FStitch: A fast and simple algorithm for detecting nascent RNA transcripts

Joseph Azofeifa
Department of Computer Science
University of Colorado
596 UCB, JSCBB
Boulder, CO 80309

Mary A. Allen
BioFrontiers Institute
University of Colorado
596 UCB, JSCBB
Boulder, CO 80309

Manuel Lladser
Department of Applied Mathematics
University of Colorado
526 UCB
Boulder, CO 80309

Robin Dowell*
Department of MCD Biology & Computer Science
BioFrontiers Institute
University of Colorado
596 UCB, JSCBB
Boulder, CO 80309

ABSTRACT

We present a fast and simple algorithm to detect nascent RNA transcription in global nuclear run-on sequencing (GRO-seq). GRO-seq is a relatively new protocol that captures nascent transcripts from actively engaged polymerase, providing a direct read-out on bona fide transcription. Most traditional assays, such as RNA-seq, measure steady state RNA levels, which are affected by transcription, post-transcriptional processing, and RNA stability. A detailed study of GRO-seq data has the potential to inform on many aspects of the transcription process. GRO-seq data, however, presents unique analysis challenges that are only beginning to be addressed. Here we describe a new algorithm, Fast Read Stitcher (FStitch), that takes advantage of two popular machine-learning techniques, a hidden Markov model (HMM) and logistic regression to robustly classify which regions of the genome are transcribed. Our algorithm builds on the strengths of previous approaches but is accurate, dependent on very little training data, robust to varying read depth, annotation agnostic, and fast.

Categories and Subject Descriptors

I.2.1 [**ARTIFICIAL INTELLIGENCE**]: Applications and Expert Systems—Medicine and science

General Terms

Algorithms, Experimentation

*corresponding author Robin.Dowell@colorado.edu

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.

http://dx.doi.org/10.1145/2649387.2649427

Keywords

Nascent Transcription, Logisitic Regression, Hidden Markov Models

1. INTRODUCTION

Almost all cellular stimulations trigger global transcriptional changes. To date, most studies of transcription have employed RNA-seq or microarrays. These assays, though powerful, measure steady state RNA levels. Consequently, they are not true measures of transcription because steady state levels are influenced by not only transcription but also RNA stability. Only recently has a method for direct measurment of transcription genome-wide become available. This technique, known as global run-on sequencing (GRO-seq), simultaneously detects the amount and direction of actively engaged polymerases at every position within the genome[7]. GRO-seq has already drastically influenced our understanding of the transcription process, as most of the genome is transcribed but rapidly degraded.

The earliest and most common approach to GRO-seq analysis is annotation centric[7, 22, 17, 14] and only two efforts have attempted to identify regions of active transcription directly from GRO-seq data[2, 11]. The first of these approaches used a two state Hidden Markov model that was parametrized based on available annotations[11]. 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. The more recent approach, called Vespucci, uses a sliding-window (specified by two user-dependent parameters) that joins transcripts together based on read depth, but requires the user to tune the algorithm with each new dataset[2]. The windowing scheme, in principle, has the benefit of not depending on annotation, however, in practice, regions of transcription are often broken into discontiguous sections, requiring the use of annotations to improve their strategy[2].

The design of our software is largely motivated by both the strengths and shortcomings of these previous efforts[11, 21, 2]. In particular, we 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[19]. Such an algorithmic technique shares a similar structure with a Maximum Entropy Markov Models[19]. Thusly, our methodology allows for parameters to adapt on the fly to new data. It is annotation agnostic, effectively identifies cohesive regions of active transcription, has a rapid runtime, and is easily parameterized by providing a small number of training examples.

2. MATERIALS AND METHODS

2.1 Algorithm Description

The GRO-seq technique measures nascent transcripts produced from actively engaged polymerase. 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 sampling of 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. But because of uneven sampling, read coverage within active regions may not be continuous. Furthermore, the sequencing and mapping process is noisy, therefore reads can also map to inactive regions.

Transcription can be modeled as a discrete time-series indexed by genomic coordinates where transcriptional activity of adjacent base-pairs are correlated. Similar to prior models of GRO-seq[11], we model this process as an ergodic first-order Markov chain where transcription oscillates between active and inactive states. Unlike previous models, which classify individual nucleotides, our model emits from each state a contig representative of an active or inactive region (Figure 1). Each contig can be described by two feature classes: contig length (maximum length of overlapping reads) and contig coverage statistics (Table 1). Active states, in general, contain a combination of long regions with high signal interspersed with short regions of relatively no signal. Hence our 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.

Many of the features of contigs are dependent on sequencing depth of a particular experiment. Therefore, we must learn the emission and transition probabilities of each state from a user provided training set. In our case, the training set corresponds to regions of active and inactive transcription. Given a training set, we learn the conditional probabilities of a state classification from the set of implicit feature vectors using logistic regression. These logistic regression predictors are interpretable as probabilities, and therefore easily embedded into our Markov chain 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.

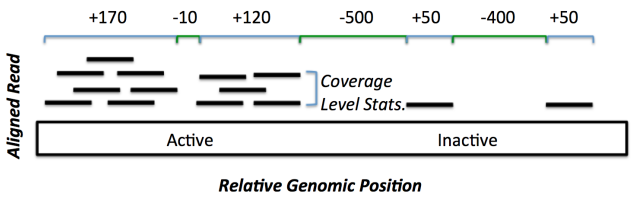

Figure 1: Contig length and coverage statistics discriminate *active* from *inactive* nascent transcription. A contig (blue) is defined by the length of overlapping reads x_1 in Table 1. Coverage statistics define mean, median, mode and variance of reads (black bars) across a contig. In places where there are no reads, a gap (green) is defined by a negative length value and all contig coverage statistics are set to zero.

2.2 Datasets

This study takes advantage of three previously published GRO-seq datasets: MCF-7 [11], IMR90 [7] and HCT116[1] as well as two previously published ChIP-Poll II datasets: HCT116[13] and MCF7 [15]. For each experiment, raw reads were mapped and aligned to the hg19 genome using Bowtie2 with the command bowtie -S -t -v 2 -best[16]. ENCODE provided H3K27ac and DNAse peak calls for IMR90 [21, 6], MCF7 [9, 12] and HCT116 [9, 24] as well as ChIA-PET peak calls for HCT116[10].

We hand annotated the entire length of chromosome 1 in the GRO-seq dataset from HCT116[1] to perform k-fold cross validation. For all testing, 95% of the labeled dataset was removed from training and used to asses model accuracy. To be clear, the entire labeled HCT116 training data contains 17,776 labeled *active* regions. In both the IMR90 and MCF7 GRO-seq datasets, 7 regions considered *active* and 7 regions considered *inactive* were labeled for parameter estimation. These training sets with genomic coordinates and labels are freely downloadable at `http://dowell.colorado.edu`.

2.3 Parameter Estimation

Both the Markov model transition probabilities and the conditional state emission probabilities 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 the parameters that define transcription from those describing read-mapping poses a difficult problem. To this end, the algorithm requires a small training dataset, provided by the user, in which regions of the genome are defined as either *active* or *inactive*. We show in section 3.1 that little training data is needed to retain high model accuracy. Intuitively, we define model accuracy as the fraction of base pairs where the user-label and classification-label agree.

Here we outline our logistic regression parameter estimation method, for a detailed exposition see Ohno-Machado's review [8]. We estimated the conditional probability $p(k \mid \vec{x})$, where $k \in \{inactive, active\}$ and $\vec{x}$ indicates our feature vector, via a labeled training set of defined genomic coordinates representing active or inactive transcription. Explicitly, Table 1 provides a complete description of the feature vector

Table 1: Feature vector $\vec{x}$ associated with a contig. y_i equals the read count at the i^{th} position between [t,t+l], where t is the genomic start of the contig and l its length. Feature vector dimensions are ordered by importance via recursive feature elimination.

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

$\vec{x}$. Clearly, $p(inactive \mid \vec{x}) = 1 - p(active \mid \vec{x})$. We express this probability as the sum of $\vec{x}$ weighted by some parameter vector $\vec{\theta}$. To treat this linear function as a probability, we bound the sum to the range [0,1] via the sigmoidal transformation as follows:

$$p(active \mid \vec{x}) = \frac{1}{1 + e^{-(x_0\theta_0 + x_1\theta_1 + ... + x_n\theta_n)}} = \frac{1}{1 + e^{-\vec{x}\cdot\vec{\theta}^T}} \quad (1)$$

A simple plot of two features, contig length (x_1) and average

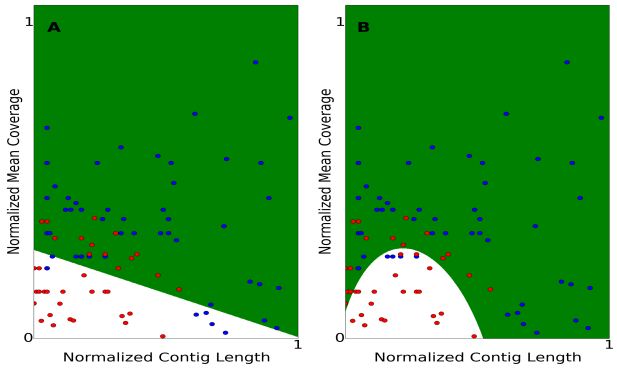

Figure 2: Read coverage features are not linearly separable.
Points colored blue represent training examples labeled *active* and those colored red indicate training examples labeled *inactive*. The green shade represents regions of the 2-D space that would be given a higher conditional probability to *active*, white *inactive*. (A) Uses a linear kernel function whereas (B) use a second-order polynomial kernel function specified in equation 2.

read coverage (x_3), shows that these features may not be linearly separable (Figure 2A). Because of this, we employ a polynomial kernel (equation 2) to learn non linear decision boundaries (Figure 2B),

$$f(\vec{x}, \vec{\theta}) = (\vec{x} \cdot \vec{\theta}^T + c)^d \quad (2)$$

The polynomial kernel function parameters c and d can be set by the user in the FStitch software package. The kernel function is incorporated into the sigmoidal transformation as follows:

$$p(active \mid \vec{x}) = \frac{1}{1 + e^{-f(\vec{x}, \vec{\theta}^T)}} \quad (3)$$

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

$$l(\vec{\theta}, D) = \sum_{i=1}^{n} \log p(k_i \mid \vec{x_i}) \quad (4)$$

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 our feature vector $\vec{x}$. k_i equals the i^{th} training label, which is either *active* or *inactive*.

We use the Newton-Raphson method[5] to iteratively update $\vec{\theta}$ until convergence. Because this techniques utilizes a second-order Taylor series approximation to the log-likelihood function, convergence is usually fast. The update rule is:

$$\vec{\theta}^{\,t+1} = \vec{\theta}^{\,t} - H^{-1} l(\vec{\theta}, D) \cdot \nabla l(\vec{\theta}, D) \quad (5)$$

Here Delta (∇) represents the gradient operator and H the Hessian operator. Finally, the most probable state sequence is estimated via the Viterbi Algorithm[23] and given by the recurrence relation:

$$v_t(k) = \max_{j \in S}(v_{t-1}(i) \cdot a_{j \to k}) \cdot p(k \mid \vec{x}) \quad (6)$$

where $a_{j \to k}$ represents the transition probability from state j to state k of the underlying Markov chain, which is estimated via Baum-Welch[19], S is the transcriptional state space {*active*, *inactive*} and $p(k \mid \vec{x})$ is given and learned independently in equations (4–5).

Learned parameters via trained data allow 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.

2.4 Detecting Enhancers as Divergent Transcription

Recent work indicates that enhancers are often transcribed, producing unstable transcripts that are detectable by GRO-seq[26]. Indeed, enhancers show a characteristic bidirectional transcription signature within GRO-seq data[21]. Only one analysis approach has, thus far, tried to leverage this bidirectional signal towards the *de novo* discovery of enhancers from GRO-seq signal[21]. A Naive Bayes classifier was trained on annotated regions in order to label unannotated 2kb windows as bidirectional, single stranded transcription, or non-transcribed[21].

Therefore, we asked whether our FStitch approach could be extended to detect enhancer RNAs (eRNAs) in an unbiased

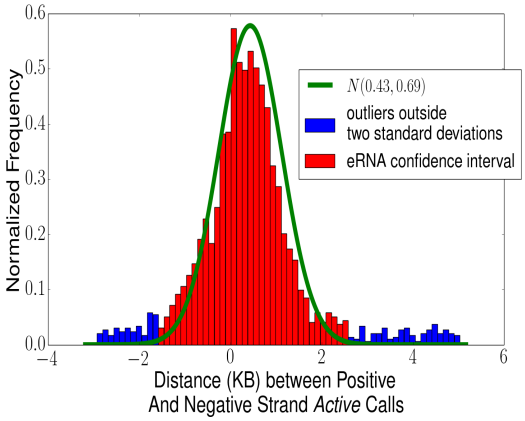

Figure 3: eRNA prediction Interval.
FStitch *active* calls that overlap both an H3K27ac and DHS chromatin marker were isolated on sense and anti-sense strand. Sense and anti-sense strand calls are paired to their nearest neighbors. The difference between the start of the positive strand call and the 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 indicates data points within two standard deviations.

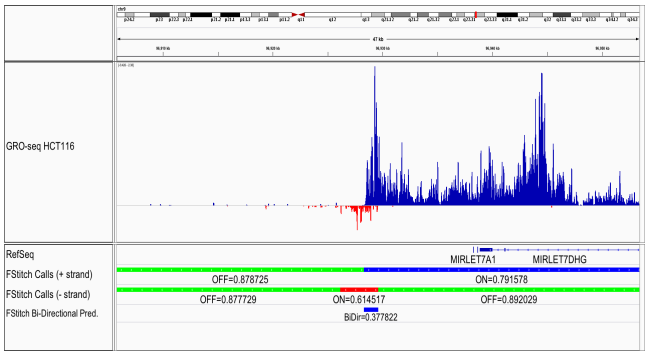

Figure 4: FStitch output.
The output of two Bed files are visualized within IGV showing a sub-region in chromosome 1. The first track shows typical GRO-seq data from the HCT116 dataset, the positive strand in blue, the negative strand in red. RefSeq annotations are shown next. FStitch output is below for each strand. Green indicates areas of inactive transcriptional activity, blue represents areas of active transcription along the positive strand and red along the negative strand. Scores are associated with each classification via the Logistic Regression and Viterbi-provided Markov state sequence. Bidirectional predictions are provided with a score via the estimated Normal Distribution confidence interval.

fashion. Conceptually, our algorithm could ask for overlapping transcription calls between the positive and negative strand as potential eRNAs similar to the Naive Bayes approach. However, it is unclear that all eRNAs show some overlap between the divergent transcripts as opposed to just relatively close proximity. Furthermore, many genes have long non-coding RNA transcripts which move anti-sense to a transcribed gene indicating that a simple overlap is not stringent enough a criterion for eRNA prediction. Because we are detecting bidirectional transcription, we expect also to find the 5'-end of many genes.

Therefore, we sought to examine the extent to which two transcripts must overlap or be adjacent in order to accurately annotate eRNAs. Using our chromosome 1 manually annotated dataset, we examined the overlap of these regions to both a DNA hypersensitivity site (DHS) and a H3K27ac mark, both indicators of enhancer activity. We then computed the distance to the nearest anti-sense FStitch call (Figure 3). We note that the displacement data show a Normal distribution. Therefore, we call a bidirectional signal where two anti-sense transcripts are within some number of standard deviations of the fitted Normal distribution. Furthermore, the confidence of bidirectional predictions can be adjusted by the user. In our subsequent analysis, divergent transcript analysis utilized a confidence interval of two standard deviations, i.e. -1.5kb to 2.25kb (Figure 3).

2.5 Algorithm Input and Output

The goal of the proposed algorithm is to segment the genome into areas of *active* and *inactive* nascent transcription and, user-friendliness was a large consideration in the design and structure of the software. FStitch accepts as input a BedGraph file of read coverage and a training set file consisting of a few segments (at least 3 segments) considered *active* versus *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 is rich in feature vectors (i.e. contig lengths and coverage statistics). As such the software design necessitates little input by the user while harnessing many training examples. FStitch has pre-labeled *active* and *inactive* segments based on house-keeping genes and gene desert regions, respectively. The user may opt for these default regions, however, care must be taken with these assumptions as the transcriptional landscape varies from experiment to experiment.

FStitch outputs two bed files (positive and negative strand classifications respectively) that can be imported into typical genome browsers (such as IGV or UCSC genome browser) to view the classifications in conjunction with read coverage files. Figure 4 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. Finally, the user may ask for divergent transcription predictions, as these are likely candidate eRNAs or 5'-ends of genes. From start to finish, FStitch takes ∼3.5 minutes to predict transcript annotations in the most deeply sequenced GRO-seq dataset, HCT116[1].

2.6 Software Availability

FStitch is written in the C/C++ programming languages and is 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 `http://dowell.colorado.edu`.

3. RESULTS

We present a fast and simple algorithm to detect nascent RNA transcription in GRO-seq that is annotation agnostic and robust to low read depth. This section is loosely divided into four categories: (1) algorithm performances and benchmarking, (2) RefSeq and previous technology comparisons, (3) validating bidirectional predictions as enhancer RNAs.

3.1 Sensitivity to depth of data

To assess the sensitivity of the algorithm to the amount of available training data, the authors hand curated the entire length of chromosome 1 in the HCT116 dataset. Regions were labeled as active-nascent or inactive-noise. Given this rich collection of labeled data, we performed K-fold cross validation. To this end, we reserved 5% of the training data for parameter estimation, 95% for testing accuracy. To assess the amount of training data needed for accurate test accuracy scores, we incrementally decreased the amount of training data.

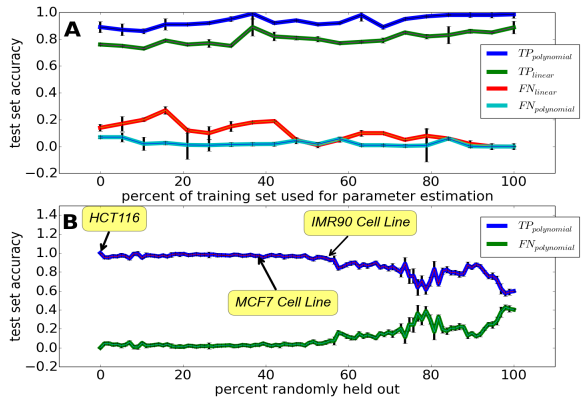

Figure 5: FStitch requires little training data and is robust to low levels of GRO-seq read coverage. (A) Graph of accuracy of classification given successively decreasing amounts of training data utilized to learn feature vector weights. (B) Graph of accuracy of classification given successively smaller sequencing depth (dataset size). In this case, we trained on 5% all available chromosome 1 labels and tested on subsamples of the original dataset (50 different subsamples of the test set). TP = *true positive rate*(blue) and FN = *false negative rate*(green). The utilized kernel is indicated (polynomial(d=2 and c=0) and linear(d=1 and c=0)).

Figure 5A shows that 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 levels of testing accuracy. The smallest successive decrease (0.1%) of the initial training dataset consists of 3 *active* and 2 *inactive* regions and maintains scores of 95% true positive and 4.3% false negative relative to the testing dataset.

Similarly, we assessed the sensitivity of FStitch to experimental sequencing depth, a rough measure of data quality. To this end, we randomly subsampled (without replacement) from the HCT116 dataset (the single experiment with the deepest read coverage). We subsampled the original dataset leaving out increasing amounts of the original dataset and re-estimated the parameters via the same training set segments. Subsequently, we reclassified *active* transcript segments and calculated training accuracy relative to the test set. Figure 5B shows that FStitch is robust to low sequencing depth of the dataset.

Table 2: Benchmarking FStitch and Vespucci
Each algorithm (FStitch, Vespucci with default parameters and Vespucci with best parameters from a grid search, G.S.) are compared to the manually annotated test set from chromosome 1 per base.

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

3.2 Benchmarking FStitch & Vespucci

We sought to evaluate our algorithm, FStitch, to the previously published windowing method Vespucci[2]. Using our hand curated test set (chromosome 1), we calculated model accuracy for Vespucci with the default parameters (Max_Edge: 500 and Density_Multiplier: 10,000) relative to our HCT116 test dataset (Table 2). In addition, we performed a grid search on a subset of ranges for both Max_Edge and Density_Multiplier combinations and reported the performance of the best (Max_Edge: 10 and Density_Multiplier: 2,000) parameters obtained for this dataset.

We asses the quality of the predictions 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 all of GRO-seq data. However, most transcribed regions are transcribed by RNA polymerase II and therefore comparison to Pol II ChIP-seq should independently verify the location of transcripts. To this end, we obtained Pol II ChIP-seq data for both MCF7 and HCT116 cell lines [13, 15]. Unfortunately, comparisons between GRO-seq and ChIP-seq are complicated as GRO-seq is strand specific whereas ChIP-seq is not. Yet, we reasoned that the summation of reads along the sense and anti-sense strand should approximate ChIP-Pol II read coverage within the same region.

Thusly, an *active* call should have a higher enrichment of RNA Pol II ChIP-seq than an inactive call. In both the HCT116 and MCF7 cell lines, we made divergent transcription, *active* and *inactive* annotations. Vespucci does not contain an unbiased divergent transcription annotator, therefore only *active* and *inactive* predictions are available. For MCF7 we utilized the published list of Vespucci annotations but for HCT116 we used the Vespucci parameters via grid search (Table 2). We note that the Vespucci approach is

less capable of distinguishing *active* from *inactive* regions as assessed by Pol II occupancy(Figure 6). We observe a statistically significant enrichment (Kolmogrov-Smirnov test) for Pol II occupancy between *active* and *inactive* FStitch regions. Indeed, we observe a high degree of Pol II occupancy at divergent transcription calls.

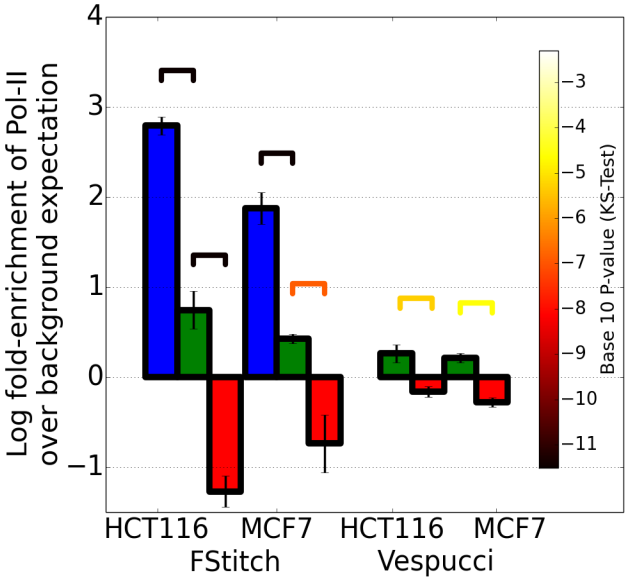

Figure 6: Correlation of GRO-seq transcript calls with Pol II ChIP-seq.
Poll-II read density was collected in regions annotated either as divergent transcription, *active* or *inactive* by either FStitch or Vespucci. Blue, green and red represent bidirectional, *active* and *inactive* calls respectively. Log fold-enrichment is relative to average Pol-II read density. Statistical significance is assessed via the Kolmogrov-Smirnov test (significance bars colored by p-value). Error bars indicate one standard deviation away from the mean.

3.3 Annotation Comparisons

We next sought to evaluate the performance of our algorithm on identifying biologically meaningful regions of active transcription by comparing the results of FStitch to RefSeq annotations. We first classify our *active* transcript calls by a wide variety of databases containing genomic annotations. Most FStitch *active* calls overlap a known annotation: gene, long non-coding RNA (lncRNA), small nucleolar RNA (snoRNA), microRNA (miRNA), transfer-RNA (tRNA) and/or Retroposon annotations (Figure 7). Interestingly, many of the miRNA and snoRNA annotations are downstream of a bidirectional transcription call (e.g. Figure 4). Of the FStitch *active* calls that do not overlap known annotations, many (most of the 57.3%) can be described as bidirectional calls that overlap an H3K27ac mark; characteristic of an eRNA.

However, 20% of these unknown *active* calls contain an open reading frame that spans 60% of the length of the call and contain a bidirectional call at the 5'-end. These may be

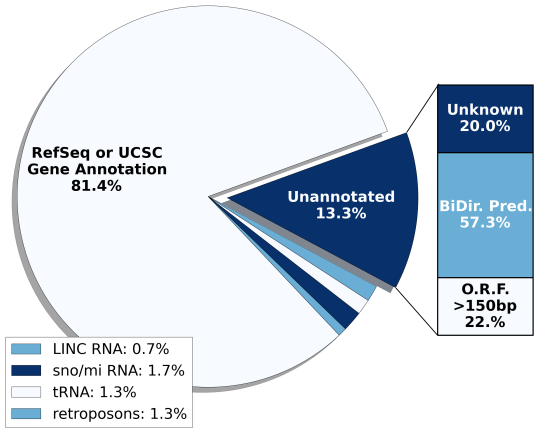

Figure 7: Active Call Characterization.
FStitch *active* calls are divided into classes based on previous genomic annotations. An *active* call is only assigned if it overlaps 95% of a previous annotation's length. Unannotated *active* calls are assigned if they overlap no previous annotations on either the positive or negative strand. Only 1% of FStitch either partially overlap a gene annotation and thus are considered neither unannotated or previously described. FStitch made 36,033 *active* annotations.

unannotated protein coding genes. We translated these regions and searched the UniProt/SwissProt database, uncovering several hits. We then isolated the statistically significant hits and tokenized the hit descriptions. More than 50% 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. It should be noted that the 3′-end of a gene annotation is the mRNA cleavage site and is not the RNA Pol-II termination site. Previous studies utilizing a Meta-gene over gene annotations detect a small peak of reads at the 3′-end in GRO-seq datasets[21]. However, this 3′ peak does not always correlate well with the exact 3′-end of the annotation likely because the annotation was not the Pol-II termination location[3]. Taking advantage of the high read coverage of the HCT116 GRO-seq dataset, we examined FStitch active calls that completely overlap a RefSeq annotation (n=2512) and averaged the read coverage within 100 uniformly distributed proportions (Figure 8) relative to the FStitch call.

This uncovered two important features of active regions: (1) the 3′-end peak is much larger than previously detected and (2) there is a corresponding small build up of reads along the anti-sense strand that mirrors the 3′-end peak. Given our algorithm does not rely on previous gene or enhancer annotations, we ask how FStitch active calls relate to known RefSeq gene annotations (Figure 9). Specifically, we measure the difference in genomic location between the end of an active call and the nearest RefSeq annotation, for both 5′ and 3′-ends. Importantly, we see a roughly 10kb elongation of GRO-seq signal past the 3′-end of annotated genes (Fig-

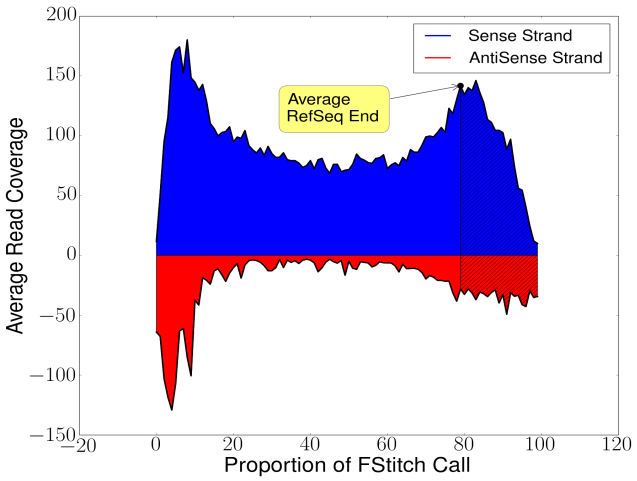

Figure 8: Average Read Coverage of FStitch active calls.
FStitch active calls on the positive strand that completely contain a Refseq annotation were used to calculate the average behavior. Blue represents positive strand coverage and red represents negative strand coverage. For each active region, read coverage was binned into 100 uniformly sized proportions.

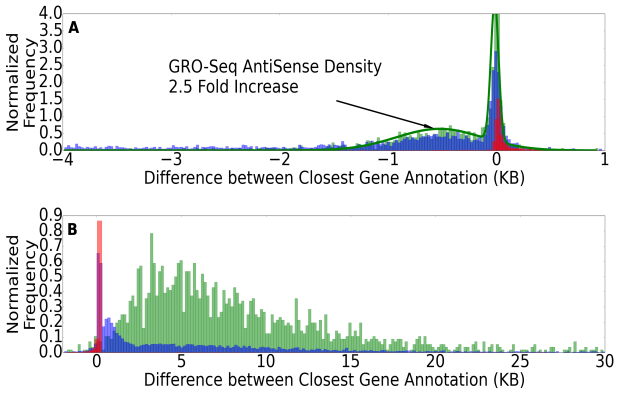

Figure 9: Histograms comparing the active region calls of FStitch to Refseq annotations.
We plot the distance between an active call and the nearest Refseq annotation for (A) 5′-ends; (B) 3′-ends; Colors *red*, *blue* and *green* are Hah, Vespucci (grid search parameters) and FStitch transcript annotations respectively.

ure 9B); consistent with the fact that polymerase proceeds far beyond the mRNA cleavage site [4].

Additionally, GRO-seq signal often begins upstream of annotated 5′ start sites of annotated Refseq genes (Figure 9A). Indeed, there appear to be two distinct populations within the 5′ starts. Therefore, we fit a mixture of two Gaussian distributions using the Expectation Maximization algorithm[20] to the 5′ histogram of active calls. We then examined the upstream Gaussian distribution for distinguishing features and found it shows a 2.5 fold enrichment of antisense transcription compared to the Gaussian centered at roughly the zero position. We suggest that many genes may have nearby overlapping upstream enhancers, which typically shows bidirectional transcription.

We then compared our correspondence to Refseq annotation to that of previously developed GRO-seq *de novo* transcript detection algorithms[11, 2]. The Vespucci algorithm, captures many of the same general trends of FStitch. But, on average, the Vespucci algorithm terminates 3′ extensions earlier than FStitch. Upon further examination, this may reflect that Vespucci's default parmeters are biased to highly expressed contigs and the 3′ extensions are often weakly transcribed. Hah's HMM was trained to match Refseq annotations and therefore is unable to identify the distinguishing features of nascent transcription at either end.

3.4 Characterizing bidirectional RNA Activity

Given that we have confidence intervals for the extent and length of overlap of bidirectional peaks many of which are eRNA events, we next sought to assess the accuracy of these predictions genome-wide. Excluding chromosome 1 (as this was our training set), we used FStitch to predict bidirectional transcription on all three cell lines: IMR90, MCF7 and HCT116. First, for each cell line we asked for the number of bidirectional peaks overlapping a DHS or H3K27ac mark (Table 3). In all cell lines, the bidirectional FStitch calls were significantly enriched, by hypergeometric test, for DNAse and H3K27ac marks indicating that a large fraction of these calls are likely eRNAs.

We hypothesize that bidirectional predictions that overlap enhancer marks will be highly transcribed, moreso than bidirectional predictions without corresponding enhancer marks (Figure 10). In all three cell lines, we see higher levels of bidirectional expression when accompanied by a chromatin enhancer mark. As proof of concept, marks which do not overlap bidirectional prediction show little read density indicating that our False-Negative rate is low. Bidirectional predictions that overlap both a gene annotations and an enhancer mark show the highest level of average expression. Moreover, we predicted 342, 241 and 198 bidirectional phenomena in the HCT116, MCF7 and IMR90 datasets, respectively, that do not overlap a chromatin enhancer mark but do show a GRO-seq expression greater than the mean GRO-seq signal of bidirectional predictions overlapping a DNase or H3K27ac mark. We believe some of these may as of yet be undiscovered enhancers.

Next, we examined the theory in gene regulation that asserts that enhancer elements are three-dimensionally connected to their gene regulatory partner. To compare GRO-seq signal with three-dimensional chromatin interactions, we utilized a chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) dataset in the HCT116 cell line[18]. ChIA-PET is an exciting new high-throughput technique that pulls down a protein of interest (in this case Pol II) and provides genomic sites on long range chromatin interactions[10]. We first confirmed overlap between FStitch *active* calls and bidi-

Table 3: 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	10,536		
DNAse	4,154†	152,768	
H3K27ac	4,554‡	13,673	32,516
HCT116			
bidirectional	14,738		
DNAse	6,750†	114,060	
H3K27ac	2,417§	13,769	57,623

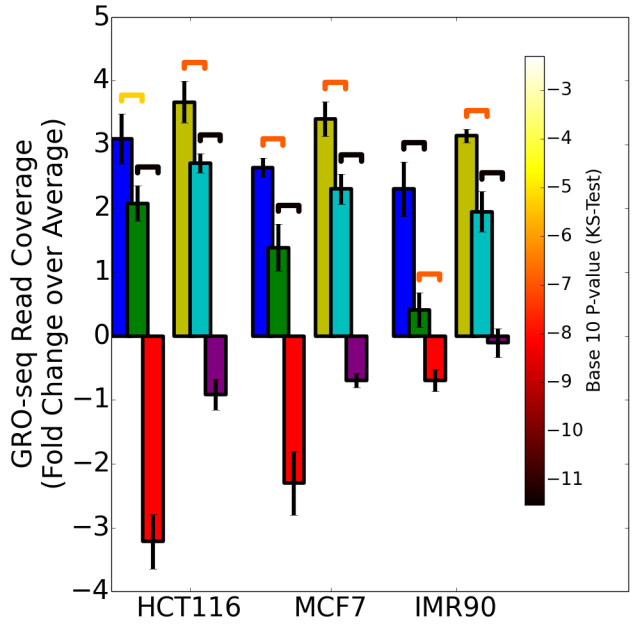

Figure 10: Bidirectional calls overlapping enhancer marks are highly transcribed

We posit that bidirectional calls that overlap enhancer marks will show strong signatures of expression. Color Description: *blue*: bidirectional Prediction (All), *green*: Active Call; *red*: Inactive Call, *army green*: bidirectional Prediction + H3K27ac + Promoter Association, *teal*: bidirectional Prediction + H3K27ac + non-promoter association, *purple* H3K27ac + no bidirectional Prediction

rectional predictions with ChIA-PET sites. We see a highly significant overlap (hypergeometric; p-value $< 10^{-10}$) between ChIA-PET calls and predictions made by FStitch.

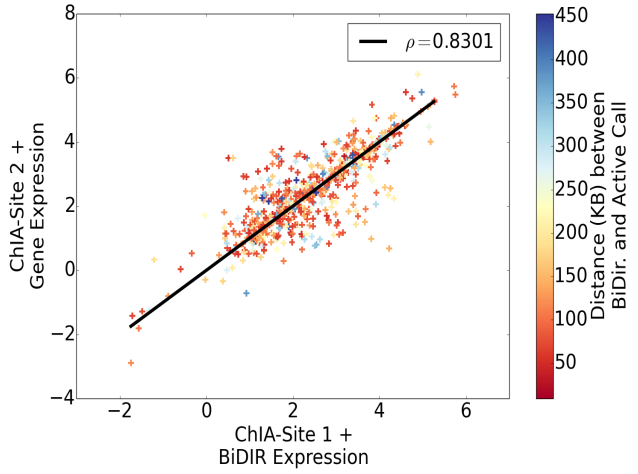

Figure 11: bidirectional predictions and *active* FStitch calls connected by a ChIA-PET call show correlated GRO-seq expression.

The GRO-seq expression level of ChIA-PET peak pairs that overlap a bidirectional Call and an *active* call on either end are plotted, demonstrating a strong correlation ($\rho = 0.8301$) in expression (as measured by GRO-seq) between the ChIA-PET pairs. Points are color according to genomic distance(KB) between bidirectional prediction and *active* call.

Given the three dimensional association implied by ChIA-PET, we next sought to ascertain if interacting chromatin sites show a correlated GRO-seq expression signal. When assaying for GRO-seq signal utilizing only ChIA-PET peak pairs, we report no correlation in expression (Pearson's correlation coefficient; ρ of 0.001). However, when we isolate ChIA-PET pairs that overlap both a bidirectional prediction and an *active* FStitch call on either end, we see a strikingly high correlation (ρ of 0.8301; Figure 11). For this analysis we do not take into account bidirectional predictions that overlap the same *active* FStitch call. Moreover, this linear relationship is completely independent of genomic distance. Figure 11 poses an obvious question: can we predict enhancer-gene interactions? Via a general linear model estimated from Figure 11, we attempted to predict enhancer-gene interactions using only GRO-seq expression level, however, only 7% of enhancer-gene interaction predictions were validated by ChIA-PET calls. This result suggests additional information is needed to create a true enhancer-to-gene predictor.

4. DISCUSSION

We present a fast and robust GRO-seq transcript annotator that is completely annotation agnostic. Parameters of the algorithm are learned from small amounts of training data and can adapt readily to even low depth of sequencing. By taking advantage of Logistic Regression we can learn a non-linear classification of the feature space. By then embedding this classifier within a Hidden Markov model framework, we are able to annotate clean, contiguous segments of active transcription. Our algorithm identifies regions of active transcription that correspond well to independently obtained

secondary datasets such as Pol II ChIP-seq and ChIA-PET. Furthermore, we can readily identify bidirectional RNAs (many likely to be enhancer RNAs) within GRO-seq data with high accuracy. FStitch is user friendly and fast, with classifications easily viewed on common genome browsers (IGV or UCSC genome browser).

In conjunction with this software release, we uncover exciting new biological phenomena. We show on a systems-wide scale that gene transcription progress much farther than the 3′-end of the mRNA cleavage site. We make many unannotated transcript discoveries that relate either to protein coding regions or unseen enhancer events. A meta-analysis of active FStitch calls shows that the 3′-end antisense and sense signal is much larger than previously appreciated. Additionally, we see a strong correlation between bidirectional predictions bearing one of the known histone marks (DNAse or H3K27ac) and higher transcription levels. We see that chromatin interactions identified by ChIA-PET are almost *always* transcribed. And most excitingly, we see that chromatin interactions that overlap a bidirectional and FStitch *active* calls are transcribed at the same level; further evidence of enhancer-to-gene interactions.

A curious user might consider extending FStitch to other high throughput read mapping datasets. As the only input to FStitch is a genome bed coverage file and training set, FStitch is not technically specific to GRO-seq data. This method may bare relevance where contiguous regions of high dense read coverage wish to be isolated; characteristic most notably in ChIP-seq Pol-II datasets. Indeed, the relevance of this algorithmic structure to ChIP-seq peak calling should be explored further.

An exciting extension of the work posed here might be the addition of an extra dimension towards the computational prediction of gene-enhancer interactions. Previous work has utilized DNA transcription factor binding motifs, chromatin marks and chromatin accessibility to predict putative enhancer events[25]. Coupled with the result that bidirectional and *active* FStitch calls correlate with three dimensional behavior of chromatin interactions, one might build a rich and interesting model to combine transcription factor binding motifs, chromatin marks and bidirectional FStitch predictions with similar GRO-seq expression profiles.

Apart from bidirectional prediction, more work is needed to better resolve the transcriptional dynamics of annotations such as the 5' and 3' peaks. The height and spread of these peaks vary from gene to gene making detection difficult. However, future work should focus on building models to better isolate this substructure and define more clearly segments within an annotated transcript. Alterations in the size and shape of the GRO-seq signal between experiments may point to distinct modes of regulation. Indeed leveraging finer substructure within GRO-seq signal may help to resolve transcriptional regulation, as genes in close proximity with relatively similar levels of expression are often grouped. The ability to isolate distinct but adjacent (or even overlapping) regions of transcription would be a powerful use of GRO-seq signal.

5. ACKNOWLEDGEMENTS

We would like to thank Aaron Odell and Josephina Hendrix for assistance with analysis of publicly available datasets. This work was funded in part by the Boettcher Foundation's Webb-Waring Biomedical Research program (RDD), a NIH training grant N 2T15 LM009451 (MAA), and an NSF IGERT 1144807 (JA). The authors acknowledge the BioFrontiers Computing Core at the University of Colorado Boulder for providing High Performance Computing resources supported by BioFrontiers IT.

6. REFERENCES

[1] M.A. Allen, Z. Andrysik, V. L. Dengler, H. S. Mellert, A. Guarnieri, J. A. Freeman, K. D. Sullivan, M. D. Galbraith, X. Luo, W. L. Kraus, R. D. Dowell, and J. M. Espinosa. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *eLife*, 3, 2014.

[2] K. A. Allison, M. U. Kaikkonen, T. Gaasterland, and C. K. Glass. Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic Acids Res.*, 42(4):2433–2447, Feb 2014.

[3] 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.

[4] 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*, 4(6), Dec 2013.

[5] 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.

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

[7] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, Dec 2008.

[8] 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.

[9] 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.

[10] 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.

[11] N. Hah, C. G. Danko, L. Core, J. J. Waterfall, A. Siepel, J. T. Lis, and W. L. Kraus. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, 145(4):622–634, May 2011.

[12] 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.

[13] 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.

[14] X. Ji, Y. Zhou, S. Pandit, J. Huang, H. Li, C.Y. Lin, R. Xiao, C.B. Burge, and X. Fu. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell*, 153(4):855–868, 2013.

[15] 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.

[16] B. Langmead. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics*, Chapter 11:Unit 11.7, Dec 2010.

[17] E. Larschan, E.P. Bishop, P.V. Kharchenko, L.J. Core, J.T. Lis, P.J. Park, and M.I. Kuroda. X chromosome dosage compensation via enhanced transcriptional elongation in drosophila. *Nature*, 471(7336):115–118, March 2011.

[18] W. Li, D. Notani, Q. Ma, B. Tanasa, E. Nunez, A. Y. Chen, D. Merkurjev, J. Zhang, K. Ohgi, X. Song, S. Oh, H. S. Kim, C. K. Glass, and M. G. Rosenfeld. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520, Jun 2013.

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

[20] 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.

[21] M. F. Melgar, F. S. Collins, and P. Sethupathy. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.*, 12(11):R113, 2011.

[22] I.M. Min, J.J. Waterfall, L.J. Core, R.J. Munroe, J. Schimenti, and J.T. Lis. Regulating rna polymerase pausing and transcription elongation in embryonic stem cells. *Genes & Development*, 25(7):742–754, 2011.

[23] 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.

[24] 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.

[25] A. Podsiado, M. Wrzesie, W. Paja, W. Rudnicki, and B. Wilczyski. Active enhancer positions can be accurately predicted from chromatin marks and collective sequence motif data. *BMC Syst Biol*, 7 Suppl 6:S16, 2013.

[26] D. Wang, I. Garcia-Bassets, C. Benner, W. Li, X. Su, Y. Zhou, J. Qiu, W. Liu, M.U. Kaikkonen, K.A. Ohgi, C.K. Glass, M.G. Rosenfeld, and X. Fu. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351):390–394, May 2011.