An Annotation Agnostic Algorithm for Detecting Nascent RNA Transcripts in GRO-Seq

Joseph G. Azofeifa, Mary A. Allen, Manuel E. Lladser, and Robin D. Dowell

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. 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, hidden Markov models and logistic regression, to classify which regions of the genome are transcribed. Given a small user-defined training set, our algorithm is accurate, robust to varying read depth, annotation agnostic, and fast. Analysis of GRO-seq data without a priori need for annotation uncovers surprising new insights into several aspects of the transcription process.

Index Terms—GRO-seq, nascent transcription, logisitic regression, hidden Markov models, algorithms, experimentation

1 Introduction

Almost all cellular stimulation triggers global transcriptional changes. To date, most studies of transcription have employed RNA-seq or microarrays, powerful measures of steady state RNA levels. Unfortunately, steady state levels can be influenced by not only transcription but also RNA stability, so these assays are not true measures of transcription. Only recently have methods for direct measurement of transcription, genome-wide, become available. A technique, known as global run-on sequencing (GRO-seq), simultaneously detects the amount and direction of actively engaged RNA polymerases at every position within the genome [1]. GRO-seq has already drastically influenced our understanding of the transcription process, as most of the genome is transcribed but rapidly degraded [2], [3], [4].

The earliest and most common approach to GRO-seq analysis is annotation centric [1], [5], [6], [7]. Yet much of transcription does not overlap protein coding annotations and appears to be noncoding [8]. In particular, one class of nascent noncoding transcripts originate from enhancers, or regulatory regions within the genome. While the ENCODE project made major inroads on identifying these critical regulatory regions [8], their precise boundaries are still difficult to ascertain, so they remain largely unannotated. The transcripts that originate from these enhancers, known as eRNAs, are unstable and lowly expressed but do appear to be critical to their regulatory activity [9], [10], [11], [12], [13]. They are detectable in GRO-seq and tend to show bidirectional transcription [14]. Therefore, the unbiased identification of all regions of transcription from GRO-seq is an important and pressing problem.

To the best of our knowledge only two efforts have attempted to identify regions of active transcription directly from GRO-seq data [15], [16], [17], though neither is fully independent of annotation. The first used a two state Hidden Markov Model (HMM) by Hah et al. that was parametrized based on available annotations [16]. This approach has the advantage of calling large contiguous regions as transcribed, but fails to call many unannotated regions because their length and transcription levels do not mimic well annotated regions. Furthermore, the approach is limited in its ability to discover transcripts that conflict with the annotation. A more recent approach, called Vespucci, uses a sliding-window (specified by two user-dependent parameters) that merges adjacent windows together based on read depth, but requires the user to tune the algorithm with each new dataset [15]. The windowing scheme, in principle, has the benefit of not depending on annotation. In practice, however, because regions of transcription are often broken into discontiguous sections, Vespucci requires the use of annotations to improve its strategy [15].

Our approach combines the strengths of these previous efforts [15], [16]. 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. This approach shares a similar structure with Maximum Entropy Markov Models [18]. Our methodology is annotation agnostic, requiring only a small number of training examples to adapt parameters to new data. It effectively identifies cohesive regions of active transcription while maintaining a rapid runtime. Furthermore, the identification of transcripts solely from the signal within

- *J.G. Azofeifa is with the Department of Computer Science, University of Colorado, Boulder, CO 80309. E-mail: joseph.azofeifa@colorado.edu.*
- *M.A. Allen is with the BioFrontiers Institute, University of Colorado, Boulder, CO 80309. E-mail: mary.a.allen@colorado.edu.*
- *M.E. Lladser is with the Department of Applied Mathematics, University of Colorado, Boulder, CO 80309. E-mail: manuel.lladser@colorado.edu.*
- *R.D. Dowell is with the Department of Molecular, Cellular, and Developmental Biology and the BioFrontiers Institute, University of Colorado, Boulder, CO 80309. E-mail: robin.dowell@colorado.edu.*

Manuscript received 22 Dec. 2014; revised 15 Dec. 2015; accepted 6 Jan. 2016. Date of publication 26 Jan. 2016; date of current version 5 Oct. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2520919

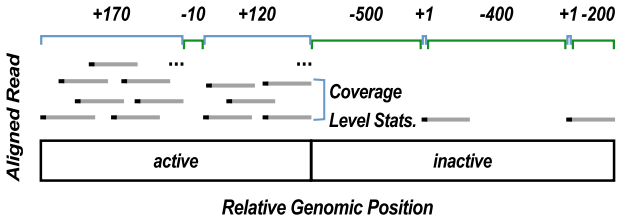

Fig. 1. *A schematic showing how contig length and coverage statistics discriminate active from inactive nascent transcription.* Regions of active transcription contain many long contigs (positive length, not drawn to scale) with significant read coverage (labeled in blue) interspersed with short regions of no coverage. Coverage statistics define mean, median, mode, and variance of reads (black bars) across a contig, see Table S1, available in the online supplemental material. In segments with no reads, a gap (labeled in green) is defined by a negative length value and all coverage statistics are set to zero. For our algorithm, reads (gray bars) are represented by only their 5′ position (black points). Therefore, a contig is also a continuous region where every base has at least one read's 5′ end at that position. Consequently, small gaps between contigs have a high probability of being in an active call.

the data uncovers distinct biological phenomena previously missed in GRO-seq analysis. Finally, user-friendliness was a large consideration in the design and structure of the software. This paper is an extended version of our earlier conference paper [19]. Here we extend upon our previous work by describing a method to compare two datasets based on the transcribed regions called by our algorithm. Using this differential transcription method, we re-analyze our earlier [20] GRO-seq dataset at both previously unannotated transcripts and annotated genes, demonstrating many of the earlier calls were annotation based artifacts. Shockingly, we demonstrate that the major response to activating p53, is increased transcription of p53's own binding site.

2 Materials and Methods

2.1 Algorithm Description

The GRO-seq technique measures nascent transcripts produced from actively engaged RNA polymerases [1]. Because splicing has not yet occurred, each transcript covers a contiguous region of the underlying genome, reflecting the extent of polymerase activity. Sequencing reads obtained from the GRO-seq protocol represent a sample from the underlying transcripts in proportion to their relative abundances. Ideally, overlapping reads could be merged into contigs, or regions of continuous read coverage, defining regions of active transcription. However, because of uneven sampling, coverage within active regions may not be contiguous. Furthermore, the sequencing and mapping process is noisy, therefore reads can also spuriously map to inactive regions.

Transcription can be modeled as a discrete time-series indexed by genomic coordinates where transcriptional activity observed at adjacent base-pairs is correlated. Similar to prior models of GRO-seq [16], 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 (Fig. 1). Each contig can be described by two feature classes: contig length (maximum length of overlapping reads) and contig coverage statistics (Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2016.2520919). 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.

We must learn the emission and transition probabilities of each state from a training set. In our case, this set corresponds to manually labeled 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. Alternative approaches to feature vector modeling, like neural networks, were considered. However, we chose to use logistic regression for three reasons: it requires little training data for parameter estimation, it quickly converges, and it readily scales with genome size. The logistic regression predictors are interpretable as probabilities, and therefore easily embedded into a HMM as emissions. After the probability transitions of the underlying Markov chain have been estimated, the well-known decoding algorithms such as Viterbi and Forward/Backward can be used to infer the most probable state sequence [18].

2.2 Datasets

This study takes advantage of three previously published GRO-seq datasets (labeled here by the underlying cell line): MCF-7 [16], IMR90 [1] and our own HCT116 (DMSO and Nutlin, wild type p53) [20], as well as three published ChIP-Pol II datasets: HCT116 [21], IMR90 [22] and MCF7 [23]. For each experiment, raw reads were mapped to the hg19 genome using Bowtie2 with the command bowtie -S -t -v 2 -best [24]. A 5′ bedgraph is then generated using BedTools's (2.16.2) genomeCoverageBed (options: -5 -bg -strand) for each strand. Additionally, the ENCODE project provided H3K27ac, H3K4me1, and DNase I hypersensitivity peak calls for IMR90 [14], [25], MCF7 [26], [27] and HCT116 [26], [28], as well as ChIA-PET peak calls for HCT116 [29]. Finally, to create a list of high confidence p53 binding sites, we combined the data from seven ChIP assays for p53 [30], [31], [32], [33] and kept only sites that were found in at least three of the seven assays.

Because most nascent transcription is unstable and therefore understudied [4], we hand annotated the entire length of chromosome 1 in our earlier HCT116 GRO-seq DMSO dataset [20] to perform k-fold cross validation. Other training datasets were considered, such as using ChromHMM or Segway calls, but we sought to capture the nuances of nascent transcription rather than the features of earlier steady state algorithms. For all testing, 95 percent of the labeled dataset was removed from training and used to assess model accuracy. To be clear, the entire labeled HCT116 training set contains 17,776 regions labeled as *active*. Based on our cross validation results, seven regions considered *active* and seven regions considered *inactive* were used for parameter estimation in both the IMR90 and MCF7 GRO-seq datasets. These training sets (with genomic coordinates and labels) are provided in Supplemental Table S3, available in the online supplemental material.

2.3 Parameter Estimation

The Markov model transition probabilities and the conditional state emission probabilities of our HMM are estimated via a user defined, labeled training set. Given that read

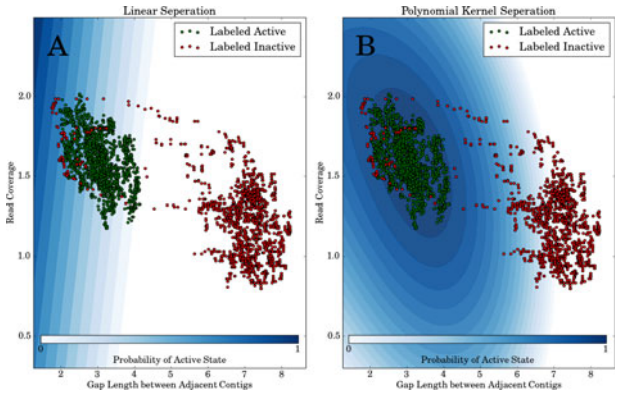

Fig. 2. *Read coverage features are not linearly separable.* Points colored green represent training examples labeled *active* and those colored red indicate training examples labeled *inactive*. The blue shading provides a contour plot of the *active* state probability given the feature's average read coverage (x_3, y-axis) and the gap length between adjacent contigs (x_1, x-axis in log nucleotides). (A) uses logistic regression with a linear kernel function (i.e., $d = 1$ in Equation (3)), whereas (B) uses a second-order polynomial kernel function (i.e., $d = 2$ in Equation (3)).

mapping can be noisy and nascent transcripts can be present at very low levels, estimating parameters that discriminate active from inactive transcription regions poses a difficult problem. However, we show in Section 3.1 that surprisingly little training data is needed to retain high model accuracy, which we define 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 [34]. We estimate the conditional probability $p(k \mid \vec{x})$, where $k \in \{inactive, active\}$ and $\vec{x}$ indicates a feature vector, via a labeled training set of defined genomic coordinates representing *active* and *inactive* transcription regions. Table S1, available in the online supplemental material, provides a complete description of the feature vector $\vec{x}$. Clearly, $p(inactive|\vec{x}) = 1 - p(active|\vec{x})$. We represent the later probability in terms of the sum of the coordinates of $\vec{x}$, weighted by some parameter vector $\vec{\theta}$. To treat this linear function as a probability, we bound the sum to the range [0,1] via the sigmoidal transformation as follows:

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

where

$$\langle \vec{x}, \vec{\theta} \rangle = \theta_0 + \sum_{i=1}^{n} x_i \cdot \theta_i, \tag{2}$$

$(n+1)$ is the dimension of the feature vector $\vec{x}$, and θ_0 is a bias term.

A simple plot of two features, gap length (x_1) and average read coverage (x_3), shows that these features may not be linearly separable (Fig. 2A). Because of this, we employ a polynomial kernel (Equation (3)) to learn non linear decision boundaries (Fig. 2B),

$$f(\vec{x}, \vec{\theta}) = \langle \vec{x}, \vec{\theta} \rangle^d + c. \tag{3}$$

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:

$$p(active \mid \vec{x}) = \frac{1}{1 + e^{-f(\vec{x}, \vec{\theta}^T)}}. \tag{4}$$

To maximize training and classification accuracy, the algorithm adjusts to the behavior of the feature space. The use of a simple second-order polynomial kernel ($d = 2$ and $c = 0$) increases the training accuracy by ~10 percent in the HCT116 GRO-seq dataset (Fig. 5). Importantly, this ~10 percent increase reflects mostly lower expressed labeled transcripts suggesting that the use of the polynomial kernel allows for greater sensitivity to under-represented, lowly transcribed regions.

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

$$l(\vec{\theta}, D) = \sum_{i=1}^{n} \log p(k_i \mid \vec{x}_i). \tag{5}$$

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

We use the Newton-Raphson algorithm [35] to iteratively update $\vec{\theta}$ until convergence. Because this techniques utilizes a second-order Taylor series approximation of the log-likelihood function, convergence is usually fast. The update rule is:

$$\vec{\theta}^{\,t+1} = \vec{\theta}^{\,t} - \left(\mathbf{H} L(\vec{\theta}, D)\right)^{-1} \cdot \nabla L(\vec{\theta}, D), \tag{6}$$

where ∇ and $\mathbf{H}$ represent the gradient and Hessian operators with respect to the vector $\vec{\theta}$, respectively. Finally, the most probable state sequence is estimated via the Viterbi Algorithm [36], using the Maximum Entropy Markov model framework [18], and is given by the recurrence relation:

$$v_t(k) = \max_{j \in S} (v_{t-1}(j) \cdot a_{j \to k}) \cdot p(k \mid \vec{x}_t), \tag{7}$$

where $a_{j \to k}$ represents the transition probability from state j to state k of the hidden Markov chain, which is estimated via Baum-Welch [18], S is the hidden transcriptional state space i.e. $S = \{active, inactive\}$, and $p(k \mid \vec{x}_t)$ is given in Equation (3) with $\vec{\theta}$ learned from the training data using the Newton-Raphson algorithm. Here $\vec{x}_t$ is either a gap or contig representation given in Table S1, available in the online supplemental material.

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

2.4 Detecting Enhancers as Divergent Transcription

Recent work indicates that enhancers are often transcribed, producing unstable bidirectional transcripts that are detectable by GRO-seq [10], [14]. Only one analysis approach has, thus far, tried to leverage this bidirectional signal towards the *de novo* discovery of enhancers from GRO-seq signal [14].

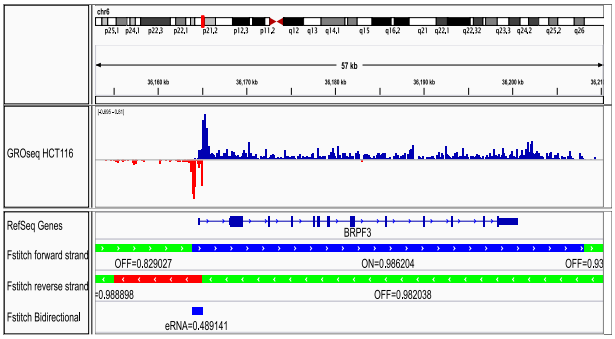

Fig. 3. *FStitch output at BRPF3.* An IGV snapshot showing a sub-region in chromosome 6 around BRPF3. The first track shows typical GRO-seq data from the HCT116 dataset, with the positive and negative strand in blue and red, respectively. RefSeq annotations are shown next. FStitch output is below for each strand with green indicating areas of *inactive* transcriptional activity, blue representing areas of *active* transcription on the positive strand and red on the negative strand. The scores associated with each classification via the Logistic Regression and Viterbi-provided Markov state sequence are also displayed. Finally, bidirectional predictions are provided at the bottom with a score via the estimated Normal Distribution confidence interval.

In that work, a Naïve Bayes classifier was trained on annotated regions in order to label unannotated 2kb windows either as bidirectional, single stranded transcription, or non-transcribed [14].

Therefore, we asked whether our FStitch approach could be extended to detect enhancer RNAs (eRNAs). Conceptually, our algorithm could simply ask for overlapping *active* calls between the positive and negative strand as potential eRNAs, similarly to the Naïve Bayes approach [14]. Unfortunately, it is unclear whether the transcripts on each strand overlap for all eRNAs as opposed to just being relatively close in proximity. Moreover, many genes have long non-coding RNA transcripts anti-sense to the gene, indicating that a simple overlap is not a stringent enough criterion for eRNA prediction. Furthermore, we expect to also detect the 5′-end of many genes because bidirectional transcription is also often observed at gene start sites [37].

Therefore, we sought to determine 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 DNase I hypersensitivity site (DHS) and a H3K27ac mark, both well known indicators of enhancer activity [27], [38]. We then computed the distance to the nearest anti-sense FStitch call. We note that the displacement data show a Normal distribution (Figure S1, available in the online supplemental material). Therefore, we make a bidirectional call when two transcripts, one on each strand, are within some number of standard deviations of the fitted Normal distribution. The confidence level of bidirectional predictions is therefore subjectively defined by the user. In our subsequent analysis, bidirectional calls utilizes a confidence interval of two standard deviations, i.e. −1.5 kb to 2.25 kb (Figure S1, available in the online supplemental material).

2.5 Algorithm Input and Output

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

FStitch outputs two bed files for positive and negative strand classifications, respectively, that can be imported into typical genome browsers such as IGV or the UCSC genome browser, to view the classifications in conjunction with read coverage files [39]. Fig. 3 shows a typical output of the algorithm. These bed files contain the genomic start and stop of each classification and an associated probabilistic score from the Viterbi algorithm (Equation (7)). From start to finish, FStitch takes ∼3.5 minutes to predict transcript annotations in the most deeply sequenced GRO-seq dataset, HCT116 (152.4 million mapped reads) [20].

2.6 Differential Transcription

One of the primary goals of many GRO-seq experiments is the identification of differentially transcribed regions between two or more conditions. As we seek to compare FStitch based differential transcription to our earlier annotation based analysis of the HCT116 dataset, we first briefly describe the experiment and its earlier analysis (see [20] for complete details). Allen et al. treated HCT116 cells with a small molecule activator of p53 known as Nutlin (or DMSO, a control) for one hour, then examined the transcriptional response by GRO-seq. Because genes are known to have a 5′ peak of read coverage that corresponds to polymerase initiation [1], [37], Allen et al. focused on differential transcription over the gene body, defined by hg19 RefSeq (downloaded Oct. 2012) annotations [40] minus the first 1 kilobase (kb). Differential transcription was determined using DESeq (v 1.4.1) [41] which runs in R (v 2.13.0) with the settings: cds < estimateSizeFactors(cds), method = 'blind', sharingMode = 'fit-only'. Genes were called as differentially transcribed if they had an adjusted p-value less than or equal to 0.1.

When using annotation, the regions of interest (typically genes) are defined a priori. Numerous methods exist for assessing the statistical significance of changes in the read depth for a given region of the genome [41], [42]. These methods are applied routinely to most short read sequencing datasets, including RNA-seq (steady state RNA measurements) and ChIP-seq. Yet, with FStitch we allow the GRO-seq data to define the regions of transcription. Given that the two experiments we wish to compare may not have precisely the same regions transcribed, the first task is to determine the coordinates of regions of interest. Intuitively we can identify three distinct means of identifying regions of interest: (1) make *active* calls in one experiment and project these coordinates to the second experiment; (2) combine the raw read data for the two experiments, make *active* calls on this joint dataset, and use the coordinates of the resulting region; or (3) make *active* calls in both experiments independently and then merge the *active* calls based on genomic

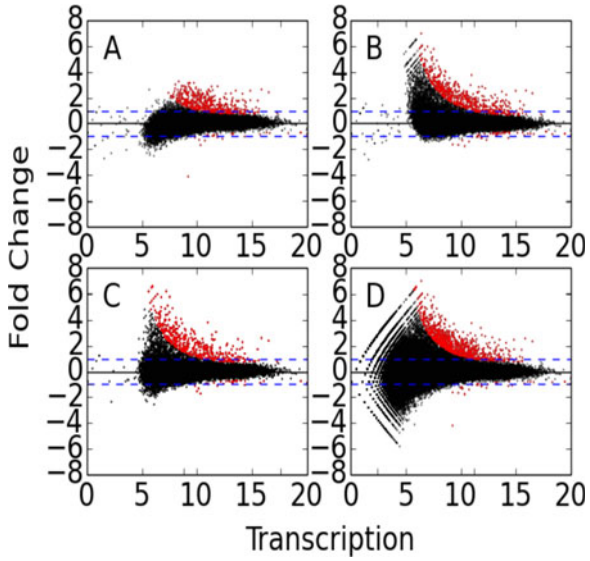

Fig. 4. *Examination of the impact of distinct approaches to identify regions of interest.* MA-plots were generated by DESeq for each of three distinct methods of determining differential transcription using FStitch *active* calls. The projection method has two variations, one starts with (A) the DMSO *active* calls and the other with (B) Nutlin *active* calls. The other methods are (C) joint and (D) the merge method. See text for details on each method.

coordinates. We refer to these options as projection, joint, or merged, respectively.

We first sought to compare the projection, joint, and merge methods of identifying regions of interest from FStitch *active* calls. For the projection method we consider both experiments as the basis for *active* calls, using only DMSO (or Nutlin) to define the regions of interest. For the joint method, we first sub-sampled the reads from the Nutlin experiment to match the depth of the DMSO experiment using samtools view (0.1.19). The Nutlin subsampled file was then combined with the DMSO reads using samtools for analysis by FStitch. Finally, for the merged method, FStitch was ran independently on the DMSO and Nutlin samples and *active* calls were combined keeping all regions called distinctly in either experiment (logical or; see Figure S2, available in the online supplemental material). Because the precise ends of an *active* call can be influenced by the read depth of the experiment, we then merged all regions smaller than 100 bp (See Figure S3, available in the online supplemental material) with an adjacent segment, unless both adjacent segments were large, meaning $>$100 bp, so as to minimizes concatenating nearby transcribed segments.

Using the same DESeq settings as Allen et al., an examination of the DESeq generated MA-plots reveals many interesting properties of each approach (Fig. 4). The projection method does not utilize all of the data to determine regions of interest which results in a bias, especially when one experiment has many more transcribed regions than the other. It is also directional and asymmetric, depending heavily on which experiment is used to define regions of interest (Figs. 4A and 4B). The joint method requires proper normalization between experiments so as to not bias the results towards the experiment with greater depth. Additionally, it forces both experiments to a common coordinate system which is problematic when the length of an *active* call changes between the experiments (Fig. 4C). Yet, a comparison of how the *active* calls shift in size between DMSO and Nutlin implies many regions change substantially (See Figure S3, available in the online supplemental material). The merged method requires a systematic means of handling arbitrarily complex overlap configurations, but has the potential to identify subregions of differential transcription. For these reasons, all subsequent analysis utilized the merge method of identifying regions of interest (Fig. 4D).

2.7 Software Availability

FStitch is written in the C/C++ programming language and complied using GNU compilers later than GCC 4.2.1. The user interface is command line, resembling many popular bioinformatics pipelines. FStitch is stand-alone and borrows from no third-party platforms, libraries or packages. The open-source software and a comprehensive manual is freely downloadable at 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) comparison to RefSeq annotation and previous methodologies, (3) validation of bidirectional predictions as enhancer RNAs, and (4) assessment of differential transcription given FStitch output.

3.1 Sensitivity to Depth of Data

To assess the sensitivity of the algorithm to the amount of training data, we hand curated the entire length of chromosome 1 in the HCT116 dataset, labeling regions as *active* or *inactive*. Our manual annotation identifies approximately 17,000 active and inactive regions, effectively labeling roughly 36 percent of chromosome 1 as active. We tested FStitch over this rich labeled data using K-fold cross validation, reserving 5 percent of the training data for parameter estimation and leveraging 95 percent for testing accuracy.

To assess the amount of training data needed for accurate classification of *active* regions, we incrementally decreased the amount of training data. Fig. 5A shows that FStitch training is robust to successive decreases in the amount of training data utilized, suggesting that very little training data is needed to achieve relatively high accuracy. The smallest training set (0.1 percent of the initial dataset) consists of three *active* and two *inactive* regions and maintains scores of 95 percent true positive and 4.3 percent false negative on the testing dataset. Furthermore, we observe that the polynomial kernel consistently outperforms the linear kernel.

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

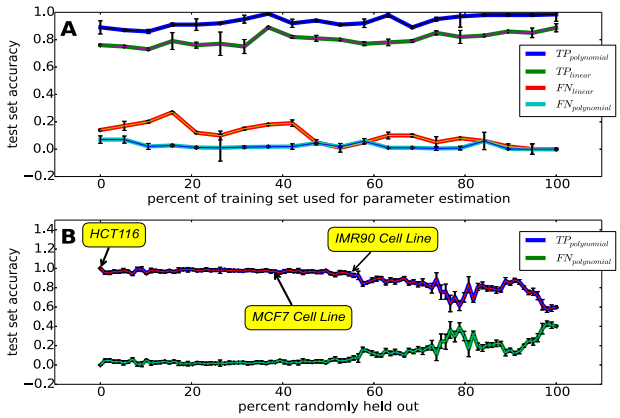

Fig. 5. *FStitch requires little training data and is robust to low levels of GRO-seq read coverage.* (A) Classification accuracy utilizing successively decreasing amounts of training data to learn feature vector weights, for the polynomial ($d = 2$ and $c = 0$; blue and teal) and linear ($d = 1$ and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

3.2 Benchmarking FStitch and Vespucci

We sought to evaluate our algorithm, FStitch, to the previously published windowing method Vespucci [15]. We calculated model accuracy for Vespucci with the default parameters over the HCT116 test dataset (Table 1). 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 parameters obtained for this dataset. Grid search optimization greatly increased Vespucci's precision and recall. FStitch outperforms Vespucci, default or grid search, in both true negative and true positive classifications.

We next assessed the quality of FStitch *active* calls to independently derived relevant biological datasets. As GRO-seq measures all actively engaged polymerase, in a strand specific fashion, there is no single alternative experiment to confirm GRO-seq data. However, RNA polymerase II is responsible for most transcribed regions and therefore comparison to Pol II chromatin immunoprecipitation (ChIP) should independently verify the location of most transcripts. To this end, we obtained previously published Pol II ChIP-seq data for MCF7, HCT116, and IMR90 cell lines [21], [22], [23]. Unfortunately, direct comparisons between GRO-seq and ChIP-seq are complicated by the fact that GRO-seq is strand specific whereas ChIP-seq is not. Yet, we reasoned that the superposition of reads along the sense and anti-sense strand within GRO-seq should approximate ChIP-Pol II read coverage within the same region.

TABLE 1
Benchmarking FStitch and Vespucci

Method	Prediction	Truth Set Label	
		Active	Inactive
FStitch	Active	98.5 percent	1.5 percent
	Inactive	0.01 percent	99.99 percent
Vespucci (default)	Active	60.7 percent	30.3 percent
	Inactive	6.03 percent	93.97 percent
Vespucci (G.S.)	Active	80.1 percent	19.9 percent
	Inactive	0.56 percent	99.44 percent

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

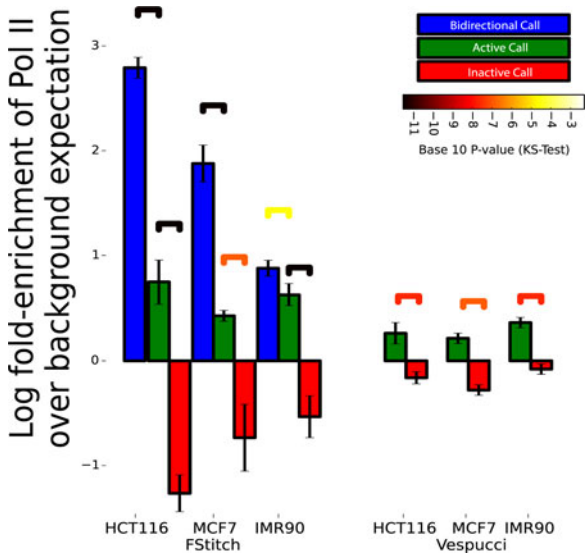

Fig. 6. *Correlation of GRO-seq transcript calls with Pol II ChIP-seq.* Pol II ChIP-seq read density was collected in regions labeled as bidirectional (blue), *active* (green) or *inactive* (red) by either FStitch (on left) or Vespucci (on right). Log fold-enrichment is relative to average Pol II ChIP-seq read density. Statistical significance is assessed via the Kolmogrov-Smirnov test (significance bars colored by p-value). Error bars indicate one standard deviation away from the mean.

Thus, an *active* call should have a higher enrichment of RNA Pol II ChIP-seq than an *inactive* call. In all three cell lines, we used FStitch to identify bidirectional, *active* and *inactive* calls. Vespucci does not contain an unbiased bidirectional transcription annotator, therefore only *active* and *inactive* predictions were obtained. For MCF7 we utilized the published list of Vespucci annotations but for both HCT116 and IMR90 we used the Vespucci parameters obtained via grid search (Table 1). We note that the Vespucci approach is less capable of distinguishing *active* from *inactive* regions as assessed by Pol II occupancy (Fig. 6). We observe a significant enrichment for Pol II occupancy between *active* and *inactive* FStitch regions. Additionally, we observe a high degree of Pol II occupancy at bidirectional calls, as expected given that enhancers are known to show significant enrichment for Pol II occupancy [9].

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 classified our *active* transcript calls on the HCT116 DMSO experiment by their overlap to genomic annotations. Most FStitch *active* calls overlap a known annotation: gene, antisense to a gene, long non-coding RNA (lncRNA), small nucleolar RNA (snoRNA), microRNA (miRNA) and transfer-RNA (tRNA) (Fig. 7). Of the 26.75 percent of FStitch *active* calls that do not overlap known annotations, many can be described as bidirectional

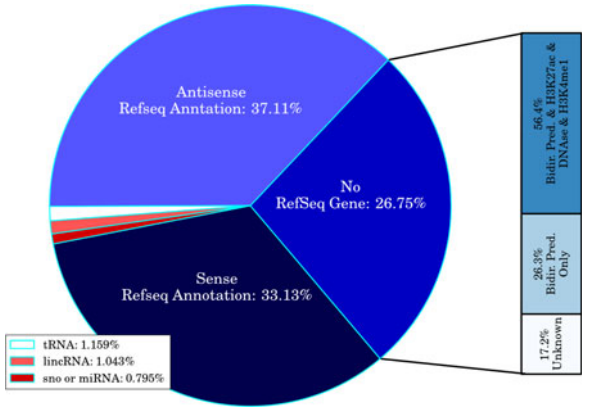

Fig. 7. *Active call characterization.* FStitch *active* calls on HCT116 DMSO are divided into classes based on overlap with genomic annotations. Unannotated *active* calls are assigned if they have no overlap to previous annotations on either strand. FStitch called 37,591 *active* regions.

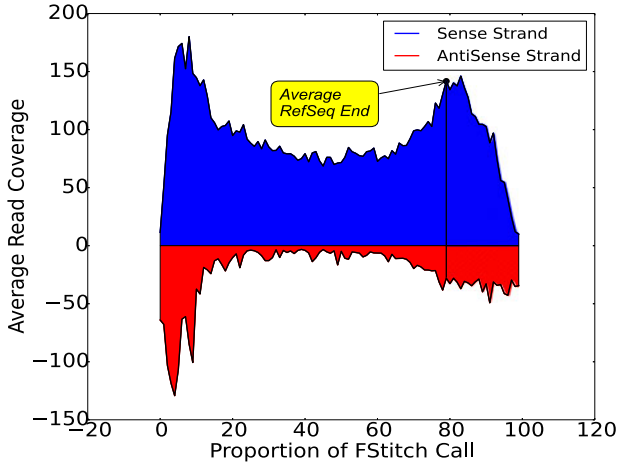

Fig. 8. *Average read coverage of FStitch active calls.* FStitch *active* calls on the positive strand that completely contain a RefSeq annotation were used to calculate the average behavior. Blue and red represent positive and negative strand coverage, respectively. For each *active* region, the length was divided into 100 uniformly sized proportions and the read coverage was averaged within each bin. The average annotated 3′ end is noted by the line and transcription beyond the annotation is shaded. Here, we require an FStitch to completely overlap a RefSeq annotation and the RefSeq annotation overlap at least 75 percent of the FStitch call.

calls that overlap an H3K27ac mark; which is characteristic of an eRNA.

Interestingly, within the unannotated *active* calls, a small fraction (9 percent) contain both an open reading frame that spans at least 60 percent of the length of the call and a bidirectional call at the 5′-end. These may be unannotated protein coding genes. We translated these regions and searched the UniProt/SwissProt protein database [43], uncovering several hits. By isolating the statistically significant hits and tokenizing the hit descriptions, we observed that more than 95 percent 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 [44]. By taking advantage of the high read coverage of the HCT116 GRO-seq dataset, we constructed a meta-gene of FStitch active calls that completely overlap a RefSeq annotation ($n = 2512$). For this analysis, we averaged the read coverage within 100 uniformly distributed proportions relative to the FStitch call (Fig. 8). This uncovered two features of active regions: (1) the 3′-end peak is much larger than previously detected [1], [14] and (2) there is a corresponding small build up of reads along the anti-sense strand that mirrors the 3′-end peak. It should be noted that the 3′ peak does not always correlate well with the exact 3′-end of the annotation [45]. This is likely because the 3′-end of a gene annotation is typically the mRNA cleavage site and not the RNA Pol II termination site.

Given that FStitch does not rely on previous annotations, we next ask how the ends (5′ and 3′) of FStitch active calls relate to known RefSeq gene annotation ends. Specifically, we measure the difference in genomic location between the 5′ end (3′ end) of an FStitch *active* call and the nearest RefSeq annotation 5′ end (3′ end), respectively. Interestingly, the GRO-seq signal often begins upstream of the annotated 5′ start site of RefSeq genes (Fig. 9A). Indeed, there appears to be two distinct populations within the 5′ ends. Therefore, we fit a mixture of two Gaussian distributions using the Expectation Maximization algorithm [46] to the difference of 5′ ends histogram. We examined the upstream Gaussian distribution for distinguishing features and found it shows a 2.5 fold enrichment of anti-sense transcription compared to the Gaussian centered at roughly the zero position. This suggests that many genes have upstream bidirectional transcription, and therefore may may have overlapping or adjacent upstream enhancers [38] or promoter upstream transcripts [47]. We note that, in these cases, the upstream region and the annotated gene are a single *active* call.

Additionally, we also see an elongation of several kilobases (average of ~8 kb) of GRO-seq signal past the 3′-end of annotated genes (Fig. 9B); consistent with the fact that polymerase proceeds far beyond the mRNA cleavage site [45], [48]. Notably, the 3′ extension is missed by earlier GRO-seq de novo transcript detection algorithms [15], [16]. Indeed, Vespucci captures many of the same general trends of FStitch, but typically terminates 3′ extensions earlier. Upon further examination, this may reflect the fact that Vespucci's default parameters are biased to highly expressed regions and the 3′ extensions are often weakly transcribed. On the other hand, the hidden Markov model of Hah et al. was trained to match RefSeq annotations and is therefore unable

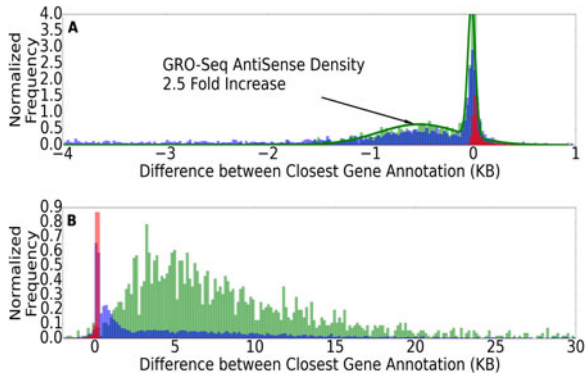

Fig. 9. *Histograms comparing the active region calls of FStitch to RefSeq annotations.* We plot the distance between the end of an *active* call and the nearest RefSeq annotation for (A) 5′-ends; (B) 3′-ends. Colors *red*, *blue* and *green* are Hah et al., Vespucci (grid search parameters) and FStitch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

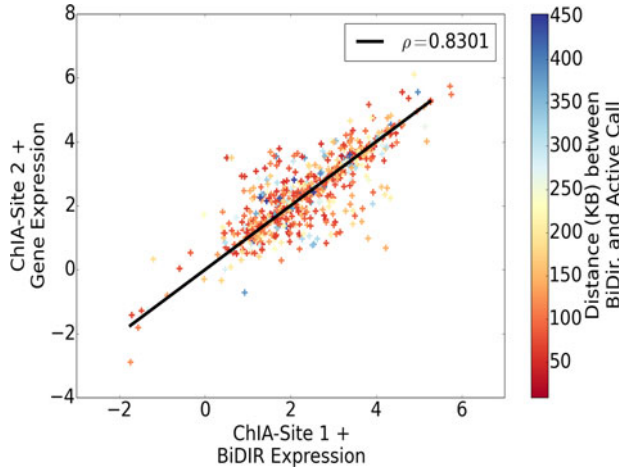

Fig. 10. *Bidirectional predictions and active FStitch calls connected by a ChIA-PET read pair show correlated GRO-seq transcription.* The GRO-seq transcription level of ChIA-PET read pairs that overlap a bidirectional call and an *active* call on either end are plotted, demonstrating a strong correlation ($\rho = 0.8301$) in transcription (as measured by GRO-seq). Points are colored according to genomic distance (kb) between bidirectional prediction and *active* call.

to identify distinguishing features of nascent transcription at either end.

3.4 Characterizing Bidirectional RNA Activity

We next sought to assess the accuracy of our bidirectional predictions genome-wide. As our goal is the identification of eRNAs, we first examined what fraction of our bidirectional calls overlap enhancer marks. For this analysis we excluded chromosome 1 (our training set) and used FStitch to predict bidirectional transcription in all three cell lines: IMR90, MCF7 and HCT116. In all cell lines, the bidirectional FStitch calls were significantly enriched for overlapping DNase I hypersensitivity sites and H3K27ac marks indicating that a large fraction of these calls are likely eRNAs (Table S2, available in the online supplemental material).

We hypothesized that bidirectional predictions that overlap enhancer marks will be highly transcribed, moreso than bidirectional predictions without corresponding enhancer marks (Figure S4, available in the online supplemental material). In all three cell lines, we see higher levels of bidirectional transcription 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 (Figure S4, available in the online supplemental material, in red). Bidirectional predictions that overlap both a gene annotation and an enhancer mark show the highest level of average transcription. 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 transcription greater than the mean GRO-seq signal of bidirectional predictions overlapping a DNase I hypersensitivity site or H3K27ac mark. These highly expressed bidirectional regions may be, as of yet, undiscovered enhancers.

Next, we examined the theory that enhancer elements are three-dimensionally connected to their gene regulatory partner, an interaction that correlates with enhancer function [9], [10], [11], [12], [13]. To compare GRO-seq signal with three-dimensional chromatin interactions, we utilized a Pol II chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) dataset in the HCT116 cell line [11]. ChIA-PET is a rather new high-throughput technique that pulls down a protein of interest (in this case Pol II) and provides information on long range chromatin interactions [29] associated with the protein. Therefore, we first examined the overlap between both FStitch *active* calls and bidirectional predictions with paired ChIA-PET reads. We see a highly significant overlap (hypergeometric; p-value $< 10^{-10}$) between ChIA-PET reads and FStitch *active* calls.

Given the three dimensional association implied by ChIA-PET, we next sought to ascertain if interacting DNA regions show a correlated GRO-seq transcription signal. When assaying for GRO-seq signal utilizing only ChIA-PET read pairs, we found no correlation in transcription level (Pearson's correlation coefficient; $\rho = 0.001$). However, when we isolate ChIA-PET read pairs that overlap both a bidirectional prediction and an *active* FStitch call on either end, we see a strikingly high correlation ($\rho = 0.8301$; Fig. 10). Note that we do not include cases where the ChIA-PET read pairs overlap the same FStitch *active* call used to make the bidirectional prediction. Moreover, this linear relationship appears completely independent of genomic distance. This poses an obvious question: can we predict enhancer-gene interactions? Using a general linear model estimated from Fig. 10, we attempted to predict enhancer-gene interactions using only GRO-seq transcription level. Unfortunately, only 7 percent of enhancer-gene interaction predictions were validated by ChIA-PET read pairs. This result suggests that while GRO-seq signal appears highly correlated between enhancers and their gene targets, additional information is needed to predict which enhancers are associated in three dimensions with particular FStitch *active* calls.

3.5 Differential Transcription at Annotated Genes: A Comparison of FStitch to Allen et al.

Finally, we sought to determine the extent to which an annotation agnostic approach (FStitch) alters our earlier annotation driven p53 GRO-seq data analysis [20]. In our earlier work we examined the direct transcriptional targets of the transcription factor p53 in HCT116 cells. In that experiment, p53 was activated by the non-genotoxic drug Nutlin (see [20] for complete details). Analysis was annotation centric but excluded the first 1 kb around the annotated start to avoid the initiation peak of polymerase. Furthermore, assessment of transcription over p53 binding sites was dependent on publicly available p53 ChIP-seq data.

We ran FStitch on the control GRO-seq (DMSO) and the p53 activated GRO-seq (Nutlin) independently. In total we found 37,591 *active* calls in DMSO and 39,097 *active* calls in the Nutlin treated sample. Many *active* calls in both DMSO and Nutlin overlap RefSeq annotated genes (annotation overlap for DMSO shown in Fig. 7). In total, 16,191 (of 23,669) genes are transcribed in at least one of the two experiments. Interestingly four large genes called as differentially transcribed in Allen et al. are not called as *active* by FStitch in either experiment. These genes appear to contain only scattered background reads (noise), but because of their massive size still contain a large total number of reads. The merged method was then used to identify regions of interest for assessing differential transcription between DMSO and Nutlin.

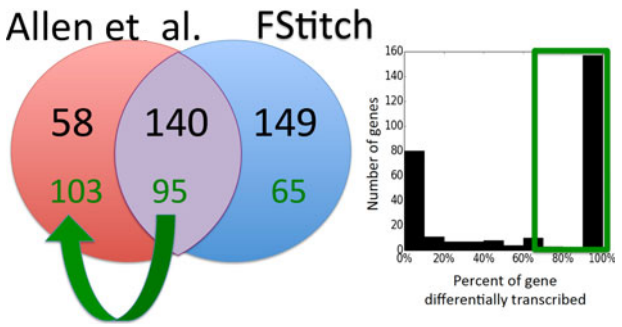

Fig. 11. *The overlap between FStitch and Allen et al. at RefSeq genes.* The gene sets called as differentially transcribed by the two methods, Allen et al. (red) and FStitch (blue), are compared at gene annotations (black numbers). The histogram on the right shows the percentage of each annotated region that is called as differently transcribed by FStitch. When the overlap to a gene is required to be $>$ 75 percent (green box), 129 genes are no longer called as differentially transcribed by FStitch, including 45 genes that were previously called by both methods.

First we sought to examine the impact of the two distinct methods of determining differential transcription, namely FStitch *active* regions versus Allen et al., at annotated genes. It is worth noting that DESeq is sensitive to the size of the input set (both in multiple hypothesis test correction and its variance estimate). Therefore to match the analysis of Allen et al., we first examined only the set of FStitch *active* regions of interest that overlap annotated genes. With this set as input to DESeq, 293 regions are differentially transcribed, overlapping 289 distinct genes (Fig. 11).

By manual inspection, we noted that many FStitch regions of interest were much shorter than the annotated gene. Therefore we next required that for each gene at least 75 percent of the gene be called as differentially transcribed. From this we conclude that many genes, including 45 called in Allen et al., do not show differential transcription along the full length of the gene. For example, PVRL4 (Fig. 12) was called as differentially transcribed in Allen et al. yet FStitch identifies that the signal for differential transcription is entirely driven by a distinct small subregion within the gene. Most of these differentially transcribed regions overlap FStitch bidirectional calls, implying that the annotation

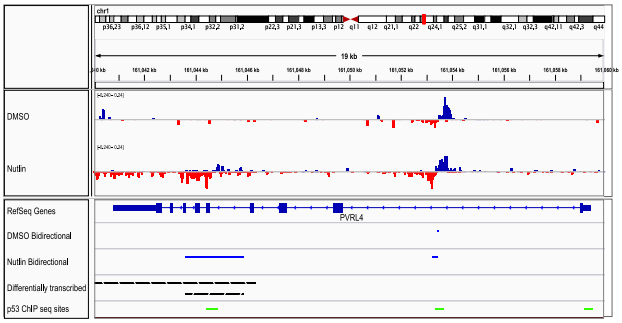

Fig. 12. *Differential transcription at PVPR4.* An IGV snapshot showing PVPR4, a negative strand gene where a small portion of the gene is differentially transcribed. The region of differential transcription (black bars) overlaps both FStitch bidirectional calls (blue bars) and p53 binding sites (green bars), indicating this may be an intragenic enhancer. The tracks, in order, are: histograms of the GRO-seq signal observed in DMSO and Nutlin, respectively (positive strand: blue; negative strand: red); RefSeq annotation for PVPR4; FStitch bidirectional calls in both DMSO and Nutlin, respectively (blue bars); FStitch differential transcription calls (black bars: top is negative strand, bottom is positive strand); location of p53 binding events (in green).

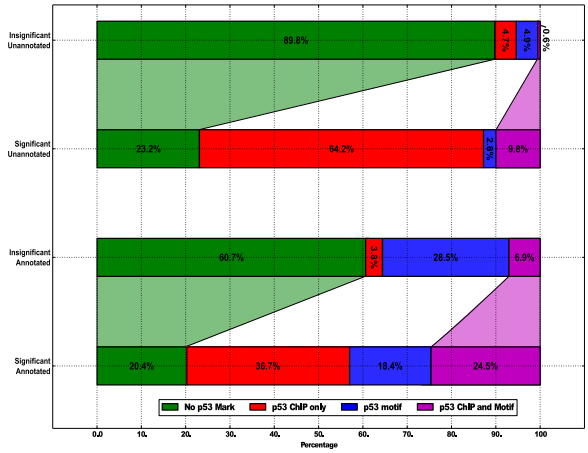

Fig. 13. *Overlap of differential transcription and p53 marks.* FStitch calls were grouped by significance of differential transcription (significant: DESeq adj. p-value $<$ 0.1) and overlap with a RefSeq annotation. From top to bottom, there are 64,899 regions without differential transcription (insignificant) and without overlapping annotation (unannotated); 782 significant-unannotated; 23,986 insignificant-annotated; and 262 significant-annotated, respectively. p53 binding site (ChIP) overlap and p53 motif presence are assessed as described in the text.

centric method was sometimes mislead by overlapping, fully contained enhancer(s).

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

3.6 Differential Transcription Using all FStitch *Active* Calls

Importantly, FStitch is able to identify unannotated regions that are differentially transcribed. When DESeq considers all FStitch regions of interest, 1,044 regions are called as differentially transcribed. Remarkably 75 percent of these regions do not overlap an annotated gene.

Because Allen et al. found differential transcription at p53 binding events, we hypothesize that a large fraction of the unannotated FStitch differentially transcribed regions would contain p53 binding events and/or p53 sequence motifs. Binding events for p53 were called as described in Allen et al., except requiring consensus from three of the seven publicly available p53 ChIP datasets [20], [31]. Presence of the motif was determined by the publicly available p53 scanner algorithm, requiring a p-value $<$ 0.01 [32]. Differentially transcribed regions, both those overlapping annotated and unannotated regions, are highly enriched for marks of p53 (either binding or motif) See Fig. 13. We note

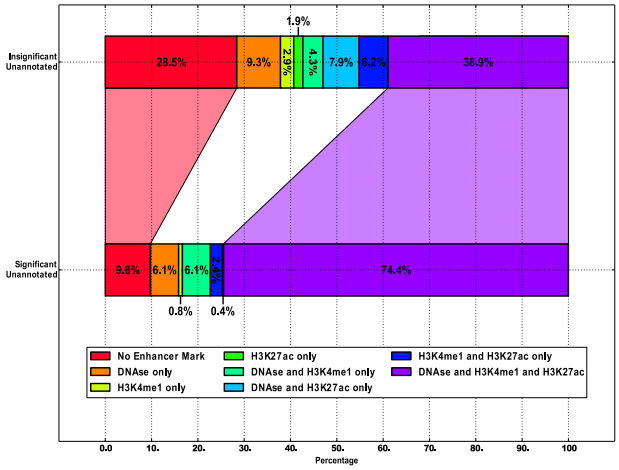

Fig. 14. *Overlap of differential transcription with enhancer marks.* FStitch calls that do not overlap any RefSeq annotation were grouped by differential transcription by DESeq (significant: adj p-value < 0.1). Regions were assessed for overlap with enhancer marks: H3k27ac, H3K4me1, and DNAse I hypersensitivity [26], [28].

that because annotated regions tend to be much longer than unannotated, they are more likely to contain a p53 motif and/or ChIP site. In fact, most regions that are differentially transcribed (73 percent) overlap an experimentally determined p53 binding event.

Lastly, we sought to determine which unannotated FStitch differential transcription calls are themselves p53 *enhancers*. To this end, we examined their overlap with known enhancer marks H3K27ac, H3K4me1 and DNAse I hypersensitivity (Fig. 14). Unannotated differentially transcribed FStitch calls are over enriched for enhancer marks, relative to background expectation. Indeed, the three enhancer marks (H3K27ac, H3K4me1 and DNAse I hypersensitivity) are more likely to co-occur in the differentially transcribed set. Interestingly, we also note that 21.2 percent of these regions are paired with another differentially transcribed FStitch call in the HCT116 ChIA-PET study. This overlap far exceeds the expectation (0.01 percent) that a random FStitch call will pair with a differentially transcribed partner by ChIA-PET.

4 Discussion

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

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

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

More work is needed to better resolve the transcriptional dynamics observed within genes, such as the 5′ and 3′ peaks. These peaks are reminiscent of patterns seen in unstranded Pol II ChIP data and likely correspond to distinct stages of RNA polymerase activity [49]. Unfortunately, the height and spread of these peaks vary from gene to gene, making their precise detection difficult. However, it may be possible to build models that can more clearly isolate this substructure within an annotated transcript. In fact, 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 could help to resolve distinct biological transcripts within *active* calls. The ability to isolate distinct but adjacent (or even overlapping) regions of transcription would be a powerful use of GRO-seq signal.

Our work demonstrates that GRO-seq is a rich and under-utilized source of insights into transcription and its regulation. Sites of bidirectional transcription are readily identified within GRO-seq data with high accuracy. These bidirectional predictions correlate strongly with known enhancer marks, implying that many are eRNAs. In fact, the single largest class of transcripts that respond (i.e. show differential transcription) when p53 is activated are bidirectional RNAs. Most of these RNAs contain p53 signals, either binding by ChIP or enrichment for the sequence motif. Interestingly, some of these differentially transcribed enhancers are intragenic, potentially confounding studies that depend on the underlying annotation.

Furthermore, when bidirectional predictions and a separate FStitch *active* call overlap chromatin interaction calls (by ChIA-PET), the two regions are transcribed at the same level; further evidence of enhancer-to-gene interaction. This finding is consistent with ENCODE reporting strong correlations between the presence of an enhancer RNA, gene expression, and promoter-enhancer interations [50], [51]. More interesting is the observation that differentially transcribed FStitch calls are three dimensionally connected via ChIA-PET to another differentially transcribed FStitch call. It remains to be seen if bidirectional FStitch predictions with similar GRO-seq transcription profiles could be combined with relevant additional information such as transcription factor binding motifs or chromatin marks to create a rich model for predicting enhancer-to-gene interactions.

It should be noted that because the only input to FStitch is a genome bedgraph file and a training set, FStitch is not technically specific to GRO-seq data. This method may bare

relevance in any experiment where contiguous regions of dense read coverage wish to be isolated; a characteristic most notably present in Pol II ChIP-seq datasets. Indeed, the relevance of this algorithmic structure to ChIP-seq peak calling should be explored further.

Acknowledgments

The authors 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 NSF ABI DBI-12624L0 (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 (NIH 1S10OD012300) supported by BioFrontiers IT.

References

[1] L. J. Core, J. J. Waterfall, and J. T. Lis, "Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters," *Science*, vol. 322, no. 5909, pp. 1845–1848, Dec. 2008.

[2] P. Kapranov, A. T. Willingham, and T. R. Gingeras, "Genome-wide transcription and the implications for genomic organization," *Nat. Rev. Genet.*, vol. 8, no. 6, pp. 413–423, Jun. 2007.

[3] B. Neymotin, R. Athanasiadou, and D. Gresham, "Determination of in vivo rna kinetics using Rate-seq," *RNA*, vol. 20, no. 10, pp. 1645–1652, Oct. 2014.

[4] C. G. Danko, S. L. Hyland, L. J. Core, A. L. Martins, C. T. Waters, H. W. Lee, V. G. Cheung, W. L. Kraus, J. T. Lis, and A. Siepel, "Identification of active transcriptional regulatory elements from GRO-seq data," *Nat. Meth.*, vol. 12, no. 5, pp. 433–438, May 2015.

[5] I. Min, J. Waterfall, L. Core, R. Munroe, J. Schimenti, and J. Lis, "Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells," *Genes Develop.*, vol. 25, no. 7, pp. 742–754, 2011.

[6] E. Larschan, E. Bishop, P. Kharchenko, L. Core, J. Lis, P. Park, and M. Kuroda, "X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila," *Nature*, vol. 471, no. 7336, pp. 115–118, Mar. 2011.

[7] X. Ji, Y. Zhou, S. Pandit, J. Huang, H. Li, C. Lin, R. Xiao, C. Burge, and X. Fu, "SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase," *Cell*, vol. 153, no. 4, pp. 855–868, 2013.

[8] T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[9] T. Kim, M. Hemberg, J. Gray, A. Costa, D. Bear, J. Wu, D. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. Worley, G. Kreiman, and M. Greenberg, "Widespread transcription at neuronal activity-regulated enhancers," *Nature*, vol. 465, no. 7295, pp. 182–187, May 2010.

[10] D. Wang, I. Garcia-Bassets, C. Benner, W. Li, X. Su, Y. Zhou, J. Qiu, W. Liu, M. Kaikkonen, K. Ohgi, C. Glass, M. Rosenfeld, and X. Fu, "Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA," *Nature*, vol. 474, no. 7351, pp. 390–394, May 2011.

[11] 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*, vol. 498, no. 7455, pp. 516–520, Jun. 2013.

[12] C. Melo, J. Drost, P. Wijchers, H. van de Werken, E. de Wit, J. Vrielink, R. Elkon, S. Melo, N. Léveillé, R. Kalluri, W. de Laat, and R. Agami, "eRNAs are required for P53-dependent enhancer activity and gene transcription," *Molecular Cell*, no. 3, pp. 524–535, Dec. 2013.

[13] N. Hah, S. Murakami, A. Nagari, C. Danko, and W. Kraus, "Enhancer transcripts mark active estrogen receptor binding sites," *Genome Res.*, vol. 23, no. 8, pp. 1210–1223, 2013.

[14] M. F. Melgar, F. S. Collins, and P. Sethupathy, "Discovery of active enhancers through bidirectional expression of short transcripts," *Genome Biol.*, vol. 12, no. 11, p. R113, 2011.

[15] 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.*, vol. 42, no. 4, pp. 2433–2447, Feb. 2014.

[16] 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*, vol. 145, no. 4, pp. 622–634, May 2011.

[17] M. Chae, C. Danko, and W. Kraus, "Grohmm: A computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data," *BMC Bioinf.*, vol. 16, no. 1, p. 222, 2015.

[18] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proc. 17th Int. Conf. Mach. Learning*, 2000, pp. 591–598.

[19] J. Azofeifa, M. A. Allen, M. E. Lladser, and R. Dowell, "FStitch: A fast and simple algorithm for detecting nascent rna transcripts," in *Proc. 5th ACM Conf. Bioinf., Comput. Biol., Health Inform.*, 2014, pp. 174–183.

[20] M. 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*, vol. 3, p. e02200, 2014.

[21] 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*, vol. 51, no. 4, pp. 493–505, Aug. 2013.

[22] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren, "A high-resolution map of the three-dimensional chromatin interactome in human cells," *Nature*, vol. 503, no. 7475, pp. 290–294, Nov. 2013.

[23] 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.*, vol. 6, p. 456, Dec. 2010.

[24] B. Langmead, "Aligning short sequencing reads with Bowtie," *Curr Protoc Bioinformat.*, vol. Chapter 11, p. Unit 11.7, Dec. 2010.

[25] L. H. Chadwick, "The NIH roadmap epigenomics program data resource," *Epigenomics*, vol. 4, no. 3, pp. 317–324, Jun. 2012.

[26] 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 TCF 7L2 can be tethered to the genome by association with GATA 3," *Genome Biol.*, vol. 13, no. 9, p. R52, 2012.

[27] 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.*, vol. 22, no. 6, pp. 1015–1025, Jun. 2012.

[28] 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*, vol. 98, no. 4, pp. 280–287, Oct. 2011.

[29] 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*, vol. 462, no. 7269, pp. 58–64, Nov. 2009.

[30] C. Wei, Q. Wu, V. Vega, K. Chiu, P. Ng, T. Zhang, A. Shahab, H. Yong, Y. Fu, Z. Weng, J. Liu, X. Zhao, J. Chew, Y. Lee, V. Kuznetsov, W. Sung, L. Miller, B. Lim, E. Liu, Q. Yu, H. Ng, and Y. Ruan, "A global map of p53 transcription-factor binding sites in the human genome," *Cell*, vol. 124, no. 1, pp. 207–219, Dec. 2006.

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

[32] L. Smeenk, S. van Heeringen, M. Koeppel, M. van Driel, S. Bartels, R. Akkers, S. Denissov, H. Stunnenberg, and M. Lohrum, "Characterization of Genome-wide P53-binding sites upon stress response," *Nucleic Acids Res.*, vol. 36, no. 11, pp. 3639–3654, 2008.
[33] L. Smeenk, S. van Heeringen, M. Koeppel, B. Gilbert, E. Janssen-Megens, H. Stunnenberg, and M. Lohrum, "Role of p53 serine 46 in p53 target gene regulation," *PLoS ONE*, vol. 6, no. 3, p. e17574, Mar. 2011.
[34] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, nos. 5/6, pp. 352–359, 2002.
[35] N. Bouguila and D. Ziou, "A hybrid SEM algorithm for High-dimensional unsupervised learning using a finite generalized Dirichlet mixture," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2657–2668, Sep. 2006.
[36] 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.*, vol. 8, no. 2, pp. 194–204, Mar. 1997.
[37] A. Seila, J. Calabrese, S. Levine, G. Yeo, P. Rahl, R. Flynn, R. Young, and P. Sharp, "Divergent transcription from active promoters," *Science*, vol. 322, no. 5909, pp. 1849–1851, 2008.
[38] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: From properties to genome-wide predictions," *Nat. Rev. Genet.*, vol. 15, pp. 272–286, Mar. 2014.
[39] H. Thorvaldsdttir, J. Robinson, and J. Mesirov, "Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration," *Briefings Bioinf.*, vol. 14, no. 2, pp. 178–192, 2013.
[40] K. Pruitt, G. Brown, S. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. Farrell, J. Hart, M. Landrum, K. McGarvey, M. Murphy, N. OLeary, S. Pujar, B. Rajput, S. Rangwala, L. Riddick, A. Shkeda, H. Sun, P. Tamez, R. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. Maglott, T. Murphy, and J. Ostell, "Refseq: An update on mammalian reference sequences," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D756–D763, 2014.
[41] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol.*, vol. 11, no. 10, p. R106, 2010.
[42] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. Mason, N. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biol.*, vol. 14, no. 9, p. R95, 2013.
[43] T. U. Consortium, "UniProt: A hub for protein information," *Nucleic Acids Res.*, vol. 43, pp. D204–D212, 2014.
[44] P. Tamayo, D. Scanfeld, B. Ebert, M. Gillette, C. Roberts, and J. Mesirov, "Metagene projection for cross-platform, cross-species characterization of global transcriptional states," *Proc. Nat. Acad. Sci.*, vol. 104, no. 14, pp. 5959–5964, 2007.
[45] 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*, vol. 4, no. 1, pp. 7–12, 2013.
[46] G. J. McLachlan and P. N. Jones, "Fitting mixture models to grouped and truncated data via the EM algorithm," *Biometrics*, vol. 44, no. 2, pp. 571–578, Jun. 1988.
[47] P. Preker, K. Almvig, M. S. Christensen, E. Valen, C. K. Mapendano, A. Sandelin, and T. H. Jensen. (2011). Promoter upstream transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters," *Nucleic Acids Res.* [Online]. Available: http://nar.oxfordjournals.org/content/early/2011/05/19/nar.gkr370.abstract
[48] 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*, vol. 4, no. 6, p. e27639, Dec. 2013.
[49] N. Fuda, M. Ardehali, and J. T. Lis, "Defining mechanisms that regulate RNA polymerase II transcription in vivo," *Nature*, vol. 461, no. 7261, pp. 186–192, Sep. 2009.
[50] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature*, vol. 489, pp. 109–113, Sep. 2012.
[51] A. Podsiadlo, M. Wrzesien, W. Paja, W. Rudnicki, and B. Wilczynski, "Active enhancer positions can be accurately predicted from chromatin marks and collective sequence motif data," *BMC Syst. Biol.*, vol. 7, no. Suppl 6, p. S16, 2013.

Joseph G. Azofeifa received the BA degree in biology from Vassar College in 2011. He is currently working toward the PhD degree in the Computer Science Department, University of Colorado at Boulder. He is also affiliated with the Interdisciplinary Quantitative (IQ) Biology program through the BioFrontiers Institute. His research focuses on the integration of topics in probability theory, machine learning, and signal processing with biological datasets.

Mary A. Allen received the BA degree in biochemistry from the University of Spring Arbor in 2000, the MS degree in cellular and molecular biology from the University of Wisconsin in 2006, and the PhD degree in molecular, cellular, and developmental biology from the University of Colorado at Boulder in 2010. She is currently a Sie Postdoctoral Fellow. She uses a combination of molecular and computational techniques to increase understanding of transcriptional regulation in cancer and Trisomy 21.

Manuel E. Lladser received the MA degree in mathematics from the University of Wisconsin in 2000, and the PhD degree in mathematics from the Ohio State University in 2003. He is currently an associate professor in the Department of Applied Mathematics, University of Colorado at Boulder. He specializes in discrete and applied probability; however, his research is in nature interdisciplinary and motivated by problems in computational biology and metagenomics.

Robin D. Dowell received two BS degrees (in genetics, and computer engineering) in 1997 from Texas A&M University, the master's degree in computer science from Washington University in St Louis in 2001, and the DSc degree in biomedical engineering from Washington University in St. Louis in 2004. She is currently an assistant professor at the University of Colorado in the BioFrontiers Institute and the Molecular, Cellular, and Developmental Biology Department. She uses machine learning approaches to better understand genomes and transcription. She has been a member of the IEEE since 2001.