Stochastic modeling of RNA polymerase predicts transcription factor activity

by

Joseph Gaspare Azofeifa

B.A., Vassar College

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2017

This thesis entitled:
Stochastic modeling of RNA polymerase predicts transcription factor activity
written by Joseph Gaspare Azofeifa
has been approved for the Department of Computer Science

Prof. Robin Dowell

Prof. Aaron Clauset

Prof. Elizabeth Bradley

Date ____________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Azofeifa, Joseph Gaspare (Ph.D., Computer Science)

Stochastic modeling of RNA polymerase predicts transcription factor activity

Thesis directed by Prof. Robin Dowell

Seventy-six percent of disease associated variants occur in non-genic sites of open chromatin suggesting that the regulation of gene expression plays a crucial role in human health. Nucleosome-free with flanking chromatin modifications, these regulatory loci are optimal platforms for transcription binding and, in fact, recruit RNA Polymerase. The subsequent transcription of these sites is an unintuitive discovery as these regulatory loci do not harbor an open reading frame.

The role these enhancer RNAs (eRNA) play in downstream gene regulation remains an open and exciting question. However, fast RNA degradation rates challenge eRNA identification, requiring non-traditional sequencing technologies. Global Run-on followed by sequencing (GRO-seq) detects non-genic transcription and thus, in theory, eRNA presence. Yet GRO-seq is not without noise and bias, predictive modeling of both the sequencing error and the stochastic nature of RNA polymerase itself is required to discover enhancer RNA transcripts.

In short, this thesis asks: what regulates eRNA transcription? To answer this question, I first develop two novel probabilistic models to unbiasedly determine eRNA location. A regression method was constructed to quickly identify all transcribed regions in GRO-seq. Based on the known enzymatic stages of RNA polymerase, a subsequent latent variable model was built to infer the precise location of eRNA initiation. With the relevant technology developed, I undertake a massive data integration project and show strong contextual relationships between TF-binding events, epigenetics and eRNA transcription. I conclude by showing that enhancer RNAs can unbiasedly quantify transcription factor activity and predict cell type.

Dedication

To my family.

Acknowledgements

First, I would like to thank my PhD committee, Dr. Aaron Clauset, Dr. Elizabeth Bradley, Dr. Michael Mozer, Dr. Katerina Kechris and my thesis advisor, Dr. Robin Dowell. Apart from shaping many of the ideas we ultimately published, Robin also trusted in my work-ethic and independence which provided me the necessary freedom I needed to thrive in graduate school. I would like to thank Dr. Manual Lladser for some of the preliminary work on the mixture model and also giving me with confidence to pursue more rigorous mathematics. Dr. Tim Read's endless phone calls helped me deeply in developing the MD-Score and correctly framing the story of eRNAs. I am very indebted to Josephina Hendrix who performed all the short read alignments ($>$ 700 datasets) for the MD-score publication. Finally, I would like to formally thank Dr. Mary Allen. Mary and I sat next to each other for 5 years and I can say, with little doubt, that I learned all my molecular biology through her.

Apart from acknowledgments related directly to my thesis, I would like to thank the IQ Biology program and the BioFrontiers Institute for funding support but also for providing a PhD training program that really fostered an atmosphere of creativity, collaboration and interdisciplinary research. Specifically, I would like to thank Dr. Thomas Cech, Dr. Jana Watson-Capps, Amber McDonell, Kim Little, Kim Kelly and, most of all, Dr. Andrea Stith. I would like to thank Sam Way and Ryan Langendorf who I feel honored to call friends; I hope to continue our lively scientific debates in the future. Finally, I would like to thank my family, Andie and Katherine Azofeifa, without whose support none of this would have been possible. I am not sure my mother realized she'd raise two scientists but I am extremely grateful for such a strong and loving family.

Contents

Chapter

1 Introduction **1**

1.1 Biological Setup 1

1.2 Regulatory Element Identification 3

1.2.1 ChIP-seq 3

1.2.2 Chromatin State and Epigenetics 5

1.2.3 Regulatory Element Identification 6

1.2.4 GRO-seq and eRNAs 9

1.3 Thesis Outline 11

1.3.1 Overview 11

1.3.2 Chapter 2: Transcribed Region Annotation 12

1.3.3 Chapter 3: Stochastic models of RNA Polymerase 12

1.3.4 Chapter 4: eRNA Profiles Predict Transcription Factor Activity 13

2 Transcribed Region Annotation **14**

2.1 Introduction 14

2.2 Nascent Transcript Model 16

2.2.1 Description 16

2.2.2 Parameter Estimation 17

2.2.3 Software Design 20

2.3 Model Accuracy . . . 21

2.3.1 Datasets . . . 21

2.3.2 Sensitivity to depth of data . . . 22

2.3.3 Benchmarking FStitch & Vespucci . . . 23

2.4 Biological Analysis . . . 25

2.4.1 Annotation Comparisons . . . 25

2.4.2 Characterizing bidirectional RNA Activity . . . 27

2.4.3 Differential transcription at annotated genes: a comparison of FStitch to Allen et. al. . . . 29

2.4.4 Differential transcription using all FStitch *active* calls . . . 30

2.5 Conclusions . . . 32

3 Stochastic models of RNA Polymerase **34**

3.1 Introduction . . . 34

3.2 Modeling RNA Polymerase . . . 36

3.2.1 Double Geometric Distribution . . . 36

3.2.2 Exponentially Modified Gaussian . . . 39

3.2.3 Poisson Point Process . . . 47

3.2.4 Mixture Models . . . 49

3.2.5 Bayesian Extensions . . . 53

3.3 Applications to GRO-seq . . . 55

3.3.1 Numerical confirmation of model inference by simulation . . . 55

3.3.2 Predicting enzymatic changes of RNAP following Experimental Perturbation 58

3.3.3 RNAP model accurately predicts marks of regulatory elements . . . 59

3.3.4 Three dimensionally paired loci display centrality and associativity based on bidirectional transcription . . . 61

4 eRNA Profiles Predict Transcription Factor Activity **64**

4.1 Introduction 64

4.2 Enhancer RNAs originate from transcription factor binding sites 65

4.3 Enhancer RNA origins mark sites of regulatory TF binding 67

4.4 eRNA origins co-localize with TF-binding motifs 68

4.5 Motif displacement scores quantify TF activity 70

4.6 MD-scores predict TF activity across cell types 72

4.7 Conclusion 73

5 Looking Forward **76**

5.1 Mixture Models 76

5.1.1 Model Selection 77

5.1.2 Integration of other Data Types 79

5.2 TF Activity Inference Models 80

5.3 Predicting Enhancer to Gene Interactions 82

5.3.1 Network Structure Prediction 82

5.3.2 Correlation Networks 83

5.3.3 Bayesian Networks 84

5.3.4 A 3D Genome 87

5.4 Thesis Conclusions 88

Bibliography **90**

Appendix

A Supplementary Material to Chapter 2 **105**

B Supplementary Material to Chapter 3 **109**

B.1 Seeding the EM 110

B.2 Datasets 111

B.3 CTCF ChIA-PET network construction 112

B.4 Software Package: Tfit 112

B.5 Numerical confirmation of model inference by simulation 112

C Supplementary Material to Chapter 4 **117**

C.1 eRNA origins 117

C.2 Genomic Feature Data Integration 118

C.3 Nascent transcription data processing 118

C.4 Tfit parameters and bidirectional prediction 119

C.4.1 Template Matching 119

C.4.2 EM Algorithm and Bidirectional Origin estimation 121

C.4.3 Footprint Estimation 121

C.5 Computation of Bimodality, ΔBIC 122

C.6 Motif Curation and Motif Scanning 123

C.7 MD-score Hypothesis Testing 124

C.7.1 The Motif Displacement score 124

C.7.2 MD-score significance under stationary model 125

C.7.3 MD-score significance under a non-stationary background model 125

C.7.4 MD-score significance between experiments 126

C.8 Cell type and TF enrichment analysis 127

C.9 Associated File Types 128

C.10 IPython Notebook 129

Chapter 1

Introduction

1.1 Biological Setup

A central goal in genetics is to understand how genotype (the unique ordering of DNA) translates to phenotype (observable qualities like height or eye color). Although some phenotypic traits are innocuous, one's genotype may influence cancer susceptibility, a predisposition to alcoholism or cognitive disabilities[47, 21, 62]. For advancements in human medicine—as well as a fundamental understanding of biology—genetics remains an exciting and active area of research.

A long way from Mendel's pea plants, whole genome sequencing makes possible the complete identification of an organism's genotype. Resolving the human genome's nearly 3.2 billion nucleotides, we now know that alterations in the gene sequence of *p53*, *kvlqt1* and *adam19* correlate with incidences of cancer, Type II diabetes and heart disease respectively[14, 173, 183]. Although genome-wide association studies (GWAS)[25] successfully link genotypic variants to phenotype, they require hundreds or even thousands of genomic samples to achieve significant correlations[88]. Yet furthermore, GWAS is unable to predict the phenotypic consequences of a novel genetic variant, unassociated with a specific phenotype. In contrast, the study of gene expression—the biochemical or molecular process by which a genotype renders a phenotype—promises to uncover *why* certain genotypes result in specific phenotypes.

To summarize briefly, gene expression begins with the enzyme RNA polymerase (RNAP) *transcribing* a messenger RNA molecule. Ribosomes may *translate* this messenger RNA into a chain of amino acids that fold into a functional unit called a protein, culminating in an observed phenotype.

As predicted by this "central dogma," phenotypically apparent genotypes occur within gene bodies (e.g. an amino acid substitution may render a new or non-functional protein product). However, the surprising fact remains: 76% of GWAS implicated loci reside within un-annotated, non-genic portions of the human genome[112]. Although previously dismissed as "junk" DNA, these cryptic regions harbor sequence motifs crucial to the *regulation* of gene expression by transcription factors (TFs).

As a heritable phenotype, population variability in gene expression may arise under two scenarios: a *trans*-effect where a coding mutation renders a change in protein activity or a *cis*-effect where, for example, a mutation eliminates TF-binding at specific locus. A common *cis*-element is a gene's promoter: a particular DNA sequence immediately upstream the coding region that initiates transcription. Yet another are enhancers: loci distal (≈100-10,000 nucleotides) from a gene's promoter.

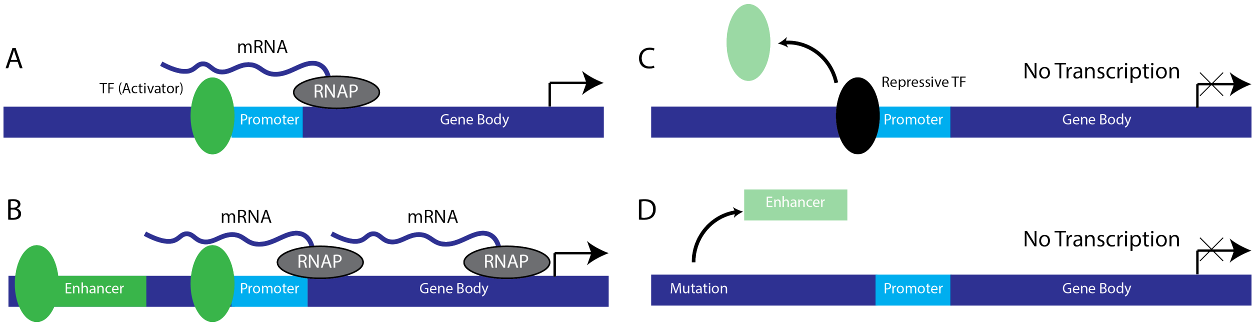


Figure 1.1: **Transcriptional Regulation by Enhancers and TFs** Panels A and B show increased transcription at some hypothetical gene by either the presence of a TF or an enhancer element respectively. (C) At promoters, loss of TF binding or the presence of TF-repressor decreases transcript output. (D) Similarly loss of an enhancer element via DNA mutation may result decreased target gene expression.

Enhancers, appropriately named, *enhance* transcription. Classical assays would artificially copy and paste these sequences near genes lacking an enhancer and observe strong increases in RNA production of that gene[22]. Only to complicate matters, there is no 1-1 mapping of enhancers to genes: one enhancer may regulate many genes and many enhancers may regulate one gene[134].

In this way, enhancers add another dimension to gene regulation, providing platforms for TFs to finely attenuate transcript levels on a larger scale (Figure 1.1).

Although crucial to organismal homeostasis[134], the research concerning enhancers has largely focused on single locus studies[133]. With the human genome's search space of 3.2 billion base pairs, the identification of candidate enhancer loci poses a significant challenge. And even with a promising locus, functional reporter assays correctly confirming enhancer activity are slow and laborious. Yet the advent of high throughput sequencing technology allows measurement of thousands of biomolecules in a signal experiment including enhancers. Even still, noise inherent to sequencing technology requires the development of computationally tractable models to identify these regulatory elements.

1.2 Regulatory Element Identification

To say without hesitation, genome sequencing has profoundly altered our view of biology. Popular in the early 2000's, an organism's genome sequence was approached simply as a code-to-be-cracked. Yet low estimates of protein coding genes ($\approx$ 20K genes or 1.9% genome content) makes explanation of the phenotypic consequences of an altered genotype difficult[82] which suggests that changes in gene regulation may explain a major portion of phenotypic diversity[87]. In large part, this propelled the assembly of a large consortium: the Encyclopedia of DNA Elements (or ENCODE)[30]. According to the most recent ENCODE release, the consortium's principle goal is to "...to build a comprehensive list of functional elements in the human genome, including regulatory elements that control cells and circumstances in which a gene is active." This section will serve in part to outline each datatype ENCODE collected, the statistical efforts used to quantify signal and how it all relates to enhancer identification.

1.2.1 ChIP-seq

Supported by an enrichment of disease associated variants surrounding regulatory elements, a hallmark of promoter and enhancer regions are *conserved* sequence features. For example, an alternating sequence of nucleotides (T,A,T,A) defines the so called promoter TATA-box found in roughly 10% of human promoter sequences. Early studies ablated the TATA sequence and showed loss of TFIID and RNA polymerase binding and a subsequent expression decrease in the downstream gene[146].

Current studies estimate 1,400 transcription factors encoded by the human genome[170]. And although they exist in transcription factor "families," different TFs bind separate DNA patterns. Referred to commonly as motifs, specific orderings of DNA (e.g. ...$ACTGGGAA$... vs ...$CTAGCCCGGGCATG$...) retain lower or higher affinities for different transcription factors (e.g. AP1 and P53 respectively).

Even still, transcription factors bind DNA in a stochastic manner: there is no single rule or specific ordering of DNA that governs where a TF will bind. The Position Weight Matrix (PWM)[107] model summarizes TF-DNA binding motifs as an ordered categorical distribution over $\{A, T, G, C\}$ [159]. Figure 1.2 displays the PWM for the transcription factor p53.

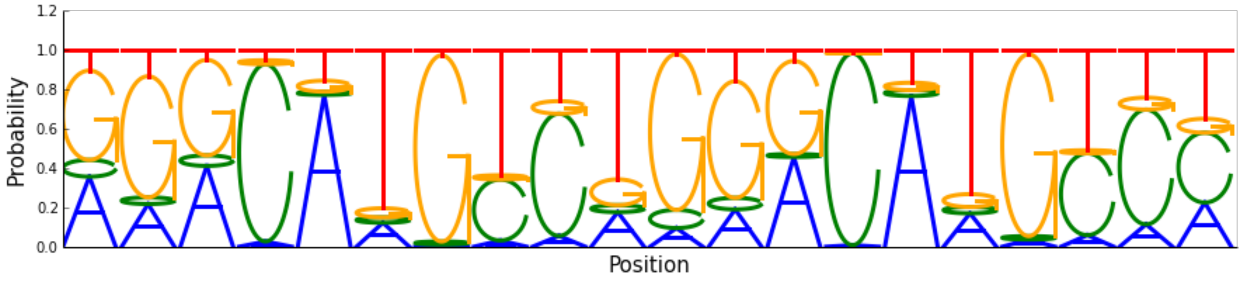


Figure 1.2: **Position Weight Matrix of the TF p53** The height of the letter indicates the probability of observing that nucleotide (y-axis) at that specific position (x-axis).

The genome may be efficiently scanned for high probable motif matches by computing a likelihood ratio between the PWM model, M, and some background model B. If $S = \{s_0, s_1, ..., s_m\}$ is an ordered set of nucleotides then the log-likelihood of S under M and B are given by equation 1.1.

$$L_M(S) = \sum_i \log(M(i, s_i)) \quad L_B(S) = \sum_i \log(B(s_i)) \tag{1.1}$$

M is a m by 4 matrix that encodes the probability of observing nucleotide j at position i and B is a vector of probabilities of observing nucleotide j independent of position (1.2). Our test statistic, $2(L_M(S) - L_B(S))$, is poorly approximated by a $\chi^2(n)$ and thus efficient dynamic programming methods have been developed to compute the 4^m possible likelihood ratios[158].

$$M = \begin{bmatrix} p_{1,a} & p_{1,c} & p_{1,g} & p_{1,t} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m,a} & p_{m,c} & p_{m,g} & p_{m,t} \end{bmatrix} \quad B = \begin{bmatrix} p_a & p_c & p_g & p_t \end{bmatrix} \tag{1.2}$$

Even at a p-value cut off of 10^{-5}, we expect over $\approx 64,000$ putative p53 binding sites across the human genome. In general, studies observe little overlap between verified TF-binding sites with high confidence motif sequences[23]. And although these sites might be bound within a different cellular context (cell type, genotype backgrounds), prediction of true TF-binding events from genomic sequence motifs alone is challenging. Yet research has focused on translating the generative PWM model into a discriminative classifier utilizing popular machine learning algorithms like linear discriminate analysis [182], support vector machines [12] and gaussian processes[149]. Such approaches have observed marginal improvements in TF binding prediction.

To identify TF binding sites within a specific cellular background, experimental assays, like Chromatin Immunoprecipitation followed by sequencing (or ChIP-seq), can capture physical protein-DNA interactions genome-wide. To summarize the experimental protocol briefly, formaldehyde is used to *cross-link* or irreversibly bind all proximal proteins and DNA sequences. DNA is then *sheared* by sonication leaving only the protein-DNA complexes. Antibodies engineered for a protein of interest isolate and *pull-down* only that DNA-bound protein and, along with it, the associated DNA. The DNA-proteins are unlinked and the protein is washed away. Under a noise free model, the remaining DNA results only from sites bound by the TF of interest. Via high throughput sequencing platforms, this DNA is mapped backed to the genome and thus provides the promised histogram of DNA-bound protein along the genome (Figure 1.3).

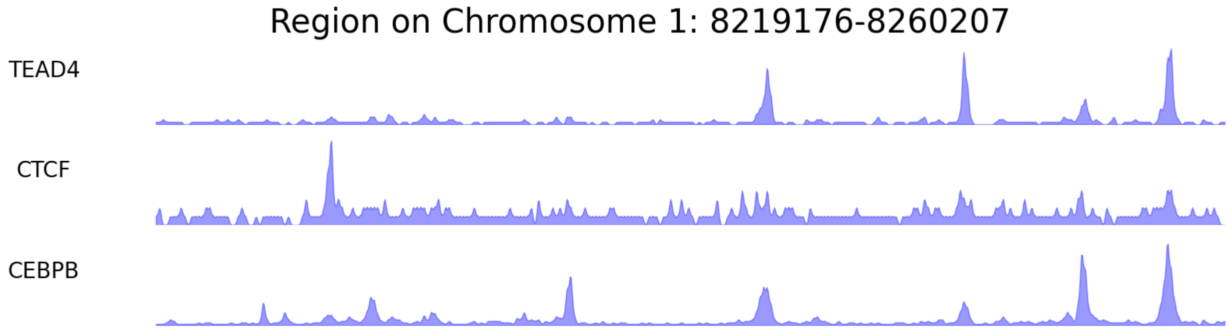


Figure 1.3: **ChIP-seq from demonstrative ENCODE experiment** The y-axis indicates the TF-relative frequency (number of mapped reads).

1.2.2 Chromatin State and Epigenetics

Regulatory elements are characterized by more than just transcription factor binding. Although Figure 1.1 provides a useful abstraction, the genome is not everywhere homogenous but wrapped around proteins, called histones, that serve as yet another layer of regulation[13].

Chromatin—the mesh complex of histones and DNA—itself is altered to provide more or less conducive platforms for TF binding. Specifically, chemical modifications to a histone changes the local distribution of electrostatic charge, either pushing neighboring histones together (inaccessible to TF binding) or away from one another (open to TF binding). In this way, a high probable motif sequence may remain unbound by a TF in a highly condensed chromatin context. There exist roughly 60 verified histone modifications and, like TF-binding assays, the frequency of any modification can be measured on a genome wide scale by ChIP-seq[130].

As an example, acetylation of the 27th lysine residue on the third histone (H3K27ac) is found ubiquitously across all actively engaged regulatory elements (promoters and enhancers alike)[38]. Yet, H3K4 mono-methylation is found in large abundance at enhancer elements but lack signal at promoter regions. Previously observed in promoters (and later at enhancer loci in chapter 4), histone modification signal shows clear bimodality and suggests a complete lack of modified histones near the center of regulatory elements. These so called "Nucleosome Free Regions (NFRs)" (where a nucleosome is a collection of eight histones) are a hallmark of active regulatory elements. Portions of the DNA lacking in tight histone packing are accessible to recognition of specific transcription factors and thus targeted regulation.

In brief summary, promoters and enhancers are marked by conserved sequence features (DNA motifs), bound proteins (transcription factors) and epigenetic modifications. Although there exists a straightforward evolutionary argument for conserved TF-DNA binding motifs, one might wonder over the mechanism underlying the deposition of histone modifications. Although enzymes such as histone acetyl and de-acetyl transferences are extensively studied[20], little is known as to their targeted specificity. To summarize, these DNA-elements, when modified, regulate the transcript levels of certain genes, the concentration of downstream protein product and the final observed phenotype.

1.2.3 Regulatory Element Identification

Although a ubiquitous technique, ChIP-seq displays a high level of noise. Within the *coss-linking* step, "proximal" may refer to not only bound transcription factors but any TFs within a potentially large radius. Yet another source of noise, if the antibody used to pull down the TF-of-interest is not completely specific, ChIP-seq signal may also emanate from other DNA-bound proteins. For these reasons, statistical modeling has focused primarily on very stringent false positive filters.

Seen prominently in Figure 1.3, mapped read coverage is non-stationary. In general, reads "peak" at critical points indicating a lack of entropy and physical DNA-TF interaction. These critical points tend to be over enriched for TF binding motifs characteristic of functional *cis*-elements. Referred to as *peak identification*[180], Model-based Analysis of ChIP-Seq data (MACs[180]) utilizes a Poisson model of read coverage to assess local (size $2d$ window) enrichment of aligned reads.

Summarized in equation 1.3, y_i refers to the number of mapped reads at genomic coordinate i.

$$\lambda_i = \sum_{j=i-d}^{i+d} y_j \qquad \lambda_i \sim \text{Poisson}(\lambda_{background}) \tag{1.3}$$

At some p-value cutoff (common 10^{-6}), neighboring coordinates with over enriched aligned reads are merged into one single call.

An extremely simple and intuitive computation, MACs inherently assumes knowledge of the correct d, $\lambda_{background}$, p-value cut-off and independence of neighboring genomic coordinates. Although some of these parameters may be estimated from the experiment itself, encoding a level of probabilistic dependence between neighboring genomic coordinates requires a fundamentally different method of segmentation. By in large, the most popular method for genomic analysis of short read alignment data are first-order Hidden Markov Models (HMMs)[49].

In most Markov contexts, state changes are indexed by time. Yet under a genomic background, Markov models evolve over and are indexed by position along the genome (i). Apart from capturing local conditional dependencies between i and $i+1$, HMMs provide nice polynomial time algorithms (Viterbi & Forward-Backward[16]) for computing the MLE sequence of state transitions under a fully specified HMM.

"Hidden" Markov models are so named because we do not actually observe the state sequence $\{x_1, x_2, ..., x_i, x_{i+1}, ..., x_n\}$, but some random variable indirectly characteristic of those states, $\{y_1, y_2, ..., y_s, y_{i+1}, ..., y_n\}$. Here, $x_i \in \{1, 2, ..., K\}$ refers to one of K states at genomic coordinate i. Regardless of interpretation, the likelihood of some state sequence X is easily computed given some set of aligned reads Y (equation 1.4).

$$p(x_{1:n}, y_{1:n}) = p(x_1)p(y_1|x_1)\prod_{s=2}^{n} p(x_s|x_{s-1})p(y_s|x_s) \tag{1.4}$$

In general, most biologists or modelers do not argue over the topology of the Markov chain: states like "enhancer locus" or "TF-binding" are all qualities of the data we wish to identify. Most disagreement falls under how to parameterize and model the probability of observing some specific count y_i conditioned on i, $p(y_i|x_i = k; \theta_i)$. Popular forms of this *emission* probability function are provided below.

$$p(y_i|x_i = k) = \begin{cases} \lambda_k^{y_i} e^{-\lambda_k}/y_i! : \text{Poisson[144]} \\ \binom{y_i+r_k+1}{y_i} p_k^{y_i}(1-p_k)^{r_k} : \text{Negative Binomial[7]} \end{cases} \tag{1.5}$$

Historically, the Poisson distribution models count data, yet Poisson modeling suffers in situations of statistical *over-dispersion* as both the expected value and variance of y_i is parameterized by signal constant λ. A negative binomial with parameters (p, r) fits nicely as a solution to over dispersion as the expected value and variance are decoupled: $E_{NB}[y_i] = \frac{pr}{1-p}$ and $Var_{NB}(y_i) = \frac{pr}{(1-p)^2}$. Although, neither the Poisson or negative binomial formulation are able to learn non-linear separability observed in other forms of high throughput datasets[10].

Although I devote a significant amount of attention to Hidden Markov models in Chapter 2, functional genomic datasets are not collected over time or space but are *sampled.* In short, the random variable of interest is not the height of the histogram bar at x_i but the probability of *observing* the genomic coordinate x_i. Owing to the memoryless property of first-order Markov chains, the statistical non-stationarity evident in high throughput sequencing datasets cannot be estimated in any meaningful way by an HMM. And in fact, traditional segmentation analysis will group nearby peaks into one long classification[5].

Mixture modeling[43] seems like a nice answer to peak identification. In the finite setting, $p(x_i)$ is broken into a sum over K distributions, each comprising their own set of parameters θ_k. The generative framework selects component k from a Multinomial($\vec{\pi}$) and draws x_i from $p(x_i|\theta_k)$.

$$p(x_i) = \sum_{k \in K} \pi_k \cdot p(x_i|\theta_k) \tag{1.6}$$

Also referred to as a latent variable model, we observe only x_i and not necessarily the component k to which x_i belongs. Given the likelihood function (equation 1.7) emits no analytic solution to the optimal Θ^*, gradient based algorithms such as Expectation Maximization (EM) as well as MCMC methods like Gibb's sampling (in the case of well defined priors) may be used to perform inference over $p(k|x_s)$.

$$\mathcal{L}(\Theta|Y[a..b]) = \prod_{i \in [a..b]} \Big[\sum_k \pi_k \cdot p(x_i|\theta_k) \Big]^{y_i} \tag{1.7}$$

This thesis devotes the entirety of Chapter 3 to mixture models and their use in high throughput sequencing datasets.

1.2.4 GRO-seq and eRNAs

Immediately following the development of ChIP-seq, antibodies specific to RNA polymerase II were engineered and RNAP binding was mapped genome-wide. As expected, RNA polymerase localized at promoters and gene bodies, yet peaks of RNAP-ChIP signal were observed at non-genic loci such as enhancers[85]. Such a result was (and remains) unintuitive: enhancers do not contain a nearby open reading frame and thus no template for a downstream protein product, so why should RNAP and enhancers co-localize?

Initially, this result was dismissed as an artifact of either the formaldehyde cross-linking or immunoprecipitation step integral to a noisy ChIP-seq experiment. Further evidence of technical noise, mRNA profiles derived from RNA-seq did not indicate significant levels of transcriptional output from non-promoter associated regulatory loci. Parallel to these RNAP-ChIP experiments, nuclear run-on assays were being used to study RNAP transcriptional rates in the yeast model organism *S. cerevisiae*[61]. A later high throughput method, global run-on followed by sequencing (GRO-seq)[36], began to detect transcript signal at enhancers (Figure 1.4).

Enhancer RNAs (or eRNAs) currently remain a confounding discovery. Although long non-coding transcripts (lncRNAs) are ubiquitous throughout the human genome ($\approx 50\%$ of RNA-seq signal emanates from non-coding genic loci in humans), lncRNAs retain sequence features more closely resembling a genic mRNA (5-prime capping, splicing, and poly-adenylation) [81]. On the other hand, eRNAs are short nascent transcripts that are unstable and actively degraded by the nuclear exosome[24]. For these reasons, RNA-seq—an assay that isolates and sequences all mature messenger RNA—is unable to detect eRNA presence given their quick degradation.

Methods of quantification, like RNA-seq, profile *steady-state* levels of mRNA where the concentration of mRNA is a function of both transcription *and* degradation rates. On the contrary, GRO-seq sequences nascent mRNA transcripts resulting from the activity of all cellular polymerase

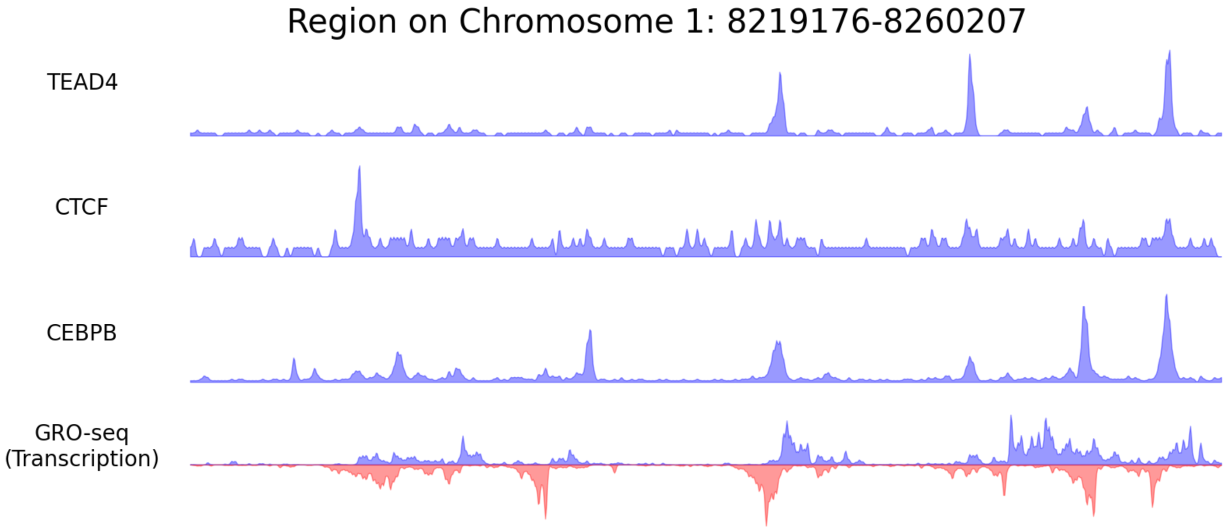


Figure 1.4: **ChIP-seq from demonstrative ENCODE experiment** The y-axis indicates the relative frequency (number of mapped reads) of either TF (top 3 panels) or nascent transcription bottom panel. GRO-seq, unlike ChIP-seq is a stranded procedure. Therefore, forward and reverse strand mapped reads are indicated by blue or red coloring respectively. As a further visual aid, reverse strand aligned reads are reflected across the x-axis (but of course these $y_i|$reverse strand is a non-negative value.)

activity, theoretically prior to RNA degradation.

To summarize the experimental protocol briefly, nuclei are isolated, frozen and starved of nucleotides effectively pausing RNAP. After administering Sarkosyl to prevent loading of other RNAP molecules, cells are then incubated with Bromouridine-triphosphate (BrUTP) and RNAP resumes transcription (run-on). With the high concentration of BrUTP, all RNA molecules with the incorporated BrUTP nucleotide result from actively transcribed RNAP. With anti-BrUTP beads, these transcripts are isolated, sequenced and mapped back to the genome, providing a unique readout on not only RNA polymerase dynamics but bona-fide transcription before RNA degradation. As an important note, GRO-seq is but one assay in a large family of nascent transcript sequencing protocols that differ only in how RNA polymerase is paused and how subsequent transcripts are pulled down.

Illustrated plainly in Figure 1.4, GRO-seq signal correlates well with TF binding events. Importantly however, peaks of mapped GRO-seq reads are not distributed like their TF-ChIP counterpart. In particular, reads aligning to the forward strand display positive skew radiating away from the TF binding event and conversely reads aligning to reverse strand appear negatively skewed. Marked by this "bidirectional" transcription, accurate identification of eRNA transcription will rely heavily on this signature.

1.3 Thesis Outline

1.3.1 Overview

Regulatory loci are vital to homeostatic gene expression. Marked by transcription binding, chromatin modifications and disease associated genetic variants, enhancers also display a new and novel feature: transcription. To understand the functional implications of this new class of regulatory elements, this thesis will build novel probabilistic models of GRO-seq and RNA polymerase to predict sites of eRNA transcription (chapters 2 and 3). These models will

(1) comprise a set of **biologically interpretable** parameters, rooted in the known behavior of RNA polymerase.

(2) guarantee **unbiased and low variance estimators** even in light of latent variables and non-exponential family distribution functions.

(3) be coupled to **computationally tractable algorithms** for parameter estimation; written for massively parallel platforms.

Subsequent to model specification, I will perform inference into eRNA location across all publicly available GRO-seq datasets (as of August 2016). From this large annotation project (chapter 4), we will learn

(1) that there exists a strong relationship between **TF binding** events, eRNA location and target gene expression.

(2) that the **co-occurrence of TF motif and eRNA location** alone can predict transcription factor activity.

(3) a set of transcription factors specific and **predictive of cancerous cell types**.

1.3.2 Chapter 2: Transcribed Region Annotation

Enhancer RNAs are most readily detected in nascent transcript sequencing data. In a noise free setup, each sequenced read originates from an actively transcribing RNA polymerase molecule. Yet in reality, many sources of the GRO-seq protocol influence read mapping errors. For example, non-nascent transcripts may be sequenced and mapped due to fragmentation biases and non-specific anti-BrUTP bead binding. More so, GRO-seq is a population based assay where collections of cells are very likely heterogeneous. Although heterogeneity may be smoothed by replications, the final GRO-seq dataset comprises noise both from technical as well as biological sources.

To annotate portions of the genome experiencing contiguous levels of nascent transcription, I developed a hidden Markov model (HMM) coupled to a logistic regression classifier. With flexibility afforded by a logistic model, I was able to embed our features into a higher dimensional space to learn non-linear decision boundaries. Benchmarking against other HMMs and window-based approaches showed that my method, Fast Read Stitcher (or FStitch), yielded the highest level of precision and recall.

Following model validation, we looked at how nascent transcripts—predicted by FStitch—change following Nutlin treatment: a drug known to activate the transcription factor p53. We discovered a significant increase in nascent transcripts overlapping known ChIP-p53 binding sites as well as chromatin marks such as H3K27ac, H34Kme[1/3]. Indeed, 81% of these induced nascent transcripts were non-genic associated indicating that a set of p53-specific enhancers were being activated. Traditional motif discovery and enrichment analysis, however, did not recover the p53 motif from these induced nascent transcribed regions.

1.3.3 Chapter 3: Stochastic models of RNA Polymerase

Transcription by RNA polymerases is a highly dynamic process involving multiple distinct stages of RNAP behavior. Although an HMM framework provides a straightforward noise filter, it is prone to over segmentation and assumes a single model genome-wide. In an effort to compensate for model degeneracies noted in Chapter 2, I built a latent variable model that infers—at signal base resolution—sites of RNAP loading and as such deconvolves individual nascent transcripts.

Within this chapter, I develop the theory and parameterization of mixture models. I present a generative, probabilistic model of RNA polymerase that fully describes loading, initiation, elongation and termination. I fit this model genome wide and profile the enzymatic activity of RNA polymerase across various loci and following experimental perturbation. I observe striking correlation of predicted loading events and regulatory chromatin marks. I provide principled statistics that compute probabilities reminiscent of traveler's and divergent ratios. I finish with a systematic comparison of RNA polymerase activity at promoter vs non-promoter associated loci.

1.3.4 Chapter 4: eRNA Profiles Predict Transcription Factor Activity

Transcription factors (TFs) exert their regulatory influence through the binding of enhancers, resulting in coordination of gene expression programs. While the functions of eRNAs remain unclear, here I show that they represent a powerful readout of transcription factor activity.

In this chapter, I identified all eRNA start sites across hundreds of publicly available global run on and sequencing (GRO-seq) datasets and showed that TF binding events co-occurring with eRNA transcription are more likely to positively influence gene expression than those that do not. By quantifying the distribution of TF binding motifs relative to sites of eRNA initiation, I derived a simple statistic capable of gauging the activity of all transcription factors, simultaneously, from a single experiment. I demonstrate how this approach can be used to indiscriminately infer the transcription factors whose activities are influenced by cellular perturbations and in doing so uncover dozens of previously unexplored links between diverse stimuli and the transcription factors that they affect.

I finally show that TF activity is extraordinarily dynamic and represents an extremely sensitive predictor of cell type. This approach constitutes a fundamentally unique strategy for discovering links between TF activity and biologically meaningful phenotypes.

Chapter 2

Transcribed Region Annotation

Portions of this chapter are adapted from:

> **J. Azofeifa**, M. Allen, M. Lladser and R. Dowell (2014) *FStitch: A fast and simple algorithm for detecting nascent RNA transcripts* Conference: 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics
>
> **J. Azofeifa**, M. Allen, M. Lladser and R. Dowell (2015) *An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-seq* IEEE/ACM Transactions on Computational Biology and Bioinformatics

2.1 Introduction

Almost all cellular stimulation triggers global transcriptional changes. To date , most studies of transcription have employed RNA-seq or microarrays, powerful measures of steady state RNA levels. Unfortunately, steady state levels can be influenced by not only transcription but also RNA stability, so these assays are not true measures of transcription. Only recently have methods for direct measurement of transcription, genome-wide, become available. A technique, known as global run-on sequencing (GRO-seq), simultaneously detects the amount and direction of actively engaged RNA polymerases at every position within the genome[33]. GRO-seq has already drastically influenced our understanding of the transcription process, as most of the genome is transcribed but rapidly degraded [80, 126, 38].

The earliest and most common approach to GRO-seq analysis is annotation centric[33, 118, 96, 74]. Yet much of transcription does not overlap protein coding annotations and appears to be noncoding[31]. In particular, one class of nascent noncoding transcripts originate from enhancers, or regulatory regions within the genome. While the ENCODE project made major inroads on identifying these critical regulatory regions[31], their precise boundaries are still difficult to ascertain, so they remain largely unannotated. The transcripts that originate from these enhancers, known as eRNAs, are unstable and lowly expressed but do appear to be critical to their regulatory activity[86, 172, 101, 116, 67]. They are detectable in GRO-seq and tend show bidirectional transcription[115]. Therefore, the unbiased identification of all regions of transcription from GRO-seq is an important and pressing problem.

At the time of this work, only two efforts have attempted to identify regions of active transcription directly from GRO-seq data[4, 66, 27], though neither is fully independent of annotation. The first used a two state Hidden Markov model by Hah et. al. that was parametrized based on available annotations[66]. This approach has the advantage of calling large contiguous regions as transcribed, but fails to call many unannotated regions because their length and transcription levels do not mimic well annotated regions. Furthermore, the approach is limited in its ability to discover transcripts that conflict with the annotation. A more recent approach, called Vespucci, uses a sliding-window (specified by two user-dependent parameters) that merges adjacent windows together based on read depth, but requires the user to tune the algorithm with each new dataset[4]. The windowing scheme, in principle, has the benefit of not depending on annotation. In practice, however, because regions of transcription are often broken into discontiguous sections, Vespucci requires the use of annotations to improve its strategy[4].

The subsequent approach combines the strengths of these previous efforts[66, 4]. In particular, I propose a fast and robust method that takes advantage of a logistic regression classifier embedded within a hidden Markov model as a means of learning non-linear decision boundaries that classify regions of active nascent transcription. This approach shares a similar structure with Maximum Entropy Markov Models[113]. This methodology is annotation agnostic, requiring only a small number of training examples to adapt parameters to new data. It effectively identifies cohesive regions of active transcription while maintaining a rapid runtime. Furthermore, the identification of transcripts solely from the signal within the data uncovers distinct biological phenomena previously missed in GRO-seq analysis. Finally, user-friendliness was a large consideration in the design and structure of the software.

Using a differential transcript methods, myself and the Dowell group (specifically Dr. Mary Allen) re-analyzed an earlier GRO-seq dataset[3] at both previously unannotated transcripts and annotated genes, demonstrating many of the earlier calls were annotation based artifacts. Shockingly, we demonstrate that the major response to activating p53, is increased transcription of p53's own binding site.

2.2 Nascent Transcript Model

2.2.1 Description

The GRO-seq technique measures nascent transcripts produced from actively engaged RNA polymerases[33]. Because splicing has not yet occurred, each transcript covers a contiguous region of the underlying genome, reflecting the extent of polymerase activity. Sequencing reads obtained from the GRO-seq protocol represent a sample from the underlying transcripts in proportion to their relative abundances. Ideally, overlapping reads could be merged into contigs, or regions of continuous read coverage, defining regions of active transcription. However, because of uneven sampling, coverage within active regions may not be contiguous. Furthermore, the sequencing and mapping process is noisy, therefore reads can also spuriously map to inactive regions.

Transcription can be modeled as a discrete time-series indexed by genomic coordinates where transcriptional activity observed at adjacent base-pairs is correlated. Similar to prior models of GRO-seq[66], I modeled this process as an ergodic first-order Markov chain where transcription oscillates between *active* and *inactive* states. Unlike previous models, which classify individual nucleotides, this model emits from each state a contig representative of an active or inactive region (Figure 2.1). Each contig can be described by two feature classes: contig length (maximum length of overlapping reads) and contig coverage statistics (Table A.1). Active states, in general, contain a combination of long regions with high signal interspersed with short regions of relatively no signal.

Hence this HMM framework allows for the classification of a continuous active region, containing one or more contigs, despite the variability in coverage of individual nucleotides that is inherent in short read sequencing data.

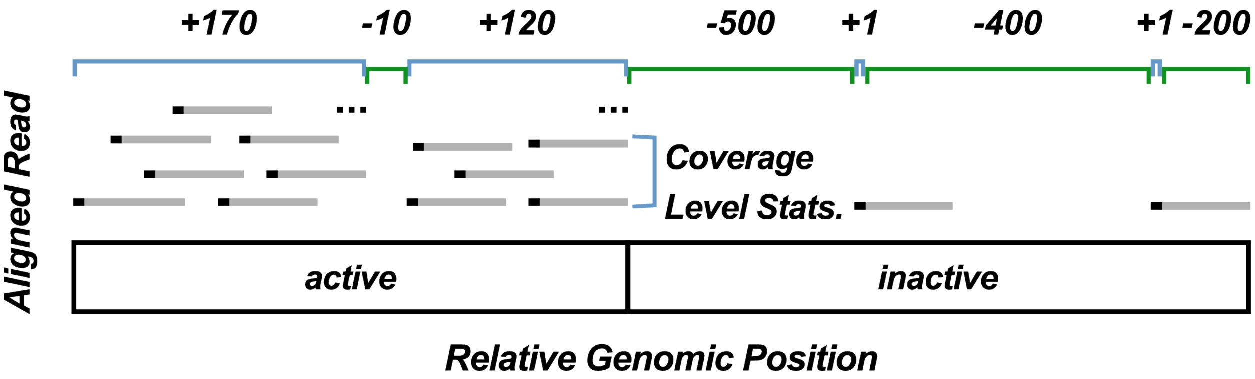


Figure 2.1: **A schematic showing how contig length and coverage statistics discriminate *active* from *inactive* nascent transcription.** Regions of active transcription contain many long contigs (positive length, not drawn to scale) with significant read coverage (labeled in blue) interspersed with short regions of no coverage. Coverage statistics define mean, median, mode and variance of reads (black bars) across a contig, see Table A.1. In segments with no reads, a gap (labeled in green) is defined by a negative length value and all coverage statistics are set to zero. Reads (grey bars) are represented by only their 5′ position (black points). Therefore a contig is also a continuous region where every base has at least one read's 5′ end at that position. Consequently, small gaps between contigs have a high probability of being in an active call.

HMM emission and transition parameter estimation requires manual labeling of active and inactive transcription. With such a training set, learning the conditional probabilities of a state classification from the set of implicit feature vectors is straightforward with logistic regression. Indeed, logistic regression comprises three main benefits: it requires little training data for parameter estimation, it quickly converges, and it readily scales with genome size. The logistic regression predictors are interpretable as probabilities, and therefore easily embedded into a HMM as emissions. After the probability transitions of the underlying Markov chain have been estimated, the well-known decoding algorithms such as Viterbi and Forward/Backward can be used to infer the most probable state sequence[113].

2.2.2 Parameter Estimation

The Markov model transition probabilities and the conditional state emission probabilities of the HMM are estimated via a user defined, labeled training set. Given that read mapping can be noisy and nascent transcripts can be present at very low levels, estimating parameters that discriminate active from inactive transcription regions poses a difficult problem. However, I show in Section 2.3.2 that surprisingly little training data is needed to retain high model accuracy—defined as the fraction of base pairs where the user-label and classification-label agree.

Here I outline the logistic regression parameter estimation method, for a detailed exposition see Ohno-Machado's review [46]. The conditional probability $p(k \mid \vec{x})$, where $k \in \{\textit{inactive}, \textit{active}\}$ and $\vec{x}$ indicates a feature vector, via a labeled training set of defined genomic coordinates representing *active* and *inactive* transcription regions. Table S1 provides a complete description of the feature vector $\vec{x}$. Clearly, $p(\textit{inactive} \mid \vec{x}) = 1 - p(\textit{active} \mid \vec{x})$. The later probability may be represented in terms of the sum of the coordinates of $\vec{x}$, weighted by some parameter vector $\vec{\theta}$. To treat this linear function as a probability, one may bound the sum to the range [0,1] via the sigmoidal transformation as follows:

$$p(\textit{active} \mid \vec{x}) = \frac{1}{1 + e^{-\langle \vec{x}, \vec{\theta} \rangle}}, \tag{2.1}$$

where

$$\langle \vec{x}, \vec{\theta} \rangle = \theta_0 + \sum_{i=1}^{n} x_i \cdot \theta_i, \tag{2.2}$$

$(n+1)$ is the dimension of the feature vector $\vec{x}$.

A simple plot of two features, gap length (x_1) and average read coverage (x_3), shows that these features may not be linearly separable (Figure 2.2A). By embedding the data into a higher dimensional space, a polynomial kernel (equation 2.3) can represent non-linear decision boundaries (Figure 2.2B),

$$f(\vec{x}, \vec{\theta}) = \langle \vec{x}, \vec{\theta} \rangle^{d} + c. \tag{2.3}$$

The polynomial kernel function parameters (c and d) can be set by the user in the FStitch software

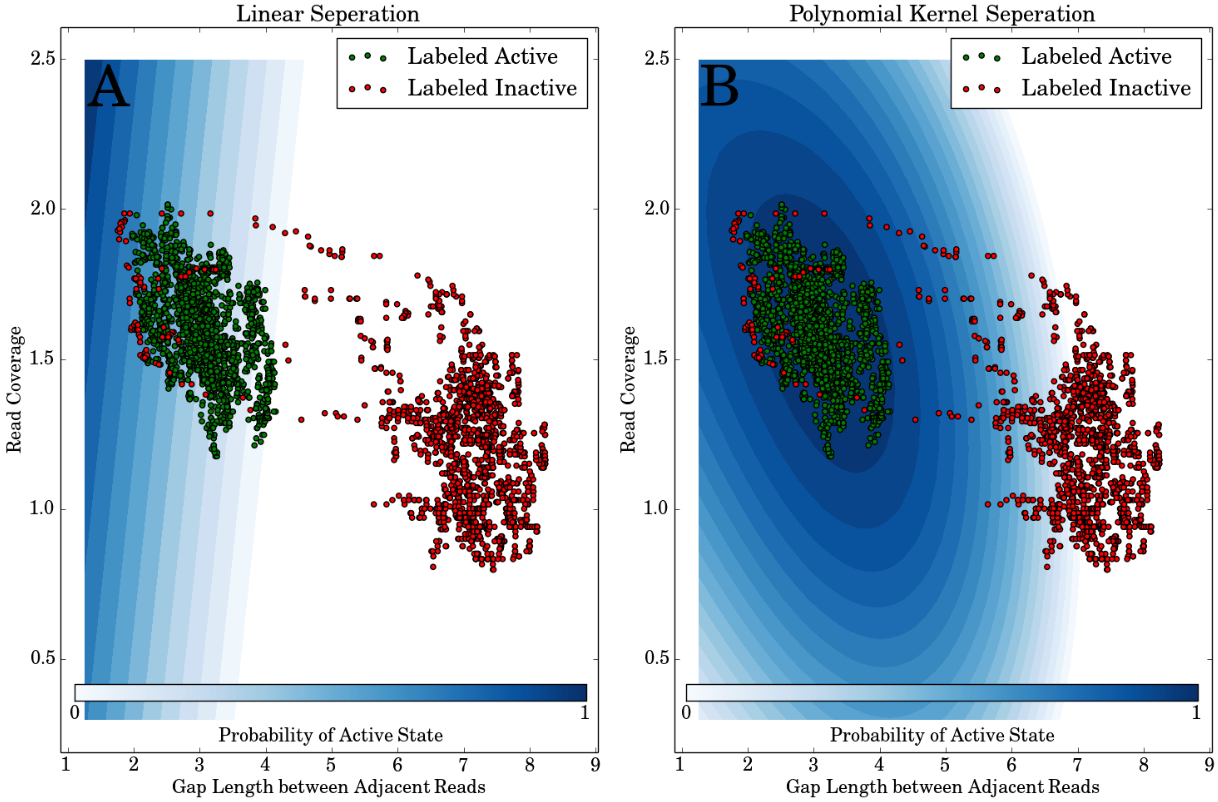


Figure 2.2: **Read coverage features are not linearly separable.** Points colored green represent training examples labeled *active* and those colored red indicate training examples labeled *inactive*. The blue shading provides a contour plot of the *active* state probability given the feature's average read coverage (x_3, y-axis) and the gap length between adjacent contigs (x_1, x-axis in log nucleotides). (A) uses logistic regression with a linear kernel function (i.e. $d = 1$ in equation 2.3), whereas (B) uses a second-order polynomial kernel function (i.e. $d = 2$ in equation 2.3).

package. The kernel function is incorporated into the sigmoidal transformation as follows:

$$p(active \mid \vec{x}) = \frac{1}{1 + e^{-f(\vec{x}, \vec{\theta}^T)}}. \tag{2.4}$$

To maximize training and classification accuracy, the algorithm adjusts to the behavior of the feature space. The use of a simple second-order polynomial kernel ($d = 2$ and $c = 0$) increases the training accuracy by ~10% in the HCT116 GRO-seq dataset (Figure 2.4). Importantly, this ~10% increase reflects mostly lower expressed labeled transcripts suggesting that the use of the polynomial kernel allows for greater sensitivity to under-represented, lowly transcribed regions.

To estimate the parameter vector $\vec{\theta}$ we maximize the log-likelihood function of the training set D:

$$l(\vec{\theta}, D) = \sum_{i=1}^{n} \log p(k_i \mid \vec{x_i}). \tag{2.5}$$

Here D can be thought of as a $N \times (n+1)$ matrix where N is the number of training examples and $(n+1)$ is the dimension of the feature vector $\vec{x}$. The i^{th} training label, k_i, is either *active* or *inactive*.

Newton-Raphson optimization routine[19] iteratively updates $\vec{\theta}$ until convergence. Because this techniques utilizes a second-order Taylor series approximation of the log-likelihood function, convergence is usually fast. The update rule is:

$$\vec{\theta}^{\,t+1} = \vec{\theta}^{\,t} - \big(\mathbf{H}L(\vec{\theta}, D)\big)^{-1} \cdot \nabla L(\vec{\theta}, D), \tag{2.6}$$

where ∇ and $\mathbf{H}$ represent the gradient and Hessian operators with respect to the vector $\vec{\theta}$, respectively. Finally, the most probable state sequence is estimated via the Viterbi Algorithm[121], using the Maximum Entropy Markov model framework[113], and is given by the recurrence relation:

$$v_t(k) = \max_{j \in S}(v_{t-1}(j) \cdot a_{j \to k}) \cdot p(k \mid \vec{x_t}), \tag{2.7}$$

where $a_{j \to k}$ represents the transition probability from state j to state k of the hidden Markov chain, which is estimated via Baum-Welch[113], S is the hidden transcriptional state space i.e. $S = \{active, inactive\}$, and $p(k \mid \vec{x_t})$ is given in equation 2.3 with $\vec{\theta}$ learned from the training data using the Newton-Raphson algorithm. Here $\vec{x_t}$ is either a gap or contig representation given in Table S1.

Using training data to learn parameters allows users to intuitively provide regions of transcriptional characterization thereby doing away with arbitrary parameter values and grid parameter search for optimization. These parameters are learned *from* the data and thus adapt accordingly.

2.2.3 Software Design

The purpose of FStitch is to segment the genome into regions of *active* and *inactive* nascent transcription. The algorithm accepts as input a $5'$ BedGraph file (each read counted only at its

5′ end) of read coverage and a training set file consisting of a few segments (at least 3 segments) labeled as *active* or *inactive* regions of nascent transcription. The training file requires only start and stop coordinates of regions considered *active* and *inactive* yet, within these regions, the data should be rich in feature vectors (i.e. contig lengths and coverage statistics). As defaults, FStitch has pre-labeled *active* and *inactive* segments for a human genome based on house-keeping genes and gene desert regions, respectively. However, care must be taken with defaults as the transcriptional landscape varies from experiment to experiment and datasets need not be human or mapped to hg19.

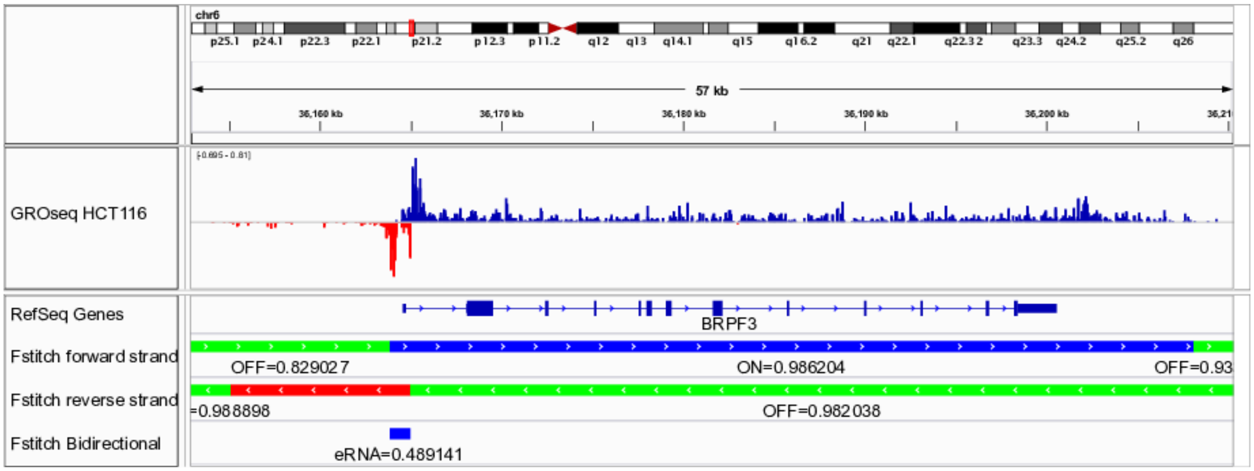


Figure 2.3: **FStitch output at BRPF3.** An IGV snapshot showing a sub-region in chromosome 6 around BRPF3. The first track shows typical GRO-seq data from the HCT116 dataset, with the positive and negative strand in blue and red, respectively. RefSeq annotations are shown next. FStitch output is below for each strand with green indicating areas of *inactive* transcriptional activity, blue representing areas of *active* transcription on the positive strand and red on the negative strand. The scores associated with each classification via the Logistic Regression and Viterbi-provided Markov state sequence are also displayed. Finally, bidirectional predictions are provided at the bottom with a score via the estimated Normal Distribution confidence interval.

FStitch outputs two bed files for positive and negative strand classifications, respectively, that can be imported into typical genome browsers such as IGV or the UCSC genome browser, to view the classifications in conjunction with read coverage files[165]. Figure 2.3 shows a typical output of the algorithm. These bed files contain the genomic start and stop of each classification and an associated probabilistic score from the Viterbi algorithm (Equation 2.7). From start to finish, FStitch takes ∼3.5 minutes to predict transcript annotations in the most deeply sequenced GRO-seq dataset, HCT116 (152.4 million mapped reads)[3].

FStitch is written in the C/C++ programming language and complied using GNU compilers later than GCC 4.2.1. The user interface is command line, resembling many popular bioinformatics pipelines. FStitch is stand-alone and borrows from no third-party platforms, libraries or packages. The open-source software and a comprehensive manual is freely downloadable at `https://github.com/azofeifa/FStitch/`.

2.3 Model Accuracy

2.3.1 Datasets

This study takes advantage of three previously published GRO-seq datasets (labeled here by the underlying cell line): MCF-7 [66], IMR90 [33] and the Dowell Group's own HCT116 (DMSO and Nutlin, wild type p53)[3], as well as three published ChIP-Pol II datasets: HCT116[73], IMR90[75] and MCF7 [77]. For each experiment, raw reads were mapped to the hg19 genome using Bowtie2 with the command bowtie -S -t -v 2 -best[93]. A $5'$ bedgraph is then generated using BedTools's (2.16.2) genomeCoverageBed (options: -5 -bg -strand) for each strand. Additionally, the ENCODE project provided H3K27ac, H3K4me1, and DNase I hypersensitivity peak calls for IMR90 [115, 26], MCF7 [58, 69] and HCT116 [58, 129], as well as ChIA-PET peak calls for HCT116[60]. Finally, to create a list of high confidence p53 binding sites, we combined the data from seven ChIP assays for p53[175, 127, 153, 152] and kept only sites that were found in at least 3 of the 7 assays.

Because most nascent transcription is unstable and therefore understudied[38], an undergraduate student within the Dowell Group (Josepine Hendrix) hand annotated the entire length of chromosome 1 from the Dowell Group's earlier HCT116 GRO-seq DMSO dataset[3] to perform k-fold cross validation. For all testing, 95% of the labeled dataset was removed from training and used to assess model accuracy. To be clear, the entire labeled HCT116 training set contains 17,776 regions labeled as *active*. Based upon the cross validation results, 7 regions considered *active* and

7 regions considered *inactive* were used for parameter estimation in both the IMR90 and MCF7 GRO-seq datasets. These training sets (with genomic coordinates and labels) are provided in Supplemental Table S3.

2.3.2 Sensitivity to depth of data

To assess the sensitivity of the algorithm to the amount of training data, I utilized the hand-annotated labeling of regions as *active* or *inactive* discussed in the previous section. The manual annotation identifies approximately 17,000 active and inactive regions, effectively labeling roughly 36% of chromosome 1 as active. FStitch was tested over this rich labeled data using K-fold cross validation, reserving 5% of the training data for parameter estimation and leveraging 95% for testing accuracy.

To assess the amount of training data needed for accurate classification of *active* regions, the amount of training data was incrementally decreased and the true and false positive scores were monitored. Figure 2.4A shows that FStitch training is robust to successive decreases in the amount of training data utilized, suggesting that very little training data is needed to achieve relatively high accuracy. The smallest training set (0.1% of the initial dataset) consists of 3 *active* and 2 *inactive* regions and maintains scores of 95% true positive and 4.3% false negative on the testing dataset. Furthermore, I observed that the polynomial kernel consistently outperforms the linear kernel.

Similarly, the sensitivity of FStitch to experimental sequencing depth was assessed. To this end, I randomly subsampled (without replacement) from the HCT116 test dataset, the single experiment with the deepest read coverage. For each subsample, I re-estimated the parameters via a fixed training set, 5% of chromosome 1 labels. Subsequently, I reclassified *active* transcript segments and calculated the training accuracy relative to the test set. Figure 2.4B shows that this method is robust to low sequencing depth of the dataset.

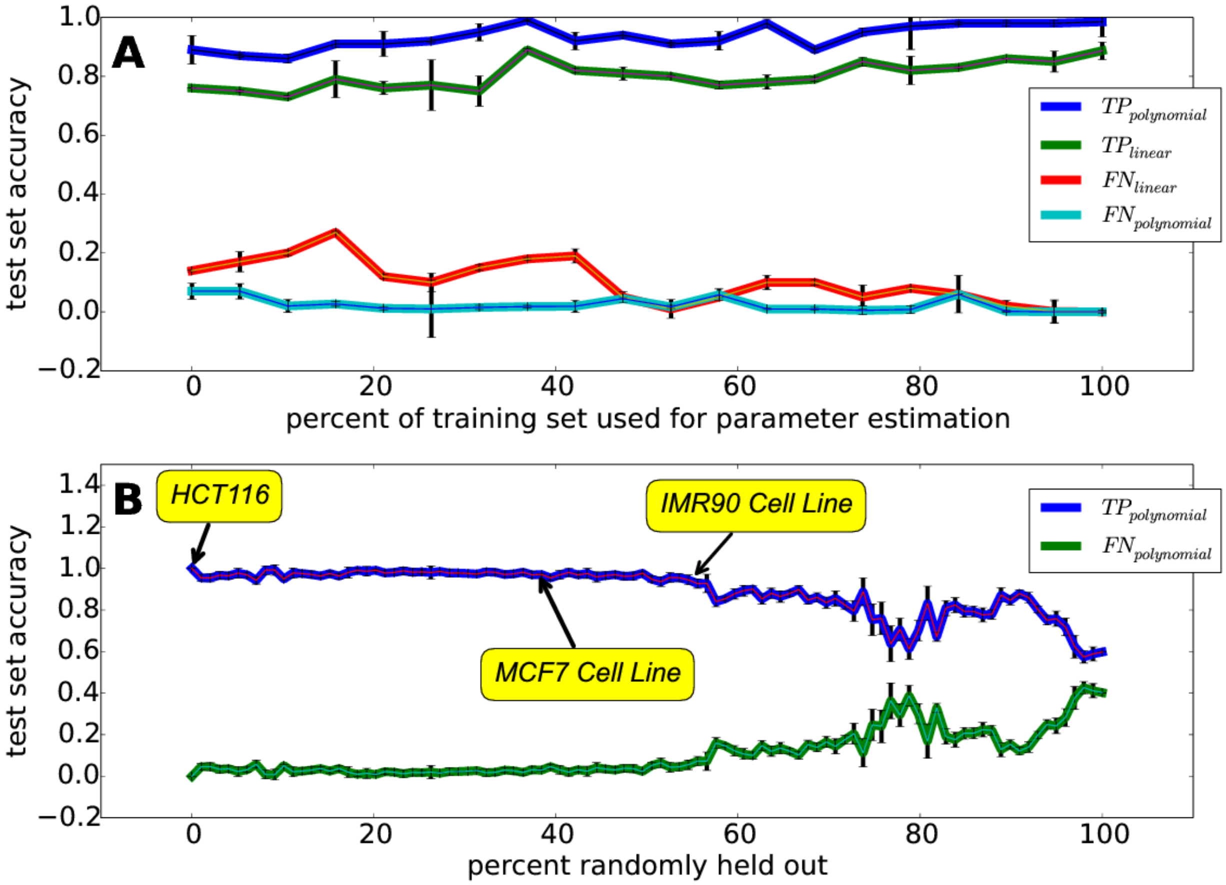


Figure 2.4: **FStitch requires little training data and is robust to low levels of GRO-seq read coverage.** (A) Classification accuracy utilizing successively decreasing amounts of training data to learn feature vector weights, for the polynomial ($d = 2$ and $c = 0$; blue and teal) and linear ($d = 1$ and $c = 0$; green and red) kernel. (B) Classification accuracy with successively less sequencing depth (dataset size). In this case, 5% of all available chromosome 1 labels was used for training and tested on 50 different subsamples of the curated dataset. TP = *true positive rate* and FN = *false negative rate.*

2.3.3 Benchmarking FStitch & Vespucci

To evaluate FStitch to the previously published windowing method Vespucci[4], model accuracy was calculated for Vespucci with the default parameters over the HCT116 test dataset (Table 2.1). In addition, a grid search was performed on a subset of ranges for both Max_Edge and Density_Multiplier combinations and reported the performance of the best parameters obtained for this dataset. Grid search optimization greatly increased Vespucci's precision and recall. FStitch

Table 2.1: **Benchmarking FStitch and Vespucci** Each algorithm, FStitch and Vespucci with default parameters (Max_Edge: 500 and Density_Multiplier: 10,000), and Vespucci with best parameters from a grid search, G.S. (Max_Edge: 10 and Density_Multiplier: 2,000), are compared on the manually annotated test set from chromosome 1. Overlap percentages are reported per base.

FStitch	Active Label	Inactive Label
Active Call	98.5%	1.5%
Inactive Call	0.01%	99.99%
Vespucci (default)	Active Label	Inactive Label
Active Call	93.97%	6.03%
Inactive Call	30.3%	60.7%
Vespucci (G.S.)	Active Label	Inactive Label
Active Call	99.44%	0.56%
Inactive Call	19.9%	80.1%

outperforms Vespucci, default or grid search, in both true negative and true positive classifications.

I next assessed the quality of FStitch *active* calls to independently derived relevant biological datasets. As GRO-seq measures all actively engaged polymerase, in a strand specific fashion, there is no single alternative experiment to confirm GRO-seq data. However, RNA polymerase II is responsible for most transcribed regions and therefore comparison to Pol II chromatin immunoprecipitation (ChIP should independently verify the location of most transcripts. To this end, previously published Pol II ChIP-seq data for MCF7, HCT116, and IMR90 cell lines[77, 73, 75] was obtained. Unfortunately, direct comparisons between GRO-seq and ChIP-seq are complicated by the fact that GRO-seq is strand specific whereas ChIP-seq is not. Yet, one can reason that the superposition of reads along the sense and anti-sense strand within GRO-seq should approximate ChIP-Pol II read coverage within the same region.

Thus, an *active* call should have a higher enrichment of RNA Pol II ChIP-seq than an *inactive* call. In all three cell lines, I used FStitch to identify bidirectional, *active* and *inactive* calls. Vespucci does not contain an unbiased bidirectional transcription annotator, therefore only *active* and *inactive* predictions were obtained. For MCF7, I utilized the published list of Vespucci annotations but for both HCT116 and IMR90; Vespucci parameters obtained via grid search (Table 2.1) were used. I noted that the Vespucci approach is less capable of distinguishing *active* from

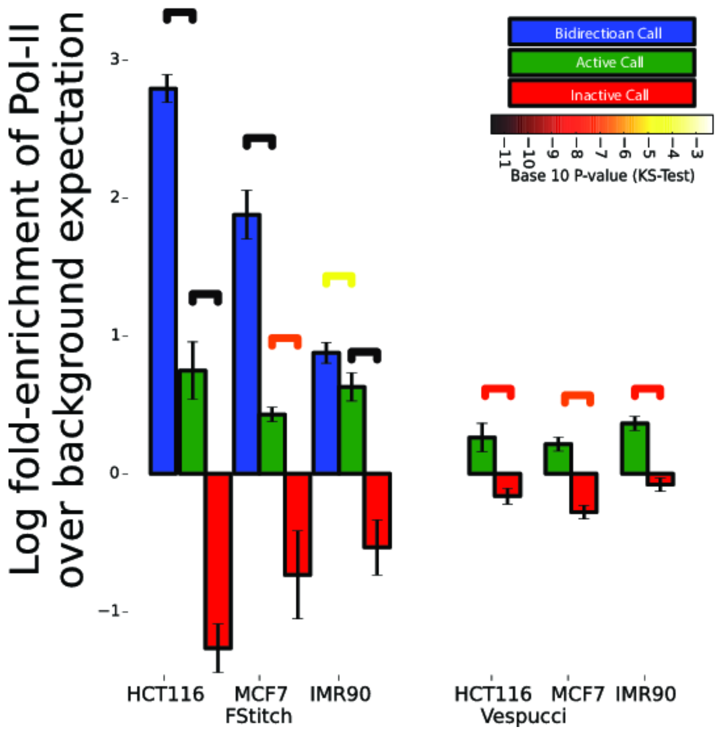


Figure 2.5: **Correlation of GRO-seq transcript calls with Pol II ChIP-seq.** Pol-II ChIP-seq read density was collected in regions labeled as bidirectional (blue), *active* (green) or *inactive* (red) by either FStitch (on left) or Vespucci (on right). Log fold-enrichment is relative to average Pol-II ChIP-seq read density. Statistical significance is assessed via the Kolmogorov-Smirnov test (significance bars colored by p-value). Error bars indicate one standard deviation away from the mean.

inactive regions as assessed by Pol II occupancy (Figure 2.5). I observed a significant enrichment for Pol II occupancy between *active* and *inactive* FStitch regions. Additionally, I observed a high degree of Pol II occupancy at bidirectional calls, as expected given that enhancers are known to show significant enrichment for Pol II occupancy[86].

2.4 Biological Analysis

2.4.1 Annotation Comparisons

To evaluate the performance of the proposed method to identify biologically meaningful regions of *active* transcription, I compared the results of FStitch to RefSeq annotations. I first classified *active* transcript calls on the HCT116 DMSO experiment by their overlap to genomic annotations. Most FStitch *active* calls overlap a known annotation: gene, antisense to a gene, long non-coding RNA (lncRNA), small nucleolar RNA (snoRNA), microRNA (miRNA) and transfer-RNA (tRNA) (Figure 2.6). Of the 26.75% of FStitch *active* calls that do not overlap known annotations, many can be described as bidirectional calls that overlap an H3K27ac mark; which is characteristic of an eRNA.

Interestingly, within the unannotated *active* calls, a small fraction (9%) contain both an open reading frame that spans at least 60% of the length of the call and a bidirectional call at the $5'$-end. These may be unannotated protein coding genes. I translated these regions and searched the UniProt/SwissProt protein database[32], uncovering several hits. By isolating the statistically significant hits and tokenizing the hit descriptions, I observed that more than 95% of all hits contained the reoccurring words *putative*, *uncharacterized* or *encode.*

Meta-gene analysis is a popular method of assessing the average behavior of an assay over gene annotations[161]. By taking advantage of the high read coverage of the HCT116 GRO-seq dataset, I constructed a meta-gene of FStitch active calls that completely overlap a RefSeq annotation (n=2512). For this analysis, I averaged the read coverage within 100 uniformly distributed proportions relative to the FStitch call (Figure 2.7). This uncovered two features of active regions: (1) the $3'$-end peak is much larger than previously detected [33, 115] and (2) there is a corresponding small build up of reads along the anti-sense strand that mirrors the $3'$-end peak. It should be noted that the $3'$ peak does not always correlate well with the exact $3'$-end of the annotation[6]. This is likely because the $3'$-end of a gene annotation is typically the mRNA cleavage site and not the RNA Pol-II termination site.

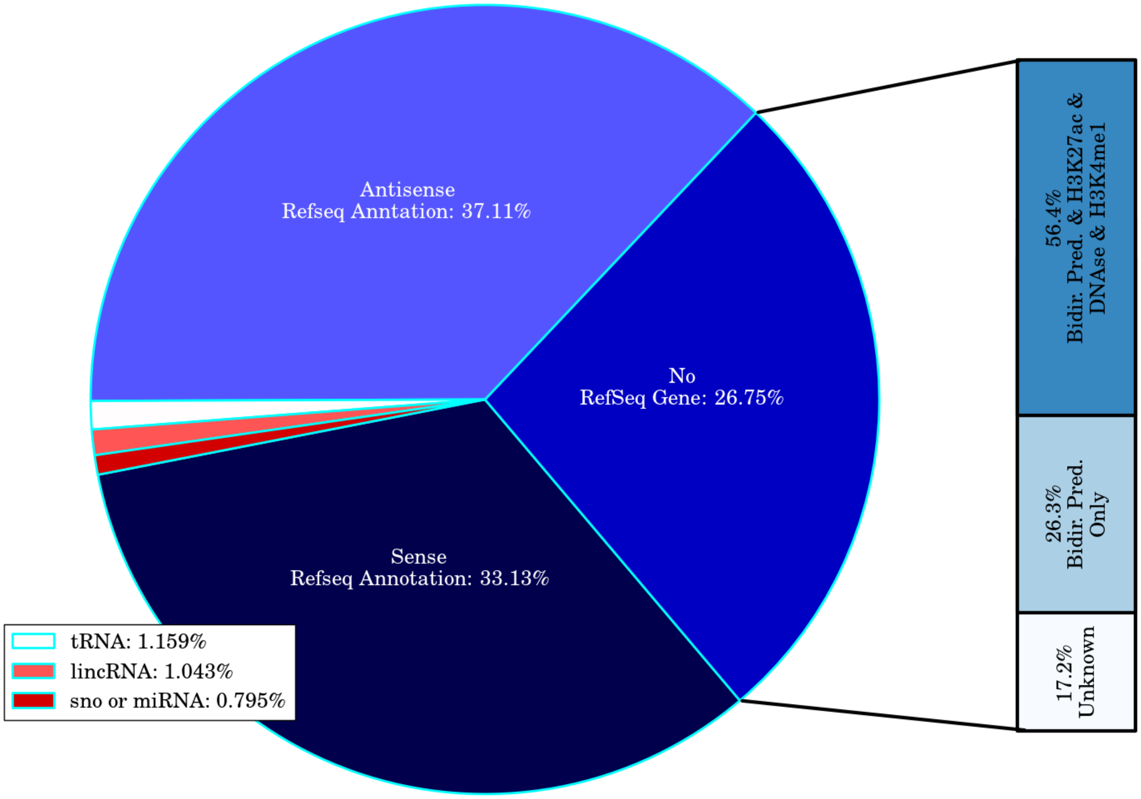


Figure 2.6: **Active Call Characterization.** FStitch *active* calls on HCT116 DMSO are divided into classes based on overlap with genomic annotations. Unannotated *active* calls are assigned if they have no overlap to previous annotations on either strand. FStitch called 37,591 *active* regions.

Given that FStitch does not rely on previous annotations, I next asked how the ends ($5'$ and $3'$) of FStitch active calls relate to known RefSeq gene annotation ends. Specifically, I measured the difference in genomic location between the $5'$ end ($3'$ end) of an FStitch *active* call and the nearest RefSeq annotation $5'$ end ($3'$ end), respectively. Interestingly, the GRO-seq signal often begins upstream of the annotated $5'$ start site of RefSeq genes (Figure 2.8A). Indeed, there appears to be two distinct populations within the $5'$ ends. Therefore, I fit a mixture of two Gaussian distributions using the Expectation Maximization algorithm[114] to the difference of $5'$ ends histogram. I examined the upstream Gaussian distribution for distinguishing features and found it shows a 2.5 fold enrichment of anti-sense transcription compared to the Gaussian centered at roughly the zero position. This suggests that many genes have upstream bidirectional transcription, and therefore may

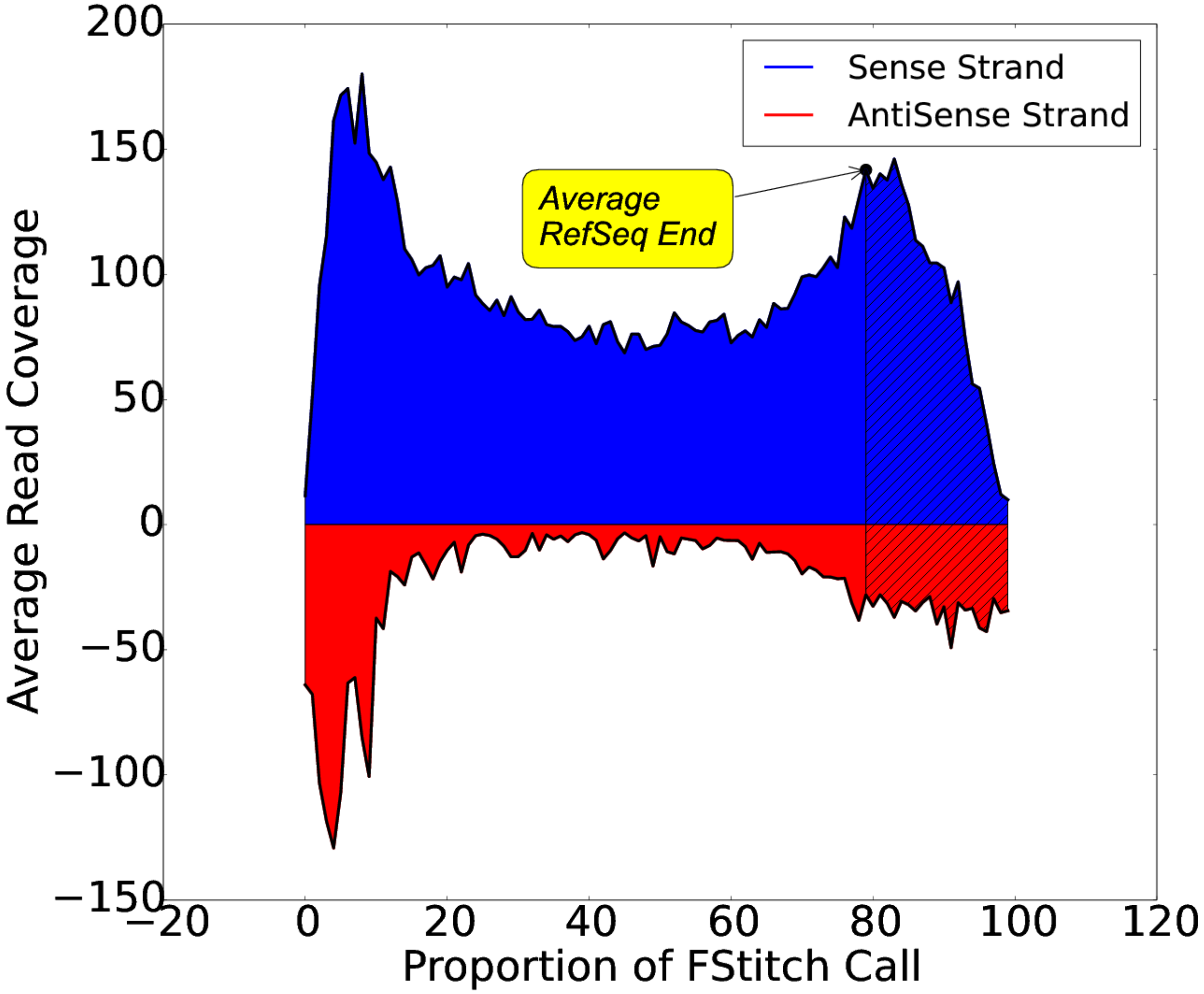


Figure 2.7: **Average Read Coverage of FStitch active calls.** FStitch *active* calls on the positive strand that completely contain a RefSeq annotation were used to calculate the average behavior. Blue and red represent positive and negative strand coverage, respectively. For each *active* region, the length was divided into 100 uniformly sized proportions and the read coverage was averaged within each bin. The average annotated 3′ end is noted by the line and transcription beyond the annotation is shaded. Here, I required an FStitch to completely overlap a RefSeq annotation and the RefSeq annotation overlap at least 75% of the FStitch call.

may have overlapping or adjacent upstream enhancers[151] or promoter upstream transcripts[136]. Notably in these cases, the upstream region and the annotated gene are a single *active* call.

Additionally, elongation traverses several kilo-bases (average of ∼8 kb) past the 3′-end of annotated genes (Figure 2.8B); consistent with the fact that polymerase proceeds far beyond the mRNA cleavage site [6, 8]. Notably, the 3′ extension is missed by earlier GRO-seq de novo transcript detection algorithms[66, 4]. Indeed, Vespucci captures many of the same general trends of FStitch, but typically terminates 3′ extensions earlier. Upon further examination, this may reflect the fact that Vespucci's default parameters are biased to highly expressed regions and the 3′ extensions are often weakly transcribed. On the other hand, the hidden Markov model of Hah et. al. was trained to match RefSeq annotations and is therefore unable to identify distinguishing features of nascent transcription at either end.

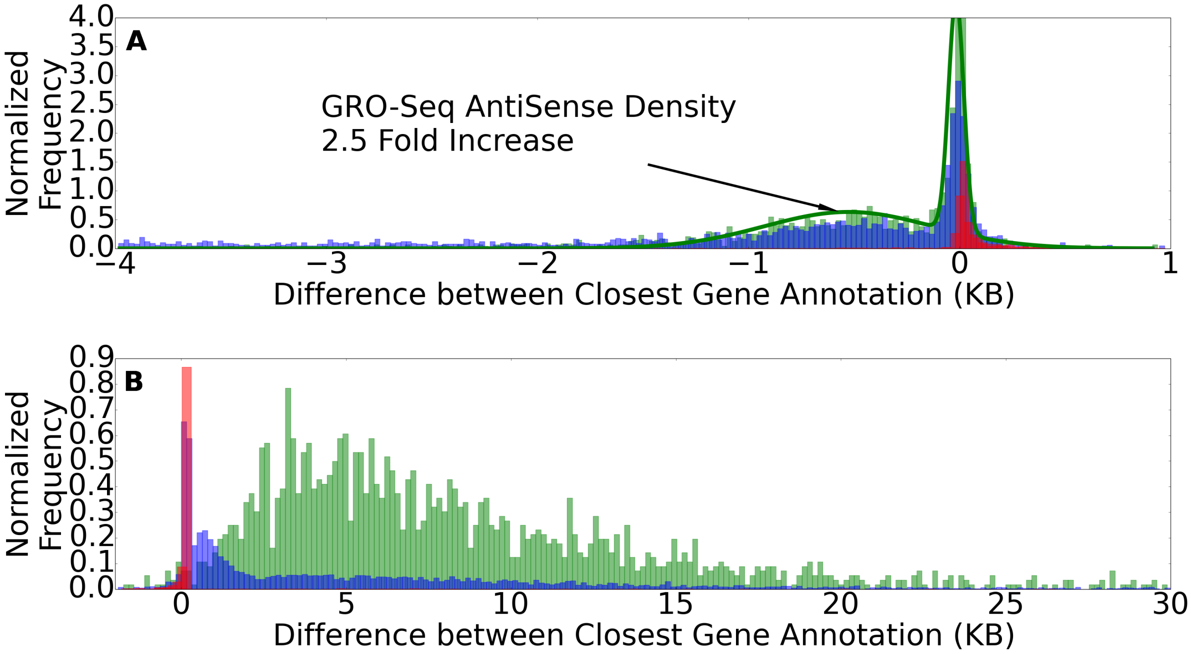


Figure 2.8: **Histograms comparing the active region calls of FStitch to RefSeq annotations.** The distance between the end of an *active* call and the nearest RefSeq annotation are compared for (A) 5′-ends; (B) 3′-ends. Colors *red*, *blue* and *green* are Hah et. al., Vespucci (grid search parameters) and FStitch *active* calls, respectively. Histograms are probability normalized.

2.4.2 Characterizing bidirectional RNA Activity

To assess the accuracy of the bidirectional predictions, I examined what fraction of the bidirectional calls overlap enhancer marks. For this analysis chromosome 1 was excluded and used

FStitch to predict bidirectional transcription in all three cell lines: IMR90, MCF7 and HCT116. In all cell lines, the bidirectional FStitch calls were significantly enriched for overlapping DNase I hypersensitivity sites and H3K27ac marks indicating that a large fraction of these calls are likely eRNAs (Table S2).

A natural hypothesis, bidirectional predictions that overlap enhancer marks should be highly transcribed, more so than bidirectional predictions without corresponding enhancer marks (Figure S4). In all three cell lines, higher levels of bidirectional transcription are observed when accompanied by a chromatin enhancer mark. As proof of concept, marks which do not overlap bidirectional prediction show little read density indicating that the false-negative rate is low (Figure S4, in red). Bidirectional predictions that overlap both a gene annotations and an enhancer mark show the highest level of average transcription. Moreover, there are 342, 241 and 198 bidirectional predictions in the HCT116, MCF7 and IMR90 datasets, respectively, that do not overlap a chromatin enhancer mark but do show a GRO-seq transcription greater than the mean GRO-seq signal of bidirectional predictions overlapping a DNase I hypersensitivity site or H3K27ac mark. These highly expressed bidirectional regions may be, as of yet, undiscovered enhancers.

To examine the theory that enhancer elements are three-dimensionally connected to their gene regulatory partner[86, 172, 101, 116, 67], GRO-seq signal was compared between three-dimensional paired chromatin interactions. I utilized a Pol II chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) dataset in the HCT116 cell line[101]. ChIA-PET is a rather new high-throughput technique that pulls down a protein of interest (in this case Pol II) and provides information on long range chromatin interactions[60] associated with the protein. Therefore, I first examined the overlap between both FStitch *active* calls and bidirectional predictions with paired ChIA-PET reads. There exists a highly significant overlap (hypergeometric; p-value $< 10^{-10}$) between ChIA-PET reads and FStitch *active* calls.

Given the three dimensional association implied by ChIA-PET, I asked if interacting DNA regions show a correlated GRO-seq transcription signal. When assaying for GRO-seq signal utilizing only ChIA-PET read pairs, no correlation is found in transcript level (Pearson's correlation

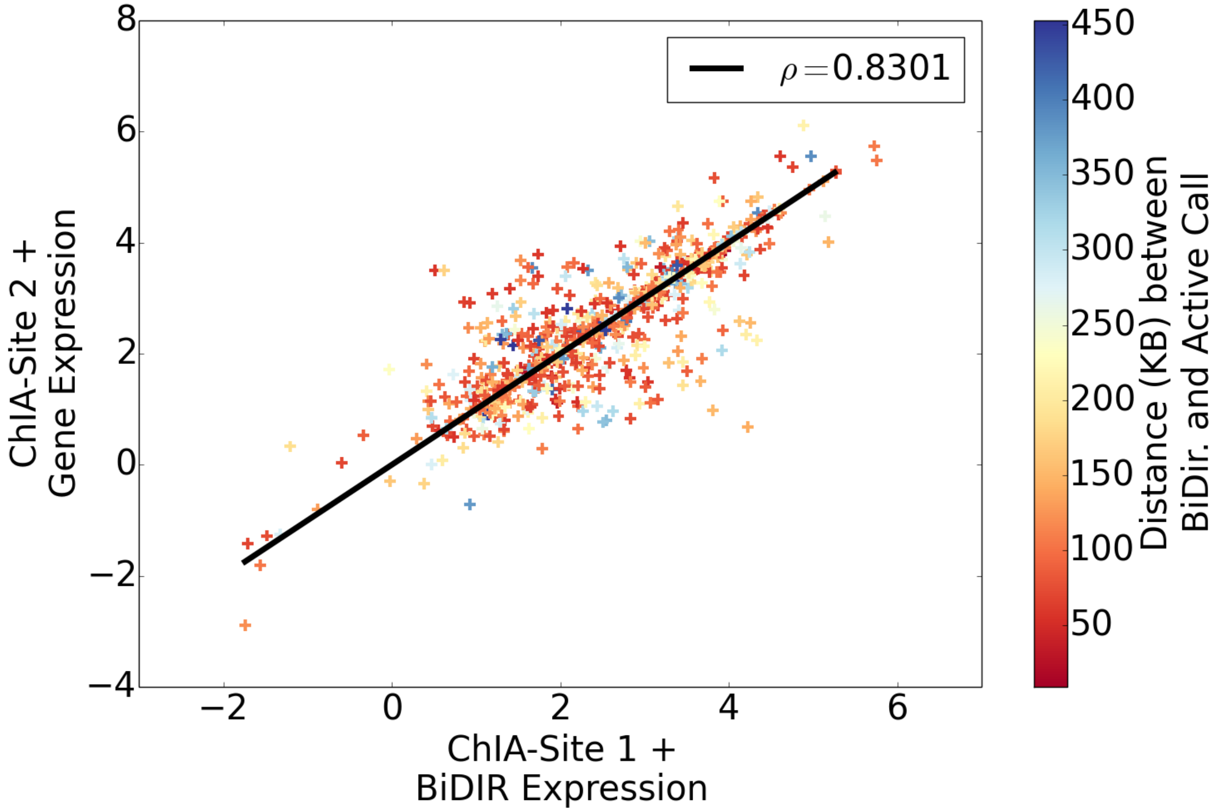


Figure 2.9: **Bidirectional predictions and *active* FStitch calls connected by a ChIA-PET read pair show correlated GRO-seq transcription.** The GRO-seq transcription level of ChIA-PET read pairs that overlap a bidirectional call and an *active* call on either end are plotted, demonstrating a strong correlation ($\rho = 0.8301$) in transcription (as measured by GRO-seq). Points are colored according to genomic distance (kb) between bidirectional prediction and *active* call.

coefficient; $\rho = 0.001$). However, when ChIA-PET read pairs are categorized by association with a bidirectional prediction, a strikingly high correlation is observed ($\rho = 0.8301$; Figure 2.9). Note that I do not include cases where the ChIA-PET read pairs overlap the same FStitch *active* call used to make the bidirectional prediction. Moreover, this linear relationship appears completely independent of genomic distance. This poses an obvious question: can one predict enhancer-gene interactions? Using a general linear model estimated from Figure 2.9, I attempted to predict enhancer-gene interactions using only GRO-seq transcription level. Unfortunately, only 7% of enhancer-gene interaction predictions were validated by ChIA-PET read pairs. This result suggests that while GRO-seq signal appears highly correlated between enhancers and their gene targets, additional information is needed to predict which enhancers are associated in three dimensions with particular FStitch *active* calls.

2.4.3 Differential transcription at annotated genes: a comparison of FStitch to Allen et. al.

In Dowell group earlier publication, the direct transcriptional targets of the transcription factor p53 were examined in HCT116 cells. In that experiment, p53 was activated by the non-genotoxic drug Nutlin (see [3] for complete details). Analysis was annotation centric but excluded the first 1 kb around the annotated start to avoid the initiation peak of polymerase. Furthermore, assessment of transcription over p53 binding sites was dependent on publicly available p53 ChIP-seq data. As an important and thankful disclaimer, Dr. Mary Allen (a post-doctoral researcher in the Dowell Group) helped greatly with the below differential expression analysis.

On the control GRO-seq (DMSO) and the p53 activated GRO-seq (Nutlin) independently, FStitch predicted 37,591 *active* calls in DMSO and 39,097 *active* calls in the Nutlin treated sample. Many *active* calls in both DMSO and Nutlin overlap RefSeq annotated genes (annotation overlap for DMSO shown in Figure 2.6). In total, 16,191 (of 23,669) genes are transcribed in at least one of the two experiments. Interestingly four large genes called as differentially transcribed in Allen et. al. are not called as *active* by FStitch in either experiment. These genes appear to contain only scattered background reads (noise), but because of their massive size still contain a large total number of reads. The merged method was then used to identify regions of interest for assessing differential transcription between DMSO and Nutlin.

By manual inspection, many FStitch regions of interest are much shorter than the annotated gene. To this end, I required that for each gene at least 75% of the gene be called as differentially transcribed. From this one can conclude that many genes, including 45 called in Allen et. al., do not show differential transcription along the full length of the gene. For example, PVRL4 (Figure 2.10) was called as differentially transcribed in Allen et. al. yet FStitch identifies that the signal for differential transcription is entirely driven by a distinct small subregion within the gene.

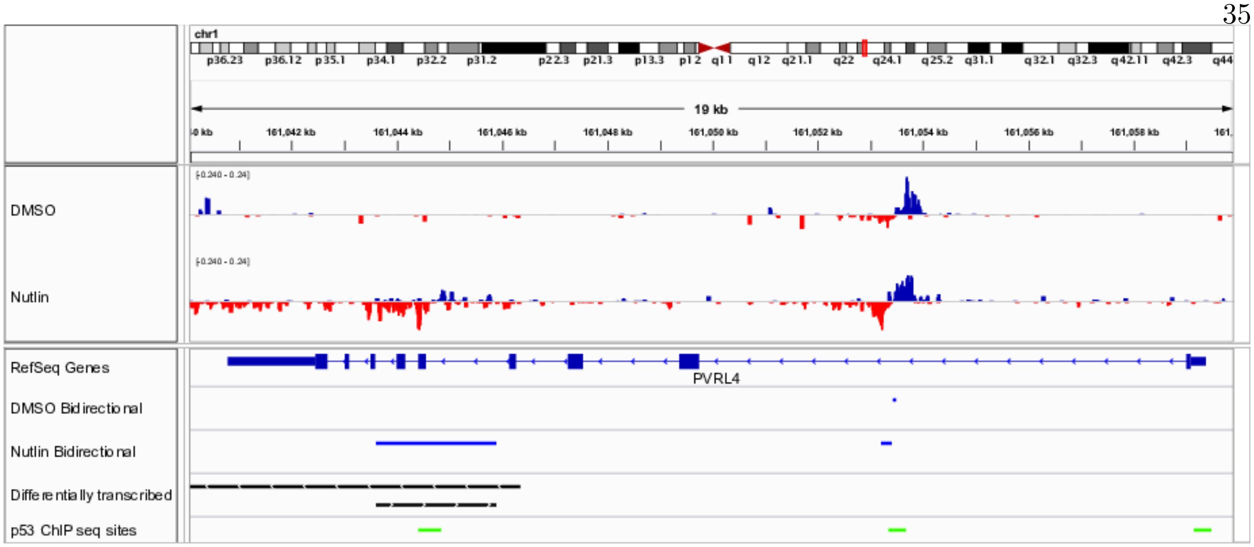


Figure 2.10: **Differential Transcription at PVPR4.** An IGV snapshot showing PVPR4, a negative strand gene where a small portion of the gene is differentially transcribed. The region of differential transcription (black bars) overlaps both FStitch bidirectional calls (blue bars) and p53 binding sites (green bars), indicating this may be an intragenic enhancer. The tracks, in order, are: histograms of the GRO-seq signal observed in DMSO and Nutlin, respectively (positive strand: blue; negative strand: red); RefSeq annotation for PVPR4; FStitch bidirectional calls in both DMSO and Nutlin, respectively (blue bars); FStitch differential transcription calls (black bars: top is negative strand, bottom is positive strand); location of p53 binding events (in green).

Most of these differentially transcribed regions overlap FStitch bidirectional calls, implying that the annotation centric method was sometimes mislead by overlapping, fully contained enhancer.

In several cases, the signal for differential transcription is not uniformly distributed across the transcribed region. The distribution of reads is not uniform, with most genes showing a $5'$ peak, corresponding to polymerase initiation that is distinct from the read distribution within the gene. The Allen et. al. analysis excluded the first 1kb of each annotated region in an effort to examine only polymerase elongation through the body of the gene. With FStitch one can consider the entirety of the *active* region. Consequently when differential transcription is driven primarily by read depth changes at the $5'$ end, the gene is called by FStitch but missed in Allen et. al. Analogously, Allen et. al. calls genes where the gene body is changing but inclusion of the $5'$ peak washes out the differential signal. Finally, there are cases where a gene is called in Allen et. al. but missed by FStitch because the *active* call overlapping the gene is much longer than the gene, a situation that arises in gene dense regions.

2.4.4 Differential transcription using all FStitch *active* calls

Importantly, FStitch is able to identify unannotated regions that are differentially transcribed. When DESeq considers all FStitch regions of interest, 1044 regions are called as differentially transcribed. Remarkably 75% of these regions do not overlap an annotated gene.

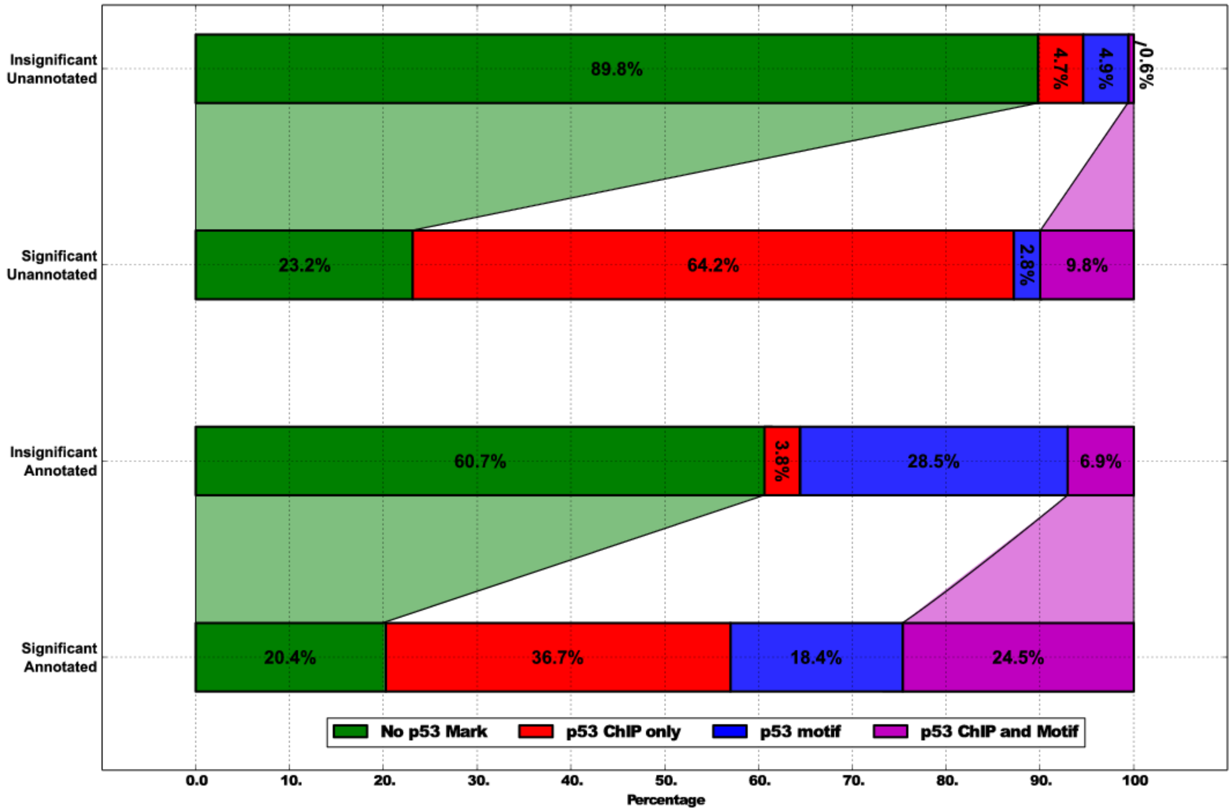


Figure 2.11: **Overlap of differential transcription and p53 marks.** FStitch calls were grouped significance of differential transcription (significant: DESeq adj. p-value < 0.1) and overlap with a RefSeq annotation. From top to bottom, there are 64,899 regions without differential transcription (insignificant) and without overlapping annotation (unannotated); 782 significant-unannotated; 23,986 insignificant-annotated; and 262 significant-annotated, respectively. p53 binding site (ChIP) overlap and p53 motif presence are assessed as described in the text.

Because Allen et. al. found differential transcription at p53 binding events, Dr. Mary Allen hypothesized that a large fraction of the unannotated FStitch differentially transcribed regions would contain p53 binding events and/or p53 sequence motifs. Binding events for p53 were called as described in Allen et. al., except requiring consensus from three of the seven publicly available p53 ChIP datasets[127, 3]. Presence of the motif was determined by the publicly available p53 scanner algorithm, requiring a p-value < 0.01[153]. Differentially transcribed regions, both those overlapping annotated and unannotated regions, are highly enriched for marks of p53 (either binding or motif). Because annotated regions tend to be much longer than unannotated, they are more likely to contain a p53 motif and/or ChIP site. In fact, most regions that are differentially transcribed (73%) overlap an experimentally determined p53 binding event.

Lastly, I sought to determine which unannotated FStitch differential transcription calls are themselves p53 *enhancers*. To this end, I examined their overlap with known enhancer marks H3K27ac, H3K4me1 and DNase I hypersensitivity (Figure 2.12). Unannotated differentially transcribed FStitch calls are over enriched for enhancer marks, relative to background expectation. Indeed, the three enhancer marks (H3K27ac, H3K4me1 and DNase I hypersensitivity) are more likely to co-occur in the differentially transcribed set. Interestingly, 21.2% of these regions are paired with another differentially transcribed FStitch call in the HCT116 ChIA-PET study. This overlap far exceeds the expectation (0.01%) that a random FStitch call will pair with a differentially transcribed partner by ChIA-PET.

2.5 Conclusions

FStitch is a fast and robust algorithm for the identification of transcripts within GRO-seq data that is annotation agnostic. Parameters of the algorithm are learned from small amounts of training data and can adapt readily to low depth of sequencing. By taking advantage of logistic regression, a non-linear classification of the feature space is learned. This classifier is then embedded within a Hidden Markov model framework, so as to identify contiguous segments of active transcription. The *active* calls from this algorithm correspond well to independently obtained secondary datasets (such as Pol II ChIP-seq and ChIA-PET) and can be used to identify sites of bidirectional transcription within a dataset or to examine differential transcription between datasets. FStitch is user friendly

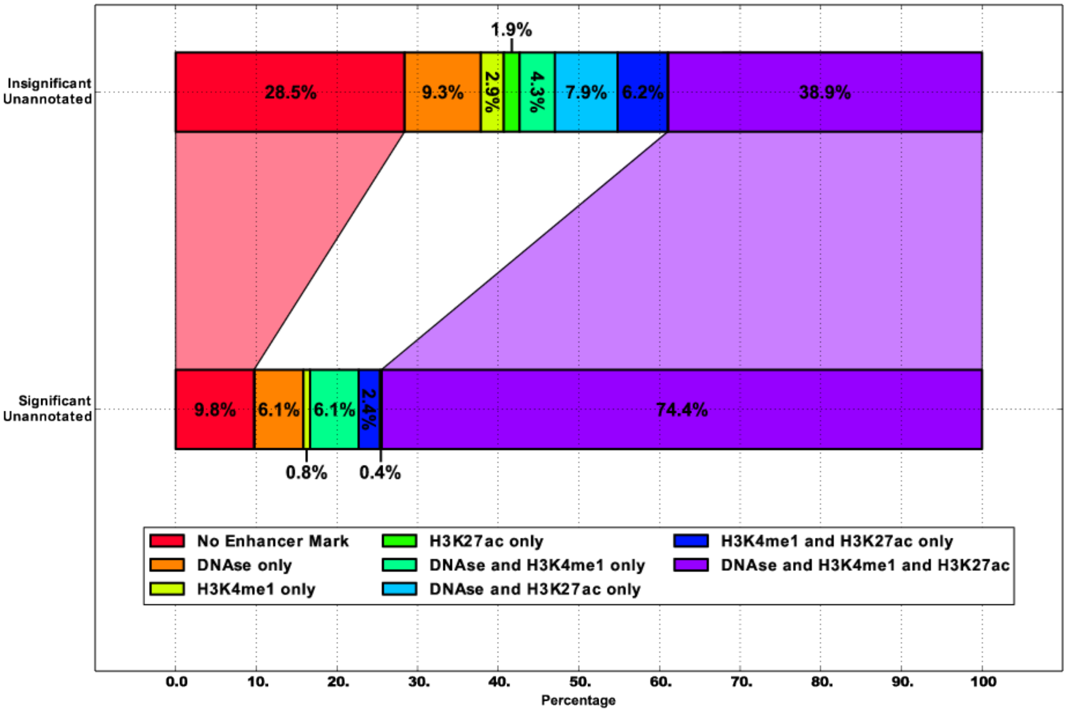


Figure 2.12: **Overlap of differential transcription with enhancer marks.** FStitch calls that do not overlap any RefSeq annotation were grouped by differential transcription by DESeq (significant: adj. p-value < 0.1). Regions were assessed for overlap with enhancer marks: H3k27ac, H3K4me1, and DNase I hypersensitivity[58, 129].

and fast, with classifications easily viewed on common genome browsers.

FStitch determines its *active* calls purely on the signal within the data. In regions of dense and/or overlapping transcription, the gaps between distinct transcripts are short to nonexistent. Consequently, FStitch makes long *active* calls that likely contain multiple transcripts. Additionally, the lack of pre-defined regions of interest complicates the assessment of differential transcription. However, the gains in insight about transcription and regulation warrants the added complexity.

Using FStitch, we learned several interesting new features of transcription at previously annotated genes. Indeed, gene transcription progresses much farther than the $3'$-end of the mRNA cleavage site. Remarkably, some of the *active* calls that are unannotated show signatures of open reading frames, implying they may be under-appreciated genes.

More work is needed to better resolve the transcriptional dynamics observed within genes, such as the $5'$ and $3'$ peaks. These peaks are reminiscent of patterns seen in unstranded Pol II ChIP data and likely correspond to distinct stages of RNA polymerase activity[59]. Unfortunately, the height and spread of these peaks vary from gene to gene, making their precise detection difficult. Chapter 3 deals almost entirely in building models to deconvolve this peak signal from gene body elongation.

Reanalysis of a Nutlin experiment displayed widespread differentially expressed FStitch calls. I showed that these differentially expressed nascent transcripts were extremely likely to overlap well established p53 binding sites (by ChIP). Given that transcription factors, such as a p53, bind DNA in a sequence specific fashion, it might be possible to combine models of GRO-seq with DNA sequence (TF motif) information to infer changes in TF-activity following a stimulus. Chapter 4 explores this hypothesis by monitoring eRNA profiles across different treatments and cell types. In conclusion, this work demonstrates that GRO-seq (in conjunction with a tool to identify nascent transcripts) can be a rich source of insights into transcription and its regulation.

Chapter 3

Stochastic models of RNA Polymerase

Portions of this chapter are adapted from:

> **J. Azofeifa** and R. Dowell (2016) *A generative model for the behavior of RNA polymerase* Bioinformatics
>
> M. Lladser, **J. Azofeifa**, M. Allen and R. Dowell (2016) *RNA Pol II transcription model and interpretation of GRO-seq data* Journal of Mathematical Biology

3.1 Introduction

Gene expression requires RNA Polymerase II (RNAP) recruitment to promoters and subsequent signaling cues to direct RNAP to fully transcribe the protein coding region [15]. With the advent of high throughput sequencing data, RNAP's location has been profiled genome-wide providing deep insight into the enzymatic stages of transcription. In brief, RNAP recruitment, initiation, pause, pause release, elongation and termination are highly controlled transcriptional stages that are distinctly regulated [91, 76, 128].

Nascent transcription assays, such as Global Run-On (GRO-seq), Precision Nuclear Run-on (PRO-seq) and Native Elongating Transcript (NET-seq), measure the production of transcripts from all cellular RNA polymerases genome-wide [33, 91, 128]. Given their high degree of resolution and strand specific nature, these assays have tremendous potential to refine my understanding of each stage of the transcription process. Indeed, these assays have been used to study the transition from paused to elongating polymerase, enhancer RNA transcription and sites of RNAP termination [54, 9]. Yet the precision of these techniques depends on an inherently noisy sequencing process with biases in both the experimental protocol [91] and read mapping strategies. To fully explore the richness of nascent transcriptional assays requires the development of biologically motivated models of RNAP that provide meaningful summary statistics of the data.

To identify nascent transcripts within nascent transcription data, work has primarily focused on segmentation based algorithms such as hidden Markov models (HMMs) [11, 27] and windowing approaches [4]. Indeed, chapter 2 of this thesis outlined a nascent transcript annotation method, FStitch. These "transcribed regions" share similar statistical properties such as comparable levels of mapped reads. However, mapped reads within these regions are distinctly non-stationary. Seen commonly in chromatin immunoprecipitation followed by sequencing (ChIP-seq) for RNAP, "peaks" of GRO-seq mapping occur at both promoter proximal and enhancer regions [124]. In fact, traditional segmentation analysis tend to group these visually distinct elements into one long classification. Consequently, "transcribed regions" often do not correspond to individual transcripts.

Within transcribed regions, the behavior of polymerase lends itself to substructure. For example, the initiating form of polymerase pauses and produces bidirectional peak signatures upstream the gene body [91, 15]. Several recent efforts have focused on identifying the bidirectional transcripts characteristic of initiating/paused RNAP using supervised learning approaches such as naive Bayes [115], support vector regression [38] or logistic regression [11]. Although each approach shows promise, these classifications lack an easy biological interpretation as learned regression coefficients do not directly represent a biological process. Furthermore, methods that focus solely on the bidirectional peak signal fail to capture the productive elongation stage of transcription. In large part, the following chapter outlines the theory and parameterization of mixture models to deconvolve the separate stages of RNA Polymerase activity.

3.2 Modeling RNA Polymerase

Evident in Figure 3.1, their exist two forms of RNAP: (1) The initiating or paused form of RNA Polymerase (immediately subsequent to RNAP loading) and (2) the elongating stage where RNAP actively transcribes the gene body into a nascent transcript. From here on out, I'll refer to

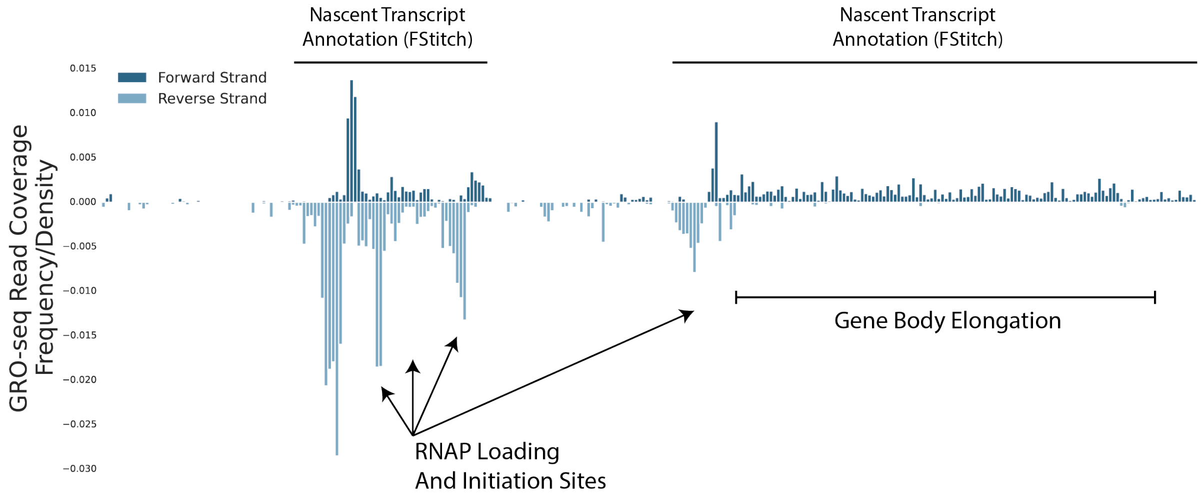


Figure 3.1: **GRO-seq and Nascent Transcript Annotation (Degenerate Case)**GRO-seq read coverage from an HCT116 cell line suggests that the nascent transcripts annotated by FStitch are not reflecting the underlying non-stationarity of the data and dynamics of RNAP.

Z as the random variable that represents the genomic position of RNA polymerase and this section will serve to outline in isolation three models of Z: the double geometric, exponentially modified gaussian and the generic poisson point process. The following section will then construct a mixture model from weighted linear combinations of these individuals models.

3.2.1 Double Geometric Distribution

3.2.1.1 Description

Depicted graphically in Figure 3.2, RNAP is first recruited to some critical site in the genome μ. The parameter μ might represent an annotated transcription start site or perhaps even the the site of sequence specific binding of a transcription factor. In any case, according to a model we (Dowell group and Dr. Manuel Lladser) published in early 2016, RNAP will always bind upstream from μ. This constraint was largely influenced by the fact that the GRO-seq data we were investigating (*Drosophila melanogaster*) did not show stranded, bidirectional transcription. Let U indicate the upstream displacement of RNAP from μ, one might model this random variable as a geometric distribution with some success probability u. Immediately following upstream loading, RNAP processes 5-prime to 3-prime, transcribing a small and abortive transcript, detectable by GRO-seq. Let D be the random variable governing this downstream walk, we chose to model this as a Markov chain or geometric distribution with success probability d.

$$Z \sim \mu - U + D \tag{3.1}$$

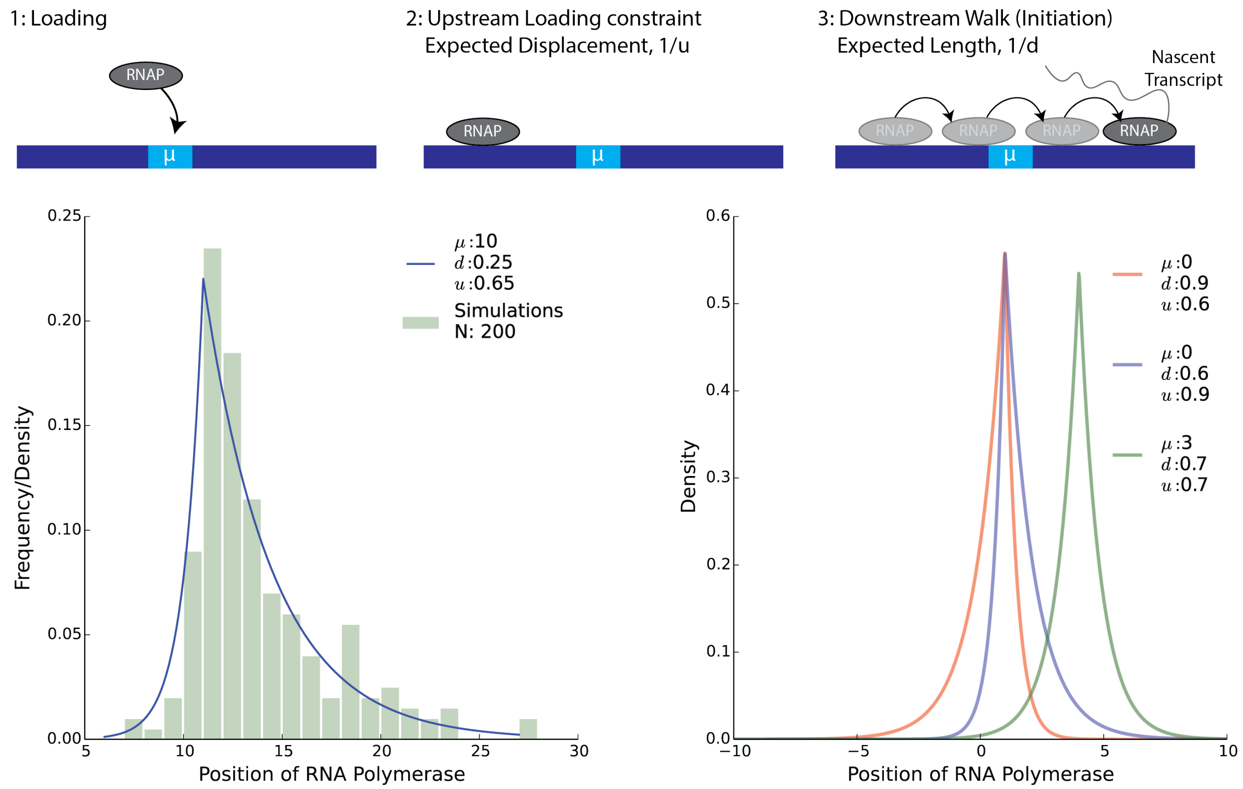


Figure 3.2: **Model of RNA polymerase loading and initiation for *Drosophila melanogaster.*** The above cartoon depicts the schematic of the two one dimensional random walks with the associated density functions referred to as an asymmetric double Geometric distribution. In this way, the height of the density function follows the probability of an RNA polymerase molecule present at that genomic coordinate.

A random variable Z is said to have an (asymmetric) Double Geometric distribution with parameters (u, d) when it has the same distribution as $(-U) + D$, where U and D are independent Geometric random variables with means $(1/u)$ and $(1/d)$, respectively. In particular, solving for the convolution of $D - U$ the probability mass function of Z is given in equation 3.2.1.1, importantly

Z takes discrete integer values i.e. $Z \in \mathbb{Z}$.

$$p_{\mu,u,d}(z) = \frac{ud}{u+d-ud} \cdot \begin{cases} (1-u)^{\mu-z} & \text{for} \quad z \leq \mu; \\ (1-d)^{z-\mu} & \text{for} \quad z \geq \mu. \end{cases}$$

3.2.1.2 Parameter Estimation

With a model of the initiating form of RNA polymerase, I would like to perform estimation of μ, u, d given GRO-seq read coverage data. Let $\mathbf{D} = \{z_1, z_2, .., z_n\}$ be a set of aligned GRO-seq reads at some genomic coordinate of interest (this could be a genic locus or nascent transcript annotation via FStitch), then the likelihood function I seek to optimize is provided in equation 3.2.

$$l(\mu, u, d|\mathbf{D}) = \prod_{i=1}^{N} p(z_i); \mathcal{L}(\mu, u, d|\mathbf{D}) = \sum_{i=1}^{N} \ln p(z_i) \tag{3.2}$$

In the case where $u = d = p$ there exists an analytic solution to the optima of $l(\mu|D, u, d)$ or $l(\mu|D, p)$.

$$\begin{aligned} \mathcal{L}(\mu, p|\mathbf{D}) &= \sum_{i=1}^{N} \ln \frac{p^2}{2p-p^2} + |\mu - z_i|(1-p) \\ \frac{\partial}{\partial \mu}\mathcal{L}(\mu, p|\mathbf{D}) &= \sum_{i=1}^{N} sgn(\mu - z_i)(1-p) \equiv 0 \\ \text{where } sgn(x) &= \mathbb{I}(x > 0) - \mathbb{I}(x < 0) \end{aligned} \tag{3.3}$$

If N is odd then $\hat{\mu}$ is the median of my data observation. If N is even, the median statistic is not uniquely defined and any number between the middle pair of $\mathbf{D}$ will optimize equation 3.3, however it is common practice to take the center value of this pair.

In the case where $d \neq u$ and μ is known, there still exists—although a somewhat more involved—analytic expression to the optima of equation 3.2. This proof largely follows that in the Lladser 2016 paper but with minor exceptions that I think makes the notation a little easier.

$$\mathcal{L}(u, d|\mathbf{D}, \mu) = \ln \frac{ud}{u+d-ud} \Big[\sum_{z_i : z_i \leq \mu} (\mu - z_i) \ln(1-u) + \sum_{z_i : z_i \geq 0} (z_i - \mu) \ln(1-d) \Big] \tag{3.4}$$

Define $\alpha_\mu = \sum_{z_i : z_i \leq \mu} z_i$ and $\beta_\mu = \sum_{z_i : z_i \geq \mu} z_i$ and I solve for the gradient in equation 3.5.

$$\begin{aligned} \frac{\partial}{\partial u}\mathcal{L}(u,d|\mathbf{D}) &= \frac{1}{u} - \frac{1-d}{u+d-ud} - \left[\frac{\alpha_\mu}{1-u}\right] \\ \frac{\partial}{\partial d}\mathcal{L}(u,d|\mathbf{D}) &= \frac{1}{d} - \frac{1-u}{u+d-ud} - \left[\frac{\beta_\mu}{1-d}\right] \end{aligned} \tag{3.5}$$

Equation 3.5 shows two coupled non-linear equations in u and d. Even still, after a fair bit of algebra I can rewrite equation 3.5 to highlight some common features between the two (equation 3.6).

$$\begin{aligned} \alpha_\mu &= (1-u)\frac{d}{u}\frac{1}{u+d-ud} \\ \beta_\mu &= (1-d)\frac{u}{d}\frac{1}{u+d-ud} \end{aligned} \tag{3.6}$$

Using the common factor $\frac{u}{d}$, we can combine these two expressions as the product of α_μ and β_μ and solve for the roots of this polynomial. Note this also assumes that invertibility of α_μ and β_μ in this way, we must assume that $d \in (0,1)$ and $u \in (0,1)$ are in the closed unit interval. Finally, the last step is just the simple use of the quadratic formula (equation 3.7).

$$\begin{aligned} \alpha_\mu\beta_\mu &= (1-u)(1-d)\frac{1}{(u+d-ud)^2} \\ 0 &= \frac{1}{(u+d-ud)^2} - \frac{1}{(u+d-ud)} - \alpha_\mu\beta_\mu \\ u+d-ud &= \frac{1+\sqrt{1+4+\alpha_\mu\beta_\mu}}{2} \equiv \gamma_\mu \end{aligned} \tag{3.7}$$

Also note that, $d - ud = \gamma_\mu - u$ and finally I can solve for u and d in isolation which finishes the MLE computation.

$$\begin{aligned} \alpha_\mu u &= (\gamma_\mu - u)\frac{1}{\gamma_\mu}; \hat{u} = \frac{1}{\alpha_\mu + 1/\gamma_\mu} \\ \beta_\mu d &= (\gamma_\mu - d)\frac{1}{\gamma_\mu}; \hat{d} = \frac{1}{\beta_\mu + 1/\gamma_\mu} \end{aligned} \tag{3.8}$$

Although the MLE solution in equation 3.8 seems a bit esoteric, upon reflection we can gain some intuition. Namely, the statistics of the data—$\alpha_\mu, \beta_\mu, \gamma_\mu$—all depend on μ. In this way, I can see a the realization of a kind of skew computation; i.e. if $\alpha_\mu \gg \beta_\mu$ then it is likely the case that $1/u \gg 1/d$ or $u \ll d$.

3.2.1.3 Discussion

The asymmetric double geometric (aDG) distribution is a useful first step towards a stochastic model of RNAP. First of all, the distribution captures a natural skew observed in GRO-seq data i.e. u is in no way required to equal d. More so, the aDG model interprets Z as a discrete random variable. This is intuitive from a data-driven standpoint as GRO-seq read coverage is indexed by position along the genome and under a genomic coordinate system these are discrete values.

Not without its short comings, the aDG approach to GRO-seq is an un-stranded model. This is a particularly severe disadvantage given that GRO-seq is in fact a stranded procedure and most organisms show bidirectional transcription. For example, the initiating peak of forward strand genes appear positively skewed and reverse strand genes appear negatively skewed. Yet an aDG model will not take such biological knowledge into account.

A product of mathematical convenience, there is no biological reason that RNAP should load *always* upstream μ. More so, a closed form or numerical solution to the estimation $\hat{\mu}$ does not exist. Within the Lladser 2016 paper, I computed $\hat{\mu}$ by brute force search over each possible μ; using the MLE computation of u and d outlined above. Taking these concerns into mind, I developed a new model of RNA polymerase II loading and initiation discussed next.

3.2.2 Exponentially Modified Gaussian

3.2.2.1 Description

Here I present a probabilistic model of transcription that captures both the position Z and bound strand S of RNAP (Fig. 3.5) while relaxing some of the assumptions made within the Double Geometric. At protein coding genes, RNAP is first recruited to the promoter region at the transcriptional start site (TSS). I model the loading position X as a Gaussian distributed random variable with parameters μ, σ^2 where μ represents the typical loading position and σ^2 the amount of error in recruitment to μ. Upon recruitment, RNAP selects and binds to either the forward or reverse strand which I characterize as a Bernoulli random variable S with parameter π.

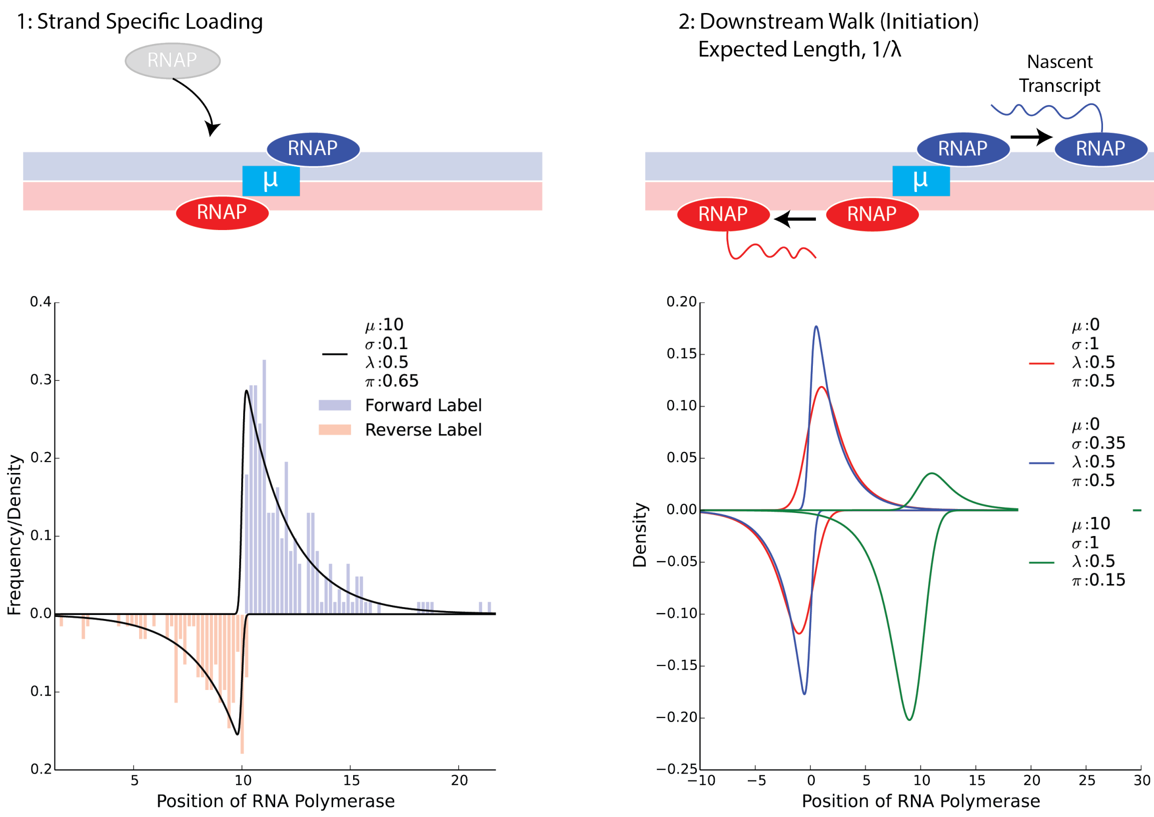


Figure 3.3: **Model of polymerase activity.** A summary of the probabilistic model (on left, see text for full description of parameters) with examples of data generated from the model (on right). Here "Loading" refers to recruitment of polymerase and pre-initiation complex formation, "Initiation" refers to initiation of transcription and promoter-proximal pausing, and "Elongation" refers to productive elongation following pause release [59, 1, 98, 76].

Following loading and pre-initiation, RNAP immediately escapes the promoter and transcribes a short distance, Y. I assume that the initiation distance (also referred to as entry length [76]), is distributed as an exponential random variable with rate parameter λ. For paused polymerase, the final genomic position Z of RNAP is a sum of two independent random variables (equation 3.9).

$$Z|S \sim X + SY \tag{3.9}$$

In equation 3.9, $S \in \{-1, +1\}$ represents the reverse and forward strand decision respectively. Since RNAP processes in a $5' \rightarrow 3'$ direction, S also encodes the signed displacement away from μ. I solve these convolutions analytically and provide a properly normalized probability distribution function (equation 3.10) governing the loading position and entry length of RNAP.

$$
\begin{aligned}
h(z,s;\mu,\sigma,\lambda,\pi) &= \lambda\phi(\frac{z-\mu}{\sigma})R(\lambda\sigma - s\frac{z-\mu}{\sigma})\mathbb{1}(s) \\
\mathbb{1}(s) &= \begin{cases} \pi & : s = +1 \\ 1-\pi & : s = -1 \end{cases}
\end{aligned}
\tag{3.10}
$$

From equation 3.10, $\phi(\cdot)$ denotes the standard normal distribution and $\mathbb{1}(\cdot)$ an indicator function. $R(\cdot)$ represents the *Mill's ratio* which is defined as $(1-\Phi(\cdot))/\phi(\cdot)$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian density. To note, the functional limits of $h(z,s)$ as $1/\lambda$ and σ tend to zero are Gaussian and exponential density functions respectively. For these reasons, $h(z,s)$ has been referred to as an exponentially modified Gaussian [142].

3.2.2.2 Moment Estimation

As is the case with the asymmetric distribution, I would like perform inference over μ,σ,λ and π given set of aligned GRO-seq reads. Let $\mathbf{D} = \{(z_1,s_1),(z_2,s_2),...,(z_n,s_n)\}$ be a independent and identically distributed set of random variables with density function outlined in equation 3.10; $h(z,s;\mu,\sigma,\lambda,\pi)$. The simplest statistic of $\mathbf{D}$—the proportion of forward strand binding RNAP—provides a straightforward moment estimate for π (equation 3.13).

$$
\begin{gathered}
N_+ = \sum_i^N \mathbb{1}(s_i > 0) \qquad N_- = \sum_i^N \mathbb{1}(s_i < 0) \\
\hat{\pi} = N_+/N
\end{gathered}
\tag{3.11}
$$

Indeed, the expected value of $\hat{\pi}$ is unbiased ($\mathbb{E}[\hat{\Theta}] - \Theta = 0$) and as $N \to \infty$ I see the variance converges to zero equation 3.12 .

$$
\begin{aligned}
\mathbb{E}[\hat{\pi}] &= \frac{1}{N}\mathbb{E}[\sum_i^N \mathbb{1}(s_i > 0)] = \frac{1}{N}\sum_i^N \mathbb{E}[\mathbb{1}(s_i > 0)] \\
&= \frac{1}{N}N\pi = \pi \\
\mathrm{Var}(\hat{\pi}) &= \frac{1}{N^2}\mathrm{Var}(\sum_i^N \mathbb{1}(s_i > 0)) = \frac{1}{N^2}\sum_i^N \mathrm{Var}(\mathbb{1}(s_i > 0)) \\
&= \frac{1}{N}\pi(1-\pi)
\end{aligned}
\tag{3.12}
$$

With Z as the sum of two independent random variables, the $\mathbb{E}[Z] = \mu + (2\pi - 1)/\lambda$ and so statistical moments (lacking strand information) will not decouple μ, σ and λ. However, because I have knowledge of strand information, s_i, I can in fact estimate μ and λ based solely on the first moment of the data.

$$\bar{Z}_+ \equiv \frac{\sum_{i=1}^N z_i \mathbb{I}(s_i > 0)}{N_+} \qquad \bar{Z}_- \equiv \frac{\sum_{i=1}^N z_i \mathbb{I}(s_i < 0)}{N_-}$$
$$\begin{aligned} \hat{\mu} =& \frac{1}{2}(\bar{Z}_+ + \bar{Z}_-) \\ 1\hat{/}\lambda =& \frac{1}{2}\Big|\bar{Z}_+ - \bar{Z}_-\Big| \end{aligned} \tag{3.13}$$

To show unbiased properties of $\hat{\mu}$ and $\hat{\lambda}$, I proceed in the same fashion as I did for $\hat{\pi}$ (equation 3.14). Note, N_+ corresponds to the number of forward strand binding RNAP enzyme where any specific molecule will bind to forward strand with probability π. Subsequently, N_+ is distributed as a binomial distribution with parameters π and $N = |\mathbf{D}|$ and conversely N_- is distributed as a binomial distribution with parameters $1 - \pi$ and $N = |\mathbf{D}|$.

$$\begin{aligned} \mathbb{E}[\hat{\mu}] &= \frac{1}{2}(\mathbb{E}[\bar{Z}_+] + \mathbb{E}[\bar{Z}_-]) \\ &= \frac{1}{2}\Big(\mathbb{E}\left[\frac{1}{N_+}\right] \sum_{i=1}^N \mathbb{E}[z_i]\, \mathbb{E}[\mathbb{I}(s_i > 0)] + \mathbb{E}\left[\frac{1}{N_-}\right] \sum_{i=1}^N \mathbb{E}[z_i]\, \mathbb{E}[\mathbb{I}(s_i < 0)]\Big) \\ \mathbb{E}[1\hat{/}\lambda] &= \frac{1}{2}|\, \mathbb{E}[\bar{Z}_+] - \mathbb{E}[\bar{Z}_-]| \\ &= \frac{1}{2}\Big|\, \mathbb{E}\left[\frac{1}{N_+}\right] \sum_{i=1}^N \mathbb{E}[z_i]\, \mathbb{E}[\mathbb{I}(s_i > 0)] - \mathbb{E}\left[\frac{1}{N_-}\right] \sum_{i=1}^N \mathbb{E}[z_i]\, \mathbb{E}[\mathbb{I}(s_i < 0)]\Big| \end{aligned} \tag{3.14}$$

Although tempting, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[x])$ or— specific to this case—$\mathbb{E}[1/N_+] \neq 1/\,\mathbb{E}[N_+]$. However by use of a first-order Taylor series expansion, I can assert that $\mathbb{E}[1/N_+] \approx 1/\,\mathbb{E}[N_+]$ (equation 3.15).

Taylor Expansion about x_0

$$g(x) = g(x_0) + \frac{g'(x_0)}{1!}(x - x_0) + \frac{g''(x_0)}{2!}(x - x_0)^2 + ...$$

Define and note:

$$g(X) = 1/X; g'(X) = -1/X^2; x_0 = \mathbb{E}[X] \tag{3.15}$$

Then:

$$\begin{aligned}
g(X) &\approx \frac{1}{\mathbb{E}[X]} - \frac{1}{\mathbb{E}[X]^2}(X - \mathbb{E}[X]) \\
\mathbb{E}[\frac{1}{X}] &\approx \mathbb{E}[\frac{1}{\mathbb{E}[X]}] - \mathbb{E}[\frac{1}{\mathbb{E}[X]^2}(X - \mathbb{E}[X])] \\
&= \frac{1}{\mathbb{E}[X]} - \frac{1}{\mathbb{E}[X]^2}\,\mathbb{E}[(X - \mathbb{E}[X])] = \frac{1}{\mathbb{E}[X]}
\end{aligned}$$

To summarize, using the fact that $\mathbb{E}[1/N_+] \approx 1/N\pi$ and $\mathbb{E}[1/N_-] \approx 1/N(1-\pi)$ I can finish the proof and show that $\hat{\mu}$ and $\hat{\lambda}$ are approximately unbiased, equation 3.16.

$$\begin{aligned}
\mathbb{E}[\hat{\mu}] &= \frac{1}{2}\Big(\frac{1}{N\pi}N\pi(\mu + 1/\lambda) + \frac{1}{N\pi}N\pi(\mu - 1/\lambda)\Big) = \mu \\
\mathbb{E}[1\hat{/\lambda}] &= \frac{1}{2}\Big|\frac{1}{N\pi}N\pi(\mu + 1/\lambda) - \frac{1}{N\pi}N\pi(\mu - 1/\lambda)\Big| = 1/\lambda
\end{aligned} \tag{3.16}$$

With $\hat{\mu}, \hat{\lambda}, \hat{\pi}$ defined as unbiased estimator, I can finally turn to estimation of σ^2; the variance in RNAP loading at μ. Define S to be the sample variance ($\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{\mu})^2$) then I can estimate σ^2 in terms of S and $1/\hat{\lambda}^2$. Note that by $1\hat{/\lambda^2} \approx 1/\hat{\lambda^2}$ by a similar taylor series expansion outlined above.

$$\begin{aligned}
\hat{\sigma^2} &= S - \frac{1}{\hat{\lambda}^2} \\
\mathbb{E}[S] &= \mathbb{E}[\sum_{i=1}^{N}(z_i - \hat{\mu})^2] = \sum_{i=1}^{N}\mathbb{E}[(z_i - \hat{\mu})^2] \\
&= \sum_{n=1}^{N}\mathbb{E}[z_i^2] - 2\,\mathbb{E}[z_i]\,\mathbb{E}[\hat{\mu}] + \mathbb{E}[\hat{\mu}]^2 = \sum_{i=1}^{N}\mathrm{Var}(z_i) \\
&= \frac{1}{N}N(\sigma^2 + 1/\lambda^2) \\
\mathbb{E}[\hat{\sigma^2}] &= (\sigma^2 + 1/\lambda^2) - 1/\lambda^2 = \sigma^2
\end{aligned} \tag{3.17}$$

Previous efforts at estimating μ, σ^2, λ required at least third-moment information from D[**?**]. However, I have presented a slightly different model where $\pi \in (0, 1)$ and thus I require only first and second-order moment information to compute unbiased estimators of μ, σ^2 and λ. Importantly however, if $\pi = 1$ or $\pi = 0$ the above expressions are completely degenerate and no longer share nice asymptotic properties of convergence in probability. Even still, these statistics are easy to compute and can be used as starting points to a more accurate numerical algorithm (Expectation Maximization) to compute MLE estimates of $\hat{\mu}, \hat{\sigma}, \hat{\lambda}$.

3.2.2.3 Maximum Likelihood Estimation

Although moment estimators display nice asymptotic properties, the estimators so far discussed will fluctuate wildly as the strand bias $\pi \to 0$ or $\pi \to 1$. Moreover, this methodology will be useless when I progress to a mixture model setting where there may be potentially many overlapping Loading/Initiation sites and elongation regions where I can make *no longer an identically distributed* assumption. In brief, this section will serve to develop a numerical method to compute maximum likelihood estimators (MLE) for $\mu, \sigma, \lambda, \pi$ in the case where $\mathbf{D}$ is i.i.d. from $h(z, s; \mu, \sigma, \lambda, \pi)$ and to prepare for the final mixture model setting.

In the usual fashion I can define the (log-)likelihood function of $\Theta = \{\mu, \sigma, \lambda, \pi\}$ given my set i.i.d. observations $\mathbf{D}$ in equation 3.18.

$$\begin{aligned} \mathcal{L}(\Theta|\mathbf{D}) &= \prod_{i=1}^{N} \lambda\phi(\frac{z_i - \mu}{\sigma})R(\lambda\sigma - s_i\frac{z_i - \mu}{\sigma})\pi^{\mathbb{I}(s>0)}(1-\pi)^{\mathbb{I}(s<0)} \\ \ln \mathcal{L}(\Theta|\mathbf{D}) &= \sum_{i=1}^{N} \ln[\lambda\phi(\frac{z_i - \mu}{\sigma})R(\lambda\sigma - s_i\frac{z_i - \mu}{\sigma})] + \mathbb{I}(s>0)\ln\pi + \mathbb{I}(s<0)\ln(1-\pi) \end{aligned} \tag{3.18}$$

Proceeding with the easiest estimator first (π) I see that maxima of equation 3.18 with respect to π (equation 3.19) is exactly the moment estimator in equation 3.13. As is the usual case with likelihood functions, maximizing the log-likelihood maximizes the likelihood function as ln(.) is a monotonic transformation.

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\Theta|\mathbf{D})}{\partial \pi} &= N_+\frac{1}{\pi} - N_-\frac{1}{1-\pi} \equiv 0 \\ &= (1-\pi)N_+ - \pi N_- = N_+ - \pi(N_+ + N_-) \\ \hat{\pi} &= \frac{N_+}{N_+ + N_-} \end{aligned} \tag{3.19}$$

Solving equation 3.18 for optimal μ, λ, σ presents however a different challenge. Most notably, the Mill's ratio which wraps all these parameters into a single expression is defined as $(1 - \Phi(\cdot))/\phi(\cdot)$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian density given in equation 3.20.

$$\Phi(x) = \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \int_{-x/\sqrt{2}}^{x/\sqrt{2}} e^{-t^2} dt \tag{3.20}$$

Indeed, the integral in equation 3.20 is the error function and emits no closed form solution, making optimization of equation 3.18 difficult. In brief, I was unable to decouple parameters in terms of statistics of the data and thus unable to find a closed form solution for this optimization problem. Fortunately however their exists a numerical procedure—the Expectation Maximization algorithm—that is successful in computing MLE estimates of μ, σ^2, λ.

3.2.2.4 EM algorithm for the non-mixture case

In constructing a numerical or iterative procedure in optimizing equation 3.18, I seek an update procedure that monotonically increases $p(\mathbf{D}|\Theta)$ with every iteration i.e. $\mathcal{L}(\Theta^{t+1}) > \mathcal{L}(\Theta^t)$. Here, Θ^t indicates some guess at the parameters $\{\lambda, \mu, \sigma, \pi\}$ and Θ^{t+1} indicates an "updated" estimate on the true Θ.

Due to the convolved normal and exponential distribution, $p(D|\Theta)$ poses a sincere challenge simply because z_i is not uniquely defined—$z_i = x_i + s_i y_i$. Yet knowledge of x_i *or* y_i would effectively decouple the convolution in $p(z_i, s_i|\Theta)$ and—since x_i and y_i are assumed independent—μ, σ and λ could be optimized for in the generic MLE case of a Gaussian and Exponential random variable. Although x_i and y_i are unobserved, I could treat them as random variables and proceed in the usual fashion of a joint distribution (equation 3.21). Due primarily to mathematical convenience, I choose to assume that y_i or $\mathbf{Y} = \{y_1, ..., y_n\}$ is known and therefore "completes" the data. To be clear , $x_i = z_i - s_i y_i$.

$$\mathcal{L}(\Theta) = p(\mathbf{D}|\Theta) = \int_{\mathbb{Y}} P(\mathbf{D}, \mathbf{Y}|\theta) \tag{3.21}$$

Since my goal is to maximize equation 3.21 as a function of Θ, I would like an algorithm that maximizes the difference between $\mathcal{L}(\Theta^{t+1}) - \mathcal{L}(\Theta^t)$. Interestingly, this is the only requirement needed to derive the EM algorithm (equation 3.22).

$$\begin{aligned} \ln \mathcal{L}(\Theta^{t+1}) - \ln \mathcal{L}(\Theta^t) &= \ln \int_{\mathbb{Y}} P(\mathbf{Z}|\mathbf{Y}, \theta^{t+1}) p(\mathbf{Y}|\Theta^{t+1}) - \ln p(\mathbf{Z}|\Theta^t) \\ &= \ln \int_{\mathbb{Y}} P(\mathbf{Z}|\mathbf{Y}, \theta^{t+1}) p(\mathbf{Y}|\Theta^{t+1}) \frac{p(\mathbf{Y}|\mathbf{Z}, \Theta^t)}{p(\mathbf{Y}|\mathbf{Z}, \Theta^t)} - \ln p(\mathbf{Z}|\Theta^t) \\ &\geq \int_{\mathbb{Y}} p(\mathbf{Y}|\mathbf{Z}, \Theta^t) \ln \frac{p(\mathbf{Z}|\mathbf{Y}, \Theta^{t+1}) p(\mathbf{Y}|\Theta^{t+1})}{p(\mathbf{Y}|\mathbf{Z}, \Theta^t) p(\mathbf{Z}|\Theta^t)} \\ &= \delta(\Theta^{t+1}, \Theta^t) \end{aligned} \tag{3.22}$$

The last inequality in 3.22 used the generalized Jensen's inequality to bring the logarithm inside the integral. In brief, since $\ln \mathcal{L}(\theta^{t+1}) \geq \ln \mathcal{L}(\theta^t) + \delta(\Theta^{t+1}, \Theta^t)$ then improving $\delta(\Theta^{t+1}, \Theta^t)$ will also increase $\mathcal{L}(\Theta^{t+1})$. And of course, I would like to make this improvement as large as possible therefore maximizing $\delta(\Theta^{t+1}, \Theta^t)$ with respect to Θ^{t+1} will complete the EM algorithm, equation 3.23.

$$\begin{aligned} \frac{\partial \delta(\Theta^{t+1}, \Theta^t)}{\partial \Theta^{t+1}} &= d\Theta^{t+1} \int_{\mathbb{Y}} p(\mathbf{Y}|\mathbf{D}, \Theta^t) \ln p(\mathbf{D}|\mathbf{Y}, \Theta^{t+1}) p(\mathbf{D}|\Theta^{t+1}) \\ &= d\Theta^{t+1} \int_{\mathbb{Y}} p(\mathbf{Y}|\mathbf{D}, \Theta^t) \ln p(\mathbf{D}, \mathbf{Y}|\Theta^{t+1}) \equiv 0 \end{aligned} \tag{3.23}$$

From equation 3.23, there are now two needed expressions $p(y_i|z_i, s_i, \theta^t)$ and $p(z_i, s_i, y|\Theta^{t+1})$. The later of these two expressions is substantially easier, so I can start with that (equation 3.24).

$$\begin{aligned} \ln p(z_i, s_i, y|\Theta^{t+1}) &= \ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z_i + s_i y - \mu)^2}{2\sigma^2}} \lambda e^{-\lambda y} \\ &= \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(z_i + s_i y_i - \mu)^2}{2\sigma^2} + \ln \lambda - \lambda y \end{aligned} \tag{3.24}$$

Expanding $(z_i + s_i y_i - \mu)^2$, I can see that coupled to the integral of $p(y|z_i, s_i, \Theta^t)$ I will need to compute two expectations $\mathbb{E}[y|z_i, s_i, \Theta^t]$ and $\mathbb{E}[y^2|z_i, s_i]$. Omitting for the moment the expression of the expectation operators, equation 3.23 can now be fully expanded.

$$\begin{aligned} \frac{\partial \delta(\Theta^{t+1}, \Theta^t)}{\partial \Theta^{t+1}} = d\Theta^{t+1} \log \frac{N\lambda}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{N} \Big[\lambda \, \mathbb{E}[Y|z_i, s_i; \theta^t] + \\ \frac{1}{2\sigma^2} \Big({z_i}^2 - s_i \, \mathbb{E}[Y|z_i, s_i; \theta^t] + \mathbb{E}[Y^2|z_i, s_i; \theta^t] - 2\mu(z_i - s_i \, \mathbb{E}[Y|z_i, s_i; \theta^t]) + \mu^2 \Big) \Big] \end{aligned} \equiv 0 \tag{3.25}$$

Completing the differentiation and solving equation 3.25 recovers the "update rules" of the EM algorithm for μ, σ, λ. I can see that in the expectation form they bare a striking resemblance to the form of the MLE in the normal and exponential case in isolation (equation 3.27).

$$
\begin{aligned}
\mu^{t+1} &:= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X|z_i, s_i; \theta^t] \\
\frac{1}{\lambda^{t+1}} &:= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[Y|z_i, s_i; \theta^t] \\
\sigma^{t+1} &:= \frac{1}{N}[\sum_{i=1}^{N} \mathbb{E}[X^2|z_i, s_i; \theta^t] - 2\mu \sum_{i=1}^{N} \mathbb{E}[X|z_i, s_i; \theta^t] + N\mu^2]
\end{aligned} \tag{3.26}
$$

In a verbose manner, equation 3.27 provides the required expectation computation, here $R(.)$ again refers to the Mill's ratio. Note that expectations over X given z_i, s_i and θ^t can be shown easily from the linearity of expectation.

$$
\begin{aligned}
\mathbb{E}[Y|z_i, s_i; \theta^t] &= s_i(z - \mu) - \lambda\sigma^2 + \frac{\sigma}{R(\lambda\sigma - s_i(z_i - \mu)/\ \sigma)} \\
\mathbb{E}[X|z_i, s_i; \theta^t] &= z_i - s_i\, \mathbb{E}[Y|g_i; \theta^t] \\
\mathbb{E}[Y^2|z_i, s_i; \theta^t] &= \lambda^2\sigma^4 + \sigma^2(2\lambda(\mu - z)s_i + 1) + (z_i - \mu)^2 - \frac{\sigma(\lambda\sigma^2 + s_i(\mu - z_i))}{R(\lambda\sigma - s_i(z_i - \mu)/\ \sigma)} \\
\mathbb{E}[X^2|z_i, s_i; \theta^t] &= \mathbb{E}[X|g_i; \theta^t]^2 + \mathbb{E}[Y^2|g_i; \theta^t] - \mathbb{E}[Y|g_i; \theta^t]
\end{aligned} \tag{3.27}
$$

3.2.3 Poisson Point Process

So far, I have discussed models of Z—the genomic position of RNAP—only during the Loading/Initiation phase. Yet subsequent to loading and initiation, RNAP will process $5'$ to $3'$ and transcribe the entire length of the gene body. This stage of RNAP dynamics is termed *elongation.* To model Z under RNAP-elongation requires a fundamentally different model than those governing loading and initiation dynamics.

Figure 3.4 displays my current understanding of RNAP polymerase elongation. Following a signaling cue, RNAP transcribes the entirety of the gene body, halting after some distance L or at some genomic coordinate l_s. Given that previous studies have shown that RNAP is very processive, I reasoned that a statistical model of RNAP should reflect stationarity between $[\mu, \mu + L]$.

Let N_t be the number of GRO-seq reads observed up until some coordinate $\mu + t$. In this way, N_t is right continuous. More so, I would like $p(N_b - N_a = k)$ to be independent of $p(N_d - N_c = k)$

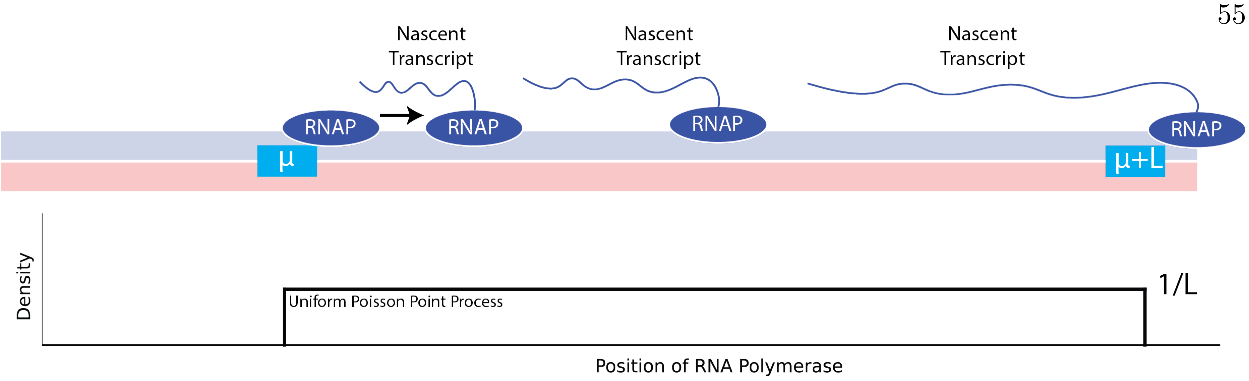


Figure 3.4: **Model of RNA polymerase elongation**. (Top) A cartoon schematic of RNAP processing through the gene body and transcribing the entire length of the protein coding region, referred to as transcriptional elongation. (Bottom) Under a Poisson point process model of RNAP dynamics, the probability that a genomic coordinate (Z) is observed by GRO-seq between $[\mu, \mu + L]$ is a uniform random variable.

given that $(a, b]$ and $(c, d]$ are disjoint intervals. To be absolutely explicit about my notion of statistical stationarity: $p(N_b - N_a = k) = p(N_{b+t} - N_{a+t} = k)$, i.e. the random variable governing nascent transcript identification or RNAP location will not depend on time. This encodes a level of memorylessness into my system.

Although previously defined as a set of i.i.d. random variables, for the moment consider Z as a Poisson process with rate parameter α. If RNAP releases from the DNA at termination sites, $l_s = \{l_f, l_r\}$ for the forward and reverse strand respectively than the position of aligned GRO-seq reads should be i.i.d from a uniform distribution (equation 3.28) between the loading (μ) and termination sites (l_s).

$$\begin{aligned} u(z, s; \mu, l_s) &= \frac{\mathbb{1}(z, s)}{s \cdot (l_s - \mu)} \\ \mathbb{1}(z, s) &= \begin{cases} 1: & sz \geq s\mu; sz \leq sl_s \\ 0: & \text{otherwise} \end{cases} \end{aligned} \tag{3.28}$$

To show that a Poisson process model of RNAP implies a uniform distribution of aligned nascent transcripts, i.e. Z, let $N(l_s)$ correspond to the number of aligned reads on strand s between $[\mu, l_s]$ or $[l_s, \mu]$ for the forward and reverse strand respectively. When $N(l_s) = 1$, this is a straightforward computation (3.29).

$$\begin{aligned} p(z \leq s|N(l_s) = 1) &= \frac{p(z \leq s, N(l_s) = 1)}{p(N(l_s) = 1)} \\ &= p([0, z) = 1, [z, l_s) = 0)/\alpha l_s e^{-\alpha l_s} \\ &= (\alpha z e^{-\alpha z})(\alpha(l_s - z)e^{-\alpha(l_s - z)})/\alpha l_s e^{-\alpha l_s} \\ &= \frac{z}{l_s} \text{ Note: a Uniform CDF} \end{aligned} \tag{3.29}$$

In the case of $N(l_s) > 1$, let $\{U_1, U_2..., U_{N(l_s)}\}$ be a sequence of uniformly distributed random variables and let $\{U_{(1)}, U_{(2)}, ..., U_{(N(l_s))}\}$ be that sequence ordered i.e. $\{U_{(1)} < U_{(2)} < ...\}$. Then the joint probability of observing some ordered list is given by 3.30. Note that there are N_{l_s} factorial ways to order $\{U_1, U_2..., U_{N(l_s)}\}$.

$$p(U_1, ..., U_{N(l_s)}) = N(l_s)!\Big[\frac{1}{l_s}\Big]^{N(l_s)} \tag{3.30}$$

Now, I want to show that in the more general setting ($N(l_s) = n$) the positions of elongating RNAP ($Z = \{z_1, ..., z_n\}$) are uniformly distributed between μ and l_s (equation 3.31) which will justify use of the uniform distribution here on out. In following with the previous notation on order statistics, let $Z = \{z_{(1)}, z_{(2)}, ..., z_{(n)}\}$ be the ordered aligned nascent transcripts and k_i be small enough such that $z_{i-1} + k_i < z_i$ for each $z_i \in Z$. Note that in this way k_i is not the same for all z_i and is itself another random variable!

$$\begin{aligned} &P(\text{one event in } z_i \text{ and } z_i + k \text{ and no events in } [z_i - k, z_i] \text{ given } N(l_s) = n) \\ &= \prod_{i=1}^{n} p(z \in [z_i, z_i + k]|N(l_s) = n) \cdot \prod_{i=1}^{n} p(z \notin [z_i - k, z_i]|N(l_s) = n) \\ &= (\alpha k_1)e^{-\alpha k_1} \cdot ... \cdot (\alpha k_n)e^{-\alpha k_n} e^{\alpha(l_s - (k_1 + ... + k_n))}/(\alpha l_s)^n e^{-(\alpha l_s)}/n! \\ &= n![\frac{1}{l_s}]^n k_1 \cdot ... \cdot k_n \end{aligned} \tag{3.31}$$

Normalizing by $k_1, ..., k_n$ and letting k_i go to zero completes the proof.

Although I discussed the relationship between Poisson Processes and Uniform distributions within the context of continuous random variables, these results hold in the case where z is integer valued and the waiting or arrival times of the point process are geometrically distributed. Indeed, the uniform density and mass functions are equivalent.

3.2.4 Mixture Models

Given that I will be discussing mixtures of RNAP dynamics where Z is no longer identically distributed, I'll refer to the Loading/Initiation/Paused and Elongating/Termination stages as LI and ET respectively.

3.2.4.1 K Exponentially Modified Gaussian and Uniform Mixtures

Nascent transcription assays serve as a readout on RNAP dynamics. Like most high throughput assays, GRO-seq (or PRO-seq) is a population averaged assay, thus providing a histogram reflecting the distribution of RNAP locations. In this way, however, GRO-seq does not directly identify whether a read originated from either the LI or ET stage of polymerase activity.

To capture these processes jointly, let k be a multinomial random variable that records a specific transcriptional component and is selected with probability w_k. Thusly, $\mathbb{K} = \{k \in \mathbb{N}^+ : k \leq \mathbf{M}\}$ represents a finite set of $\mathbf{M}$ transcriptional components. With this in mind, $p(z, s; \Theta)$ represents a mixture distribution describing an arbitrary number of initiation and elongation components (equation 3.32, Figure 3.5)

$$p(z, s; \Theta) = \sum_{k \in \mathbb{K}} w_k f(z, s, k; \theta_k) \tag{3.32}$$

Importantly, $f(z, s, k)$ in equation 3.32 may represent either $h(z, s)$ or $u(z, s)$ (equations 3.10, 3.28 respectively). If $\mathbb{K}_P$ represents the set of LI components and $\mathbb{K}_E$ represents the set of ET components, then $\forall k_e \in \mathbb{K}_E$ there exists a $k_p \in \mathbb{K}_P$ such that μ_k lower or upper bounds the support of k_e depending on the strand orientation of k_e. In this way, LI and ET components are directly linked.

Under a finite $\mathbf{M}$-mixture model, I wish to perform model inference over Θ given nascent transcription data, $\mathbf{D}$ where $N = |\mathbf{D}|$. In total, I seek to identify a parameter set Θ^* under which

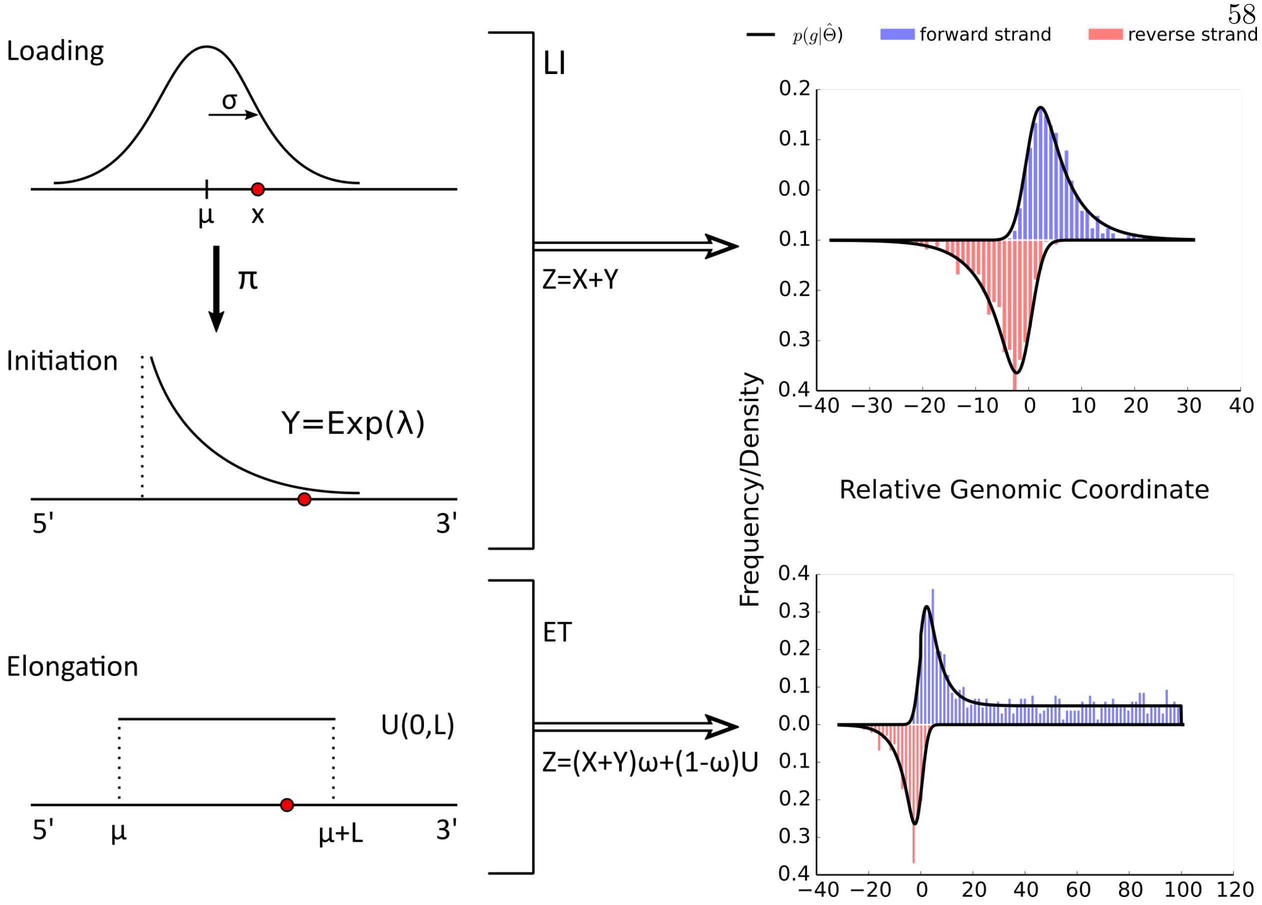


Figure 3.5: **Model of polymerase activity.** A summary of the probabilistic model (on left, see text for full description of parameters) with examples of data generated from the model (on right). Here "Loading" refers to recruitment of polymerase and pre-initiation complex formation, "Initiation" refers to initiation of transcription and promoter-proximal pausing, and "Elongation" refers to productive elongation following pause release [59, 1, 98, 76].

D is most probable, $\mathcal{L}(\Theta|\mathbf{D})$ (equation 3.33), i.e. the maximum likelihood estimate (MLE).

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{N} \sum_{k \in \mathbb{K}} w_k p(g_i; \theta_k) \tag{3.33}$$

Without specifying the set of transcriptional component identifiers (**K**) associated with **D**, equation 3.33 does not emit a closed form solution. Even still, $\{\mathbf{D}, \mathbf{K}\}$ does not fully specify $\hat{\mu}_k$, $\hat{\lambda}_k$ or $\hat{\sigma}_k$ as z_i equals the sum of two latent random variables: x_i (loading position) and y_i (initiating length). However, observing the set of initiating lengths (**Y**) effectively decouples the convolution in equation 3.10 allowing for a straightforward computation of the MLE for a Gaussian and an exponential distribution. Taken together, let the *complete* data be $\mathbf{C} = \{\mathbf{D}, \mathbf{K}, \mathbf{Y}\}$. It follows easily that $\mathcal{L}(\Theta|\mathbf{C})$ has a closed form solution given the assumed independence of z_i, s_i, k_i, y_i.

Although I do not observe $\mathbf{K}$ or $\mathbf{Y}$, I can treat k_i and y_i as random variables and perform iterative optimization of equation 3.33 by the Expectation Maximization algorithm (EM). The EM algorithm alternates between two steps: the **E-step** computes the conditional expectation of latent variables $\{k_i, y_i\}$ given observed variables $g_i = \{z_i, s_i\}$ (equation 3.34) and the **M-step** performs a gradient step along this expectation.

$$\mathbb{E}\Big[\log p(\mathbf{C}|\theta)|\mathbf{D}, \theta^t\Big] = \int\limits_{y \in \mathbb{R}^+} \sum_{k \in \mathbb{K}} \sum_{i=1}^{N} \log p(k_i, y_i, g_i; \theta) \prod_{j=1}^{N} p(y_j, k_j | g_j; \theta^t) \tag{3.34}$$

Admittedly daunting, simplification of equation 3.34 can be achieved in a number of ways. First, I assume that k_i and y_i are independent therefore $p(y_j, k_j|g_j; \theta^t) = p(k_j|g_j; \theta^t) \cdot p(y_j|g_j; \theta^t)$. Furthermore, $p(y_j|g_j; \theta^t)$ integrates to one across $\mathbb{R}^+$ and $\sum_{k \in \mathbb{K}} p(k|g_j; \theta^t)$ sums to one over all $k \in \mathbb{K}$ components. Finally, I need not consider $p(y_i|g_i)$ for mixture components involving elongating polymerase. Therefore, I can see that the complete data log-likelihood function depends only on three quantities: y_i, y_i^2 and k_i.

The probability a component k given a data point g_i (equation 3.35) follows immediately from Bayes' Theorem and for succinctness I define this term as r_i^k. Commonly referred to as the *responsibility term* [17], r_i^k measures the extent to which g_i belongs to some component k.

$$r_i^k = p(k|g_i; \theta_k^g) = \frac{w_k \cdot p(g_i; \theta_k^g)}{\sum_{k \in \mathbb{K}} w_k \cdot p(g_i; \theta_k^g)} \tag{3.35}$$

With the necessary conditional expectations defined, I solve for the maximum of equations 3.34

and 3.25. Equation 3.36 provides the "update rules" for the EM algorithm.

$$
\begin{aligned}
w_k^{t+1} &:= \frac{r_k}{r} \\
\pi_k^{t+1} &:= \frac{\sum_{i=1}^N r_i^k I(s_i = 1)}{r_k} \\
\mu_k^{t+1} &:= \frac{1}{r_k} \sum_{i=1}^N \mathbb{E}[X|z_i, s_i; \theta^t] r_i^k \\
\frac{1}{\lambda_k^{t+1}} &:= \frac{1}{r_k} \sum_{i=1}^N \mathbb{E}[Y|z_i, s_i; \theta^t] r_i^k \\
\sigma_k^{t+1} &:= \frac{1}{r_k} \Big(\sum_{i=1}^N \mathbb{E}[X^2|z_i, s_i; \theta^t] r_i^k - 2\mu_k \sum_{i=1}^N \mathbb{E}[X|z_i, s_i; \theta^t] r_i^k \Big) + \mu_k^2
\end{aligned}
\tag{3.36}
$$

In keeping with the traditional notation of mixture models in equation 3.36, I define $r_k = \sum_{i=1}^N r_i^k$ and $r = \sum_{k \in \mathbb{K}} r_k$.

Due to the finite nature of uniform distributions, my EM update rules (equation 3.36) assume that l_s is fixed, presumably to the minimal or maximal order statistics, g_0 and g_n of $\mathbf{D}$. However, the length of elongation or exact site of termination varies throughout the genome [42]. In this way, a fixed l_s is an unattractive modeling assumption of RNAP.

To optimize l_s requires an adjusted EM algorithm. In brief, I want to preserve the *contractive map* property of the EM namely $|\Theta^{t+1} - \Theta^*| \leq \beta |\Theta^t - \Theta^*|$ where $0 < \beta < 1$ and Θ^* refers to a fixed point of the EM map. Yet, moving l_s away from the max and min order statistics will result in some $g_i \in \mathbf{D}$ having no probability mass ($\mathcal{L}(\Theta) \to 0$) or $\mathcal{L}(\Theta)$ to monotonically decrease.

To estimate for an optimal l_s, I place a uniform distribution over $\mathbf{D}$ with support between $[a, b]$ and $p(s) = 0.5$ for either $s = +1$ or $s = -1$. The mixing weight (w_k) remains fixed at $\min(E/|\mathbf{D}|, 1)$ where E represents the expected number of mapping errors across $[a, b]$. Under a binomial error model assumption, $E = p|G||b - a|/S$ where S refers to the length of the genome, $|G|$ the total number of mapped reads and p represents the probability that a read maps by chance alone. With this addition, l_s is no longer confined to the min and max order statistics of $\mathbf{D}$. For illustrative purposes, I provided a pseudo-code description of my MLE methodology in the Appendix (Algorithm B1).

3.2.4.2 Model Selection

An important limitation of finite mixture models is a-priori knowledge of $|\Theta|$, the number of transcription components. To perform model selection over potentially many component sizes, I utilize penalized Bayesian Information Criterion (BIC), equation 3.37.

$$\mathrm{BIC}(\Theta, \mathcal{L}) = \alpha|\Theta| \log N - 2\log \mathcal{L} \tag{3.37}$$

$\mathcal{L}$ represents the likelihood function evaluated at Θ^*, α is the penalty term, $|\Theta|$ is the number of free parameters within a specific model topology (e.g. one initiation component, one forward strand and one reverse strand elongation component contains 8 free parameters) and N is the total number of data points within $\mathbf{D}$. In brief, BIC penalizes model complexity while balancing improvement in $\mathcal{L}$. Unless otherwise specified, α is set to one for all subsequent analysis.

3.2.4.3 Seeding the EM

Like many gradient based optimization methods, the EM algorithm is subject to local maxima across the likelihood landscape. A common approach to handle this issue is to use many random initializations of $\Theta^{t=0}$, yet this approach is inherently time consuming. Seen commonly with Gaussian mixture models [143], both the EM's linear rate of convergence and final parameter quality, $\mathcal{L}(\Theta^*)$, vary most as a function of μ_k^0: the random initialization of the LI component center.

To this end, I propose an approximate windowing method to scan for regions of local likelihood for an LI component within $\mathbf{D}$. Let $\mathbf{D}_{[a,b]} = \{z_i : a \leq z_i \leq b\}$ be a specific ordered collection of $\mathbf{D}$. At $\mathbf{D}_{[a,b]}$ I compute the ratio of BIC scores of a fully specified single component mixture model (parameters Θ_B) to a uniform distribution with support across $[a, b]$ (equation B.1).

$$\mathrm{LL}(\mathbf{D}_{[a,b]}) = \frac{\log N_{[a,b]} - 2N_{[a,b]} \log(0.5/(b-a))}{8\alpha \log N_{[a,b]} - 2\log \mathcal{L}(\Theta_B)} \tag{3.38}$$

It should be noted that the elongation l_s components l_+, l_- are set to b, a respectively. Overlapping windows of size $b-a$ are merged if LL(.) exceeds 1. Unless otherwise specified, $\alpha = 1$. These merged windows are then used for the random initialization μ_k as they constitute a heuristic estimate to the expected value of an LI component.

3.2.5 Bayesian Extensions

Of particular interest might be to incorporate biologically relevant information on $\mu, \lambda, \pi, \vec{w}$. Information like gene annotation location, gene strand information and previous travelers ratio calculations can all be incorporated in a meaningful way to the model outlined thus far. Here I show that I can perform maximum a-posteriori inference even in light of non-exponential family component densities and provide new update rules for Θ given the incorporation of conjugate priors.

It is simple to show that if X and Y are independent, exponential family random variables, then their product is exponential family. A random variable is exponential family if it can be rewritten as

$$h(x) \exp\Big[\eta^T T(x) - A(\eta)\Big] \tag{3.39}$$

where $h(x)$ is the base measure, $\eta(\theta)$ the natural parameter set, $A(\eta)$ is the log partition function and ensures that the probability function sums to one and $T(x)$ is the sufficient statistic of the data. If X and Y are independent than their joint density is an exponential random variable.

$$p(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda e^{-\lambda y} = \frac{\lambda}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \lambda y\Big] \tag{3.40}$$

Thusly,

$$T(x) = \begin{bmatrix} x \\ x^2 \\ y \end{bmatrix}; \eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \\ \lambda \end{bmatrix}; A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2); h(x) = \frac{\lambda}{\sqrt{2\pi\sigma^2}} \tag{3.41}$$

When μ, σ^2 are unknown for a normal distribution a Normal-Inverse Gamma $(m, \kappa, \alpha, \beta)$ distribution is conjugate. I only have the addition of an extra parameter λ which is completely independent from the loading event X. A Gamma distribution (τ, υ) is fully conjugate for an exponential random variable. A Dirichlet distribution is fully conjugate for $\vec{w}$.

With a simple modification to equation 3.23 I can treat my parameters as random variables and incorporate $p(\Theta)$.

$$R(\Theta, \Theta^g) = \int_{y\in\mathbb{Y}} \sum_{k\in\mathbb{K}} \sum_{i=1}^{N} \log p(k_i, y_i, z_i, s_i|\Theta) \cdot p(\Theta) \prod_{j=1}^{N} p(y_j, k_j|z_j, s_j; \Theta^g) \tag{3.42}$$

I note that the E-step remains unchanged, thus I simply modify the M-step to provide my new updates. Sense the joint log likelihood is exponential family and the probability of loading (X) is independent of (Y) I can use all, fully conjugate priors.

For the mixing components w_k, I use a symmetric Dirichlet prior of the form $\frac{1}{B(\alpha_0)} \prod_i^{2M} w_i^{\alpha_0 - 1}$, a beta prior for the strand probability π of the form $\frac{1}{B(\alpha,\beta)} \pi^{\alpha_{l.e}} (1-\pi)^{\beta_{l,e}}$, a Normal-Inverse Gamma distribution for the prior on μ, σ^2 of the loading event x_i, $\frac{\sqrt{\tau}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{\sigma^2}^{\alpha+1} \exp(-\frac{2\beta+\tau(\mu-m_0)^2}{2\sigma^2})$ and a simple gamma distribution for the prior on initiating length parameter λ. In this way, it is relatively straightforward to incorporate stronger priors on the strand and component weights that may reflect gene annotation.

$$
\begin{aligned}
\mu_k &:= \frac{m_0\tau + \sum_i r_{i,k}\, \mathbb{E}[X|z_i, s_i; \theta^g]}{r^k + \tau} \\
\sigma_k^2 &:= \frac{\sum_{i=1}^N \mathbb{E}[X^2|z_i, s_i; \theta^g] \cdot r_i^k - 2\mu \sum_{i=1}^N \mathbb{E}[X|z_i, s_i; \theta^g] \cdot r_i^k + \mu^2 r^k + 2\beta + \tau(\mu - m_0)^2}{\alpha + 3 + r^k} \\
\frac{1}{\lambda_k} &:= \frac{\sum_{i=1}^N \mathbb{E}[Y|z_i, s_i; \theta^g] \cdot r_i^k + \beta}{r^k + \alpha} \\
w_k &:= \frac{r_k + \alpha_0}{r_k + K\alpha_0} \\
\pi_k &:= \frac{\sum_{i=1}^N r_i^k \cdot I(s_i = 1) + \alpha_l - 1}{r_k + \alpha_l + \beta_l - 2}
\end{aligned}
\tag{3.43}
$$

3.3 Applications to GRO-seq

3.3.1 Numerical confirmation of model inference by simulation

I have presented a probabilistic generative model of RNAP that provides a straightforward method to draw a collection of random samples **D** from $p(z, s; \Theta)$. As shown in Figure 3.5, this model governing RNAP location is generative. The two distinct stages of RNAP, Loading/Initiation (LI) and Elongation/Termination (ET), are illustrated by 250 random draws from $p(z, s; \Theta)$. Qualitatively, this simulated data resembles GRO-seq read pile up seen in previous studies [3, 102, 105, 92].

Using synthetic read data generated from $p(z, s; \Theta)$, I first tested the accuracy and correctness of my proposed parameter estimation procedure. I drew varying sizes of **D** from 10 to 500 data points at equally spaced intervals of 20. Under each sample size, I collected 25 unique subsets from $p(z, s|\Theta)$ and computed the standard error $|\Theta - \hat{\Theta}|$. As shown in Extended Data Figure B.1A, I observe a significant and fast decrease in standard error as $\mathbf{D}$ increases in size.

Lastly, to monitor accuracy in model selection by Bayesian Information Criteria (BIC), I drew from $p(z, s|\Theta)$ under varying levels of model complexity, $|\mathbb{K}| \in \{1..20\}$ (Extended Data Figure B.1B). To quantify variability in accuracy, I drew 25 sets of $\mathbf{D}$ from $p(g|\Theta)$ where $|\mathbf{D}| = 100$. Under each set, I computed MLE estimates for every model topology between 1 to 20 and selected the model with the minimum BIC score. A maximum of 20 is chosen arbitrarily and for computational convenience. In short, my model selection procedure correlates well with the true $|\mathbb{K}|$.

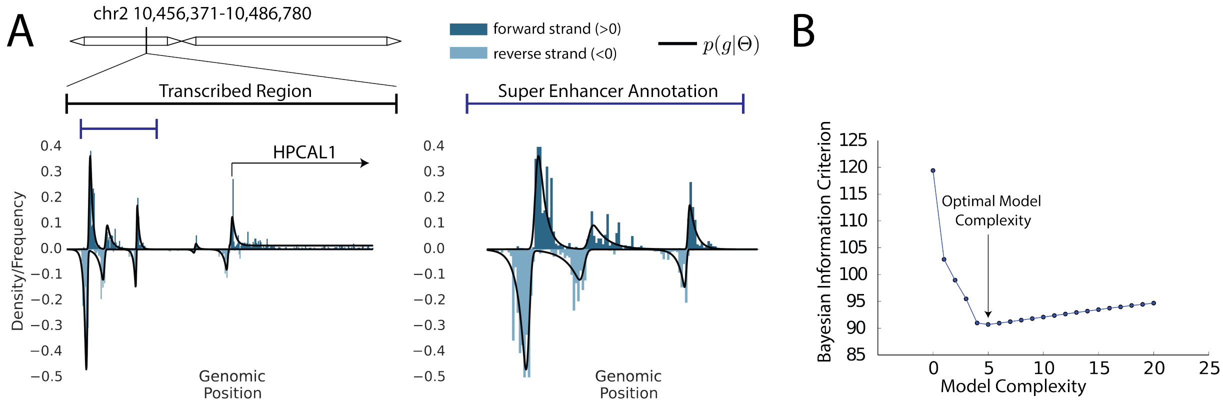


Figure 3.6: **Characteristic loci showing RNAP inference.** (A) shows the final inferred density function at characteristic transcribed and super enhancer regions defined by FStitch [9] and dbSuper [84] respectively. (B) shows the BIC calculation across 20 mixture models. A model complexity of 5 is shown to be minimal and thus considered optimal.

To perform model inference of RNAP location requires an interval $[a, b]$ where GRO-seq read mapping data ($\mathbf{D}$) can be collected. To this end, I utilized Fast Read Stitcher (FStitch) that implements a maximum entropy Markov model to segment the genome into "transcribed regions" [11, 9]. In a HCT116 GRO-seq dataset [3], FStitch classified 19,709 transcribed regions. With these regions in hand, I computed Θ^* by my MLE methodology across mixtures containing $|\mathbb{K}| \in \{1, 2, ..., 20\}$ for each interval independently and selected a final Θ^* by the minimum BIC score.

Table 3.1: **Genome wide summary statistics of inferred $\hat{\Theta}$.** Computations of $\mu - TSS$ and l_s- 3‘end are specific to transcribed regions overlapping a single isoform gene. $\hat{w_p}$ was recomputed for a 2KB window surrounding $\hat{\mu_k}$ as $|l_s - \mu_k|$ will artificially correlate with w_p.

	Mean	Median	Standard Deviation
$\lvert\mathbb{K}_p\rvert$	2.8	2.1	±1.5
$\mu - TSS$	-42	-15	±120
l_s- 3‘end	6325	6194	±1200
σ	38.84	21.46	±65.73
λ	170.15	137.57	±130.15
π	0.51	0.51	±0.29
w_p	0.73	0.79	±0.13

Figure 3.6A shows a transcribed and super enhancer region with reference to the estimated density function. As a final illustrative example, Figure 3.6B displays the associated Bayesian Information Criterion scores as a function of model complexity Supplemental Table B.1 provides a collection of statistics describing the distribution of fitted parameters across all transcribed regions.

To address the accuracy of my model complexity procedure, I reasoned that at active single isoform genes should predict only one LI component while at transcriptionally inactive regions (by FStitch) I should predict no components. Extended Data Figure B.2 highlights the accuracy of my RNAP inference model to discriminate between active and inactive transcribed regions based solely on LI component presence (AUC$\approx$ 0.95). At a FDR of 0.05, I observe that the distribution of $|\mathbb{K}_p|$ at single isoform genes contains a clear and prominent mode at $|\mathbb{K}_p| = 1$ (Extended Data Figure B.2).

The distribution of $|\mathbb{K}_p|$ at both single isoform genes and all FStitch-defined transcribed regions is heavy tailed, suggesting an appreciable number of transcribed regions contain more than one RNAP loading event (Extended Data Figure B.2). To assess whether the model is incorrectly introducing extra LI component centers to compensate for data poorly described by an exponentially modified Gaussian distribution, I compared the distribution of $|\hat{\mu_i} - \hat{\mu_j}|$ $(i \neq j)$ at loci harboring $|\mathbb{K}_p| > 1$ to the distribution of LI component standard deviation $(\hat{\sigma} + 1\hat{/}\lambda)$. I observe that median pairwise LI component center distance far exceeds what I would expect under the variability of LI component sizes, indicating that $\{\hat{\mu}_1, .., \hat{\mu}_k\}$ describe independent portions of the data.

Suggested by the associated standard deviations in Supplemental Table B.1, $\hat{\Theta}$ varies from locus to locus: some genes experience a large degree of initiation ($1/\lambda \gg 0$) or considerable strand bias $\pi \neq 0.5$. Whether this variability relates to experimental noise or actual biological structure, may be addressed by the reproducibility and consistency of $\hat{\Theta}$ across biological replicates. To this end, FStitch defined transcribed regions between replicate one and two were merged and model inference was performed in each replicate independently.

I observe strong correlation in model selection (Extended Data Figure B.4A) and exceedingly high correlation between identical parameters (e.g. $\rho(\lambda_{rep1}, \lambda_{rep2}) = 0.95$, Extended Data Figure B.4B), suggesting that estimated parameters display low variance. Apart from w and l_s, I observe little to no correlation between differing parameters (e.g. λ_{rep1} and π_{rep2}, Extended Data Figure B.4C), suggesting that there is no confounding dependencies between parameters.

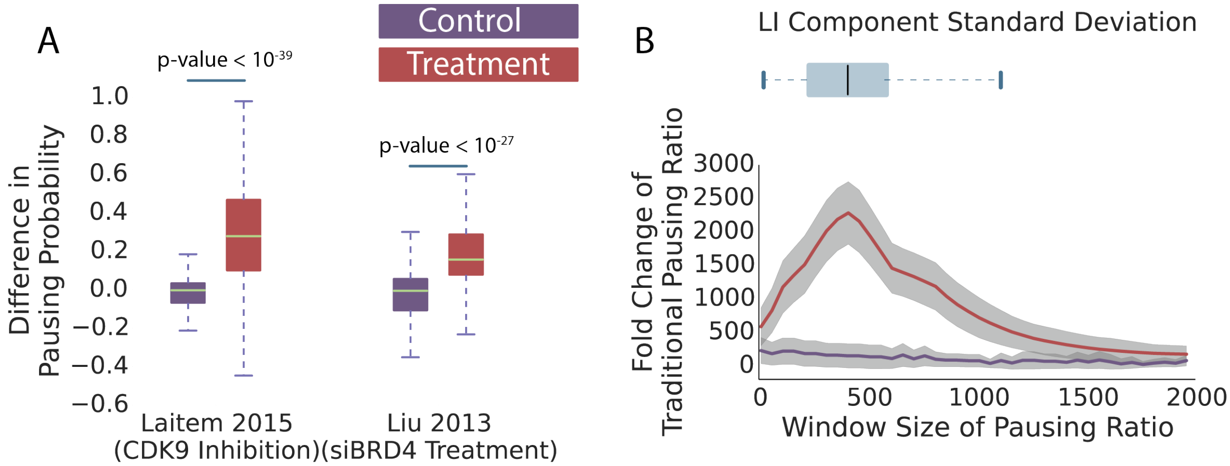


Figure 3.7: **Changes in promoter proximal pausing are correctly identified by a generative model of RNAP.** Mixture model inference was performed over RefSeq gene annotations between control and treated cell lines in two independently derived datasets: Laitem 2015 and Liu 2013. (A) shows the predicted difference in LI mixing weights between control and treated cells. Under a normal assumption, mean mixing weights were compared using a t-test. (B) shows the distribution of LI length (blue box) with the traditional computation of the pausing ratio as a function of window size plotted for the control (purple line) and treated (red line) cell lines of the Laitem 2015 dataset. Grey shading indicates one tenth of one standard deviation.

Estimates of w_p sufficiently larger than the population average (two standard deviations, Supplemental Table B.1) are significantly lacking in an overlapping transcription start site (p-value numerically indistinguishable from zero, Hypergeometric test). Intuitively, this is expected as transcription over enhancer regions or non-coding regulatory loci do not harbor downstream gene bodies. I observed that the loading strand bias (π) tracks closely with the strand orientation of the underlying RefSeq gene (Extended Data Figure B.5A). Particularly, $\bar{\pi} \gg 0.5$ and $\bar{\pi} \ll 0.5$ for forward and reverse strand gene annotations respectively. Loading events lacking an annotated TSS display no appreciable strand bias, $\bar{\pi} \approx 0.5$.

Given my model predicts μ as the site of RNAP loading and l_s as the site of elongating termination, I compared the location of μ and l_s to estimates of annotated transcriptional start sites (TSS) and termination sites respectively. I observed a high degree of correlation between μ and the TSS, noting a significant ≈ 40 base pair upstream displacement of μ from the TSS (Extended Data Figure B.5B). This displacement is in line with estimates from other groups using independent methods [76]. Similar to previous estimates of transcription termination [55, 9], I observed l_s to be $\approx$ 6KB downstream the polyadenylation site (Extended Data Figure B.5C).

3.3.2 Predicting enzymatic changes of RNAP following Experimental Perturbation

Of significant importance to transcriptional studies is to monitor changes in RNAP activity following experimental perturbation. Specifically, the transition between promoter-proximal pausing into the subsequent RNAP elongation constitutes a highly regulated process of tremendous interest [1]. A popular metric to quantify changes in RNAP pausing, the "pausing ratio" computes mapped reads under some TSS-centered window divided by mapped reads under some gene body-centered window. Given that my model directly infers LI and ET RNAP stages from data alone, I ask whether I can correctly identify changes in RNAP activity following experimental perturbation known to affect promoter proximal pausing.

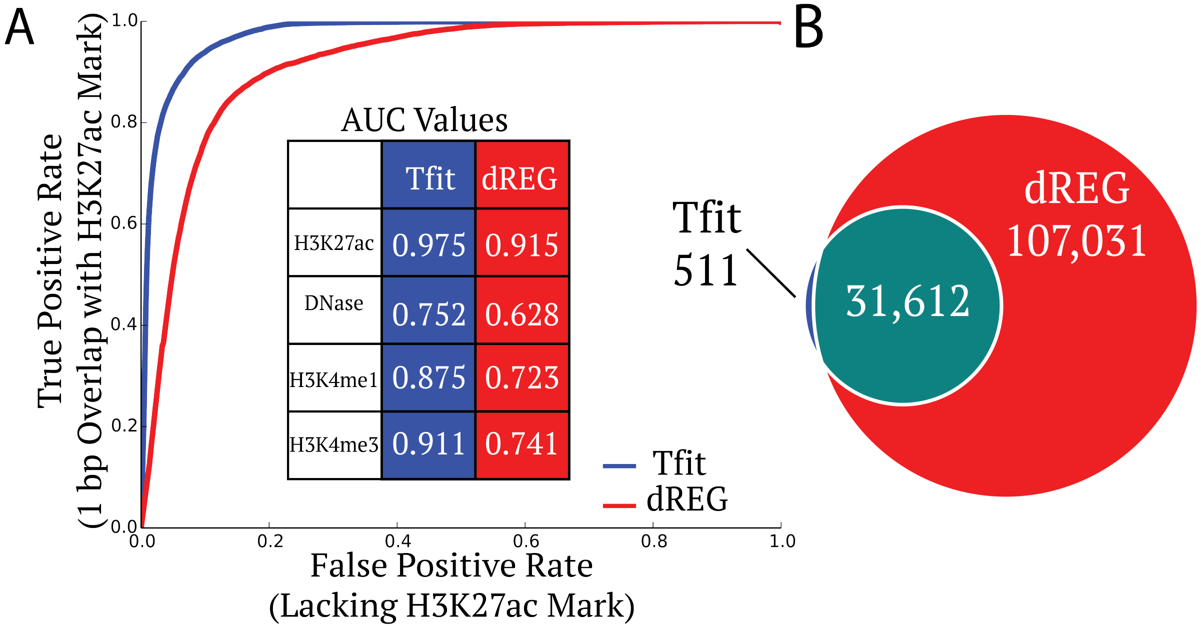
I reanalyzed data from two studies that utilized GRO-seq to probe RNAP pausing activity. Specifically, one study knocked down bromodomain-containing protein 4 (Brd4) in HEK293 cells and observed global changes in RNAP pausing ratios suggesting a critical role in RNAP pause release [105]. Yet another study noted global shifts in RNAP pausing ratios by cyclin-dependent kinase (CDK) 9 inhibition in HeLa cells [92]. With these datasets, I hypothesis that my model should accurately reproduce the observed changes in pausing ratios by appropriate changes in the LI and ET mixing weights.

For each dataset, I performed model estimation and computed changes in estimated parameters between the untreated and treated cells. Specifically, I fit a single component mixture model (one LI component and two ET components) at each gene annotation region. I then compared pairwise fold change in LI component mixing weights first between untreated biological replicates and then between untreated and treated experiments. In both studies, I observed highly significant global changes in LI component mixing weights relative to untreated replicates (Figure 3.7A).

Although easily computable, the standard pausing ratio calculations rely on ad hoc methods of window sizes and distance thresholds [1]. To highlight this point explicitly, I examined the impact of TSS-centered window size on the pausing ratio (Figure 3.7B). Intuitively, window sizes that are either too small or too large dramatically reduce the observable differences between treated and untreated cells. For comparison, I provide the distribution of LI standard deviation obtained from my model.

3.3.3 RNAP model accurately predicts marks of regulatory elements

Beyond annotated genes, it is well known that key chromatin marks are associated with transcription in different parts of the genome [163, 164]. Moderate levels of H3K4me1/2 and high levels of H3K27ac mark active enhancers whereas high levels of H3K4me3 and H3K27ac mark areas of active promoters. Recent studies show that these marks harbor transcription [102] and show a characteristic "bidirectional" signature, where forward and reverse strand read coverage appear positively and negatively skewed respectively. To study the interplay between enhancer transcription and chromatin landscape requires the development of models to accurately identify bidirectional transcripts. Although my model of RNAP does not implicitly assume LI components to appear bidirectional, certain parameter combinations (e.g. $\pi \approx 0.5$ and $1/\lambda \gg 0$) will show both



	Tfit	dREG
H3K27ac	0.975	0.915
DNase	0.752	0.628
H3K4me1	0.875	0.723
H3K4me3	0.911	0.741

Figure 3.8: **RNAP model accurately profiles for bidirectional transcription.** (A) A receiver operating characteristic (ROC) curve displaying the relationship between true and false positive rates of H3K27ac prediction from bidirectional transcription alone. The area under the ROC curve (AUC) values are summarized for multiple marks. As a Venn diagram, (B) shows the overlap in bidirectional transcription classifications between dREG and Tfit at false discovery rate of 0.05 relative to the H3K27ac prediction.

positive and negative skew emanating from μ. To profile for bidirectional transcripts genome wide, I hereafter utilized my EM seeding method (complete description in Supplement; Θ_B set to the parameter values in Supplemental Table B.1).

To asses the accuracy of my bidirectional transcript classifications, I monitored how well a regulatory mark (DNase I HS, H3K27ac, H3K4me1/3) may be predicted from bidirectional transcription alone. With an HCT116 GRO-seq dataset [3], I benchmarked my method against the current state-of-the-art bidirectional detection algorithm, dREG [38]. The BIC penalty α (Tfit) and support vector regression score (dREG) was varied. True positives were considered as an overlap with chromatin mark peaks (HCT116 [164], MACS broad peak settings) and the resulting bidirectional transcript prediction. To assess false positives, I randomly selected an equivalent number of 2KB loci that do not overlap MACS peak calls. Thus, a false positive is a bidirectional prediction overlapping a negative example. I observed improvements over dREG across all regulatory marks (Figure 3.8A-B). Both dREG and Tfit predict H3K27ac marks exceedingly well from bidirectional transcript presence alone suggesting that H3K27ac signal reflects nascent transcription.

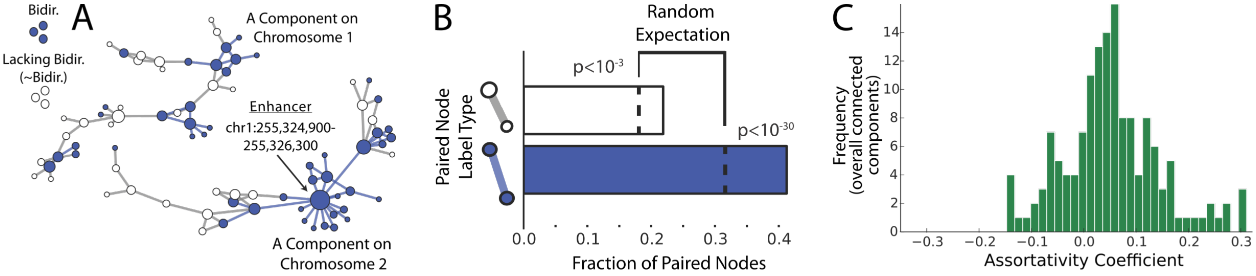


Figure 3.9: **CTCF paired loci network displays assortativity by bidirectional presence** (A) displays two characteristic connected components on chromosome 1 and 2 from a CTCF ChIA-PET dataset derived from the K562 cell line. Nodes are colored as to whether a Tfit prediction overlaps a paired loci by one base pair or not; Bidir. and ∼Bidir. respectively . The circumference of the node is proportional to the degree. (B) shows the proportion of edges containing a similar label, significance is calculated by a Binomial test. (C) shows the distribution of the assortativity coefficient across all connected components, > 0 indicates modularity.

3.3.4 Three dimensionally paired loci display centrality and associativity based on bidirectional transcription

The role of transcription at enhancer elements (defined commonly as non-TSS associated H3K27ac presence) remains an open and exciting question. Correlation in both transcript levels and three dimensional proximity of enhancer elements and target genes [97, 3, 11] point prominently to the functional importance of enhancer transcripts. To begin to address the question of enhancer RNA function, I present an analysis that demonstrates both the utility of my predicted RNAP loading events and the intriguing relationship between chromatin interaction datasets and nascent transcription assays.

Given that the insulator protein CTCF has been implicated as a key player in enhancer to gene looping events [154, 135], I examined the loci-loci pair interaction network defined by Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) for CTCF derived from the K562 cell line [164]. With a cell line matched GRO-seq dataset [34], I compared network attributes of loci containing or lacking bidirectional transcript predictions ($\gamma = 1$, FDR=0.05 according to H3K27ac prediction). Figure 3.9A displays two illustrative connected component examples, built as described in the Supplement.

Enhancers are implicated in defining key cellular phenotypes such as cell fate [37, 176] and tumorigenesis [137, 72] and thus may play a central role in three dimensional looping. my constructed network reflects locations in the genome (nodes) connected by the CTCF ChIA-PET data. With this in mind, I grouped nodes by whether they lacked or contained an association with an annotated TSS or Tfit prediction and computed common measures of node centrality. I observed the highest degree of node centrality at non TSS associated bidirectional transcripts across all criteria (Supplemental Table S2). This result suggests that enhancer RNAs play a central role in the 3D configuration of the genome.

With the advent of chromosome conformation capture technology, extensive chromosomal looping has been observed at so called transcription factories [41]. These discrete nuclear sites of transcription allow for rapid expression of many three dimensionally proximal genes [50]. To this end, I investigated evidence of modularity or network homophily based on lacking or containing bidirectional transcript presence. Indeed, the proportion of edges linking bidirectionally transcribed nodes is much higher than by chance alone (Figure 3.9B). Furthermore, computation of network modularity (a measure of network label clustering) shows weak assortativity across most connected components (Figure 3.9C).

Discussion

I described a probabilistic, generative mixture model that is founded on a biologically motivated description of RNAP behavior. This model is inspired by the current understanding of polymerase behavior [59, 91, 98, 34] at protein coding genes and provides a principled mathematical approach to GRO-seq data analysis. To perform model inference, I derived a parameter estimation scheme based on the theory of maximum likelihood and the Expectation Maximization algorithm. When applied to GRO-seq, position of RNAP loading (μ) corresponded well with previous genomic annotations such as marks of active chromatin (H3K27ac and DHS1) as well as transcription start sites (RefSeq gene promoters). Taken together, this model of RNAP polymerase appears as a parsimonious explanation of GRO-seq read coverage.

Although strongly enriched for RefSeq promoter annotations, the majority of model predictions arose at sites laking a gene transcription start site. The functional role of eRNAs, and even their presence alone, remains an interesting and open question. Apart from a strand bias (π), $\hat{\theta}$ was not drastically altered between promoter or enhancer loci suggesting little difference in transcriptional dynamics. In some ways, this data suggests that the only difference between an enhancer and a promoter is a downstream gene body. Chapter 4 examines eRNA presence by this model across more varied datasets, revealing patterns of eRNA localization driven by cell type, drug perturbation and TF-activity.

I also showed that pausing probabilities or the proportion of elongation to initiation signal ($\vec{w}$) correlated well with experiments directly probing RNAP pausing on a global scale. Challenged by both the multiple levels of model latency, I was unable to asses significant changes in Θ on per gene basis between two different conditions. Given that $\vec{w}$ and π represent a proportion it might be possible to apply Fischer's exact test or a contingency χ^2 table model. However hypothesis testing involving μ, σ, λ requires a fundamentally different approach as these parameters are directly affected by the convolution of the exponentially-modified Gaussian distribution.

Moreover, it is very likely that gene body transcription is not well approximated by a uniform distribution which assumes a homogenous, stationary poisson point process over RNAP location. In fact, qualitative inspection reveals ebbs and flows in the GRO-seq read coverage signal over gene bodies indicating, possibly, local RNAP pausing in elongation. It would be very interesting to investigate models more commonly used in transportation sciences (car-traffic-jam models) to investigate any evidence of RNAP elongation pausing.

Although model selection by bayesian information criteria (BIC) appeared accurate at single isoform genes, the extent to which this model selection procedure can be assumed at enhancer loci needs to be investigated. Even so, model selection by this method is deeply limited by computational time (iterating from $1 - 20$ models is CPU intensive). Bayesian non-parametric methods, like an infinite mixture model, would be a natural extension to this work and require little to no re-tooling of the model.

In summary, with the advent of high throughput sequencing transcriptional assays like GRO-seq, RNAP is now being studied in increasingly exciting and precise ways. One of the key goals of this mixture model was to provide a set of biologically interpretable parameters that capture alterations in polymerase behavior induced by changes in regulatory proteins. I believe this mixture model is one of the first steps in building a comprehensive predictive model of RNAP.

Chapter 4

eRNA Profiles Predict Transcription Factor Activity

Portions of this chapter are adapted from

> **J. Azofeifa**, M. Allen, J. Hendrix, T. Read, J. Rubin and R. Dowell (2016) *eRNA Profiling Predicts Transcription Factor Activity* Nature 2017 (in review)

4.1 Introduction

To ultimately understand the contextual and functional consequences of eRNA presence, this chapter investigates heavily the epigenetic and transcription factor (TF)-binding flanking or localized precisely with Tfit-predicted sites of RNAP loading ($\hat{\mu}$). TFs function as major determinants of cell state[160, 140]. Despite their critical importance for controlling cellular phenotypes, no reliable method for ascertaining TF activity exists to date. Chromatin immunoprecipitation (ChIP-seq) studies have identified binding sites for many of the approximately 1,400 transcription factors encoded within the human genome[171], allowing estimation of a consensus DNA-binding motif for more than 600 factors[89]. However, studies comparing TF-binding events to RNA expression levels have revealed that many TF-binding sites have no apparent effect on nearby transcription[103, 53, 141]. Distinguishing such "silent" TF-binding events from those with regulatory capacity is a fundamental challenge.

Given that eRNAs are most readily detected in Global Run-on following by sequencing (GRO-seq), chapters 2 & 3 developed predictive models of both GRO-seq read coverage and RNA Polymerase II dynamics to first discover eRNAs genome-wide. With these models in hand, we are now prepared to predict eRNAs across a variety of tissue types and following experimental perturbation. By placing eRNAs in the context of extremely stratified sources of data (771 nascent transcript datasets), I will be able to make more fundamental claims as to the precise spatial and temporal relationship governing eRNA presence and absence.

4.2 Enhancer RNAs originate from transcription factor binding sites

In order to profile enhancer transcription genome-wide, I identified 39,633 putative sites of bidirectional transcription in a K562 GRO-cap dataset,[34] of which 30,324 were not associated with an annotated promoter (Figure 4.1, Extended Data Figure C.1 & C.2).

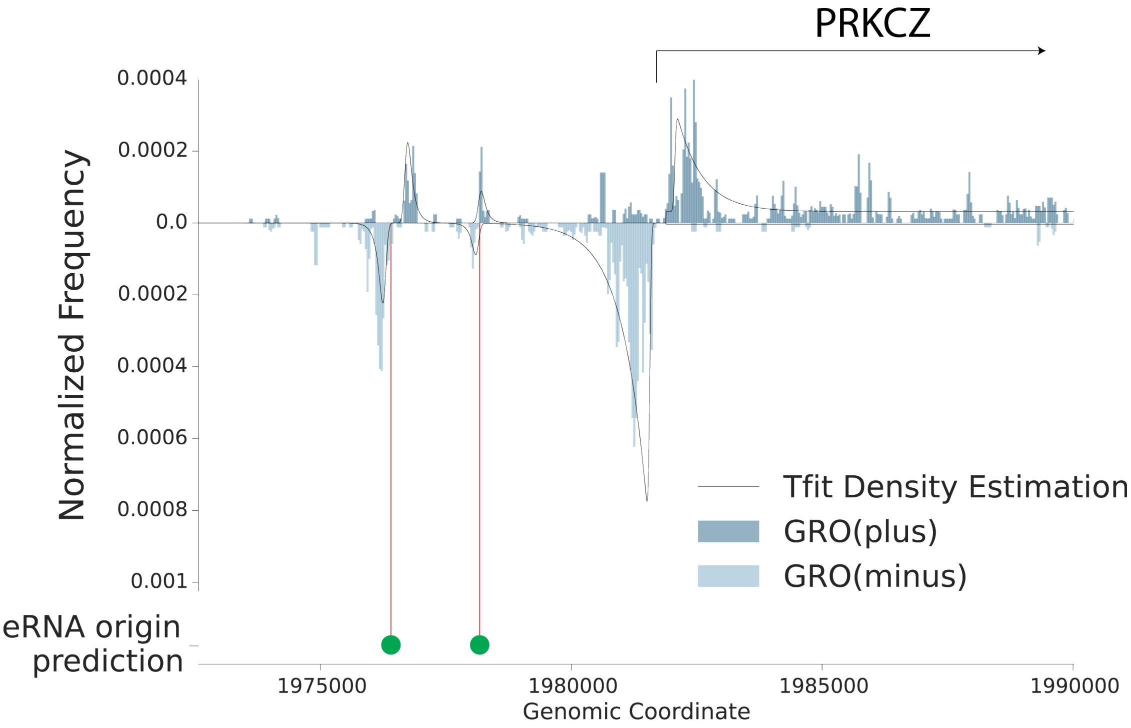


Figure 4.1: **eRNA Origins** An example locus displaying nascent transcript sequencing read coverage (HCT116 GRO-seq[3]) with the overlain density estimation via Tfit and the associated eRNA origin predictions (green dots).

As previously observed[38, 9], DNase I hypersensitivity (DHS), histone 3 lysine 27 acetylation (H3K27ac), and histone 3 lysine 4 mono-, di-& tri-methylation significantly associate with non-promoter bidirectional transcription (Extended Data Fig. C.3). Indeed, histone modifications are displaced from bidirectional centers supporting the presence of a nucleosome-free region localized precisely at the origins of bidirectional transcript initiation (Figure 4.2). Given their overwhelming co-occurrence with marks of active and open chromatin, as well as their distal location relative to annotated promoters, I refer to these transcripts as enhancer RNAs (eRNAs).

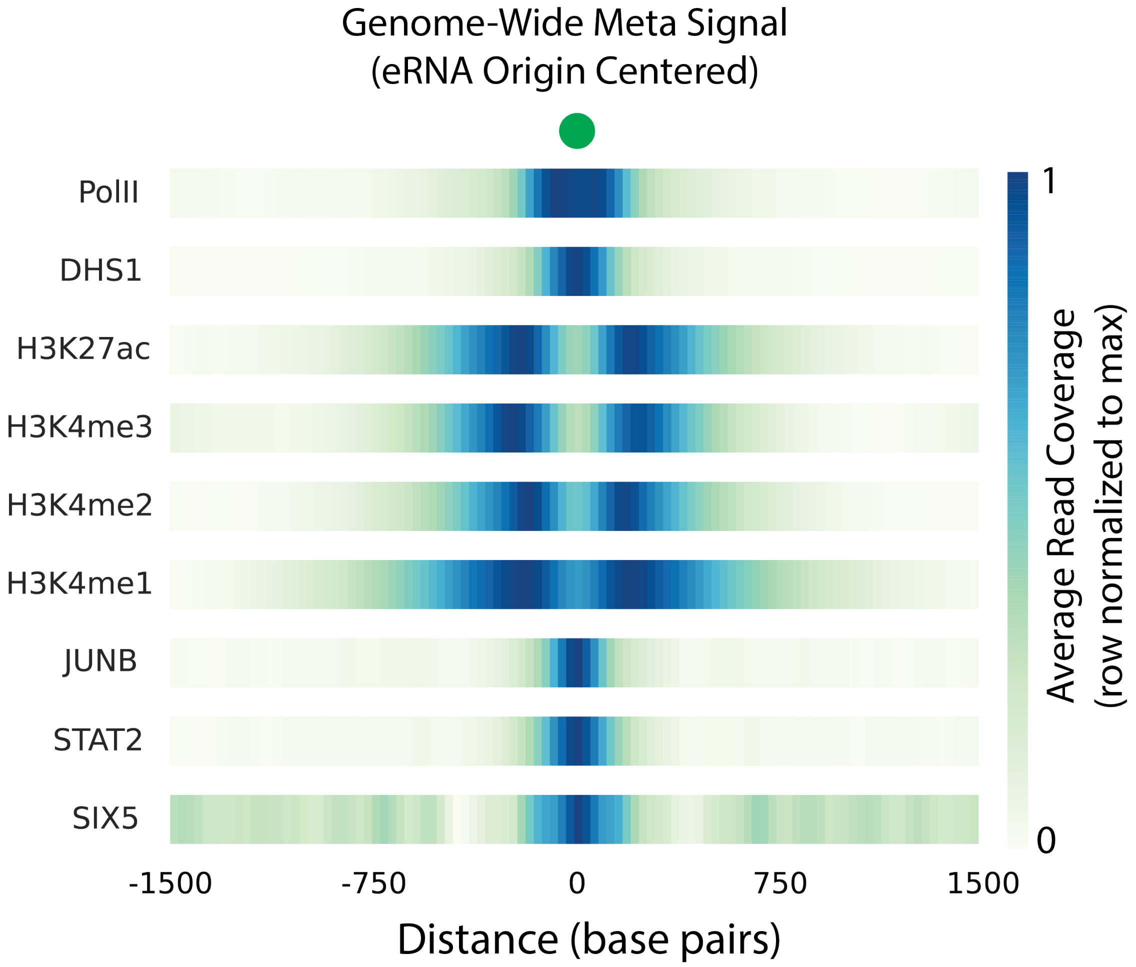


Figure 4.2: **Marks of Active Chromatin associate with eRNA Origins** Genome-wide meta-signal for marks of active chromatin aligned to eRNA origins inferred by Tfit in a K562 GRO-cap dataset[34](marks in Supplementary Table C.1).

Links between the location of eRNA transcription and specific TF-binding sites (e.g. p53, TNFα, ESR1) have been observed under a variety of experimental and cellular contexts[3, 106, 35]. To assess eRNA and TF-binding co-occurrence in a more systematic fashion, I integrated our set of eRNA origins with the genomic binding locations of 139 proteins profiled by ChIP-seq, also in K562 cells (Supplementary Table C.1). Consistent with previous results[38], 98% of eRNAs are bound by at least one regulator, where an average of 52.9 regulators localize at any one eRNA (Extended Data Fig. C.4A-B). In fact, I observed three distinct patterns of TF binding (Figure 4.3A): 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). In summary, unique sets of TFs bind to specific eRNA origins.

For the set of eRNA origins that overlap TF-binding sites, I next examined the co-localization of TFs relative to eRNA origins (Figure 4.3B). I 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 (Supplementary Note C.5, Extended Data Fig. C.4C). For example, factors such as RBB5, PHF8 and CDH1 are significantly displaced an average of 150, 200, and 398 base pairs away from the eRNA origin, respectively (Extended Data Fig. C.4D). 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 Figure 4.2 (p-value $< 10^{-6}$).

In addition to chromatin state, TF-combinatorial control also plays a pivotal role in downstream gene regulation[18]. In general, the number of TFs co-localized at sites of open chromatin is larger when an eRNA is present than not (Figure 4.4A). Furthermore, TF co-association dramatically increases when considering eRNA presence (Figure 4.4B). Taken together, the localization of diverse binding complexes at eRNA-associated TF-binding sites suggests that eRNAs may be markers of functional transcription factor binding.

4.3 Enhancer RNA origins mark sites of regulatory TF binding

Although the vast majority of eRNA origins localize with TF-binding, only a fraction of TF-binding sites overlap eRNA origins (Figure 4.5A). Similarly, only a fraction of TF-binding

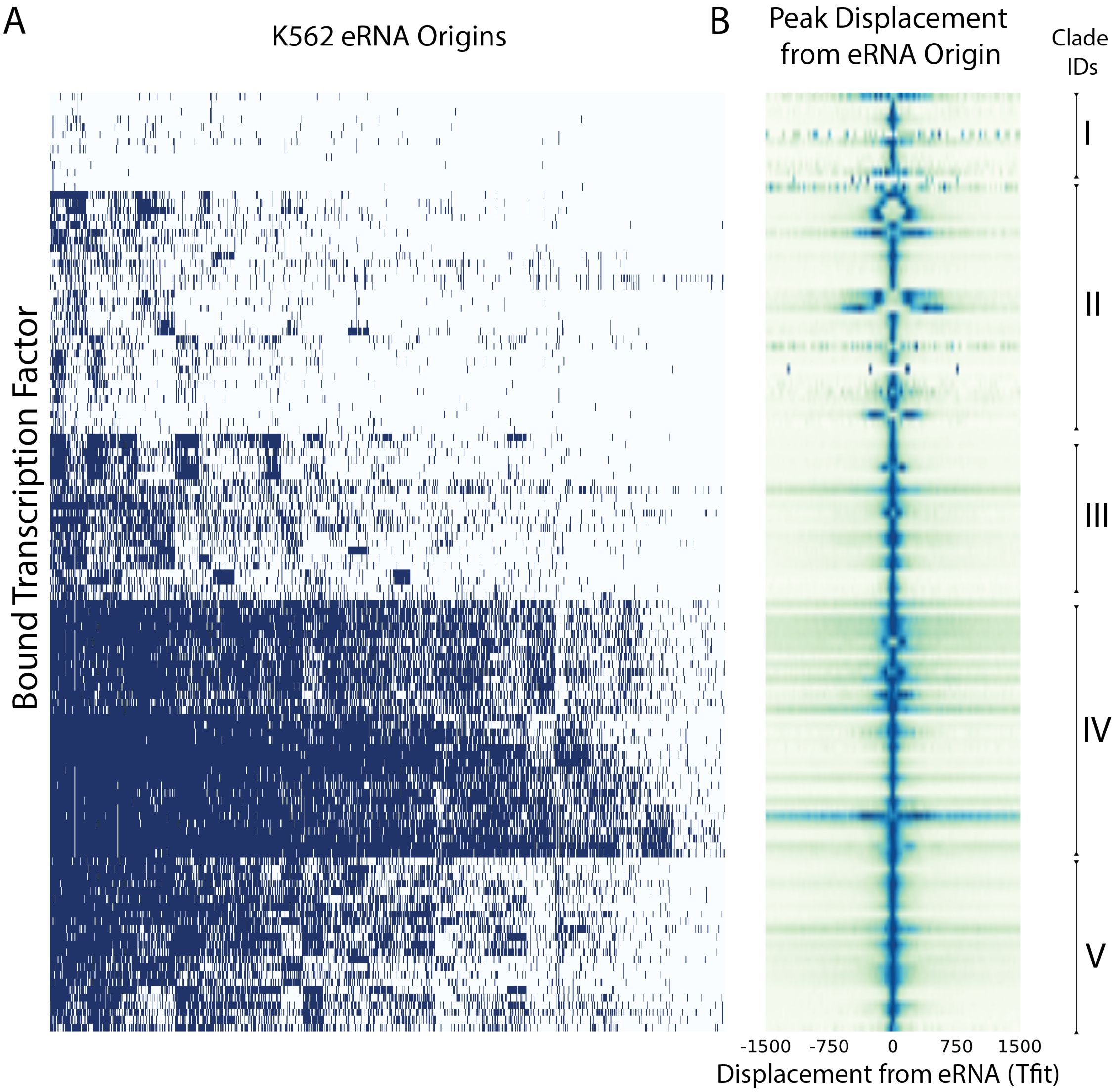


Figure 4.3: **eRNAs originate from sites of TF-binding** (A) The overlap of eRNA origins (columns) with 139 regulatory proteins (rows) (Tfs in Supplementary Table C.1). A blue tick indicates the presence of TF-binding site within 1.5 KB of the Tfit inferred origin; sorted by hierarchical clustering. (B) Histogram of the spatial displacement of the TF-binding peak from eRNA origins (heat is normalized to min/max of the histogram).

sites result in a concomitant change in nearby gene expression[147]. Given the strong relationship between active chromatin and eRNA transcription, I asked whether eRNAs discriminate “silent” from “active” TF-binding. In support of this hypothesis, TF-binding sites occurring at sites of eRNA origination display a significantly increased overlap with canonical marks of active chromatin relative to non-eRNA associated TF binding (Figure 4.5B). Moreover, no statistical difference is

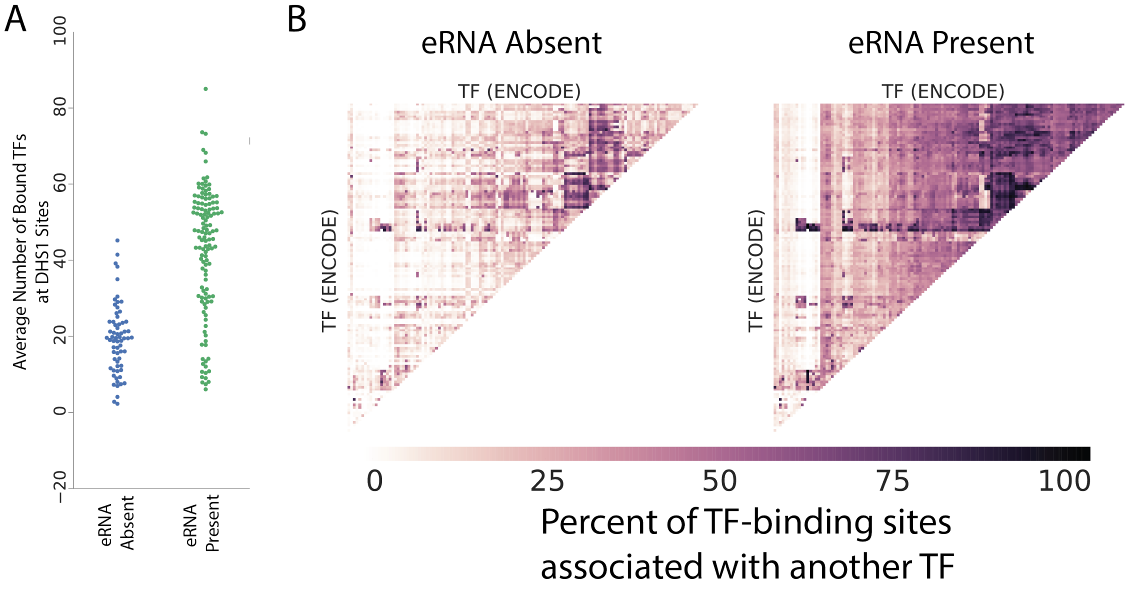


Figure 4.4: **eRNAs originate from sites of TF-binding** (A) The overlap of eRNA origins (columns) with 139 regulatory proteins (rows) (Tfs in Supplementary Table C.1). A blue tick indicates the presence of TF-binding site within 1.5 KB of the Tfit inferred origin; sorted by hierarchical clustering. (B) Histogram of the spatial displacement of the TF-binding peak from eRNA origins (heat is normalized to min/max of the histogram).

detected between these categories for repressive chromatin marks.

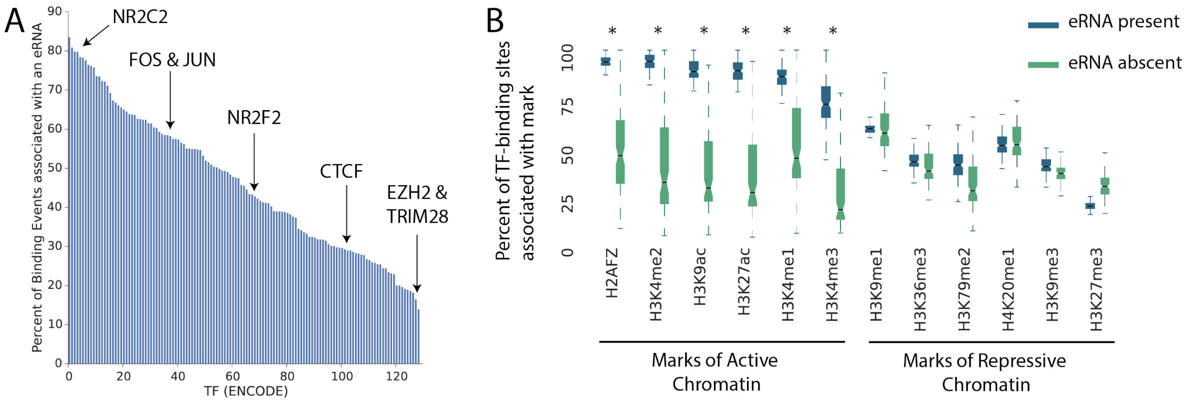


Figure 4.5: **TF-binding displays spectrum of eRNA association** (A) The y-axis indicates the proportion of a TFs ChIP peaks associated < 1.5KB with an eRNA origin. The x-axis is one of the 129 TFs profiled by ENCODE in K562 cells. (B) TF-binding peaks (Supplementary Table **??**) were grouped according to eRNA association. A box-and-whiskers displays the median/variability in proportion of histone mark association between the groups across all TFs (Supplementary Table C.1). Asterisks indicate a p-value $< 10^{-10}$ by z-test. All data in A-B are K562 cells.

Although regulatory TF binding is often enriched for open and active chromatin, functional

TF binding must ultimately lead to a change in gene expression. To this end, I considered TF-binding events conserved between two cell types but differing in terms of eRNA presence with the hypothesis that neighboring gene expression would be elevated in the eRNA-harboring cell type (Figure 4.6A). There are 95 TFs profiled in at least 2 cell types for which cell type-matched nascent transcription is available (Supplementary Table C.2). For example, binding of the transcription factor NR2F2 was profiled in both K562 and MCF7 cell lines, yielding 30,618 and 16,678 binding peaks respectively, with 3,491 peaks shared between the two cell types (Figure 4.6B). Of these cell type invariant peaks, 25% harbor an eRNA origin in both cell types, 7% only in K562, 12% only in MCF7 and 56% do not harbor an eRNA origin in either cell type. Measuring the transcription level of nearby target genes (TF-binding site $<$ 10 KB of gene promoter) revealed that eRNA presence is significantly correlated with elevated local gene expression (p-value $< 10^{-6}$). After making a total of 263 possible pairwise cell type comparisons (96 TF, 5 cell types), I noted that 73% of these comparisons display such dynamics (Figure 4.6C, Extended Data Fig. C.5).

4.4 eRNA origins co-localize with TF-binding motifs

Given that many TFs bind DNA in a sequence specific manner, I next sought to determine the precise spatial relationship between TF-binding motifs and eRNA expression. To this end, I examined a K562 GRO-cap dataset and measured the distance of TF motifs to eRNA origins. I observed a stark co-localization of the motif with the eRNA origin specifically in the TF-bound fraction of eRNAs (Extended Data Fig. C.6A), suggesting that the motif is present at the precise point of eRNA origination. This led to the speculation that the genome-wide patterns of motif to eRNA co-occurrence could identify the set of active transcription factors directly regulating eRNA transcription, even when ChIP data is not available.

To investigate this hypothesis systematically requires a measurement of motif-eRNA co-localization. With this in mind, I devised a simple statistic—the Motif Displacement score (MD-score)—which computes the proportion of TF sequence motifs within an $h-$radius of eRNA origins relative to a larger local H-radius (Supplementary Note C.7.1). Consistent with the average length

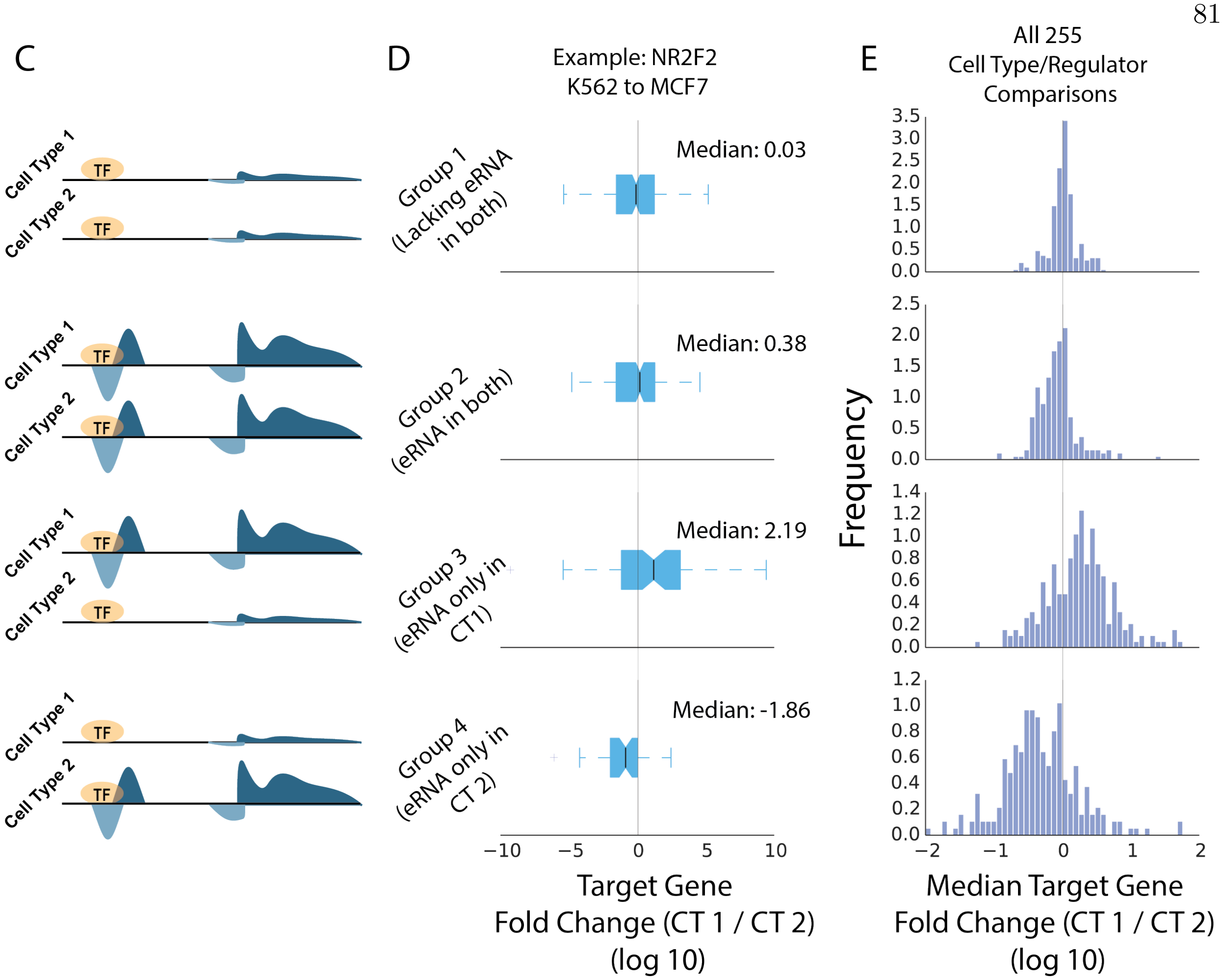


Figure 4.6: **eRNA presence marks the active subset of TF-binding** (A) Pairwise cell-type associated TF-binding peaks were grouped according to eRNA presence from matched cell types (Supplementary Table C.2). A gene was considered "neighboring" by a distance less than 10 KB. (B) Log base 10 FPKM fold change of "neighboring" genes related to eRNA-grouped NR2F2 binding peaks. (C) Histogram of Log base 10 FPKM fold change of "neighboring" genes for all possible eRNA-grouped TF ChIP-seq datasets (n=255).

of a nucleosome free region[177], I set the $h-$radius based on the average estimated distance between the forward and reverse strand transcript peaks at eRNA origins (h =150bp, H =1500bp; Extended Data Fig. C.6B). Consistent with our previous qualitative analysis, the MD-score is elevated in the bound set of eRNAs relative to the not bound set (Extended Data Fig. C.6C).

In order to expand our approach to include TFs for which no ChIP-seq is available, I leveraged a hand-curated database of TF-binding motif models (HOCOMOCO[89]) and measured the

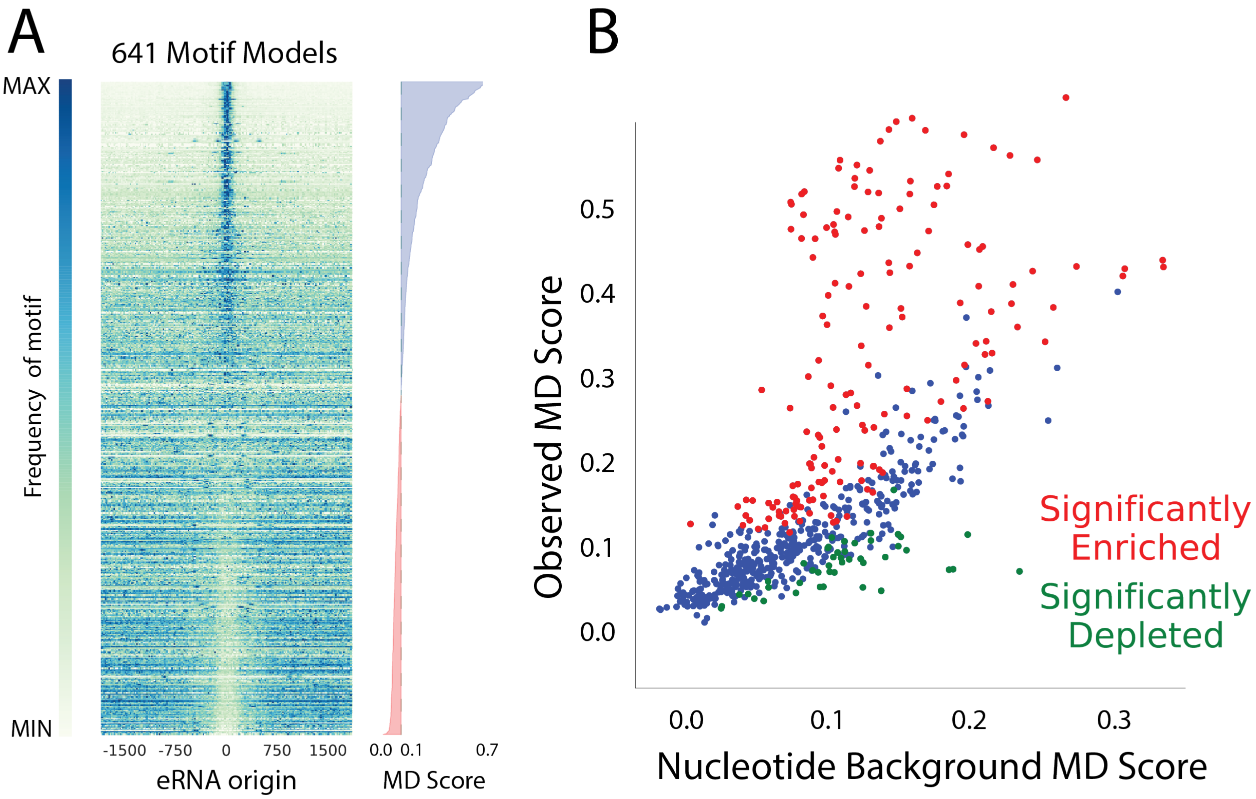


Figure 4.7: **TF-binding motifs localize at sites of eRNA transcription** (A) Each row is a TF motif model and each column is a bin of a histogram (100) where heat is proportional to the frequency of identify a motif at that distance from an eRNA origin. (B) A comparison between the expected MD-score for a motif model (x-axis) and the observed MD-score in a K562 GRO-cap experiment[34]. Red and green dots indicate a p-value $< 10^{-6}$ above or below expectation hypothesis tests, respectively.

distribution of 641 possible motif sequences proximal to K562 eRNA origins (Figure 4.7A). Under a uniform nucleotide background model, 32% of the motif models co-localize significantly with eRNAs (p-value $< 10^{-6}$; Supplementary Note C.7.2). However, similar to gene promoters and TF-binding motifs, eRNAs (e.g. enhancers) exhibit heightened GC content, which may artificially induce GC-rich motif presence at eRNA origins (Extended Data Fig. C.7A). To control for local sequence bias in our co-localization metric, I developed a simulation based method to perform empirical hypothesis testing of the MD-score (Extended Data Fig. C.7B, Supplementary Note C.7.3). I observed that —even in light of a significant nucleotide bias— 27% of motif models remain significantly co-localized with eRNA origins (Figure 4.7B).

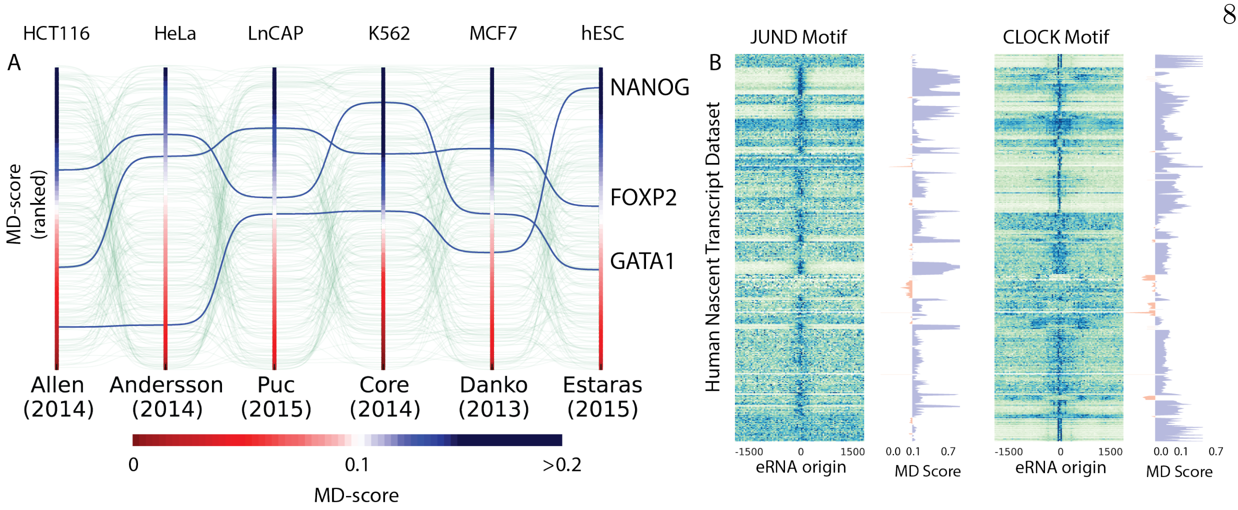


Figure 4.8: **MD-scores are variable across nascent transcript datasets** (A) MD-scores were computed and ranked under 6 nascent transcript datasets. (B) Each row corresponds to a nascent dataset and each column relates to motif frequency. These are shown for two demonstrative examples and the associated MD-scores, sorted by publication.

To determine whether this significant co-localization was an artifact of a single experiment, I examined MD-scores of all motif models across a set of nascent transcript datasets from six unique cell types. Our analysis revealed wide fluctuations in eRNA and motif co-localization across experiments (Figure 4.8A). Furthermore, I observed that the MD-score associated with cell type specific TF motifs are elevated in their known lineage of activity. For example, NANOG is elevated in embryonic stem cells consistent with its role in maintaining pluripotency[119]. Additionally, GATA1 is elevated in K562 cells consistent with its role in leukemia[150]. When I predicted eRNA origins in all publicly available nascent transcription datasets to date (67 publications, 34 cell types & 205 treatments), I uncovered that the spatial relationship between eRNA transcription and motif location is exceedingly dynamic (Extended Data Fig. C.8) as exemplified by the JUND and CLOCK motif models (Figure 4.8B). Overall, 78% of motif models in HOCOMOCO are significantly co-localized with eRNA origins in at least one dataset. Taken together, these results imply that these MD-score fluctuations reflect changes in transcription factor activity.

4.5 Motif displacement scores quantify TF activity

To investigate whether MD-score changes reflect alterations in TF activity, I turned to experiments where the activity of individual TFs is perturbed. In previous work, Dr. Mary Allen utilized the drug Nutlin-3a to activate p53 in HCT116 cells[3]. Here I observed a significant increase in the co-localization of the p53 sequence motif and eRNA origins following Nutlin-3a treatment (ΔMD-score 0.17, p-value$< 10^{-33}$). In fact, of the 641 available TF-motif models, only p53 and p63, which have nearly identical motifs, display altered MD-scores following Nutlin-3a treatment (p-value $< 10^{-6}$, Figure 4.9A). As expected, differential MD-score analysis between biological replicates revealed no significant shifts in motif to eRNA co-localization, indicating that our false discovery rate is low (Extended Data Fig. C.9A).

A number of other studies have specifically activated TFs: TNFα activates NFκB[106], estradiol activates ESR1[67]. Even in light of distinct nascent transcription protocols, I observed dramatic shifts in the MD-score for the transcription factor(s) known to be activated by each stimulus (Figure 4.9B,C). As before, differential MD-score analysis between biological replicates revealed no significant shifts in motif-eRNA co-localization (Extended Data Fig. C.9B,C). Despite that fact that treatments involving Nutlin-3a, TNFα, and estradiol are known to modulate gene expression[3, 67, 106], I observed no detectable differences in MD-scores when considering only promoter-associated bidirectional transcript sites (Extended Data Fig. C.10). In all three cases (Figure 4.9A-C), TF activation resulted in the production of new eRNAs that are uniquely enriched for the relevant motif, effectively elevating the TF's MD-score (Extended Data Fig. C.11).

In each case, nascent transcription was assessed at short time points (45 minutes or 1 hour). Therefore, I next sought to determine whether MD-scores could capture transcription factor activity across broader time frames. First, I observed that changes in TF activity are rapid, as exemplified by flavopiridol (a CDK9 inhibitor) treated mouse embryonic cells[92] which display a dramatic and monotonic increase in the MD-score of p53 and E4F1 (Figure 4.10A). For a number of TFs, MD-scores trend upward at 12.5 minutes and show significant changes within 25 minutes of exposure.

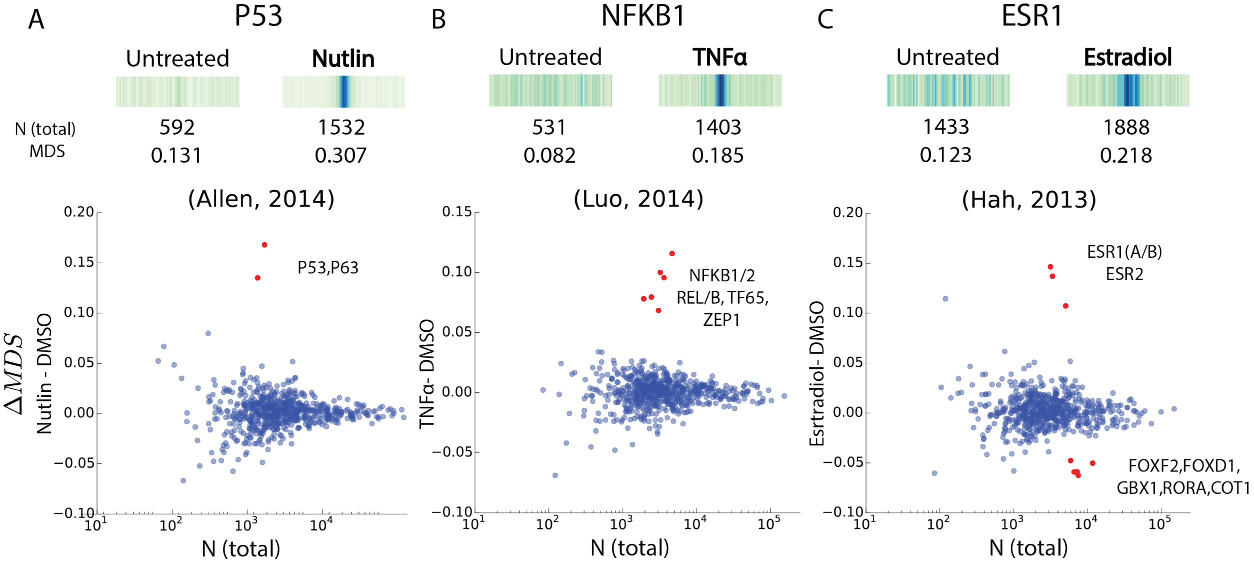


Figure 4.9: **Motif displacement scores predict TF activity following short treatment time points** (A) Top: The motif displacement distribution, MD-score and the number of motifs within 1.5 KB of any eRNA origin before and after stimulation with Nutlin3 on P53[3], the transcription factor known to be activated. Bottom: For all motif models (each dot), the change in MD-score following perturbation (y-axis) relative to the number of motifs within 1.5 KB of any eRNA origin (x-axis). Red points indicate significantly increased and/or decreased MD-scores, respectively (p-value $< 10^{-6}$). Similar analysis for (B) TNFα activation of NFκB[106] and (C) Estradiol activation of estrogen receptor (ESR1)[66, 67].

A longer time course, a Kdo2-lipid A (a highly specific TLR4 agonist) treated mouse T-cells[79], shows dynamic and time-ordered shifts in MD-scores for a number of key transcription factors (Figure 4.10B), including interferon (IRF7) and STAT2. Collectively, these results indicate that profiles of eRNA transcription—when combined with motif models—identifies shifts in TF activity in response to perturbation.

4.6 MD-scores predict TF activity across cell types

Stimulus responsive TF activity is detectable by motif displacement over eRNA origins, but transcription factors also play a pivotal role in cell fate and identity[119]. With this in mind, I examined a differentiation time series where human embryonic stem cells were differentiated into pancreatic tissue[174]. In this scenario, I observed a substantial decrease in MD-score for OCT4, SOX2, PO52 and NANOG immediately following differentiation to endoderm, concordant with

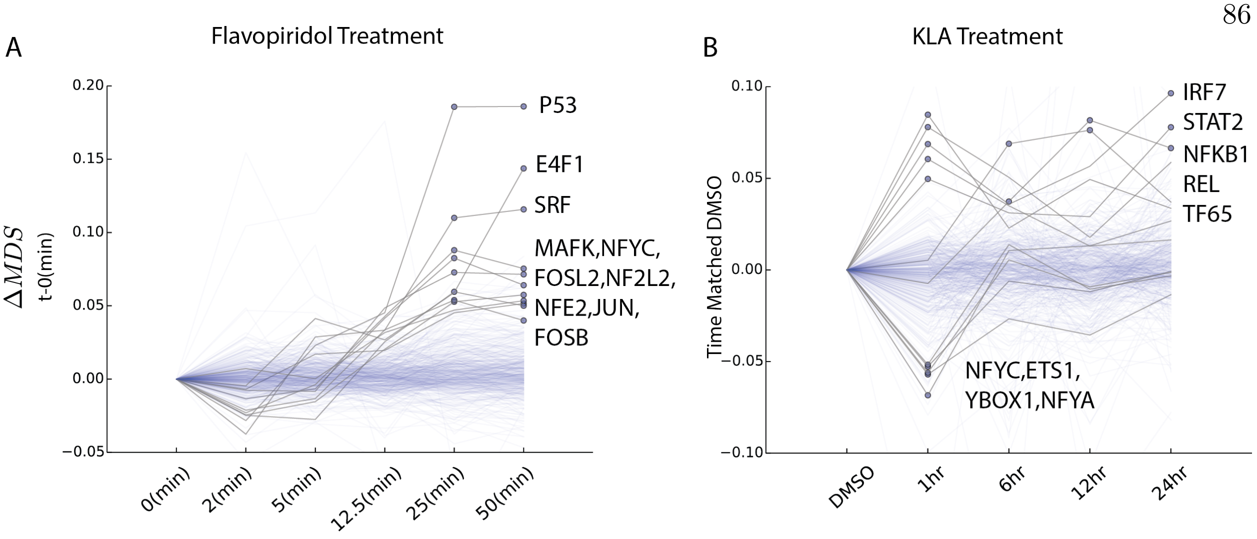


Figure 4.10: **Motif displacement scores predict TF activity across long time series** (A) A time series dataset following treatment with Flavopiridol. The y-axis indicates the MD-score change relative to time point 0. Blue dots indicate a MD-score difference $< 10^{-6}$. A darker shaded line indicates a time trajectory with at least one significant MD-score. (B) Time series dataset following treatment with Kdo2-lipid A (KLA) where each time point is normalized to time-matched DMSO. Therefore, the y-axis indicates MD-score difference relative to the time point matched DMSO sample. GEO SRR numbers of these comparisons are outlined in Supplementary Table C.3.

their role as embryonic stem cell markers (Figure 4.11A). Furthermore I observe RFX4, which has the same motif as the pancreatic islet specific RFX6[2], is elevated late in the time series.

I next sought to determine whether differential MD-scores could be utilized to identify, without prior knowledge, the key TFs that define distinct cell types. Consistent with this goal, an examination of eRNA/motif co-localization between two different cell types—GM1278 (Myeloid) and K562 (Leukemia) cell lines— resulted in higher activity of the GATA family of transcription factors in leukemia (K562) cells, whereas the interferon-regulatory family of TFs were more active in Myeloid (GM1278) cells (Figure 4.11B). When comparing more closely related cells, specifically fetal-lung tissue (IMR90) and lung carcinoma (A549), I noted a strong increase in the transcription factors related to the BACH1/MAFK pathway in the lung carcinoma samples (Figure 4.11C). While these TFs show altered activity between the cell types, they are not necessarily cell type

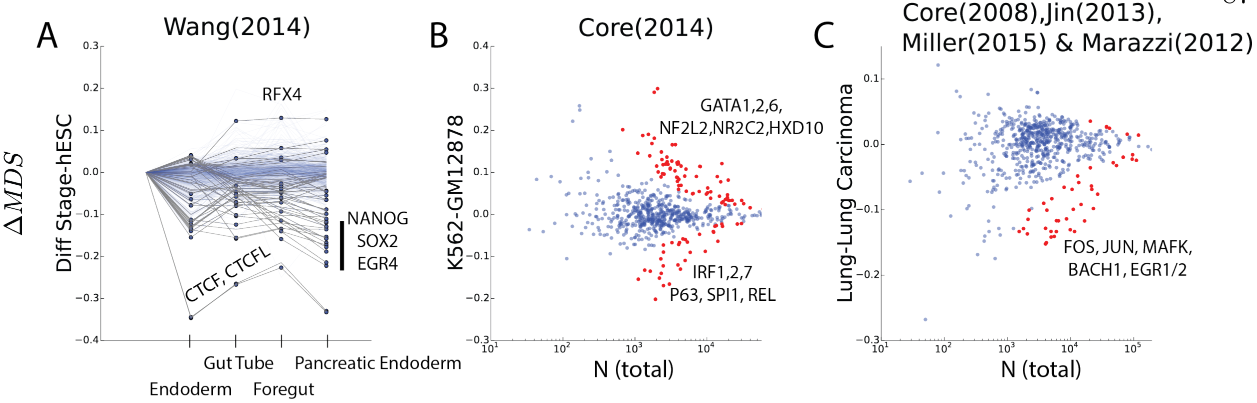


Figure 4.11: **Motif displacement scores predict TF activity across long time series** (A) A time series dataset following treatment with Flavopiridol. The y-axis indicates the MD-score change relative to time point 0. Blue dots indicate a MD-score difference $< 10^{-6}$. A darker shaded line indicates a time trajectory with at least one significant MD-score. (B) Time series dataset following treatment with Kdo2-lipid A (KLA) where each time point is normalized to time-matched DMSO. Therefore, the y-axis indicates MD-score difference relative to the time point matched DMSO sample. GEO SRR numbers of these comparisons are outlined in Supplementary Table C.3.

specific TFs, reflecting a limitation of the pairwise comparison.

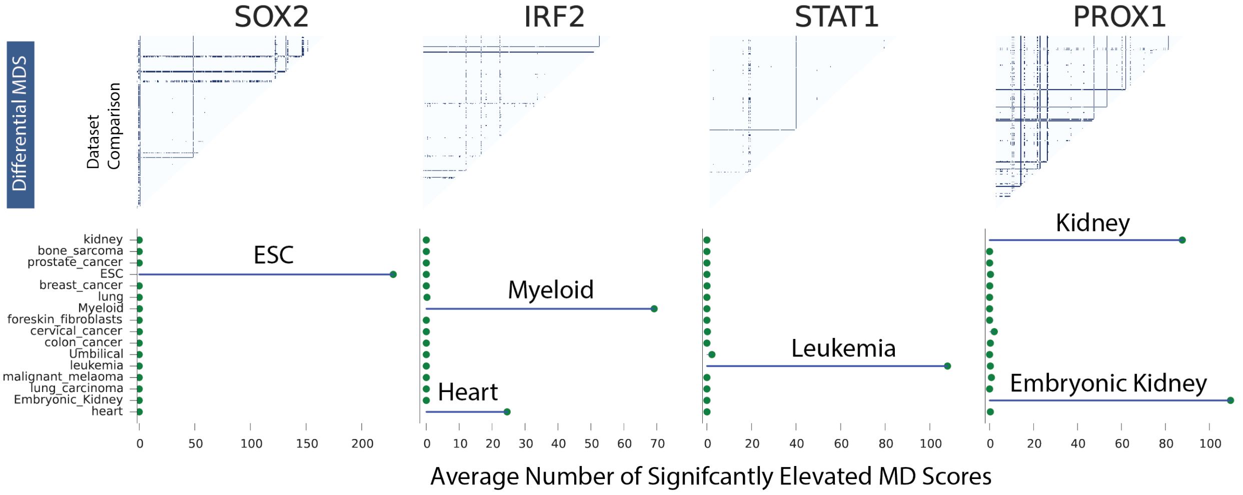


Figure 4.12: **Motif displacement scores predict TF activity across long time series** (A) A time series dataset following treatment with Flavopiridol. The y-axis indicates the MD-score change relative to time point 0. Blue dots indicate a MD-score difference $< 10^{-6}$. A darker shaded line indicates a time trajectory with at least one significant MD-score. (B) Time series dataset following treatment with Kdo2-lipid A (KLA) where each time point is normalized to time-matched DMSO. Therefore, the y-axis indicates MD-score difference relative to the time point matched DMSO sample. GEO SRR numbers of these comparisons are outlined in Supplementary Table C.3.

To move beyond isolated pairwise comparisons, I asked for significantly altered MD-scores between all possible human, untreated dataset pairs (124,251 pairwise comparisons, Extended Data Fig.C.12). Cell type-matched datasets revealed lower numbers of significantly altered MD-scores than cell type-differing comparisons suggesting a cell type effect on motif/eRNA association (Extended Data Fig. C.13). Indeed, the SOX2 motif model, a well established cell type-specific TF, differentially co-occur with eRNAs in 7,554 comparisons, 99.8% of which are embryonic stem cells (Figure 4.12). Following in suit, motif models for IRF2, STAT1 and PROX1 appear significantly enriched in Myeloid, Leukemia and Kidney cells.

Finally, I identified cell type-specific TF activity across all tissue types for which nascent transcription data is available. Examination of the co-association network reveals many well-documented links between cell type and TF activity (e.g. Retinoic Acid Receptors & Cervical cancer). Even still, this analysis uncovered dozens of previously unknown cell type-unique transcription factors whose mechanistic contribution to cellular identity has yet to be investigated (e.g. Supplementary Table C.4). In addition to cell type-unique associations, this analysis also reveals inter-cell type relationships: MCF7 breast cancer and embryonic stem cells share activity of COT and SMAD family TFs consistent with recent evidence linking stem cell like behavior to breast cancer[51]. As the diversity and quantity of nascent transcription data increases, eRNA profiling will precisely define the biological systems where individual TFs exert their regulatory influence.

4.7 Conclusion

By an overwhelming support of TF-binding events, enrichment of DNA motifs and changes in nearby gene expression, this chapter strongly suggests that TFs regulate eRNA transcription. Although an important step in understanding the basic biology of eRNAs, this chapter raised the important claim that eRNA predictions—in conjunction with TF/DNA-binding motifs— can infer TF-activity for over 641 transcription factors within a single experiment. Although functional experiments are needed to fully justify this result, there is little doubt that eRNAs represent a new and exciting read out on TF-activity.

An extraordinarily simple statistic, the motif displacement (MD) score predicted alterations in TF-activity that matched both the underlying drug treatment and cell type. Even so, the MD-score makes assumptions that should be relaxed in future modeling efforts. To measure co-occurrence between TF-binding motifs and locations of eRNA transcription, I choose two parameters h, H to compute local enrichment of TF-binding motifs near eRNA origins. Although h was estimated from the average distribution of the RNAP footprint, it is very likely that h should be adapted for each motif model independently. Yet more so, bimodal motif displacement distributions would not be captured well by this method. Other measures of one-dimensional co-occurrence that do not require h or H—like nearest neighbor—might also be investigated, although they will likely provide wildly different motif displacement distributions and statistics.

Apart from MD-score heuristics, the MD-score required genome-wide profiles of eRNA presence to make statements about TF-activity. However, transcription factors bind hundreds to thousands of separate and individual genomic loci. An interesting challenge might be to assign, for each eRNA, the TF (or set of TFs) that uniquely regulate its expression. In this way, it is likely this model will be hierarchical: first, we would wish to claim that the TF is active in a given experiment (so to regulate any eRNA) and then, predict the set of eRNAs that the TF regulates. With eRNA and motif information alone, this proposal might pose a significant challenge as upwards of ≈40 TFs bind anyone eRNA.

Indeed, the MD-score measures the activity of a single TF however transcription factors are well known to bind DNA in a combinatorial and context-dependent manner. Future efforts might monitor patterns of TF-binding motifs over eRNAs to build a network of TF co-association conditioned on eRNA presence. Moreover, simple pairwise correlations of MD-scores across nascent transcription datasets might reveal novel co-regulated transcription factors. In either of these analysis however, similarity in the position specific scoring matrix must certainly be take into account as similar PSSMs will likely map to similar genomic positions by chance thereby artificially inflating correlation scores.

Looking ahead, this chapter's observation that eRNAs predict modulations in TF-activity may have important implications in understanding the phenotypic consequences of non-coding regulatory mutations. Possibly, mutations over TF-binding motifs at eRNA origins may render transcription factor binding—and subsequent enhancer activity—obsolete, leading to alterations in important gene expression programs. An integrative analysis of disease-associated genotypes with sites of eRNA transcription may shed light on the complex problem of non-coding genomic variants.

Chapter 5

Looking Forward

Fundamentally, this thesis addressed the question: what regulates eRNA transcription? To answer, I developed novel probabilistic models of both GRO-seq read coverage and RNA Polymerase II dynamics to predict the precise locations of eRNA transcription genome-wide (chapters 2 & 3). With these predictions in hand, chapter 4 demonstrated that eRNA profiles represent a powerful readout on the dynamics and changes in TF activity. Although we have learned a lot through nascent transcriptional datasets, these modeling efforts open more questions than they answer. This final chapter will outline some of these questions in areas related to model improvements, data type inclusion and analysis ideas that may continue to shed light on the fascinating roles that eRNAs play in biology.

5.1 Mixture Models

The exponentially-modified Gaussian mixture model discussed in chapter 3 provided a mathematical explanation of GRO-seq read coverage based on the known enzymatic stages of RNAP. Although a parsimonious explanation of GRO-seq, this model was numerically expensive to compute necessitating parallel programing development on a large compute cluster. The major bottle neck in computation involved model selection (iterating over many different mixture topologies to EM convergence). In addition, current estimation of Θ (the parameters governing RNAP dynamics) is limited only to GRO-seq however many biochemical assays (e.g. ChIP-seq chromatin modifications) provide read coverage diagnostic of regulatory DNA. Disparate datatype inclusion may allow for more accurate estimates on Θ. In short, this section will serve to outline a few ideas on improvements to mixture model parameter estimation scheme discussed in chapter 3.

5.1.1 Model Selection

To revisit the model selection discussed in chapter 3 briefly, let $f(g;\theta)$ be some density function where $g = \{z, s\}$ represents a position along the genome (z) and the associated strand orientation (s). Furthermore, f may vary from simple models (Gaussian or Laplace) to more complicated ones (double geometric or an exponentially-modified Gaussian density). Because for given a loci there may be multiple eRNAs (characteristic of super enhancers), I described $h(g)$ as a function formed by the weighted linear combination of K densities (equation 5.1).

$$h(g;\Theta) = \sum_{k=1}^{K} \pi_k f(g;\Theta_k) \tag{5.1}$$

As with most mixture, clustering or topic model settings, much effort is placed in choosing and adequately modeling K (in our biological setting this corresponds to the number of distinct and different RNAP loading events). In chapter 3, K was estimated via Bayesian Information Criteria (BIC) by iterating over $k \in \{1, ..., 20\}$ and running the EM algorithm to full convergence. The final decision in K used a function proportional to the number of data points, log-likelihood improvement and model complexity penalty.

By no stretch of the imagination, such a framework does not scale well with increasing data for which you may very well expect more RNAP loading events (perhaps even to exceed 20 events). Such a computational burden has led me to revisit the field of Bayesian non-parametrics which promises to estimate both Θ and K simultaneously.

Unlike the finite mixture model, Bayesian non-parametrics treats K as a random variable with support between $[1, \infty]$. Of course, given a fixed sample data size (N), K can only exist between $[1, N]$ however the infinite mixture model setup allows K to grow with the data $(\mathbb{E}[K] = \alpha \log N)$. With the so called “Chinese Restaurant Process” or “Polya urn scheme”, the number of clusters (or RNAP loading events) is no longer fixed but follows a stochastic, non-parametric process. Equation

5.2 provides the agglomerative algorithm for this generative model.

$$p(c_n = k|\mathbf{c}_{1:n-1}) = \begin{cases} \frac{m_k}{n-1+\alpha} : & k \leq K_c \\ \frac{\alpha}{n-1+\alpha} : & \text{otherwise} \end{cases} \tag{5.2}$$

Here, K_c refers to the number of current components at the n^{th} draw, m_k refers to the number of data points associated with that mixture component and finally $\mathbf{c_{1:n-1}}$ provides for each data point the k^{th} component ownership. Given the creation of a new component or group, θ_k is drawn from some prior distribution $G(\Theta)$ and finally $g = \{z, s\}$ is drawn from the simulation algorithm outlined in chapter 3. Figure 5.1 shows three samples from this "Dirichlet process" where $G(\Theta)$ refers to the conjugate priors for $\mu, \sigma^2, \lambda, \pi$ outlined in chapter 3.

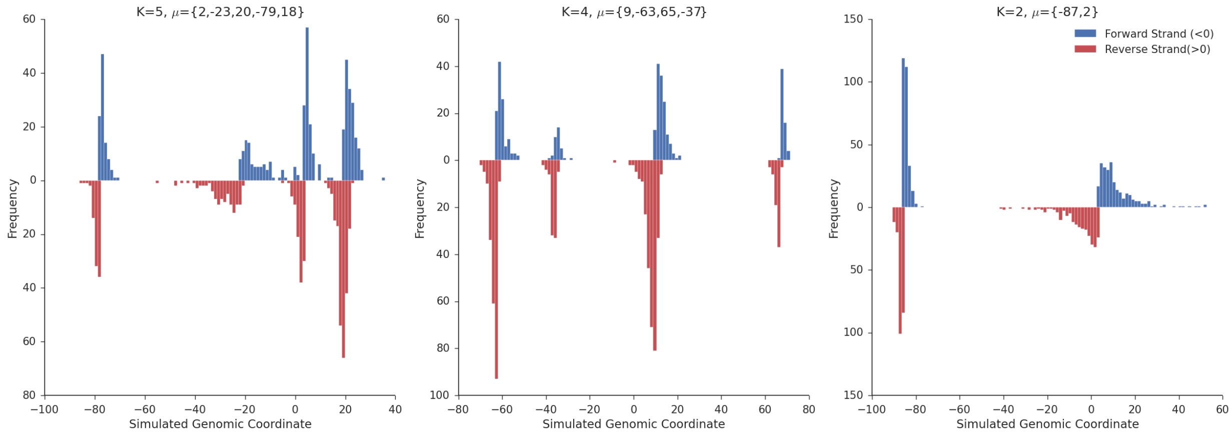


Figure 5.1: **Bayesian non-parametric simulations of the exponentially-modified Gaussian distributions.** Three draws from a Dirichlet process with a concentration parameter $\alpha = 0.1$ and within each draw $N = 1000$ samples.

The generative story underlying Bayesian non-parametric extensions to mixture modeling is particularly appealing within the context of eRNAs. As sequencing becomes deeper, the emergence of low-level or *very* quickly degraded eRNAs will likely arise. With this in mind, these infinite mixture models appear as a natural extension to tackle increasing model complexity in GRO-seq datasets.

Even so, modeling fitting (at least within the exponentially-modified Gaussian setting) might be challenging given that chapter 3 required derivation of the EM algorithm not only for the mixture model latency but also for the convolved density function. To take this one step further, the most naive solution to estimation of θ_k—under an infinite mixture set up—would be to let Gibb's sampling approximate $p(k|g)$ and then run the EM algorithm to full convergence over data points only associated with component k at each step in the sampling scheme. Given that MCMC sampling techniques can require thousands of iterations to converge and burn-in properly this does not appear like a computationally feasible solution. Another option might be to actually use the moment estimators discussed in early chapter 3 to perform updates on θ_k. These estimators can be computed quickly in terms of running sums. Of course, previously discussed degeneracies in moment estimation (e.g. $\pi \to 1$ or $\pi \to 0$) would need to be addressed, perhaps by appealing to strong priors (e.g. $\hat{\pi} = \frac{N_+ + \alpha}{N + \beta}$ where $\beta, \alpha \gg 0$).

5.1.2 Integration of other Data Types

Although Global Run-on followed by sequencing is a powerful readout on RNAP dynamics, there exist many other datasets which also capture signal diagnostic of non-coding regulatory loci. For example, DNase Hypersensitivity (DHS1) profiling or ATAC-seq both provide information on nucleosome free regions which is a hallmark of TF-binding and regulatory regions. Discussed heavily in chapter 4, ChIP-seq with antibodies specific to chromatin modifications like H3K27ac and H3K4me1 display bimodal signal flanking $\hat{\mu}$ (the expected value of RNAP loading). The bimodal frequency of chromatin modifications surrounding $\hat{\mu}$ follows intuition as the base pairs immediately proximal to $\hat{\mu}$ are lacking in nucleosomes. An interesting extension of the work presented in chapters 2 & 3 might be a model that attempts to incorporate all such data in a single or hierarchical framework to better inform estimates on θ (and specifically μ).

To extend the model of RNAP loading to incorporate data such as ATAC-seq or DHS1, we might make the fair assumption that ATAC-seq is distributed as a Gaussian random variable with expected value μ and variance σ^2. In keeping with previous notation, $g = \{z, s\}$ is a random variable representing a genomic coordinate and the associated strand orientation from GRO-seq. Yet now, we let $g' = \{z, u\}$ where $u = \{1, 2, 3\}$ represents the datatype of interest, either forward strand GRO-seq, reverse strand GRO-seq or ATAC-seq. We can then rewrite our density function f to allow for data type integration possibilities (equation 5.3).

$$f(g,s;\theta) = \begin{cases} \lambda\phi(\frac{z-\mu}{\sigma})R(\lambda\sigma - \frac{z-\mu}{\sigma})w_1 & s = 1; \text{Forward Strand GRO-seq} \\ \lambda\phi(\frac{z-\mu}{\sigma})R(\lambda\sigma + s\frac{z-\mu}{\sigma})w_2 & s = 2; \text{Reverse Strand GRO-seq} \\ \phi(\frac{z-\mu}{\sigma})w_3 & s = 3; \text{ATAC-seq} \end{cases} \tag{5.3}$$

Importantly, w_s is just the proportion of data arising from a specific datatype (equation 5.4) and can be estimated from $\mathbf{D} = \{g_1, g_2, ..., g_n\}$. In this way, $\sum_s w_s = 1$.

$$w_s = \sum_{g \in \mathbf{D}} \frac{\mathbb{1}(g \cap \{s\} \neq \emptyset)}{|\mathbf{D}|} \tag{5.4}$$

Because each datatype is pairwise independent, the EM algorithm derived in chapter 3 can be rewritten in almost an identical fashion to include multiple sources of data. With the addition of ATAC-seq, you are likely to get a better estimate on (μ, σ) but (λ, π) will be unaltered as the constitute GRO-seq specific parameters. Inclusion of epigenetic marks like ChIP-seq for H3K27ac or H3K4me1/2/3 into this modeling framework will likely be more challenging given the significant displacement away from $\hat{\mu}$. Within this in mind, one will need to appeal to a slightly different density function (e.g. the displaced Laplace mixture model discussed in Appendix C for estimated TF-binding event localization to $\hat{\mu}$).

5.2 TF Activity Inference Models

Discussed heavily in chapter 4, I showed that eRNA predictions in conjunction with TF-binding motifs may be used to unbiasedly monitor changes in TF-activity following an experimental perturbation and across different cell types. Although a rich analysis technique, this model inherently assumes that eRNAs occur under a birth/death process (eRNAs are either present or absent). Seen in a multiple publications[3, 106, 67], this binary categorization does not adequately explain an eRNA present in both samples but differing in GRO-seq read intensity. This section will serve to outline an idea to incorporate GRO-seq coverage *levels* into the TF-activity profiling.

To begin with simplest method first, one could take all the eRNAs that are significantly differentially expressed between two conditions (p-value $< \alpha$) and count the number of times some TF-binding motif was within a $h-$radius of $\hat{\mu}$. To assess significance, one could appeal to an enrichment model that takes into account both the number of differentially expressed eRNAs, number of genome wide motifs and total number of eRNAs (i.e. a hypergeometric model). Of course such an analysis relies heavily on assumptions such as the correct p-value cutoff (α) and the correct $h - radius$ (a limitation of the MD-score discussed in chapter 4 as well).

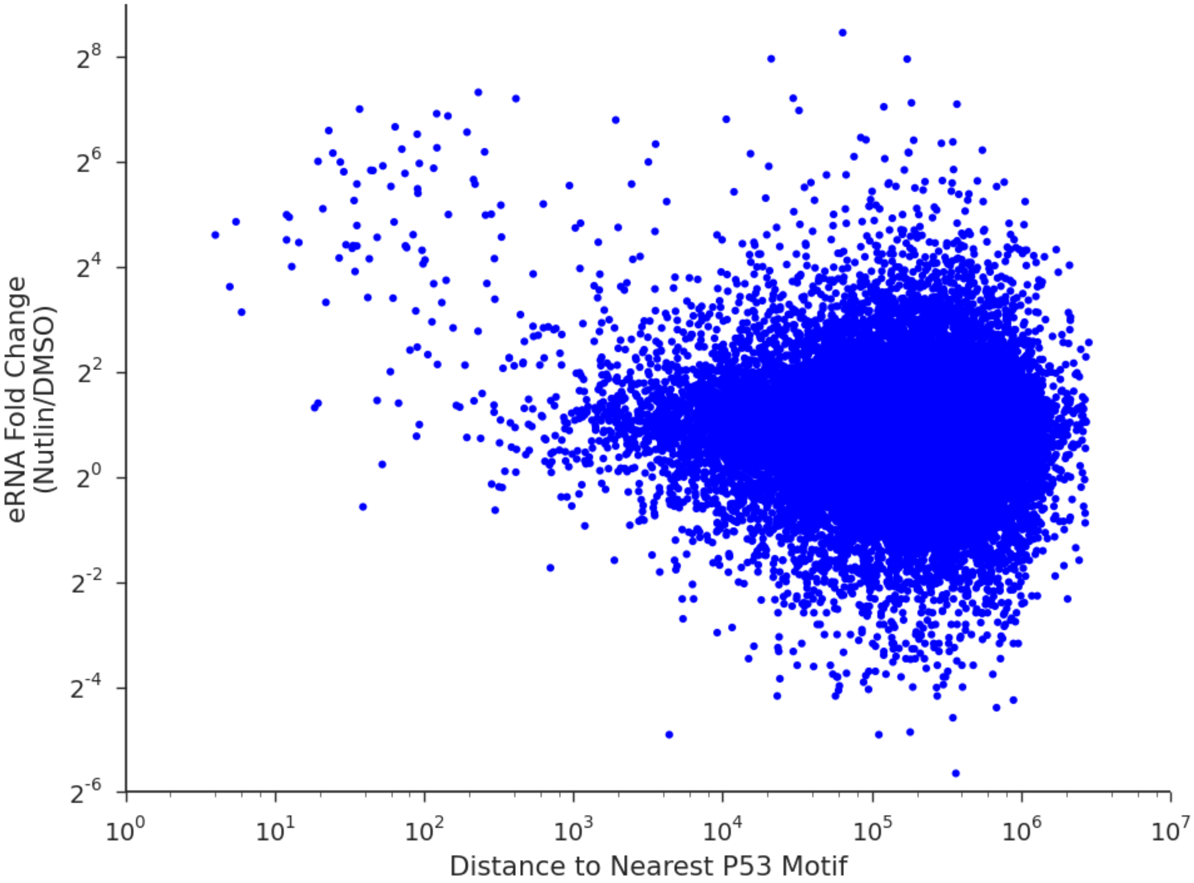


Figure 5.2: **A comparison between fold change of eRNAs following nutlin treatment and distance to nearest p53-binding motif** eRNAs were profiled using Tfit in DMSO control and Nutlin treated HCT116 cells, overlapping Tfit calls were merged and number of reads in both samples were collect (forward and reverse strand reads were summed together).

To address these nuisance parameters in a more systematic fashion, we might instead look for gross correlations in eRNA and TF-binding motif distances and their differential expression state. To be consistent with the notation discussed in Appendix C, let $E = \{e_1, e_2, ...\}$ be the set of Tfit-identified $\hat{\mu}$ genomic locations and $M_i = \{m_1, m_2, ...\}$ be the set of TF-binding motif genomic locations for transcription factor i. Let $D_i = \{d_1, d_2, ..., d_{|E|}\}$ be set of nearest-neighbor distances where d_j refers to the distance of the nearest TF-binding motif i to eRNA j (equation 5.5).

$$d_j = \min\{|m_1 - e_j|, |m_2 - e_j|, |m_3 - e_j|, ...\} \tag{5.5}$$

And finally for complete clarity, let $F = \{f_1, f_2, ..., f_{|E|}\}$ be the set of $\log_2$ fold changes (or some other measure of differential expression) in GRO-seq read coverage at each enhancer locus between control and treatment experiments.

Figure 5.2 shows the relationship between D and F where HCT116 cells were treated with Nutlin (a drug that specifically activates p53); as such, nearest neighbor distances were computed against the p53 binding motif. Seen qualatatively in Figure 5.2, a significant population of eRNAs experienced a large increase in transcription (relative to DMSO) following Nutlin treatment and are all within a 100 base pairs of a p53 motif.

Although qualitatively diagnostic of TF-activity, manual inspection of all 641 TF-binding motifs across many different control/treatment pairs is unrealistic. To quantitate the effect in Figure 5.2, we might appeal to rank order statistics. We would first order eRNAs based on their differential expression state between the control/treatment pair. Subsequently, we would ask if TF-binding motifs are *significantly* nearer eRNAs at the high (or low) ends of the ranked list. However, to assess significance requires a fundamentally different method than the random walk models and KS-statistics discussed within the Gene Set Enrichment Analysis literature since we wish to take into account both eRNA differential expression *and* a measure of genomic co-occurrence (nearest TF-binding motif). Even still, I am hopeful that such a method can be developed since their exists such strong signal observed in Figure 5.2.

5.3 Predicting Enhancer to Gene Interactions

Chapters 2 & 3 focused primarily on labeling or annotating portions of the genomes. Whether these annotations involved enhancer, *cis*-regulatory elements or nascent transcripts, the goal of each algorithm was fundamentally about classification. Of course labeling is really just the most primitive step to understanding the functional nature of a specific genomic element. As shown in chapter 4, enhancers RNA harbor upwards of 40 TF-binding events and are primed with epigenetic modifications but how do they directly correlate with heightened local gene expression? A method to predict these enhancer-to-gene-interactions would have extremely important implications on understanding the phenotypic consequences of a non-coding regulatory mutation at an enhancer. By using techniques from polymer physics and Bayesian network reconstruction, this next section serves as an outline for an idea on linking enhancers to genes using GRO-seq data alone.

5.3.1 Network Structure Prediction

With nascent transcriptional data, we can measure the transcriptional output of enhancers with as much accuracy as their gene counterparts. In fact in this way, we might look to the well-established field of gene regulatory inference to predict enhancer/gene pairs. The so called “guilt-by-association” models—where co-regulated genes show correlated transcriptional output—may also be applicable to enhancer-promoter interactions as we might also hypothesis matched transcriptional signal.

To be formal, let v_i indicate some gene (or *cis*-element) and let $e_{i,j}$ indicate the presence of a regulatory pressure from v_i to v_j. For example, if $e_{i,j} > 0$ then we might say v_i activates v_j or conversely $e_{i,j} < 0$ implies an inhibitory effect. With little effort, the set of enhancers/genes (V) and the set of regulatory pressures (E) make a graph, $G = (V, E)$. By allowing the directionality and magnitude of $e_{i,j}$ to vary, we can encode a signed, directed and weighted graph that describes a system wide relationship between all functional genomic loci.

The complexity of regulatory networks can not be understated. If the genome consists of

n functional loci and the regulatory network is fully connected, undirected and simple (i.e. no self loops and no multi-edges), there are $2^{\frac{1}{2}n(n-1)}$ possible graph structures. The exponential explosion over network topologies makes inference in any mathematical context difficult. Even beyond network structure, modeling $e_{i,j}$ (how v_i exerts pressure over v_j) may result in unrealistic data requirements or simplifying assumptions rendering the regulatory model useless.

To fully specify network structure learning with high throughput sequencing assays, take an example a dataset d to be two experimental conditions c_1, c_2. These conditions may represent a gene knockout experiment, treatment with a drug or samples over a time-series. To summarize, for every $v_i \in V$ there exist a before and after measure of activity. Since non-coding *cis*-elements can only be measured by transcriptional or ChIP-seq assays, v_i specifically represents the sum of read counts across the genomic loci $[a, b]$, $v_i = \sum_{s \in \{a..b\}} y_s$. Taken together, $\mathbf{D} = \{d_1, ..., d_n\}$ represents the collection or set of all these perturbation experiments.

5.3.2 Correlation Networks

One approach to network structure creation is by correlation thresholding[111]. The assumption is that loci who are conditionally dependent should be correlated across experimental perturbations and time series. Let $f(v_i, v_j)$ be some function that maps the similarity between v_i and v_j across $\mathbf{D}$. Common forms of $f(.)$ are given in equation 5.6.

$$f(v_i, v_j) = \begin{cases} E[(v_i - \mu_{v_i})(x_j - \mu_{v_j})] \text{ covariance} \\ \frac{cov(v_i, v_j)}{\sigma_{v_i}\sigma_{v_j}} \text{ correlation} \\ \sum_{v_i, v_j \in \{0,1\}} p(v_i, v_j) \log \frac{p(v_i, v_j)}{p(v_i)p(v_j)} \text{ mutual information} \\ ||v_i - v_j|| \text{ euclidean distance} \end{cases} \tag{5.6}$$

Here μ_{v_i} and σ_{v_i} represent the expected value and standard deviation of v_i across $\mathbf{D}$. Housed under a similarity matrix S, $S_{i,j}$ represents some measure of distance or correlation between v_i, v_j. Tp represent an undirected-unweighted graph, an adjacency matrix A can be easily constructed under some threshold ρ, $A_{i,j} = \mathbb{1}(S_{i,j} > \rho)$.

The sparsity or density of the inferred graph varies as a function of ρ, pointing to the somewhat arbitrariness of ρ. Confidence intervals surrounding any of the above distance metrics can be computed by randomly sampling $\mathbf{D}$ with replacement (i.e. bootstrapping) but a p-value cut-off of 0.05 still retains some level of carelessness.

In brief, thresholding over correlation does not explicitly model the biological mechanism of $e_{i,j}$ and thus the final network gives up regulatory interpretability. Even so, correlation networks are ubiquitous and very useful in systems biology: primordial germ cell specification[110], osteosarcoma cell proliferation[64], circadian clock oscillators[120], micro-biome community detection[39] and diagnostic expression signatures for Alzheimer's[178], sarcoidosis[109] and Crohn's[168] diseases phenotypes (to name only a few of the hundreds of success stories).

5.3.3 Bayesian Networks

To combat some of the heuristics observed with correlation-thresholding, a popular scheme for representing a high dimensional joint distribution is a Bayesian network (BaN)[56] where conditional dependencies within a set of random variables are represented by a directed-acyclic graph (DAG), G.

$$\begin{aligned} p(v_1, v_2, ..., v_n) &= p(v_1) \cdot p(v_2|v_1) \cdot p(v_3|v_2, v_1) \cdot .... \cdot p(v_n|v_{n-1}, ..., v_1) \\ &= \prod_{i=1}^{N} p(v_i|Pa(v_i); \theta_i) \qquad (5.7) \end{aligned}$$

where $Pa(v_i)$ represent the parents of v_i from G

Modeling $p(v_i|Pa(v_i); \theta_i)$ varies throughout the literature. By discretization ($v_i \to \{0, 1, 2, ..., r\}$), a multinomial provides mathematical convenient conjugate priors[57]. In the case of binary discretization, $p(v_i|Pa(v_i))$ might be viewed as specifying *soft* logical statements[122]. Furthermore, linear Gaussian functions where the expected value of $p(v_i|Pa(v_i); \theta_i)$ becomes a linear combination of $Pa(v_i)$ grants a continuous representation of v_i[70].

Maximum likelihood or *a posteriori* methods support estimation of the conditional distribution parameter set $\hat{\Theta} = \{\hat{\theta_1}, .., \hat{\theta_n}\}$ from $\mathbf{D}$[71] yet require a fully specified DAG. With our goal to predict directed relationships between enhancer loci and sets of target genes, we seek estimation of both $\hat{\Theta}|\hat{G}$ and $\hat{G}$ simultaneously. Ignoring for the moment the computational complexity of enumerating all possible DAGs given $|V|$ vertices, we would like a function that maps how well some candidate G fits the perturbation data $\mathbf{D}$. Estimation of G from the likelihood of the data (equation 5.8) results in a fully connected network, symptomatic of over-fitting[123]. To this end, an important component to BaN structure learning is model selection where we wish to enforce complexity penalties that grows monotonically with $|G|$. Equation 5.8 presents only a popular few "scoring functions:" Akaike Information[95] and Bayesian Information[95].

$$\begin{gathered}\mathcal{L}(G,\Theta|\mathbf{D}) = \prod_{i=1}^{|\mathbf{D}|}\prod_{j=1}^{|V|} p(x_{i,j}|Pa(v_j)=x_i); K = \sum_{j=1}^{|V|}|Pa(v_j)| \\ Score(G) = \begin{cases} 2\cdot K - 2\cdot\ln\mathcal{L} : \text{AIC} \\ |\mathbf{D}|\cdot K - 2\cdot \ln\mathcal{L} : \text{BIC} \end{cases}\end{gathered} \tag{5.8}$$

Yet, none of these model selection criterions make explicit use of a prior belief over some DAG, i.e. $p(G)$. A benefit to incorporating non-coding *cis*-elements like enhancers is that their regulatory mode is very well understood biologically: enhancers *enhance*. In this way, we can place strong priors on the *sign* of the directed arc from $v_i \to v_j$ given that v_i is an enhancer. With this in mind, we would like to collect statistics (either in closed-form or via sampling) from the posterior of all DAG structures given our data of experimental perturbations $\mathbf{D}$ (equation 5.9).

$$\begin{gathered}p(G|\mathbf{D}) = p(\mathbf{D}|G)p(G)/p(\mathbf{D}) \\ p(\mathbf{D}|G) = \frac{p(\Theta|G)p(\mathbf{D}|\Theta,G)}{p(\Theta|\mathbf{D},G)} = \int_{\Theta} p(D|G,\Theta)d\Theta \textbf{ marginal likelihood}\end{gathered} \tag{5.9}$$

From MacKay[108], the marginal likelihood in equation 5.9 indirectly encodes Occam's razor (a simpler model should be favored to a complex one under equal evidence) and thus implicitly performs model selection.

Regardless of a specific scoring function, an algorithm to search over all DAG structure falls under the class of NP-complete algorithms given the exponential explosion in the number graphs with N vertices[28]. Common heuristic optimization algorithms like hill-climbing[166], beam-search[70] or even ant-colony algorithms[40] have been used to find a local MAP estimates of $\hat{G}$ under $p(G|\mathbf{D})$. A simple greedy algorithm[71] deletes, inverts or adds an edge to $\hat{G}$ and accepts the edge if the marginal likelihood increases. Common initializations to this greedy algorithm are the best DAG tree ($|E| - 1 = |V|$) where a globally optimal MAP estimate can be computed in polynomial time[29].

A greedy algorithm however highlights some important limitations with BaNs: two different DAGs can express the same marginal likelihood or model selection criterion score[167, 131]. Briefly, a BaN specifies a set of conditional independence statements of the form: $v_i \perp v_k|v_j$ or $p(v_i, v_k|v_j) = p(v_i|v_j) \cdot p(v_k|v_j)$. Yet such statements can arise under many different DAG topologies.

$$(1)\ v_i \rightarrow v_j \rightarrow v_k \implies v_i \perp v_k|v_j \qquad (3)\ v_i \leftarrow v_j \rightarrow v_k \implies v_i \perp v_k|v_j$$

$$(2)\ v_i \leftarrow v_j \leftarrow v_k \implies v_i \perp v_k|v_j \qquad (4)\ v_i \rightarrow v_j \leftarrow v_k \implies v_i \not\perp v_k|v_j$$

The above rules of "d-separation" highlight the shortcomings of both Bayesian and correlation networks. Given a similarity matrix S computed from N samples drawn from $p(v_i, v_j, v_k)$, we would observe a high degree of similarity between v_i, v_j, v_k (cases 1-3) yet the *causal* process giving rise to this correlation might be extraordinarily different.

Intuitively, this discriminative resolution allows us to define equivalence classes between DAGs: G, G' are equivalent if the set of conditional independence states (CI) defined by G and G' are equal, $G = G' \iff CI_G = CI_{G'}$[131]. As a modification to the simple greedy optimization algorithm outlined above, we would like to add/remove/invert edges such that the resulting *equivalence classes* change not just the underlying DAG.

Although DAG identifiability poses a sincere challenge in linking eRNAs to specific gene promoters, Bayesian networks have become exceedingly popular in modeling gene regulatory networks. Due in part to their robustness to noise and minimal data requirements[132], BaNs have successfully predicted signaling pathways in bacteria[52, 104], yeast[117, 179, 78] and humans[138, 63, 45].

5.3.4 A 3D Genome

Pervasive in genome biology is a bias to understand the genome as a one dimensional track but this notion of distance is deceptive. DNA exists within a fluid, where proteins like CTCF help fold and package the DNA three dimensionally. *cis*-elements like enhancers can be located many thousands of nucleotides away from their target gene (even located on different chromosomes), but three dimensionally these loci must be quite close so to exert their regulatory effect.

Very recently, scientists developed assays to measure three dimensional chromosomal proximity. Techniques like chromosome conformation capture (3C)[44] or chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)[99] glue together portions of the genome near one another, isolate these regions and sequence them. After lots of post-processing, 3C outputs an interaction matrix A where each cell $[i, j]$ corresponds to the number of observed three dimensional interactions between locus i and locus j. Via a binomial test, we can threshold A under some α-level to an undirected graph.

Among the many discoveries of 3C-like techniques are *transcriptional factories*: sets of connected loci that display similar transcript levels[50]. And as shown in chapter 2, transcript levels of 3D-linked enhancer RNAs and genes show correlated RNA production[11]. To highlight the important connection between enhancer RNA transcription and 3D-linkage analysis, Figure 3.9 showed two connected components generated from a Leukemia cell line (K562). By the simplest measure of node centrality (degree number), we observe that loci with enhancer transcriptional signatures (defined by the exponentially-modified Gaussian mixture model) also contain a larger degree distribution.

Evident at least qualitatively, the CTCF networks in Figure 3.9 display week modularity by transcriptional presence. The extent to which transcription *flows* through these networks remains unanswered and would require accurate null-modeling. Although biologically, 3D transcriptional flow makes intuitive sense. If a set of genes require a similar level of transcription than RNA Polymerase might traverse these loci far more efficiently when termination sites are linked to *cis*-

elements like promoters or enhancers.

Network visualization of A is still a lower-dimensional projection or discretization of the true 3D structure. To this end, some modeling effort has focused on inferring a three dimensional folding of a chromosome in some arbitrary xyz space[181]. The key insight to these methods requires translating the interaction frequency matrix A to a distance matrix, D. The intuition for this conversion is simple: the more observations between locus i and j the closer they should be. Equation 5.10 describes an inverse relationship between frequency and distance by some constant α.

$$D_{i,j} = \begin{cases} (1/A_{i,j})^{\alpha} : A_{i,j} > 0 \\ \infty : A_{i,j} = 0 \end{cases} \tag{5.10}$$

Given α and assuming that $\vec{x_i}$ refers to the 3D embedding of locus i, we would like $||x_i - x_j||$ to be as close to $D_{i,j}$ as possible. As a non-convex, non-linear optimization problem, 3D inference falls under the class of NP-complete algorithms[169]. Yet heuristic methods exist. By taking into account some constraints with stiff polymer physics, Monte Carlo sampling methods can recover meaningful chromosome structures[145]. A cleverer solution exists by relaxing the solution space from $\vec{x_i} \to \mathbb{R}^3$ to $\vec{x_i} \to \mathbb{R}^n$ where n is the number of loci. In this was, the optimization becomes a convex semi-definite programming problem for which global minimizers can be computed in polynomial time[181].

No matter the method, I believe strongly that any predictive model of enhancer-gene interactions must take into account the three dimensional architecture of DNA as it is through enhancer-gene looping that regulation by TF-binding can take place. As I see it, the clearest path forward to predicting enhancer-gene interactions might be a combined approach of Bayesian network modeling with priors informed by stiff polymer physics.

5.4 Thesis Conclusions

Hopefully reflected by each proceeding chapter, this thesis took a multi-disciplinary approach to science. Requiring biological context, software engineering and mathematical modeling, each project gave me the opportunity to participate deeply with people from many disparate areas of science. Specifically, I developed a regression-based method and a novel mixture modeling framework to identify enhancer RNA transcripts in GRO-seq. Having proven that the aforementioned algorithms were mathematically rigorous, computationally tractable and accurate, I applied these to a large swath of GRO-seq like datasets and discovered important relationships between eRNAs, TF-binding motifs and cell type.

The era of "big-data" has certainly had an effect on Biology. With a 3.2 billion base pair genome and assays that can profile each one of those bases, new computational approaches are being created daily to tackle these challenges. Algorithm development, however, is not the bottleneck of the bioinformatics field but rather a systematic approach to interpreting the output of each new predictive model. As computational-minded scientists continue to innovate within molecular biology, I feel that emphasis should always be placed on building *interpretable* algorithms and less so on vanilla implementations of "machine-learning."

Instead, we might consider modeling what each piece of data represents both in terms of technological limitations but also their unique scientific consequences. Models should optimize not only for quality metrics, like F1-precision scores, but also for intelligibility as an interpretable model is more likely to give rise to a testable scientific hypothesis. Indeed to optimize for intelligibility, modelers need to consider the biological or physical phenomena underlying the data. Although some might argue, data-*science* is not a field of engineering but a branch of science.

Looking forward, enhancer RNAs or non-coding *cis* regulatory elements are likely to play a fundamental role in personalized medicine. With the result that 76% of disease associated genotypes occur within or near enhancer loci, these cryptic non-coding portions of the genome might be used to match patients to a specific and optimized treatment options. As we start to move away from viewing the genome as a linear track and more as a complex shoe-string in a non-stationary fluid, transcriptional regulation will likely take on new meaning with important therapeutic treatments following suit.

Bibliography

[1] Karen Adelman and John T. Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet, 13(10):720–731, 10 2012.

[2] Syed Aftab, Lucie Semenec, Jeffrey Shih-Chieh Chu, and Nansheng Chen. Identification and characterization of novel human tissue-specific rfx transcription factors. BMC Evolutionary Biology, 8(1):226, 2008.

[3] Mary A Allen, Hestia Mellert, Veronica Dengler, Zdenek Andryzik, Anna Guarnieri, Justin A Freeman, Xin Luo, William L Kraus, Robin D Dowell, and Joaquín M Espinosa. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. eLife, 3:e02200, 2014.

[4] Karmel A. Allison, Minna U. Kaikkonen, Terry Gaasterland, and Christopher K. Glass. Vespucci: a system for building annotated databases of nascent transcripts. Nucleic Acids Research, 2013. PMID: 24304890.

[5] Karmel A Allison, Minna U Kaikkonen, Terry Gaasterland, and Christopher K Glass. Vespucci: a system for building annotated databases of nascent transcripts. Nucleic acids research, 42(4):2433–2447, 2014.

[6] K. Anamika, A. Gyenis, and L. Tora. How to stop: The mysterious links among RNA polymerase II occupancy 3′ of genes, mRNA 3′ processing and termination. Transcription, 4(1):7–12, 2013.

[7] Simon Anders. Analysing rna-seq data with the deseq package. Mol Biol, 43(4):1–17, 2010.

[8] A. G. Arimbasseri, K. Rijal, and R. J. Maraia. Comparative overview of RNA polymerase II and III transcription cycles, with focus on RNA polymerase III termination and reinitiation. Transcription, 5(1):e27639, Dec 2013.

[9] J. Azofeifa, M. Allen, M. Lladser, and R. Dowell. An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-seq. IEEE/ACM Transactions on Computational Biology and Bioinformatics, PP(99):1–1, 2016.

[10] Joseph Azofeifa, Mary A Allen, Manuel Lladser, and Robin Dowell. An annotation agnostic algorithm for detecting nascent rna transcripts. In IEEE Transactions in Computational Biology. ACM, 2015.

[11] Joseph Azofeifa, Mary A. Allen, Manuel E. Lladser, and Robin Dowell. FStitch: A fast and simple algorithm for detecting nascent RNA transcripts. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14, pages 174–183, New York, NY, USA, 2014. ACM.

[12] Timothy L Bailey and William Stafford Noble. Searching for statistically significant regulatory modules. Bioinformatics, 19(suppl 2):ii16–ii25, 2003.

[13] Jordana Bell, Athma Pai, Joseph Pickrell, Daniel Gaffney, Roger Pique-Regi, Jacob Degner, Yoav Gilad, and Jonathan Pritchard. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology, 12(1):R10, 2011.

[14] William P Bennett, S Perwez Hussain, Kirsi H Vahakangas, Mohammed A Khan, Peter G Shields, and Curtis C Harris. Molecular epidemiology of human cancer risk: gene–environment interactions and p53 mutation spectrum in human lung cancer. The Journal of pathology, 187(1):8–18, 1999.

[15] David L. Bentley. Coupling mRNA processing with transcription in time and space. Nat Rev Genet, 15(3):163–175, 03 2014.

[16] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute, 4(510):126, 1998.

[17] Jeff A Bilmes et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 4(510):126, 1998.

[18] Yonatan Bilu and Naama Barkai. The design of transcription-factor binding sites is affected by combinatorial regulation. Genome Biology, 6, 2005.

[19] N. Bouguila and D. Ziou. A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture. IEEE Trans Image Process, 15(9):2657–2668, Sep 2006.

[20] Albert A Bowers, Nathan West, Tenaya L Newkirk, Annie E Troutman-Youngman, Stuart L Schreiber, Olaf Wiest, James E Bradner, and Robert M Williams. Synthesis and hdac inhibitory activity of largazole analogs: Alteration of the zinc-binding domain and macrocyclic scaffold. Organic letters, 11(6):1301–1304, 03 2009.

[21] Paul Brennan, Sarah Lewis, Mia Hashibe, Douglas A Bell, Paolo Boffetta, Christine Bouchardy, Neil Caporaso, Chu Chen, Christiane Coutelle, Scott R Diehl, et al. Pooled analysis of alcohol dehydrogenase genotypes and head and neck cancer: a huge review. American journal of epidemiology, 159(1):1–16, 2004.

[22] Kathryn J Brown, Susan F Maynes, Anna Bezos, Deborah J Maguire, Miriam D Ford, and Christopher R Parish. A novel in vitro assay for human angiogenesis. Laboratory investigation; a journal of technical methods and pathology, 75(4):539–555, 1996.

[23] Martha L Bulyk. Computational prediction of transcription-factor binding site locations. Genome biology, 5(1):201, 2003.

[24] Eliezer Calo and Joanna Wysocka. Modification of enhancer chromatin: what, how, and why? Molecular cell, 49(5):825–837, 2013.

[25] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing gwas results: a review of statistical methods and recommendations for their application. The American Journal of Human Genetics, 86(1):6–22, 2010.

[26] L. H. Chadwick. The NIH Roadmap Epigenomics Program data resource. Epigenomics, 4(3):317–324, Jun 2012.

[27] Minho Chae, Charles G Danko, and W Lee Kraus. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. BMC bioinformatics, 16(1):222, 2015.

[28] David Maxwell Chickering. Learning bayesian networks is np-complete. In Learning from data, pages 121–130. Springer, 1996.

[29] David Maxwell Chickering, David Heckerman, and Christopher Meek. A bayesian approach to learning bayesian networks with local structure. In Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence, pages 80–89. Morgan Kaufmann Publishers Inc., 1997.

[30] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. Science, 306(5696):636–640, 2004.

[31] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74, 09 2012.

[32] The UniProt Consortium. Uniprot: a hub for protein information. Nucleic Acids Research, 2014.

[33] Leighton Core and John Lis. Transcription regulation through promoter-proximal pausing of RNA polymerase II. Science, 319:1791, 2008. PMID: 18369138.

[34] Leighton J Core, Andre L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet, 46(12):1311–1320, 12 2014.

[35] Leighton J. Core, Joshua J. Waterfall, Daniel A. Gilchrist, David C. Fargo, Hojoong Kwak, Karen Adelman, and John T. Lis. Defining the status of RNA polymerase at promoters. Cell Reports, 2(4):1025 – 1035, 2012.

[36] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322(5909):1845–1848, 2008.

[37] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences, 107(50):21931–21936, 2010.

[38] Charles G Danko, Stephanie L Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Hyung Won Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Meth, 12(5):433–438, 05 2015.

[39] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature, 505(7484):559–563, 2014.

[40] Luis M De Campos, Juan M Fernandez-Luna, José A Gámez, and José M Puerta. Ant colony optimization for learning bayesian networks. International Journal of Approximate Reasoning, 31(3):291–311, 2002.

[41] Binwei Deng, Svitlana Melnik, and Peter R Cook. Transcription factories, chromatin loops, and the dysregulation of gene expression in malignancy. In Seminars in cancer biology, volume 23, pages 65–71. Elsevier, 2013.

[42] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research, 22(9):1775–1789, 2012.

[43] Ivo D Dinov. Expectation maximization and mixture modeling tutorial. Statistics Online Computational Resource, 2008.

[44] Josée Dostie, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D. Green, and Job Dekker. Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. Genome Research, 16(10):1299–1309, 2006.

[45] Karen G Dowell, Allen K Simons, Hao Bai, Braden Kell, Zack Z Wang, Kyuson Yun, and Matthew A Hibbs. Novel insights into embryonic stem cell self-renewal revealed through comparative human and mouse systems biology networks. Stem Cells, 32(5):1161–1172, 2014.

[46] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform, 35(5-6):352–359, 2002.

[47] Alison M Dunning, Catherine S Healey, Paul DP Pharoah, M Dawn Teare, Bruce AJ Ponder, and Douglas F Easton. A systematic review of genetic polymorphisms and breast cancer risk. Cancer Epidemiology Biomarkers & Prevention, 8(10):843–854, 1999.

[48] Sascha H.C. Duttke, Scott A. Lacadie, Mahmoud M. Ibrahim, Christopher K. Glass, David L. Corcoran, Christopher Benner, Sven Heinz, James T. Kadonaga, and Uwe Ohler. Human promoters are intrinsically directional. Molecular Cell, 57(4):674 – 684, 2015.

[49] Sean R Eddy. Hidden markov models. Current opinion in structural biology, 6(3):361–365, 1996.

[50] Lucas Brandon Edelman and Peter Fraser. Transcription factories: genetic programming in three dimensions. Current Opinion in Genetics & Development, 22(2):110 – 114, 2012. PMID: 22365496.

[51] Uche I Ezeh, Paul J Turek, Renee A Reijo, and Amander T Clark. Human embryonic stem cell genes oct4, nanog, stellar, and gdf3 are expressed in both seminoma and breast carcinoma. Cancer, 104(10):2255–2265, 2005.

[52] José P Faria, Ross Overbeek, Fangfang Xia, Miguel Rocha, Isabel Rocha, and Christopher S Henry. Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. Briefings in bioinformatics, 15(4):592–611, 2014.

[53] William W Fisher, Jingyi Jessica Li, Ann S Hammonds, James B Brown, Barret D Pfeiffer, Richard Weiszmann, Stewart MacArthur, Sean Thomas, John A Stamatoyannopoulos, Michael B Eisen, et al. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proceedings of the National Academy of Sciences, 109(52):21330–21335, 2012.

[54] Nova Fong, Kristopher Brannan, Benjamin Erickson, Hyunmin Kim, Michael A. Cortazar, Ryan M. Sheridan, Tram Nguyen, Shai Karp, and David L. Bentley. Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA polymerase II termination suggest widespread kinetic competition. Molecular Cell, 60(2):256–267, 2015/10/31 2015.

[55] Nova Fong, Hyunmin Kim, Yu Zhou, Xiong Ji, Jinsong Qiu, Tassa Saldi, Katrina Diener, Ken Jones, Xiang-Dong Fu, and David L. Bentley. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes & Development, 28(23):2663–2676, 2014.

[56] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. Machine learning, 29(2-3):131–163, 1997.

[57] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. Journal of computational biology, 7(3-4):601–620, 2000.

[58] S. Frietze, R. Wang, L. Yao, Y. G. Tak, Z. Ye, M. Gaddis, H. Witt, P. J. Farnham, and V. X. Jin. Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. Genome Biol., 13(9):R52, 2012.

[59] Nicholas J. Fuda, M. Behfar Ardehali, and John T. Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature, 461(7261):186–192, Sep 2009. PMID: 19741698.

[60] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462(7269):58–64, Nov 2009.

[61] José Garcıa-Martınez, Agustın Aranda, and José E Pérez-Ortın. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Molecular cell, 15(2):303–313, 2004.

[62] Doron Gothelf, Stephan Eliez, Tracy Thompson, Christine Hinard, Lauren Penniman, Carl Feinstein, Hower Kwon, Shuting Jin, Booil Jo, Stylianos E Antonarakis, et al. Comt genotype predicts longitudinal cognitive decline and psychosis in 22q11. 2 deletion syndrome. Nature neuroscience, 8(11):1500–1502, 2005.

[63] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. Nature genetics, 47(6):569–576, 2015.

[64] A Grilli, M Sciandra, M Terracciano, P Picci, and K Scotlandi. Integrated approaches to mirnas target definition: time-series analysis in an osteosarcoma differentiative model. BMC medical genomics, 8(1):1, 2015.

[65] Aric Hagberg, Dan Schult, Pieter Swart, D Conway, L Séguin-Charbonneau, C Ellison, B Edwards, and J Torrents. Networkx. high productivity software for complex networks. Webová strá nka https://networkx. lanl. gov/wiki, 2004.

[66] Nasun Hah, CharlesG. Danko, Leighton Core, JoshuaJ. Waterfall, Adam Siepel, JohnT. Lis, and W.Lee Kraus. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell, 145(4):622 – 634, 2011. PMID: 21549415.

[67] Nasun Hah, Shino Murakami, Anusha Nagari, Charles G. Danko, and W. Lee Kraus. Enhancer transcripts mark active estrogen receptor binding sites. Genome Research, 23(8):1210–1223, 2013.

[68] Florian Halbritter. Geneprof manual, 2013.

[69] H. H. He, C. A. Meyer, M. W. Chen, V. C. Jordan, M. Brown, and X. S. Liu. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res., 22(6):1015–1025, Jun 2012.

[70] David Heckerman. Bayesian networks for data mining. Data mining and knowledge discovery, 1(1):79–119, 1997.

[71] David Heckerman. A tutorial on learning with Bayesian networks. Springer, 1998.

[72] Denes Hnisz, Jurian Schuijers, Charles Y Lin, Abraham S Weintraub, Brian J Abraham, Tong Ihn Lee, James E Bradner, and Richard A Young. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Molecular cell, 58(2):362–370, 2015.

[73] D. Hu, E. R. Smith, A. S. Garruss, N. Mohaghegh, J. M. Varberg, C. Lin, J. Jackson, X. Gao, A. Saraf, L. Florens, M. P. Washburn, J. C. Eissenberg, and A. Shilatifard. The little elongation complex functions at initiation and elongation phases of snRNA gene transcription. Mol. Cell, 51(4):493–505, Aug 2013.

[74] Xiong Ji, Yu Zhou, Shatakshi Pandit, Jie Huang, Hairi Li, Charles?Y Lin, Rui Xiao, Christopher?B Burge, and Xiang-Dong Fu. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. Cell, 153(4):855–868, May 2013. PMID: 23663783.

[75] Fulai Jin, Yan Li, Jesse R. Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature, 503(7475):290–294, 11 2013.

[76] Iris Jonkers and John T Lis. Getting up to speed with transcription elongation by RNA polymerase II. Nature Reviews Molecular Cell Biology, 16(3):167–177, 2015.

[77] R. Joseph, Y. L. Orlov, M. Huss, W. Sun, S. L. Kong, L. Ukil, Y. F. Pan, G. Li, M. Lim, J. S. Thomsen, Y. Ruan, N. D. Clarke, S. Prabhakar, E. Cheung, and E. T. Liu. Integrative model of genomic factors for determining binding site selection by estrogen receptor. Mol. Syst. Biol., 6:456, Dec 2010.

[78] Aashiq H Kachroo, Jon M Laurent, Christopher M Yellman, Austin G Meyer, Claus O Wilke, and Edward M Marcotte. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science, 348(6237):921–925, 2015.

[79] Minna U. Kaikkonen, Nathanael J. Spann, Sven Heinz, Casey E. Romanoski, Karmel A. Allison, Joshua D. Stender, Hyun B. Chun, David F. Tough, Rab K. Prinjha, Christopher Benner, and Christopher K. Glass. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular Cell, 51(3):310 – 325, 2013.

[80] Philipp Kapranov, Aarron T. Willingham, and Thomas R. Gingeras. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet, 8(6):413–423, Jun 2007. PMID: 17486121.

[81] Hideya Kawaji, Jessica Severin, Marina Lizio, Alistair R. R. Forrest, Erik van Nimwegen, Michael Rehli, Kate Schroder, Katharine Irvine, Harukazu Suzuki, Piero Carninci, Yoshihide Hayashizaki, and Carsten O. Daub. Update of the fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. Nucleic Acids Research, 39(suppl 1):D856–D860, 2011.

[82] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, et al. Defining functional dna elements in the human genome. Proceedings of the National Academy of Sciences, 111(17):6131–6138, 2014.

[83] W James Kent, Ann S Zweig, G Barber, Angie S Hinrichs, and Donna Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics, 26(17):2204–2207, 2010.

[84] Aziz Khan and Xuegong Zhang. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research, 44(D1):D164–D171, 2016.

[85] Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295):182–187, 2010.

[86] Tae-kyung Kim, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F. Worley, Gabriel Kreiman, and Michael E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295):182–187, May 2010. PMID: 20393465.

[87] Mary-Claire King and Allan C Wilson. Evolution at two levels in humans and chimpanzees. 1975.

[88] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. Plant methods, 9(1):1, 2013.

[89] Ivan V. Kulakovskiy, Yulia A. Medvedeva, Ulf Schaefer, Artem S. Kasianov, Ilya E. Vorontsov, Vladimir B. Bajic, and Vsevolod J. Makeev. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Research, 41(D1):D195–D202, 2013.

[90] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Anastasiia V Soboleva, Artem S Kasianov, Haitham Ashoor, Wail Ba-alawi, Vladimir B Bajic, Yulia A Medvedeva, Fedor A Kolpakov, et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic acids research, 44(D1):D116–D125, 2016.

[91] Hojoong Kwak, Nicholas J Fuda, Leighton J Core, and John T Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (New York, N.Y.), 339(6122):950–953, 02 2013.

[92] Clélia Laitem, Justyna Zaborowska, Nur F Isa, Johann Kufs, Martin Dienstbier, and Shona Murphy. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes. Nat Struct Mol Biol, 22(5):396–403, 05 2015.

[93] Ben Langmead, Kasper Hansen, and Jeffrey Leek. Cloud-scale rna-sequencing differential expression analysis with myrna. Genome Biology, 11(8):R83, 2010.

[94] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. Nat Meth, 9(4):357–359, Apr 2012. PMID: 22388286.

[95] Pedro Larrañaga, Ramon Etxeberria, José A Lozano, and José M Peña. Combinatorial optimization by learning and simulation of bayesian networks. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pages 343–352. Morgan Kaufmann Publishers Inc., 2000.

[96] Erica Larschan, Eric P. Bishop, Peter V. Kharchenko, Leighton J. Core, John T. Lis, Peter J. Park, and Mitzi I. Kuroda. X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. Nature, 471(7336):115–118, Mar 2011. PMID: 21368835.

[97] Thien P. Le, Miao Sun, Xin Luo, W. Lee Kraus, and Geoffrey L. Greene. Mapping ERβ genomic binding sites reveals unique genomic features and identifies EBF1 as an ERβ interactor. PLoS ONE, 8(8):e71355, 08 2013.

[98] Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. Cell, 152(6):1237 – 1251, 2013.

[99] Guoliang Li, Melissa Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, Chia-Lin Wei, Yijun Ruan, and Wing-Kin Sung. Chia-pet tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biology, 11(2):R22, 2010.

[100] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14):1754–1760, 2009. PMID: 19451168.

[101] Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, Soohwan Oh, Hong-Sook Kim, Christopher K. Glass, and Michael G. Rosenfeld. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature, 498(7455):516–520, 06 2013.

[102] Wenbo Li, Dimple Notani, and Michael G. Rosenfeld. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet, 17(4):207–223, 04 2016.

[103] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol, 6(2):e27, 2008.

[104] Ziv Lifshitz, David Burstein, Kierstyn Schwartz, Howard A Shuman, Tal Pupko, and Gil Segal. Identification of novel coxiella burnetii icm/dot effectors and genetic analysis of their involvement in modulating a mitogen-activated protein kinase pathway. Infection and immunity, 82(9):3740–3752, 2014.

[105] Wen Liu, Qi Ma, Kaki Wong, Wenbo Li, Kenny Ohgi, Jie Zhang, Aneel Aggarwal, and Michael G Rosenfeld. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. Cell, 155(7):1581–1595, 12 2013.

[106] Xin Luo, Minho Chae, Raga Krishnakumar, Charles G Danko, and W Lee Kraus. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. BMC Genomics, 15:155–155, 2014.

[107] Kenzie D MacIsaac, Ting Wang, D Benjamin Gordon, David K Gifford, Gary D Stormo, and Ernest Fraenkel. An improved map of conserved regulatory sites for saccharomyces cerevisiae. BMC bioinformatics, 7(1):1, 2006.

[108] David JC MacKay. Information theory, inference and learning algorithms. Cambridge university press, 2003.

[109] Jeroen Maertzdorf, January Weiner, Hans-Joachim Mollenkopf, TBornotTB Network, Torsten Bauer, Antje Prasse, Joachim Müller-Quernheim, Stefan HE Kaufmann, Oswald Bellinger, Roland Diel, et al. Common patterns and disease-related signatures in tuberculosis and sarcoidosis. Proceedings of the National Academy of Sciences, 109(20):7853–7858, 2012.

[110] Erna Magnúsdóttir, Sabine Dietmann, Kazuhiro Murakami, Ufuk Günesdogan, Fuchou Tang, Siqin Bao, Evangelia Diamanti, Kaiqin Lao, Berthold Gottgens, and M Azim Surani. A tripartite transcription factor network regulates primordial germ cell specification in mice. Nature cell biology, 15(8):905–915, 2013.

[111] Florian Markowetz and Rainer Spang. Inferring cellular networks–a review. BMC bioinformatics, 8(Suppl 6):S5, 2007.

[112] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. Science, 337(6099):1190–1195, 2012.

[113] A. McCallum, D. Freitag, and F. Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation. 17th International Conf. on Machine Learning, 2000.

[114] G. J. McLachlan and P. N. Jones. Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics, 44(2):571–578, Jun 1988.

[115] Michael Melgar, Francis Collins, and Praveen Sethupathy. Discovery of active enhancers through bidirectional expression of short transcripts. Genome Biology, 12(11):R113, 2011.

[116] CarlosA. Melo, Jarno Drost, PatrickJ. Wijchers, Harmen vandeWerken, Elzo deWit, JoachimA.F.Oude Vrielink, Ran Elkon, SniaA. Melo, Nicolas L©veill©, Raghu Kalluri, Wouter deLaat, and Reuven Agami. eRNAs are required for p53-dependent enhancer activity and gene transcription. Molecular Cell, 49(3):524 – 535, 2013.

[117] Andreas Milias-Argeitis, Ana Paula Oliveira, Luca Gerosa, Laura Falter, Uwe Sauer, and John Lygeros. Elucidation of genetic interactions in the yeast gata-factor network using bayesian model selection. PLOS Comput Biol, 12(3):e1004784, 2016.

[118] Irene M. Min, Joshua J. Waterfall, Leighton J. Core, Robert J. Munroe, John Schimenti, and John T. Lis. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes & Development, 25(7):742–754, 2011. PMID: 21460038.

[119] Kaoru Mitsui, Yoshimi Tokuzawa, Hiroaki Itoh, Kohichi Segawa, Mirei Murakami, Kazutoshi Takahashi, Masayoshi Maruyama, Mitsuyo Maeda, and Shinya Yamanaka. The homeoprotein nanog is required for maintenance of pluripotency in mouse epiblast and es cells. cell, 113(5):631–642, 2003.

[120] Alexandra Montagner, Agata Korecka, Arnaud Polizzi, Yannick Lippi, Yuna Blum, Cécile Canlet, Marie Tremblay-Franco, Amandine Gautier-Stein, Rémy Burcelin, Yi-Chun Yen, et al. Hepatic circadian clock oscillators and nuclear receptors integrate microbiome-derived signals. Scientific reports, 6, 2016.

[121] S. Moon and J. N. Hwang. Robust speech recognition based on joint model and feature space optimization of hidden Markov models. IEEE Trans Neural Netw, 8(2):194–204, 1997.

[122] Kevin Murphy, Saira Mian, et al. Modelling gene expression data using dynamic bayesian networks. Technical report, Technical report, Computer Science Division, University of California, Berkeley, CA, 1999.

[123] Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

[124] Gioacchino Natoli and Jean-Christophe Andrau. Noncoding transcription at enhancers: General principles and functional models. Annual Review of Genetics, 46(1):1–19, 2012. PMID: 22905871.

[125] Mark EJ Newman. Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23):8577–8582, 2006.

[126] Benjamin Neymotin, Rodoniki Athanasiadou, and David Gresham. Determination of in vivo rna kinetics using rate-seq. RNA, 20(10):1645–1652, 10 2014.

[127] F Nikulenkov, C Spinnler, H Li, C Tonelli, Y Shi, M Turunen, T Kivioja, I Ignatiev, A Kel, J Taipale, and Selivanova G. Insights into p53 transcriptional function via genome-wide chromatin occupancy and gene expression analysis. Cell Death and Differentiation, 19:1992–2002, 2013.

[128] Takayuki Nojima, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J Proudfoot. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. Cell, 161(3):526–540, 2015.

[129] K. Ogoshi, S. Hashimoto, Y. Nakatani, W. Qu, K. Oshima, K. Tokunaga, S. Sugano, M. Hattori, S. Morishita, and K. Matsushima. Genome-wide profiling of DNA methylation in human cancer cells. Genomics, 98(4):280–287, Oct 2011.

[130] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics, 10(10):669–680, 2009.

[131] Judea Pearl. Belief networks revisited. Artificial intelligence in perspective, pages 49–56, 1994.

[132] Judea Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, 2014.

[133] Len A Pennacchio, Nadav Ahituv, Alan M Moses, Shyam Prabhakar, Marcelo A Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D Lewis, et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature, 444(7118):499–502, 2006.

[134] Len A Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A Nobrega, and Gill Bejerano. Enhancers: five essential questions. Nature Reviews Genetics, 14(4):288–295, 2013.

[135] Jennifer E. Phillips and Victor G. Corces. CTCF: Master weaver of the genome. Cell, 137(7):1194–1211, 2015/05/16 2009.

[136] P. Preker, K. Almvig, M. S. Christensen, E. Valen, C. K. Mapendano, A. Sandelin, and T. H. Jensen. Promoter upstream transcripts share characteristics with mrnas and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Research, 2011.

[137] Jason Qian, Qiao Wang, Marei Dose, Nathanael Pruett, Kyong-Rim Kieffer-Kwon, Wolfgang Resch, Genqing Liang, Zhonghui Tang, Ewy Mathé, Christopher Benner, et al. B cell super-enhancers and regulatory clusters recruit aid tumorigenic activity. Cell, 159(7):1524–1537, 2014.

[138] Víctor Quesada, Laura Conde, Neus Villamor, Gonzalo R Ordóñez, Pedro Jares, Laia Bassaganyas, Andrew J Ramsay, Sílvia Beà, Magda Pinyol, Alejandra Martínez-Trillos, et al. Exome sequencing identifies recurrent mutations of the splicing factor sf3b1 gene in chronic lymphocytic leukemia. Nature genetics, 44(1):47–52, 2012.

[139] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841–842, 2010.

[140] Owen J L Rackham, Jaber Firas, Hai Fang, Matt E Oates, Melissa L Holmes, Anja S Knaupp, The FANTOM Consortium, Harukazu Suzuki, Christian M Nefzger, Carsten O Daub, Jay W Shin, Enrico Petretto, Alistair R R Forrest, Yoshihide Hayashizaki, Jose M Polo, and Julian Gough. A predictive computational framework for direct reprogramming between human cell types. Nat Genet, 48(3):331–335, 03 2016.

[141] Timothy Read, Phillip A Richmond, and Robin D Dowell. A trans-acting variant within the transcription factor RIM101 interacts with genetic background to determine its regulatory capacity. PLoS Genet, 12(1):e1005746, 2016.

[142] William J Reed and Murray Jorgensen. The double pareto-lognormal distributiona new parametric model for size distributions. Communications in Statistics-Theory and Methods, 33(8):1733–1753, 2004.

[143] Douglas A Reynolds. Automatic speaker recognition using Gaussian mixture speaker models. In The Lincoln Laboratory Journal. Citeseer, 1995.

[144] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, 2010.

[145] Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. BMC bioinformatics, 12(1):414, 2011.

[146] Sarah Sainsbury, Carrie Bernecky, and Patrick Cramer. Structural basis of transcription initiation by rna polymerase ii. Nature Reviews Molecular Cell Biology, 16(3):129–143, 2015.

[147] Daniel Savic, Brian S Roberts, Julia B Carleton, E Christopher Partridge, Michael A White, Barak A Cohen, Gregory M Cooper, Jason Gertz, and Richard M Myers. Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/enhancer-binding protein beta binding sites. Genome research, pages gr–191593, 2015.

[148] Sequence Read Archive Submissions Staff. Using the SRA Toolkit to convert .sra files into other formats, 2011.

[149] Richard I Sherwood, Tatsunori Hashimoto, Charles W O'Donnell, Sophia Lewis, Amira A Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling dnase profile magnitude and shape. Nature biotechnology, 32(2):171–178, 2014.

[150] Takashi Shimamoto, K Ohyashiki, JH Ohyashiki, K Kawakubo, T Fujimura, H Iwama, S Nakazawa, and K Toyama. The expression pattern of erythrocyte/megakaryocyte-related transcription factors gata-1 and the stem cell leukemia gene correlates with hematopoietic differentiation and is associated with outcome of acute myeloid leukemia. Blood, 86(8):3173–3180, 1995.

[151] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet, 15(4):272–286, April 2014.

[152] L. Smeenk, S.J. van Heeringen, M. Koeppel, B. Gilbert, E. Janssen-Megens, H.G. Stunnenberg, and M. Lohrum. Role of p53 serine 46 in p53 target gene regulation. PLoS ONE, 6(3):e17574, 03 2011.

[153] L. Smeenk, S.J. van Heeringen, M. Koeppel, M.A. van Driel, S.J.J. Bartels, R.C. Akkers, S. Denissov, H.G. Stunnenberg, and M. Lohrum. Characterization of genome-wide p53-binding sites upon stress response. Nucleic Acids Research, 36(11):3639–3654, 2008.

[154] Erik Splinter, Helen Heath, Jurgen Kooren, Robert-Jan Palstra, Petra Klous, Frank Grosveld, Niels Galjart, and Wouter de Laat. CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. Genes & development, 20(17):2349–2354, 2006.

[155] Rodger Staden. Staden: Searching for Motifs in Nucleic Acid Sequences, pages 93–102. Springer New York, Totowa, NJ, 1994.

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

[157] John D. Storey, Jennifer Madeoy, Jeanna L. Strout, Mark Wurfel, James Ronald, and Joshua M. Akey. Gene-expression variation within and among human populations. The American Journal of Human Genetics, 80(3):502–509, 2015/04/23 2007.

[158] Gary D Stormo. Dna binding sites: representation and discovery. Bioinformatics, 16(1):16–23, 2000.

[159] Gary D. Stormo. Introduction to Protein-DNA Interactions: Structure, Thermodynamics, and Bioinformatics. Cold Spring Harbor Laboratory Press, 2013.

[160] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell, 126(4):663–676, 2006.

[161] P. Tamayo, D. Scanfeld, B.L. Ebert, M.A. Gillette, C.W.M. Roberts, and J.P. Mesirov. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. Proceedings of the National Academy of Sciences, 104(14):5959–5964, 2007.

[162] Ge Tan. Cne identification and visualisation, 2016.

[163] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447(7146):799–816, June 2007. PMID: 17571346.

[164] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74, Sep 2012. PMID: 22955616.

[165] H. Thorvaldsdttir, J.T. Robinson, and J.P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14(2):178–192, 2013.

[166] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. Machine learning, 65(1):31–78, 2006.

[167] TS Vermal J udea Pearl. Equivalence and synthesis of causal models. In Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence, pages 220–227, 1991.

[168] Atle van Beelen Granlund, Arnar Flatberg, Ann E Østvik, Ignat Drozdov, Bjørn I Gustafsson, Mark Kidd, Vidar Beisvag, Sverre H Torp, Helge L Waldum, Tom Christian Martinsen, et al. Whole genome gene expression meta-analysis of inflammatory bowel disease colon mucosa demonstrates lack of major differences between crohn's disease and ulcerative colitis. PLoS One, 8(2):e56818, 2013.

[169] Robert J Vanderbei and David F Shanno. An interior-point algorithm for nonconvex nonlinear programming. Computational Optimization and Applications, 13(1-3):231–252, 1999.

[170] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics, 10(4):252–263, 2009.

[171] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. Nat Rev Genet, 10(4):252–263, Apr 2009. PMID: 19274049.

[172] Dong Wang, Ivan Garcia-Bassets, Chris Benner, Wenbo Li, Xue Su, Yiming Zhou, Jinsong Qiu, Wen Liu, Minna U. Kaikkonen, Kenneth A. Ohgi, Christopher K. Glass, Michael G. Rosenfeld, and Xiang-Dong Fu. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature, 474(7351):390–394, Jun 2011. PMID: 21572438.

[173] Q Wang, M_E Curran, I Splawski, TC Burn, JM Millholland, TJ VanRaay, J Shen, KW Timothy, GM Vincent, T De Jager, et al. Positional cloning of a novel potassium channel gene: Kvlqt1 mutations cause cardiac arrhythmias. Nature genetics, 12(1):17–23, 1996.

[174] Siyu Wang, Jinbo Xu, and Jianyang Zeng. Inferential modeling of 3d chromatin structure. Nucleic Acids Research, 2015.

[175] C. Wei, Q. Wu, V. Vega, K. Chiu, P. Ng, T. Zhang, A. Shahab, H. Yong, Y. Fu, and Z. Weng. A global map of p53 transcription-factor binding sites in the human genome. Cell, 124(1):207–219, January 2006. PMID: 16413492.

[176] Warren A Whyte, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, Michael H Kagey, Peter B Rahl, Tong Ihn Lee, and Richard A Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell, 153(2):307–319, 2013.

[177] Adam N Yadon, Daniel Van de Mark, Ryan Basom, Jeffrey Delrow, Iestyn Whitehouse, and Toshio Tsukiyama. Chromatin remodeling around nucleosome-free regions leads to repression of noncoding RNA transcription. Molecular and cellular biology, 30(21):5110–5122, 2010.

[178] Hong Yue, Bo Yang, Fang Yang, Xiao-Li Hu, and Fan-Bin Kong. Co-expression network-based analysis of hippocampal expression data associated with alzheimer's disease using a novel algorithm. Experimental and Therapeutic Medicine, 11(5):1707–1715, 2016.

[179] Yan Zhang, Hye Kyong Kweon, Christian Shively, Anuj Kumar, and Philip C Andrews. Towards systematic discovery of signaling networks in budding yeast filamentous growth stress response using interventional phosphorylation data. PLoS Comput Biol, 9(6):e1003077, 2013.

[180] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). Genome biology, 9(9):1, 2008.

[181] ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3d chromosome modeling with semi-definite programming and hi-c data. Journal of computational biology, 20(11):831–846, 2013.

[182] Wenxuan Zhong, Peng Zeng, Ping Ma, Jun S Liu, and Yu Zhu. Rsir: regularized sliced inverse regression for motif discovery. Bioinformatics, 21(22):4169–4175, 2005.

[183] Hong-Ming Zhou, Gisela Weskamp, Valérie Chesneau, Umut Sahin, Andrea Vortkamp, Keisuke Horiuchi, Riccardo Chiusaroli, Rebecca Hahn, David Wilkes, Peter Fisher, et al. Essential role for adam19 in cardiovascular morphogenesis. Molecular and cellular biology, 24(1):96–104, 2004.

Appendix A

Supplementary Material to Chapter 2

Table A.1: **Feature vector $\vec{x}$ associated with a contig.** Let $[t, t+l]$ denote the interval of genomic positions covered by a contig; in particular, l is the length of the contig. For each position i in this interval, let y_i denote the read count at i. Feature vector coordinates are ordered by importance via recursive feature elimination.

$\vec{x}$-coordinate	description	definition
x_0	bias term	1
x_1	length (contig or gap)	l
x_2	total count	$\sum_{i=t}^{t+l} y_i$
x_3	mean count	$\frac{1}{l}\sum_{i=t}^{t+l} y_i$
x_4	median count	$median(y_t, ..., y_{t+l})$
x_5	max count	$max(y_t, ..., y_{t+l})$
x_6	min count	$min(y_t, ..., y_{t+l})$
x_7	count variance	$\frac{1}{l-1}\sum_{i=t}^{t+l}(y_i - x_3)^2$

Table A.2: **Evaluation of bidirectional predictions as eRNAs** On the diagonal are total events in this category. The intersection of a row and column indicates the total overlap between these events. Significance of the overlap was assessed by hypergeometric and p-value are indicated as follows: $^{\S}10^{-3}$, $^{\dagger}10^{-4}$, $^{\|}10^{-8}$, $^{\ddagger}10^{-9}$

IMR90	bidirectional	DNAse	H3K27ac
bidirectional	5,177		
DNAse	1,892§	140,803	
H3K27ac	1,874$^{\|}$	20,673	57,623

MCF7	bidirectional	DNAse	H3K27ac
bidirectional	10,536		
DNAse	4,154†	152,768	
H3K27ac	4,554‡	13,673	32,516

HCT116	bidirectional	DNAse	H3K27ac
bidirectional	14,738		
DNAse	6,750†	114,060	
H3K27ac	2,417§	13,769	57,623

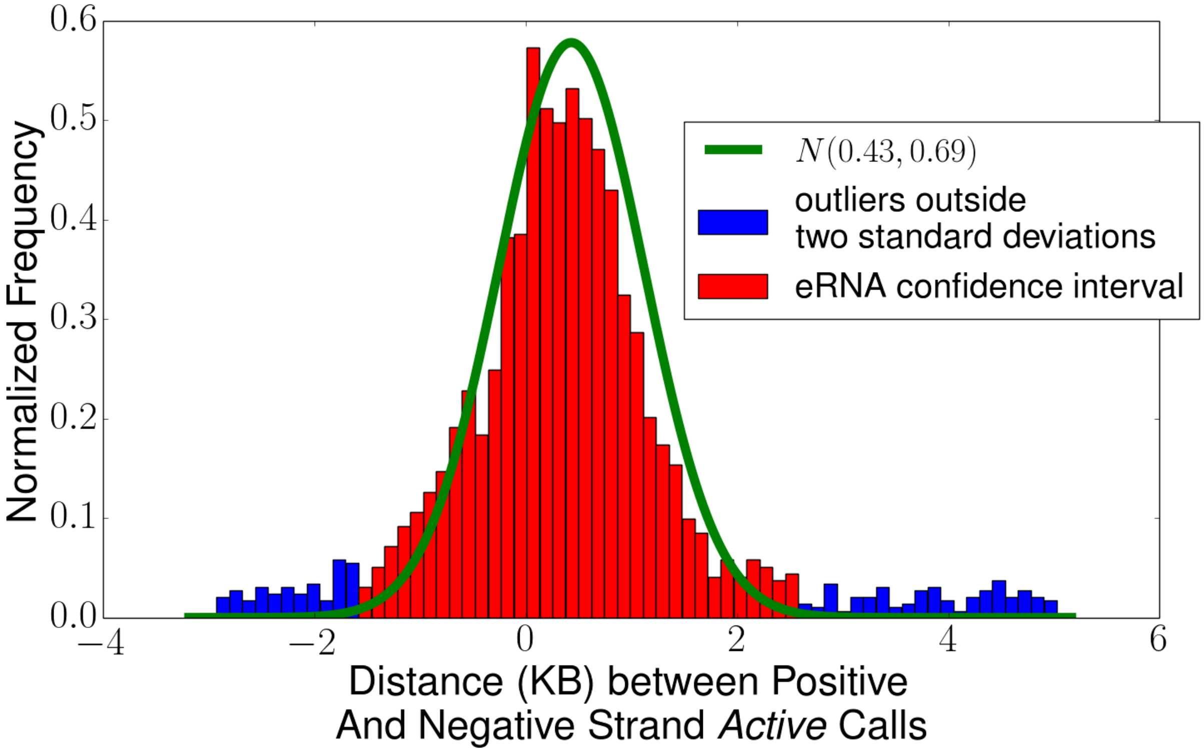


Figure A.1: **eRNA prediction Interval.** FStitch *active* calls that overlap both an H3K27ac and DNase I hypersensitivity chromatin marker were isolated on both strands. Sense and anti-sense strand calls are paired to their nearest neighbors. The difference between the start of the positive and the negative strand call are shown here. A positive value corresponds to overlap and negative to separation. The green line is a fitted Normal and the red bars indicate data points within two standard deviations.

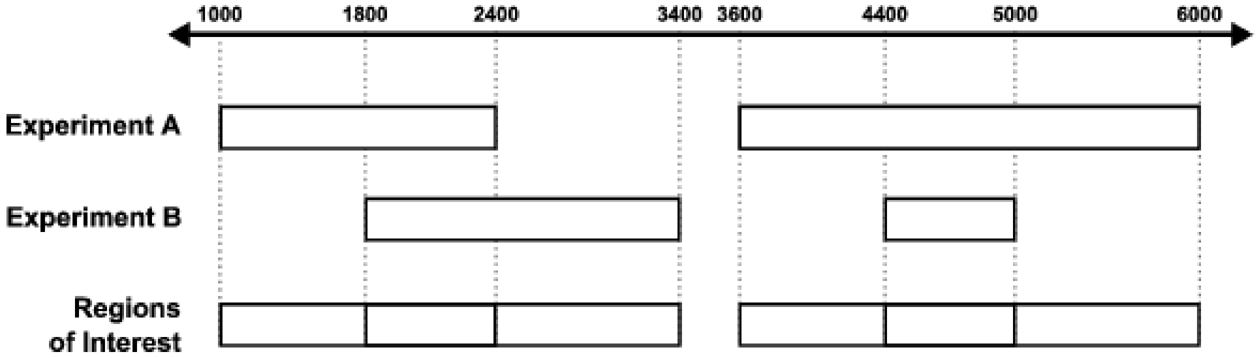


Figure A.2: **Schematic of the merged method of identifying regions of interest.** In this region (1000 to 6000) there are two *active* regions called in both Experiment A and Experiment B. The first *active* call in Experiment A (1000 to 2400) overlaps the first *active* call in Experiment B (1800 to 3400). All distinct regions of the overlap are retained in the regions of interest resulting in three segments: (1000 to 1800), (1800 to 2400), and (2400 to 3400). The example on the right shows how behavior is similar when a *active* call in Experiment B (4400 to 5000) is contained within a *active* call in Experiment A (3600 to 6000).

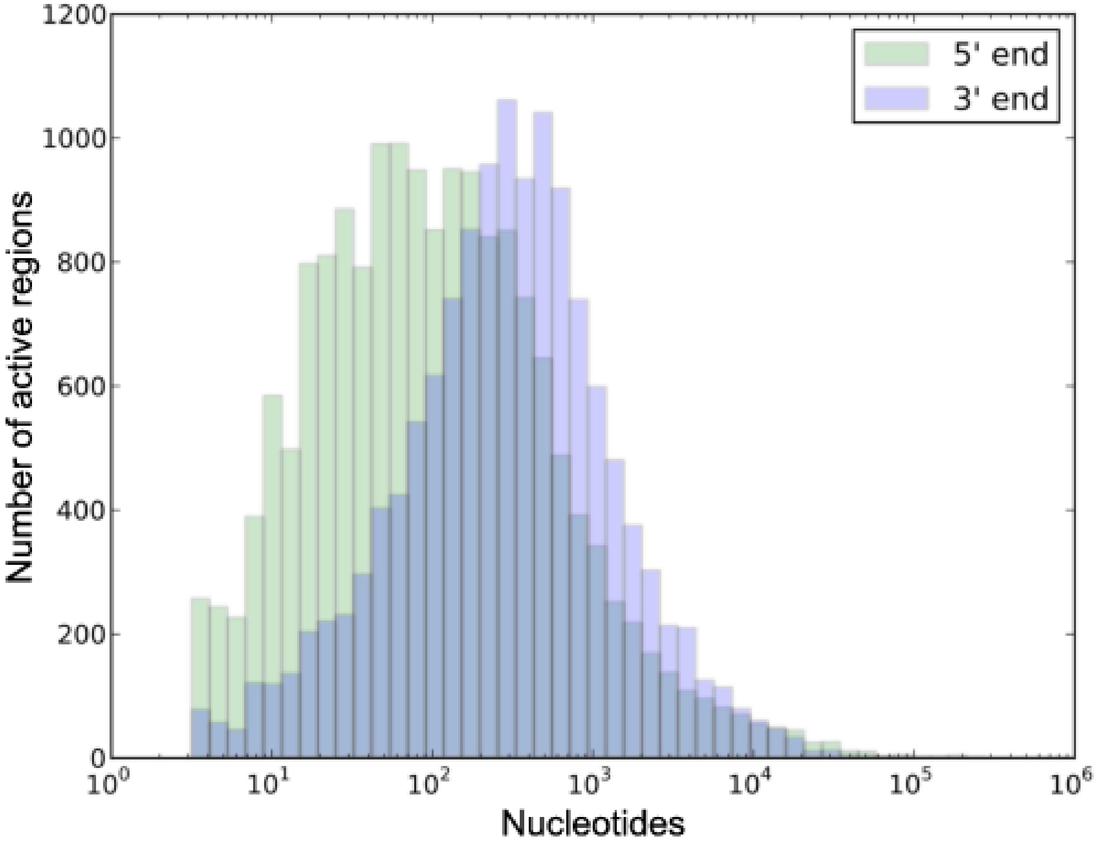


Figure A.3: **Changes in coordinates of 5′ or 3′ end of *active* call.** The distance between the 5′ end of an FStitch *active* call in DMSO compared to the nearest 5′ end in Nutlin (green). (Requiring the Nutlin call to be used only on the nearest DMSO call.) Similarly the 3′ end is shown in blue. Based on these distributions, we merged all regions smaller than 100 bp with an adjacent segment. It is also clear from the heavy tail that many regions change substantially in length between the two experiments.

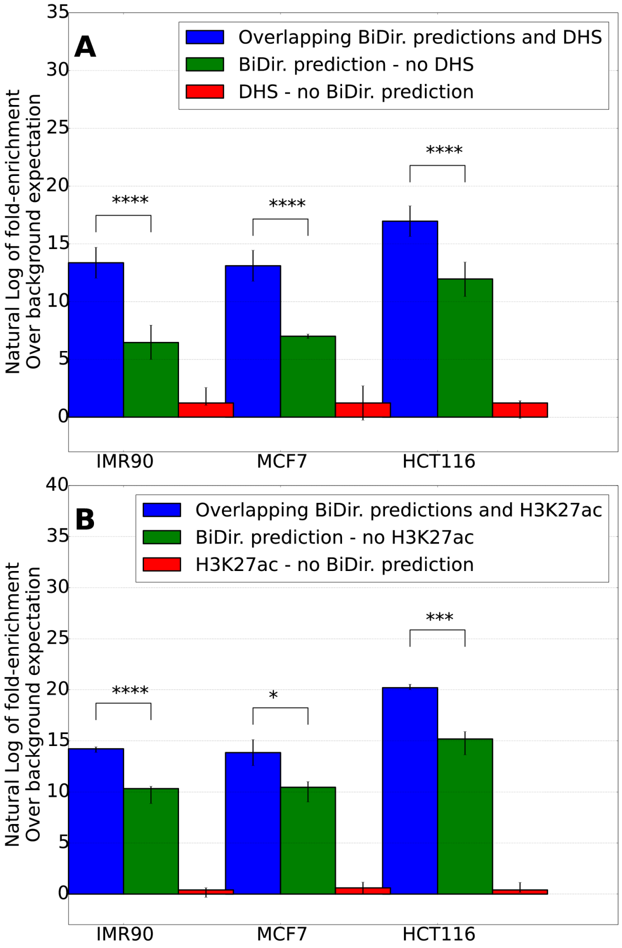


Figure A.4: **Bidirectional calls overlapping enhancer marks are highly transcribed.** We posit that bidirectional calls that overlap enhancer marks (A) the DNase I hypersensitivity (DHS) mark and (B) H3K27ac will show strong signatures of transcription. Blue: Overlap between the enhancer mark and a bidirectional call; Green: Bidirectional calls without overlap to an enhancer mark; Red: Enhancer marks without a corresponding overlapping bidirectional call.

Appendix B

Supplementary Material to Chapter 3

Algorithm 1 Adjusted EM Algorithm for RNAP Mixture Model Inference

```
Require: $\mathbf{D}$, $K$, $\epsilon$, $\tau$
Ensure: $K > 1$, $|\mathbf{D}| > 1$
Randomly Initialize $\Theta$
while $\mathcal{L}(\Theta^{t+1}) - \mathcal{L}(\Theta^{t}) > \epsilon$ do
    E-Step: Compute Expectations eq. 7 and 9
    for $i = 1 \rightarrow |\mathbf{D}|$ do
        for $k = 1 \rightarrow K$ do
            $r_i^k, \mathbb{E}[X|g_i;\theta^t], \mathbb{E}[X^2|g_i;\theta^t], \mathbb{E}[Y|g_i;\theta^t]$
        end for
    end for
    M-Step: Maximize Conditional Expectations eq. 10
    for $k = 1 \rightarrow K$ do
        $\theta_k^{t+1} := \underset{\theta}{\operatorname{argmax}} \mathbb{E}\left[\log p(\mathbf{C}|\theta_k)|\mathbf{D}, \theta_k^t\right]$
    end for
    $L_{old} = \mathcal{L}(\Theta^{t+1})$
    for $k = 1 \in \mathbb{K}_e$ do
        $l_{s,new} = l_{s,k} + \mathcal{N}(0, \tau)$
        if $\mathcal{L}(\Theta_{new}) > L_{old}$ then
            $L_{old} := \mathcal{L}(\Theta_{new})$
            $l_{s,k} := l_{s,new}$
        end if
    end for
end while
```

B.1 Seeding the EM

Like many gradient based optimization methods, the EM algorithm is subject to local maxima across the likelihood landscape. A common approach to handle this issue is to use many random initializations of $\Theta^{t=0}$, yet this approach is inherently time consuming. Seen commonly with Gaussian mixture models [143], both the EM's linear rate of convergence and final parameter quality, $\mathcal{L}(\Theta^*)$, vary most as a function of μ_k^0: the random initialization of the LI component center.

To this end, we propose an approximate windowing method to scan for regions of local likelihood for an LI component within $\mathbf{D}$. Let $\mathbf{D}_{[a,b]} = \{g_i : a \leq z_i \leq b\}$ be a specific ordered collection of $\mathbf{D}$. At $\mathbf{D}_{[a,b]}$ we compute the ratio of BIC scores of a fully specified single component mixture model (parameters Θ_B) to a uniform distribution with support across $[a,b]$ (equation B.1).

$$\text{LL}(\mathbf{D}_{[a,b]}) = \frac{\log N_{[a,b]} - 2N_{[a,b]} \log(0.5/(b-a))}{8\alpha \log N_{[a,b]} - 2\log \mathcal{L}(\Theta_B)} \tag{B.1}$$

It should be noted that the elongation l_s components l_+, l_- are set to b, a respectively. Overlapping windows of size $b-a$ are merged if LL(.) exceeds 1. Unless otherwise specified, $\alpha = 1$. These merged windows are then used for the random initialization μ_k as they constitute a heuristic estimate to the expected value of an LI component.

B.2 Datasets

We utilize previously published GRO-seq datasets to perform RNAP inference. Data from four separate publications were obtained as FASTQ files from the short read archive: SRP035278 [3], SRP031885 [105], SRR1552484 [34] and ERP009673 [92]. Reads were mapped by bowtie2 under the "−−very-sensitive" parameter settings. Mapped reads were converted to Bedgraph format via samtools and bedtools, versions 0.1.19 and 2.22 respectively.

The SRP035278 dataset [3], from HCT116 cells, was compared to publicly available ENCODE [163, 164] peak data on the same cell line: DNase hypersensitivity (ENCFF001WIJ), H3K27ac (ENCFF001VCO), H3K4me1 (ENCFF001VCN) and H3K4me3 (ENCFF001XBW). The SRR1552484 dataset [34], from K562 cells, were compared to ENCODE's CTCF ChIA-PET data on the same cell line (ENCFF001THV). Gene annotations with associated transcription start and termination sites were collected from the UCSC table browser (Hg19).

B.3 CTCF ChIA-PET network construction

The ENCODE provided file defining a list CTCF chromatin interactions (output type: "long range chromatin interactions") was used for the network analysis. In brief, one base pair overlapping CTCF loci were merged to construct an adjacency matrix, $A_{i,j}$, where i corresponds to some merged loci and $A_{i,j}$ evaluates to one if merged regions i and j contain a chromatin interaction. The network homophily or modularity score sorted by LI component presence is defined by equation B.2 [125].

$$Q = \frac{1}{2m}\sum_{i,j}(A_{i,j} - \frac{k_i k_j}{2m})I(i,j) \tag{B.2}$$

In equation B.2, $I(i,j)$ evaluates to one if locus i and j both contain or lack a RNAP loading event prediction, otherwise $I(i,j)$ evaluates to zero. k_i refers to the number of chromatin interactions involving locus i and $m = \sum_i k_i$. Measures of node centrality (degree, eigenvector and closeness) were computed using the python package Networkx [65].

B.4 Software Package: Tfit

Transcription Fit (Tfit) is a freely available, open source software package written in the C/C++ program language that requires GNU compilers 4.7.3 or greater. Tfit uses the popular openMPI framework to perform massive parallelization via multithreading on multiple core, single node systems or multiple core, multiple node compute clusters. Users are asked to provide nascent transcription data in the bed graph file format. Users may (optionally) specify parameters such as the number of random seeds to EM, the expected error in read mapping (p) and model complexity bounds ($\mathbf{M}$).

B.5 Numerical confirmation of model inference by simulation

We present a probabilistic generative model of RNAP that provides a straightforward method to draw a collection of random samples $\mathbf{D}$ from $p(g;\Theta)$. As shown in Fig. 1, our model governing RNAP location is generative. The two distinct stages of RNAP, Loading/Initiation (LI) and Elongation/Termination (ET), are illustrated by 250 random draws from our mixture model. Qualitatively, our simulated data resembles GRO-seq read pile up seen in previous studies [3, 102, 105, 92].

Using data generated directly from our model, we first test the accuracy and correctness of our parameter estimation procedure. To study the robustness of our parameter estimation methodology, we drew varying sizes of $\mathbf{D}$ from 10 to 500 data points at equally spaced intervals of 20. Under each sample size, we collect 25 unique subsets from $p(g|\Theta)$ and compute the standard error $|\Theta - \hat{\Theta}|$. As shown in SFig. 1A, we observe a significant and fast decrease in standard error as $\mathbf{D}$ increases in size.

Lastly, to monitor accuracy in model selection by Bayesian Information Criteria (BIC), we drew from $p(g|\Theta)$ under varying levels of model complexity, $|\mathbb{K}| \in \{1..20\}$ (SFig. 1B). To quantify variability in accuracy, we drew 25 sets of $\mathbf{D}$ from $p(g|\Theta)$ where $|\mathbf{D}| = 100$. Under each set, we computed MLE estimates for every model topology between 1 to 20 and select the model with the minimum BIC score. A maximum of 20 is chosen arbitrarily and for computational convenience. We note that users can modify this value as necessary. We observe that our model selection procedure correlates well with the true $|\mathbb{K}|$.

Table B.1: **Genome wide summary statistics of inferred $\hat{\Theta}$.** Computations of $\mu - TSS$ and l_s- 3‘end are specific to transcribed regions overlapping a single isoform gene. $\hat{w_p}$ was recomputed for a 2KB window surrounding $\hat{\mu_k}$ as $|l_s - \mu_k|$ will artificially correlate with w_p.

	Mean	Median	Standard Deviation
$\lvert\mathbb{K}_p\rvert$	2.8	2.1	±1.5
$\mu - TSS$	-42	-15	±120
l_s- 3‘end	6325	6194	±1200
σ	38.84	21.46	±65.73
λ	170.15	137.57	±130.15
π	0.51	0.51	±0.29
w_p	0.73	0.79	±0.13

Table B.2: **Measures of node centrality: Degree, Eigenvector and Closeness**. Nodes were grouped according to one base pair overlap with either a Ref-Seq TSS annotation or a bidirectional transcript (Bidir.) classification. Note: ∼ indicates nodes lacking a TSS or Bidir. one base pair overlap. The mean centrality measure (with associated standard deviation) across all node labeling types is reported in each cell. Bold lettering indicates a significant increase (p-value $< 10^{-}5$) in mean relative to ∼ TSS and ∼ Bidir. as assessed by a bootstrapped permutation of node labeling to provide an empirical distribution over each centrality measure.

Centrality	∼TSS	TSS	TSS	∼TSS
Measure	∼Bidir.	∼Bidir.	Bidir	Bidir
Degree	1.81±0.3	2.01±0.6	2.20±1.1	**2.44±1.6**
Eigenvector	5.31±10.1	7.77±12.3	8.06±30.1	**43.53±80.7**
Closeness	3.89±2.12	3.19±5.12	**5.14±1.12**	**5.81±2.5**

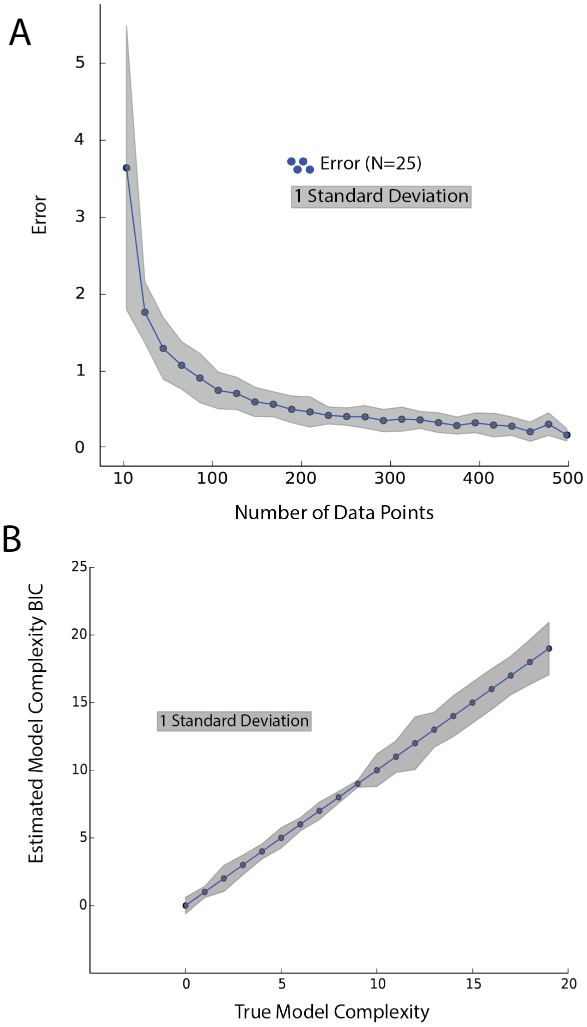


Figure B.1: **Performance on simulated data.** (A) shows the effect of error, defined as $|\Theta - \Theta^*|$, with increasing data size $|\mathbf{D}|$. (B) quantifies the relationship between true and estimated model complexity. Data size of simulation grows linearly with model size.

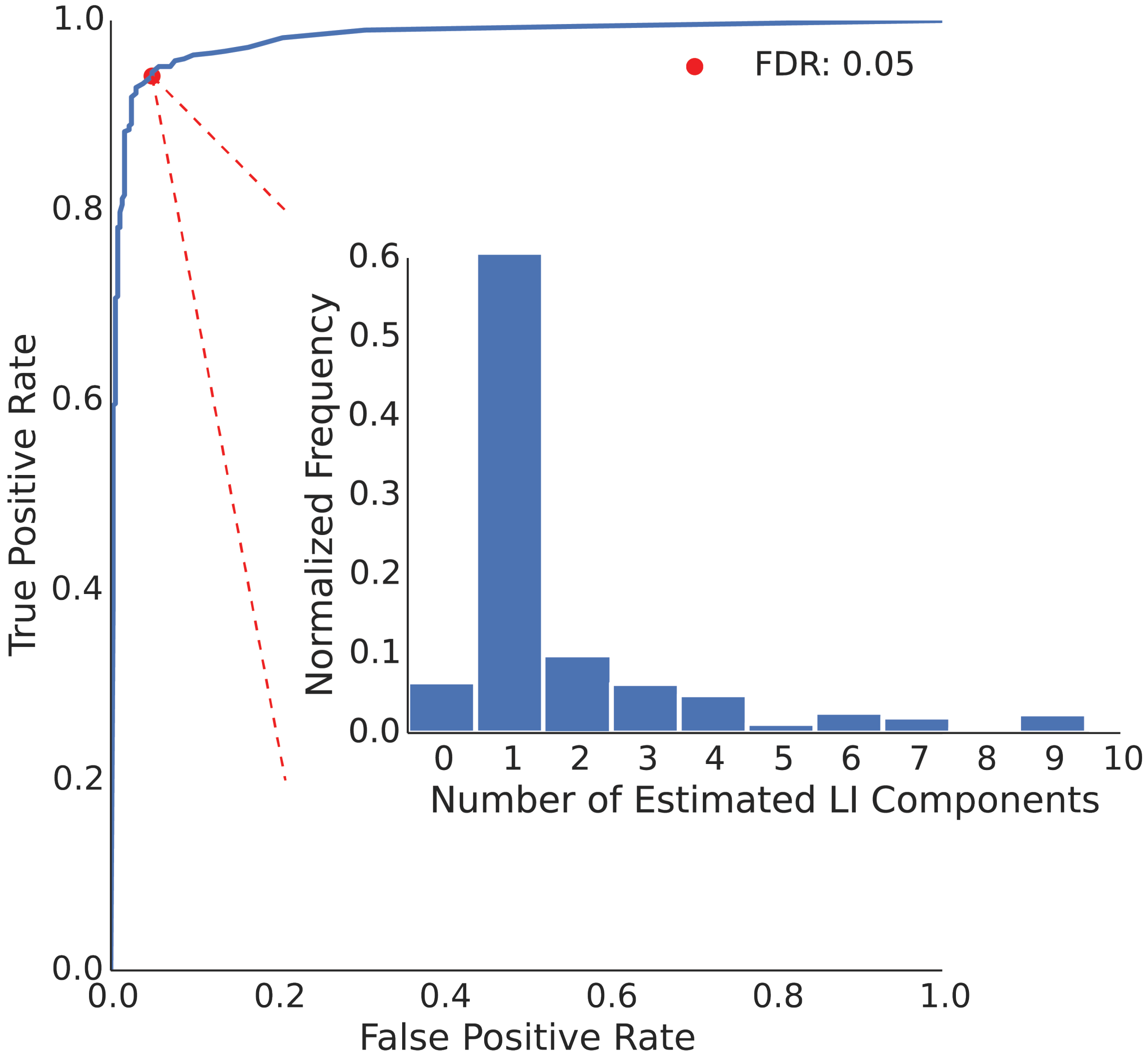


Figure B.2: **Distribution of model complexity at transcribed single isoform genes** To perform ROC analysis, we consider a true positive to be an RNAP mixture model fit where $|\mathbb{K}_p|$ exceeds zero over FStitch overlapping single isoform genes. We consider a false positive to be an RNAP mixture model fit where $|\mathbb{K}_p|$ exceeds zero over FStitch-defined transcription inactive regions of the genome. We vary the Bayesian Information Criteria penalty γ to assess the relationship between true and false positive rates. At a false positive rate of 0.05, we present the distribution of $|\mathbb{K}_p|$ at single isoform genes. We observe that the mode of this distribution lies prominently at one.

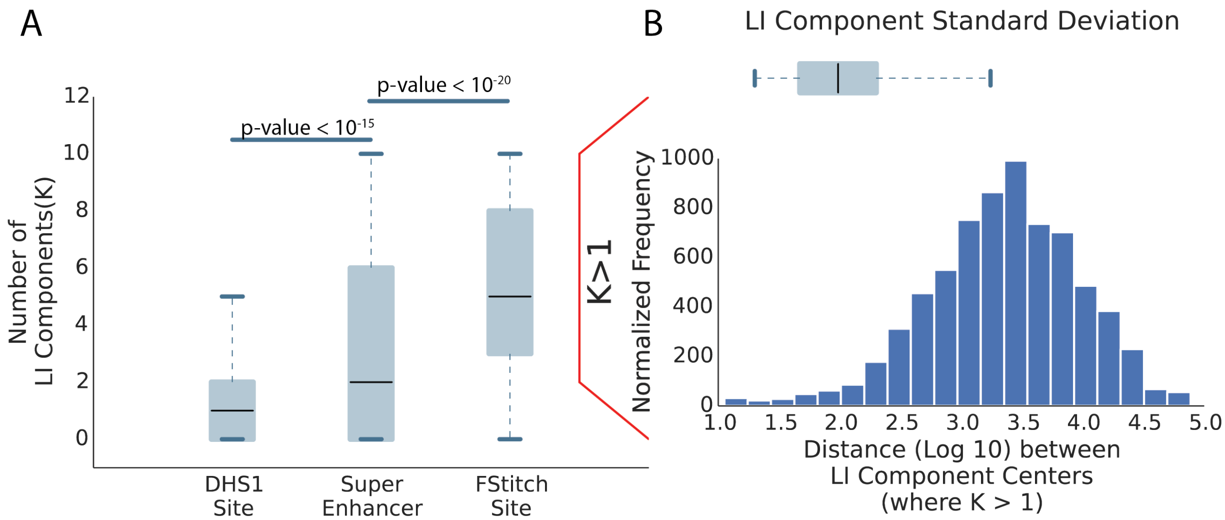


Figure B.3: **Distribution of model complexity at various regulatory loci and pairwise distances between LI component centers** (A) We performed inference in RNAP LI component locations over three types of regulatory loci: DNase I Hypersensitivity sites (DHS1) profiled by ENCODE [164], Super Enhancers profiled by dbSuper [84] and nascent transcriptional regions profiled by FStitch [11]. (B) At loci where $|\mathbb{K}_p|$ exceeds one, we asked for the pairwise distances between all LI component centers $|\mu_i - \mu_j|$ where $i \neq j$. Finally, we place this pairwise distance histogram in reference to distribution of LI component standard deviations defined by $\sqrt{\sigma^2 + 1/\lambda^2}$.

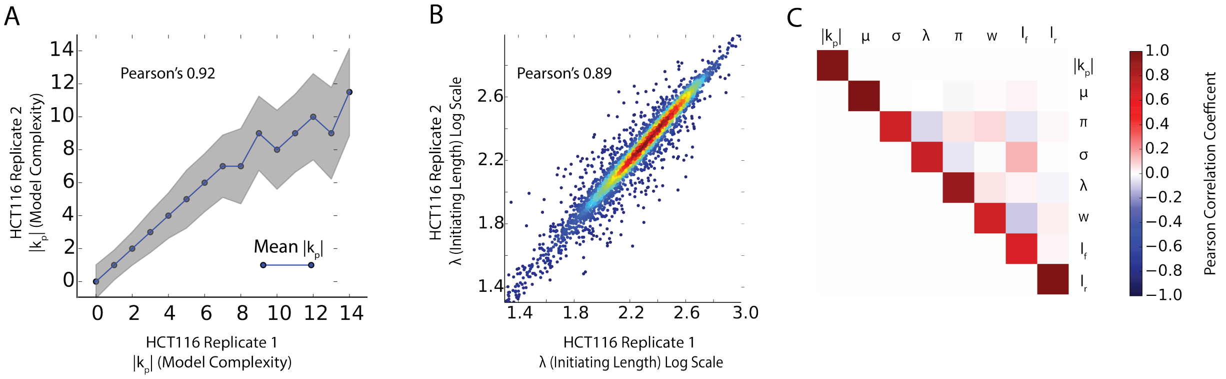


Figure B.4: **RNAP inference procedure computes consistent parameter estimates across biological replicates.** (A) quantifies the number of LI components predicted across each transcription unit as a histogram. Blue line and scatter indicate the mean $|\mathbb{K}_p|$ of replicate 2 conditioned on replicate 1. Grey shading incates one standard deviation (B-C) shows the correlation between biological replicates of model complexity prediction and initiating parameter $\hat{\lambda}$ estimation respectively. (D) shows all pairwise correlations between inferred mixture model inference between biological replicates. Note, that Θ_{rep1} and Θ_{rep2} were ordered by μ_k and this θ_{rep1} and θ_{rep2} were compared only if they shared equivalent rank.

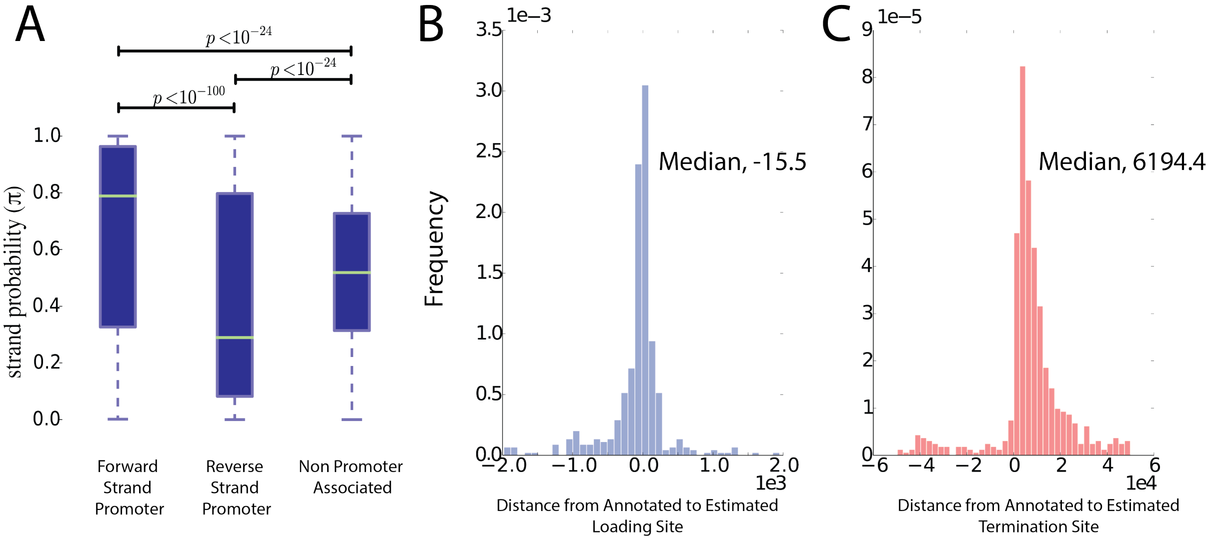


Figure B.5: **Model predictions track closely with known estimates of gene annotations.** (A) shows the distribution of estimated strand bias $\hat{\pi}$ from LI classifications grouped by forward, reverse strand or no annotated gene loci. Under a normality assumption of $\hat{\pi}$, sample averages of $\hat{\pi}$ grouped by RefSeq annotation were computed via a t-test. (B) shows the distance between $\hat{\mu} - TSS$ defined by RefSeq gene annotations. Importantly, this analysis was only computed at single isoform genes such that a one-to-one correspondence between μ and the annotated TSS is expected. (C) shows the distance $\hat{L}-$ 3-prime termination site defined by RefSeq gene annotations.

Appendix C

Supplementary Material to Chapter 4

C.1 eRNA origins

In prior work[9] we leveraged the known behavior of RNA polymerase II to identify individual transcripts within nascent transcription data. Although our model[9], known as Transcription fit (Tfit), does not implicitly assume polymerase initiation will be bidirectional, we observed bidirectional transcription at both promoters[48] and enhancers[38]. Whether bidirectional (2 transcripts) or unidirectional (1 transcript), our model precisely infers the point of RNA polymerase loading, e.g. the origin point of transcription. Formally, this origin point (μ) represents the expected value of a Gaussian (Normal) random variable discussed in great detail in our previous publication[9] and later within this supplement.

As reported previously[9], we observed that super enhancers[84] and regions annotated as transcribed [4, 11, 27, 38] often contain multiple origins. For example, as shown in Extended Data Figure C.1, the entirety of this super enhancer annotated region is transcribed and identified as a single region by most nascent transcription analysis algorithms[4, 11, 27, 38]. However, since Tfit looks for site of initiation rather than transcribed regions, our model identifies three origins of transcription in 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 an eRNA origin, and the resulting transcripts (in this case six) as individual eRNAs.

C.2 Genomic Feature Data Integration

We examine the relationship between genomic features such as TF-binding peaks, chromatin modifications, DNA sequence, TF-binding motifs and enhancer RNA presence. Frequently, we compare two (or more) datasets 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 their feature is within 1500 base pairs of each other. For example, let some TF-binding peak be located on chromosome 1 with a start coordinate of 10000 and stop coordinate of 10405 and an eRNA origin at chromosome 1 with a start coordinate of 10200 and stop coordinate of 10201. 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 1500 base pair cutoff is justified in Extended Data Fig C.2. Furthermore, all genomic coordinates refer to hg19 or mm10 for human or mouse datasets, respectively.

C.3 Nascent transcription data processing

SRA files were downloaded from the NCBI Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`). Accession numbers are provided in Supplementary Table C.5. The SRA files were converted into fastq format using fastq-dump 2.3.2-5 in the SRA Toolkit[148]. Studies utilizing a second strand synthesis kit were reverse complemented using fastx-reverse-complement[68] -Q33. Human and mouse fastx files were mapped to the hg19 and mm10 genomes, respectively, using bowtie2[94] version 2.0.2 -very-sensitive. The resulting sam files were converted to sorted bam files using samtools[100] 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[139] 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[83]. The hg19.chrome.sizes and mm10.chrome.sizes input files were made using fetchChromSizes[162] from UCSC and the hg19 and mm10 genome files, respectively.

C.4 Tfit parameters and bidirectional prediction

Transcription fit (Tfit) is a finite mixture model that utilizes a model of RNA polymerase II behavior to identify and characterize all transcripts in nascent transcription data[9]. Here we modify our previous approach in two ways. First, to rapidly identify all sites of transcription initiation genome-wide, we compute a likelihood ratio statistic between a fully specified exponentially-modified Gaussian (equation C.1, the loading/initiation/pausing phase of our earlier Tfit model[9]) against a Uniform distribution background model (equation C.2) at some genome interval $[a, b]$. We hereafter refer to this approach as Template Matching. Second, we amend our earlier estimate of the loading step of polymerase activity to permit variable distances between the forward and reverse strand transcripts, hereafter referred to as a polymerase footprint. For completeness, we now describe both modifications in full detail.

C.4.1 Template Matching

The loading/initiation/pausing portion of our earlier model, fully specified in [9], describes the initial activity of RNA polymerase II (RNAP) and captures initiating transcription, which is often bidirectional, genome-wide. Briefly, our model assumes RNAP is first recruited and binds to some genomic coordinate X as a Gaussian distributed random variable with parameters μ, σ^2 where μ might represent the typical loading position (e.g. origin of any resulting transcript either TSS or enhancer locus) and σ^2 the amount of error in recruitment to μ. Upon recruitment, RNAP selects and binds to either the forward or reverse strand, which we characterize as a Bernoulli random variable S with parameter π. Following loading and pre-initiation, RNAP immediately escapes the promoter and transcribes a short distance, Y. We assume that the initiation distance, is distributed as an exponential random variable with rate parameter λ. In this way, the final genomic position Z of RNAP is a sum of two independent random variables (X + SY) where the density function (resulting from the convolution/cross-correlation) is given in equation C.1. Note that, in keeping with traditional notation, we let upper case, non-greek alphabet letters represent random variables and the associated lower case refer to instances or observations of the stochastic process.

$$\begin{aligned} h(z,s;\mu,\sigma,\lambda,\pi) &= \lambda\phi(\frac{z-\mu}{\sigma})R(\lambda\sigma - s\frac{z-\mu}{\sigma})\mathbb{1}(s) \\ \mathbb{1}(s) &= \begin{cases} \pi & : s=+1 \\ 1-\pi & : s=-1 \end{cases} \end{aligned} \tag{C.1}$$

Above, $R(.)$ refers to the Mill's ratio and $\phi(.)$ refers to the standard normal density function.

In contrast, reads obtained outside of initiation regions are captured by a Uniform distribution (equation C.2).

$$u(z;a,b) = \frac{\hat{\pi}}{b-a} \tag{C.2}$$

Where π refers to the maximum likelihood estimator for the strand bias (equation C.3).

$$\hat{\pi} = \sum_{i=1}^{N} I(s_i > 0)/N \tag{C.3}$$

Finally, the (log-)likelihood of the exponentially modified Gaussian (LL_{emg}) and Uniform (L_u) distribution computed at a genomic interval $[a,b]$ using aligned read counts is given in equation C.4.

$$\begin{aligned} LL_{emg} &= \sum_{i=a}^{b} \log h(z_i, s_i; \hat{\mu}, \hat{\sigma}, \hat{1/\lambda}, \hat{\pi}) \\ LL_u &= \sum_{i=a}^{b} I(s_i > 0)\log\frac{\hat{\pi}}{b-a} + I(s_i < 0)\log\frac{1-\hat{\pi}}{b-a} \\ LLR &= LL_{emg} - LL_u \end{aligned} \tag{C.4}$$

Here, $\hat{\mu}$ refers to the center of the window. Based on our previous study[9], we set $\{\hat{\sigma}, \hat{1/\lambda}, \hat{w}, \hat{\pi}\} = \{34.2, 391.7, 0.358, 0.501\}$.

The algorithm is a simple sliding window of LLR computations. Overlapping (1 base pair) regions of interest ($LLR > \tau$) are merged. In every study profiled for bidirectional transcription by Tfit, $\tau = 10^3$. More information on running and using Tfit output is available at `github.com/azofeifa/Tfit`.

C.4.2 EM Algorithm and Bidirectional Origin estimation

On its own however, the template matching module of Tfit does not provide an exact estimate over Θ (the parameters associated with a single loading position). To perform optimization over Θ and specifically μ (the origin of bidirectional transcription), we derived the Expectation Maximization algorithm (outlined in detail in our previous publication[9]) to optimize the likelihood function of equation C.4. In brief, we used the following EM-specific parameters at each loci: the number of random re-initializations per loci was set to 64,the threshold at which the EM was said to converge, $|ll_t - ll_{t+1}|$, was set to 10^{-5}. Finally for computational tractability, the EM algorithm halted after maximum of 5000 iterations.

At each window predicted by the sliding window algorithm, we perform inference over μ, σ, λ,and π by the EM algorithm. Details of the derivation, model selection and algorithm design can be found in our previous report [9].

C.4.3 Footprint Estimation

Importantly, our previous effort at parameter estimation of the finite mixture model assumed that RNA polymerase II behaved as a point source[9]. Consequently, we could not incorporate a systematic approach to estimate observed gaps between the forward and reverse strand peaks which deviate more than could be explained by an exponentially-modified Gaussian density function. Here, we amend our model only slightly to estimate this behavior. We call the distance between the forward and reverse strand peaks, the *footprint* of RNA polymerase II or fp. In brief, fp amounts to adding or removing a constant to z_i. Assuming that $fp > 0$ then the above equations remain valid by a simple transformation to z_i.

$$z_i := z_i - s_i \cdot fp$$

As in our previous effort[9], we insert this new parameter into the conditional expectation of the latent variables given the observed random variables and perform a gradient step. This allows us to optimize for fp (equation C.5).

$$\hat{fp}_k := \frac{1}{r_k}\sum_{i=1}^{N}(s_i(z_i-\mu) - E[Y|z_i,s_i;\theta^g])\cdot r_i^k \tag{C.5}$$

The interested reader should refer to our previous paper [9] where derivation of the EM algorithm and fitting the Tfit model are discussed heavily. For complete clarity, the full expression of the expectation operators is given below.

$$\begin{aligned} E[Y|g_i;\theta^t] &= s_i(z-\mu) - \lambda\sigma^2 \quad + \frac{\sigma}{R(\lambda\sigma - s_i(z_i-\mu)/\ \sigma)} \\ r_i^k &= p(k|g_i;\theta_k^g) = \frac{w_k\cdot p(g_i;\theta_k^g)}{\sum_{k\in\mathbf{K}} w_k\cdot p(g_i;\theta_k^g)} \\ r_k &= \sum_{i=1}^{N} r_i^k \end{aligned} \tag{C.6}$$

C.5 Computation of Bimodality, ΔBIC

To assess whether the distribution of ChIP peaks or TF-binding motifs around an eRNA origin is bimodal, we developed and employed a pairwise distribution test. We define the ΔBIC score (in equation C.8) to be the difference in BIC scores between a single Laplace-Uniform mixture centered at zero (unimodal) and a two component Laplace-Uniform mixture with displacement away from 0, i.e. c (bimodal). The density function of a Laplace distribution with parameters (c, b) is provided in equation C.7 and we use the formulation for the Uniform distribution of equation C.2.

$$p(d;c,b) = \frac{1}{2b}\exp -\frac{|d-c|}{b} \tag{C.7}$$

Here D refers to the set of distances, $d_i \in [-1500, 1500]$, either the center of the TF-binding peaks obtained from MACS or the center of TF-binding motifs from PSSM scanner relative to eRNA origin. If $\Delta\text{BIC} \gg 0$, we assume bimodality in TF peak location relative to the eRNA origin.

$$\begin{aligned}
\mathcal{L}_0(D;\Theta^*) &= \prod_{i=1}^{N} \frac{1}{3000} \\
\mathcal{L}_1(D;\Theta^*) &= \prod_{i=1}^{N} w\frac{1}{2b}\exp\{-\frac{|d_i|}{b}\} + \frac{1-w}{3000} \\
\mathcal{L}_2(D;\Theta^*) &= \prod_{i=1}^{N} \frac{w}{4b}\exp\{-\frac{|d_i - c|}{b}\} + \frac{w}{4b}\exp\{-\frac{|d_i + \mu|}{b}\} + \frac{1-w}{3000}
\end{aligned} \tag{C.8}$$

$$\Delta\text{BIC} := -2(\log\mathcal{L}(D)_1 - \log\mathcal{L}(D)_2) + k\log(|D|)$$

Θ^* is optimized again by the Expectation Maximization algorithm where the update rules are given in equation C.9.

$$\begin{aligned}
d_{t+1} &= \frac{1}{2(r^a + r^b)}(\sum_{i=1}^{n} r_i^a d_i + \sum_{i=1}^{n} r_i^a d_i) \\
b_{t+1} &= \frac{1}{2(r^a + r^b)}(\sum_{i=1}^{n} r_i^a |d_i| + \sum_{i=1}^{n} r_i^b |d_i|) \\
w_{t+1} &= \frac{r^a + r^b}{r} \\
r_i^a &= \frac{p(d_i; c, b)}{p(d_i; c, b) + p(d_i; -c, b) + u(d_i; -1500, 1500)} \\
r_i^b &= \frac{p(d_i; -c, b)}{p(d_i; c, b) + p(d_i; -c, b) + u(d_i; -1500, 1500)} \\
r_i^u &= 1 - r_i^a + r_i^b \quad r^x = \sum_{i=1}^{N} r_i^x \quad r = r^a + r^b + r^u
\end{aligned} \tag{C.9}$$

We refer to a signal as bimodal when $\Delta\text{BIC} > 500$, estimated from the distribution in Extended Data Fig. C.4D.

C.6 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[89, 90] database of hand-curated transcription factor binding motifs 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 was performed by the algorithm outlined by Staden[155]. False discovery rate (FDR) was quantified by the approach outlined by Storey[156]. Only sequences were the FDR did not exceed 10^{-4} were considered a significant TF-sequence motif, 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 `github.com/azofeifa/MDS`.

C.7 MD-score Hypothesis Testing

C.7.1 The Motif Displacement score

The Motif Displacement score (MD-score) relates the proportion of significant motif sites within some window $2 * h$ divided by the total number of motifs against some larger window $2 * H$ centered at all bidirectional origin events. It is calculated on a per PWM binding model basis.

Let $X_j = \{x_1, x_2, ...\}$ be the set of bidirectional origin locations genome wide for some experiment j. Let $Y_i = \{y_1, y_2, ...\}$ be the set of all significant motif sites for some TF-DNA binding motif model i genome wide, which is static as it only depends on the genome build of interest. Therefore the set of all MD-scores is calculated by equation C.10.

$$\begin{aligned} g(X_j, Y_i; a) &= \sum_{x \in X_j} \sum_{y \in Y_i} \delta(|x - y| < a) \\ md_{j,i} &= g(X_j, Y_i; h)/g(X_j, Y_i; H) \\ md_{j,i} &\in [0, 1) \text{ if } h < H \end{aligned} \tag{C.10}$$

Here $\delta(.)$ is a simple indicator function that returns one if the condition (.) evaluates true otherwise to zero. The double sum, i.e. $g(a)$, is naively $O(|X||Y|)$ however data structures like interval trees reduce time to $O(|X| \log |Y|)$.

To be clear, there exist 641 TF-DNA binding models in the HOCOMOCO database and therefore 641 MD-scores exist for some experiment j. Let md_i be the MD-score computed for some TF-DNA binding motif model. Therefore let $MD_j = \{md_1, md_2, ..., md_{641}\}$ be the vector of all MD-scores for some dataset j.

C.7.2 MD-score significance under stationary model

If y_i and x_i are uniformly distributed throughout the genome, i.e. following a homogeneous Poisson point process, then $g(h)$ is distributed as a binomial distribution with parameters p, N (equation C.11).

$$\begin{aligned} g(h) \sim& \mathrm{B}(n,p) \\ \mathrm{B}(k;n,p) =& \binom{n}{k}(p)^k(1-p)^{n-k} \\ \text{where } n = G(H) &\text{ and } p = h/H \end{aligned} \tag{C.11}$$

In cases where $g(H) \gg 0$, the binomial is well approximated by a Gaussian distribution and hypothesis testing under some α level can proceed in the typical fashion. In brief, significantly increased MD-scores (by a binomial test) is diagnostic of heightened motif frequency surrounding eRNA origins.

C.7.3 MD-score significance under a non-stationary background model

Motifs, however, are not distributed uniformly throughout the genome. Specifically, particular regions, such as gene promoters of the genome are known to exhibit significance sequence bias. Indeed, the localized GC content is highly non-stationary at eRNAs (Extended Data Figure C.7A). Consequently, a binomial test—which assumes a homogeneous Poisson process of motif locations genome wide—may be a too liberal a null model (e.g. the wrong background assumption). In this paper we assume H = 1500 bps.

To control for this non-stationarity, we propose a simulation based method to compute p-values for MD-scores under an empirical CDF, i.e. a localized background model. Let p be a $4x2H$

matrix where each column corresponds to a position from an origin and each row corresponds to a probability distribution over the DNA alphabet $\{A, C, G, T\}$. To be clear, $p_{0,0}$ corresponds to the probability of an A at position -H from any bidirectional origin, similarly $p_{2,1500}$ corresponds to the probability that a G occurs at exactly the point of the bidirectional origin.

Therefore, the simulation based method of the background model is simple. Given an experiment of X_j bidirectional origin locations, we simulate $|X_j|$ sequences following this non-stationary GC content bias. We then iterate over all PWM models and look for significant motif hits, as discussed in Supplemental Note C.6. We then compute summary statistics about the displacement of the motif relative to the set of synthetic sequences, i.e. $MD = \{md_1, md_2, ..., md_{641}\}$. It should be noted that, in this dataset, any motif match is by complete chance alone. We iterate this process 10,000 times to compute a random distribution over md_i, i.e. $\vec{md_i}$, and thus we can assess the probability of our observed (i.e. from real data) md_i relative to our empirically simulated $\vec{md_i}$. Example simulations are shown in Extended Data Figure C.7B.

C.7.4 MD-score significance between experiments

The MD-score constitutes a proportion and as long as h is upper bounded by H then $md_{j,i}$ will always exist within the semi-open interval $[0, 1)$. An important question is whether $md_{j,i}$ has significantly shifted between two experiments j, k as a function of X_j and X_k. This analysis is straightforward under the two proportion z-test. Specifically we are testing the null and alternative hypothesis tests in equation C.12.

$$\begin{aligned} H_0 &: md_{j,i} = md_{k,i} \\ H_1 &: md_{j,i} \neq md_{k,i} \end{aligned} \tag{C.12}$$

We can then compute the pooled sample proportion (p_i) and standard error (SE) as shown in equation C.13. Therefore our test statistic z (equation C.14) is normally distributed with mean 0 and variance 1.

$$\begin{aligned} p_i &= \frac{(md_{j,i} \cdot g(X_j, Y_i; H) + md_{k,i} \cdot g(X_k, Y_i; H))}{g(X_j, Y_i; H) + g(X_k, Y_i; H)} \\ SE &= \frac{p(1-p)}{(1/g(X_j, Y_i; H) + 1/g(X_k, Y_i; H))} \end{aligned} \tag{C.13}$$

$$z = \frac{md_{j,i} - md_{k,i}}{\sqrt{SE}} \sim N(0,1) \tag{C.14}$$

Computation of the p-value can be assessed in the normal fashion under some α level. In all comparisons, we utilize multiple hypothesis correction outlined by Storey[157].

C.8 Cell type and TF enrichment analysis

The sections serves to outline the rational for determining if heightened MD-scores correlate with a specific cell type category. More traditional approaches such as a one-way ANOVA test (MD-scores computed from similar cell types are grouped and within group variance is assessed via a F-distribution) will not adequately account for MD-scores with little support (i.e. motif hits that overlap very few eRNAs). To overcome this, we propose a relatively straightforward method that relies on performing hypothesis testing on all pairwise experimental comparisons.

Let j and k be two nascent transcription datasets of interest, then $mds_{j,i}$ and $mds_{k,i}$ refer to MD-scores for some TF-motif model (i) for which we can perform hypothesis testing over as outlined in Supplementary Note C.7. If we let α be the threshold at which we consider $mds_{j,i} - mds_{k,i}$ to significantly increase, than we expect on average $\alpha \cdot N - 1$ false positives when considering a single experiment against the rest of the corpus of size N.

Put another, if we let the random variable $S_{j,i}$ refer to the number of times we consider $mds_{j,i} - mds_{k,i}$ to significantly increase in a dataset comparison then $S_{j,i}$ is binomial distributed with parameters $N-1$ and α (equation C.15) assuming that there is not a relationship between the motif model i and the experiment j.

$$S_{j,i} = \sum_{k=1}^{N} \mathbb{I}(p(mds_{j,i} > mds_{k,i}) < \alpha) \tag{C.15}$$

In practice we set α to 10^{-6} and $\mathbb{I}$ refers to an indicator function which returns 1 in the case where the statement evaluates to truth otherwise 0.

Naively, we could now ask for all the datasets annotated as some cell type *ct* and then perform hypothesis testing on S_{ct} (the sum of $S_{j,i}$'s where experiment j belongs to the *ct* cell type set). Importantly, we only consider dataset pairs for which i and j belong to different cell type sets. Unfortunately, a single experiment within the cell type set might show strong association with a TF (i.e. 90% of the $N-1$ comparisons significantly deviate from zero) where the rest of the cell-types show small numbers of significant deviations. By a binomial test, this is unlikely—even when considering the expansion induced by the cell type set—but intuitively does not fit into our notion of cell type association.

To this end, we define a final random variable $A_{ct,i}$ to be the number of times motif i is significantly enriched for a dataset j and that dataset j belongs to some cell type (equation C.16).

$$A = \sum_{j=1}^{N} p(S_{j,i} > S) < 10^{-6}\mathbb{I}(j \in CT) \tag{C.16}$$

where CT refers to the set of experiments that are annotated as cell type *ct*. From there it is easy to assess A across cell types and motifs under a contingency model using Fisher's exact test.

C.9 Associated File Types

We provide here three important sets of tables: 1) a folder of Tfit predicted eRNA origins for every publicly available GRO-seq dataset to date (771); 2) the genomic locations of 641 significant TF-binding motifs (HOCOMOCO) and 3) a histogram of motif locations surrounding eRNA origins for each of the 771 nascent transcript datasets and 641 motif models. These files types are discussed in the Supplementary Data Tables **??** for Tfit, TF-binding motifs and motif displacements histograms respectively.

C.10 IPython Notebook

Given the shear size of the data analyzed (771 nascent transcription datasets, 641 TF motif models), we could not explore all possible comparisons. However, 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 is downloadable at `https://github.com/azofeifa/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.

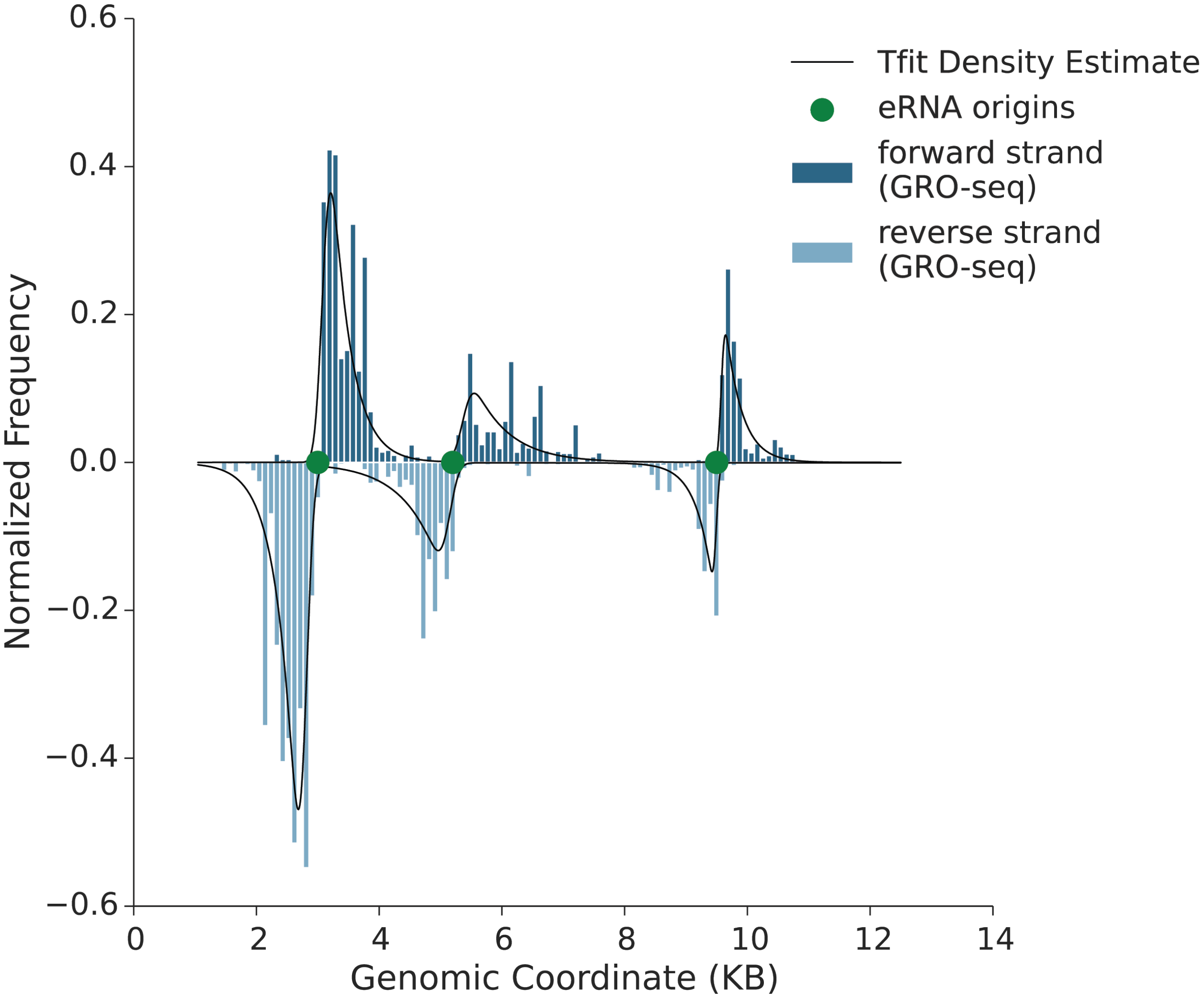


Figure C.1: **Annotated super enhancer region showing RNAP inference.** An annotated super enhancer (starting at chr2:10,456,371), GRO-seq read coverage from an HCT116 dataset**??** and the final inferred density function obtained by Tfit[9]. Via Bayesian model selection, three distinct eRNA origins are identified. Green dots indicate the eRNA origin estimate.

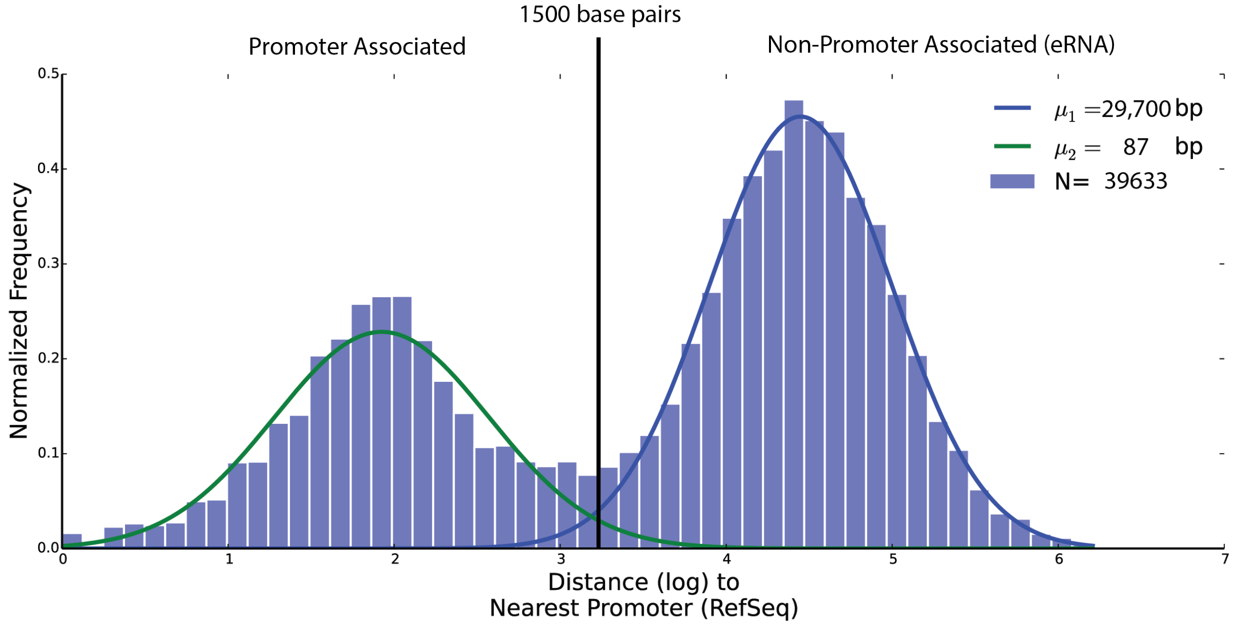


Figure C.2: **Most sites of bidirectional transcription identified by Tfit[9] lack association with promoter regions.** Sites of bidirectional transcription were profiled in a K562 GRO-cap[35] (SRR1552480) experiment using Tfit parameters (Supplementary Note C.4). A promoter is defined as the region associated with an annotated gene's start site using RefSeq release 76. Bimodiality was estimated via a two component Gaussian mixture model fit with the EM algorithm.

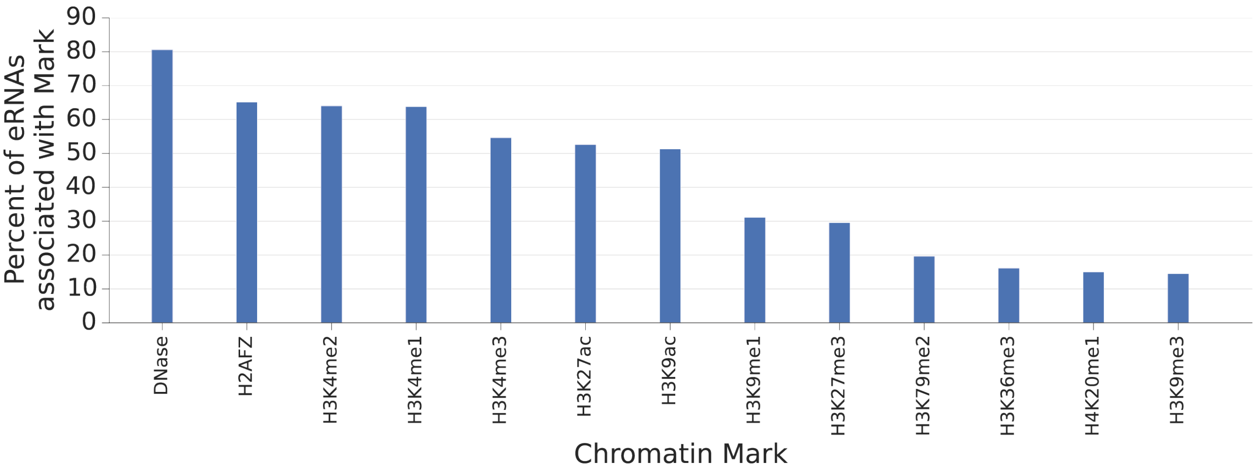


Figure C.3: **Sites of non-promoter associated bidirectional transcription overlap marks of regulatory DNA.** The percentage of eRNAs in a K562 GRO-cap dataset (SRR1552480[35]) that associate with specific chromatin marks. Promoter associated Tfit predictions were removed from this analysis. Chromatin mark peaks were gathered from ENCODE for the K562 cell line (Supplementary TableC.1).

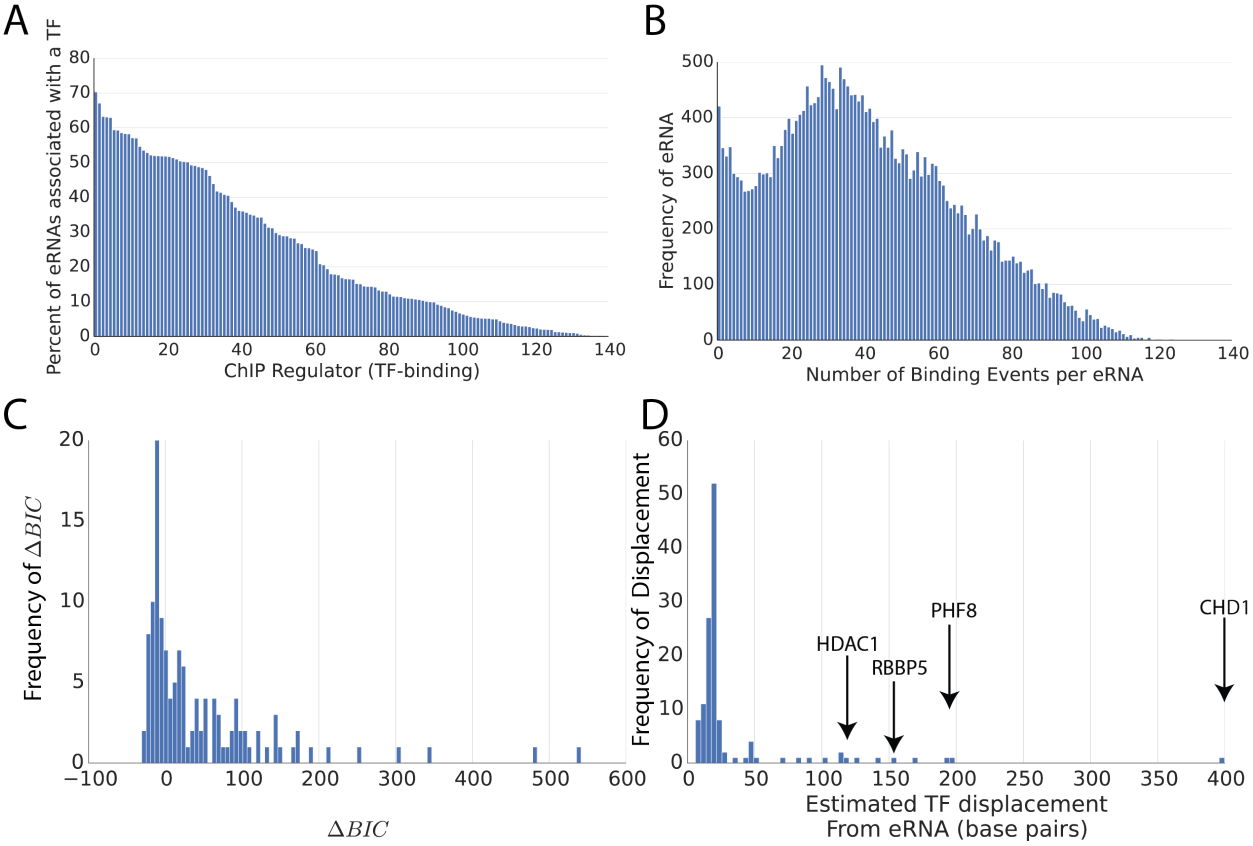


Figure C.4: **Sites of non-promoter associated bidirectional transcription overlap sites of TF-binding.** (A) The proportion of eRNAs associated with a given transcription factor. (B) The number of unique TF-binding peaks occurring at individual eRNAs. (C) TF displacement data was calculated within a 1.5KB radius around eRNA locations and bimodal model selection was performed via a Laplace-Uniform mixture (Supplementary Note C.5). 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[35] (SRR1552480) and ENCODE ChIP-seq peaks (Supplementary TableC.1).

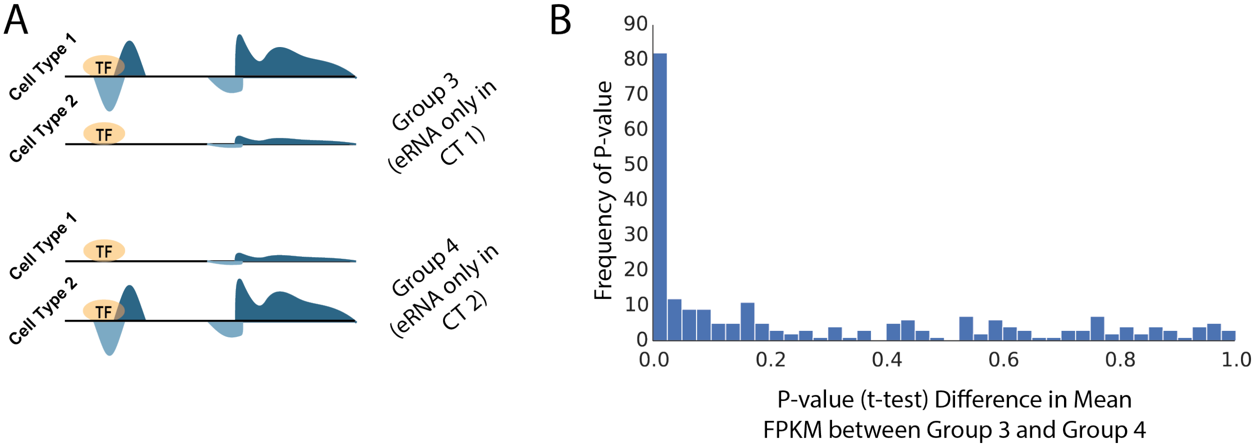


Figure C.5: **TF-binding sites associated with cell type unique eRNAs modulate local gene expression**. (A) A schematic of the analysis. Briefly, TF-binding sites (by ChIP) conserved between two cell types were identified. These (non-promoter associated) genomic loci were further categorized as associated with an eRNA in cell type 1 (CT 1) and lacking an eRNA in cell type 2 (CT 2) or vice versa. Finally, $\log_1 0$ fold chance in FPKM of genes near these sites (< 10 KB) was collected and a two-tailed t-test was used to assess a difference in means. (B) Histogram of p-values following this hypothesis test.

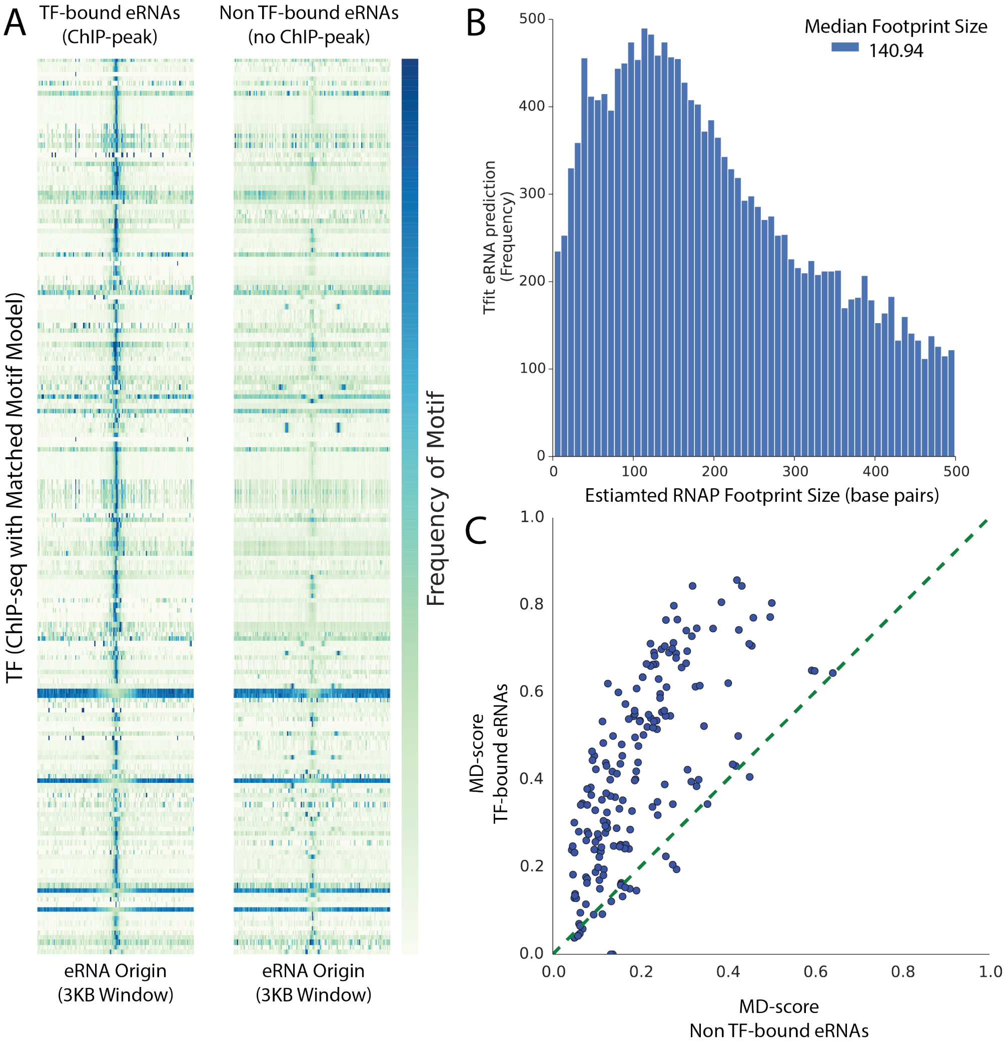


Figure C.6: **TF-binding motifs and eRNA localization reflect TF-binding**. (A) Heatmaps display the frequency of TF-binding motifs centered at the origin eRNAs predicted by Tfit from a K562 GRO-cap[35] (SRR1552480) experiment. eRNAs were further separated by association with or distal to a TF-binding peak. Motifs were gathered from the HOCOMOCO database[89] of hand curated PSSM models. PSSM- and ChIP-matched datasets 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). Refer to Supplementary Note C.4.3 for discussion on the maximum likelihood estimation of RNAP footprint. (C) The co-association of motif with eRNA origin is elevated at bound sites. X-axis: the MD-score computed from eRNAs that are not bound, y-axis: the MD-score computed from TF-bound eRNAs.

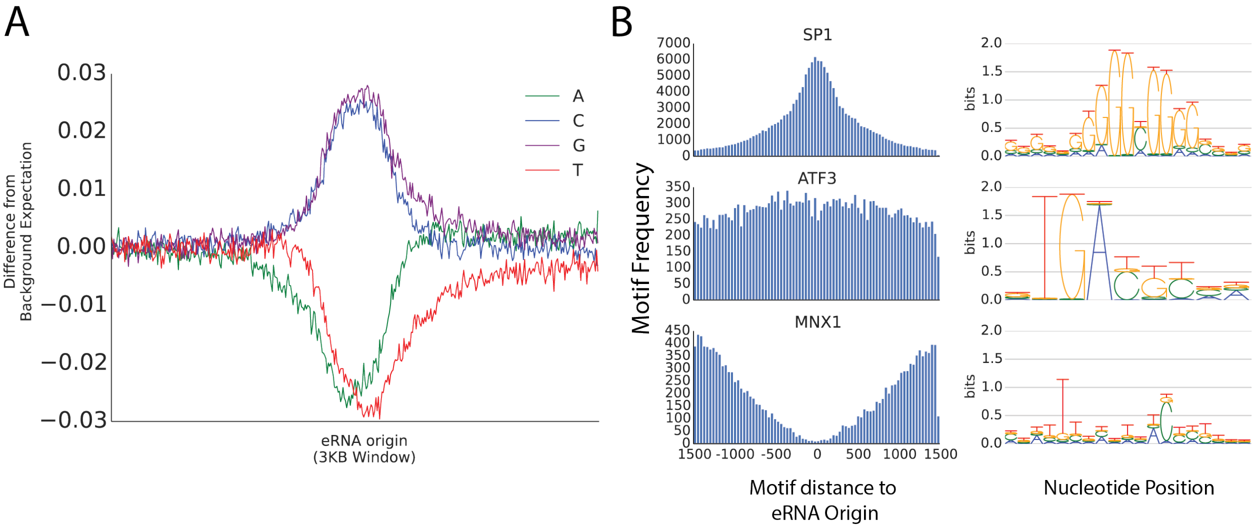


Figure C.7: **eRNA GC content bias.** eRNAs were predicted by Tfit from a K562 GRO-cap[35] (SRR1552480) experiment and sequence from the hg19 human genome build was collected within a 3KB window centered at eRNA origins. (A) Background expectation ACGT was computed from the entire hg19 genome yielding 24.19%, 25.72%, 24.31%, 25.76% for A,C,T,G nucleotides respectively. (B) 10^9 3KB sequences were simulated from the empirical ACGT frequency. Column panels show the distribution of the location of significant PSSM matches $< 10^{-7}$ within simulated data for three demonstrative transcription factors: SP1, ATF3 and MNX1. Adjacent to each motif distribution, the associated PSSM in terms of information content (bits).

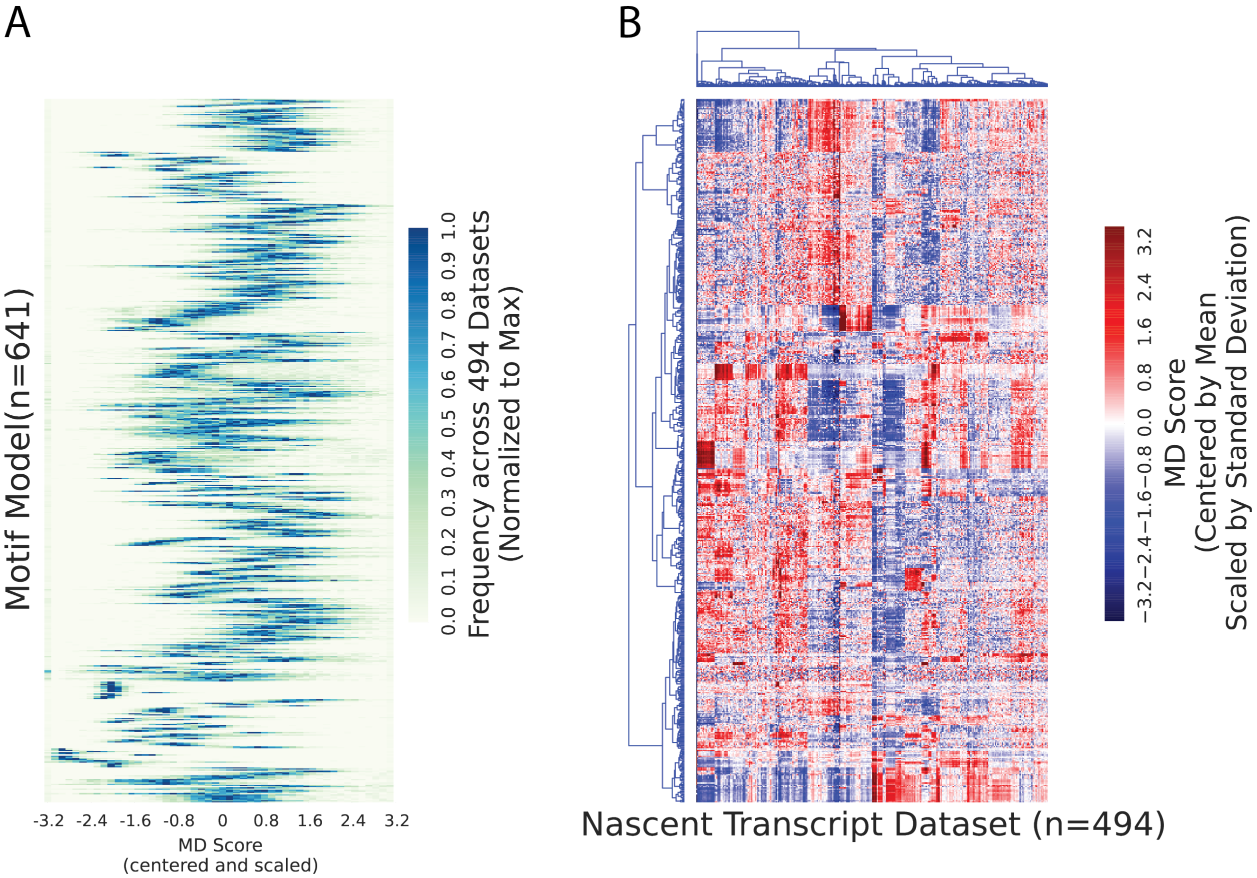


Figure C.8: **MD-scores display wide variability across all publicly available nascent transcript datasets.** Sites of bidirectional transcription were profiled by Tfit across all nascent transcript datasets allowing computation of the 641 (HOCOMOCO) MD-scores. (A) Each row is a motif model and each column represents the frequency of that MD-score across all nascent transcript datasets. For display and comparison, MD-scores were centered by the mean and scaled by the standard deviation. (B) Each row represents a motif model and each column represents an experiment. Heat indicates higher MD-scores (relative to the mean). Rows and Columns were separately sorted by hierarchical clustering.

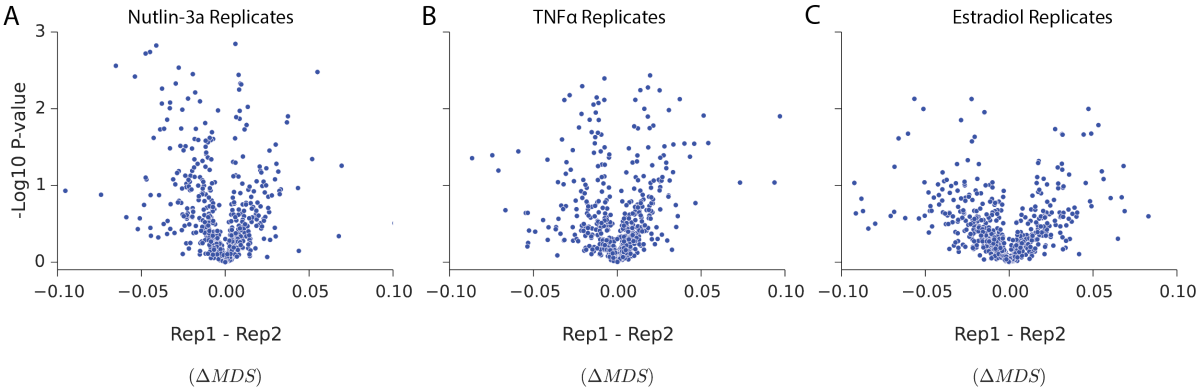


Figure C.9: **No significant differences in MD-scores between biological replicates.** Experiments annotated as biological replicate pairs: (SRR1105738, SRR1105739), (SRR1015589, SRR1015590), (SRR653425, SRR653426) were used to study differences in MD-scores for the Nutlin-3a[3], TNFα[106] and estradiol[67], respectively. Y-axis indicates the negative $\log_{10}$ p-value (two-tailed proportion test) in MD-score change. x-axis provides the change in MD-score.

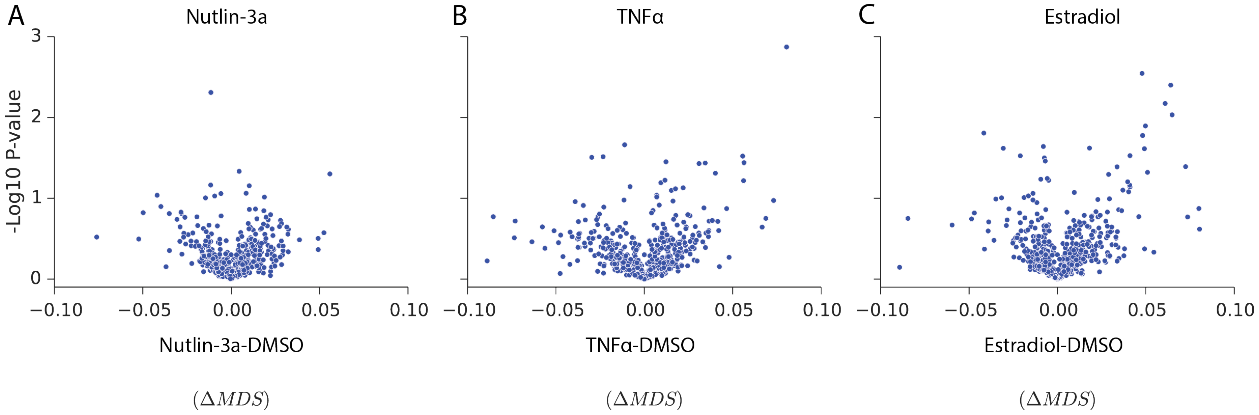


Figure C.10: **No significant differences in MD-scores when considering only promoter associated bidirectional transcripts**. Experiments annotated as treatment/control pairs: (SRR1105737, SRR1105739), (SRR1015583, SRR1015589), (SRR653421, SRR653425) were used to study differences in MD-scores following treatment with Nutlin-3a, TNFα and estradiol for the Nutlin-3a[3], TNFα[106] and estradiol[67] respectively. MD-scores were computed over promoter associated bidirectional transcripts. Y-axis indicates a the negative $\log_{10}$ p-value (two-tailed proportion test) in MD-score change (x-axis).

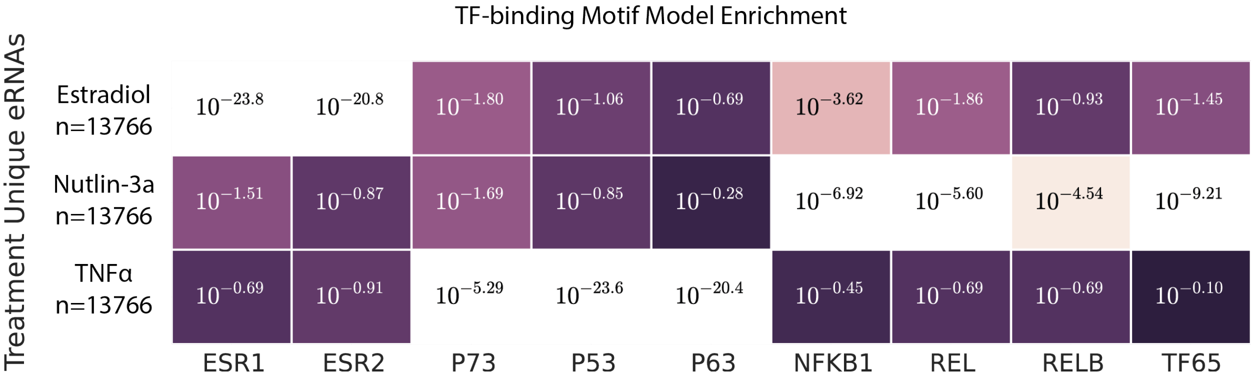


Figure C.11: **Treatment-unique eRNAs are enriched for specific TF-binding motifs**. Treatment-specific eRNAs were isolated from experiments annotated as treatment/control pairs: (SRR1105738, SRR1105739), (SRR1015589, SRR1015590), (SRR653425, SRR653426) for Nutlin-3a[3], TNFα[106] and estradiol[67] respectively. A treatment-unique eRNA is considered if the eRNA origin is not within 1500 base-pairs of control-present eRNA. A treatment-unique eRNA is considered associated with a motif if the motif center is within 150 base pairs of the eRNA origin. Significance of motif over representation is assessed via a one-tailed hypergeometric distribution. Cell color is proportional to $\log_{10}$ of the p-value where ligher color indicates greater significance.

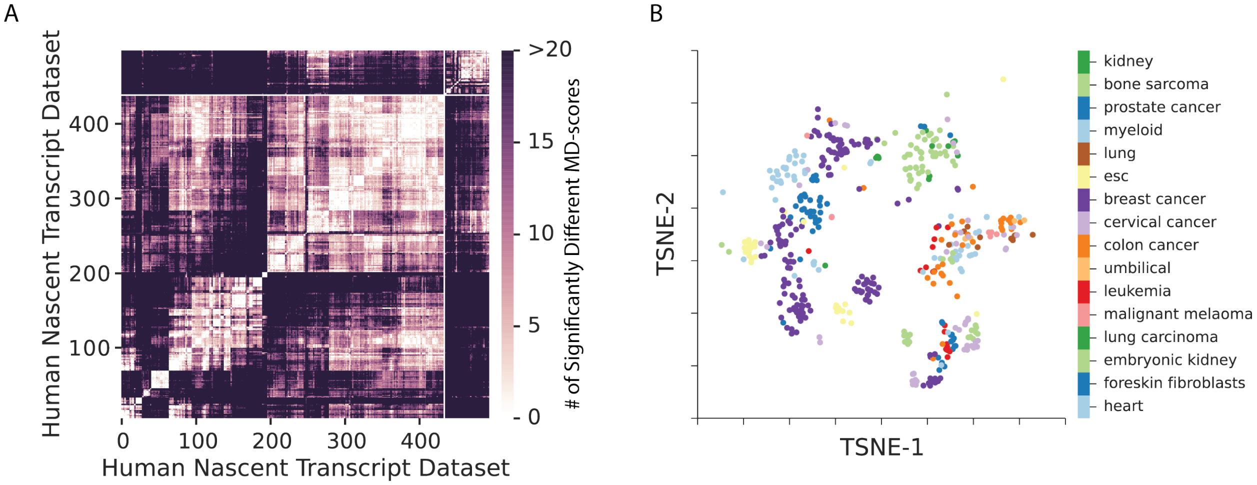


Figure C.12: **Cell type influences MD-score similarity.** MD-scores were computed for all human nascent transcript datasets. (A) Distance matrix where each cell's heat is proportional to the number of significantly different MD-scores ($p(\Delta MDS \neq 0) < 10^{-6}$). Rows and Columns are sorted by Ward hierarchical clustering via euclidean distance metric. (B) Dimensionality reduction by t-Distributed Stochastic Neighbor Embedding (TSNE) of the distance matrix in panel A. Therefore, a point represents a dataset. Color is by publication annotated cell-type.

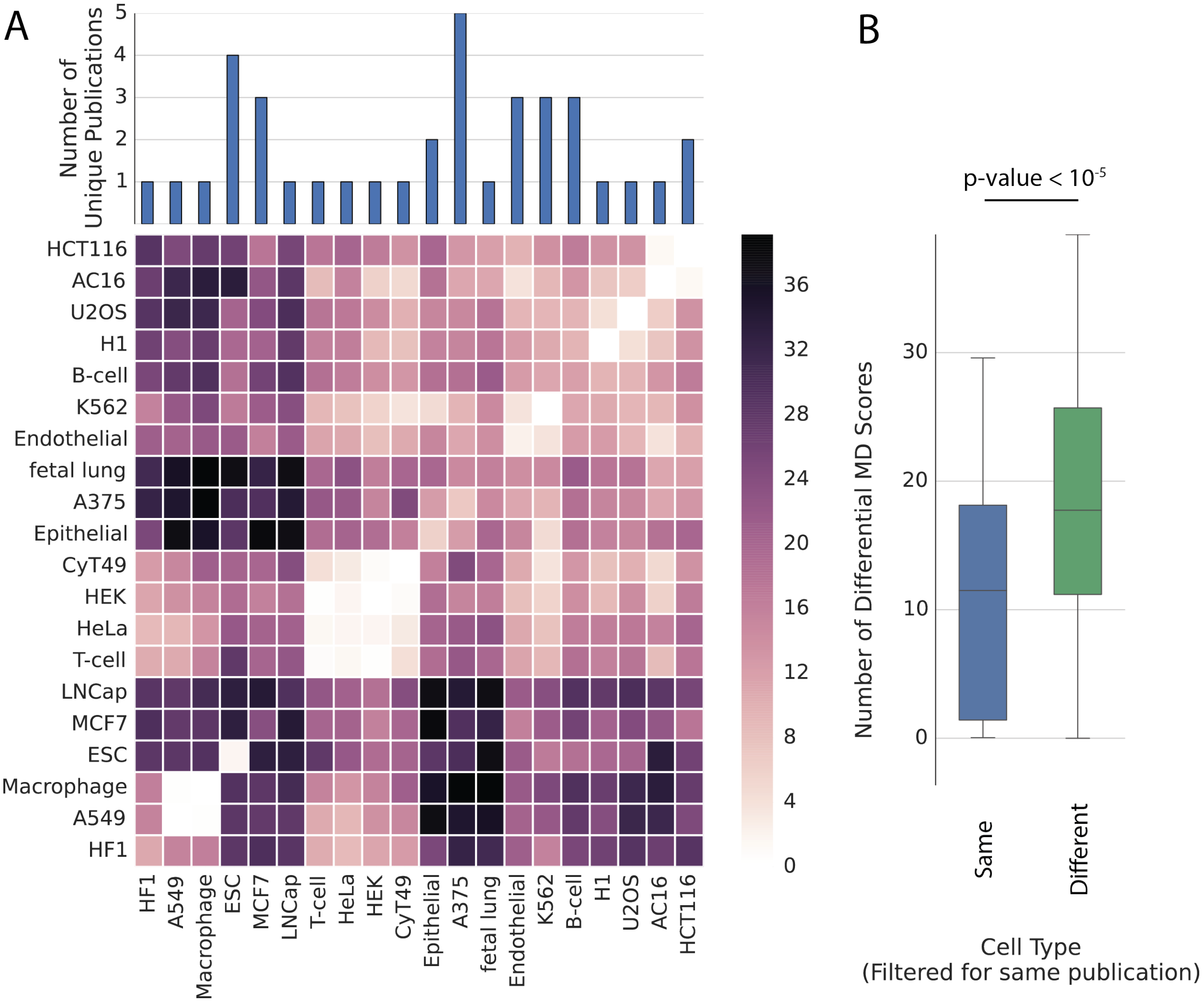


Figure C.13: **Fewer MD-scores are significantly different within similar than across different cell types.** Each untreated nascent transcript dataset ($n = 158$) was independently compared and assessed for significant changes in MD-scores (possible comparisons = 12403). (A) each cell indicates the 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 datasets were of different publications.

Supplementary Table C.1: **ENCODE identifiers for K562 TF-binding peak and chromatin modification peak data utilized**

Each line corresponds to a distinct ChIP-seq dataset profiled by the ENCODE consortium in K562 cells. Within this comma separated file there are three columns: ENCODE_ID, common_name, peak_total. *ENCODE_ID* provides the unique ENCODE identifier (e.g. ENCFF001WBD is a DNase MACS peak bed file). *common_name* which is, in general, the antibody used for the ChIP experiment, otherwise DNase; And, *peak_total* which provides the total number of peaks within that experiment. This description corresponds to the Supplementary Table ENCODE_DATA.csv associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al.

Supplementary Table C.2: **Cell Type Invariant TF-binding sites and eRNA profiles**

This description corresponds to the Supplementary Table STable_to_2D_PreservedTFBindingOverlapStats.csv associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al. Each line within this comma separated file corresponds to a distinct comparison made in main Figure 2D: TF-binding peaks profiled by ChIP/MACS in two different cell types and their association with eRNAs. In this way, there are 12 fields: TF Name, ENC1, ENC2, cell-type 1, cell-type 2,ct1 total,ct2 total,Overlap Total,#[00],#[01],#[10],#[11]. *TF Name* is simply the antibody, *ENC1* and *ENC2* refer to the ENCODE ID for the MACS peaks profiled in two different cell types, *cell-type 1* and *cell-type 2* give the name of the cell type. *ct1 total* and *ct2 total* provide the total number of non-promoter associated MACS peak calls. *Overlap total* provides the number of cell type invariant TF-binding peaks. Finally, *#[00]* indicates the number of ct-invariant binding events that lacked a bidirectional transcript in both cases. In following with this logic, *#[01]*, *#[10]*, and *#[11]* provide the number of ct-invariant binding peaks where an eRNA was only in ct2, only in ct1 or in both respectively.

Supplementary Table C.3: **Nascent transcript Data set usage in pairwise comparisons**

The purpose of this table is just to outline the SRA# datasets utlized in motif distribution comparisons within, between pairs or across all datasets. Specific to Figures 4A-C, Group A and Group B refer to control and treatment. Figures 5B-C these refer to separate cell types

Figure Number	Group A	Group B
4A	SRR1105737	SRR1105739
4B	SRR1015583	SRR1015587
4C	SRR653421	SRR653425
4D	SRR935093	SRR935093 (none) SRR935097 (2 minutes) SRR935101 (5 minutes) SRR935105 (12.5 minutes) SRR935109 (25 minutes) SRR935113 (50 minutes)

4E	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)

5A	SRR1145801	SRR1145801 (hESC) SRR1145808 (endoderm) SRR1145815 (primitive gut tube) SRR1145822 (posterior foregut) SRR1145829 (pancreatic endoderm)
5B	SRR1552482 SRR1552483 SRR1552485	SRR1552480 SRR1552481 SRR1554311
5C	SRR639050 SRR014286	SRR1041870 SRR1041871 SRR408117

Supplementary Table C.4: **Cell Type and TF association table**

This description corresponds to the Supplementary Table celltype_tf_association.csv associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al. Each line within this comma separated value file corresponds to a TF/cell-type comparison. This file contains 7 fields: motif,ct,bin_p,bin_n,obs,pvalue,ENR. *motif* and *ct* refer to the motif model and cell type name respectively. *bin_p* refers to the number of times that motif model was differently elevated in a dataset comparison divided by the total number of comparisons (Note: if there are 491 human datasets then there are 491 choose 2 or 120,295 possible comparisons). *bin_n* refers to the number of dataset comparisons associated with that cell type. *pvalue* results from a binomial test. *ENR* refers to the random expectation (bin_p·bin_n) over the actual cell type and motif count.

Supplementary Table C.5: **Conditions, Cell type and SRA Dataset Table**

This description corresponds to the Supplementary Table conditions_table.csv associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al. This study analyzed 771 nascent transcript datasets which span different organisms, cell types, treatments and conditions. To this end, we provide a meta table .csv format where each row corresponds to some nascent transcript dataset. The columns in this table proceed in this order: SRAnumber, organism, tissue, general_celltype, specific_celltype, treatment_code, treated_or_like_treated, repnumber, keyword, exptype, mapped_reads, total_reads, percent_mapped, TSS, bidirectionals. Hopefully evident from the column identifier, SRAnumber provides the unique GEO identifier which was queried to pull down the original fastq files. Terms such as tissue, general_celltype, specific_celltype and treatment_code were populated by reviewing the publication associated with GEO SRA number. Fields such as mapped_reads, total_reads and percent_mapped refer to quality metrics output from running bowtie2. Finally bidirectionals refer to the total number of bidirectional origins predicted by Tfit and TSS refers the proportion of those associated with a transcription start site ($\mu < 1500$ of RefSeq TSS annotation)

Supplementary Table C.6: **Tfit bidirectional prediction table**

This description corresponds to supplementary tables labeled as SRA#.csv within the tar ball "tfit_predictions" associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al. These tables are comma separated files generated for each dataset in the Supplementary Table C.5 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 ($\mu < 1500$) a RefSeq transcriptional start site annotation. In this way, the set of eRNAs are those bidirectionals where tss = 0.

Supplementary Table C.7: **Sites of TF-binding Motifs**

This description corresponds to supplementary tables labeled as motif_model.bed within the tar ball labeled "motif_models" associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al. At the time of this publication, there exists 641 HOCOMOCO motif models. The genome was scanned for these nucleotide sites (as described in Supplementary Note C.6) and significant hits are represented in bed file format. Files are named according to the unique HOCOMOCO motif model identifier.

Supplementary Table C.8: **Motif Displacement Histograms**

This description corresponds to Supplementary tables labeled as SRA#.csv within the tar ball name "Motif_Displacements" associated with *Enhancer RNA Expression Predicts Transcription Factor Activity* by Azofeifa et al. For each dataset in the conditions table, 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 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 the specified motif.