Bioinformatics, 33(2), 2017, 227–234
doi: 10.1093/bioinformatics/btw599
Advance Access Publication Date: 23 September 2016
Original Paper

OXFORD

Gene expression

A generative model for the behavior of RNA polymerase

Joseph G. Azofeifa[1] and Robin D. Dowell[1,2,3,*]

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

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on May 29, 2016; revised on August 31, 2016; accepted on September 12, 2016

Abstract

Motivation: Transcription by RNA polymerase is a highly dynamic process involving multiple distinct points of regulation. Nascent transcription assays are a relatively new set of high throughput techniques that measure the location of actively engaged RNA polymerase genome wide. Hence, nascent transcription is a rich source of information on the regulation of RNA polymerase activity. To fully dissect this data requires the development of stochastic models that can both deconvolve the stages of polymerase activity and identify significant changes in activity between experiments.
Results: We present a generative, probabilistic model of RNA polymerase that fully describes loading, initiation, elongation and termination. We fit this model genome wide and profile the enzymatic activity of RNA polymerase across various loci and following experimental perturbation. We observe striking correlation of predicted loading events and regulatory chromatin marks. We provide principled statistics that compute probabilities reminiscent of traveler's and divergent ratios. We finish with a systematic comparison of RNA Polymerase activity at promoter versus non-promoter associated loci.
Availability and Implementation: Transcription Fit (Tfit) is a freely available, open source software package written in C/C++ that requires GNU compilers 4.7.3 or greater. Tfit is available from GitHub (https://github.com/azofeifa/Tfit).
Contact: robin.dowell@colorado.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Regulation of gene expression plays crucial roles in diseased and healthy cellular phenotypes. Gene expression requires RNA Polymerase II (RNAP) recruitment to promoters and subsequent signaling cues to direct RNAP to fully transcribe the protein coding region (Bentley, 2014; Fuda *et al.*, 2009). 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 (Fuda *et al.*, 2009; Jonkers and Lis, 2015; Kwak *et al.*, 2013; Nojima *et al.*, 2015).

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 (Core and Lis, 2008; Kwak *et al.*, 2013; Nojima *et al.*, 2015). Given their high degree of resolution and strand specific nature, these assays have tremendous potential to refine our 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 (Azofeifa *et al.*, 2016; Fong *et al.*, 2015; Hah *et al.*, 2013; Wang *et al.*, 2011). Yet the precision of these techniques depends on an inherently noisy sequencing

process with biases in both the experimental protocol (Kwak *et al.*, 2013) 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 transcripts within nascent transcription data, work has primarily focused on segmentation based algorithms such as hidden Markov models (HMMs) (Azofeifa *et al.*, 2014; Chae *et al.*, 2015) and windowing approaches (Allison *et al.*, 2013). 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 (Natoli and Andrau, 2012). In fact, traditional segmentation analysis tends 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 of the gene body (Bentley, 2014; Kwak *et al.*, 2013). Several recent efforts have focused on identifying the bidirectional transcripts characteristic of initiating/paused RNAP using supervised learning approaches such as naive Bayes (Melgar *et al.*, 2011), support vector regression (Danko *et al.*, 2015) or logistic regression (Azofeifa *et al.*, 2014). 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.

Our first approach to a unified model of RNAP behavior described paused RNAP as an asymmetric double geometric followed by elongation signal defined as a homogeneous Poisson point process (Lladser *et al.*, 2016). Although a significant step forward, model inference was constrained to single isoform genes as the parameter estimation method supports only one loading event. Additionally, the model was single stranded and therefore most applicable to organisms such as *Drosophila melanogaster* where paused RNAP does not show bidirectional transcript signal.

To address these limitations, we propose a novel generative model of RNAP that describes both initiating/paused and elongating RNAP. The model accounts for signal on both strands simultaneously, capturing the behavior of RNA Polymerase II genome-wide. Even in light of the non-exponential family distribution functions, we develop a parameter estimation method based on the theory of maximum likelihood. With our model in hand, we perform inference into RNAP activity and assay changes in loading event locations and pausing probabilities across conditions.

2 Algorithm

2.1 Model description

Eukaryotic gene expression is a highly coordinated stochastic process involving the enzymatic synthesis of RNA by RNAP. The precise location of RNAP along DNA can be measured either by chromatin immunoprecipitation or nascent transcription assays. Conceptually, in the absence of noise, each read originates from an actively engaged RNAP molecule. Here we present a unified probabilistic model of transcription that captures the position Z of RNAP (Fig. 1).

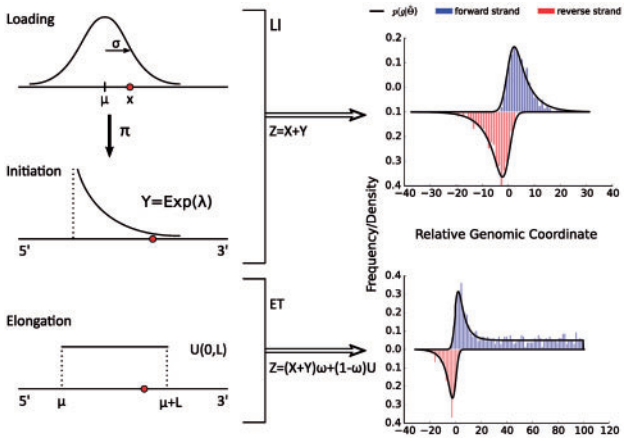

Fig. 1. Model of polymerase activity. A summary of the probabilistic model (on left, see text for full description of parameters) with examples of data generated from the model (on right). Here 'Loading' refers to recruitment of polymerase and pre-initiation complex formation, 'Initiation' refers to initiation of transcription and promoter-proximal pausing, and 'Elongation' refers to productive elongation following pause release (Fuda *et al.*, 2009; Adelman and Lis, 2012; Lee and Young, 2013; Jonkers and Lis, 2015) (Color version of this figure is available at *Bioinformatics* online.)

At protein coding genes, RNAP is first recruited to the promoter region at the transcriptional start site (TSS). We 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 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 (also referred to as entry length (Jonkers and Lis, 2015)), 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 1).

$$Z|S \sim X + SY \tag{1}$$

In Equation 1, $S \in \{-1, +1\}$ 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. We solve these convolutions analytically and provide a properly normalized probability distribution function (Equation 2) governing the loading position and entry length of RNAP.

$$h(z, s; \mu, \sigma, \lambda, \pi) = \lambda\phi\left(\frac{z-\mu}{\sigma}\right)R\left(\lambda\sigma - s\frac{z-\mu}{\sigma}\right)\mathbb{1}(s)$$
$$\mathbb{1}(s) = \begin{cases} \pi & : s = +1 \\ 1-\pi & : s = -1 \end{cases} \tag{2}$$

From Equation 2, $\phi(\cdot)$ 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 (Reed and Jorgensen, 2004).

Following recruitment, strand decision and initiation, RNAP may transition to productive elongation (e.g. pause release) moving

$5' \rightarrow 3'$ to transcribe the full length of the coding sequence. And finally, RNAP releases from the DNA at termination sites, $l_s = \{l_f, l_r\}$ for the forward and reverse strand respectively. We describe the location of elongating polymerases as a homogeneous Poisson point process (Lladser *et al.*, 2016). By well-known conditioning results for Poisson point processes (Kingman, 1992), the positions of elongating polymerases should be uniformly distributed (Equation 3) between the loading (μ) and termination sites (l_s).

$$u(z, s; \mu, l_s) = \frac{\mathbb{1}(z, s)}{s \cdot (l_s - \mu)}$$
$$\mathbb{1}(z, s) = \begin{cases} 1: & sz \geq s\mu; sz \leq sl_s \\ 0: & \text{otherwise} \end{cases} \tag{3}$$

In equation 3, $\mathbb{1}(\cdot)$ represents an indicator function that directly encodes the biological constraint that elongating RNAP must first load at μ. Given that Z describes the location of RNAP, it originates either from the initiating/paused, $h(z, s)$, or elongating, $u(z, s)$, stage of transcription. For brevity, we refer to the Loading/Initiation/Paused and Elongating/Termination stages as LI and ET respectively.

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, $\mathbb{K} = \{k \in N^+ : k \leq \mathbf{M}\}$ 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 4).

$$p(z, s; \Theta) = \sum_{k \in \mathbb{K}} w_k f(z, s, k; \theta_k) \tag{4}$$

Importantly, $f(z, s, k)$ in Equation 4 may represent either $h(z, s)$ or $u(z, s)$ (Equations 2, 3 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.

2.2 Model inference

Under a finite $\mathbf{M}$-mixture model, we wish to perform model inference over Θ given nascent transcription data. Let G be the set of aligned reads across the entire genome where each $g \in G$ consists of a genomic coordinate z and strand identifier s. At some genomic locus $[a, b]$, let $\mathbf{D} = \{g \in G : a \leq z \leq b\}$ and $N = |\mathbf{D}|$. In total, we seek to identify a parameter set Θ^* under which $\mathbf{D}$ is most probable, $\mathcal{L}(\Theta|\mathbf{D})$ (Equation 5), i.e. the maximum likelihood estimate (MLE).

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{N} \sum_{k \in \mathbb{K}} w_k p(g_i; \theta_k) \tag{5}$$

Without specifying the set of transcriptional component identifiers ($\mathbf{K}$) associated with $\mathbf{D}$, Equation 5 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 ($\mathbf{Y}$) effectively decouples the convolution in Equation 2 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 $z_i, s_i, k_i,$ *and* y_i.

Although we do not observe $\mathbf{K}$ or $\mathbf{Y}$, we can treat k_i and y_i as random variables and perform iterative optimization of Equation 5 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 6) and the **M-step** performs a gradient step along this expectation.

$$\mathbb{E}[\log p(\mathbf{C}|\theta)|\mathbf{D}, \theta^t] = \int_{y \in \mathbb{R}^+} \sum_{k \in \mathbb{K}} \sum_{i=1}^{N} \log p(k_i, y_i, g_i; \theta) \prod_{j=1}^{N} p(y_j, k_j | g_j; \theta^t) \tag{6}$$

Admittedly daunting, simplification of Equation 6 can be achieved in a number of ways. First, we 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, we need not consider $p(y_i|g_i)$ for mixture components involving elongating polymerase. Therefore, we 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 7) follows immediately from Bayes' Theorem and for succinctness we define this term as r_i^k. Commonly referred to as the *responsibility term* (Bilmes *et al.*, 1998), r_i^k measures the extent to which g_i belongs to some component k.

$$r_i^k = p(k|g_i; \theta_k^g) = \frac{w_k \cdot p(g_i; \theta_k^g)}{\sum_{k \in \mathbb{K}} w_k \cdot p(g_i; \theta_k^g)} \tag{7}$$

Turning to $\mathbf{Y}$, the convolution of Z induced by the simultaneous loading and initiation of RNAP requires a more involved computation to the expected value of $\mathcal{L}(\mathbf{D} \cup \mathbf{Y})$ given the incomplete data (Equation 8). Fortunately however, knowledge of $Z = z_i$ reveals conditional dependence between X and Y and thus a way forward for iterative optimization of $\mu, \sigma,$ and λ.

$$\begin{aligned} &\mathbb{E}[\log p(\mathbf{D} \cup \mathbf{Y}|\theta)|\mathbf{D}, \theta^t] = \\ &\sum_{i=1}^{N} \int_0^{\infty} \log p(y_i, g_i; \theta) p(y_i|g_i; \theta^t) dy = \\ &\log \frac{N\lambda}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{N} [\lambda \mathbb{E}[Y|g_i; \theta^t] + \\ &\frac{1}{2\sigma^2} \left(z_i^2 - s_i \mathbb{E}[Y|g_i; \theta^t] + \mathbb{E}[Y^2|g_i; \theta^t] - 2\mu(z_i - s_i \mathbb{E}[Y|g_i; \theta^t]) + \mu^2\right)] \end{aligned} \tag{8}$$

To complete the **E-step**, we define the conditional expectation of the initiating length Y conditioned on θ^t, Equation 9.

$$\begin{aligned} \mathbb{E}[Y|g_i; \theta^t] &= s_i(z - \mu) - \lambda\sigma^2 \\ &\quad + \frac{\sigma}{R(\lambda\sigma - s_i(z_i - \mu)/\sigma)} \\ \mathbb{E}[X|g_i; \theta^t] &= z_i - s_i \mathbb{E}[Y|g_i; \theta^t] \\ \mathbb{E}[Y^2|g_i; \theta^t] &= \lambda^2\sigma^4 + \sigma^2(2\lambda(\mu - z)s_i + 1) \\ &\quad + (z_i - \mu)^2 - \frac{\sigma(\lambda\sigma^2 + s_i(\mu - z_i))}{R(\lambda\sigma - s_i(z_i - \mu)/\sigma)} \\ \mathbb{E}[X^2|g_i; \theta^t] &= \mathbb{E}[X|g_i; \theta^t]^2 + \mathbb{E}[Y^2|g_i; \theta^t] \\ &\quad - \mathbb{E}[Y|g_i; \theta^t] \end{aligned} \tag{9}$$

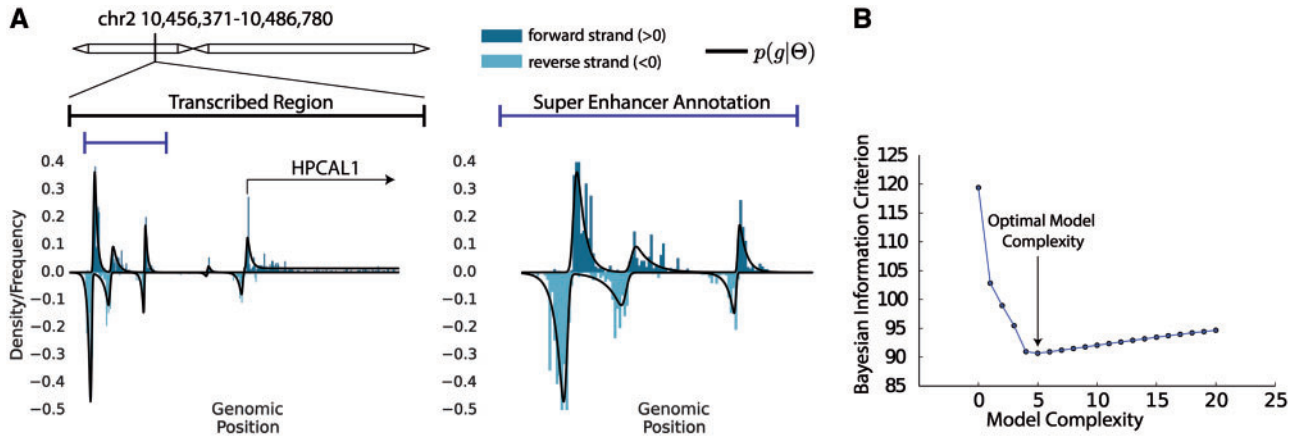

Fig. 2. Characteristic loci showing RNAP inference. (**A**) The final inferred density function at characteristic transcribed region and the super enhancer region contained therein, defined by FStitch (Azofeifa *et al.*, 2016) and dbSuper (Khan and Zhang, 2016) respectively. (**B**) The BIC calculation across 20 mixture models. A model complexity of 5 is shown to be minimal and thus considered optimal (Color version of this figure is available at *Bioinformatics* online.)

Expectations over X (the random variable governing RNAP loading position) given g_i and θ^t can be shown easily from the linearity of expectation.

With the necessary conditional expectations defined, we solve for the maximum of Equations 6 and 8. Equation 10 provides the 'update rules' for the EM algorithm.

$$w_k^{t+1} := \frac{r_k}{r}; \pi_k^{t+1} := \frac{\sum_{i=1}^{N} r_i^k I(s_i = 1)}{r_k}$$
$$\mu_k^{t+1} := \frac{1}{r_k}\sum_{i=1}^{N}\mathbb{E}[X|g_i;\theta^t]r_i^k; \frac{1}{\lambda_k^{t+1}} := \frac{1}{r_k}\sum_{i=1}^{N}\mathbb{E}[Y|g_i;\theta^t]\cdot r_i^k \quad (10)$$
$$\sigma_k^{t+1} := \frac{1}{r_k}\left[\sum_{i=1}^{N}\mathbb{E}[X^2|g_i;\theta^g]r_i^k - 2\mu_k\sum_{i=1}^{N}\mathbb{E}[X|g_i;\theta^t]r_i^k\right] + \mu^2$$

In keeping with the traditional notation of mixture models in Equation 10, we 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, our EM update rules (Equation 10) assume that l_s is fixed, presumably to the minimal or maximal order statistics, g_0 and g_n of **D**. However, the length of elongation or exact site of termination varies throughout the genome (Derrien *et al.*, 2012). 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, we 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, we place a uniform distribution over **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 **D**. We provide a pseudo-code description of our MLE methodology in the Supplement (Algorithm 1).

2.3 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, we utilize penalized Bayesian Information Criterion (BIC), equation 11.

$$\mathrm{BIC}(\Theta, \mathcal{L}) = \alpha|\Theta|\log N - 2\log\mathcal{L} \quad (11)$$

$\mathcal{L}$ 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 **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 Results

3.1 Estimation of RNAP location from GRO-seq

After confirming our model inference method using simulated data (see Supplementary text and Fig. S1), we assess our model on publicly available biological data. To perform model inference of RNAP location requires an interval $[a, b]$ where GRO-seq read mapping data (**D**) can be collected. To this end, we utilized Fast Read Stitcher (FStitch) that implements a maximum entropy Markov model to segment the genome into 'transcribed regions' (Azofeifa *et al.*, 2014, 2016). In a HCT116 GRO-seq dataset (Allen *et al.*, 2014), FStitch classified 19 709 transcribed regions. With these regions in hand, we computed Θ^* by our MLE methodology across mixtures containing $|\mathbb{K}| \in \{1, 2, \ldots, 20\}$ for each interval independently and selected a final Θ^* by the minimum BIC score.

Figure 2A shows a transcribed region with reference to the estimated density function. As a final illustrative example, Figure 2B displays the associated Bayesian Information Criterion scores as a function of model complexity. Supplementary Table S1 provides a collection of statistics describing the distribution of fitted parameters across all transcribed regions.

To address the accuracy of our model complexity procedure, we reasoned that at active single isoform genes we should predict only one LI component while at transcriptionally inactive regions (by FStitch) we should predict no components. Supplementary Figure S2 highlights the accuracy of our RNAP inference model to discriminate between active and inactive transcribed regions based solely on LI component presence (AUC≈0.95). At a FDR of 0.05, we observe that the distribution of $|\mathbb{K}_p|$ at single isoform genes contains a clear and prominent mode at $|\mathbb{K}_p| = 1$ (Supplementary Fig. S2).

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 (Supplementary Fig. S3). To assess whether the model is incorrectly introducing extra LI component centers to compensate for data poorly described by an exponentially modified Gaussian distribution, we compared the distribution of $|\widehat{\mu}_i - \widehat{\mu}_j|$ $(i \neq j)$ at loci harboring $|\mathbb{K}_p| > 1$ to the distribution of LI component standard deviation $(\widehat{\sigma} + \widehat{1/\lambda})$. We observe that median pairwise LI component center distance far exceeds what we would expect under the variability of LI component sizes, indicating that $\{\widehat{\mu}_1, .., \widehat{\mu}_k\}$ describe independent portions of the data.

Suggested by the associated standard deviations in Supplementary Table S1, $\widehat{\Theta}$ 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 $\widehat{\Theta}$ across biological replicates. To this end, FStitch defined transcribed regions between replicate one and two were merged and model inference, by Tfit, was performed in each replicate independently.

We observe strong correlation in model selection (Supplementary Fig. S4A) and exceedingly high correlation between identical parameters (e.g. $\rho(\lambda_{rep1}, \lambda_{rep2}) = 0.95$, Supplementary Fig. S4B), suggesting that estimated parameters display low variance. Apart from w and l_s, we observe little to no correlation between differing parameters (e.g. λ_{rep1} and π_{rep2}, Supplementary Fig. S4C), suggesting that there is no confounding dependencies between parameters.

Estimates of w_p sufficiently larger than the population average (two standard deviations, Supplementary Table S1) 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. We observed that the loading strand bias (π) tracks closely with the strand orientation of the underlying RefSeq gene (Supplementary Fig. S5A). 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 our model predicts μ as the site of RNAP loading and l_s as the site of elongating termination, we compared the location of μ and l_s to estimates of annotated transcriptional start sites (TSS) and termination sites respectively. We observed a high degree of correlation between μ and the TSS, noting a significant ≈40 base pair upstream displacement of μ from the TSS (Supplementary Fig. S5B). This displacement is in line with estimates from other groups using independent methods (Jonkers and Lis, 2015). Similar to previous estimates of transcription termination (Azofeifa *et al.*, 2016; Fong *et al.*, 2014), we observed l_s to be ≈6KB downstream the polyadenylation site (Supplementary Fig. S5C).

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 (Adelman and Lis, 2012). 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

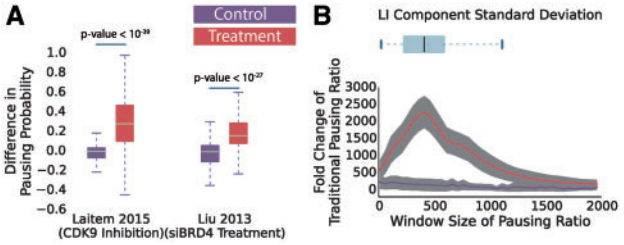

Fig. 3. Changes in promoter proximal pausing are correctly identified by a generative model of RNAP. Mixture model inference was performed over RefSeq gene annotations between control and treated cell lines in two independently derived datasets: Laitem 2015 and Liu 2013. (**A**) The predicted difference in LI mixing weights between control (on left) and treated cells. Under a normal assumption, mean mixing weights were compared using a *t*-test. (**B**) The distribution of LI length (box) with the traditional computation of the pausing ratio as a function of window size, plotted for the control (bottom line) and treated (top line) cell lines of the Laitem 2015 dataset. Grey shading indicates one tenth of one standard deviation (Color version of this figure is available at *Bioinformatics* online.)

that our model directly infers LI and ET RNAP stages from data alone, we ask whether we can correctly identify changes in RNAP activity following experimental perturbation known to affect promoter proximal pausing.

We 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 (Liu *et al.*, 2013). Yet another study noted global shifts in RNAP pausing ratios by cyclin-dependent kinase (CDK) 9 inhibition in HeLa cells (Laitem *et al.*, 2015). With these datasets, we hypothesis that our model should accurately reproduce the observed changes in pausing ratios by appropriate changes in the LI and ET mixing weights.

For each dataset, we performed model estimation and computed changes in estimated parameters between the untreated and treated cells. Specifically, we fit a single component mixture model (one LI component and two ET components) at each gene annotation region. We 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, we observed highly significant global changes in LI component mixing weights relative to untreated replicates (Fig. 3A).

Although easily computable, the standard pausing ratio calculations rely on ad hoc methods of window sizes and distance thresholds (Adelman and Lis, 2012). To highlight this point explicitly, we examined the impact of TSS-centered window size on the pausing ratio (Fig. 3B). Intuitively, window sizes that are either too small or too large dramatically reduce the observable differences between treated and untreated cells. For comparison, we provide the distribution of LI standard deviation obtained from our model.

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 (The ENCODE Project Consortium, 2007, 2012). 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 (Li *et al.*, 2016) 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 our 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 positive and negative skew emanating from μ. To profile for bidirectional transcripts genome wide, we hereafter utilized our EM seeding method (complete description in Supplement; Θ_B set to the parameter values in Supplementary Table S1).

To assess the accuracy of our bidirectional transcript classifications, we 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 (Allen *et al.*, 2014), we benchmarked our method against the current state-of-the-art bidirectional detection algorithm, dREG (Danko *et al.*, 2015). 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 (The Encode Project Consortium, 2012), MACS broad peak settings) and the resulting bidirectional transcript prediction. To assess false positives, we 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. We observed improvements over dREG across all regulatory marks (Fig. 4A, B). Both dREG and Tfit predict H3K27ac marks exceedingly well from bidirectional transcript presence alone, suggesting that H3K27ac signal reflects nascent transcription.

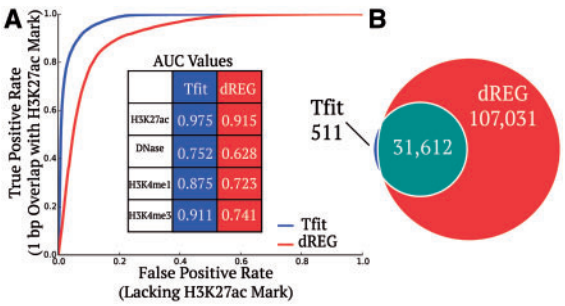

Fig. 4. RNAP model accurately profiles for bidirectional transcription. (**A**) A receiver operating characteristic (ROC) curve displaying the relationship between true and false positive rates of H3K27ac prediction from bidirectional transcription alone. The area under the ROC curve (AUC) values are 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 (Color version of this figure is available at *Bioinformatics* online.)

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 (Allen *et al.*, 2014; Azofeifa *et al.*, 2014; Le *et al.*, 2013) point prominently to the functional importance of enhancer transcripts. To begin to address the question of enhancer RNA function, we present an analysis that demonstrates both the utility of our 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 (Phillips and Corces, 2009; Splinter *et al.*, 2006), we 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 (The Encode Project Consortium, 2012). With a cell line matched GRO-seq dataset (Core *et al.*, 2014), we compared network attributes of loci containing or lacking bidirectional transcript predictions ($\gamma = 1$, FDR = 0.05 according to H3K27ac prediction). Figure 5A displays two illustrative connected component examples, built as described in the Supplement.

Enhancers are implicated in defining key cellular phenotypes such as cell fate (Creyghton *et al.*, 2010; Whyte *et al.*, 2013) and tumorigenesis (Hnisz *et al.*, 2015; Qian *et al.*, 2014) and thus may play a central role in three dimensional looping. Our constructed network reflects locations in the genome (nodes) connected by the CTCF ChIA-PET data. With this in mind, we grouped nodes by whether they lacked or contained an association with an annotated TSS or Tfit prediction and computed common measures of node centrality. We observed the highest degree of node centrality at non TSS associated bidirectional transcripts across all criteria (Supplementary 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 (Deng *et al.*, 2013). These discrete nuclear sites of transcription allow for rapid expression of many three dimensionally proximal genes (Edelman and Fraser, 2012). To this end, we 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 (Fig. 5B). Furthermore, computation of network modularity (a measure of network label

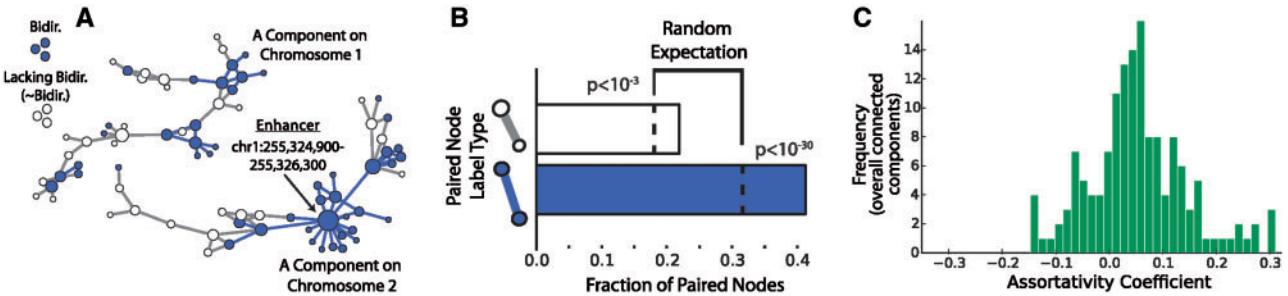

Fig. 5. CTCF paired loci network displays assortativity by bidirectional presence (**A**) displays two characteristic connected components on chromosome 1 and 2 from a CTCF ChIA-PET dataset derived from the K562 cell line. Nodes are colored as to whether a Tfit prediction overlaps a paired loci by one base pair or not; Bidir. and ~Bidir. respectively. The circumference of the node is proportional to the degree. (**B**) The proportion of edges containing a similar label, significance is calculated by a Binomial test. (**C**) The distribution of the assortativity coefficient across all connected components, > 0 indicates modularity (Color version of this figure is available at *Bioinformatics* online.)

clustering) shows weak assortativity across most connected components (Fig. 5C).

4 Discussion

We described a probabilistic, generative mixture model that is founded on a biologically motivated description of RNAP behavior. Our model is inspired by the current understanding of polymerase behavior (Core *et al.*, 2014; Fuda *et al.*, 2009; Kwak *et al.*, 2013; Lee and Young, 2013) at protein coding genes and provides a principled mathematical approach to GRO-seq data analysis. To perform model inference, we derived a parameter estimation scheme based on the theory of maximum likelihood and the Expectation Maximization algorithm. When applied to GRO-seq, our model can not only identify individual transcripts but also quantify their characteristics, as evidenced by the fact that the loading site predictions of our model correlate well with RefSeq annotation.

The relatively recent and unexpected discovery that enhancers are themselves transcribed challenges our conventional view of transcriptional regulation (Li *et al.*, 2016). Whether the underlying region is a promoter or an enhancer, our model explicitly assumes a singular behavior of RNAP genome-wide. The ability of our model to fit well to transcribed regions that are unannotated suggests that the unified model recently proposed in the literature (Andersson *et al.*, 2015; Core *et al.*, 2014) and assumed here is appropriate. Turning our generative model to a discriminative classifier, we observed that our bidirectional transcript predictions precisely recover sites of marked regulatory chromatin. Indeed, our method outperforms the current state of the art algorithm (Danko *et al.*, 2015) for bidirectional transcript classification. Furthermore, our observation that non-TSS associated bidirectional transcripts constitute a central role in CTCF ChIA-PET networks points yet again to the increasing importance of enhancer RNAs.

The parameters of our model provide meaningful summary statistics of nascent transcription data. Changes in these statistics across cell types, experimental conditions and/or perturbations reflect biologically meaningful alterations in RNAP behavior. Consistent with this idea, our inference procedure confirmed a dramatic change in pausing probability following experimental perturbations for two independently derived datasets. We anticipate that our model will be used to assign particular regulatory proteins to the distinct stage of polymerase they regulate. In the case of a single experiment, correlations between the parameters of our model and other high throughput datasets will inform on the underlying regulatory process. For example, it is of great interest to monitor the co-occurrence of transcription factor binding events and motifs with initiation site predictions obtained from our model.

As we learn more about how polymerase is regulated, it will be possible to extend our model accordingly. Close inspection of GRO-seq read coverage reveals oscillatory behavior within the gene body, as such a homogeneous Poisson point process may not be an appropriate model governing transcriptional elongation. Pol II elongation rates vary and influence a number of co-transcriptional processes (Jonkers and Lis, 2015; Jonkers *et al.*, 2014), suggesting that the undulations observed in the elongation region could be biologically informative. However, the extent to which heightened levels of mapped reads correspond to mapping biases or biologically meaningful points of transcription regulation is unclear. Additionally, RNAP is thought to pause proximal to its termination (Bentley, 2014). Indeed, an interesting and prominent 3′ peak is often observed a few kilobases upstream of the end of the elongation region (Fong *et al.*, 2014). More detailed studies of both elongation and transcriptional termination, both experimentally and computationally, are needed to shed light onto these processes (Fong *et al.*, 2014; Jonkers and Lis, 2015). Furthermore, an improved understanding of these process may also require us to re-evaluate the interpretation of particular parameters.

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 our mixture model was to provide a set of biologically interpretable parameters that capture alterations in polymerase behavior induced by changes in regulatory proteins. To this end, a key next step will be the development of rigorous statistical methods for detecting potentially small, but meaningful changes in model parameters between experiments. We believe our proposed mixture model is one of the first steps in building a comprehensive predictive model of RNAP.

Acknowledgement

We would like to thank Mary A. Allen, Ryan E. Langendorf and Samuel F. Way for discussions related to the model design and Josephina Hendrix for assistance with analysis of publicly available datasets. The authors acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing High Performance Computing resources (NIH 1S10OD012300) supported by BioFrontiers IT.

Funding

This work was supported by the National Science Foundation [DBI-1262410; DGE-1144807].

Conflict of Interest: none declared.

References

Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.*, **13**, 720–731.

Allen,M.A. *et al.* (2014) Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *eLife*, **3**, e02200.

Allison,K.A. *et al.* (2013) Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic Acids Res.*, PMID: 24304890.

Andersson,R. *et al.* (2015) A unified architecture of transcriptional regulatory elements. *Trends Genet.*, **31**, 426–433.

Azofeifa,J. *et al.* (2014). 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, pp. 174–183. ACM, New York, NY, USA.

Azofeifa,J. *et al.* (2016) An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-seq. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, doi: 10.1109/TCBB.2016.2520919.

Bentley,D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, **15**, 163–175.

Bilmes,J.A. *et al.* (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Int. Comput. Sci. Inst.*, **4**, 126.

Chae,M. *et al.* (2015) groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*, **16**, 222.

Core,L. and Lis,J. (2008) Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, **319**, 1791–PMID: 18369138.

Core,L.J. *et al.* (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.

Creyghton,M.P. *et al*. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci*., **107**, 21931–21936.

Danko,C.G. *et al*. (2015) Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods*, **12**, 433–438.

Deng,B. *et al*. (2013) Transcription factories, chromatin loops, and the dysregulation of gene expression in malignancy. *Semin. Cancer Biol*., **23**, 65–71.

Derrien,T. *et al*. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*., **22**, 1775–1789.

Edelman,L.B. and Fraser,P. (2012) Transcription factories: genetic programming in three dimensions. *Curr. Opin. Genet. Dev*., **22**, 110– 114. PMID: 22365496.

Fong,N. *et al*. (2014) Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev*., **28**, 2663–2676.

Fong,N. *et al*. (2015) Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA polymerase II termination suggest widespread kinetic competition. *Mol. Cell*, **60**, 256–267.

Fuda,N.J. *et al*. (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, **461**, 186–192. PMID: 19741698.

Hah,N. *et al*. (2013) Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res*., **23**, 1210–1223.

Hnisz,D. *et al*. (2015) Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell*, **58**, 362–370.

Jonkers,I. and Lis,J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol*., **16**, 167–177.

Jonkers,I. *et al*. (2014) Genome-wide dynamics of pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, **3**,

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

Kingman,J.F.C. (1992) *Poisson Processes*, vol. **3**. Clarendon Press, Oxford.

Kwak,H. *et al*. (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science (New York, N.Y.)*, **339**, 950–953.

Laitem,C. *et al*. (2015) CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes. *Nat. Struct. Mol. Biol*., **22**, 396–403.

Le,T.P. *et al*. (2013) Mapping ERβ genomic binding sites reveals unique genomic features and identifies EBF1 as an ERβ interactor. *PLoS ONE*, **8**, e71355.

Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.

Li,W. *et al*. (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet*., **17**, 207–223.

Liu,W. *et al*. (2013) Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell*, **155**, 1581–1595.

Lladser,M.E. *et al*. (2016) RNA Pol II transcription model and interpretation of GRO-seq data. *J. Math. Biol*., doi: 10.1007/s00285-016-1014-4.

Melgar,M. *et al*. (2011) Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol*., **12**, R113.

Natoli,G. and Andrau,J.C. (2012) Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet*., **46**, 1–19. PMID: 22905871.

Nojima,T. *et al*. (2015) Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell*, **161**, 526–540.

Phillips,J.E. and Corces,V.G. (2009) CTCF: Master weaver of the genome. *Cell*, **137**, 1194–1211.

Qian,J. *et al*. (2014) B cell super-enhancers and regulatory clusters recruit aid tumorigenic activity. *Cell*, **159**, 1524–1537.

Reed,W.J. and Jorgensen,M. (2004) The double pareto-lognormal distributiona new parametric model for size distributions. *Commun. Stat.-Theory Methods*, **33**, 1733–1753.

Splinter,E. *et al*. (2006) CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. *Genes Dev*., **20**, 2349–2354.

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

The Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74. PMID: 22955616.

Wang,D. *et al*. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390–394. PMID: 21572438.

Whyte,W.A. *et al*. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.