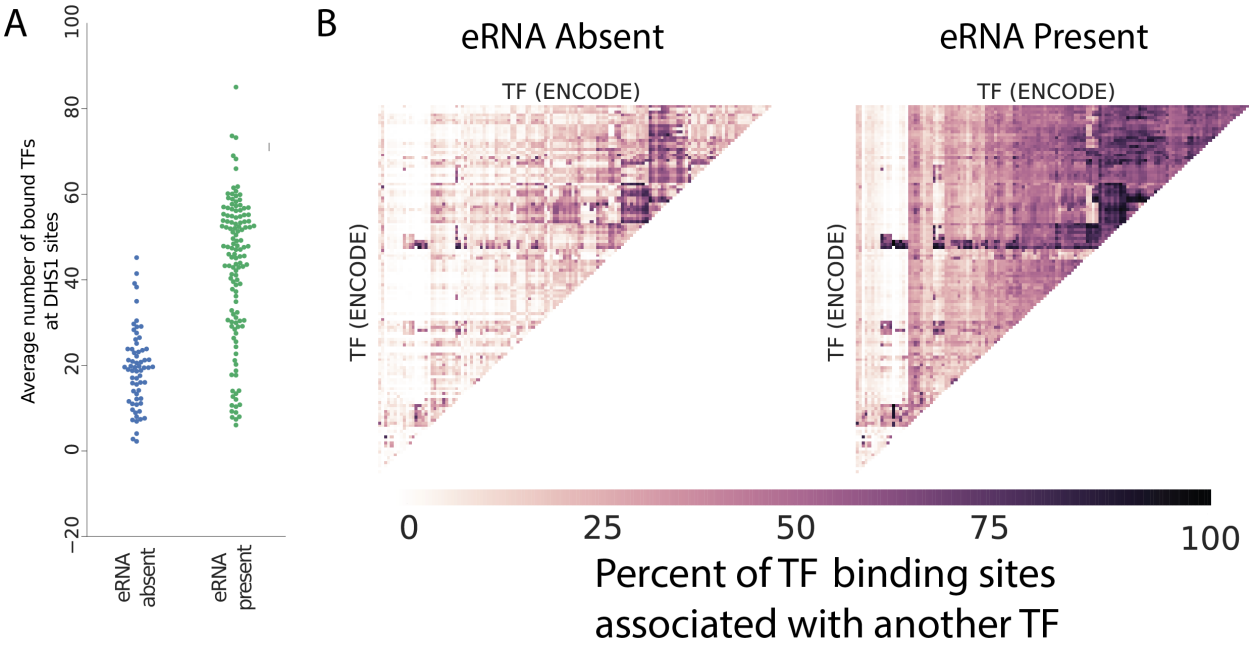
Supplemental Figure 3: **Sites of non-promoter associated bidirectional transcription overlap marks of regulatory DNA.** (A) An example locus displaying nascent transcription read coverage (HCT116 GRO-seq (Allen et al., 2014)) with the overlaid density estimation via Tfit and the associated eRNA origin predictions (green dots). (B) Genome-wide meta-signal for marks of active chromatin aligned to eRNA origins inferred by Tfit in a K562 GRO-cap data set (Core et al., 2014) (marks in Supplemental Table S1). (C) The percentage of eRNAs in a K562 GRO-cap data set (SRR1552480 (Core et al., 2012)) that associate with specific chromatin marks, defined by ENCODE (Supplemental Table S1). Promoter associated Tfit predictions were removed from this analysis.



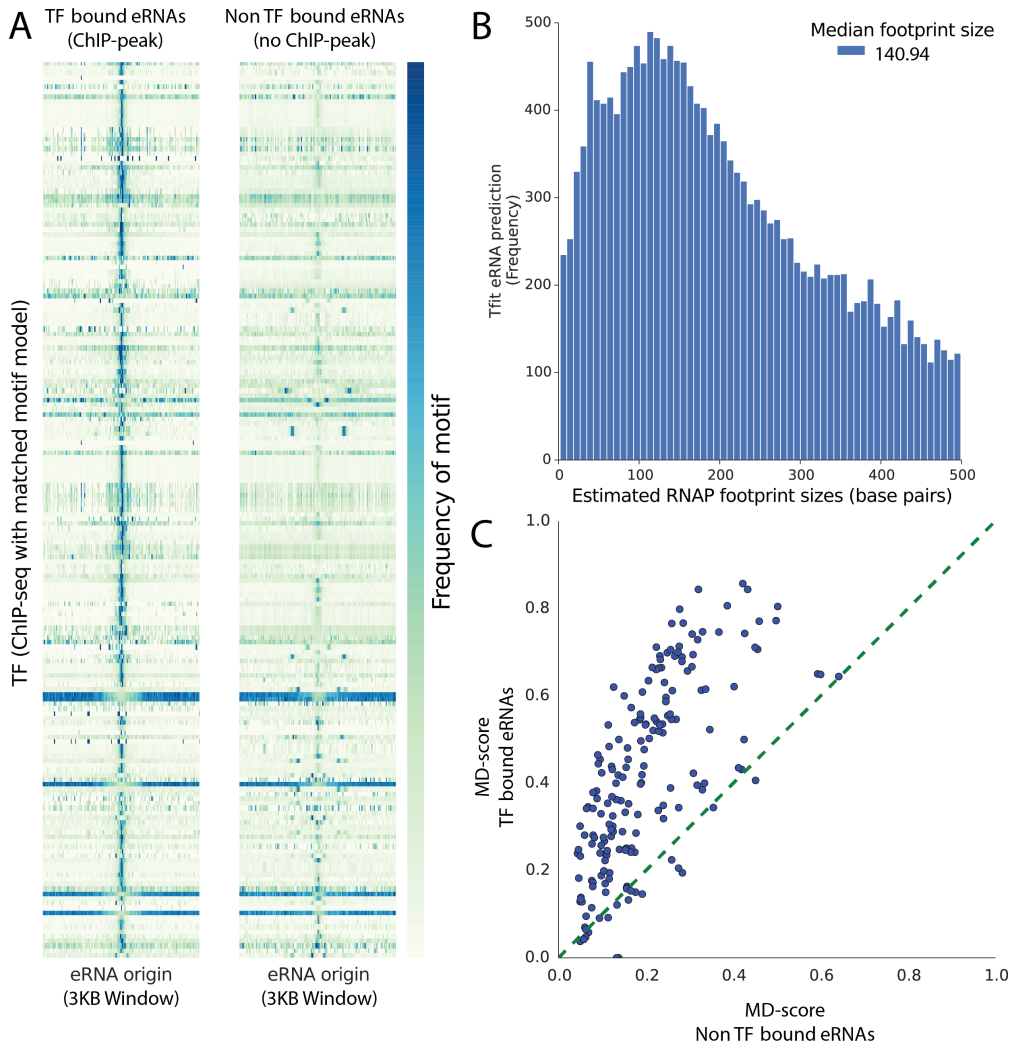
Supplemental Figure 4: **Enhancer RNAs originate from sites of regulatory protein binding.** (A) The overlap of eRNA origins (columns) with 139 regulatory proteins (rows) (Supplemental Table S1). A blue tick indicates the presence of regulatory protein binding site within 1.5kb of the Tfit inferred origin; sorted by hierarchical clustering. (B) Histogram of the spatial displacement of the regulatory protein binding peak from eRNA origins (heat is normalized to min/max of the histogram).



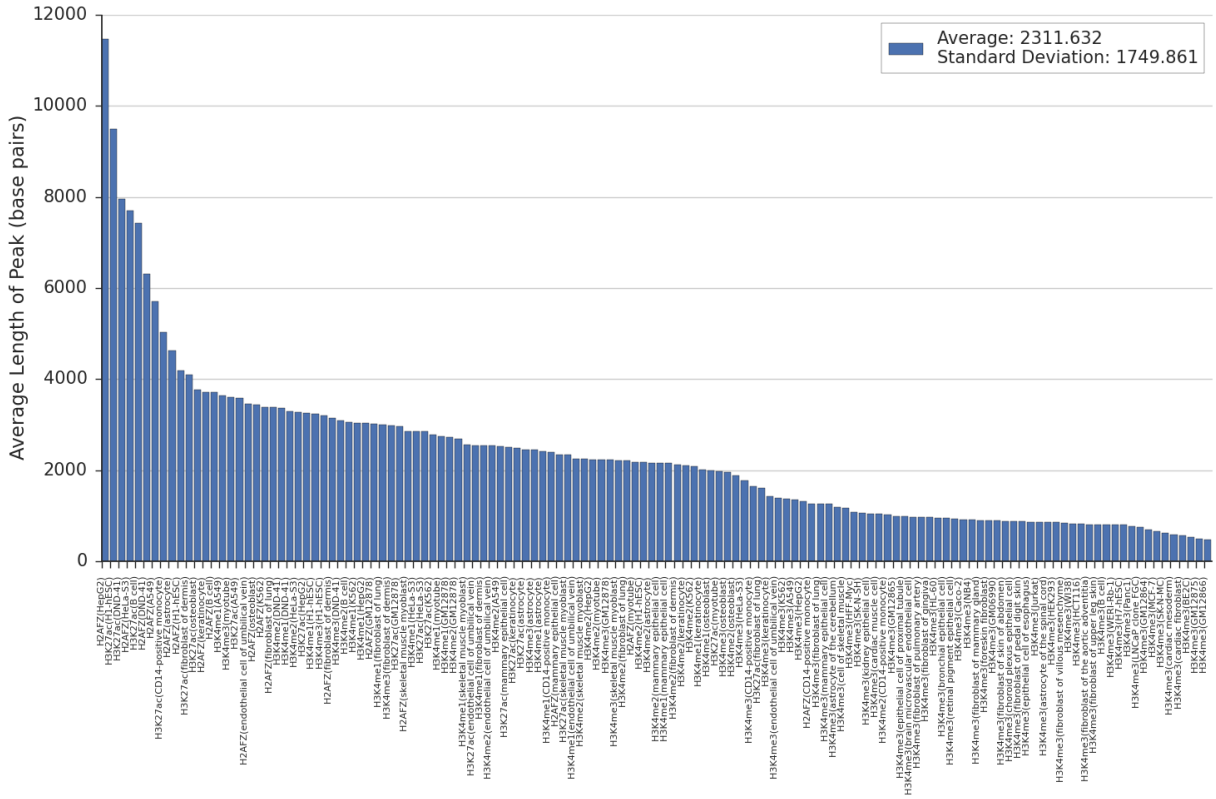
Supplemental Figure 5: **Sites of non-promoter associated bidirectional transcription overlap sites of TF binding.** (A) The proportion (y-axis) of a TFs ChIP peaks associated < 1.5kb with an eRNA origin. The x-axis is one of the 129 TFs profiled by ENCODE in K562 cells (Supplemental Table S1). (B) The number of unique TF binding peaks occurring at individual eRNAs. (C) TF displacement data was calculated within a 1.5kb radius around eRNA locations and bimodal model selection was performed via a Laplace-Uniform mixture. Briefly, a larger ΔBIC value indicates greater support for bimodal TF peak displacement. (D) The distribution of estimated peak displacements. All data from K562 cells, both nascent transcription (Core et al., 2012) (SRR1552480) and ENCODE ChIP-seq peaks (Supplemental Table S1).



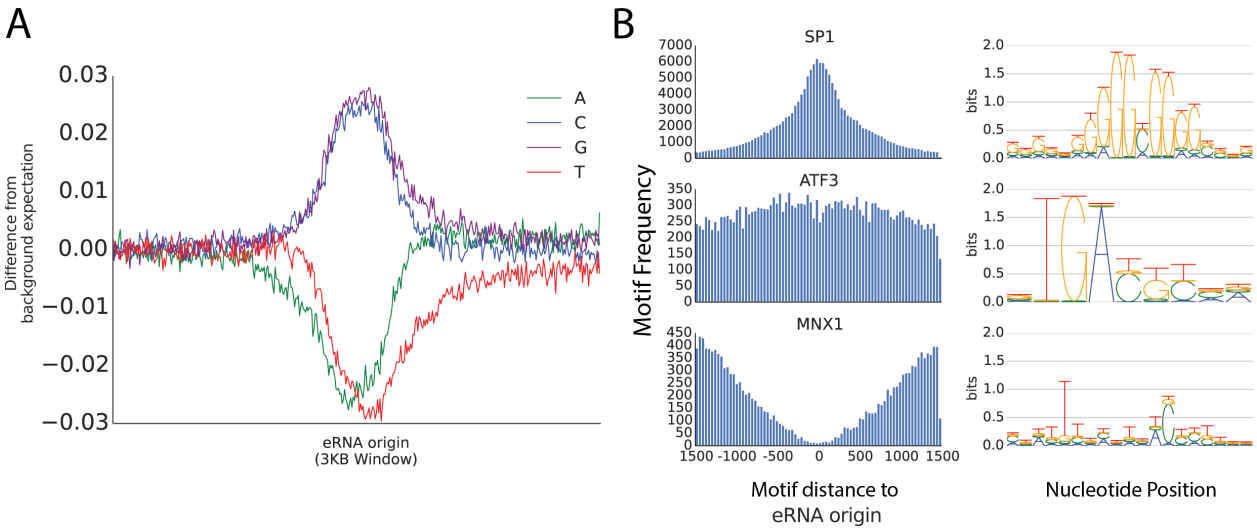
Supplemental Figure 6: **Enhancer RNA presence increases TF binding co-associativity.** (A) A swarm plot displaying the number of bound TFs at sites of open chromatin grouped by eRNA association. (B) A pairwise co-association map where increased heat indicates a greater proportion of TFs binding sites also bound by another TF; categorized by eRNA association. Data is K562 (Supplemental Table S1).



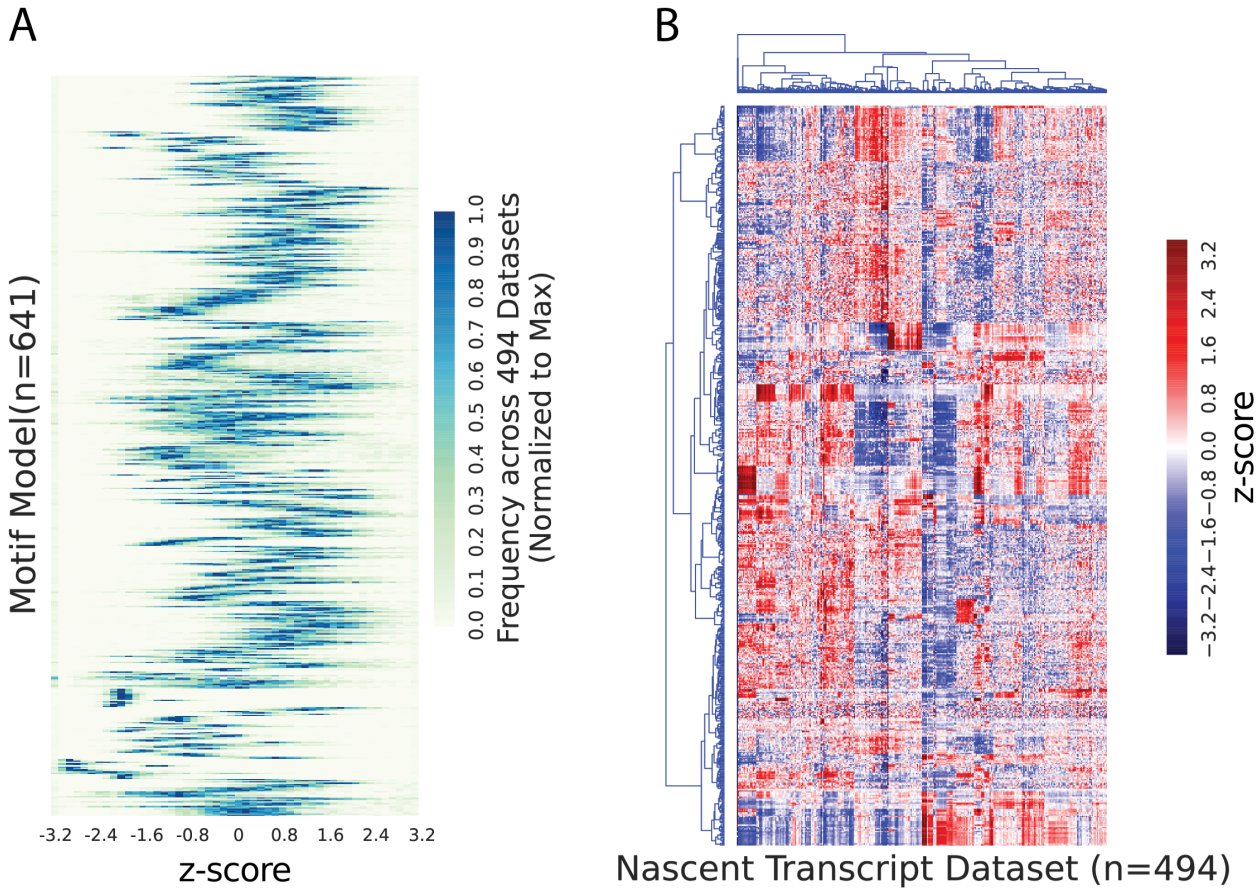
Supplemental Figure 7: **Instances of TF motif and eRNA localization reflect TF binding**. (A) Heatmaps display the frequency of TF motif instances centered at the eRNA origin predicted by Tfit from a K562 GRO-cap (Core et al., 2012) (SRR1552480) experiment. eRNAs were further separated by association with or distal to a TF binding peak (by ChIP). Motif models (Kulakovskiy et al., 2013) and ChIP-matched data sets yielded 57 unique transcription factors and 187 separate peak files. (B) The distribution of estimated RNAP footprint size (distance between forward and reverse strand peaks) for Tfit predicted eRNAs (K562). (C) The co-association of instances of the motif with eRNA origin is elevated at bound sites. MD-score computed from x-axis: eRNAs that are not bound, and y-axis: TF-bound eRNAs.



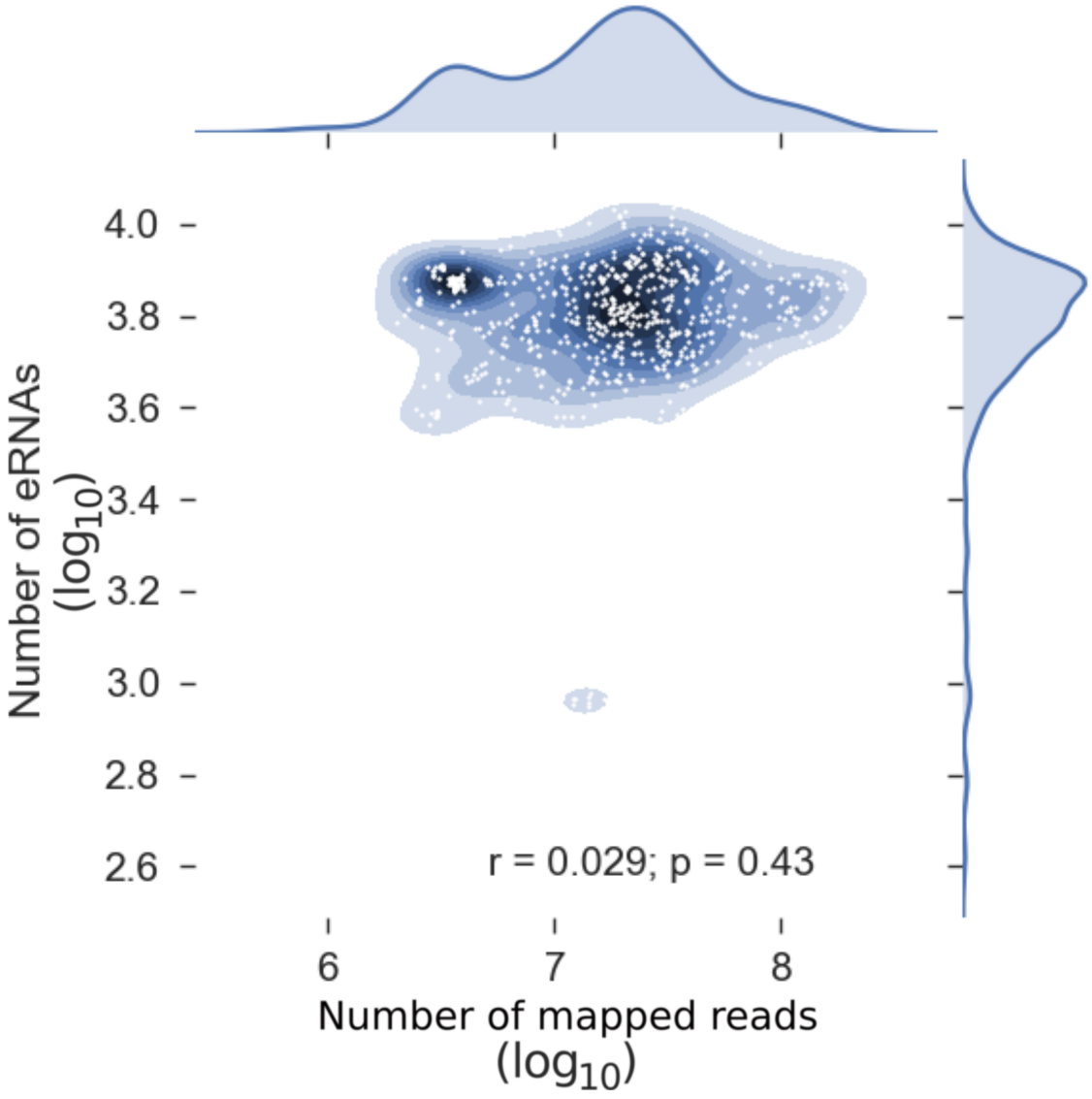
Supplemental Figure 8: **Peak length distribution across ENCODE chromatin mark data.** ENCODE-supplied peaks calls were gathered for ChIP-seq data sets where the antibody was H3K27ac, H3K4me1, H3K4me2, H3K4me3 or H2AFZ, as these correspond to sites of transcription initiation (promoters and/or enhancers). Each bar indicates the average peak length within the specified data set. Across all data sets, average peak width is 2312 bps with a standard deviation of 1750 bps. Cell type of origin noted in parenthesis for each data set.



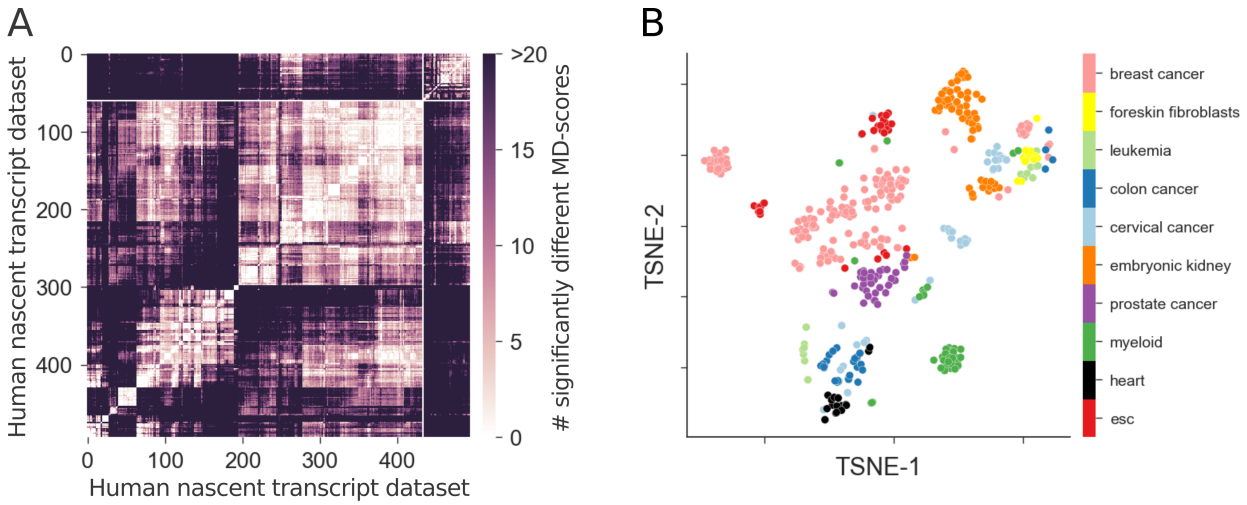
Supplemental Figure 9: **Enhancer GC content bias.** (A) Position specific bias surrounding eRNA predictions. eRNAs were predicted by Tfit from a K562 GRO-cap (Core et al., 2012) (SRR1552480) experiment and a 3kb window (centered at eRNA origin) of sequence from the hg19 human genome build was collected. Background expectation was computed from the entire hg19 genome yielding 24.19%, 25.72%, 24.31%, 25.76% for A, C, T, and G nucleotides, respectively. (B) 10^9 3kb sequences were simulated from the empirical ACGT frequency shown in panel A. The distribution of motif instances (significant PSSM matches $< 10^{-7}$) within simulated data is shown for three demonstrative transcription factors: SP1, ATF3 and MNX1 (as rows). Adjacent to each motif distribution, the associated PSSM in terms of information content (bits).



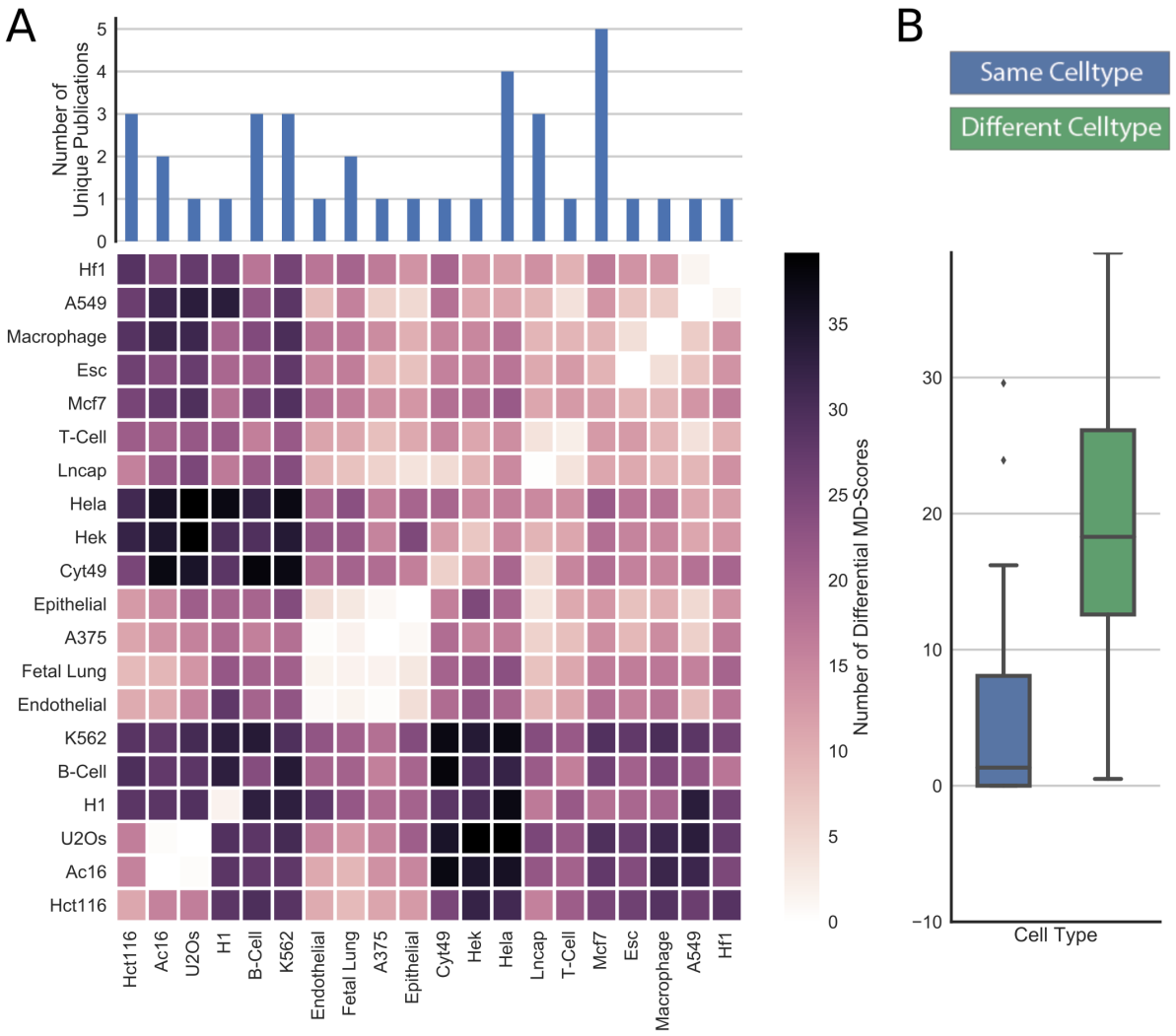
Supplemental Figure 10: **MD-scores display wide variability across all publicly available nascent transcription data sets.** (A) Sites of bidirectional transcription were profiled by Tfit across the full compendium of nascent transcription data sets allowing computation of the 641 (HOCOMOCO) MD-scores. Each row is a single motif model, plotted as the histogram of z-scores (MD-scores were centered by the mean and scaled by the standard deviation). (B) Each row represents a motif model and each column represents a nascent transcription data set. Heat indicates higher MD-scores (relative to the mean). Rows and columns were separately sorted by hierarchical clustering.



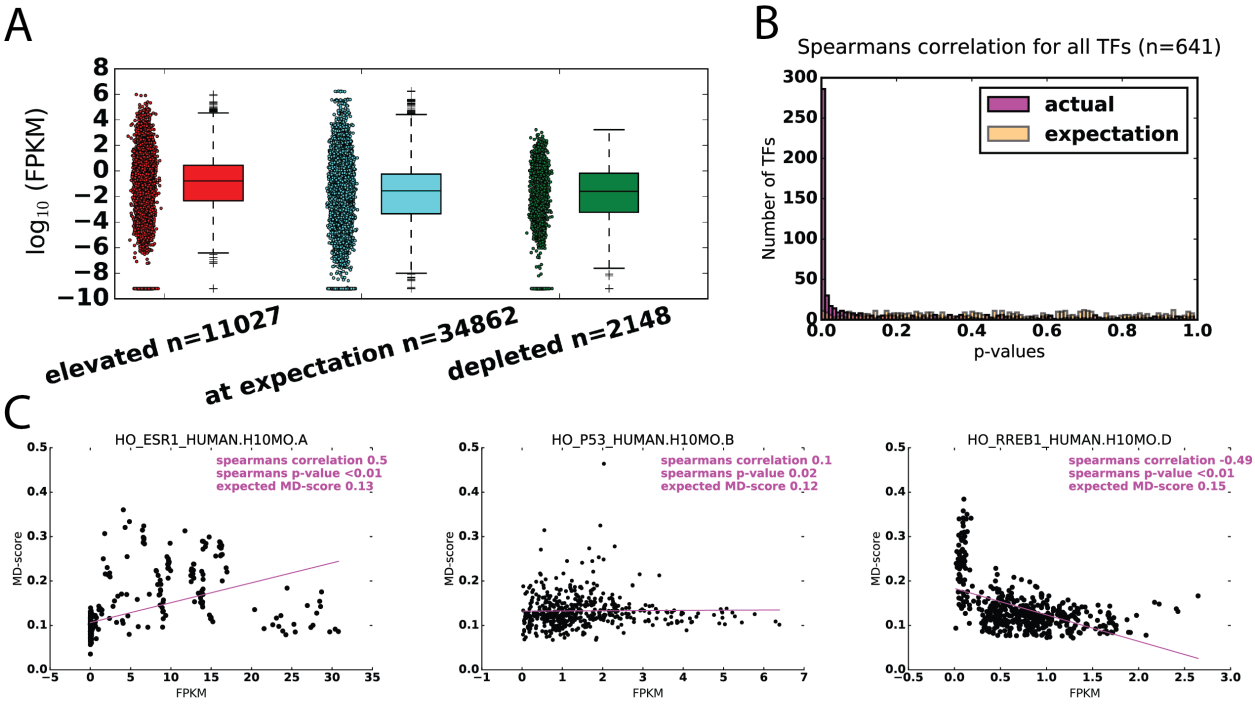
Supplemental Figure 11: **Sequencing depth loosely correlates with the number of predicted eRNA origins by Tfit.** Contour plot of relationship between sequencing depth (x-axis, logarithm (base 10) of number of aligned reads) and number of eRNA origins predicted by Tfit (y-axis, logarithm base 10). Each white cross, indicates a nascent RNA sequencing data set in our study. Contour lines are draw to represent sensitivity as well as the marginal KDE estimates on the top and right panels. Pearson's correlation and the associated p-value (which the null represent $\rho = 0$) are indicated.



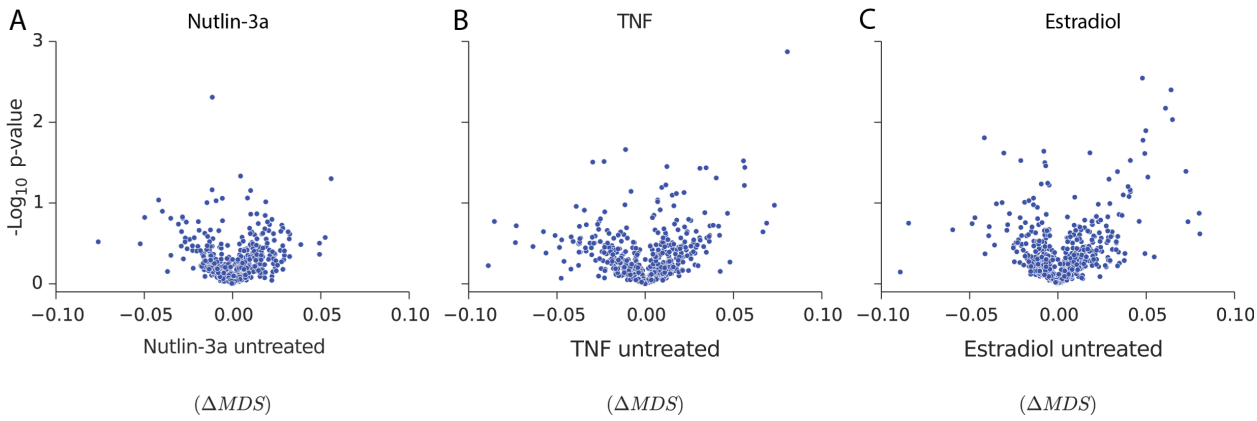
Supplemental Figure 12: **Cell type influences MD-score similarity.** MD-scores were computed for all 491 human nascent transcription data sets in the compendium. (A) Pairwise comparison of each data set shown as a distance matrix where each cell's heat is proportional to the number of significantly different MD-scores ($p(\Delta MDS \neq 0) < 10^{-6}$). Rows and columns are sorted by Ward hierarchical clustering via euclidean distance metric. (B) Dimensionality reduction by t-Distributed Stochastic Neighbor Embedding (TSNE) of the distance matrix in panel A. Only publication annotated cell types with at least ten data sets are shown, each data set is a point.



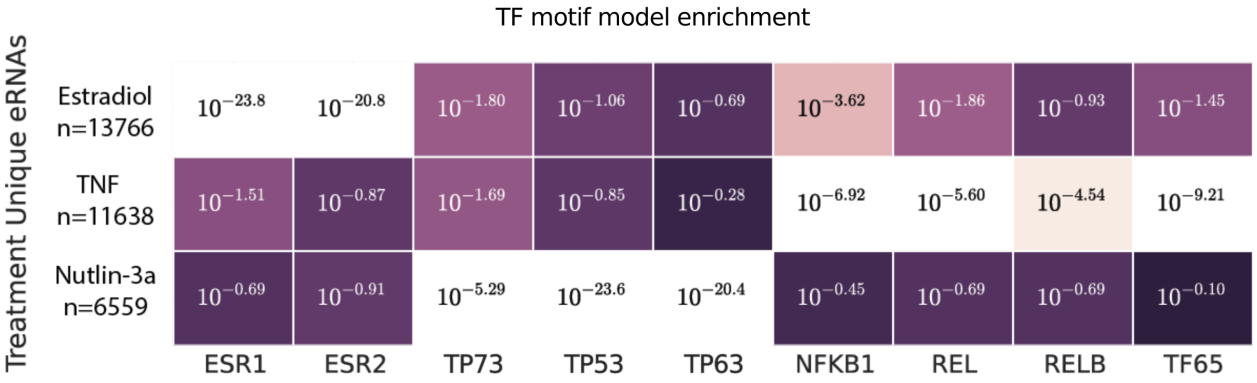
Supplemental Figure 13: **MD-scores within cell types are similar.** Each untreated human nascent transcription data set ($n = 158$) was independently compared and assessed for significant changes in MD-scores (possible comparisons = 12403). (A) Heatmap showing the average number of significantly altered MD-scores (p-value$< 10^{-6}$) between any two experiments that are annotated as the associated cell type. (B) The distribution of the number of significantly different MD-scores grouped by comparison type: same (e.g. ESC to ESC) or different (e.g. HeLa vs LnCAP) cell type. Hypothesis testing on the means of these distributions was performed by the standard t-test. Specific to panel (B), comparisons were only made if the data sets were from different publications.



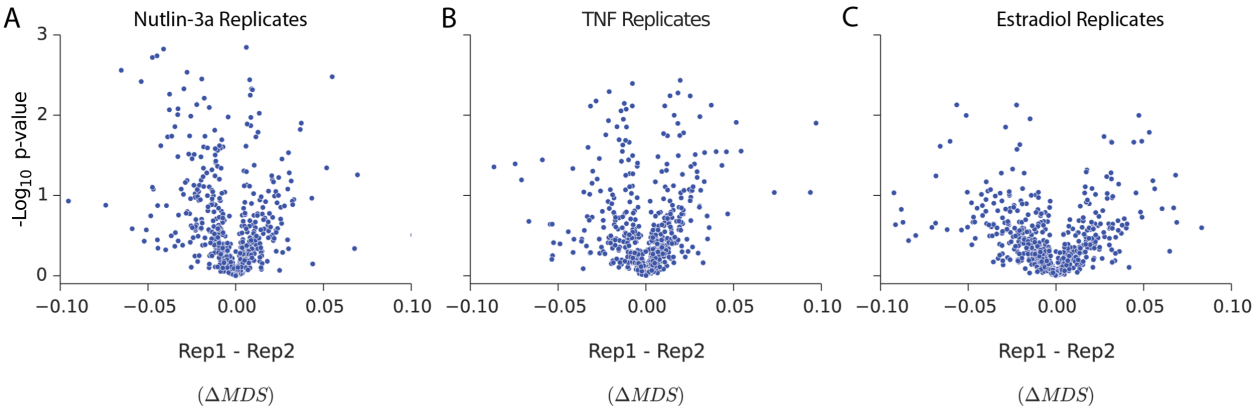
Supplemental Figure 14: **MD-scores and transcription of the gene encoding the TF.** (A) The FPKM of the gene encoding the TF for TFs with MD-scores classified as elevated (red), at-expectation (blue), or depleted (green) are plotted as box-and-whiskers. (B) Spearman rank correlation and associated p-values between FPKM and MD-scores were calculated for each TF across the 491 human nascent transcription data sets. Using a p-value cutoff of 0.01, 286 TFs (magenta) have significant correlation, which exceeds the expectation obtained by permutation testing (orange). (C) Example scatter plots for ESR1, TP53 and RREB1 that show the relationship between TF FPKM and MD-scores. TFs with a strong positive correlation coefficient are likely activators, such as ESR1. TFs with a strong negative correlation coefficient are likely repressors, such as RREB1. However, many TFs do not show strong correlations. For example, in the case of TP53, the TF is known to be post-translationally regulated (Dai and Gu, 2010) and is observed to have high MD-scores only in samples either treated by TP53 activating drugs or under stress.



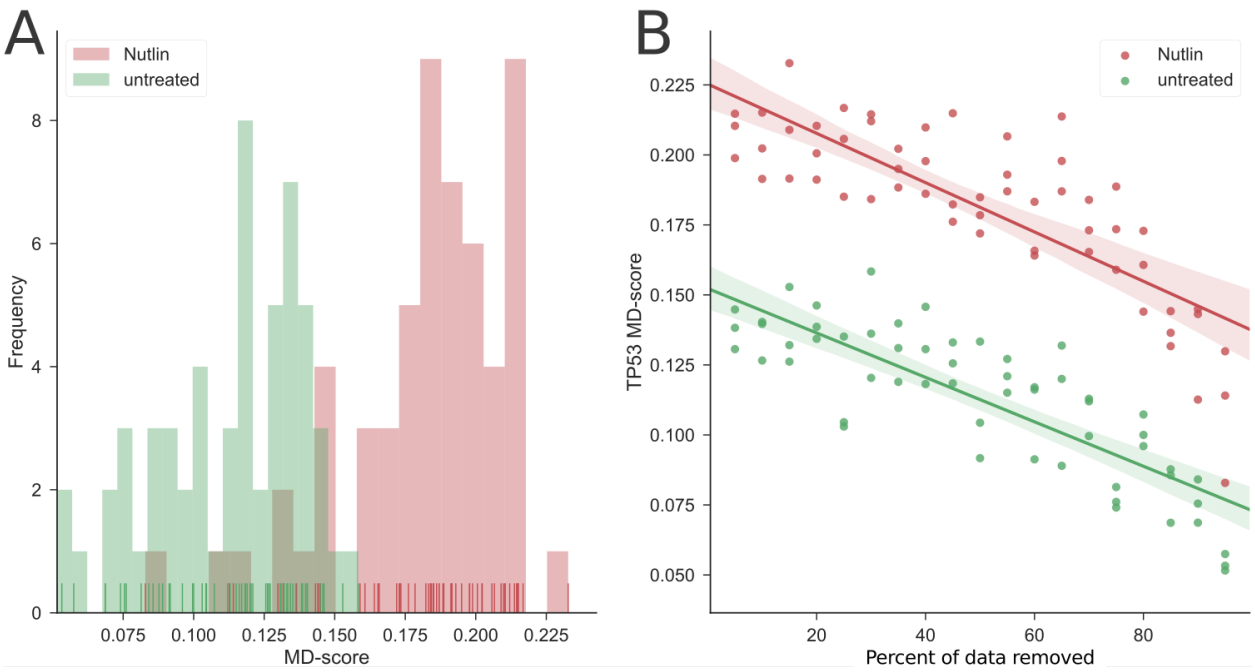
Supplemental Figure 15: **No significant differences in MD-scores is observed when considering only promoter associated bidirectional transcripts**. Experiments annotated as treatment/Untreated pairs: Nutlin-3a (Allen et al., 2014) (SRR1105737, SRR1105739), TNF (Luo et al., 2014) (SRR1015583, SRR1015587) and estradiol (Hah et al., 2013) (SRR653421, SRR653425) were used to study differences in MD-scores following treatment. MD-scores were computed over only promoter associated bidirectional transcripts. Change in MD-score (x-axis) is plotted against the negative $\log_{10}$ p-value (two-tailed proportion test) in MD-score change (y-axis).



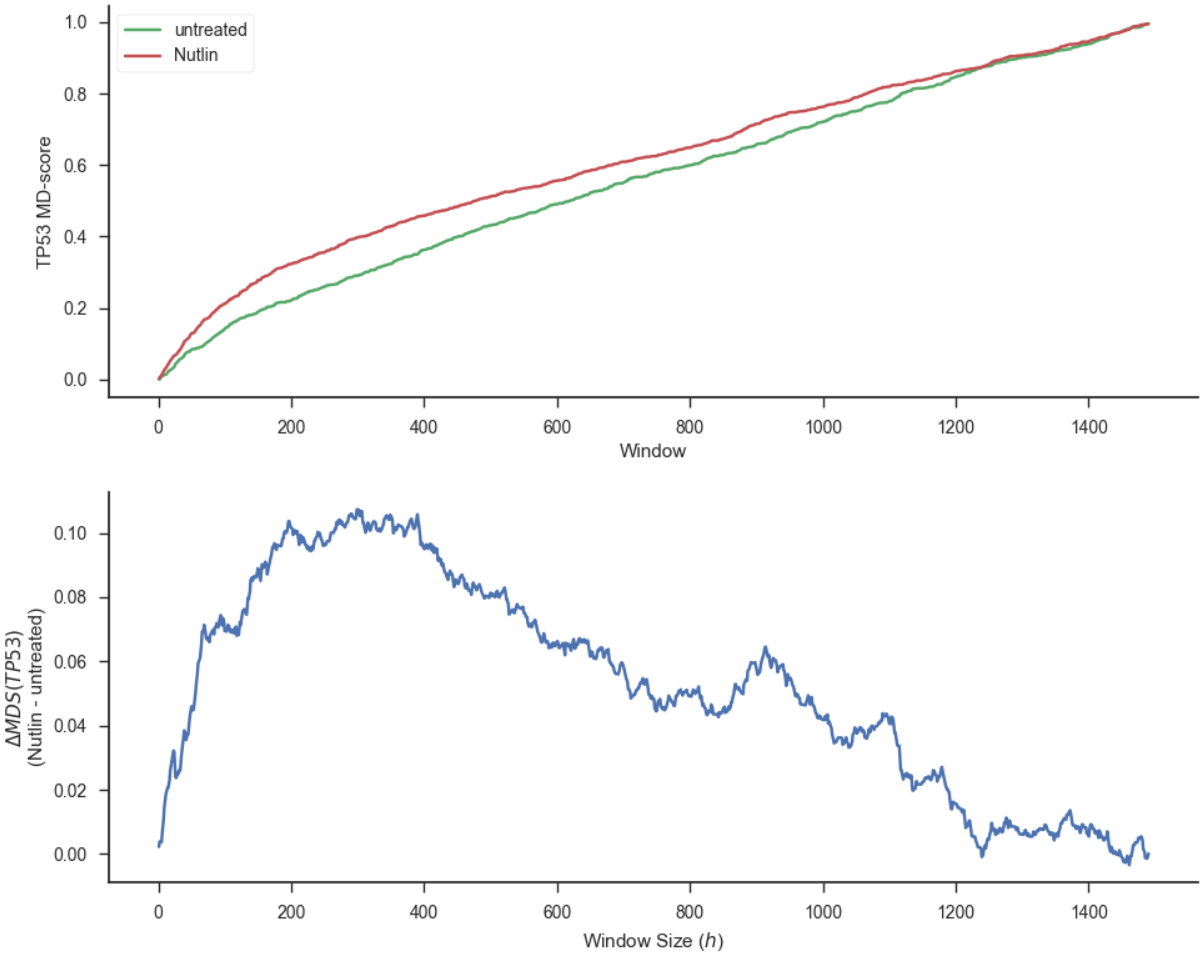
Supplemental Figure 16: **Treatment-unique eRNAs are enriched for specific TF motif models**. Treatment-specific eRNAs were inferred from experiments annotated as treatment/untreated pairs: Nutlin-3a (Allen et al., 2014) (SRR1105737, SRR1105739), TNF (Luo et al., 2014) (SRR1015583, SRR1015587) and estradiol (Hah et al., 2013) (SRR653421, SRR653425). An eRNA is considered treatment-unique if the eRNA origin is not within 1500 base-pairs of an eRNA origin detected within the untreated sample. Motif model enrichment considers only instances of the motif within 150bps of an eRNA origin. Significance of motif enrichment is assessed via a one-tailed hypergeometric distribution. Cell color is proportional to $\log_{10}$ of the p-value where lighter color indicates greater significance.



Supplemental Figure 17: **No significant difference in MD-scores is observed between biological replicates.** Experiments annotated as biological replicate pairs: Nutlin-3a (Allen et al., 2014) (SRR1105738, SRR1105739), TNF (Luo et al., 2014) (SRR1015587, SRR1015588) and estradiol (Hah et al., 2013) (SRR653425, SRR653426) were used to study differences in MD-scores across replicates. Change in MD-score (x-axis) is plotted against the negative $\log_{10}$ p-value (two-tailed proportion test) in MD-score change (y-axis).



Supplemental Figure 18: **Sequencing depth does not affect differential MD-score analysis**. MD-scores are computed for the transcription factor TP53 before (untreated, red) and after Nutlin-3a treatment (green) at increasingly smaller sub-sets of the data. Data removed varied between 5% to 95% at intervals of 5% with three replicates each. (A) Histogram of the subsampled MD-scores shows the mean of these two populations (untreated and Nutlin) are significantly different (p-value $< 10^{-10}$). (B) Relationship between depth and MD-score for each sample. The slopes of the two regression lines are significantly different between untreated and Nutlin conditions (p-value $< 10^{-11.2}$).



Supplemental Figure 19: **Impact of h-radius on differential MD-score analysis**. We varied the h-radius between zero and 1500bps (x-axis) to compute MD-score for the transcription factor TP53 (y-axis, top graph) for both before (untreated, green line) and after Nutlin-3a treatment (red line). The ΔMD-score (blue line, bottom graph) is also shown.

Supplemental Table 4: Nascent transcript data set usage in pairwise comparisons
The purpose of this table is to outline the SRA# data sets utilized in motif distribution comparisons within, between pairs or across all data sets. Specific to Figures 3A-C, Group A and Group B refer to untreated and treatment.

Figure Number	Group A	Group B
3A	SRR1105737	SRR1105739
3B	SRR1015583	SRR1015587
3C	SRR653421	SRR653425
3D	SRR935093	SRR935093 (none) SRR935097 (2 minutes) SRR935101 (5 minutes) SRR935105 (12.5 minutes) SRR935109 (25 minutes) SRR935113 (50 minutes)
3E	SRR930649(DMSO;1 hour) SRR930653(DMSO;6 hour) SRR930655(DMSO;12 hour) SRR930657(DMSO;24 hour)	SRR930659 (KLA;1 hour) SRR930663 (KLA;6 hour) SRR930665 (KLA;12 hour) SRR930667 (KLA;24 hour)

Supplemental Data File 1: Tfit bidirectional predictions

This description corresponds to supplementary files labeled as SRA#.csv within the zipped tar ball "Supplemental Data S1" available at (`http://dowell.colorado.edu/pubs/MDscore`) associated with *Enhancer RNA Profiling Predicts Transcription Factor Activity* by Azofeifa et al. These tables are comma separated files generated for each data set in the Supplementary Table S3 with four columns: chrom, start, stop, tss. The field chrom refers to the chromosome location of the bidirectional origin, the start and stop refer to the genomic location on that chromosome and tss will either return 1 or 0 depending on whether that bidirectional origin overlapped ($\mu < 1,500$) a RefSeq transcription start site annotation. In this way, the set of eRNAs are those bidirectionals where tss $= 0$.

Supplemental Data File 2: Motif Displacement Histograms

This description corresponds to Supplementary files labeled as SRA#.csv within the tar ball name "Supplemental Data S2" available at (`http://dowell.colorado.edu/pubs/MDscore`) associated with *Enhancer RNA Profiling Predicts Transcription Factor Activity* by Azofeifa et al. For each data set in Supplemental Table S3, there exists a comma separated file where the first column refers to motif model ID from HOCOMOCO, the second column refers to whether or not this motif displacement distribution was computed using tss associated or non-tss associated (eRNA) Tfit bidirectional predictions. The final 3001 columns provide the position away from eRNA origin and the number of motifs observed at that position. This constitutes the empirically observed motif displacements histogram for each motif.

References

Allen, M. A., Mellert, H., Dengler, V., Andryzik, Z., Guarnieri, A., Freeman, J. A., Luo, X., Kraus, W. L., Dowell, R. D., and Espinosa, J. M., *et al.*, 2014. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *eLife*, **3**:e02200. doi: 10.7554/eLife.02200.

Allison, K. A., Kaikkonen, M. U., Gaasterland, T., and Glass, C. K., 2013. Vespucci: a system for building annotated databases of nascent transcripts. *NAR*, **42**(4):2433–47.

Azofeifa, J., Allen, M. A., Lladser, M. E., and Dowell, R., 2014. FStitch: A fast and simple algorithm for detecting nascent RNA transcripts. In *Proc. 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 174–183, New York, NY, USA. ACM.

Azofeifa, J. G. and Dowell, R. D., 2017. A generative model for the behavior of RNA polymerase. *Bioinformatics*, **33**(2):227–234.

Bilu, Y. and Barkai, N., 2005. The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biology*, **6**:R103. doi: 10.1186/gb-2005-6-12-r103.

Chae, M., Danko, C. G., and Kraus, W. L., 2015. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*, **16**(1):222. doi: 10.1186/s12859-015-0656-3.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T., 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet*, **46**(12):1311–1320.

Core, L. J., Waterfall, J. J., Gilchrist, D. A., Fargo, D. C., Kwak, H., Adelman, K., and Lis, J. T., 2012. Defining the status of RNA polymerase at promoters. *Cell Reports*, **2**(4):1025–1035.

Dai, C. and Gu, W., 2010. p53 post-translational modification: deregulated in tumorigenesis. *Trends in molecular medicine*, **16**(11):528–536.

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., Cheung, V. G., Kraus, W. L., Lis, J. T., and Siepel, A., *et al.*, 2015. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Meth*, **12**(5):433–438.

Hah, N., Murakami, S., Nagari, A., Danko, C. G., and Kraus, W. L., 2013. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Research*, **23**(8):1210–1223.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D., 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**(17):2204–2207.

Khan, A. and Zhang, X., 2016. dbSUPER: a database of super-enhancers in mouse and human genome. *NAR*, **44**(D1):D164–D171.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B., and Makeev, V. J., 2013. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *NAR*, **41**(D1):D195–D202.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-alawi, W., Bajic, V. B., Medvedeva, Y. A., Kolpakov, F. A., *et al.*, 2016. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *NAR*, **44**(D1):D116–D125.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat Meth*, **9**(4):357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al., *et al.*, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16):2078–2079.

Luo, X., Chae, M., Krishnakumar, R., Danko, C. G., and Kraus, W. L., 2014. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. *BMC Genomics*, **15**:155–155.

Quinlan, A. R. and Hall, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6):841–842.

Staden, R., 1994. *Staden: Searching for Motifs in Nucleic Acid Sequences*, pages 93–102. Springer New York, Totowa, NJ.

Storey, J. D., 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3):479–498.