Inferring High Resolution Transcription Elongation Dynamics from Native Elongating Transcript Sequencing (NET-seq)

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365034

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Inferring High Resolution Transcription Elongation Dynamics from Native Elongating Transcript Sequencing (NET-seq)

Robert D. Shear

A Thesis in the Field of Biology
for the Degree of Master of Liberal Arts in Extension Studies

Harvard University
March 2020
Abstract

Regulatory phenomena occur in every phase of eukaryotic transcription. Recent improvements in the measurement of nascent RNA measurements have led to insights in elongation dynamics. Native elongating transcript sequencing (NET-seq) reports genome-wide RNA polymerase (RNAP) occupancy at nucleotide level resolution revealing phenomena associated with pausing, the control of anti-sense transcription elongation and others.

Here we report that in *S. cerevisiae* elongation is characterized by regions of constant RNAP occupancy punctuated by abrupt reductions in occupancy. Furthermore, many of these locations are in the vicinity of covalent histone modifications, particularly H3K4 di- and tri-methylation. Additionally, these locations have a characteristic nucleotide type signature.
Dedication

To the memory of my father, Frank Shear.
Acknowledgements

This work would not have been possible without the support, patience, and advice of my thesis director, Dr. Stirling Churchman and the entire Churchman Lab. I am particularly grateful for the generosity and thoughtful suggestions of current and former Churchman Lab members Heather Landry Drexler, Kevin Harlen, Kate Lachance, and Robert Ietswaart. I am grateful to my research advisor, Dr. James Morris for his guidance and encouragement.
# Table of Contents

Dedication ........................................................................................................ iv
Acknowledgements ......................................................................................... v
List of Tables .................................................................................................... viii
List of Figures ................................................................................................... ix
Chapter I. Introduction ..................................................................................... 1
  Eukaryotic Transcription .............................................................................. 4
  Initiation ......................................................................................................... 4
  Elongation ....................................................................................................... 6
  Termination ..................................................................................................... 7
  Expression Regulation is Complicated ......................................................... 7
  Nucleosomes ................................................................................................. 8
  Regulation of Elongation .............................................................................. 8
  Elongation Assays ......................................................................................... 9
  NET-seq ......................................................................................................... 11
  Elongation Rates .......................................................................................... 12
  Nascent RNA Statistical Models ................................................................. 13
  Changepoint Analysis ................................................................................... 16
Chapter II. Materials and Methods ................................................................. 20
  Biological Materials ..................................................................................... 20
  The Computation Pipeline .......................................................................... 20
  Sequence-to-occupancy pipeline .............................................................. 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changepoint Identification</td>
<td>22</td>
</tr>
<tr>
<td>Changepoint Metrics</td>
<td>23</td>
</tr>
<tr>
<td>Detecting Changepoint Biological Correlates</td>
<td>24</td>
</tr>
<tr>
<td>Chapter III. Results</td>
<td>29</td>
</tr>
<tr>
<td>Changepoint Identification</td>
<td>29</td>
</tr>
<tr>
<td>Comparison of Changepoints Between Samples</td>
<td>33</td>
</tr>
<tr>
<td>Biological Correlates to Changepoint Location</td>
<td>38</td>
</tr>
<tr>
<td>Chapter IV. Discussion</td>
<td>45</td>
</tr>
<tr>
<td>References</td>
<td>47</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Changepoint symbols used in this work.</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>Feature sources.</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>Classifier algorithms investigated.</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>Consensus changepoint dataset summary.</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>Number of changepoints for all genes by sample.</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>Similarity between samples.</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>Pairwise changepoint count distances.</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>Number of changepoints per gene: Correlation between sample pairs.</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>Mean occupancy per gene: Correlation between sample pairs.</td>
<td>36</td>
</tr>
<tr>
<td>10</td>
<td>Mean occupancy of final segment per gene: Correlation between sample pairs.</td>
<td>38</td>
</tr>
<tr>
<td>11</td>
<td>Detection of changepoint correlates: comparison of model performance.</td>
<td>41</td>
</tr>
</tbody>
</table>
List of Figures

1  RNAPII occupancy as reported by NET-seq. ........................................ 3
2  From gene to protein. ................................................................. 5
3  Schematic depiction of NET-seq assay. ........................................... 12
4  Symbolic example of changepoint analysis ....................................... 18
5  Genes selected for analysis. ......................................................... 21
6  Schematic of the computation pipeline. .......................................... 21
7  How regions of interest (ROIs) are defined for ML ............................ 26
8  Features derived from sequences ................................................... 28
9  Profile of the first changepoint. ..................................................... 30
10 Direction of cliffs. ........................................................................ 31
11 Heatmap of metagene anchored the gene body end points. ................. 32
12 Proportion of genes by number of changepoints. ............................... 34
13 Pairwise consistency of changepoints between sample pairs. .............. 37
14 Comparison of changepoint overlaps to negative controls. ................. 39
15 Distribution of mean distances between consensus changepoints. ......... 40
16 Model performance. ....................................................................... 42
17 Relative variable importance. ......................................................... 43
18 A,T content near changepoints. ...................................................... 44
Chapter I
Introduction

The regulation of genetic expression is crucial to normal cellular function. It underlies virtually every life process: the transformation of nutrients into energy and cellular building blocks, defense against environmental stress, damage repair, and reproduction. For complex organisms, it also governs tissue differentiation and morphogenesis. When this regulation goes awry, the result is frequently disease or death.

The first step in genetic expression is transcription, the synthesis of RNA from a DNA template by RNA polymerase (RNAP).

In particular, the synthesis of messenger RNA (mRNA) is catalyzed by RNA polymerase II (RNAPII). Transcription is, in turn, divided into three distinct phases: initiation, elongation and termination.

Until recently, transcription initiation was thought to be the principal regulatory control point of transcription. Elongation and termination were viewed as processes that, while necessary to RNA synthesis, were of little importance to regulation (Buc et al., 2009). But as evidence of the complexities of elongation has accumulated, so has interest in regulatory phenomena during elongation. Now there is now strong evidence for several such regulatory phenomena (Kwak & Lis, 2013). Nonetheless, many questions remain to be answered (Saunders, Core, & Lis, 2006; Harlen & Churchman, 2017; F. X. Chen, Smith, & Shilatifard, 2018). The answers to these questions will lead to a more complete understanding of expression regulation and will play a role in the development of new therapies.
Characterizing the probable genomic locations of active RNAPII reveals a variety of previously hidden elongation phenomena. During last decade several new methods have been developed for identifying the specific genomic location of active RNAPII. (Mayer, Landry, & Churchman, 2017). Among these methods is native elongating transcript sequencing (NET-seq) which affords single nucleotide resolution of in vivo RNAPII genomic location (Churchman & Weissman, 2011).

NET-seq produces a snapshot of nascent transcripts synthesized by RNAPII which is then processed, reported by high-throughput sequencing (HTS) and aligned to a reference genome. The 3’-ends of these aligned transcripts yield an instantaneous genome-wide occupancy profile of RNAPII density (Figure 1a). The assay has the extraordinary ability to identify the precise nucleotide at which in vivo productive RNAPII is located. The profile of occupancy counts by location can be interpreted as corresponding to the probability of finding productive RNAPII at those locations. If the mean occupancy decreases at a certain genomic region, then either rate of elongation has increased or a portion of the RNAPII has dissociated from the DNA resulting in premature termination.

Certain insights into RNAPII activity, such as widespread antisense transcription, are apparent by inspection of the shape of occupancy plots. Wide excursions from background occupancy counts are easily recognized as the precise locations locations where transcription has paused.

But other valuable inferences require a more precise quantification of density. For example, the abrupt increase in transcription efficiency at a specific location ought to be reflected in a corresponding decrease in occupancy density starting at that location. These inferences have frequently been based on computing statistics based on sliding windows or on a fixed partition into “tiles” (Figure 1b and 1c). Such statistics necessarily lose resolution for small features while losing statistical power for larger features.
Figure 1. RNAPII occupancy as reported by NET-seq.

(A) The $\log_2$ count of 3’-end alignments for *S. cerevisiae* gene YAR015W from a single sample (GEO accession GSM617027, Churchman & Weissman, 2011). (B) The same sample divided into 8 “tiles” of equal width. The blue vertical lines depict the boundaries of each region. The black horizontal lines show the mean occupancy for that region. (C) The same data with 20 tiles. (D) The same data, but with regions defined by changepoint analysis, assuming negative binomial distribution and computed by R package “Exact Bayesian Segmentation (EBS)” (Cleynen, 2016). This clearly reveals two abrupt drops in mean occupancy (changepoints) located 171 and 550 nucleotides downstream from the annotated TSS. Plausible interpretations include an increase in transcription rate or premature termination at these changepoints.
Eukaryotic Transcription

Transcription is the second of three coupled processes that transform the information encoded in DNA into proteins (figure 2). The purpose of transcription is to synthesize RNA. While the details of eukaryotic transcription are still being elaborated, the processes discovered so far reveal an enormous intricacy and complexity. For concise current reviews, refer to Cramer (2019) and Roeder (2019). These processes perform a range of functions, but for the relevant processes for the present purpose are those which bear on transcription regulation, that is the rate at which RNA is produced. Because mature RNA is subject to degradation, the rate of productive transcription is crucial to maintaining the appropriate concentration of RNA molecules in the cell.

The three phases of transcription are sequential. Elongation can only take place by an RNAP complex that has undergone successful initiation. And the synthesis of a complete mRNA molecule can only take place after the successful completion of elongation. Consequently, the rate of mature RNA production is a conjunction of the rates for all three processes.

Initiation

The first of the three phases of transcription is initiation. It commences when proteins known as transcription factors bind to specific regions upstream from a gene (promoters). These factors in turn recruit several others and form a “pre-initiation complex” (PIC) that includes the RNAP (Sainsbury, Bernecky, & Cramer, 2015). A sequence of processes causes the DNA double helix to open (“melt”) so that the RNAP can envelop the template strand. The composition of the initiation complex continue to evolve and eventually commences active elongation. The RNAP complex is now referred
Figure 2. From gene to protein.

This figure depicts the three major biochemical processes in the synthesis of protein. DNA macromolecules reside in each cell and encode specific sequences (genes) of amino acid residues that comprise a specific peptide or protein. When a cell divides, the DNA macromolecules are faithfully duplicated in the daughter cell by replication. The DNA is not capable of directing the synthesis of the protein itself. Instead, there is a two step process. First, RNA is synthesized by transcription that functionally encodes the same sequence for a given protein as the DNA. After various post-transcription processes, the third step translation occurs where ribosomes read the RNA code to synthesize the amino acid sequence that will become a protein.
to as the transcription elongation complex (TEC).

The biochemical processes that take place during initiation depend on signals in the form of the presence (or absence) of various proteins and RNA species. These signals can be evoked by internal changes in the needs of the cell – metabolic, developmental or otherwise. They may also be evoked by external changes, such as stress or change in ambient nutrients. These signals control the rate of initiation for the transcription of each gene and consequently the rate of accretion of productive TECs. They therefore are the first step in the network of gene expression that governs the composition of the cell.

As initiation is the first step in the synthesis of mRNA, the rate of initiation is the least upper bound of the rate of synthesis. There appears to be two stochastic mechanisms that drive the rate of initiation. The first, generally associated with constitutive genes, produces TECs at independent random times and produces a Poisson distribution of TECs. In the second, known as “bursty” initiation (Mcadams & Arkin, 1997), initiation alternates at independent random time between silence and a high level initiation. This is characterized by overdispersion. This distinction can be used to discriminate between the two modes of initiation by comparing the mean number of initiation events to its variance (Ghosh, 2015).

Elongation

Once the TEC has been assembled and starts to read-out the template strand and synthesize an RNA copy of the sequence, elongation has begun. One nucleotide at a time, RNAP acquires one nucleotide bound bind to the 3’-end of the nascent RNA chain and then translocates to the next DNA nucleotide to repeat the process. RNAP can also backtrack one or a few nucleotides to undo transcription errors (Mustaev, Roberts, & Gottesman, 2017) or perhaps to pause. Elongation may prematurely terminate because
RNAP dissociates from the template DNA. Otherwise, it will continue to add nucleotides to the nascent RNA chain until it encounters a stop codon at which time the termination phase of transcription takes commences.

Termination

During termination, a number of steps take place, prominently polyadenylation and cleavage of the mature mRNA from RNAP. The mature RNA must be transported from the site of transcription to the relatively faraway cell location where translation into protein will take place. Several additional steps occur before the RNAP dissociates from the DNA, many of which are not yet well understood (Anamika, Gyuens, & Tora, 2014). The failure of any of these steps to proceed to success will reduce the rate of RNA production. Such a failure may be general or for a specific gene being transcribed.

Expression Regulation is Complicated

From the above discussion of transcription, it should be clear that regulatory phenomena may occur at any point in the synthesis process. There are also several known post-transcriptional processes that can modulate the rate of accretion of active mRNA molecules for any given gene. Additionally, the quantity of mRNA present for a given gene is a function not only of the synthesis but also of the rate of degradation. This degradation is also subject to a range of regulatory phenomena (Beelman & Parker, 1995; Dori-Bachash, Shema, & Tiros, 2011). Beyond the mean accretion rate, there are dynamic phenomena, including the time to respond to a change in control signals, stability of feedback mechanism, and changes in the statistical distribution of each component.

Yet, the current work is concerned specifically with the processes that specifically occur during elongation, which may reflect changes in the rate of synthesis or the number
of RNAPIIs that are active at specific locations.

The results of this study may support certain hypotheses regarding elongation regulatory mechanisms but cannot be assumed to directly affect the ultimate level of genetic expression.

Nucleosomes

By definition, the difference between eukaryotes and prokaryotes, such as bacteria, is the presence of a nucleus that contains the cell’s DNA in the form of chromosomes. To fit the enormous amount of DNA into the tiny nucleus, eukaryotic DNA is tightly packed into a highly organized complex called chromatin. The basic building block of chromatin is the nucleosome which consists of 146 base pairs of DNA tightly wrapped around a disk-shaped protein to form a coil of 1.67 turns. This disk is comprised of a pair of 4 quarter-disk shaped histone subunits.

While this solves the space problem, it impedes access to nucleosome bound DNA. How RNAP is able to overcome the nucleosome barrier has been extensively studied. While much has been learned, open questions remain (Kireeva et al., 2005). Persistent specific covalent modifications to histone residues are associated with the control of expression of certain biological processes (Weiner et al., 2015; Wu et al., 2015; Kujirai & Kurumizaka, 2020). This will prove to be of interest in the present study.

Regulation of Elongation

While the regulation of elongation is only one element of gene expression, it is an essential piece. Regulation of transcription during elongation is modulated by several processes. Prominent among these is a progressive sequence of biochemical changes in the transcription elongation complex (TEC). The TEC configuration affects transcription
dynamics as well as co-transcriptional processes, such as splicing and chromatin remodeling (Hahn & Young, 2011).

Nucleosome configuration can inhibit transcription by protecting its associated DNA from being transcribed. This can be reversed by nucleosome relocation or eviction (Iyer, 2012). Furthermore, certain covalent modifications of specific histone residues are associated with sequence-specific regulatory phenomena (Rando & Winston, 2012).

The state of RNAP carboxy-terminal domain (CTD) phosphorylation is also a significant, gene-wide influence on elongation dynamics. The CTD is comprised of a repeating heptapeptide. The phosphorylation levels of the CTD changes dynamically and in a specific order as the TEC traverses the gene body. This appears to provide a general control signal that modulates elongation dynamics (Harlen et al., 2016).

Additionally, elongation may be blocked by anti-sense transcription, possibly by collision between active sense and anti-sense RNAP (Gelfand et al., 2011; Pelechano & Steinmetz, 2013).

Regardless of the regulatory mechanism, the effect is either a change in transcription rate or the loss of some productive RNAP. That change of rate might be complete arrest, that is a pause (Mayer et al., 2017). In metazoans, “promoter proximal pausing” prominently occurs just downstream from the transcription start site. (Adelman & Lis, 2012).

Beyond pausing, transcription rates vary. Mean transcription rate in eukaryotes have been measured at between $1 \times 10^3$ and $6 \times 10^3$ nucleotides min$^{-1}$, reviewed in (Ardehali & Lis, 2009). An extreme rate of $50 \times 10^3$ nucleotides min$^{-1}$ as been reported in human cells (Maiuri et al., 2011).
Elongation Assays

Observation of elongation processes is difficult. The consequences of elongation are sandwiched between initiation and termination. The tiny scale and large complexity of these processes render the molecular dissection of elongation daunting.

But progress has been made. Numerous methods have been employed to directly measure transcription rate (Coulon, Chow, Singer, & Larson, 2013; Stavreva, Varticovski, & Hager, 2012). Kinetic ensemble assays rely on activation by stress or drug after which time point samples are analyzed by ChiP, qRT-PCR or GRO-seq. In fluorescence in situ hybridization (FISH), fluorescently labeled DNA is hybridized to RNA on fixed cell. Other applicable visualization techniques include fluorescence recovery after photobleaching (FRAP), fluorescence resonance energy transfer (FRET), and FCS (fluorescence correlation spectroscopy). These methods allow the microscopic visualization of fluorescent nascent RNA and mature mRNA which can then be counted or tracked.

Single molecule methods reveal actual rates but only provide data on a few genes. But there are probably heterogeneous regimes of regulation between cells even within a homogeneous sample (Palangat & Larson, 2012). Consequently, while such methods may precisely report the behavior of a particular cell, it is infeasible to collect reliable rates over the entire sample. So phenomena that occur in a subpopulation may be overlooked.

RNAPII occupancy assays are a particularly powerful tool for studying elongation dynamics, reviewed in (Mayer et al., 2017). Such assays capture in vivo RNAPII locations and produce counts of RNAPII occupancy at specific genomic locations. This signal can be transformed into the probability of finding engaged RNAPII at given locations.

The general strategy for these assays is to capture and process active DNA/RNAP/RNA complexes. The complexes are purified and then their genomic
location is inferred from the sequence of either the DNA or the nascent RNA associated with each complex, either by DNA microarray or by high-throughput sequencing.

Permanganate footprinting captures single-stranded DNA in the transcription bubble and is consequently strand-specific. The resolution of the resulting signal is limited by the size of the bubble as well as noise from single-stranded DNA in the sample from sources other than the transcription bubble.

Chromatin Immunoprecipitation uses immunoprecipitation to pull down DNA-RNAP-chromatin complexes followed by purification, amplification and measurement of the captured DNA. The measurement may be realized by microarray (ChiP-chip) or HTS (ChiP-seq). It is strand-specific, but suffers from high noise levels. It reports the presence of RNAPII at specific locations but not whether the RNAPII was actively elongating at that time.

Transcription run-on methods extract transcription complexes, after which transcription is restarted in vitro with labeled nucleotides. The resulting nascent RNA will have labeled bases for that elongation that occurred in vitro. GRO-seq reports the proportion of nascent RNA in vitro while PRO-seq reports the 5’-most labeled nucleotide, reflecting the location at which transcription was halted. It is strand-specific, but only reports nascent RNA’s which have been restarted under non-physiological conditions.

NET-seq

NET-seq similarly reports the relative occupancy of RNAPII at genomic locations, but has the benefit of directly reporting the genomic location of the TEC without the need to restart elongation in vitro. Additionally, NET-seq readouts are extraordinarily precise.

The NET-seq protocol starts with a in vivo flash frozen sample followed by the extraction and purification of the nascent RNA-DNA-RNAPII complex (figure 3). Then
the 40 nucleotide nascent RNA at the 3’-end of the nascent RNA chain (nearest the RNAPII) is released, amplified and sequenced (Churchman & Weissman, 2011). The location of the RNAPII is strand-specific and precise to roughly the single nucleotide level. Furthermore, because the nascent RNA is measured as it was captured in vivo, the resultant signal reflects actual productive elongation.

In principle, NET-seq is particularly well suited to the study of transcription dynamics, as it directly measures genome-wide RNAPII occupancy profile at single nucleotide resolution. Some insights into RNAPII behavior, such as widespread antisense transcription, follow directly from the observed presence of RNAPII on the antisense strand. Other results, such as pause locations during transcription, show up as wide excursions from background occupancy counts (Churchman & Weissman, 2011).

While conspicuous phenomena, such as pausing and cryptic unstable transcripts, are readily observed in the NET-seq signal, there are several challenges in inferring more subtle phenomena. These include determining appropriate statistical models, correcting for biases, accommodating short-read and other limitations in alignment and employing statistical methods which are appropriate to the characteristics of the data.

Elongation Rates

By reporting occupancy counts at each nucleosome, NET-seq quantifies the relative change in TEC density along the genome. But this change in density could imply either a change in the rate of productive RNAPII molecules passing each location or a contrary change in the elongation rate or some combination of the two.

Transcription rates have nonetheless been inferred by other means. For example, Pelechano, Chá Vez, and Pérez-Ortín (2010) assume a transcription rate of 25 nt/s based on previously published data to study 4,670 S. cerevisiae genes in YPD log growth
Figure 3. Schematic depiction of NET-seq assay.
(Figure from Churchman & Weissman, 2011).

conditions. Using an extension of GRO-seq and adjustments from several sources in they derive median occupancy estimates of 0.078 molecules per kilobase or 0.096 RNAP per gene. Only the most highly expressed genes (< 1%) carry more than one active RNAP simultaneously. For all other genes, except for pausing and assuming there is no premature termination, the relative occupancy for each position in the gene will be proportional to the dwell time for that position unless there is premature termination.

Larson, Zenklusen, Wu, Chao, and Singer (2011) employed a real-time single-molecule method to measure the duration of initiation process as well as total dwell time for two specific genes, GLT1 and MDN1. They then computationally separate initiation from the the total time and conclude that initiation stochastic process while elongation proceeds at a constant rate, which they measured as 20 ± 8 nt/s. The rate coefficient for MDN1 varied with the cell cycle phase.

Investigating Arabidopsis thaliana with single-molecule fluorescence in situ hybridization (smFISH), Ietswaart, Rosa, Wu, Dean, and Howard (2017) report absolute
elongation rates of 2-20 nt/s.

Nascent RNA Statistical Models

Because so much biological observation is well modeled by the Gaussian (normal) distribution, many statistical models for transcription explicitly or implicitly assume this distribution. While this is frequently appropriate for many of the measurements of transcription, it is less so for count data, especially count data with a wide dynamic range. NET-seq occupancy data is such a distribution and thus is not well modeled by a Gaussian distribution (Hilbe & Greene, 2007).

Applying statistical inference methods using a linear model and a normal error distribution, \(Y = \beta_0 + \beta X + \epsilon, \epsilon \sim N(\mu, \sigma^2)\), will produce misleading results, frequently including a high level of type I (false positive) errors.

In principle, count data that arises from are well modeled by the Poisson distribution. But the Poisson distribution has the variance equal to the mean, while real world count data, including NET-seq occupancy data, exhibits variance that grows faster than the mean. Such a dataset is said to be overdispersed. Such data is frequently well fit by the negative binomial distribution model (NB), which is the basis of much current genetic count data analysis (Anders & Huber, 2010; Hilbe, 2014).

Because of the high cost of sequencing, most HTS count data has few replicates. Furthermore, though the number of reads is large, so are the number of genes (for RNA-seq) or even more extreme, the number of genomic locations to which a read may be assigned (NET-seq).

A common bias in count data is the introduction of spurious zero valued observations, known as zero inflation. In the case of NET-seq, a prominent source of zero inflation arises from the practice of discarding ambiguous alignments arising from
paralogs and other non-unique sequences in the reference genome. If a reasonably unbiased estimate of the extent of zero inflation is available, the distribution can be modeled as a mixture distribution of the underlying distribution and a distribution of constant zeros (Hilbe, 2014). The MDSeq R package (Ran & Daye, 2017) combines a re-parameterization of NB with a zero inflation model to generate improved estimates of overdispersion.

In current practice, statistics such as mean and standard are derived from subsets of observations constructed from “sliding windows” (Solaimanpour, Sarmiento, & Mrázek, 2015) or “fixed windows” also known as “tiles” or “bins”. Both involve selecting the window width – a fixed length for each sample. With sliding windows, the statistics are computed at each nucleotide position from a sample that is the window width in size and centered at each successive nucleotide, while with fixed windows the genome is divided into adjacent segments each the width of the window.

The window size is left to the judgement of the analyst. Different window sizes will produce different results. The smaller the window size, the weaker statistical power. Conversely, larger windows obscure the features that should be discoverable when employing a very high resolution assay.

These considerations are increasingly important with the increasing use of HTS. While its original application was the efficient elaboration of DNA sequences and the assembly of whole genomes, it has been rapidly adapted to a variety of transcriptome related applications (Reuter, Spacek, & Snyder, 2015). Prominent among these is RNA-seq, the most common application of which is to characterize differential expression (DE) of specific genes (Wilhelm et al., 2008).

With the growing use of RNA-seq, computational methods for the analysis of RNA-seq have received wide attention (Anders & Huber, 2010; Love, Hogenesch, & Irizarry, 2016). As there are certain common computational considerations between
RNA-seq and NET-seq, some of the methods that have been developed to improve RNA-seq results are potentially applicable to NET-seq analysis. These include reduced signal distortion due to better normalization methods, control over GC-content, PCR and sequencing bias. Furthermore, improved characterization of statistical distribution assumptions and inference techniques may be helpful in improving NET-seq analysis.

Early methods for normalization of read counts between samples were based on scaling to some common factor. For example, “reads per million mapped reads” (RPM) which simply the number of reads associated with a given feature (gene in the case or RNA-seq) by the number of reads in the sample dataset (in megabases). “Reads per kilobase per million mapped reads” (RPKM) adjusts not only by the number of reads but also by the length (in kilobases) of the feature under consideration. (Anders et al., 2013) Such methods produce misleading results because when mapping the entire genome, there is frequently a high degree of variability in the read counts between samples for features that are not part of the experiment. (Li et al., 2017). This results in distortions in between-sample comparisons and a loss of resolution when samples are pooled.

In addition, applying a scaling factor to counts will distort the relationship between variance and mean for reasons that will be described below.

Changepoint Analysis

While the identification of an improved model for the distribution of NET-seq may allow a clearer statistical focus, it does not break significant new ground by itself. As depicted in figure 1 and mentioned at the outset of this introduction, “tiling” and similar approaches to the statistical analysis of occupancy sequences have inherent limitations that become more pronounced as the precision of the occupancy signal increases.

Changepoint Analysis (CPA) is a method that can reveal additional structure in
nascent RNA signals. It seeks to identify regions of statistical homogeneity within
sequences of independent observations (Page, 1955). Its applications are wide-ranging,
including meteorology, geology, economics, finance, manufacturing, history, and literature
(J. Chen & Gupta, 2012; Little & Jones, 2011). CPA has been employed in various
biology applications including ecology, epidemiology, neurology, and cardiology. These
observations are frequently a time series, but might also, for example, be a sequence of
physical locations, such as a successive kilometer stretches of a river or genomic
coordinates.

Recent reports of CPA applications to genetics include the annotation of regions of
loss of heterozygosity (Albert, Hunsberger, Nan, & Taylor, 2004), copy number variation
(Girimurugan et al., 2018), regions of altered GC content (Yildirim, Singh, & Doucet,
2013), exon/intron boundaries (Cleynen, Dudoit, & Robin, 2014), isoforms (Cleynen &
Robin, 2016), and gene damage by the analysis of microarray based comparative genomic
hybridization (Muggeo & Adelfio, 2011; Priyadarshana, Polushina, & Sofronov, 2015).

Regardless of the application, the CPA model is typically formulated as follows
(J. Chen & Gupta, 2012). To model a sequence of N observations, construct a vector \( \mathbf{X} \)
comprised of independent random variables \( X_i \)

\[
\mathbf{X} \equiv \{X_1, X_2, \ldots, X_N\} \quad N \in \mathbb{Z}^+
\]

where \( X_i \) are independent random variables of probability distribution \( \mathcal{G} \), \( X_i \sim \mathcal{G}(\theta_i) \) and
each \( \theta_i \) is a vector of parameters for \( \mathcal{G} \).

The simplest question can be posed as the hypothesis that there is a single abrupt
change in the parameter vector \( \theta \) and compared to the null hypothesis \( H_0 \) that \( \mathbf{X} \) is
homogeneous, while the hypothesis that there is a single abrupt, two-sided transition at
observation \( i \) might be may be written

\[
H_0: \quad \theta_1 = \theta_2 = \cdots = \theta_N,
\]

\[
H_A: \quad \theta_1 = \theta_2 = \ldots = \theta_i \neq \theta_{i+1} = \theta_{i+2} \cdots = \theta_N.
\] (1)

Position \( \tau = i + 1 \) is called the changepoint (or breakpoint) and the two segments may be denoted as \( X_i^i \) and \( X^N_{i+1} \).

Numerous approaches to the test statistic for equation (1) have been studied, including the maximum likelihood estimate (MLE) ratio test (Rigaill, Lebarbier, & Robin, 2012), information theoretic methods (Biernacki, Celeux, & Govaert, 2000) and Bayesian methods (Chernoff & Zacks, 1964).

Once the test statistic has been computed, changepoint can be identified by comparing the test statistic \( T_0 = T(X) \) with \( T_i = \text{argmax}_{1<i<N} T(X_i^i; X^N_{i+1}) \), the time complexity of which is \( \mathcal{O}(N) \).

The model may be extended for more than one changepoint, that is \( k > 2 \) segments, for a fixed \( k \) (figure 4).

\[
H_A: \quad \theta_{\tau_0} = \ldots = \theta_{\tau_{2-1}} \neq \theta_{\tau_2} = \ldots = \theta_{\tau_{3-1}} \neq \theta_{\tau_3} = \ldots = \theta_{\tau_{k-1}} = \ldots = \theta_{\tau_k}.
\] (2)

\begin{figure}
\centering
\begin{tikzpicture}
\node[vertex] (x1) at (0,0) {\( X_1 \)}; \node[vertex] (x2) at (1,0) {\( X_2 \)}; \node[vertex] (x3) at (2,0) {\( X_3 \)}; \node[vertex] (x4) at (3,0) {\( \ldots \)}; \node[vertex] (xn-1) at (5,0) {\( X_{n-1} \)}; \node[vertex] (xn) at (6,0) {\( X_n \)};
\node[vertex] (t0) at (0,-1) {\( \tau_0 \)}; \node[vertex] (t1) at (1,-1) {\( \tau_1 \)}; \node[vertex] (t2) at (2,-1) {\( \tau_2 \)}; \node[vertex] (t3) at (3,-1) {\( \tau_3 \)};
\node[vertex] (theta1) at (1,-2) {\( \theta_1 \)}; \node[vertex] (theta2) at (2,-2) {\( \theta_2 \)}; \node[vertex] (theta3) at (3,-2) {\( \theta_3 \)};
\end{tikzpicture}
\caption{Symbolic example of changepoint analysis}
\end{figure}
Now, \(\{\tau_j\}_0^k, \quad 1 = \tau_0 < \tau_2 \ldots < \tau_k = N + 1\) is a partition of \(X\) into \(k\) segments, where each segment \(S_j\) is comprised of observations \(S_j = X^{\tau_j(j+1)^{-1}}\) (Eckley, Fearnhead, & Killick, 2011). For convenience, the number changepoints, which is always one less than the number of segments, will be denoted by \(\kappa = k - 1\).

In the multiple changepoint case (equation 2), the maximum number of segments \(k_{max}\) is typically estimated by analyst. Then the CPA solution is computed for each \(k \in \{1, \ldots, k_{max}\}\), and then \(\hat{k}\) is selected by optimizing the test statistic. In this context, some sort of multiple test penalty is needed, the appropriate formulation of which is unclear. Likewise, if attempting a Bayesian solution, the optimization the comparison between even two posterior distributions can easily produce unusable results.

Unlike the single changepoint problem, the multiple changepoint problem does not lend itself to an exhaustive search, as its time complexity grows with the factorial of the maximum number of changepoints, \(O(k!)\).

Numerous algorithms have been developed to find approximate but efficient solutions to the multiple changepoint problem.

While the information theoretic approach directly addresses this scenario, the commonly used penalty functions, Akaike information criterion (AIC) and Bayesian information criterion (BIC) tend to overstate \(k\) when segment sizes are of significantly different length. A promising improvement is the Integrated Completed Likelihood (ICL) which is the BIC penalized by the mean entropy (Bertoletti, Friel, & Rastelli, 2015).

Within this general framework, a wide range of problems can be formulated. Examples include: Identification of changepoint \(\tau\) in the mean assuming constant but unknown variance; Identification of changepoints where either the mean or variance changes; Estimation the number of changepoints \(k\) in the mean mean assuming constant variance.

Changepoint analysis has been applied to ribosome profiling with useful results.
(Zupanic, Meplan, & Grellscheid, 2014)

Most of the available parametric CPA software packages address Gaussian distributions. But, as mentioned earlier, NET-seq measurements are count data and are better modeled with an NB distribution. In addition, the statistical models implicit in such computations may contain unfounded assumptions, such as Gaussian distribution or homoscedasticity, that can lead to misleading inferences (Hilbe, 2014).
Chapter II
Materials and Methods

Biological Materials

Four replicates of *S. cerevisiae* strain BY4741 were prepared in accordance with
the NET-seq protocol (Churchman & Weissman, 2012), enhanced with unique molecular
identifiers (UMIs) (Kivioja et al., 2012; Mayer et al., 2015).

The range of each gene body to be analyzed was taken to be the range of the
annotated open reading frame (ORF) in the sacCer3 build of the *S. cerevisiae* genome as
annotated in the *Saccharomyces* Genome Database (SGD), retrieved from
https://downloads.yeastgenome.org (Cherry, 2015). Genes were limited to those
annotated as having a verified ORF. Genes having introns were eliminated. That left 4,891
genes as the universe to be analyzed.

The Computation Pipeline

Unless otherwise noted, all statistical computations were performed using version
3.5.3 of R language (R Core Team, 2018) and version 3.8 of Bioconductor (Lawrence et
al., 2013). The H2O machine learning platform, version 3.26.0.2 was used when needed
(LeDell et al., 2019).

The transformation of raw sequence read files into conveniently manipulable
arrays of gene bodies and their computed change points is accomplished by a chain of
Figure 5. Genes selected for analysis.
After filtering out genes with introns and genes that were not annotated as having verified ORFs, 4,891 genes remained as our universe to study.

three sub-pipelines (Figure 6).

Sequence-to-occupancy pipeline

The sequence-to-occupancy pipeline follows the basic flow that has been generally used in NET-seq analysis. The pipeline input a raw FASTQ file. The sequence of program steps is given below.

1. The adapter is removed from the 5’-end of each read with program cutadapt (Compeau et al., 2013).

2. Quality control transformation is performed by program PRINSEQ (Schmieder & Edwards, 2011).

3. The UMIs are removed from the read and added to the FASTQ record by a custom script.

4. The reads are aligned to the sacCer3 genome with the STAR alignment tool (Dobin
Figure 6. Schematic of the computation pipeline.
This figure depicts the general flow of the computation pipeline, starting from raw FASTQ files to human readable results.
et al., 2013).

5. Duplicate reads are detected as reads that align to the same location and have the same UMI. The duplicates are eliminated.

The primary output of this pipeline is in bedGraph format. Each input data set of raw reads for a specific sample produces two bedGraph files, one for each strand. Each record in one of the bedGraph files gives a specific nucleotide location and the adjusted occupancy count at that location. Unreferenced genomic locations implicitly have an occupancy of zero.

Changepoint Identification

The readout from NET-seq is a count at each genomic location called “occupancy”. The same the term to refer to the sequence of counts in a genomic region. It is understood that such a sequence is strand specific. A summary of changepoint symbols used in this work is given in table 1.

The CE.ZINB function of the R breakpoint package (Priyadarshana et al., 2015; Priyadarshana & Sofronov, 2016) was used to identify changepoints. This package was chosen from among 7 candidate packages by evaluating its performance on simulated data sets (data not shown).

Changepoint Metrics

An abrupt change in mean occupancy at changepoint \( i \), called here the cliff at \( i \), may be measured as the relative change in occupancy: \( c_i = \frac{\mu_{i+1}}{\mu_i} \). So, a decrease in occupancy at a changepoint will have a cliff \( 0 < c_i < 1 \), while an increase will have a cliff \( c_i > 1 \). The cliff value will also be expressed as \( \log_2(\text{cliff}) \), the log fold change (LFC).
Table 1

Changepoint symbols used in this work.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>An observation with respect to a genomic region of interest. The sequence of occupancies at each position or, depending on context, the sequence of segments $S_i$.</td>
</tr>
<tr>
<td>$N$</td>
<td>The length of $X$</td>
</tr>
<tr>
<td>$S_j$</td>
<td>The segment starting at $\tau_i$, that is the sequence of $x$ observations or random variables $x_i$, where $\tau_i \leq i &lt; \tau_{i+1}$</td>
</tr>
<tr>
<td>$\mu_i, \sigma_i$</td>
<td>Mean and standard deviation for segment $S_j$</td>
</tr>
<tr>
<td>$k$</td>
<td>The number of segments in the changepoint partition of occupancy sequence $X$</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>The number of changepoints, $\kappa = k - 1$</td>
</tr>
<tr>
<td>$\tau_i$</td>
<td>The boundaries of the segments: The sequence $X$ begins at $\tau_0 = 1$, for changepoint $i$, $1 \leq i \leq \kappa$, and the sequence ends at $\tau_{\kappa} = N$</td>
</tr>
<tr>
<td>$X^{(u)}$</td>
<td>$X$ for sample $u$</td>
</tr>
<tr>
<td>$X^{(u,v)}$</td>
<td>$X$ for sample $u$ and region of interest (gene) $v$</td>
</tr>
</tbody>
</table>

Consensus changepoints in a set of samples $S$ are identified by the following algorithm. First, compute $k_{\text{min}} = \min_{s \in S}(k_s)$. Now compute the minimum euclidean distance for all possible arrangements of $k_{\text{min}}$ changepoints for all of the samples. This allows the comparison between samples of having the same $k_{\text{min}}$. The consensus changepoint is taken to be the mean position over all of the samples for each of the $k_{\text{min}}$ changepoints.

While the cliff is an accurate measure of the relative change in occupancy, it is not possible to distinguish between an increase in velocity and a decrease in flux. Doubling the mean velocity within a gene will reduce the occupancy by a factor of 2, as will the loss of half the active RNAP with no change in velocity. A number of papers assume that there is premature termination and consequently assumes that changes in occupancy necessarily correspond to changes in velocity.
Detecting Changepoint Biological Correlates

Supervised machine learning (ML) was used in to detect biological correlates to the change points. For our purposes, the function such an ML system is to learn from from a number of observations and their true classification how to classify new observations. Such systems are known as “classifiers”. There are several popular approaches to the implementation of classifiers, each with its own advantages and drawbacks.

More formally, following the terminology of (Shalev-Shwartz & Ben-David, 2014), each of the observations (or examples) are members of a domain set, $\mathcal{X}$. Each of the observations will be represented as a vector of features. The set of possible classifications is the label set and its elements are labels. For the current application, the label set will be $\mathcal{Y} = \{0, 1\}$, where the label 1 indicates that the feature vector is measured at a changepoint while label 0 indicates that the feature vector is measured at a location other than a changepoint. That is to say, it is a negative control.

The training data or training set is comprised of a sequence of ordered pairs $S = \{ (x_i, y_i) \} \subset \mathcal{X} \times \mathcal{Y}$, where each of the labels $y_i$ represents the true classification for the corresponding observation $x_i$, where $x_i$ is understood to be a feature vector. The function of the ML system can now be formally described as the creation of predictive function $h: \mathcal{X} \rightarrow \mathcal{Y}$. If every $h(x_i) = \hat{y}_i$ is identical to the corresponding $y_i$, then the function $h$ is perfectly accurate. Of course, this may arise because the model is overfitted. To protect against this, a portion of the training set may be set aside as a validation set. When the validation set, which has known label values, produces similar efficacy to the training set, is evidence of the quality of the predictive function.

There are several ways to quantify the performance of the model. In this work, we will employ the receiver operating characteristics area under the curve, or AUC. If the false positive rate is plotted on the x axis and the true positive rate on the y axis, then the
integral over the range \([0, 1]\) is the AUC. One represents a perfect prediction and 0.5 represents pure chance. In spite of certain limitations, this metric has the benefit of being simple to understand and being widely employed in comparing the performance of prediction systems.

Establishing biological correlate candidates in this setting is a matter of defining plausible features. The features can then then be evaluated, separately or in combination, by the AUC to which each of them gives rise.

The foundation for each feature is the “region of interest” (figure 7).

**Figure 7.** How regions of interest (ROIs) are defined for ML

The blue region at the top of this figure represents a single gene, with changepoints at \(\tau_i, 1 \leq i \leq k - 1\). The sole parameter is the window, \(w\), which indicates the number of nucleotides on either side of the changepoint that will be part of the region. The pink regions in this figure represent ROIs with a positive label, \(y_i = 1\). Preliminary investigation indicated that the distribution of nucleotide types is distinctly different within the gene body and on the sense strand. Consequently, taking negative controls from random locations on the genome overestimated learning. So the green ROIs above are constructed so that they are within the same gene body but do not overlap the positive observations.

Specific feature sources are described in table 2. Several of the feature sources produce a vector. Where a feature vector is comprised of more than one feature source, the feature source vectors are concatenated to produce the feature vector.

The distribution random forest (DRF) algorithm was employed in the initial search
Table 2

Feature sources.

<table>
<thead>
<tr>
<th>Source Code</th>
<th>Variable Names</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTD</td>
<td>CTD.Ser2P, 5P, 7P</td>
<td>maximum&lt;sup&gt;a&lt;/sup&gt;</td>
<td>RNAPII carboxyl-terminal domain phosphorylation at serines 2, 5, and 7. (Vinayachandran et al., 2018)</td>
</tr>
<tr>
<td>Kmers/ï²&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Kmer.AA ... Kmer.TT&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Count</td>
<td>Occurrences of each K-mer within the window</td>
</tr>
<tr>
<td>NcLoc</td>
<td>NcLoc</td>
<td>distance in nt</td>
<td>Distance between change-point and closest nucleosome center per (Brogaard, Xi, Wang, &amp; Widom, 2012)</td>
</tr>
<tr>
<td>NcMod</td>
<td>NcMod&lt;sub&gt;mod&lt;/sub&gt;&lt;sup&gt;d&lt;/sup&gt;</td>
<td>mean value</td>
<td>Normalized histone modification level (Weiner et al., 2015)</td>
</tr>
<tr>
<td>HM</td>
<td>NcModMax&lt;sub&gt;mod&lt;/sub&gt;</td>
<td>maximum value</td>
<td>Normalized histone modification level</td>
</tr>
<tr>
<td>NTP</td>
<td>NTP.1 ... NTP.w&lt;sup&gt;e&lt;/sup&gt;</td>
<td>A, C, G, T</td>
<td>Nucleotide types</td>
</tr>
<tr>
<td>PctA ...</td>
<td>PctA, ... PctT&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Proportion</td>
<td>Count of nt of type n ÷ w</td>
</tr>
<tr>
<td>PctCG</td>
<td>PctCG</td>
<td>PctC + PctG</td>
<td>GC ratio</td>
</tr>
<tr>
<td>PctACG</td>
<td>PctA, PctC, PctG</td>
<td>Proportion</td>
<td></td>
</tr>
</tbody>
</table>

Note. Each of these features is understood to operate within the context of a region of interest.

<sup>a</sup> Arbitrary units

<sup>b</sup> ï² - width of k-mer.

<sup>c</sup> See figure 8.

<sup>d</sup> mod - identifier for specific modification, for example NcMod.h3k4me3.

<sup>e</sup> w - width of ROI.
Figure 8. Features derived from sequences

This figure shows how features are derived from sequence information within a window. These features are derived either as a proportion or as a count. For proportion, \( PctGC \) reports the classic GC-content – the proportion of cytosine or guanine within the sequence. \( PctA, PctC, PctG \) and \( PctT \) report the proportion for the corresponding individual nucleotides. \( PctACG \) report each of the nucleotide type proportions individually except for thymine. For k-mers, the fixed width \( k \) must be specified, in this example \( Kmer/4 \) which generates \( 2^k = 256 \) variables.

for relevant features. This algorithm is a robust detector of interactions between factors. Screening all of the proposed factors using the DRF resulted in the exclusion of ineffective features.

The performance of the classifier depends not only on the model but also the selection of the classifier algorithm. Furthermore, these algorithms have “hyper-parameters” which influence the performance of the classifier.

Once useful features were identified, the performance several classifiers (table 3) as well as certain hyper-parameter choices were compared to produce a final combination of model, algorithm, and parameters.
Table 3

Classifier algorithms investigated.

<table>
<thead>
<tr>
<th>H2O Id</th>
<th>Name</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRF</td>
<td>Distributed Random Forest</td>
<td>Generates a forest of decision trees for the model. Useful in detecting hidden relationships between features.</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalized Linear Model</td>
<td>Regression model a range of common statistical distributions.</td>
</tr>
<tr>
<td>XGBoost</td>
<td>eXtreme Gradient Boosting</td>
<td>Another alternative.</td>
</tr>
<tr>
<td>GBM</td>
<td>Gradient Boosting Machine</td>
<td>A classification algorithm with heuristics.</td>
</tr>
<tr>
<td>DeepLearning</td>
<td>Neural Networks</td>
<td>Highly configurable. Excels at the detection of patterns.</td>
</tr>
</tbody>
</table>
Chapter III
Results

Changepoint Identification

A striking pattern is revealed when occupancy at each nucleotide position per gene as reported by NET-seq is sorted by the first changepoint location relative to the TSS (figure 9). Relative mean occupancy appears to be flat before the first change point and then abruptly drops. The location of this changepoint is specific to the gene.

Furthermore, this reduction in mean occupancy is not limited to the first changepoint. Figure 10 depicts log fold change cliff values \( \log_2 c_i, c_i = \mu_{i+1}/\mu_i \) from all of the genes from a single sample. Overall, 94.2\% of changepoints reflect downward shift in occupancy.

In figure 11 is similar to a metagene depiction. But instead of aligning the genome to the transcription start site or the center of the gene, each segment for each gene is adjusted to be of uniform length for each gene. For example, all genes with two segments will have their changepoint positioned in the middle of the x-axis. The y-axis indicates the mean occupancy while the z-axis (fill color) reflects the number of genes that have that mean occupancy at that x position. The monotonic downward trend over the gene body is conspicuous.
Figure 9. Profile of the first changepoint.

The profile of gene body NET-seq occupancy from a typical sample. The lower 15% of genes by mean occupancy have been omitted as well as genes with a first change point beyond the range of the plot (1,200 nt). Observed occupancy has been smoothed by a running 3 nucleotide mean, then scaled so that for each gene the mean occupancy is 1. (A) Relative occupancy profile of genes ordered by first change point. Genes are represented by successive rows and sorted by the distance from the start of the gene body to the first changepoint. Occupancy has been transformed by log2 and then color coded. (B) The metagene centered on first change point. The mean occupancy for all genes in the sample with change points that have a margin of at least 100 nucleotides to the start and end of the gene body.
Figure 10. Direction of cliffs.

The y-axis is the log$_2$ change in occupancy at each changepoint. In this sample, 94.2% (5,658 out of 6,005) were negative, indicating a downward shift in occupancy. The x-axis shows the partition of all changepoints into quartiles based on the mean occupancy of the segment preceding the changepoint and indicates that the negative cliff magnitudes remain similar regardless of comparative density. Genes were trimmed based on total occupancy so that the lowest and highest 5% were eliminated from consideration. In addition individual cliffs of the highest and lowest 2% were discarded.
Figure 11. Heatmap of metagene anchored the gene body end points.

The y-axis indicates the mean occupancy. (A) The distribution of densities of the first segment for each gene. (B) Within each gene, the segments are scaled to be of equal length. The z-axis (heat level) indicates that relative number of genes at that point in the metagene (x-axis) and that mean occupancy per nucleotide. (C) The distribution of densities for the last segment of each gene. From these three panels the overwhelming pattern of monotonically decreasing occupancy is clear.
Comparison of Changepoints Between Samples

The comparison of changepoints within genes and between replicates not only provides evidence regarding the reproducibility of the changepoint computation but also information on the biological variability between the replicates. To study these comparisons, consensus changepoints were computed for the complete dataset. Additionally, a pairwise analysis was conducted for each of the 6 possible pairs of samples.

Table 4 depicts the basic facts of dataset. Table 5 indicates the number of genes with with $\kappa$ changepoints, $1 \leq \kappa \leq 4$. While measurements such as ICL are a principled approach to selecting the most likely ensemble of changepoints for a specific sample, it is inadequate for evaluation of computed changepoints in the context of transcriptional elongation. In particular, it is not adequate to estimate consensus changepoints. To answer such questions a two-part heuristic approach is adopted. First the number of changepoints is considered and then the distances between the changepoints are evaluated.

Table 4

Consensus changepoint dataset summary.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universe of genes</td>
<td>4,399</td>
</tr>
<tr>
<td>Genes with consensus changepoints</td>
<td>3,424</td>
</tr>
<tr>
<td>N samples</td>
<td>4</td>
</tr>
<tr>
<td>Number of changepoints in universe</td>
<td>24,726</td>
</tr>
<tr>
<td>Mean number of changepoints / gene in universe</td>
<td>1.81</td>
</tr>
<tr>
<td>Mean of changepoints in universe per sample</td>
<td>6,182</td>
</tr>
<tr>
<td>Number consensus changepoints</td>
<td>6,305</td>
</tr>
<tr>
<td>Mean number consensus changepoints / gene</td>
<td>1.84</td>
</tr>
</tbody>
</table>

*Note.* The universe of genes includes those with no detected changepoints. The consensus changepoints were computed using the method described in Materials & Methods, Changepoint Metrics.
Table 5 shows the number of genes in each sample with $\kappa$ changepoints, $1 \leq \kappa \leq 4$. The relative proportions are depicted in figure 12. While the maximum number of changepoints dataset is 13, for most genes (85.6%) none of the data points have more than 4 changepoints. This leaves the impression that samples WT-3 and WT-4 are more similar than WT-1 and WT-2. This impression is further amplified by examining the distance between the samples shown in table 6.

**Table 5**

**Number of changepoints for all genes by sample**

<table>
<thead>
<tr>
<th></th>
<th>WT-1</th>
<th>WT-2</th>
<th>WT-3</th>
<th>WT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,505</td>
<td>1,785</td>
<td>1,065</td>
<td>928</td>
</tr>
<tr>
<td>2</td>
<td>1,549</td>
<td>1,240</td>
<td>1,656</td>
<td>1,804</td>
</tr>
<tr>
<td>3</td>
<td>291</td>
<td>305</td>
<td>549</td>
<td>566</td>
</tr>
<tr>
<td>4</td>
<td>79</td>
<td>94</td>
<td>154</td>
<td>126</td>
</tr>
</tbody>
</table>

*Figure 12.* Proportion of genes by number of changepoints.

A sample pairwise tabulation of the number of genes with specific differences between the number of changepoints (table 7) shows that most pairs vary at most by one
Table 6

Similarity between samples

<table>
<thead>
<tr>
<th></th>
<th>WT-1</th>
<th>WT-2</th>
<th>WT-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT-2</td>
<td>1690</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WT-3</td>
<td>1754</td>
<td>2078</td>
<td></td>
</tr>
<tr>
<td>WT-4</td>
<td>1666</td>
<td>2126</td>
<td>1506</td>
</tr>
</tbody>
</table>

Note. This table depicts the Manhattan ($L_1$) distance between sample pairs based on the changepoint count for each gene. The pair WT-3/WT-4 is conspicuously lower than the other combinations and WT-2 is furthest from the other three samples.

Table 7

Pairwise changepoint count distances.

<table>
<thead>
<tr>
<th></th>
<th>WT-1v2</th>
<th>WT-1v3</th>
<th>WT-1v4</th>
<th>WT-2v3</th>
<th>WT-2v4</th>
<th>WT-3v4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2,014</td>
<td>1,956</td>
<td>2,008</td>
<td>1,753</td>
<td>1,670</td>
<td>2,157</td>
</tr>
<tr>
<td>1</td>
<td>1,161</td>
<td>1,209</td>
<td>1,191</td>
<td>1,312</td>
<td>1,432</td>
<td>1,048</td>
</tr>
<tr>
<td>2</td>
<td>218</td>
<td>232</td>
<td>200</td>
<td>311</td>
<td>272</td>
<td>199</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>27</td>
<td>25</td>
<td>48</td>
<td>50</td>
<td>20</td>
</tr>
</tbody>
</table>

Note. The rows represent the distance between changepoints and the values are the number of genes for the corresponding sample pair that have that distance.

The correlation of number of changepoints between sample pairs (table 8) is also consistent with the proposition that samples WT-3 and WT-4 are biologically closer to each other than to WT-1, and still further from WT-2. Interestingly, the correlation of mean RNAP density by gene (table 9) appears to be much less sensitive than the corresponding changepoint counts.

Looking at the mean occupancy in the last segment of each gene (that part of the gene body from the last changepoint to its end) might give a more accurate estimate of completed transcription, as it reflects occupancy just before elongation ends and
termination begins. The corresponding correlation (table 10) shows more variability than
is reported in the correlation (table 9) of whole gene body mean occupancy.

Table 8

Number of changepoints per gene: Correlation between sample pairs.

<table>
<thead>
<tr>
<th></th>
<th>WT-1</th>
<th>WT-2</th>
<th>WT-3</th>
<th>WT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.418</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.469</td>
<td>0.387</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.493</td>
<td>0.376</td>
<td>0.548</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Note. Spearman’s rank correlation coefficient $\rho$.

Table 9

Mean occupancy per gene: Correlation between sample pairs.

<table>
<thead>
<tr>
<th></th>
<th>WT-1</th>
<th>WT-2</th>
<th>WT-3</th>
<th>WT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.980</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.992</td>
<td>0.980</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.992</td>
<td>0.976</td>
<td>0.996</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Note. Spearman’s rank correlation coefficient $\rho$.

In addition to examining the consistency between number of changepoints, the
consistency of changepoint location can be measured by pairwise counting of the number
of changepoints that overlap with respect to a window width, where window width is
measured as a number of nucleotides. Within a gene, the maximum number of overlaps
for any sample is taken as the denominator to give a proportion of overlaps for any gene.
Summing all of the overlaps and dividing by the maximum possible number of overlaps
yields a proportion of overlaps for the sample pair (figure 13).

Figure 14 depicts the same data but also shows for each case a box plot of overlap
**Figure 13.** Pairwise consistency of changepoints between sample pairs.

The changepoints of each sample pair are considered to overlap if they are within $w$ nucleotides of each other, where $w$ is the window width. The $y$-axis gives the proportion of overlaps as the quotient of the number of overlaps and the maximum number of overlaps for each gene in the sample pair (linear scale). (A) The sample pairs depicted in a grid comparing window widths for each pair. (B) The same data but organized by window width to compare relative proportion of overlaps by sample pair.
Table 10

Mean occupancy of final segment per gene: Correlation between sample pairs.

<table>
<thead>
<tr>
<th></th>
<th>WT-1</th>
<th>WT-2</th>
<th>WT-3</th>
<th>WT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.880</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.895</td>
<td>0.861</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.908</td>
<td>0.880</td>
<td>0.896</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Note. Spearman’s rank correlation coefficient $\rho$. The value that is correlated is the mean occupancy for the region of each gene body starting at the last changepoint and ending at the annotated transcription termination site.

proportions for ten simulations that match the segment count $k$ for each gene but segments in length but randomly assign changepoints within the gene.

Figure 15 shows the distribution of mean distances between consensus changepoints.

Biological Correlates to Changepoint Location

The candidates for correlation with identified changepoints include several features derivative from the consensus sequence, CTD phosphorylation measurements, nucleosome location, and histone modification measurements (table 2). Attempts to correlate any of these features with changepoint location using direct statistical models were unsuccessful, so supervised machine learning algorithms were employed.

As an initial screen, various permutations of these features were examined with DRF algorithm as described in the “Materials and Methods” section above. The features that yielded a meaningful prediction were CTD phosphorylation (CTD), covalent histone modifications (HM), and the density of adenine (A) and thymine (T) bases in the vicinity of changepoints.

Having selected these 4 features, an automated search was conducted for the best
**Figure 14.** Comparison of changepoint overlaps to negative controls.
The points are the same computed overlap proportions depicted in figure 13. The y-axis is proportion of overlaps as the prior figure, but with a log scale. These are compared to sets of 10 negative controls depicted as boxplots. The negative controls for each sample and gene has the same changepoint count as the corresponding sample, but the changepoint locations are randomly selected within the gene body.
Figure 15. Distribution of mean distances between consensus changepoints.
Candidate consensus changepoints were retained regardless of distance in nucleotides. Application of the consensus changepoints should be conditioned on a maximum mean distance. (A) The density of mean distance to the consensus changepoint through the ninth decile. (B) The density of mean distance to the consensus changepoint through a mean distance of 30 nucleotides.

performing combination of features, window size (figure 7), and ML algorithm (table 3). Table 11 indicates the highest scoring algorithm for each of several candidate models. Features that had no significant affect on model performance are excluded from this table. The highest AUC (0.814) was achieved with only the HM feature, window size of 24 nt and the DRF algorithm. The model comprised of all 4 features, the same 24 nt window and the “xgboost” algorithm was almost as good (AUC 0.803). But the model that omitted the CTD signal and a 12 nt window, only half the size of the leader, had an AUC only 4.4% lower (AUC 0.769).

This suggests that the biological correlates to changepoint location are (1) covalent histone modification marks (Weiner et al., 2015) and (2) the relative amount of adenine and thymine near the changepoint.

Figure 16 provides more information on model performance. The AUC of 0.769 (panel A) is reasonably strong evidence of a relationship between the changepoints and
Table 11

Detection of changepoint correlates: comparison of model performance.

<table>
<thead>
<tr>
<th>CTD</th>
<th>HM</th>
<th>A</th>
<th>T</th>
<th>Window</th>
<th>Algorithm</th>
<th>Training.AUC</th>
<th>Test.AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>drf</td>
<td>0.744</td>
<td>0.814</td>
</tr>
<tr>
<td>X</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>gbm</td>
<td>0.982</td>
<td>0.803</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>24</td>
<td>xgboost</td>
<td>1.000</td>
<td>0.777</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>16</td>
<td>drf</td>
<td>0.744</td>
<td>0.775</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>12</td>
<td>gbm</td>
<td>0.982</td>
<td>0.769</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>12</td>
<td>gbm</td>
<td>0.984</td>
<td>0.762</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>16</td>
<td>drf</td>
<td>0.736</td>
<td>0.761</td>
</tr>
<tr>
<td>X</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>xgboost</td>
<td>1.000</td>
<td>0.715</td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>drf</td>
<td>0.744</td>
<td>0.679</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>16</td>
<td>gbm</td>
<td>0.982</td>
<td>0.666</td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>gbm</td>
<td>0.985</td>
<td>0.652</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>24</td>
<td>gbm</td>
<td>1.000</td>
<td>0.500</td>
</tr>
</tbody>
</table>

Notes. In order of decreasing Test AUC. The columns CTD, HM, A, and T indicate these features are included for description of Model, Feature, and Window see table 2, table 3, and figure 7, respectively. The algorithm in each row was the highest performer for that model. The table is sorted by in descending order by Test AUC.
the two predictors. Note that the confusion matrix (panel B) predicts many false positives. I hypothesize that these cases may include locations which the potential of a changepoint but where some factor is missing for its realization.

\[ \text{True Positive Rate} \]
\[ \text{False Positive Rate} \]

**Figure 16. Model performance.**

For the validation frame with features HM, A, T and window size 12 nt. \( N = 2636 \).

(A) Receiver Operating Characteristic Curve. \( AUC = 0.769 \), GINI coefficient = 0.538.

(B) Confusion Matrix. Condition \( C \) denotes regions of interest centered at changepoints, while \( X \) are negative control regions.

The contribution of each variable of the feature vector to the overall predictive power of the model is depicted in figure 17. Notably the two nitrogenous base levels and the methylation state of H3 residue lysine 4 comprise over 25% of the overall model predictive power.

Not only are the adenine and thymine levels predictive of changepoints, but also when viewed in aggregate at nucleotide resolution aligned to the computed changepoints, a distinctive pattern emerges (figure 18). The adenine population dominates the thymine population on the sense strand over the gene body, except when close to a changepoint. The thymine level then rises rapidly near then changepoint and then rapidly drops after the
Figure 17. Relative variable importance.
The same model as figure 16. The model is comprised of 28 variables: relative maximum values for 26 covalent histone marks and the mean levels of adenine and thymine in the window. This chart depicts the importance of each variable as a proportion of the most important variable.
changing. In aggregate, the adenine population then returns to dominance roughly 30 bases after the changepoint.

\[ \text{prop} \]

\[ \text{change point} \]

(A) The mean proportions of A and T within 100 base locations of the changepoint. (B) Same data as (A), but only the 20 locations nearest to the changepoint. (C) The relative information content (entropy) for each of the 20 nearest positions to the changepoint. Rendered with ggseqlogo (Wagih, 2017).

Figure 18. A,T content near changepoints.
Chapter IV
Discussion

The changepoint evidence presented here is consistent with the hypothesis that throughout the *S. cerevisiae* genome, transcription elongation exhibits widespread regions of constant average occupancy punctuated by abrupt decreases. Reduction in mean occupancy over the gene body has been previously observed by Cohen, Zafrir, and Tuller (2018) among others. But in the absence of changepoint observations there was no reason to suspect anything other than a gradual decrease in occupancy. Single molecule studies report constant rates of elongation but not abrupt changes, for example (Hocine, Raymond, Zenklusen, Chao, & Singer, 2012). However, premature termination might be difficult to detect when observing one elongation event at a time (Wu et al., 2015).

Mean occupancy alone is inadequate to determine whether a change in occupancy is due to increased velocity or premature termination. For example Cohen et al. (2018); Kawamura, Koyama, and Yoshida (2019); Jordán-Pla, Pérez-Martínez, and Pérez-Ortí (2019) assume explicitly or implicitly that premature termination is not significant and therefore that occupancy changes imply velocity changes.

The hypothesis that changepoint locations correspond to premature termination is at least as plausible as the increased velocity alternative. The latter would suggest abrupt highly localized velocity increases of similar magnitude whether they occurred once or several times per gene. Furthermore, a simple increase in velocity would not change the overall RNAP flux and consequently would be less likely to be relevant to the regulatory processes that follow elongation.
Valuable insights into transcription processes have been inferred by the analysis of statistical distributions over the gene body (Choubey, Kondev, & Sanchez, 2015).
Applying this approach to changepoint data may provide valuable clues to resolve the premature termination versus velocity question.

Additional avenues for further research include the dissection of gene versus histone modification data by biological function, the search for piecewise-constant nascent RNA occupancy in other species, and further characterization of the structure of sequences near changepoints.

The comparison of changepoint data between samples exhibits more sensitivity to variation than does total nascent RNA count and therefore may be an immediate application in the analysis of differential transcription.


51


