Abstract

Mechanisms of Yeast Gene Definition

Carl de Boer
Doctor of Philosophy
Graduate Department of Molecular Genetics
University of Toronto
2014

The yeast *Saccharomyces cerevisiae* is a prevalent system for studying gene regulation because of the ease of experimental methods and the simplicity of its gene structure. Here, I describe my work that aims to identify the sequences and factors responsible for demarcating genes within the genome sequence. With comparative genomics and RNA-Seq, we are quite adept at identifying gene structure. However, the cell does not have access to this kind of information. Instead, it uses the specificities of DNA- and RNA-binding proteins to read and interpret the sequence of the genome; it is this process that I have studied in my thesis.

In the first chapter, I describe my work collecting yeast transcription factor specificities. I evaluated these specificities using available confirmatory data to determine which one best represents the transcription factor; this gave me a high-confidence description of what DNA sequences yeast transcription factors recognize.

Next, I look for over- and under-represented DNA words within and surrounding gene structures and attempt to explain these in terms of the specificities of known factors or other known biological phenomena. I found that the sequences in the 5' and 3' gene ends are very similar and can often be explained by similar phenomena. I also provide evidence that several factors may be involved in regulating transcription in non-canonical ways.

In the final chapter, I describe my efforts to build a model that uses my collection of
transcription factor specificities as well as DNA structural features to identify gene structure as we think the cell would. This model is comprised of two classifiers that identify mRNA initiation and termination sites, and these are used to provide evidence to a hidden Markov model that predicts gene structure. I test that the predicted determinants of promoter structure are sufficient to initiate transcription, and that the transcription arising from randomly-generated DNA is correctly predicted. Overall, my work demonstrates that the sequence elements demarcating yeast genes are relatively simple in nature, which has implications for how transcription is regulated and how genes evolve.
Dedication

This thesis is dedicated to my friends, family, and especially my wife Charlene, who have all supported me on my quest to avoid getting a "real" job.
Acknowledgements

I would first like to acknowledge all the Hughes Lab members, past and present, who have helped my work. I would like to thank Desiree Tillo for getting me started on my project, Kate Cook, Harm van Bakel, and Debasish Ray for many helpful discussions, and all other past and present Hughes Lab members with whom I have had the privilege to work for their helpful advice and friendship. I would also like to thank Kyle Tsui (formerly of Corey Nislow’s lab) for protocols and advice, Pinay Kainth (formerly of Brenda Andrews’s lab) for plasmids and advice, and everyone else who has aided me with experimental and other advice. I would also like to thank my committee, Jack Greenblatt, Quaid Morris, and Alan Moses, for their advice, support, and guidance. Finally, I would like to thank my supervisor, Tim Hughes, who allowed me to work at my own pace but was always available for advice when needed, and without whom I would not be as rigorous a scientist as I am today.

This work was supported by a University of Toronto Open Fellowship, an Ontario Graduate Scholarship, and a Natural Sciences and Engineering Research Council of Canada Postgraduate scholarship (doctoral).
Contents

1 Introduction .......................................................... 1
  1.1 The chromatin context ........................................... 3
    1.1.1 Nucleosome positioning .................................. 3
    1.1.2 Histone modifications and variants ....................... 4
  1.2 Transcript initiation ........................................... 7
    1.2.1 The General Transcription Factors and PIC formation ... 7
    1.2.2 Transcript initiation and promoter release ............... 9
    1.2.3 Promoters initiate bidirectionally ....................... 10
    1.2.4 Transcription factors ..................................... 10
    1.2.5 Transcription-mediated regulation ....................... 14
  1.3 Transcript elongation .......................................... 15
    1.3.1 The RNA polymerase II CTD cycle ......................... 16
    1.3.2 Process of nucleotide addition .......................... 17
    1.3.3 Transcription through nucleosomes ....................... 18
    1.3.4 Pausing and arrest ....................................... 19
  1.4 Pre-mRNA splicing ............................................. 20
  1.5 Transcript termination ......................................... 20
    1.5.1 Cleavage and polyadenylation ............................ 21
    1.5.2 Termination of transcription ............................. 22
    1.5.3 Non-CPA termination mechanisms ......................... 24
1.5.4 Gene loops ................................................. 25
1.6 Predicting gene structure .................................... 26
1.6.1 Predicting translational units .............................. 27
1.6.2 Predicting promoters ...................................... 28
1.6.3 Predicting cleavage sites .................................. 28
1.6.4 Predicting gene expression ................................. 29
1.7 Evolution of genes ............................................ 30
1.7.1 De novo gene birth ........................................ 30
1.8 Summary and thesis outline ................................. 31

2 A compendium of transcription factor specificities 33
2.1 Summary .................................................... 34
2.2 Motivation/introduction ...................................... 34
2.3 Generation of the database .................................. 36
2.4 Evaluation of motifs ........................................ 38
2.4.1 ChIP-chip enrichment ..................................... 39
2.4.2 Correlation with gene expression data ................. 42
2.4.3 GO term enrichment ...................................... 43
2.4.4 Inter-study concurrence .................................. 44
2.5 Manual evaluation of motifs ................................ 44
2.6 Content and information retrieval .......................... 46
2.6.1 Tables ..................................................... 47
2.6.2 Cart ....................................................... 47
2.6.3 Downloads ............................................... 49
2.7 Analysis tools ............................................... 49
2.7.1 Sequence scanning ....................................... 49
2.7.2 Motif similarity search .................................. 51
2.7.3 Genome browser ......................................... 51
4.4.1 Promoter-defining elements drive transcription \textit{in vivo} . . . . . . . . . 111

4.5 A unified model improves computational gene identification . . . . . . . . 116
4.5.1 Creation of the UM . . . . . . . . . . . . . . . . . . . . . . . . . . 116
4.5.2 Evaluation of the model . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

4.6 Predictions of the UM . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
4.6.1 Predicted transcript species are transcribed . . . . . . . . . . . . . 126
4.6.2 Stable transcripts are produced by random sequence . . . . . . . . . 130
4.6.3 Prevalence of bidirectional CPA sites . . . . . . . . . . . . . . . . . . 135

4.7 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136
4.7.1 Interactions between features in \textit{cis} . . . . . . . . . . . . . . . . . 136
4.7.2 Promoter and terminator directionality . . . . . . . . . . . . . . . . . 137
4.7.3 Synthetic regulatory elements and the evolution of genes . . . . . . . 141

5 Conclusions 143

5.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 143

5.2 Implications . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 145
5.2.1 Gene definition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 145
5.2.2 Antisense transcription and gene looping . . . . . . . . . . . . . . . 148
5.2.3 The role of the Nrd1 complex . . . . . . . . . . . . . . . . . . . . . 149

5.3 Future directions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150
5.3.1 GRF function . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 150
5.3.2 Terminator-promoter feedback . . . . . . . . . . . . . . . . . . . . . 151
5.3.3 Promoter directionality . . . . . . . . . . . . . . . . . . . . . . . . . 151
5.3.4 Determinants of nucleosome binding . . . . . . . . . . . . . . . . . 152
5.3.5 Factor specificities . . . . . . . . . . . . . . . . . . . . . . . . . . . 152
5.3.6 Unexplained motifs . . . . . . . . . . . . . . . . . . . . . . . . . . . 153
5.3.7 Improvements to the UM . . . . . . . . . . . . . . . . . . . . . . . . 154

5.4 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 157
List of Tables

2.1 Example motif search queries. ........................................ 37
2.2 Examples of potential TF cofactors. .................................. 46

3.1 Top 50 motifs enriched and depleted in the region from 100 bp upstream of the intron to the 5’ SS. ................................. 65
3.2 Top 50 motifs enriched and depleted in the region from the 5’ SS to 100 bp into the intron. .................................................. 66
3.3 Top 50 motifs enriched and depleted in the region from 100 bp upstream of the 3’ SS to the 3’ SS. ....................................... 67
3.4 Top 50 motifs enriched and depleted in the region from the 3’ SS to 100 bp downstream of the intron. ................................. 68
3.5 Top 50 motifs enriched and depleted in 5’ UTRs. ...................... 70
3.6 Top 50 motifs enriched and depleted in ORFs. ......................... 78
3.7 Association of GO slim terms with k-mer occurrence within ORFs. . 80
3.8 Top 50 motifs enriched and depleted in 3’ UTRs. ...................... 83

4.1 Known features of promoters and their occurrence in the initiation classifier bins. ..................................................... 100
4.2 UM transition probabilities. ............................................ 118
4.3 Final UM observation distributions. ................................... 118
List of Figures

1.1 Nucleosome occupancy and modification profiles across genes. . . . . . . . 5
1.2 Relating motif occurrence to TF binding and expression output. . . . . . 14
1.3 PolIII CTD marks across genes. . . . . . . . . . . . . . . . . . . . . . . 17

2.1 Comparison of the scores of motifs for the different evaluation criteria be-
fore and after trimming low information content (IC ≤ 0.2) flanking bases. 39
2.2 Comparison of ChIP and PBM-derived motifs. . . . . . . . . . . . . . . . 40
2.3 Comparison of the distributions of actual motif evaluation scores with
randomized data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
2.4 The “Motif” table of YeTFaSCo. . . . . . . . . . . . . . . . . . . . . . . 48
2.5 Sequence scan of GAL1-10 promoter region. . . . . . . . . . . . . . . . . 50

3.1 Overall approach to identifying sequence features associated with transcripts. 61
3.2 TA-repeats within 5’ UTRs represent TATA boxes. . . . . . . . . . . . . . 71
3.3 The distribution of AAAAA, TTTTT, AAGAA, and TTCTT motifs across
TSSs and CPA sites. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
3.4 The effect of AANAA motifs on nucleosome occupancy \textit{in vivo}. . . . . 73
3.5 Base bias of yeast ORFs. . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
3.6 The most enriched/depleted k-mers within ORFs show functional enrich-
ment. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
3.7 Hrp1 binding sites are clustered and in phase in terminator regions. . . 84
3.8 Hrp1 binding sites are bound by the TATA-binding protein. ......... 87

4.1 Design, refinement, and performance of the classifiers. ............... 97
4.2 Feature selection and feature importance for the initiation, elongation, and termination classifiers. ................................................. 105
4.3 Resolution of the classifiers and of cellular transcription. ............ 108
4.4 Terminators are bidirectional. ............................................... 109
4.5 Initiation classifier can predict the genes that will be affected by trans-acting factor mutants. ......................................................... 112
4.6 Construction and analysis of the combinatorial promoter library. .... 115
4.7 A genome-scale yeast transcript model. .................................... 117
4.8 Structure and example predictions of RNA-Seq-based transcript identifiers. 121
4.9 Performance of the UM. ..................................................... 127
4.10 Novel predicted transcripts are often transcribed. ....................... 129
4.11 Predicted transcript structure and measured expression of the four randomly-generated 6 kb fragments. .............................. 134
4.12 Gene definition model. ..................................................... 138
4.13 Context dependence of promoter identity. ............................... 139
4.14 Predictions of models across non-traditional transcripts. ............. 140
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>AUPRC</td>
<td>area under the precision recall curve</td>
</tr>
<tr>
<td>AUROC</td>
<td>area under the ROC curve</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>CDF</td>
<td>cumulative distribution functions</td>
</tr>
<tr>
<td>CFIA</td>
<td>cleavage factor IA</td>
</tr>
<tr>
<td>CFIB</td>
<td>cleavage factor IB</td>
</tr>
<tr>
<td>CFII</td>
<td>cleavage factor II</td>
</tr>
<tr>
<td>CPF</td>
<td>cleavage and polyadenylation factor</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CPA</td>
<td>cleavage and polyadenylation</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain</td>
</tr>
<tr>
<td>CUT</td>
<td>cryptic unstable transcript</td>
</tr>
<tr>
<td>DBDs</td>
<td>DNA-binding domains</td>
</tr>
<tr>
<td>EE</td>
<td>efficiency element</td>
</tr>
<tr>
<td>ER</td>
<td>error rate</td>
</tr>
<tr>
<td>GRF</td>
<td>general regulatory factor</td>
</tr>
<tr>
<td>GTF</td>
<td>general transcription factor</td>
</tr>
<tr>
<td>HMM</td>
<td>hidden Markov model</td>
</tr>
<tr>
<td>ML</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NDR</td>
<td>nucleosome depleted region</td>
</tr>
<tr>
<td>NFR</td>
<td>nucleosome free region</td>
</tr>
<tr>
<td>NT</td>
<td>nucleotide</td>
</tr>
<tr>
<td>NTP</td>
<td>nucleoside triphosphate</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PBM</td>
<td>Protein Binding Microarray</td>
</tr>
<tr>
<td>PE</td>
<td>positioning element</td>
</tr>
<tr>
<td>PFI</td>
<td>polyadenylation factor I</td>
</tr>
<tr>
<td>PFM</td>
<td>position frequency matrix</td>
</tr>
<tr>
<td>PIC</td>
<td>pre-initiation complex</td>
</tr>
<tr>
<td>PolII</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>PPI</td>
<td>protein-protein interaction</td>
</tr>
<tr>
<td>PWM</td>
<td>Position Weight Matrix</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA-binding protein</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>TFBS</td>
<td>transcription factor binding site</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
<tr>
<td>TTS</td>
<td>transcription termination site</td>
</tr>
<tr>
<td>UM</td>
<td>unified model</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The process of transcription is central to the cell’s interpretation of the genome, and is the first stage in the expression of genomic sequence. Despite a great deal of research regarding specific mechanisms of transcription, there has been no systematic approach to identifying the sequence and protein factors responsible for demarcating gene boundaries in genomic DNA. There are many programs available that identify specific gene structures in the genome sequence, but few rely on transcriptional mechanisms or test their predictions. We also have numerous experimental approaches that allow us to identify gene structure, and so, although we can readily identify genes in the genome, we are unsure how the cell performs the same task.

Here, I take a bottom-up approach in three parts: create an index of what DNA-binding factors recognize what DNA sequences; identify sequence elements associated with yeast transcripts; and, finally, build a computational model of transcription that identifies transcripts using cellular factors. I use the yeast *Saccharomyces cerevisiae* as a model system because of its simple gene structure and our relatively advanced understanding of the factors involved in transcription of its genes, although it appears that many transcriptional mechanisms work similarly in higher eukaryotes.

In order to accurately model the process of transcription we must first consider what
is known about how transcripts are generated, including: how transcripts initiate, are transcribed, and terminate; how the stages of transcription affect each other; how the promoters and termination sites are recognized; and what sequence elements are involved and what are the factors that recognize them. Transcription is mediated by a DNA-dependent RNA polymerase, which copies the information encoded in the DNA as a messenger RNA (mRNA). In eukaryotes, synthesis of mRNAs initiates in promoter regions and continues until the cleavage and polyadenylation site is reached, at which point the nascent RNA is cleaved, a poly-A tail is added, and the polymerase stops transcribing a short time later. Eukaryotic DNA is generally wrapped around proteins called histones which serve to compact the DNA and regulate DNA-dependent processes. For instance, yeast promoter sequences, where transcripts initiate, are generally devoid of nucleosomes.

Transcript initiation can be affected by multiple factors, including sequence elements in the promoter and the factors that recognize them, as well as epigenetic factors reflecting the chromatin state of the region. All stages of transcription can be regulated to control gene expression level, but transcript initiation is arguably the most important since all subsequent stages of regulation depend on it. The transcript termination site also contains sequences that are specifically recognized by the RNA-binding components of the cleavage machinery, which recruit the factors necessary for mRNA cleavage, polyadenine tail addition, and termination of transcription. There are also non-poly-A termination mechanisms that, instead of resulting in a stable mRNA, result in transcript degradation, and these too are mediated by sequence signals. Transcription can also regulate itself in cis. For instance, polymerase molecules can collide as they transcribe the DNA in a process called transcriptional interference. Further, the act of transcription results in modifications to the histone proteins packaging the DNA, which seems to act as feedback to reinforce transcriptional decisions and repress spurious transcription.

Transcription is clearly a complex process with many sequences and proteins involved,
and if we hope to understand how the cell uses these factors to identify genes, we potentially have to account for all this.

1.1 The chromatin context

The entire process of transcription takes place in the context of chromatin; the DNA of the genome is wrapped around histone octamers (two each of histones H2A, H2B, H3, and H4), forming nucleosomes in a beads-on-a-string configuration, with each nucleosome encompassing about 147 bp of DNA [326]. As described in more detail in the individual sections, the chromatin state affects nearly all stages of transcription, and the act of transcription also affects chromatin. Transcription factors can bind and displace nucleosomes [20, 187, 146], recruit chromatin remodellers [105], and provoke histone modifications [234, 283, 235]. The act of transcription can also affect chromatin, for instance, by displacing histones [320, 435, 109] and chemically modifying the histone tails [70].

1.1.1 Nucleosome positioning

Genes have a distinctive chromatin structure (Figure 1.1A and B). In the promoter, there is a characteristic nucleosome free region (NFR), followed by a well positioned nucleosome where transcription begins (the +1 nucleosome) [550, 344]. Some people prefer the term “nucleosome depleted region” (NDR) to NFR because the nucleosome occupancy measurements are derived from populations of cells and so are depleted across the population. However, within a given cell, at a given position in the genome, a nucleosome is either there or it is not; an NDR corresponds to a region that is free of nucleosomes in a relatively high proportion of the population. Hence, I will use these terms interchangeably. The +1 nucleosome has an array of well-positioned nucleosomes packed against it in the downstream direction and is eventually followed by another NFR at the 3’ gene end [550, 344]. The 5’ NFR is linked to transcription; nucleosomes have
been shown to form a barrier to initiation \textit{in vitro} [320], and so it is not surprising
that promoters lacking an NFR tend to be more tightly regulated, compared with NFR-
containing promoters that tend to be constitutively expressed [64]. The 3' NFR is of
unknown significance but has been postulated to be caused by the cleavage site \textit{cis}-
elements being refractory to nucleosomes, an artifact of MNase digestion, and linked to
3' end formation [7, 49, 88, 128, 462, 242].

Part, but not all, of the nucleosome structure characteristic of genes is determined
directly from the DNA sequence. Although histones do not read the sequence of the DNA
directly [326], some sequences more readily form nucleosomes by virtue of their structure
[322, 440]. Models have been created that predict nucleosome occupancy, given the DNA
sequence [322, 439, 242, 294], and these frequently rely on two primary features: G/C-
content and poly-dAdT tracts [385, 488, 131]. Poly-dAdT tracts are known to exclude
nucleosomes by virtue of their rigid structures [550, 440, 215]. When comparing \textit{in vivo}
nucleosome positions to DNA sequence, it is clear that transcription factors (TFs) also
play a role in determining nucleosome positions, as certain binding sites are especially
enriched in NFRs (see section on the General Regulatory Factors) [242]. However, the
specificity of nucleosomes appears to play a dominant role in a minority of nucleosome
positions [562, 242, 559, 207]. Most other nucleosome positions are controlled by the
action of chromatin remodellers, which use the energy of ATP hydrolysis to move or
eject nucleosomes [90]. This flow of nucleosomes can be restricted by boundary elements
and well-positioned nucleosomes [273]. For instance, the densely packed nucleosomes
within gene bodies are a product of chromatin remodelling enzymes, which use ATP to
shift the positions of nucleosomes [213, 504].

1.1.2 Histone modifications and variants

Histones can also undergo various covalent modifications of their N- and C-terminal “tail”
domains, including acetylation, mono-, di-, and trimethylation, phosphorylation, sumo-
Chapter 1. Introduction

Figure 1.1: Nucleosome occupancy and modification profiles across genes. (A) Nucleosome structure across a PolII-transcribed gene. (B) The average nucleosome occupancy profile, averaged across genes. Nucleosomes proximal to the 5' NFR tend to have the histone variant H2A.Z. (C) The approximate distribution of selected histone modifications as they tend to occur across genes. Figures adapted from Jiang and Pugh, 2009 [224], and Smolle and Workman, 2013 [457].

lation, isomerization, and ubiquitylation (see [457]). These modifications often appear in distinct regions of genes, when viewed in aggregate, with different marks appearing to label the 5' gene ends, gene bodies, and 3' gene ends, and some marks differing between highly and lowly expressed genes [396, 315] (Figure 1.1). These histone marks are written and erased by chromatin modifiers that are recruited in a multitude of ways, including through TFs [50, 404] and concomitantly with the different stages of transcription [232, 537, 182]. They are also influenced by the marks themselves, which can recruit enzymes that spread or modify the mark [285, 184, 25, 70]. These marks can be
maintained through DNA replication using proteins that both recognize and place the same mark [184]. Examples of positioned histone marks include the H3K4 and H3K36 methylation marks that occur within gene bodies of actively transcribed genes [37, 466] (Figure 1.1), which are placed there by the Set1 and Set2 methyltransferases, respectively [418, 37]. The H3K36me3 mark seems to have a repressive effect on transcription, preventing spurious transcription from cryptic promoter-like elements within gene bodies [70]. The histone acetylation frequently found in promoters is often placed there by the histone acetyltransferases recruited by TFs [169, 145]. Histone marks are read by various proteins and can affect diverse processes, including DNA replication [31, 507], the remodelling of chromatin [67], transcription [507], and splicing [325]. For example, the RSC chromatin remodelling complex preferentially remodels acetylated histones in promoters [67].

In addition, there are also several histone variants that are encoded by separate histone genes and appear to have slightly different functions in the cell [549]. For instance, H2A.Z seems to be primarily associated with unstable nucleosomes around promoters [406, 5] (see Figure 1.1B) and may help to facilitate nucleosome loss through a decreased stability in inactive promoters [556, 301]. While higher eukaryotes also have DNA methylation as a chromatin mark, *S. cerevisiae* lacks this feature [402]. These various marks are termed “epigenetic” because they are not reflected as changes to the genome sequence, but can nonetheless have functional consequences in the cell.

Overall, there are many examples of epigenetic chromatin marks that label the transcriptional status and gene structure of yeast and, although there are compelling examples of chromatin marks controlling gene expression in higher eukaryotes (e.g. [415]), overall, this appears to be rare, especially in yeast. In general, the chromatin state cannot be the primary means by which the cell interprets the genome, since naked DNA (e.g. plasmids) can be inserted into the cell and recapitulate many of the same features as the endogenous genes (e.g. [309, 222, 117, 170]). Further, the facts that mammalian fibroblast
cells can be reprogrammed into a stem-cell like state by expressing specific transcription factors \cite{473} and many chromatin modifiers are not essential (e.g. \cite{405, 156, 296}) argue that, in general, it is the TFs that regulate both transcription and the chromatin state. I think that, rather than dictating cell state directly, these histone marks help to buffer and reinforce the transcriptional decisions made by the complement of sequence-specific transcription factors in the cell.

1.2 Transcript initiation

Transcription is initiated in the promoter region, which is recognized by a multitude of competing DNA-binding proteins: nucleosomes occlude the DNA, the general transcription factors (GTFs) load polymerase onto the DNA, and TFs alter the efficiency of polymerase loading and initiation. A great deal of work has been done, both \textit{in vivo} and \textit{in vitro}, on the GTFs from both humans and yeast. However, these factors and their functions are highly conserved within metazoans \cite{59, 374}, and so the knowledge gained from either organism often applies to both.

1.2.1 The General Transcription Factors and PIC formation

By itself, purified RNA polymerase II (PolII) is incapable of specifically initiating transcription from a promoter on a purified template \textit{in vitro}, but can initiate upon the addition of a crude cellular extract \cite{520}. When this was discovered, several groups identified the active components by fractionation of cellular extracts. This resulted in the identification of the GTFs, which are named after the fraction in which they were purified (A, B, D, E, F, and H) (see \cite{486} for review). These factors play various roles in assembly of the preinitiation complex (PIC) and initiation of transcription. Although capable of sequential assembly \textit{in vitro}, where each GTF assembles into the PIC in turn \cite{133, 56}, it is unclear how prevalent this is in the cell, as PolII holoenzyme can be purified as a
pre-assembled complex that already includes many of the general transcription factors [486]. However, it is generally thought that the first stage of PIC formation requires the recognition of the promoter sequence by TFIID and, in particular, binding of the TATA box by a TFIID component, the TATA-binding protein (TBP) [56, 535, 59, 453]. The PIC is formed when the other GTFs and PolIII join TFIID at the promoter. Some eukaryotes have binding sites recognized by other GTFs [486], but the TATA box and initiator sequence appear to be the only ones present in yeast. The initiator sequence, recognized by TFIIB, PolIII, and/or Taf2 [394, 304, 60, 73], has a consensus of YR [560] and so is found frequently throughout the genome.

The GTFs are thought to play various roles in PIC assembly and function, including structural, enzymatic, and regulatory. TFIID recognizes the promoter, nucleates PIC assembly through interactions with other TFs [59, 508], and enzymatically modifies chromatin and other protein factors involved in transcription [358, 332, 390]. TFIIA stabilizes TBP-promoter complexes through direct binding of TBP, the promoter, and other GTFs [287, 93, 18, 155, 476] and can act as a co-activator with other TFs [266]. TFIIB appears to act primarily by stabilizing the TBP-promoter complex [563, 528] and acting as an interface among TFIID, PolIII, and the other GTFs [57, 177], but also acts as a co-activator [189, 431] and plays a role in start site selection [394, 383, 127]. TFIIIE interacts with activators [431, 568, 372] and GTFs, recruits [137, 345, 372, 547, 516] and stimulates the activity of TFIIH [371, 442, 370], and may facilitate promoter melting via binding of single stranded DNA [547, 201, 282]. TFIIF facilitates PIC formation by binding to PolIII and the other GTFs [460, 400, 136, 416] and plays a role in start site selection and promoter escape [160, 540]. TFIIH is an ATPase, DNA helicase, and kinase [371, 312] that interacts with activators and repressors [571], acts to facilitate promoter melting and escape [312, 201, 129], and is responsible for setting up the initial phosphorylation marks on the PolIII C-terminal domain (CTD) [130]. Together, the GTFs, PolIII, and the promoter comprise the PIC.
In addition to the GTFs, the Mediator complex is also required for the \textit{in vivo} activation of many genes [212, 154]. This complex appears to act as an interface between transcriptional activators and the transcription machinery [41, 335], translating TF binding to transcriptional activation. Mediator can also interact with elongation [472, 112] and splicing factors [205] and may play additional roles in elongation through its control of CTD phosphorylation [225, 112, 44]. In addition to elongation, the Mediator complex has also been shown to play a role in termination, although the mechanism behind this remains opaque [363].

1.2.2 Transcript initiation and promoter release

Once a stable PIC has formed, TFIIH, through its ATP-dependent helicase activity, transitions the PIC into the open complex [130, 200, 201, 259, 513], so named because the DNA strands around the TSS are not base paired, making a “transcription bubble”. After formation of the open complex, several short transcripts are generated as the DNA template is repeatedly transcribed in a process known as abortive initiation; these can dissociate from the polymerase [200, 247], but seem to be stabilized as more bases are transcribed and the strength of the RNA-DNA hybrid increases [247]. When the transcripts are short and weakly bound, TFIIB can stabilize the RNA-DNA hybrid [60], but at longer lengths (>5 bp), TFIIB starts to get in the way [60]. Because the polymerase remains bound to the promoter and DNA is being fed into the PIC by TFIIH [173], the DNA between the promoter-DNA contacts and the current DNA template remains single stranded and bunched up within the polymerase in a process termed "scrunching", which allows the PIC to build up the energy needed to escape the promoter [411, 240]. Meanwhile, TFIIH phosphorylates the PolIII CTD on Ser5 (this process is described in further detail below), which may contribute to PolIII-mediator dissociation [443], but more generally appears to only be required for 5’ RNA capping [239, 438]. Eventually, TFIIB is pushed out of the way, allowing for the generation of longer transcripts that make their
way out the RNA-exit tunnel [553, 525], while the energy stored in the scrunched DNA allows the polymerase to break free of the PIC [247]. This final stage appears to often be the rate-limiting step to productive elongation [280, 518].

1.2.3 Promoters initiate bidirectionally

Many promoters are capable of initiating transcription in both orientations. While specific examples of bidirectional promoters resulting in two stable transcripts have been known for years (e.g. [492, 532, 353, 16, 449, 429]), it was not until recently that the true extent of bidirectional transcription was understood. It appears that most yeast genes initiate transcription bidirectionally, but, under normal growth conditions, the divergent transcripts are undetectable [367, 538, 89, 100]. In order to detect these transcripts, called CUTs (cryptic unstable transcripts), the enzyme that degrades them must first be disabled [536]. These bidirectional promoters tend to share an NDR and, when analyzed as a whole, the TSSs of the two divergent transcripts occur about 180 bp away from each other [538]. Studies in other organisms have found similar results [68, 448, 398, 482, 243], and so it seems that promoters may inherently be bidirectional.

1.2.4 Transcription factors

Transcription factors (TFs) are proteins that bind to DNA in a sequence-specific manner and affect transcription either positively or negatively. In contrast to the GTFs, TFs are not required for promoter-specific transcription of templates in vitro and generally only affect transcription of a subset of genes [87, 204, 149]. The binding sites of many TFs, which the TFs recognize through their DNA-binding domains (DBDs), are enriched in promoters, relative to background [123]. The most common DBDs in yeast include zinc-cluster (e.g. Gal4) and C2H2-zinc fingers (e.g. Zap1). Because of a common structure and related amino acid sequence, TFs within a common DBD class often share a related binding specificity, and the specificity generally diverges with divergence of the amino
acid sequence [8]. The specificity of a TF can be derived in both direct and indirect ways. In yeast, most of the specificities available are derived from PBM experiments (an *in vitro* method designed to measure specificities) [564, 20], or by searching for repeated sequences in bound ChIP-chip probes (e.g. [138, 329, 293, 185]). Specificities (or “motifs”) can be represented in many different forms, but by far the most prevalent are position weight matrices (PWMs) and IUPAC nucleotide codes (which can be translated into PWMs). PWMs are most often log-odds matrices of the frequency of observing each base at each position in a binding site, relative to the background, giving a score which, if above 0, is more like the motif than the background. With these motif representations, DNA can easily be scanned, and potential binding sites for the TFs identified.

TFs seem to have a plethora of mechanisms by which they affect transcription. These include interaction with the GTFs (e.g. [291, 382]), the mediator complex (e.g. [53, 483]), and chromatin modifiers (e.g. [50, 404]). Many TFs have similar acidic activation domains (e.g. [565, 116]), although other types exist (e.g. [17, 125]). A common feature of acidic activation domains is that they are largely unstructured [491] and are often not conserved in amino acid sequence [339] but maintain their function in heterologous systems [179]. Together, this suggests that activating interactions are relatively transient and unstable. Indeed, Gcn4-Mediator interactions seem to be relatively unstable and can occur in multiple conformations [54], but this appears to be compensated for by multiple activation domains interacting with multiple activator-binding interfaces [333].

Some TFs are repressors rather than activators. These can also function through varying mechanisms, including recruitment of histone deacetylases (e.g. [534, 517, 424]), stabilization of nucleosomes (e.g. [401]), and competition with activators (e.g. [567]). There are also cases where a TF can be both an activator and a repressor. For instance, Leu3 masks its own activation domain in the absence of alpha-isopropylmalate to become a repressor [263].

Some TFs seem to work cooperatively, where binding of a pair of TFs is more efficient
than expected given the binding of either TF on its own. This action seems to be mediated by two primary mechanisms: protein-protein interactions (PPIs) (either direct or indirect) between DNA-binding components that stabilize TF-DNA interactions; and competition of the two factors with nucleosomes. There are many examples of TFs that bind cooperatively through direct PPIs, primarily through homo/heterodimerization (e.g. [124, 34, 551, 475]); however, indirect PPIs are less common. Examples of this would include interacting with different components of the transcription or chromatin remodelling machinery. Since there is no obvious theoretical reason why this should be uncommon, the relative lack of evidence for this kind of cooperativity may simply reflect relatively small effect sizes, increased difficulty in studying these systems, and/or an increased number of possible interactions (i.e. direct PPIs are limited by the ability of two TFs to interact, but any pair of TFs binding different interfaces on a third protein or complex can interact indirectly). Cooperativity mediated by nucleosomes has been relatively well-characterized both \textit{in vivo} and \textit{in vitro} (e.g. [389, 505, 354, 2]). Here, both factors compete with a common nucleosome for binding to the DNA, shifting the equilibrium condition to having one or both TFs bound. The general regulatory factors (GRFs; see below) are a classic example of this, wherein binding of the GRF opens chromatin, allowing other TFs to bind [46, 548].

In addition to competition with nucleosomes, in cases where binding sites overlap, TFs must compete with each other for binding to the DNA. Sometimes TFs compete by binding partly overlapping or directly adjacent spots on the DNA [392], such that only one factor can bind at a time. Other times, the binding sites are identical; these often represent cases where the DBDs are related, and so they bind a similar DNA sequence (e.g. [567]).
The General Regulatory Factors

The General Regulatory Factors (GRFs) are abundant, essential transcription factors that appear to have the ability to effectively compete with nucleosomes for binding DNA. The canonical GRFs include Abf1, Reb1, and Rap1, while other factors exhibit GRF-like properties, including Rsc3, Tbf1, Mcm1, and Cbf1 [20, 108, 494, 141]. These factors had been implicated in altering the chromatin structure of specific promoters (e.g. [289, 548]), but recent genome-wide measurements of nucleosome occupancy have shown that the binding sites for these factors are generally depleted of nucleosomes in vivo [294, 242]. Further, upon depletion of these factors, their binding sites become bound by nucleosomes [543, 20, 187, 146]. Although these factors have been implicated in many cellular processes, including recombination, telomeric chromatin maintenance, and rRNA transcription [447, 361], their primary function in promoters appears to be through increasing the accessibility of DNA for other transcription factors [548, 542].

Mapped interaction networks

In contrast to the GRFs, most TFs tend to regulate the transcription of a relatively small number of genes, which often share a related function [204]. Substantial effort has been expended applying genomics approaches to mapping the regulatory interactions between TFs and their gene targets. Using genetic perturbation (e.g. TF deletion, overexpression, mutation, etc.), several studies have attempted to recreate regulatory networks [204, 87]. Essentially, these studies attempt to identify targets by measuring expression changes when a TF is perturbed, with the idea that the expression of targets of that TF will change upon TF perturbation. An alternative approach attempts to map regulatory networks by identifying physical TF-DNA interactions. The primary method implemented in yeast is ChIP-chip, wherein epitope-tagged TFs are chemically cross-linked to DNA in the cell, the tagged TFs are affinity purified, and the associated DNA is isolated and quantified using a microarray [410]. Two major datasets have emerged using
this technology, one encompassing the other [185, 293]. One of the major conclusions of these studies is that most TFs regulate few genes, with exceptions including the GRFs, although most of these experiments were done in one or few conditions in which the TF may not be active [293, 204]. By identifying the set of targets of a TF, the DNA motif to which the TF binds can often be identified because of the tendency for targets to contain binding sites within their promoter regions [185, 204, 293, 87, 420]. These genome-wide approaches are important in providing the data necessary for global analyses of gene regulation, including how motif presence, physical binding, and expression output relate to one another (Figure 1.2).

![Figure 1.2: Relating motif occurrence to TF binding and expression output. The DNA contains potential binding sites for TFs. Some of these are bound by TFs under a given condition and some binding events give rise to a change in a nearby gene’s expression level. If we know any one of these, we can partially infer the other two. For instance, if a TF mutant causes an expression change in a gene (and the gene is a direct target of the TF), then the TF binds to the gene’s promoter, and, where the TF binds, there is likely an instance of the TF’s binding site. Figure adapted from Hughes and de Boer, 2013 [208].](image)

### 1.2.5 Transcription-mediated regulation

It is possible for the transcription emanating from one promoter to impact the expression of an adjacent promoter. This can take several forms. For instance, adjacent genes are co-expressed more than pairs of randomly selected genes [92], and the genes with the most
highly correlated expressions are in the tandem orientation [278], indicating that transcription of the upstream gene may be influencing transcription of the downstream gene. Further, there are well-documented cases of regulatory RNAs that enhance or inhibit the transcription of a gene in cis (e.g. [66, 203, 55, 203]). It is not always clear what the mechanism of regulation is for individual examples, but two themes have emerged: regulation by histone modification, and regulation by transcriptional interference. For instance, Spt6-mediated nucleosome assembly and Set2-mediated histone methylation have been shown to repress transcription from promoter-like sequences within genes [241, 70, 250], and there are specific examples of ncRNA (non-coding RNA) transcription regulating the promoters of ORF-containing genes through repressive histone marks [203, 55]. Transcriptional interference, on the other hand, describes the ability of elongating polymerase to interfere directly with the transcription of another gene, either by collisions between polymerases or by occlusion of the promoter region [348]. The fact that adjacent promoters can influence each other in these ways adds an additional layer of complexity to gene regulation, but it is unknown how significant these phenomena are on a genome-wide scale.

1.3 Transcript elongation

Transcript elongation is the iterated process of NTP incorporation into the nascent RNA, followed by subsequent translocation of the polymerase to the next base. This process begins in the promoter where the first nucleotides are added and the polymerase assumes a transcription-competent form. The RNA continues to be elongated until the transcript is terminated and the polymerase is recycled. During the various stages of transcription, there is an evolution of PolII forms, mediated through the C-terminal domain (CTD) of PolII.
1.3.1 The RNA polymerase II CTD cycle

The PolII CTD is comprised of 26 repeats (52 in human) of a heptapeptide whose consensus is YSPTSPS [9]. Although by no means completely understood, it appears that the primary role of the CTD is to regulate the proteins associated with PolII at the various stages of transcription, including the factors responsible for pre-mRNA processing [350].

PolII is initially recruited to the promoter in an unphosphorylated state [323]. In the promoter, the Kin28 subunit of TFIIH phosphorylates both Ser5 and Ser7 [4, 130, 268]. When the polymerase is released from the promoter, Bur1/Bur2 and Ctk1 phosphorylates Ser2 [84, 230, 403]. As PolII elongates, phosphatases (e.g. Scp1) are recruited that remove Ser5-P [546], but the kinase Bur1 maintains Ser7-P levels [487]. Following cleavage and polyadenylation, the Ser2-P and other phosphate marks must be removed (by Fcp1, Ssu72, and others) so that PolII can be recycled [86, 275, 271, 188]. Consequently, there is a characteristic profile of CTD marks across the length of the gene, where Ser5-P is high in the promoter, but quickly decreases to a low level at the cleavage site, Ser2-P increases through the length of the gene up to the termination site, and Ser7-P is high at the promoter and remains relatively high across the length of the gene [487, 347, 268] (Figure 1.3). Here, I have only described the most well studied of the CTD marks; however, there are numerous other modifications, including Thr4-P, Tyr1-P (see Figure 1.3), cis/trans proline isomerization, and glycosylation [191]. Since proteins are likely to recognize different combinations of marks with different affinities, the protein factors that associate with the CTD at a given time are likely to depend on the amount and co-occurrence of these marks. For instance, the protein Nrd1 recognizes a combination of Ser5-P and a cis-conformation of Pro6 [279]. However, the many possible combinations of CTD marks and the corresponding interactions with CTD binding factors remain largely unstudied.

Each CTD state is selectively bound by one or more proteins at the different stages of mRNA processing. For instance, the Ser5-P in the promoter facilitates binding of the
Figure 1.3: PolIII CTD marks across genes. The distributions of several histone marks are shown as they occur, on average, across genes. Figure adapted from Heidemann et al., 2013 [191].

proteins responsible for 5’G mRNA capping and 5’G methylation [85, 349, 393, 434]. This MeG cap protects the nascent mRNA from degradation by the cellular 5’->3’ ribonucleases, facilitates translation, and acts as a platform on which a host of mRNA processing and regulatory factors are recruited [144, 446, 270, 181]. Upon promoter escape, the polymerase is free to transcribe the template DNA. Splicing also depends on the CTD [139] and its phosphorylation state [196, 38]. At the 3’ end of the gene, the CTD is required for proper mRNA cleavage and polyadenylation [195] and the cleavage and polyadenylation machinery selectively binds high Ser2-P low Ser5-P [308, 3]. Moreover, Tyr1-P prevents the association of cleavage and polyadenylation factors [346].

1.3.2 Process of nucleotide addition

When in the elongation-competent form, PolIII must repeatedly incorporate NTPs into the nascent mRNA while negotiating the chromatin landscape. With each nucleotide, the polymerase undergoes three steps: addition of the new NTP; formation of a phos-
phodiester bond between the incoming NTP and the existing RNA (and release of the pyrophosphate); and finally translocation of the polymerase to align the next base pair with the catalytic site [52]. In vivo, PolIII manages to do this about 4000 times a minute [98]. However, the process of nucleotide addition and translocation is non-uniform across the DNA template. The rate of PolIII elongation (that is, the number of bases added per minute) can be affected by the DNA sequence being transcribed and NTPs being incorporated into the mRNA, as a function of both the concentration of nucleotide [342], and likely the properties of the individual nucleotide (e.g. purine vs. pyrimidine). In turn, the rate of elongation can affect pre-mRNA processing steps, such as cleavage and polyadenylation and splicing [102, 395], likely reflecting the fact that trans-acting factors have longer to recognize cis-elements when elongation rate is reduced. Further, the polymerase frequently stalls as it transcribes the DNA template [409, 89], with causes including DNA-damage [111], misincorporation of NTPs [223], and transcription through nucleosomes [89, 198].

1.3.3 Transcription through nucleosomes

Unwinding nucleosomes appears to form a sufficient barrier to transcription to cause polymerase stalling [89]. Although nucleosomes form a significant barrier to transcription in vitro [217, 264, 216, 198], its effects are less profound in vivo, perhaps because of the actions of chromatin remodellers [451, 217, 36, 450, 153, 456, 436]. Nevertheless, PolII frequently pauses upon collision with nucleosomes and tends to stop upon reaching the nucleosome dyad [89, 198], consistent with the dyad position forming particularly strong histone-DNA interactions [180]. As PolII transcribes the chromatin template, it must unwind nucleosomes and displace them. This may be facilitated by spontaneous partial unwinding of the nucleosome [302]. In its wake, PolII must reassemble this chromatin and tends to replace many of the same nucleosomes that were displaced. This process appears to be inefficient, however, as the degree of re-deposition of histones is inversely correlated
with transcriptional activity [437, 485, 109, 423]. One mechanism for this nucleosome reassembly appears to be the formation of a transcriptional bubble that moves around the nucleosome, allowing the transcribed DNA to reassemble on the histones and thereby reforming the nucleosome as transcription proceeds [198, 262, 36]. Histones H2A and H2B may be lost during transcription; the rates of H2A/B turnover are much higher than H3/H4 turnover [485, 220], but more efficient re-incorporation may be facilitated by H2B ubiquitylation [134]. This may allow a preservation of the histone modifications present on H3 and H4. The increased turnover on highly expressed genes may be due in part to the increased density of PolIII on the gene (i.e. if one polymerase directly follows another, it will prevent nucleosome reassembly after passage of the first) [281]. Multiple concurrent polymerases transcribing the same template can also act synergistically to overcome the nucleosomal barriers by preventing backtracking [227, 281]. In the single polymerase case, pausing is commonplace in the transcription of a gene, and several factors help to prevent pausing and speed polymerase recovery.

1.3.4 Pausing and arrest

PolIII pausing is a transient halt to the addition of NTPs, but can transition to an arrest when PolIII elongation is blocked [497]. This arrested complex is characterized by the polymerase backtracking along the DNA, resulting in a misalignment of the active site and the RNA’s 3’ hydroxyl [511, 269]. TFIIF can help prevent this backtracking from occurring by transiently interacting with the elongation complex and stabilizing the post-translocation elongation complex [217, 251, 553, 555]. When arrest occurs, TFIIS can rescue it by initiating mRNA cleavage, producing a new 3’ OH on which PolIII can begin nucleotide addition [512, 422]. TFIIS does this by inserting a long linker into the PolIII active site and stimulating the inherent hydrolytic activity of the polymerase [19, 253]. Once the mRNA is cleaved and a new 3’ hydroxyl is ready for nucleotide addition, the polymerase remains paused for some time [89], indicating that cleavage of
the transcript is not enough to restart transcription. One source of this arrest is through the misincorporation of NTPs, which, by requiring TFIIS to stimulate cleavage of the RNA containing the misincorporated nucleotide, acts as a proofreading mechanism for mRNA transcription [223].

1.4 Pre-mRNA splicing

Splicing, the process of removing sections of RNA called introns from the pre-mRNA, is highly conserved amongst eukarya. In S. cerevisiae, fewer than 300 genes are spliced. However, in the few genes that are spliced, the sequence elements responsible for mediating splicing are highly conserved [421]. There are three known sequence signals contributing to intron recognition. The 5’ splice site is recognized by the U1 snRNA within the U1 snRNP [40] and has the consensus sequence gGUAUGU [421]. The underlined G residue of the 5’ splice site undergoes a transesterification reaction with the hydroxyl group on an A residue within the branch site [110, 426]. The branch site is recognized by the U2 snRNA within the U2 snRNP [40] and has the consensus sequence UACUAACA [426], where the A residue that forms the first trans-ester bond is underlined. This forms a free 3’ end on the upstream exon and a “lariat” structure within the intronic RNA [417, 110]. Finally, the free 3’ OH from the upstream exon can attack the first AG motif downstream of the branch site, resulting in covalent linkage between exons and removal of the intronic lariat RNA [455]. However, the efficiency of 3’ splice site selection can be influenced by the presence of a U-rich sequence upstream of the AG [384].

1.5 Transcript termination

There are several complementary pathways that can lead to termination of PolII transcription. These usually involve the cleavage of the nascent mRNA and appropriate modification of the PolII CTD. Following release of the RNA, the polymerase is de-
phosphorylated by two phosphatases (Fcp1 and Ssu72 [275, 86, 267]), falls off the DNA [305], and is recycled back into the pool of unphosphorylated PolIII [86].

1.5.1 Cleavage and polyadenylation

For termination of most mRNAs, PolIII transcribes the sequence elements that direct cleavage and polyadenylation (CPA). Specific factors recognize the CPA site in the RNA and recruit other factors that initiate mRNA cleavage, degrade the nascent uncapped RNA, and terminate transcription.

Factors involved

The transcript termination machinery consists of several distinct sub-complexes. These include cleavage factor IA (CFIA), cleavage factor IB (CFIB), and the cleavage and polyadenylation factor (CPF), which can be further subdivided into cleavage factor II (CFII), polyadenylation factor I (PFI), and several accessory factors [355]. CFIA contains Rna14 (which dimerizes to allow simultaneous binding of two Rna15 and Hrp1 molecules per CFI [165] and is also required for the formation of gene loops [351]; see below), Rna15 (an RNA-binding protein [172]), Clp1, and Pcf11 (a protein that interacts with the PolIII CTD [29, 308]) [152]. CFIB contains a single protein, Hrp1, which is an RNA-binding protein [387]. CPF comprises many proteins, including Yth1 and Yhh1 (RNA-binding proteins [27, 470]), as well as Pap1 (the poly-A polymerase [313]) [152]. These proteins have various structural and enzymatic functions [336], but for the purposes of this thesis, I will concentrate on the RNA-binding components and the sequences they recognize.

Sequence elements

In yeast, there are five sequence elements that are thought to contribute to cleavage and polyadenylation. These include the efficiency element (EE), the upstream U-rich element (UUE), the A-rich positioning element (PE), the downstream U-rich element (DUE), and
the cleavage site [175, 355, 107]. The EE consists of an UA-rich motif and is bound by Hrp1 [252, 80, 498, 387]. This protein prefers to bind three UA repeats, but will also bind UAYRUA (where Y is a pyrimidine and R is a purine) [408]. The PE is an A-rich motif thought to be bound by Rna15 [172], although this protein appears to not be sequence specific in the absence of Rna14 [172], with which it dimerizes [171, 165]. Hrp1 and Rna15 are the RNA-binding components of CFI and, although the complex contains two of each of these proteins [165], the two pairs are unlikely to bind the adjacent sites on the RNA simultaneously, since they are quite separate in space [30]. However, it is possible that these proteins bind to more distant binding sites simultaneously, which would be expected to greatly increase the efficiency of cleavage. The UUE and DUE are both U-rich motifs and are recognized by Yth1 and/or Yhh1 [106, 470, 26]. Finally, the cleavage site itself generally has the consensus \( Y_A^n \) [192] and is probably recognized by the enzyme responsible for mRNA cleavage, Ysh1 [337, 148].

1.5.2 Termination of transcription

Following recognition of the cleavage site and mRNA cleavage, PolIII continues transcribing the DNA template until termination occurs. There are two non-mutually exclusive pathways through which PolIII is thought to terminate. The allosteric model stipulates that, upon reaching a cleavage site, the polymerase is modified to a less processive state leading to termination [319], whereas the torpedo model predicts that a 5’→3’ ribonuclease degrades the nascent RNA until it reaches the polymerase and destabilizes it [94]. The evidence for these mechanisms is discussed further below.

**Torpedo/allosteric models**

Upon mRNA cleavage, the nascent mRNA now lacks a 5’ MeG cap, and so is a potential substrate for the torpedo complex, which includes Rtt103, Rat1, and Rai1. Rtt103 includes a CTD-interacting domain and has a preference for binding Ser2-P CTD [257].
Rat1 is the 5'->3' ribonuclease [228] responsible for degrading the nascent RNA [257], whereas Rai1 binds and stabilizes the other components [539] and includes non-methylated 5'G de-capping activity [226]. Together, these proteins are localized at the 3' ends of genes and cause termination defects when mutated or deleted [257]. Further evidence for this model comes from the fact that termination of transcription is coupled to mRNA cleavage, but not polyadenylation [39, 526, 419, 159]. A similar pathway has also been demonstrated in humans [523], indicating that this mechanism is likely conserved. However, Rat1 is not required for termination of sn/snoRNA genes [258] and is insufficient to terminate transcription of mRNA genes by itself [328]. Since termination can occur without mRNA cleavage [427], another termination pathway must be present.

The alternative pathway is the allosteric model for transcript termination that predicts that transcription across a cleavage site recruits factors that modify the processivity of the polymerase complex. This could be mediated by covalent modification, for instance through modification of the CTD, or through association or dissociation of factors that affect processivity (or both). The facts that termination can be uncoupled from cleavage [375, 427, 376, 14], that the CTD is required for termination [350], and that Pcf11 can dismantle elongation-competent PolII \textit{in vitro} [561] support this pathway. However, because of the conflicting evidence depending on the system being studied, it seems likely that both the allosteric and torpedo mechanisms are used, with different factors being more or less important at individual genes, with factors including CTD state and \textit{cis}-elements present in the RNA determining which pathway predominates. Indeed, one study showed that these pathways are not mutually exclusive, but can act in concert; Rat1 and Pcf11 interact to couple termination and nascent RNA degradation [328]. Altogether, this indicates that termination of transcription for mRNA genes can depend on many factors, but requires the transcription of termination signals.
1.5.3 Non-CPA termination mechanisms

In addition to the standard mode of mRNA termination, there are several other mechanisms the cell can use to terminate transcription that can either result in stable or unstable mRNAs. These are described in more detail below.

Nrd1 complex

The Nrd1 complex mediates two alternate forms of transcript termination, one resulting in a stable RNA product [464] and the other an RNA targeted for immediate degradation [484, 536, 15]. Both pathways involve detection of sequence motifs in the RNA by RNA-binding members of the Nrd1 complex while the PolIII CTD contains the marks characteristic of the beginnings of transcripts [120]. The members of the Nrd1 complex include Nrd1, Nab3, and Sen1 [464]. Nrd1 and Nab3 are both RRM-domain containing RNA-binding proteins that recognize GTAR and UCUU sequences in the RNA, respectively [69]. Nrd1 also contains a CTD-interacting domain with which it can bind Ser5-phosphorylated PolIII CTD [552, 279]. Sen1 is an RNA-helicase and is thought to be involved in unwinding DNA-RNA-hybrids (R-loops) in the wake of the elongating polymerase [255]. Sen1 also appears to play a role in poly-A dependent termination, potentially unwinding R-loops following cleavage to facilitate degradation of the nascent RNA [246].

For unstable transcripts that are terminated by the Nrd1 complex, the TRAMP (Trf4, Air1/2, Mtr4) complex is recruited, which initiates RNA cleavage, adds a short poly-A tail to the 3' end of the RNA, and facilitates RNA degradation by the exosome complex [503, 284, 536], a nuclear 3'-&gt;5' exonuclease [356]. Under normal cellular conditions, this process takes place so quickly that the short RNAs degraded by this mechanism have an extremely low abundance. Indeed, these RNAs, referred to as cryptic unstable transcripts (CUTs), were discovered in an exosome mutant where they were seen to accumulate in the cell [536].
There are also transcripts for which the Nrd1 complex plays a role in maturation but not destruction of the transcript, including sn/snoRNAs [464]. Here, these factors play nearly the same role, except that the RNA is not completely degraded by the exosome. Instead, the mature sn/snoRNPs appear to form concomitantly with transcription [541, 24, 359], which protects the mature 3’ end from complete degradation by the exosome.

**Other termination mechanisms**

There are also other mechanisms for mediating mRNA 3’ end formation in the cell, although usage of these alternative mechanisms is presumed to be rare. For instance, the human histone genes include a complex RNA structure at the 3’ end that is recognized by SLBP (stem-loop binding protein) that initiates cleavage following this motif [341]. Presumably, this allows the histone genes to be strictly regulated because the non-polyadenylated RNAs are so unstable. In yeast, the ribonuclease III homolog (Rnt1) is involved in rRNA processing [121], as well as 3’ end formation for snRNAs [1, 76] and certain mRNA genes [159]. This protein recognizes hairpin loops in the RNA and cleaves the RNA immediately following the stem loop. Rnt1 has frequently been used experimentally to cause transcriptional termination without polyadenylation [159, 419, 478].

**1.5.4 Gene loops**

Recently, it has been discovered that the genome has many stable long-range interactions [113]. One such structure is a so-called “gene-loop”, which is a physical connection between the promoter and terminator regions of a gene. The existence of these loops was first suggested by the observation that certain promoter-specific factors (such as GTFs) could also be found in terminator regions [366, 65, 256, 347, 344]. This was later confirmed to be due to a structure linking these two regions [377, 12]. These structures have been shown to depend on various factors, including proteins involved in PIC formation (e.g. TFIIB [477]) and cleavage and polyadenylation (e.g. Ssu72, Pta1, Rna14,
Pap1, Pcf11 [12, 351]), and seem to require proper 3' end formation, since cleavage and
termination by a non-poly-A mechanism does not result in gene loop formation [478].

The physiological role of these gene loops remains obscure. However, one study
showed that gene loops can facilitate a kind of “transcriptional memory”, wherein a gene
induced by a stimulus becomes activated more rapidly upon a second exposure to the
stimulus, and that this response is dependent on the maintenance of a gene loop structure
[286]. Gene loops have also been implicated in maintaining promoter directionality: when
a gene loop is disrupted, transcription of the sense mRNA is decreased, and production of
an antisense RNA originating from the same promoter increased [478, 338]. This suggests
the existence of a feedback mechanism wherein successful polyadenylation reinforces the
choice of promoters.

1.6 Predicting gene structure

Predicting how DNA sequences are interpreted by the transcription machinery is integral
to the study of the genome, since it is a direct test of our understanding of the sequences
and factors involved. Much work has been done in creating computational models to
predict transcription based on gene sequence. These have primarily focused on four
areas: predicting ORFs, predicting promoters, predicting cleavage sites, and predicting
gene expression. Historically, these areas have been treated separately, even though there
is significant evidence to suggest that there is cross talk among some of these processes
in the cell (see, for example, sections on chromatin and gene looping). If our aim is to
understand how the cell identifies genes, mechanistic models must be created, restricting
the features to those that can realistically be used by the cell. One of the current
limitations in the literature is that, even when these mechanistic models are created, the
predictions of the models are not tested, and so we remain unsure as to whether the
features recognized by the models are properties of genes that only we recognize or the
gene-defining sequences recognized by the cell. However, the aim of many of these models (particularly the non-mechanistic models) is not to understand how the cell defines its genes, but to identify genes for later study, and, in this way, these models have been very successful.

1.6.1 Predicting translational units

A great deal of work has been done in predicting translational units, given the sequence of the genome. This started in bacteria, where it is relatively straightforward to identify many genes in genomic sequence. These take advantage of several features that are conserved across both genes and genera, including the presence of uninterrupted open reading frames (ORFs) and ribosome entry sites [277, 327]. In eukaryotes, the problem is significantly harder because most genes lack ribosome entry sites, the genes are interrupted by introns, and the search space is generally larger. In spite of this, eukaryotic gene finders have been quite successful. Identifying ORFs using gene-finders is a common first step to annotating a new genome sequence (e.g. [74, 47]). Many of the most successful eukaryotic gene-finders rely on hidden Markov models (HMMs) [6, 334, 58]. An HMM is a type of probabilistic graphical model where the states represent the possible classification states (e.g. start codon, stop codon) and the set of observations (e.g. nucleotide sequence) informs the model as to what state it is in at every position. HMMs are ideal models for predicting gene structure because they can capture ORF structural elements, such as the start codon, splice sites, and stop codons, and work directly from the DNA-sequence. However, the cell does not identify translational units directly: no known mechanism exists for the cell to recognize ORFs in the DNA sequence. This means that, as far as we know, the primary gene-identification task the cell undertakes at the level of the genome is identifying what DNA to transcribe. Thus, while these are excellent tools for identifying genes in genomic DNA, they are unable to tell us how the cell performs the same task.
1.6.2 Predicting promoters

Predicting promoters from the DNA sequence is a well-studied area of research. Many programs exist to detect promoters in eukaryotic DNA. These rely on various features, including: GTF binding sites [115, 352]; conservation [458]; nucleotide content [115, 22, 459, 352]; TF binding sites [399, 352]; and k-mer counts [23, 459, 142]. However, it is often unclear what exactly the features correspond to in the cell, complicating our understanding of cellular promoter recognition and making hypothesis derivation and testing difficult. For that reason, it is important to ask how well we can identify promoters given the tool set of the cell, namely, the GTFs, TFs, and G/C-content (through its effect on nucleosome occupancy). Megraw et al. used mechanistic features to identify human promoters in genomic sequence, but restricted their input set to a small number of well-defined promoters and relied heavily on detecting the CpG islands that characterize many human promoters [352]. In this study, the TATA box, G/C-content, and initiator motif seemed to be the most important features. Importantly, neither the Megraw study, nor any of the others, have tested their predictions \textit{in vivo}.

1.6.3 Predicting cleavage sites

Models have been created that predict cleavage sites using various features, including predicted nucleosome occupancy [197], nucleotide content [197, 238], the canonical sequence elements [168, 295, 238], and k-mer counts and other \textit{de novo} motifs [316, 83, 469]. However, few have targeted yeast cleavage sites, which differ significantly in architecture from those in other eukaryotes [355]. One study using yeast as a model system employed an HMM with the canonical termination sequence elements to identify cleavage sites, using the presence/absence of the canonical sequence elements as inputs, and thus including few input features [168]. This model met with some success, but was not systematically evaluated, and the predictions of the model appeared to be quite noisy. Further, neither this nor the other CPA site prediction models evaluated the mechanistic predictions of
1.6.4 Predicting gene expression

Many groups have created models of gene expression. Some attempt to predict the expression of individual genes [62, 306], but others aim to determine which of several expression profiles characterize a given gene [35, 441]. The plethora of genomic data available in yeast has made it an ideal system with which to attempt to predict gene expression given gene sequence. At its core, the gene expression problem attempts to predict the abundance of the regulated gene product (e.g. mRNA levels) by identifying its relationship to the regulating proteins. Many data types can be used to infer the values of the variables at different stages in regulation, including the expression levels of regulators [45], specificities of TFs for target promoter sequences [35], sequences common to co-regulated genes [248, 514, 62], measured binding of TFs to target promoters (e.g. ChIP) [147, 306], and expression data of targets [147, 62, 306]. Both linear (e.g. [147]) and non-linear (e.g. [35]) modelling approaches have been used and have met with comparable success, although the linear models tend to be simpler. However, there is no reason to suspect that gene expression is simple in the cell. Indeed, several of these studies have stressed the importance of combinatorial interactions among TFs and low affinity binding sites [479, 157, 35]. If these phenomena are widespread, it may be impossible to learn the rules of gene regulation with only the native complement of promoters, since the number of possible interactions between TFs greatly outnumbers the number of native promoters.

One strategy to circumvent this problem is to make synthetic promoters (e.g. [157, 407, 445]). These can be employed to help learn cis-regulatory logic by providing many more examples from which to learn. One study used a synthetic library of yeast promoters (made by concatenating TF binding sites) to create a thermodynamic model of gene regulation, and used the model to predict expression of native genes [157]. Another demonstrated, by testing the expression of designed promoter sequences, that a large
number of variables seem to affect transcription, including distance from the TSS, binding site number, binding site orientation, proximity to nucleosome disfavouring sequences, and binding face on the double helix [445]. If true, all of these factors will need to be incorporated into future models to faithfully capture gene regulation. In general, these approaches use a basal promoter [157, 407, 445], and it remains to be seen if the rules learned using these constructs translate to other promoters.

1.7 Evolution of genes

An important evolutionary question asks how new genes arise. Fundamental to this question is defining what constitutes a gene in the transcriptional sense. There are several mechanisms for the derivation of new genes, including the duplication and divergence of existing genes by errors in DNA replication, recombination, or repair [174], which can include both genes and their regulatory regions. Another mechanism is through retroposition [236], which could include some of the cis elements required for cleavage and polyadenylation and most often does not include promoter sequences (although exceptions have been found [373]). A third possibility is through de novo gene birth, wherein genes arise from essentially random intergenic DNA.

1.7.1 De novo gene birth

Although de novo gene birth was originally thought to be extremely uncommon because of the statistical improbability of deriving a functional unit from non-functional DNA [218], recent evidence suggests it may be more common than previously appreciated. Specific examples of genes that appear to have originated from non-functional transcribed DNA have been found [63, 300], but evidence is increasing that this phenomenon is widespread [566, 71]. A recent study used ribosome profiling data [210] to show that many putative ORF-containing transcripts appear to be translated, and many of these showed signs of
selection across yeast species, with a continuum of conservation, suggesting a continuum of functionality [71]. This suggests that proto-genes may first acquire the ability to be transcribed, gain the ability to be translated, and finally gain a useful function which can be selected for [63]. However, because the mechanisms by which the cell defines transcriptional units is not fully known, the first stage of this process remains opaque.

1.8 Summary and thesis outline

Transcription is thought to rely on the detection of specific sequence elements by cellular factors. Initiation of transcripts occurs in the promoter region, which is characterized by having TF binding sites, the TATA box, an NFR, and certain histone marks. At the initiation stage, the PolIII CTD is modified into a state characteristic of elongation. Once elongation begins, nucleotides are added in bursts, interrupted by PolII pausing, and the CTD continues to be modified across the length of the gene. Meanwhile, nucleosomes are replaced on the DNA in the wake of PolII transit and the histones comprising them are chemically modified. The yeast terminator contains binding sites for the RNA-binding components of the cleavage machinery, which initiate transcript cleavage if the CTD is in a termination-competent state. Alternative (non-poly-A) termination pathways function similarly, but rely on different sequence and CTD signals. Following mRNA cleavage, PolIII terminates transcription, the CTD is dephosphorylated, and the polymerase is recycled. These stages of transcription can interact in various ways. (e.g. through histone modification, transcriptional interference, CTD modification, and gene looping). Altogether, transcription appears to be a complex process with many factors involved and many potential interactions.

In this thesis, I describe my work that aims to determine what sequences and features are involved in demarcating genes in the genomic sequence. In Chapter 2, I describe my collection of high-confidence yeast TF specificities, created by combing the literature
for TF motifs and evaluating them to see which best represent the specificities of the TFs. This provided me with a high-confidence, low-redundancy set of specificities to use for modelling and identifying putative regulators. In Chapter 3, I describe my efforts to identify the sequence elements associated with yeast transcripts. I then describe potential roles these sequences may play in defining gene structure. In Chapter 4, I create computational models of transcription in yeast, simulating the decisions made by the transcription machinery by relying solely on features that could be used by the cell, such as the TF collection described in Chapter 2. This model predicts that gene structure is relatively simple in nature and provides evidence that the interactions between structural elements in cis is of widespread importance.
Chapter 2

A compendium of transcription factor specificities

A modified version of this chapter is encompassed in the following publication, which is available under the terms of the Creative Commons Attribution Non-Commercial License:


I performed all the work described except for the expert curation step, which was performed almost entirely by Tim Hughes. I had help from Harm van Bakel, Kate Cook, Mihai Albu, and Matt Weirauch when setting up the website, and Matt Weirauch provided me with his DBD predictions which I used in conjunction with my own. The writing was performed by myself and Tim, but I am also grateful to Jack Greenblatt, Alan Moses, Harm van Bakel, and Matt Weirauch for comments on the original manuscript.
2.1 Summary

The yeast *Saccharomyces cerevisiae* is a prevalent system for the analysis of transcriptional networks. As a result, multiple DNA-binding sequence specificities (motifs) have been derived for most yeast transcription factors (TFs). However, motifs from different studies are often inconsistent with each other, making subsequent analyses complicated and confusing. Here, I describe YeTFaSCo (The Yeast Transcription Factor Specificity Compendium, http://yetfasco.ccbr.utoronto.ca/), an extensive collection of *S. cerevisiae* TF specificities. YeTFaSCo differs from related databases by being more comprehensive (including 1709 motifs for 256 proteins or protein complexes in version 1.00), and by evaluating the motifs using multiple objective quality metrics. The metrics include correlation between motif matches and ChIP-chip data, gene expression patterns, and GO terms, as well as motif agreement between different studies. YeTFaSCo also features an index of “expert-curated” motifs, each associated with a confidence assessment. In addition, the database website features tools for motif analysis, including a sequence scanning function and precomputed genome-browser tracks of motif occurrences across the entire yeast genome. Users can also search the database for motifs that are similar to a query motif.

2.2 Motivation/introduction

The yeast *S. cerevisiae* is a powerful model for the study of gene regulation, and one in which numerous computational and experimental approaches to the study of transcriptional networks have been field-tested and applied on a large scale [185, 410, 293, 20, 150, 569, 140]. As a result, there has been some level of characterization of the sequence specificity of most yeast transcription factors (TFs). A TF’s sequence specificity, or “motif”, is frequently represented as a Position Weight Matrix (PWM) whose entries represent the log-odds ratio of bases being part of the motif, relative to the background
sequence, which is generally taken to represent the relative preference of the corresponding protein for that sequence [465]. It is desirable to have a comprehensive collection of yeast TF motifs for use in a variety of purposes, including the computational analysis of transcriptional networks (e.g. [35]) and study of genome evolution (e.g. [495]). However, different published motifs for the same TF often conflict and may not represent the TF’s true intrinsic sequence preferences, thus potentially confounding many studies that use the motifs.

Here, I have created YeTFaSCo, a database of yeast TF sequence specificities, obtained from diverse sources. Tim Hughes and I have evaluated the motifs’ predictive power and consistency with a variety of sources, including genome-wide studies, knowledge of the types of sites that different structural classes of TFs can and cannot bind, and detailed studies from the literature. To my knowledge, no similar resources exist. UniPROBE [368] contains only Protein Binding Microarray (PBM) data, while YEASTRACT [481] does not contain any PBM data. YPA [77] and MYBS [493] have collected motifs from several different sources, but concentrate on using these motifs to predict genomic binding sites and regulatory associations. Perhaps the most commonly-used index of yeast TF motifs - the MacIsaac collection [329] - contains only motifs from ChIP-chip data. While Transfac [343] and JASPAR [428] compile motifs from the literature, neither these nor the aforementioned collections evaluate the motifs for predictive power. Other recent studies [79, 138, 360, 379, 570] have also compiled motifs, and in some cases evaluated them for consistency with external information, but do not aim to comprehensively survey the literature, or to evaluate the motifs against each other and against multiple data types. Our evaluation methods go beyond previous studies because (a) I evaluate the motifs using four independent criteria, and (b) we include a manual curation step, producing a collection of non-redundant motifs that are annotated with expert confidence ratings (as evaluated by Tim). YeTFaSCo also incorporates tools for scanning new sequences for PWM matches, browsing the genome for potential binding
sites, and comparing among motifs. I anticipate that, as a unique resource, YeTFaSCo will be invaluable to a wide variety of researchers.

2.3 Generation of the database

YeTFaSCo has two central tables that are related to each other. One is a table of all genes/proteins and their encoded DNA-binding domains (DBDs), if any, and the other is a table of motifs assigned to these (and other) proteins. There are often multiple motifs associated with each TF, but typically only a single TF associated with each distinct motif, unless the TF binds as a part of a complex. I considered any yeast protein with either a DBD or an associated DNA-binding motif to be a potential TF. The current version of YeTFaSCo contains 264 known and putative TFs (248 with motifs + 16 with DBDs, but no motifs yet described) and eight TF complexes (with motifs).

To assign DBDs, the union of three sets of domain predictions was taken, including those from Badis et al. [20], those from Weirauch et al. [522], and my own predictions made by scanning all yeast genes with the HMMs from the Pfam [132], SMART [298], and SUPFAM [381] databases that correspond to the Weirauch DBD set [522]. Similarly, I populated the database with motifs using several approaches. First, existing databases containing motifs were used to direct the search for motifs [20, 368, 481, 138, 360, 379, 570, 122]. I re-extracted motifs in these databases from the primary literature and documented the assays used to derive them. Next, I performed general searches of the literature, looking for papers that had derived motifs for any yeast proteins (Table 2.1 - Query Type “General”). Next, a reference-directed approach was taken. Here, for each popular biochemical assay from which motifs can be derived, I searched for publications that included the original publication to describe this method as a reference. This limited my search to those papers which were likely to derive motifs using these methods (Table 2.1 - Query Type “Reference-directed”). These methods included ChIP-chip [410], BIACore
Finally, for every putative TF in the collection that had zero or one motifs derived for it, I performed a directed search, looking for motifs specifically for these factors (Table 2.1 - Query type “Gene-directed”).

<table>
<thead>
<tr>
<th>Query Type</th>
<th>Engine</th>
<th>Example Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Scholar</td>
<td>yeast OR saccharomyces OR cerevisiae motif OR pfm OR logo OR pwm &quot;transcription factor&quot; OR &quot;DNA-binding&quot; -intitle:human -intitledrosophila</td>
</tr>
<tr>
<td></td>
<td>Scholar</td>
<td>motif specificity sequence &quot;transcription factor&quot; saccharomyces OR cerevisiae &quot;DNA binding&quot; -intitle:arabidopsis -intitlesubtilis -intitledrosophila -intitle:human -intitle:prokaryotic -intitle:mouse -intitle:albicans logo OR PWM OR PSSM OR PFM</td>
</tr>
<tr>
<td></td>
<td>Scholar</td>
<td>&quot;transcription factor&quot; motif specificity cerevisiae &quot;DNA-binding&quot; -intitledrosophila -intitle:human -intitle:plant -intitle:mammal</td>
</tr>
<tr>
<td></td>
<td>Pubmed</td>
<td>transcription factor (motif OR specificity OR pwm OR PFM) (cerevisiae OR yeast) &quot;DNA-binding&quot;</td>
</tr>
<tr>
<td>Reference-directed</td>
<td>Scholar</td>
<td>DNA binding cerevisiae &quot;transcription factor&quot; PWM OR PFM OR PSAM OR PSSM OR &quot;sequence specificity&quot; motif</td>
</tr>
<tr>
<td>Gene-directed</td>
<td>Scholar</td>
<td>&lt;GeneName&gt; OR &lt;SysName&gt; † DNA binding cerevisiae &quot;transcription factor&quot; motif OR PWM OR PFM OR PSAM OR PSSM OR consensus OR &quot;sequence specificity&quot; OR &quot;binding site&quot;</td>
</tr>
</tbody>
</table>

† The “intitle:” term is used to exclude papers with a given term in the title.
‡ &lt;GeneName&gt; and &lt;SysName&gt; were replaced with the gene and systematic name of the gene being searched for.

For each motif, I converted from the provided form to a position frequency matrix (PFM), the standard motif form used in this database. The entries in a PFM represent the frequency of observing each base at a given position in the motif, and thus the frequencies for each position in the motif sum to 1. I chose this motif form because it is a simple, yet robust model of TF specificity, most other motif representations can be converted to PFMs with relative ease, and there are many tools developed which use either PFMs or PWMs. PFMs can easily be converted to PWMs by dividing each entry by the corresponding background base frequency and changing to a log scale [465].
The PWM format facilitates scanning DNA sequences by, for every possible alignment of the motif to the sequence, providing the log-odds ratio of the subsequence being an instance of the motif, vs. part of the genomic background [465]. Thus, scores above zero represent sequences more likely to be an instance of the motif than random DNA. As noted above, the same calculation is widely taken to represent the relative affinity of a TF to a particular sequence the same width as the PFM/PWM. Some of the motifs in the database have flanking bases with low information content. Trimming the motifs to remove low information content bases did not on average improve the motifs by our criteria, however (Figure 2.1), so the motifs were left in their original form.

In total, YeTFaSCo (version 1.02, the current version) contains 1880 motifs for 301 proteins or protein complexes, which were derived from 137 publications. As many as 445 motifs come from a single publication [138], and individual TFs have as many as 41 different associated motifs (Ste12). The motifs present in the database were derived from diverse data types, but ChIP-chip has by far the greatest number of motifs (1193), with many of these motifs being derived from the same source data using different algorithms [185], resulting in many TFs with multiple ChIP-derived motifs (Figure 2.2A).

### 2.4 Evaluation of motifs

I used four objective criteria to give a confidence value for the accuracy of each motif. These criteria include the correlation of predicted binding sites and ChIP-chip data, the correlation between predicted binding sites in promoters and expression changes in TF mutants, the enrichment of GO terms in genes whose promoters have binding sites, and the agreement between different studies. These criteria are described in more detail here.
Figure 2.1: Comparison of the scores of motifs for the different evaluation criteria before and after trimming low information content (IC ≤ 0.2) flanking bases.

2.4.1 ChIP-chip enrichment

For 212 proteins in the database, genome-wide chromatin immunoprecipitation data is available [185, 529, 474, 506]. I used these data to test the quality of the motifs by calculating the Spearman correlation coefficient between the relative probe intensities and the probability of the probe region being bound by the motif. When the ChIP experiments used entire intergenic regions as microarray probes, the “probe region” to be scanned was defined as the entire probe sequence, and when the probes represented
Figure 2.2: Comparison of ChIP and PBM-derived motifs. (A) Histogram showing the number of motifs per TF broken down by derivation method. (B) Comparison of summary scores (see text) for the highest-scoring motifs derived independently by both PBMs and ChIP-chip for 112 TFs common to both. Each point represents a single TF for one of the four scoring measures, the total score, or the average of all ChIP data, excluding the Harbison et al. [185] data, from which most of the ChIP motifs were derived.
equally sized short sequences, the “probe region” was defined as the maximum range of DNA around the probe that could fully hybridize to the microarray probe given the published upper size limit of the sheared ChIPed DNA (e.g. If the DNA was sheared to a maximum size of 300 bp, and the probe was 60 bp and started at \( x \), the “probe region” was taken as \( x - 240 \) to \( x + 300 \)). The probability of the probe region being bound is calculated as the probability that at least one binding site in the sequence is occupied, given the motif \cite{82}. The P-value of the correlation between relative probe intensities and the probability of that probe region being bound was calculated using the Edgeworth series approximation method \cite{87}. As a summary score for each motif, I calculated the average \(-\log(P\text{-value})\) for the correlations over all ChIP-chip datasets. The distribution of these scores, in comparison to the scores for 1,000 permutations of the probe intensities for each ChIP experiment, is shown in Figure 2.3A. Only 1% of the randomized data scored above 1.4, in contrast to 67% of the actual data. Thus, we use a cutoff of 1.4 to distinguish motifs that significantly correlate with ChIP data from those that do not, representing an empirically-determined 1% FDR.

Many of the motifs in the database were derived from the same ChIP-chip data being used to evaluate it \cite{185}. This circularity would be expected to bias the analyses in favour of ChIP-chip motifs. However, comparison of the highest-scoring motif for each TF derived by ChIP-chip to the highest-scoring motif derived by PBM (the second most abundant motif derivation method in the database) revealed that, for the 112 TFs with motifs derived by both methods, PBMs perform slightly better overall; 60 of the 112 motifs have a higher total score for PBMs (Figure 2.2B). This is in spite of the fact that there is a much larger pool of ChIP-chip motifs from which to choose the best motif (Figure 2.2A). One possible explanation is that motif derivation from ChIP-chip data faces the inherent difficulty of searching for short motifs in long stretches of non-random DNA. However, I cannot exclude a bias in the evaluation criteria, or the binding model in favour of PBM-derived motifs.
2.4.2 Correlation with gene expression data

Several studies have used microarrays to systematically examine the effects of TF overexpression and/or deletion [204, 194], and many others have analyzed one or a few TF mutants. I downloaded expression data from systematic studies [204, 194] and individual studies included in the SPELL collection [340], giving us data from 58 sources that include mutant expression data for 212 of the TFs in YeTFaSCo. These data are useful
for evaluating motif quality since genes with TF binding sites in their promoters would be expected to have their expression perturbed in the corresponding TF mutant. I scanned promoters (taken from -500 to +100 relative to the transcription start site) with each motif using the same binding model described for the ChIP-chip Enrichment criterion to yield a probability of each promoter being bound by the TF. Similar to the ChIP-chip metric, I then calculated the Spearman correlation coefficient between the probabilities of the TF binding each promoter and the log expression changes in the corresponding genes, with a P-value being derived as above. The summary score for this criterion is the mean of the correlation -log(P-value)s for all available mutant expression datasets. The distribution of these scores, in comparison to the scores derived from 1,000 permutations of the fold expression changes for each experiment, is shown in Figure 2.3B. Only 1% of the randomized data scored above a threshold of 1.3, compared with 36% of the actual data.

2.4.3 GO term enrichment

TFs often regulate specific pathways and processes [204]. To test for enrichment of binding sites in promoters of functionally-related genes, I calculated the probability of the TF binding to each promoter for each motif, and performed AUROC and rank sum tests for each motif-GO slim term combination to ask whether binding probabilities differ between genes that are annotated with the GO term and those that are not. Both enriched and depleted ROCs are considered because, in addition to having certain TFs responsible for activating certain processes, it is possible that there are certain processes which specifically lack certain motifs. In general, one would expect this latter case to be uncommon. Indeed, using a 1% FDR, only 49 motifs show a significant depletion in a GO slim term, compared with 216 that are enriched. Most of these former motifs are depleted in either “ribosome biogenesis” (18/49) or “RNA metabolic process” (13/49), and many (5/18 and 12/13, respectively) of these are very similar to AGGGG, the “stress response
element” and known Msn2/Msn4 binding site [166]. Since Msn2 and Msn4 are activated in stress conditions [72], it is not surprising that these sites are generally absent from genes involved in ribosome biogenesis, since these genes are generally repressed under times of environmental stress [149, 176]. The score for this criterion is the -log(P-value) of the rank sum test for the most significantly enriched/depleted GO term. The distribution of these scores, in comparison to 1,000 permutations of the GO term labels, is shown in Figure 2.3C. Only 1% of this randomized data scored above 4.0, in contrast to 25% of the actual data.

2.4.4 Inter-study concurrence

I used the concurrence between independent studies as a measure of a motif’s reliability. For instance, if two studies independently characterize the same specificity for the same protein, this is strong evidence that the motif is correct. Using Tomtom (version 4.1.0) [330], I compared each motif to the other motifs derived for the same TF (from independent data) using the Euclidean distance metric, yielding a P-value representing how likely each match is, given the similarity of the two motifs. This P-value, negated and logged, is used as the score for this metric. The distribution of these scores, in comparison to 1,000 permutations of the gene labels for the motifs, is shown in Figure 2.3D. While 1% of this randomized data scored above a threshold of 5.3, 26% of the actual data surpassed this same threshold.

2.5 Manual evaluation of motifs

For most purposes, a library of TF motifs would ideally include the single most accurate motif for each TF. For example, in computational modelling, it is desirable to have as few features as possible, while still having a comprehensive list of relevant features. With this motivation, we created an “Expert Curated” motif set, which consisted of one motif
for each TF, or more than one if the TF appears to have multiple binding modes (e.g. some GAL4-class TFs appear to tolerate more than one spacing between the half-sites, and/or can bind as either monomers or dimers [531], while some bZIP proteins appear to tolerate different spacings between the half-sites, and can heterodimerize in different combinations).

Because the data available differs for each TF, and in many cases the data is not easily comparable among TFs or even for the same TF, no one-size-fits-all objective system can be applied automatically to all TFs; this fact motivated the expert curation step. For each TF, Tim Hughes used the evaluation criteria to select the best representative motif(s) and rated the confidence in the motif as “high”, “medium”, or “low”. Consequently, the expert curations are subjective to a degree. However, the following procedures were employed. First, using the Table tools described below, Tim manually examined all of the motifs and scores for each TF in the database and also considered additional information in the literature (such as DNaseI footprinting and reporter assays), as well as knowledge regarding the types of sites that are normally bound by the different structural classes of DNA-binding domains. For example, monomeric GAL4-class zinc clusters typically bind sequences containing CGG, and these proteins also often bind as dimers that prefer a specific spacing and orientation of the two CGG subsites [531]. Tim attempted to avoid selecting motifs that were likely due to cofactors (e.g. some of the motifs that have been reported for Ace2, Fhl1, Yap5, Pdr1, and Sfp1 are, in fact, Rap1 motifs; see Table 2.2). He also considered whether or not the motifs were supported by more than one criterion (e.g. if a motif was scoring highly simply because of a high correlation to ChIP-chip data, but did not perform well by any other measure), and whether the data used to derive the motif reflects the protein binding directly to the DNA. In addition to the confidence score, some putative TFs were also assigned “Dubious” status in the database, due to lack of evidence that they bind DNA directly or in a sequence-specific manner; motifs for these TFs are not included in the expert curated set. The full details
Table 2.2: Examples of potential TF cofactors. All motifs were derived from ChIP-chip data.

<table>
<thead>
<tr>
<th>Potential Co-factor</th>
<th>Motif IDs</th>
<th>TF known to bind motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ash1</td>
<td>648, 932</td>
<td>Mcm1</td>
</tr>
<tr>
<td>Rlm1</td>
<td>1079</td>
<td>Mcm1</td>
</tr>
<tr>
<td>Ndk1</td>
<td>366</td>
<td>Mcm1</td>
</tr>
<tr>
<td>Arg81</td>
<td>1507</td>
<td>Mcm1</td>
</tr>
<tr>
<td>Arg80*</td>
<td>1483</td>
<td>Mcm1</td>
</tr>
<tr>
<td>Fhl1</td>
<td>406, 629, 893, 1196, 1504, 1618</td>
<td>Rap1</td>
</tr>
<tr>
<td>Sfp1</td>
<td>357, 621, 1100, 1710</td>
<td>Rap1</td>
</tr>
<tr>
<td>Pdr1</td>
<td>899</td>
<td>Rap1</td>
</tr>
<tr>
<td>Yap5</td>
<td>896</td>
<td>Rap1</td>
</tr>
<tr>
<td>Ace2</td>
<td>918</td>
<td>Rap1</td>
</tr>
<tr>
<td>Cha4</td>
<td>1607</td>
<td>Rap1</td>
</tr>
<tr>
<td>YDR026C*</td>
<td>408, 696, 1160, 1581, 1921</td>
<td>Reb1</td>
</tr>
<tr>
<td>Stb2</td>
<td>710</td>
<td>Reb1</td>
</tr>
<tr>
<td>Rpm4</td>
<td>1090</td>
<td>Reb1</td>
</tr>
<tr>
<td>Pho2</td>
<td>1680</td>
<td>Abf1</td>
</tr>
</tbody>
</table>

*Homologous to TF known to bind motif.

of the Expert Curated set are available from the “Expert Curation” link on the side bar of the YeTFaSCo home page. In total, the expert curated set (version 1.02) contains 231 motifs for 193 TFs (most of which are a 1-to-1 mapping; there are five instances of protein complexes and 33 TFs that have multiple motifs, due to evidence for multiple binding modes such as monomeric vs. dimeric). 143 of the motifs are high confidence, 67 are medium confidence, and 21 are low confidence. If we consider the set of all 246 known+putative yeast TFs to include those that either have a motif in the database or contain one of the canonical eukaryotic TF DBDs [522], and are not “Dubious” in our judgement, then 78% of all known+putative yeast TFs have a motif in the “expert curated” set, 58% of which are high confidence.

2.6 Content and information retrieval

From the main YeTFaSCo page (http://yetfasco.ccbr.utoronto.ca), there are links to the five table views, a downloads page, a place for users to submit data, a cart, help pages, a website tour, and sequence analysis tools (described below). The table views contain
all the data in the database and provide links between tables for easy navigation. While navigating the database tables, the cart can be used for keeping track of specific motifs of interest, which can then be downloaded. Alternatively, the downloads page provides access to all the motifs in the database in multiple formats. These three features are described here in more detail.

2.6.1 Tables

The data in the database is accessible through several tables, which are the primary means of browsing the motifs. The “Motifs” table displays an abbreviated version of the data in the database (Figure 2.4). This is the primary method of seeing which motifs are in the database and displays the TF names, motif IDs, a logo representation of the motif [463], the score for that motif, any DBDs that TF might have, the study responsible for derivation of the motif, what biochemical approach was taken to generate the motif, and the expert confidence in the motif (if it is in the “expert curated” set). This page links to several other tables in the database, including the “Gene” table, which shows more information about the genes, the “Expert Curation” table, which shows the details of the expert curation, the “Reference” table, which shows more details about the particular study from which the motif is derived, and the “Motifs - Details” table which shows all the information of the “Motifs” table, with additional details of how that motif scored for the various evaluation criteria. Additionally, the “Motifs - Details” table provides links to the detailed breakdown of the ChIP-chip and expression enrichment scores, broken down by dataset. Each of these tables can be filtered and sorted by adding criteria to the filter bar and by clicking the header links, respectively.

2.6.2 Cart

In the “Motifs” and “Motifs - Details” tables, the final column has buttons to allow the user to add motifs to their cart. This cart is useful for downloading only a subset of
Figure 2.4: The “Motif” table of YeTFaSCo. For each of the database tables, there is a link to the help pages associated with that table in the upper right (1). In the help, there is an explanation of the meanings of each column. For all tables, results can be filtered by entering criteria into the filter bar (2) and pressing “Filter!”. The sort order can be changed by clicking on the various header links (3). For the motif table views, there is an option to add (or remove) motifs from the cart (4). There are also links provided for downloading the table data in various formats (5). The “To Detailed View...” button (6) switches to the “Motifs - Details” view, with the same filters applied. This view can also be reached for an individual motif entry by clicking the logo (7). The table also links to several different pages, including the expert curation (8) where additional details of the expert curation are provided, the gene entry (9) where details of the particular gene can be accessed and a link to SGD [122] is provided, and the reference entry (10) where more details of the study can be viewed, together with links to the corresponding PubMed entry. In addition, the DBDs (11) link to the corresponding entry in Interpro [209] (or other domain databases).

motifs as PFMs. The contents of the cart can be viewed through the “View Cart” link in the header, which leads to a page where its contents can be edited or downloaded; in addition, users can scan a custom DNA sequence with the motifs contained in the cart.
2.6.3 Downloads

There are several downloads available from the downloads page, including all the motifs in the database as IUPAC, PFM, PWM (using the S. cerevisiae base content), and logo [463] representations. In addition, PFM and PWM sets are available for the “Expert Curated” set, which should facilitate the use of the motifs in the database for custom sequence analysis. The download page also provides access to all microarray and ChIP-chip data used in this study, as well as the \textit{in vivo}, \textit{in vitro}, and predicted nucleosome occupancy tracks, and conservation track present in the genome browser (see below).

2.7 Analysis tools

The YeTFaSCo website provides several analysis tools, which include sequence scanning with user-defined sequences, a utility for comparing a user-specified motif to the motifs in the database, and precomputed genome-wide TF binding sites available in a Genome Browser.

2.7.1 Sequence scanning

The sequence scanner identifies potential TF binding sites in user-defined sequences by scanning with different subsets of motifs in the database for sites that are more like the motif than like the background. This can be customized to be more or less stringent by specifying a percent of the maximum PWM score to use as a threshold (e.g. a threshold of 100\% would show only perfect motif matches, while 0\% shows all potential matches). By default this threshold is set to 75\%. In addition, the user can also change the background base content to suit their needs.

There are three ways to choose subsets of motifs with which to scan. The first is to use one of the provided motif sets (e.g. “all motifs” or “Expert Curated”). The second is to select individual motifs by adding them to the cart, and then scanning with the cart
Figure 2.5: Sequence scan of GAL1-10 promoter region. Only binding sites that achieve >75% of the maximum PWM score are shown. Previously characterized binding sites [249, 135] are boxed in red and blue. The numbers next to the TF name represent the motif ID, while the +/- represents the binding site orientation.
2.7.2 Motif similarity search

The YeTFaSCo website also has a tool for finding motifs that are similar to a user-provided motif. I anticipate this will be useful for instances where a potential regulatory motif is found, but the trans-acting factor is not known. Indeed, I myself am a frequent user of this tool for this exact purpose. To use this tool, users can input an IUPAC consensus motif, sequence alignment, or PFM. Using Tomtom [330] (as before), the most similar motifs are found and provided in table format in descending order of significance (until $P > 0.05$).

2.7.3 Genome browser

I scanned the yeast genome with the “Expert Curated” set using an 80% of the maximum PWM score threshold (except in cases where there were fewer than 1000 or greater than 20,000 binding sites genome wide, in which cases we, respectively, repeatedly lowered or raised the thresholds by 5% until between 1000 and 20,000 sites were found or 0% or 100% were reached). YeTFaSCo provides these results in an implementation of the GBrowse genome browser [463]. In addition to genome wide TF binding sites, YeTFaSCo provides tracks for in vivo, in vitro, and predicted nucleosome occupancy [242, 488, 294]. These tracks are provided as a reference because TFs are known to preferentially bind nucleosome free regions [317, 550]. I have also included a track representing the degree of conservation between closely related yeast species [143] because functional binding sites are more likely to be conserved [91]. These data could help users to identify binding sites which are used in vivo. The YeTFaSCo genome browser uses version 64 (2011-02-03) of the S. cerevisiae genome.
2.7.4 Identifying potential regulators of genes

This tool allows you to discover potential regulators of your favourite genes and was motivated by the fact that I had already collected much of the relevant data for evaluation of the motifs. This tool uses the data in the database to calculate how well TF-gene interactions correlate with query data. The data can take the form of either a gene list (e.g. genes in a pathway) or a dataset of gene-value pairs (e.g. expression data). The regulatory interaction data to which the query is compared can take three forms, and has one genome-wide dataset for each TF: gene expression data for TF mutants, ChIP-chip data of TFs in promoters, or predicted binding sites of TFs in promoters. In practise, this entails performing a rank sum test to see if the gene list provided has significantly different regulatory scores than the gene background, or, for quantitative data, comparing the correlation between the datasets. As an example, say we are interested in phosphate metabolism genes, and so we take all genes annotated with the appropriate GO terms, and use this as the query set. This tool can be used to identify the TFs that likely regulate the query genes by identifying the TFs that bind (ChIP) or have the potential to bind (motifs) their promoters, and the TFs which, when perturbed, result in changes to their expression.

The most direct way of inferring a regulatory interaction is using the expression data for TF mutants, since this represents transcriptional changes upon perturbation of a transcription factor. However, the identified interactions in this data may not represent direct targets, since secondary effects can be present. Expression data has the added advantage that you can infer whether the regulating TF is activating or repressing depending on the nature of the mutant and relationship identified to the query genes. ChIP-chip data provides a complementary approach because the interactions identified represent direct binding of the TF to promoters, although not all of these may represent functional binding events. However, the ChIP data is limited by the availability of data and the pertinence of the condition under which the ChIP experiment was performed.
By comparing your data to the occurrences of motifs in the promoters of genes, potential regulatory interactions can be identified that are less-specific, but are not restricted by experimental conditions.

The tool performs rank sum/ROC tests for gene lists and Pearson/Spearman correlation for continuous gene data. The rank sum test yields two values for each dataset: the P-value and the area under the ROC curve, which represents whether your gene list had on average higher (ROC>0.5) or lower (ROC<0.5) values in the data. Using Pearson or Spearman correlation yields a P-value and the correlation coefficient.

2.8 Future directions

In the process of constructing the YeTFaSCo database and manually curating the motif collection, we compiled a list of additional features and further analyses to incorporate into future versions of the database. One particularly important step will be to revisit the motif derivation steps. Browsing YeTFaSCo, it is clear that different motif-finding algorithms can yield dramatically different motifs from the same ChIP-chip data. The same may be true of motifs from PBMs and possibly also MITOMI: a recent analysis described a method for obtaining motifs that are demonstrably more accurate than those derived from previous approaches [564]. More sophisticated motif evaluation methods might also yield higher correspondence between data sets; for example, correspondence between TF motifs and ChIP-chip or expression data may be higher if nucleosome occupancy over the motif match is considered, as well as the presence of General Regulatory Factor (GRF) binding sites in proximity [242, 317]. It is known that open chromatin is a major determinant of TF binding in vivo [550], suggesting that most TFs rely on additional cues - some of which are known and can be incorporated into computational models.

There are still five putative TFs that have no motif. In addition, 92 proteins that were
previously annotated as known or putative TFs we labelled “dubious” and excluded them from the final manually-curated list, because there is, as yet, no formal demonstration that these proteins have intrinsic sequence-specific DNA-binding activity - although there is at least some suggestion that they may. Thus, the sequence specificities of yeast TFs will, I hope, remain an active area of research, and future iterations of YeTFaSCo will incorporate emerging data. Many of these “dubious” TFs with motifs assigned to them are known to be an upstream signalling component or downstream effector. For at least some such cases, the motif derived for these proteins corresponds to a known co-acting TF, suggesting that the signalling/effector protein is specific to this TF (e.g. all ChIP-chip-derived Fhl1 motifs are in fact binding sites for Rap1; see Table 2.2 and [164] for additional examples). It would likely be valuable for the mapping and mechanistic understanding of transcriptional networks to have, in addition to an index of TF sequence specificities, an index of which cofactors and chromatin factors are recruited by each of the individual TFs, or are involved in its recruitment.

2.9 Conclusions

As a unified and comprehensive resource of manually-curated TF motifs, YeTFaSCo addresses a fundamental need in the analysis of yeast transcriptional networks. I anticipate that this database will be an extremely useful resource for the yeast community and will facilitate a greater understanding of transcriptional regulation.
Chapter 3

Sequence properties of yeast transcripts

This chapter has not yet been submitted for publication. In its current form, I performed all the described work and wrote all the text (with the single exception of the paragraph on potential combinatorial binding of Hrpl, which Tim contributed to as it was included in the publication described in Chapter 4). However, I am also grateful to Kate Cook for helpful discussions.
3.1 Summary

Identifying over and underrepresented DNA motifs associated with genomic features has previously proven useful in identifying cis-regulatory elements. Here, I analyze sequence biases associated with yeast ORFs, introns, 5' UTRs, and 3' UTRs. I describe a modified approach to identify DNA words that are enriched or depleted relative to the words containing and contained within them. I further demonstrate how this approach can be modified to search for functional cis-elements in evolutionarily constrained sequences, such as ORFs. I apply this method to yeast introns, 5' UTRs, 3' UTRs, and coding sequences and confirm that my method is specific and sensitive by identifying the known splicing motifs, as well as motifs associated with Y' elements and ribosomal genes (which are overrepresented among spliced genes). Many enriched motifs are common to both 5' and 3' UTRs, and I show that one such motif is bound and recognized by the Nrd1 complex in both locations. In 3' UTRs, I find an enrichment of TA-repeats that corresponds to the known binding site of the termination factor Hrp1, but I also show that these elements can initiate antisense transcription by acting as a TATA box. I find that many of the most enriched motifs within ORFs correspond to antisense binding sites for the Nrd1 complex, suggesting a widespread role for Nrd1 in suppressing antisense transcription. Altogether, I confirm the role of many known functional cis-elements, provide hypotheses for the function of others, and confirm several of these hypotheses.

3.2 Motivation/introduction

Relating genome sequence to gene function is an integral part of modern molecular biology. The genome sequence of the yeast S. cerevisiae is extensively studied, and the functions of many of its genomic sequence elements are known.

In the promoter, many transcription factor (TF) binding sites are overrepresented [480, 123]. The TATA-binding protein (TBP) is one such factor, and it binds to the TATA
box that is present about 80 bp upstream of the TSS in approximately 20% of promoters [123, 32]. In addition, the promoter also contains an overrepresentation of poly-dA:dT tracts [468, 244], which exclude nucleosomes by virtue of their rigid structure [215]. In addition to stable mRNAs, many promoters produce cryptic unstable transcripts (CUTs), which are relatively short transcripts that are recognized by the Nrd1 complex shortly after initiation, leading to degradation of the RNA by the exosome [219]. The Nrd1 complex contains the RNA-binding proteins (RBPs) Nrd1 and Nab3, which recognize GUAR and UCUU sequences in the RNA, respectively [219].

The open reading frames (ORFs) contained within mRNAs are generally thought to harbour few functional sequence elements because they are highly constrained by the protein coding sequence. However, functional elements can exist within the confines of the genetic code. For instance, the coding sequence has been shown to include binding sites for RBPs, TFs, and microRNAs in higher eukaryotes [311, 414, 114]; however, few examples of these secondary elements are known in yeast (e.g. [163, 199]).

Several studies have analyzed the sequences associated with both constitutive [558, 509] and alternative [28, 545, 51, 509] splicing in higher eukaryotes. For the few genes whose pre-mRNAs are spliced in yeast, there are highly conserved sequences at the 5’ splice site (5’ SS), 3’ splice site (3’ SS), and branch site that are recognized by the splicing machinery [421]. In addition, poly-U tracts have been shown to help the efficiency of splicing [95, 384]. The 5’ SS and 3’ SS mark the boundaries of the intron, and the branch site and poly-U sequences are contained between these two.

The 3’ ends of genes often contain TA repeats as well as poly-A and poly-T sequences [244] that, when transcribed, are bound by the cleavage and polyadenylation machinery, which cleaves the nascent transcript and adds a poly-A tail [355]. These RNA sequences are recognized by various cleavage and termination factors: the RBP Hrp1, which recognizes UA-repeats [252]; Rna15, which recognizes A-rich sequences [172]; and Yth1 and/or Yhh1, which recognize U-rich sequences [106, 470, 26]. Together, these sequence elements
are thought to mark the cleavage site.

Numerous methods have been developed to computationally identify functional elements in the DNA sequence. Many of these aim to derive a position weight matrix (or set of matrices) representing the motif(s) present in the query sequences. Common approaches for this include expectation maximization (e.g. [21]), Gibbs sampling (e.g. [290, 420]), and biophysical modelling (e.g. [564]); (see also [490]). However, these methods generally cannot identify depleted motifs, are often non-deterministic, and cannot exhaustively test the motif space. Markov chain analysis [391] and related statistical methods (e.g. [500, 186, 183, 501]) aim to derive enrichment and/or depletion scores for all DNA words, and so exhaustively examine the possible motifs and provide a deterministic score for each word.

Because the sequence of the genome is contains many biases, it is often beneficial to compare the DNA of interest to a set of background sequences to eliminate the motifs that are enriched, but not functionally relevant to the question being asked (e.g. [530]). For instance, when analyzing protein coding sequences, it is often useful to compare them to a codon-shuffled background to account for the presence of the ORF (e.g. [214, 245]). In addition, many motifs show a non-random distribution around genomic landmarks (e.g. the TATA box occurs approximately 50-80 bp upstream of the transcription start site in yeast [123]) and this information can also be helpful in identifying functional elements, but few approaches currently take advantage of this information (e.g. [501, 510]).

Most previous analyses finding motifs by any method in yeast have concentrated on promoter regions, often aiming to identify process-specific regulators. These have found various motifs, many of which correspond to known TF binding sites, TATA box-like motifs, as well as A- and T-rich motifs (e.g. [500, 61, 461, 452, 183]). However most such studies do not distinguish between the promoter and the 5' UTR sequences. In the 3' UTR, studies have found an enrichment of the Hrp1 binding site and variants thereof, A- and T-rich sequences, as well as many other sequences [167, 501, 444]. I
am unaware of any studies that have computationally searched protein coding sequences or the sequence surrounding introns for enriched/depleted sequences in yeast. Further, in most of the studies that identified over and underrepresented sequence motifs, their functional roles remain untested.

Here, I use several methods to analyze the sequence properties of yeast gene components, including 5' and 3' UTRs, introns, and coding sequences. I first analyze fluctuations in strand specific base content relative to genomic features of interest to gain a general picture of positional sequence biases. I then identify over and underrepresented DNA words using a modified algorithm described previously [186]. Wherever possible, I make use of appropriate background sequences to control for motifs that are not specific to my feature of interest.

I recapitulate many known functional sequence elements, discover several elements with previously unknown function, and demonstrate the role of some of the trans-acting factors implicated in binding over/underrepresented cis-elements. In 5' and 3' UTRs, I find an overrepresentation of TA-repeats, which I show can act as binding sites for TBP, forming alternate promoters in 5' UTRs and initiating antisense transcription within 3' UTRs. I also find that binding sites for the Nrd1-complex member Nab3 are enriched on both strands within 5' and 3' UTRs, and appear to be functional. Within ORFs, I find that many of the most enriched motifs correspond to antisense Nrd1-complex binding sites.

3.3 Overall approach

In my analysis, I wanted to concentrate on four different parts of transcripts: 5' UTRs, the coding sequence, introns, and 3' UTRs (Figure 3.1A). I obtained the positions of ORFs and introns from SGD, and UTRs as described in Section 4.3.1. To analyze these features, I took two complementary approaches. First, I plotted the average base content
across features, aligned to the boundaries of the feature (Figure 3.1B-J). This facilitates the identification of general trends in the sequence and illustrates the sequence motifs that precisely align with feature boundaries. I then identify enriched DNA words (k-mers) in the different regions of interest. In order to gauge k-mer enrichment/depletion, I counted the number of observed instances of each k-mer and compared that to the number expected. Where appropriate, I also compared to an appropriate background when enriched k-mers were likely to be dominated by other effects (for instance, when the region analyzed is also a coding sequence).

### 3.3.1 Calculating Z-scores for enrichment/depletion

In order to calculate over/underrepresentation of words, I modified a measure used previously [186]. Briefly, for each word of length k (k-mer), I first ask if the k-mer is enriched/depleted given the two sub-words ((k-1)-mers) contained within it. For example, for the 4-mer TAGC, I compare the number of TAGCs in the sequence \(O_{TAGC}\) to the number expected given the number of TAG 3-mers observed and the frequency of C \(O_{TAG} \times F_C\), and to the number of AGC’s observed and the frequency of T \(O_{AGC} \times F_T\). Next, I ask how enriched/depleted each word is compared to all super-words ((k+1)-mers) containing it. For example, I compare the number of occurrences of TAGCN and NTAGC to the number expected given the observations of TAGC and the frequency of N \(O_{TAGC} \times F_N\), for all N=A,T,G,C. This way, I can identify only those enriched motifs for which adding an additional base of information does not add to the enrichment, but subtracting a base of information reduces the enrichment. This results in Z-scores for how enriched/depleted each word is relative its sub-words (Sub-Z) and how enriched/depleted its super-words are relative it (Super-Z). Intuitively, positive Z-scores indicate enrichment and negative Z-scores indicate depletion, while a high magnitude Sub-Z score in combination with a lower Super-Z score indicates that the DNA word is more likely to be functional than its sub-words or super-words.
Figure 3.1: Overall approach to identifying sequence features associated with transcripts. (A) The regions analyzed in this paper: 5’ UTRs, introns, 3’ UTRs and the coding sequence. (B-J) The average base content for each region. In each plot, the middle lines for each curve represents the mean estimate of the average base content for each nucleotide, while upper and lower bounds for each curve represent the 90th and 10th percentile mean estimates derived from sampling the data 100 times, illustrating the uncertainty in the base content. (B-D) The base content averaged across each region, scaled such that the starts and ends of features align and smoothed over 20 bp for each of (B) 5’ UTRs, (C) introns, and (D) 3’ UTRs. (E-J) Average base content at boundaries of features, un-scaled and un-smoothed, for each of (E) TSSs, (H) ORF starts, (F) 5’ SSs, (I) 3’ SSs, (G) ORF ends, and (J) cleavage sites.
Mathematically, the Z-score for enrichment/depletion of a word equals the number of times that word was observed minus the number expected, divided by the square root of the number expected \( ((O_w - E_w)/\sqrt{E_w}) \), where the number expected is defined by the abundance of each of the two possible (k-1)-mers in the sequence, as described previously [186]. However, instead of taking the expected number of k-mers as the average of the numbers expected given the two possible (k-1)-mers, as previously described, I took the maximum of the two when the number observed is greater than both (e.g. \( E_{ATGC} = \max(O_{ATG} \times F_C, O_{TGC} \times F_A) \) if \( O_{ATGC} > O_{ATG} \times F_C \) and \( O_{ATGC} > O_{TGC} \times F_A \)), and the minimum when the number observed is less than both (e.g. \( E_{ATGC} = \min(O_{ATG} \times F_C, O_{TGC} \times F_A) \) if \( O_{ATGC} < O_{ATG} \times F_C \) and \( O_{ATGC} < O_{TGC} \times F_A \)). If the number observed was between the two ( \( O_{TGC} \times F_A < O_{ATGC} < O_{ATG} \times F_C \) or \( O_{ATG} \times F_C < O_{ATGC} < O_{TGC} \times F_A \)), I set the number expected to the number observed, yielding a Z-score of 0. My reasoning behind this change was that if, as an example, the true motif is ATGC (and this motif is enriched), then the word ATGCA is not enriched relative to ATGC, but is enriched relative to TGCA because TGCA does not contain the whole true motif. Thus, taking the average of the expected numbers, given the two sub-motifs, would falsely indicate that the motif ATGCA is enriched. Taking the maximum (or minimum, for depletion) corrects for this. Meanwhile, there would still be more occurrences of ATGC than expected, given either ATG and TGC, and so the Sub-Z score would correctly indicate enrichment.

The Super-Z score provides a score indicating how much more or less enriched a word gets by adding an additional base of information and is calculated similarly to the Sub-Z score; the Super-Z score for an enriched motif is the maximum Z-score for all (k+1)-mers containing that word, while for depleted k-mers, it is the minimum k+1 Z-score. Here, the expected number of each (k+1)-mer is the k-mer count times the frequency of the additional base of information (e.g. \( E_{ATGCA} = O_{ATGC} \times F_A \)).

When identifying enriched/depleted words, it is sometimes desirable to compare the
motif occurrences to a set of background sequences. For instance, the exon component of intron-exon junctions is also generally coding sequence, and so many of the over/underrepresented words in this region may be due to the coding sequence, not the intron. By using the coding sequence as a background, this can be accounted for. When a background is used, enrichment/depletion of a query sequence is measured in relation to a background. In this case, the Z-score equals the difference between the number of times the word is observed and the number expected in the query sequence, minus the same value for the background, all divided by the square root of the number expected in the query sequence \( \frac{(O_q - E_q) - (O_{bg} - E_{bg})}{\sqrt{E_q}} \). In this way, only sequences that are enriched or depleted relative to the background yield a high Z-score.

I also make use of the K-S statistic \( K \) to describe the positional bias of motifs across each region of interest. This statistic describes the maximum distance between the observed and expected cumulative distribution functions (CDF) for motif instances across the region. The expected CDF is calculated directly from the average base content of the region, smoothed over 5 bases. For instance, the expected frequency of GATC at position one is the product of the observed frequencies of G, A, T, and C at positions one to four, respectively. The K-S statistic is the maximum difference between the observed and expected CDFs, multiplied by the square root of the number of words observed, so that the larger the statistic, the greater the positional bias of the motif.

### 3.4 Validation of the method: Yeast introns

In order to validate my approach, I analyzed the sequences in and around yeast introns. Looking at the overall base composition, the A content tends to be relatively high upstream of the 5’ SS, goes down within the intron body while T content increases, and is similar to the bias observed in coding sequence following the 3’ SS (Figure 3.1C, compare to Figure 3.5A). As expected, the splice donor (GTATGT [421]) and splice acceptor
(AG [421]) recognition motifs are universally present at the 5′ and 3′ ends of introns, respectively (Figure 3.1F and I).

I next calculated the Z-scores of enrichment/depletion and K-S statistics for the positional bias of k-mers (from k=2 to 8) across four 100 bp regions: upstream and downstream of the 5′ and 3′ SSs (Tables 3.1- 3.4). For regions encompassing coding sequence, I used the coding sequences of non-spliced genes as a background. I restricted my analysis to the top 50 most enriched/depleted sequences (by Sub-Z), first filtering for sequences with a higher Sub-Z than Super-Z (i.e. those for which the k-mer is more enriched/depleted than all (k+1)-mers containing it) for all k-mers up to k=7. I retained 8-mers for which Sub-Z<Super-Z because 9-mers were not included in the analysis, so these would otherwise not be represented.

I recovered all the known splicing sequence elements (i.e. the 5′ SS, 3′ SS, branch point, and poly-U tracts [421, 384]), which are all enriched and show a strong positional bias (Tables 3.2 and 3.3), although the 3′ SS is so short that it was recovered as the optimal 4-mer containing the consensus (ACAG; Table 3.3). The dinucleotide AG was not significantly enriched, but still showed a strong positional preference around the 3′ SS (Sub-Z=1.62, K-S=4.43). The majority of the remaining enriched k-mers correspond to repeats common to ribosomal genes or genes lying in the Y′ elements of subtelomeric regions. These Y′ elements are repetitive DNA sequences that come in long and short varieties [75], both containing related DNA sequences [321]. Of the 273 spliced genes, 13 lie in Y′ elements, 7/13 lie in long Y′ elements (the other 6 in short), and 11/13 are putative helicases. The homology between these regions likely accounts for the overrepresented k-mers. Of the remaining 260 spliced genes, 89 include ribosomal genes and the k-mers that are enriched in these genes may be due to the high sequence similarity between duplicated ribosomal genes. As expected, I did not find any strong candidates for novel splicing cis-elements.
Table 3.1: Top 50 motifs enriched and depleted in the region from 100 bp upstream of the intron to the 5' SS, compared to a coding sequence background, with K-S statistics calculated for the 200 bp around the 5' SS. RC(k-mer) represents the reverse complement of the k-mer.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>K-S</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched</td>
<td>AA</td>
<td>TT</td>
<td>7.08</td>
<td>6.39</td>
<td>0.49</td>
<td>2.01</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CCACGAAG</td>
<td>CTTTGTGG</td>
<td>5.42</td>
<td>1.93</td>
<td>3.49</td>
<td>2.32</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CTAACGCC</td>
<td>GGCCATAG</td>
<td>4.43</td>
<td>3.17</td>
<td>1.26</td>
<td>2.14</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>AATGG</td>
<td>CATT</td>
<td>4.42</td>
<td>2.85</td>
<td>1.58</td>
<td>2.77</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GACTAACG</td>
<td>GCTTAGTC</td>
<td>4.35</td>
<td>3.85</td>
<td>0.49</td>
<td>1.31</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TAA</td>
<td>TTA</td>
<td>4.29</td>
<td>0.93</td>
<td>3.35</td>
<td>1.36</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TTATACC</td>
<td>CGATTAGA</td>
<td>4.16</td>
<td>1.20</td>
<td>2.96</td>
<td>1.35</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GAAGCTTG</td>
<td>CACGCTTC</td>
<td>4.16</td>
<td>3.17</td>
<td>0.99</td>
<td>3.14</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>AGT</td>
<td>ACT</td>
<td>4.14</td>
<td>2.66</td>
<td>1.48</td>
<td>2.21</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>TTG</td>
<td>4.13</td>
<td>3.36</td>
<td>0.77</td>
<td>1.85</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TTTTTTT</td>
<td>AAAAAAA</td>
<td>4.13</td>
<td>3.95</td>
<td>0.18</td>
<td>3.07</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>CGACTAAG</td>
<td>CTTATCGC</td>
<td>4.05</td>
<td>3.85</td>
<td>0.20</td>
<td>1.76</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CTTTTCAC</td>
<td>GTGAAAAA</td>
<td>4.03</td>
<td>2.15</td>
<td>1.88</td>
<td>1.65</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CTTGGTTTC</td>
<td>GAAACAGG</td>
<td>3.97</td>
<td>2.84</td>
<td>1.14</td>
<td>1.67</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TTTTTTTT</td>
<td>AAAAAAA</td>
<td>3.94</td>
<td>3.33</td>
<td>0.62</td>
<td>2.83</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>CTATAAC</td>
<td>GGTITTAC</td>
<td>3.89</td>
<td>4.41</td>
<td>-0.52</td>
<td>1.99</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GCCGTGTC</td>
<td>GATCACGC</td>
<td>3.89</td>
<td>3.17</td>
<td>0.71</td>
<td>2.19</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>ATA</td>
<td>TAT</td>
<td>3.83</td>
<td>2.71</td>
<td>1.12</td>
<td>1.17</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CCGCTGAG</td>
<td>CTCAGCCC</td>
<td>3.74</td>
<td>3.94</td>
<td>-0.20</td>
<td>1.85</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GCCGTGAG</td>
<td>CTCAGCCC</td>
<td>3.74</td>
<td>3.94</td>
<td>-0.20</td>
<td>1.83</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTATTCCC</td>
<td>GAAGGTCAC</td>
<td>3.72</td>
<td>3.69</td>
<td>0.03</td>
<td>1.31</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTATCCCT</td>
<td>AGGTACC</td>
<td>3.70</td>
<td>2.29</td>
<td>1.41</td>
<td>1.49</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CGATTGC</td>
<td>CAACTCGG</td>
<td>3.64</td>
<td>3.59</td>
<td>0.05</td>
<td>1.50</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTACCAC</td>
<td>TGGGTAC</td>
<td>3.64</td>
<td>3.59</td>
<td>0.05</td>
<td>1.84</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GATCTATG</td>
<td>CATAGATC</td>
<td>3.63</td>
<td>4.41</td>
<td>-0.78</td>
<td>2.15</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CGATTTGG</td>
<td>CCCAATCG</td>
<td>3.62</td>
<td>4.41</td>
<td>-0.79</td>
<td>1.45</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTACCAAG</td>
<td>CGTGTCAC</td>
<td>3.62</td>
<td>3.17</td>
<td>0.45</td>
<td>1.76</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GCCCAAG</td>
<td>CGTGGGC</td>
<td>3.62</td>
<td>3.17</td>
<td>0.45</td>
<td>2.10</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>AAATGTG</td>
<td>CCATTTTT</td>
<td>3.61</td>
<td>1.66</td>
<td>1.95</td>
<td>2.19</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CGATTTGG</td>
<td>CCAATCGG</td>
<td>3.58</td>
<td>4.13</td>
<td>-0.55</td>
<td>1.39</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTGATC</td>
<td>GATCAC</td>
<td>3.56</td>
<td>1.50</td>
<td>2.06</td>
<td>2.45</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CTTTCTTC</td>
<td>GAAGAACG</td>
<td>3.56</td>
<td>1.89</td>
<td>1.66</td>
<td>1.06</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTGATAC</td>
<td>GATGAC</td>
<td>3.53</td>
<td>2.49</td>
<td>1.04</td>
<td>1.12</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GGACCG</td>
<td>CGGTCC</td>
<td>3.52</td>
<td>2.38</td>
<td>1.14</td>
<td>1.02</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CAAAAAGG</td>
<td>CTTTTTTT</td>
<td>3.50</td>
<td>1.52</td>
<td>1.97</td>
<td>1.09</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CACATTC</td>
<td>GAAGTGTC</td>
<td>3.44</td>
<td>2.04</td>
<td>1.41</td>
<td>0.89</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CGTGAAG</td>
<td>CTTACGCC</td>
<td>3.44</td>
<td>3.11</td>
<td>0.33</td>
<td>1.23</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TTTGATAC</td>
<td>GCACCAAA</td>
<td>3.42</td>
<td>3.85</td>
<td>-0.43</td>
<td>1.57</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CAGATCCC</td>
<td>GGATTCTG</td>
<td>3.38</td>
<td>2.31</td>
<td>1.07</td>
<td>1.85</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TAAACC</td>
<td>GGGTTTA</td>
<td>3.37</td>
<td>3.16</td>
<td>0.21</td>
<td>1.65</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CAAAACGC</td>
<td>CGGTCCTT</td>
<td>3.37</td>
<td>2.55</td>
<td>0.82</td>
<td>0.93</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CCAAGTAG</td>
<td>CTTACCTC</td>
<td>3.37</td>
<td>3.94</td>
<td>-0.58</td>
<td>1.72</td>
<td>Y' element</td>
</tr>
<tr>
<td>Depleted</td>
<td>GG</td>
<td>CC</td>
<td>-5.12</td>
<td>-5.09</td>
<td>0.03</td>
<td>2.01</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>GG</td>
<td>-4.81</td>
<td>-3.23</td>
<td>1.58</td>
<td>0.91</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TG</td>
<td>CA</td>
<td>-4.74</td>
<td>-3.65</td>
<td>1.00</td>
<td>2.02</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GATGAGTC</td>
<td>CATCATTG</td>
<td>-4.51</td>
<td>-0.61</td>
<td>3.91</td>
<td>0.48</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>AT</td>
<td>AT</td>
<td>-4.37</td>
<td>-3.31</td>
<td>1.07</td>
<td>1.76</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GGTG</td>
<td>CACC</td>
<td>-4.19</td>
<td>-1.54</td>
<td>2.64</td>
<td>1.21</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GACG</td>
<td>CGTC</td>
<td>-3.86</td>
<td>-1.05</td>
<td>2.81</td>
<td>1.57</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GATG</td>
<td>CATC</td>
<td>-3.47</td>
<td>-2.11</td>
<td>1.36</td>
<td>0.87</td>
<td>Y' element</td>
</tr>
</tbody>
</table>
Table 3.2: Top 50 motifs enriched and depleted in the region from the 5’ SS to 100 bp into the intron, with K-Ss calculated for the 200 bp around the 5’ SS.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>K-S</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched</td>
<td>GTATG</td>
<td>CATA</td>
<td>21.18</td>
<td>13.93</td>
<td>7.25</td>
<td>7.37</td>
<td>5’ SS</td>
</tr>
<tr>
<td></td>
<td>CTAAC</td>
<td>GTTAT</td>
<td>16.40</td>
<td>13.08</td>
<td>3.32</td>
<td>6.50</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>GTATGT</td>
<td>ACATAC</td>
<td>14.09</td>
<td>4.35</td>
<td>9.74</td>
<td>7.97</td>
<td>5’ SS</td>
</tr>
<tr>
<td></td>
<td>ACTAAC</td>
<td>GTTATG</td>
<td>13.22</td>
<td>8.71</td>
<td>4.51</td>
<td>7.23</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAA</td>
<td>11.52</td>
<td>9.28</td>
<td>2.23</td>
<td>4.35</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTTTT</td>
<td>AAAA</td>
<td>9.57</td>
<td>6.43</td>
<td>3.14</td>
<td>4.18</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TACTAAC</td>
<td>GTTATGTA</td>
<td>8.82</td>
<td>4.08</td>
<td>4.74</td>
<td>6.30</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>TTTTTT</td>
<td>AAAAAA</td>
<td>7.73</td>
<td>7.33</td>
<td>0.40</td>
<td>3.07</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTTTTTTT</td>
<td>AAAAAAAA</td>
<td>7.46</td>
<td>6.72</td>
<td>0.74</td>
<td>2.83</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>ATG</td>
<td>CAT</td>
<td>7.22</td>
<td>7.02</td>
<td>0.20</td>
<td>4.39</td>
<td>5’ SS</td>
</tr>
<tr>
<td></td>
<td>TTTTTTTTT</td>
<td>AAAAAAAAA</td>
<td>6.82</td>
<td>5.75</td>
<td>1.07</td>
<td>2.58</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>GAA</td>
<td>TTC</td>
<td>6.16</td>
<td>3.98</td>
<td>2.19</td>
<td>2.51</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>GCTT</td>
<td>5.56</td>
<td>4.12</td>
<td>1.55</td>
<td>2.42</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>CCCGCCG</td>
<td>GCCGCCG</td>
<td>5.36</td>
<td>2.99</td>
<td>2.57</td>
<td>1.82</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCTGCC</td>
<td>GGCAGC</td>
<td>5.03</td>
<td>4.60</td>
<td>0.43</td>
<td>1.25</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GTACGT</td>
<td>ACGTAC</td>
<td>4.76</td>
<td>3.21</td>
<td>1.55</td>
<td>2.63</td>
<td>5’ SS</td>
</tr>
<tr>
<td></td>
<td>CGCTTAC</td>
<td>GTTAAGCC</td>
<td>4.70</td>
<td>4.32</td>
<td>0.38</td>
<td>2.01</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGGCTCTC</td>
<td>GAGACGCC</td>
<td>4.70</td>
<td>4.32</td>
<td>0.38</td>
<td>1.30</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCTGCCC</td>
<td>GGCGAC</td>
<td>4.64</td>
<td>2.34</td>
<td>2.30</td>
<td>1.30</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GTATGTG</td>
<td>AACATAC</td>
<td>4.46</td>
<td>1.02</td>
<td>3.44</td>
<td>5.98</td>
<td>5’ SS</td>
</tr>
<tr>
<td></td>
<td>GACCAGG</td>
<td>CGCGGTC</td>
<td>4.36</td>
<td>4.67</td>
<td>-0.31</td>
<td>1.65</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCTTACCG</td>
<td>CGGTAAAGC</td>
<td>4.36</td>
<td>4.67</td>
<td>-0.31</td>
<td>2.00</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCTTACCG</td>
<td>CGCGTACG</td>
<td>4.36</td>
<td>2.70</td>
<td>1.65</td>
<td>1.82</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>GTT</td>
<td>4.28</td>
<td>3.30</td>
<td>0.98</td>
<td>2.60</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>GTGTTCGGT</td>
<td>CAGCACAC</td>
<td>4.26</td>
<td>5.11</td>
<td>-0.85</td>
<td>1.17</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGGACTAC</td>
<td>GTAGTCCG</td>
<td>4.21</td>
<td>2.42</td>
<td>1.79</td>
<td>1.14</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGGAAATC</td>
<td>GATTTTCG</td>
<td>4.21</td>
<td>4.18</td>
<td>0.03</td>
<td>1.36</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>TACTAACA</td>
<td>TGTATGTA</td>
<td>4.16</td>
<td>2.31</td>
<td>1.85</td>
<td>4.81</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>GAAA</td>
<td>TTC</td>
<td>4.15</td>
<td>1.92</td>
<td>2.23</td>
<td>1.99</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CGGTTATC</td>
<td>GATAACCG</td>
<td>4.13</td>
<td>4.67</td>
<td>-0.54</td>
<td>1.87</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGTTATCC</td>
<td>GGATACGC</td>
<td>4.13</td>
<td>4.67</td>
<td>-0.54</td>
<td>2.08</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGGGCCGG</td>
<td>CGCGCTTG</td>
<td>3.90</td>
<td>3.87</td>
<td>0.03</td>
<td>1.36</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GACAGTCG</td>
<td>CGACTGTC</td>
<td>3.90</td>
<td>4.18</td>
<td>-0.28</td>
<td>1.86</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCAGGCGGC</td>
<td>CGGCCTGC</td>
<td>3.90</td>
<td>3.87</td>
<td>0.03</td>
<td>1.33</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GTACCGAC</td>
<td>GTCGGTAC</td>
<td>3.90</td>
<td>4.18</td>
<td>-0.28</td>
<td>1.12</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CAGGGCC</td>
<td>CGGCGC</td>
<td>3.90</td>
<td>3.87</td>
<td>0.03</td>
<td>1.34</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GGCCTTC</td>
<td>GAGGACC</td>
<td>3.90</td>
<td>1.51</td>
<td>2.39</td>
<td>1.09</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GAGCTTAC</td>
<td>GTAGTGC</td>
<td>3.81</td>
<td>4.10</td>
<td>-0.29</td>
<td>1.61</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGGGCTC</td>
<td>GAGCGCGG</td>
<td>3.67</td>
<td>4.67</td>
<td>-1.00</td>
<td>1.58</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCATATGCG</td>
<td>GCATATGG</td>
<td>3.64</td>
<td>2.18</td>
<td>1.46</td>
<td>1.11</td>
<td>Y’ element</td>
</tr>
<tr>
<td>Depleted</td>
<td>TA</td>
<td>TA</td>
<td>-6.77</td>
<td>-6.44</td>
<td>0.33</td>
<td>2.31</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>TAG</td>
<td>CTA</td>
<td>-6.25</td>
<td>-1.68</td>
<td>4.57</td>
<td>1.43</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GGT</td>
<td>ACC</td>
<td>-5.17</td>
<td>-2.28</td>
<td>2.88</td>
<td>4.84</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CTGG</td>
<td>CAAG</td>
<td>-4.99</td>
<td>-1.66</td>
<td>3.34</td>
<td>1.90</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GTC</td>
<td>GAC</td>
<td>-4.55</td>
<td>-2.60</td>
<td>1.94</td>
<td>1.92</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>TAAG</td>
<td>CTTA</td>
<td>-4.53</td>
<td>-1.45</td>
<td>3.08</td>
<td>2.66</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>AGT</td>
<td>ACT</td>
<td>-4.41</td>
<td>-3.37</td>
<td>1.04</td>
<td>2.21</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CG</td>
<td>CG</td>
<td>-4.25</td>
<td>-3.14</td>
<td>1.11</td>
<td>2.00</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>TGTA</td>
<td>TACA</td>
<td>-3.87</td>
<td>-2.45</td>
<td>1.42</td>
<td>3.13</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CATAT</td>
<td>ATAG</td>
<td>-3.73</td>
<td>-2.13</td>
<td>1.60</td>
<td>1.29</td>
<td>Y’ element</td>
</tr>
</tbody>
</table>
### Table 3.3: Top 50 motifs enriched and depleted in the region from 100 bp upstream of the 3' SS to the 3' SS, with K-Ss calculated for the 200 bp around the 3' SS.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>K-S</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched</td>
<td>CTAAC</td>
<td>GTTAG</td>
<td>25.49</td>
<td>18.98</td>
<td>6.51</td>
<td>8.19</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>ACTAAC</td>
<td>GTTAGT</td>
<td>19.17</td>
<td>13.23</td>
<td>5.94</td>
<td>8.71</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>TACAA</td>
<td>GTTAGTA</td>
<td>13.40</td>
<td>8.88</td>
<td>4.52</td>
<td>7.59</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>TTT</td>
<td>AAAA</td>
<td>13.38</td>
<td>10.77</td>
<td>2.62</td>
<td>1.66</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAAA</td>
<td>11.10</td>
<td>7.55</td>
<td>3.55</td>
<td>3.00</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>GTATG</td>
<td>CATAC</td>
<td>9.04</td>
<td>7.51</td>
<td>1.33</td>
<td>5.25</td>
<td>5' SS</td>
</tr>
<tr>
<td></td>
<td>TACTAAC</td>
<td>GTTAGTA</td>
<td>9.02</td>
<td>5.04</td>
<td>3.98</td>
<td>6.18</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAAA</td>
<td>7.79</td>
<td>7.03</td>
<td>0.76</td>
<td>2.42</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>GTATG</td>
<td>ACATA</td>
<td>7.60</td>
<td>3.01</td>
<td>4.59</td>
<td>4.99</td>
<td>5' SS</td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>GTT</td>
<td>7.37</td>
<td>7.11</td>
<td>0.26</td>
<td>4.19</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>AACA</td>
<td>TGTT</td>
<td>7.30</td>
<td>7.18</td>
<td>0.12</td>
<td>3.79</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAAAAA</td>
<td>7.21</td>
<td>5.40</td>
<td>1.82</td>
<td>1.85</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTACTAAC</td>
<td>GTTAGTAA</td>
<td>6.31</td>
<td>7.04</td>
<td>-0.74</td>
<td>5.48</td>
<td>branch site</td>
</tr>
<tr>
<td></td>
<td>GCTTACCGG</td>
<td>CGGTAAGC</td>
<td>6.20</td>
<td>3.68</td>
<td>2.52</td>
<td>2.24</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GACGCTTAC</td>
<td>GTAGCTGC</td>
<td>5.90</td>
<td>6.16</td>
<td>-0.26</td>
<td>2.21</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GAGCGACGG</td>
<td>GTGCTGTC</td>
<td>5.90</td>
<td>2.80</td>
<td>3.09</td>
<td>2.62</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GACCGGCGG</td>
<td>CCGCCGTC</td>
<td>5.88</td>
<td>5.25</td>
<td>0.63</td>
<td>2.37</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAAAAA</td>
<td>5.54</td>
<td>5.07</td>
<td>0.47</td>
<td>1.27</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>CGCGGCTC</td>
<td>GAGCCGGC</td>
<td>5.29</td>
<td>5.25</td>
<td>0.04</td>
<td>2.25</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CCGCGGC</td>
<td>CCGCGGC</td>
<td>5.29</td>
<td>3.25</td>
<td>2.04</td>
<td>2.33</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CCGCGG</td>
<td>CCGCGG</td>
<td>5.29</td>
<td>5.25</td>
<td>0.04</td>
<td>2.30</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>TTTTT</td>
<td>AAAA</td>
<td>5.18</td>
<td>4.65</td>
<td>0.53</td>
<td>1.42</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>ACAG</td>
<td>CTTG</td>
<td>4.91</td>
<td>4.07</td>
<td>0.84</td>
<td>4.03</td>
<td>preferred 3' SS</td>
</tr>
<tr>
<td></td>
<td>CGCGTATC</td>
<td>GATAACGG</td>
<td>4.80</td>
<td>5.25</td>
<td>-0.45</td>
<td>1.30</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CGTTATCC</td>
<td>GGATAACG</td>
<td>4.80</td>
<td>5.25</td>
<td>-0.45</td>
<td>1.34</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GCTTAC</td>
<td>GATAGC</td>
<td>4.53</td>
<td>3.89</td>
<td>0.64</td>
<td>1.72</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GTGCTGCC</td>
<td>GCCAGACAC</td>
<td>4.47</td>
<td>4.43</td>
<td>0.03</td>
<td>1.64</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CGGTCGCC</td>
<td>GCCACGC</td>
<td>4.47</td>
<td>1.95</td>
<td>2.52</td>
<td>1.60</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GACAGTCGG</td>
<td>CGACTGTCT</td>
<td>4.45</td>
<td>3.97</td>
<td>0.48</td>
<td>1.69</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GGGGACAG</td>
<td>CTGTCGCC</td>
<td>4.45</td>
<td>2.46</td>
<td>1.99</td>
<td>1.81</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GATACGC</td>
<td>CGCTATC</td>
<td>4.06</td>
<td>2.58</td>
<td>1.48</td>
<td>1.76</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GAGATA</td>
<td>GTATCTC</td>
<td>4.03</td>
<td>3.25</td>
<td>0.78</td>
<td>2.32</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>CGCTGTGG</td>
<td>CAAACACG</td>
<td>4.00</td>
<td>3.97</td>
<td>0.03</td>
<td>1.09</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GAGACATG</td>
<td>GATGTCTC</td>
<td>4.00</td>
<td>3.97</td>
<td>0.03</td>
<td>1.77</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GCGTCTGC</td>
<td>GCCAGACG</td>
<td>4.00</td>
<td>2.10</td>
<td>1.89</td>
<td>1.46</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GGACAGAC</td>
<td>GTCTGTCC</td>
<td>4.00</td>
<td>4.41</td>
<td>-0.42</td>
<td>1.76</td>
<td>Y' element</td>
</tr>
<tr>
<td></td>
<td>GGTGCCAC</td>
<td>GTGGACACC</td>
<td>4.00</td>
<td>3.97</td>
<td>0.03</td>
<td>1.62</td>
<td>Y' element</td>
</tr>
<tr>
<td>Depleted</td>
<td>AT</td>
<td>AT</td>
<td>-8.94</td>
<td>-4.17</td>
<td>4.77</td>
<td>1.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TA</td>
<td>TA</td>
<td>-7.12</td>
<td>-5.18</td>
<td>2.04</td>
<td>3.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAA</td>
<td>TAAA</td>
<td>-5.42</td>
<td>-1.69</td>
<td>3.73</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>GA</td>
<td>-5.10</td>
<td>-4.71</td>
<td>0.39</td>
<td>4.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAGA</td>
<td>TCTA</td>
<td>-4.97</td>
<td>-1.62</td>
<td>3.34</td>
<td>2.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GT</td>
<td>AC</td>
<td>-4.87</td>
<td>-4.49</td>
<td>0.38</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAGT</td>
<td>ACTA</td>
<td>-4.74</td>
<td>-2.29</td>
<td>2.45</td>
<td>1.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAT</td>
<td>ATTA</td>
<td>-4.50</td>
<td>-1.40</td>
<td>3.10</td>
<td>1.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TCTA</td>
<td>TAGA</td>
<td>-4.45</td>
<td>-2.17</td>
<td>2.28</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTC</td>
<td>GAC</td>
<td>-4.36</td>
<td>-1.30</td>
<td>3.06</td>
<td>2.71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CG</td>
<td>CG</td>
<td>-4.13</td>
<td>-2.66</td>
<td>1.47</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAA</td>
<td>TTTA</td>
<td>-4.10</td>
<td>-2.73</td>
<td>1.37</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAG</td>
<td>CTTA</td>
<td>-3.98</td>
<td>-1.10</td>
<td>2.88</td>
<td>1.73</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.4: Top 50 motifs enriched and depleted in the region from the 3’ SS to 100 bp downstream of the intron, compared to a coding sequence background, with K-Ss calculated for the 200 bp around the 3’ SS.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>K-S</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched</td>
<td>AAG</td>
<td>CTT</td>
<td>7.73</td>
<td>6.71</td>
<td>1.02</td>
<td>4.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGT</td>
<td>ACC</td>
<td>6.75</td>
<td>4.36</td>
<td>2.38</td>
<td>3.33</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>GGTAAG</td>
<td>CTTACC</td>
<td>5.59</td>
<td>2.69</td>
<td>2.90</td>
<td>1.36</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>GCT</td>
<td>AGC</td>
<td>5.11</td>
<td>3.41</td>
<td>1.70</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAGACC</td>
<td>GGTCTT</td>
<td>5.07</td>
<td>1.95</td>
<td>3.12</td>
<td>2.26</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CAAG</td>
<td>CTTG</td>
<td>5.01</td>
<td>4.48</td>
<td>0.53</td>
<td>3.07</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CGTATAC</td>
<td>CGTATAC</td>
<td>4.99</td>
<td>4.23</td>
<td>0.76</td>
<td>1.38</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>AGCG</td>
<td>GCGCT</td>
<td>4.96</td>
<td>2.57</td>
<td>2.40</td>
<td>1.66</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGCAAAG</td>
<td>CGTTGCGG</td>
<td>4.87</td>
<td>4.68</td>
<td>0.18</td>
<td>1.10</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCACTG</td>
<td>CAGTTGG</td>
<td>4.86</td>
<td>1.87</td>
<td>2.98</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TATACGT</td>
<td>AGCTATA</td>
<td>4.74</td>
<td>4.63</td>
<td>0.11</td>
<td>1.31</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCCGATCG</td>
<td>CGTTGCGC</td>
<td>4.72</td>
<td>4.23</td>
<td>0.49</td>
<td>1.21</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCCGAT</td>
<td>ACCGC</td>
<td>4.70</td>
<td>1.56</td>
<td>3.14</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGAAAG</td>
<td>CTTTCT</td>
<td>4.70</td>
<td>2.31</td>
<td>2.39</td>
<td>1.32</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>GTATACGT</td>
<td>AGCTATAC</td>
<td>4.51</td>
<td>4.94</td>
<td>-0.43</td>
<td>1.51</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGAGT</td>
<td>GCTCAC</td>
<td>4.42</td>
<td>2.93</td>
<td>1.49</td>
<td>1.64</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCTCGA</td>
<td>GTCCAGG</td>
<td>4.41</td>
<td>3.27</td>
<td>1.14</td>
<td>1.03</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CAACGT</td>
<td>AGCTTG</td>
<td>4.39</td>
<td>2.12</td>
<td>2.27</td>
<td>1.72</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GATAACAC</td>
<td>GTGGTATC</td>
<td>4.37</td>
<td>3.65</td>
<td>0.72</td>
<td>1.48</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GGTTACCT</td>
<td>AGTACC</td>
<td>4.36</td>
<td>1.95</td>
<td>2.41</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TATACGT</td>
<td>GACGTATA</td>
<td>4.27</td>
<td>5.09</td>
<td>-0.82</td>
<td>1.61</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>TTCTAGCG</td>
<td>CCGATGAA</td>
<td>4.27</td>
<td>3.65</td>
<td>0.62</td>
<td>1.75</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CAACGTAT</td>
<td>ATACCGTG</td>
<td>4.22</td>
<td>5.09</td>
<td>-0.86</td>
<td>1.54</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>AGGGT</td>
<td>ACCCTT</td>
<td>4.16</td>
<td>1.39</td>
<td>2.77</td>
<td>0.75</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>AGAACA</td>
<td>TGGTTCT</td>
<td>4.16</td>
<td>2.65</td>
<td>1.51</td>
<td>1.63</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CCAAGAAG</td>
<td>CTTTCTGG</td>
<td>4.15</td>
<td>1.88</td>
<td>2.27</td>
<td>1.11</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CATCGGAG</td>
<td>CTTCGATG</td>
<td>4.07</td>
<td>4.73</td>
<td>-0.66</td>
<td>1.54</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCAAG</td>
<td>CTTTGG</td>
<td>4.04</td>
<td>1.55</td>
<td>2.49</td>
<td>1.27</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>GTCGTCG</td>
<td>CGACGAC</td>
<td>4.02</td>
<td>2.21</td>
<td>1.81</td>
<td>0.78</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>GGTTCACT</td>
<td>GATGAAAC</td>
<td>3.96</td>
<td>3.88</td>
<td>0.08</td>
<td>1.77</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCAAGG</td>
<td>CTTGG</td>
<td>3.94</td>
<td>2.84</td>
<td>1.10</td>
<td>2.13</td>
<td>ribosomal</td>
</tr>
<tr>
<td></td>
<td>CTGGATTC</td>
<td>GAATCGAG</td>
<td>3.91</td>
<td>3.27</td>
<td>0.64</td>
<td>2.17</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGCAGA</td>
<td>ATCTGGCGG</td>
<td>3.88</td>
<td>5.09</td>
<td>-1.20</td>
<td>1.02</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCTACGG</td>
<td>CGCTAGGG</td>
<td>3.82</td>
<td>3.99</td>
<td>-0.17</td>
<td>0.84</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CTACAGC</td>
<td>CGCTGAGG</td>
<td>3.82</td>
<td>3.99</td>
<td>-0.17</td>
<td>1.52</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCTAGATG</td>
<td>CATCTAAG</td>
<td>3.82</td>
<td>1.92</td>
<td>1.90</td>
<td>1.75</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GACGGCTT</td>
<td>CAAGCGTC</td>
<td>3.82</td>
<td>2.72</td>
<td>1.10</td>
<td>1.62</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GAGCCTCG</td>
<td>CGAGGCTC</td>
<td>3.82</td>
<td>3.79</td>
<td>0.03</td>
<td>1.34</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GACGATTC</td>
<td>GAATCTGC</td>
<td>3.82</td>
<td>3.99</td>
<td>-0.17</td>
<td>1.11</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCTACAGC</td>
<td>GCTGAGG</td>
<td>3.82</td>
<td>3.79</td>
<td>0.03</td>
<td>1.54</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CAACCTTG</td>
<td>CCAAGGGG</td>
<td>3.82</td>
<td>1.00</td>
<td>2.83</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGAGCCTC</td>
<td>GAGGGTCC</td>
<td>3.82</td>
<td>3.99</td>
<td>-0.17</td>
<td>1.35</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CCAAGTT</td>
<td>AACCTTGG</td>
<td>3.72</td>
<td>1.22</td>
<td>2.50</td>
<td>1.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GACTAACAG</td>
<td>CTTCTAGC</td>
<td>3.71</td>
<td>2.68</td>
<td>1.03</td>
<td>1.41</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>CGTCAGA</td>
<td>CGTCCAGC</td>
<td>3.70</td>
<td>2.25</td>
<td>1.45</td>
<td>1.14</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>ATACGT</td>
<td>CGACGTAT</td>
<td>3.70</td>
<td>4.23</td>
<td>-0.53</td>
<td>1.43</td>
<td>Y’ element</td>
</tr>
<tr>
<td></td>
<td>GCAA</td>
<td>TACCTGGG</td>
<td>3.69</td>
<td>5.36</td>
<td>-1.67</td>
<td>1.14</td>
<td>Y’ element</td>
</tr>
<tr>
<td>Depleted</td>
<td>AGT</td>
<td>ACT</td>
<td>-5.16</td>
<td>-4.55</td>
<td>0.61</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATA</td>
<td>TAT</td>
<td>-4.83</td>
<td>-2.66</td>
<td>2.18</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAT</td>
<td>ATTA</td>
<td>-4.12</td>
<td>-2.36</td>
<td>1.76</td>
<td>1.36</td>
<td></td>
</tr>
</tbody>
</table>
3.5 5’ UTRs

From Figure 3.1B, it is clear that the base content surrounding the TSS appears highly non-random. As expected, the initiator motif (YR [560]) is clearly visible at the 5’ UTR, and a sequence similar to the Kozak sequence (accATGg [274]) is visible at the translational start (Figures 3.1E and H). Around the TSS, there is an overall increase in the number of As and Gs and a decrease in the number of Ts, and overall there are many more As than Ts within the UTR (Figure 3.1B). In the promoter region, there is a preference for As around -150 relative to the TSS and a preference of Ts at around -60 relative to the TSS, as noted previously [118].

3.5.1 Over/underrepresented k-mers in 5’ UTRs

I next wanted to ask what motifs were enriched or depleted in 5’ UTRs, and so applied my method on all 5’ UTR sequences, while calculating K-S statistics for the positional bias of motifs surrounding the TSS. As expected, the sequence of the start codon (ATG) is significantly depleted in 5’ UTRs. The majority of k-mers that are both enriched relative to their submotifs (high Sub-Z score) and well-positioned (high K-S statistic) include poly-Ts, poly-As, TA-repeats, and TTCTT and its reverse complement (and variants thereof) (Table 3.5). These enriched motifs are similar to those found previously, analyzing the region upstream of ORF ATGs [183], indicating that some may be specific to 5’ UTRs.

Although the TATA box frequently contains TA-repeats (which may account for the high K-S statistic for this motif), I was unaware of a role for these sequences within UTRs. I hypothesized that some of these may represent TATA boxes for alternative promoters. To test this, I looked at TFIIB ChIP-exo data [412], as well as RNA-Seq reads specific to TSSs [386]. I isolated these data surrounding TATATA motifs within 5’ UTRs and compared this to the same data for control sequences. The control sequences were randomly
Table 3.5: Top 50 motifs enriched and depleted in 5' UTRs, with K-Ss calculated for the region from -500 to +100 relative the TSSs.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>K-S</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched</td>
<td>TT</td>
<td>AA</td>
<td>50.49</td>
<td>41.12</td>
<td>9.37</td>
<td>10.29</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTT</td>
<td>AAAA</td>
<td>42.60</td>
<td>32.42</td>
<td>10.18</td>
<td>14.53</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAAA</td>
<td>33.39</td>
<td>24.82</td>
<td>8.57</td>
<td>15.83</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>CT</td>
<td>28.60</td>
<td>17.02</td>
<td>11.58</td>
<td>11.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTTT</td>
<td>AAAA</td>
<td>25.48</td>
<td>21.42</td>
<td>4.06</td>
<td>14.92</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTTTTT</td>
<td>AAAAAAAA</td>
<td>21.89</td>
<td>18.62</td>
<td>3.27</td>
<td>13.22</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TTTTTTTT</td>
<td>AAAAAAAAA</td>
<td>18.97</td>
<td>14.29</td>
<td>4.68</td>
<td>11.72</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>GC</td>
<td>GC</td>
<td>15.87</td>
<td>11.74</td>
<td>4.13</td>
<td>7.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG</td>
<td>CTT</td>
<td>18.24</td>
<td>8.17</td>
<td>10.08</td>
<td>4.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>TG</td>
<td>17.02</td>
<td>15.88</td>
<td>1.14</td>
<td>10.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGAA</td>
<td>TTCTT</td>
<td>16.77</td>
<td>13.74</td>
<td>3.03</td>
<td>3.89</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>AAAAAAAA</td>
<td>TTAAAAAAA</td>
<td>15.94</td>
<td>13.83</td>
<td>2.12</td>
<td>22.24</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>AAAA</td>
<td>TTTT</td>
<td>15.34</td>
<td>12.84</td>
<td>2.49</td>
<td>21.51</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>TTCTT</td>
<td>AAGA</td>
<td>14.74</td>
<td>12.36</td>
<td>2.38</td>
<td>2.57</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>TTTTTTTTT</td>
<td>AAAAAAAAA</td>
<td>14.55</td>
<td>13.43</td>
<td>1.12</td>
<td>10.02</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>ATA</td>
<td>TAT</td>
<td>14.29</td>
<td>3.44</td>
<td>10.85</td>
<td>6.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAGAA</td>
<td>TTCTT</td>
<td>14.27</td>
<td>8.41</td>
<td>5.86</td>
<td>5.26</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>AAAAAAAA</td>
<td>TTAAAAAAA</td>
<td>14.13</td>
<td>12.77</td>
<td>1.36</td>
<td>21.40</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>CATT</td>
<td>AATG</td>
<td>12.70</td>
<td>7.05</td>
<td>5.65</td>
<td>1.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTCTTT</td>
<td>AAGA</td>
<td>12.65</td>
<td>9.49</td>
<td>3.16</td>
<td>3.52</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td>TG</td>
<td>12.61</td>
<td>10.75</td>
<td>1.86</td>
<td>4.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTGTG</td>
<td>ACAA</td>
<td>12.45</td>
<td>8.07</td>
<td>4.38</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATATA</td>
<td>TATAT</td>
<td>12.18</td>
<td>8.63</td>
<td>3.55</td>
<td>6.19</td>
<td>poly-TA</td>
</tr>
<tr>
<td></td>
<td>ATATATAT</td>
<td>ATATATAT</td>
<td>12.16</td>
<td>12.66</td>
<td>-0.50</td>
<td>3.62</td>
<td>poly-TA</td>
</tr>
<tr>
<td></td>
<td>TTTCTT</td>
<td>AAGA</td>
<td>11.45</td>
<td>9.74</td>
<td>1.71</td>
<td>3.96</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>AGCA</td>
<td>TGCT</td>
<td>11.32</td>
<td>4.17</td>
<td>7.15</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TATATATA</td>
<td>TATATATA</td>
<td>10.79</td>
<td>13.57</td>
<td>-2.78</td>
<td>3.71</td>
<td>poly-TA</td>
</tr>
<tr>
<td></td>
<td>GTG</td>
<td>CAC</td>
<td>9.82</td>
<td>3.21</td>
<td>6.62</td>
<td>2.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTCTTTT</td>
<td>AAGA</td>
<td>9.69</td>
<td>7.17</td>
<td>2.52</td>
<td>3.86</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>CTGC</td>
<td>GCAG</td>
<td>9.54</td>
<td>5.40</td>
<td>4.14</td>
<td>2.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGAG</td>
<td>CTCT</td>
<td>9.47</td>
<td>6.10</td>
<td>3.36</td>
<td>3.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TATTT</td>
<td>AAATA</td>
<td>9.21</td>
<td>7.72</td>
<td>1.49</td>
<td>1.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGGAA</td>
<td>TTCT</td>
<td>9.08</td>
<td>6.79</td>
<td>2.29</td>
<td>1.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTCTTT</td>
<td>AAGA</td>
<td>8.95</td>
<td>6.66</td>
<td>2.29</td>
<td>2.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAAG</td>
<td>CTTT</td>
<td>8.95</td>
<td>6.12</td>
<td>2.83</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AACAA</td>
<td>TTGT</td>
<td>8.91</td>
<td>4.95</td>
<td>3.97</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAGAAA</td>
<td>TTCTTT</td>
<td>8.78</td>
<td>6.89</td>
<td>1.89</td>
<td>5.23</td>
<td>TTCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>TTTATT</td>
<td>AATA</td>
<td>8.39</td>
<td>6.97</td>
<td>1.42</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTGC</td>
<td>GCAA</td>
<td>8.38</td>
<td>6.37</td>
<td>2.01</td>
<td>2.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTGATT</td>
<td>AATGAA</td>
<td>8.34</td>
<td>4.47</td>
<td>3.87</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTGATT</td>
<td>AACAAA</td>
<td>8.31</td>
<td>5.81</td>
<td>2.50</td>
<td>2.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTCTTT</td>
<td>AAGA</td>
<td>8.22</td>
<td>8.19</td>
<td>0.04</td>
<td>4.16</td>
<td>TTCTT/AAGAA</td>
</tr>
</tbody>
</table>

| Depleted  | ATG            | CAT       | -56.56| -5.28   | 51.28        | 6.92 | start codon |
|           | TA             | TA        | -31.46| -13.52  | 17.94        | 7.31 |             |
|           | CG             | CG        | -21.79| -7.99   | 13.80        | 6.84 |             |
|           | TAA            | TTA       | -12.32| -10.95  | 1.37         | 2.43 |             |
|           | GAC            | GTC       | -12.00| -8.26   | 3.74         | 2.26 |             |
|           | GTC            | GAC       | -11.19| -2.61   | 8.58         | 4.55 |             |
|           | ATTTA          | TAAAT     | -9.35 | -5.01   | 4.35         | 2.39 |             |
|           | CTA            | TAG       | -8.58 | -7.06   | 1.52         | 5.38 |             |
selected from 5’ UTRs, with the constraint that the distribution of distances from the TSS were preserved. I found that the area downstream of the TATATA motifs generally has more downstream TFIIB reads and fewer reads overlapping the motif, compared with the controls, consistent with these motifs being occupied by TBP and recruiting TFIIB (Figure 3.2A and B). In addition, overall, there are more TSS-specific RNA-Seq reads in the region from +40 to +90, relative the control (Figure 3.2C), consistent with the normal distance from TATA boxes to the TSS [123]. This is unlikely to be a result of misannotated TSSs because the majority of the TFIIB-specific reads lie upstream of the TATATA motifs and the control sequences do not display the same downstream TSS bias (Figure 3.2A and C).

I noted that TTCTT/AAGAA and poly-A/poly-T motifs were also enriched in the 3’ UTR (see below). The fact that these two motifs are very similar motivated me to look at their distribution around TSSs and CPA sites. I found that the occurrence of AAGAA and TTCTT closely mirror the occurrence of AAAAA and TTTTT, respectively, at both
TSSs and CPA sites (Figure 3.3), implying that they might share a common function (Pearson R = 0.443 and 0.0839 for AAAAA and AAGAA at TSSs and CPAs, respectively and R=0.754 and 0.562 for TTTTT and TTCTT across TSSs and CPAs, respectively).

Figure 3.3: The distribution of AAAAA, TTTTT, AAGAA, and TTCTT motifs across TSSs and CPA sites. Frequencies represent the probability of a motif starting at a particular base. (A) Motifs across TSSs. (B) Motifs across CPA sites. (C) and (D) Zoomed in versions of A and B, respectively.

Because of the role poly-A sequences play in nucleosome occupancy [215], I next asked if TTCTT/AAGAA motifs played a similar role. To address this, I identified instances of AANAA across the genome and compared the nucleosome occupancy [294] surrounding
these motifs for N=A, T, G, and C (Figure 3.4). I found that, although AAAAA and AATAA were depleted of nucleosomes in vivo, AAGAA (and AACAA) were not, implying that the prevalence of AAGAA/TTCTT motifs in 5’ and 3’ UTRs is not due to an effect on nucleosome occupancy.

Figure 3.4: The effect of AANAA motifs on nucleosome occupancy in vivo [294]. I identified all instances of AANAA for each nucleotide N across the yeast genome and then averaged the nucleosome occupancy [294] across these motifs and their reverse complements. Upper and lower bounds for each curve represent the 90th and 10th percentile estimates of the mean derived from sampling the data 100 times.

I next considered whether these AAGAA/TTCTT motifs could represent binding sites for Nab3, whose binding site is TCTT [219]. I first compared the binding of Nab3
measured by PAR-CLIP [221] for 5' UTRs with differing numbers of Nab3 binding sites, and these were found to have a significant correlation on both the sense (Spearman R=0.263, P<10^{-76}) and antisense (R=0.108, P<10^{-13}) strands. I reasoned that if these motifs represent functional Nab3 targets then their expression should increase in exosome mutants, where the CUTs recognized (in part) by Nab3 are no longer degraded. I found that expression changes in an exosome mutant [433] correlated significantly with the occurrence of the TCTT motif within 5' UTRs on both strands (Spearman R=0.158, P<10^{-27}, and R=0.0615, P<2×10^{-5} for sense and antisense, respectively). This indicates that the occurrence of these motifs within the 5' UTR is at least partly explained by their action as Nab3 binding sites.

3.6 The coding sequence

I next wanted to ask if there were any interesting sequence properties associated with yeast ORFs. However, the coding sequence is the most highly constrained part of genes because of the proteins these sequences encode. I therefore wanted to account for the coding constraints in my analysis, allowing me to ask what sequence signals are associated with the coding sequence, given the restrictions of the ORF and codon bias.

3.6.1 Base bias of coding sequence

I first wanted to analyze the base bias across coding sequence as a whole. In general, the coding sequence contains a strong preference of A over T and a slight preference of G over C (Figure 3.5A). There is also a strong bias in the base content at the start and end of the ORF. However, this does not account for the constraint of the coding sequences; these are especially biased at the starts and ends because all ORFs start with ATG and end with a stop codon (TAA, TGA, TAG). Further, the yeast genome as a whole has a preference of As and Ts over Gs and Cs. This motivated me to ask how much of this
bias could be accounted for by the coding sequence and overall base content. I ignored the codon bias for this analysis because it would be cyclic: the codon usage necessarily includes the base bias. I calculated the observed-expected base content for each base across ORFs, given the constraints of the coding sequence and the G/C content of the yeast genome. For instance, for an asparagine residue encoded by AAT, I observe one T at position 3, but expect 0.62 Ts since both AAT and AAC encode for asparagine and the yeast genome is 31% T, whereas the bases at positions 1 and 2 must be an A to still encode for asparagine, and so the number observed and expected are the same for these positions. I then averaged this data across ORFs yielding an observed-expected base content profile across ORFs (Figure 3.5B-D).

I found that, by accounting for the constraints of the coding sequence and overall base content, I could account for much of the bias in the coding sequence. Further, many of the trends observed at the beginning and ends of ORFs are almost entirely accounted for (compare Figure 3.5A and B). For verified ORFs, I find that the base bias is partly reversed, with there actually being more Ts than expected, approximately as many As as expected, and slightly fewer Gs and Cs than expected, but little difference between Gs and Cs across the transcript (Figure 3.5B). This may be accounted for by codon bias; if I consider uncharacterized ORFs, the trend is much closer to random (Figure 3.5B and C), indicating that the selection causing the base bias in verified ORFs is less strong in uncharacterized ORFs. Accordingly, these genes are generally much more lowly expressed compared with verified genes (10.02 vs 216.9 read count per million, on average [314]), and so would have less biased codon usage. A stronger bias occurs in dubious ORFs, but it is unlike the bias in verified ORFs (Figure 3.5D) and may be accounted for by the fact that many dubious ORFs are encoded antisense to other ORFs.
3.6.2 Over/underrepresented k-mers in coding sequence

Next I looked for motifs that were over or underrepresented in the coding sequence when compared to a background of randomized ORFs. In order to generate an appropriate set of background sequences, I shuffled the positions of the codons, while preserving both the numbers of each codon and the coding sequence of each ORF. I shuffled each ORF 20 times this way and used these shuffled ORFs to get a robust estimate of the expected
Chapter 3. Sequence properties of yeast transcripts

frequencies of each k-mer. Note that the enrichment/depletion of these motifs cannot be
due to codon bias because I am comparing the observed k-mer counts to those expected
in shuffled ORFs, which maintain the same codon bias.

The single most enriched sequence within ORFs is CAAGA (Table 3.6), which cor-
responds to the reverse complement binding site for the Nrd1 complex member, Nab3.
In fact, nine of the 32 most enriched k-mers contain reverse complements of either Nrd1
or Nab3 binding sites (Table 3.6), and these binding sites are consistent with a recent
study showing that Nab3 and Nrd1 may prefer to bind UCUUG and WGTAAG, respec-
tively [397]; these extended motifs are represented in the top two most enriched motifs
containing Nrd1 and Nab3 binding sites, for both factors.

Several of the enriched k-mers contain CAG/AGC/GCA repeats. About 70% of these
repeats (≥9 bp) are located in the CAG reading frame, indicating that they generally
encode for glutamine. CAG repeats encoding poly-glutamine are very well character-
ized repeat elements that are frequently found in transcription factors and other highly-
interacting proteins [432], and are implicated in many human diseases [190]. These
repeats are likely overrepresented because they form by expansion of the repeat [467]
and thus contain only the less frequent of the two possible glutamine codons.

In addition to enriched sequences, several sequences are depleted in ORFs. A potential
TF binding site was depleted in ORFs: CGCG, which corresponds to the binding site
for the chromatin remodelling factor, Rsc3 [20]. Since this sequence is identical on both
DNA strands, Rsc3 could potentially bind these sites in either orientation. Interestingly,
both GTCA G and its reverse compliment are also depleted, suggesting that another TF
may bind this sequence; however I was unable to find any TFs with matching specificities
[101]. In addition, poly-A sequences were found to be amongst the most depleted k-mers
in ORFs, and although poly-T sequences are also depleted, they are much less depleted
than poly-As (Sub-Z scores of -12.72 and -11.73 for A7 and A8 vs. -3.60 and -3.02 for T7
and T8). This difference is unlikely to be an artifact resulting from the coding sequence
Table 3.6: Top 50 motifs enriched and depleted in ORFs, compared with a randomized ORF background.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enriched</td>
<td>CAAGA</td>
<td>TTCTTG</td>
<td>25.21</td>
<td>24.86</td>
<td>0.35</td>
<td>RC(Nab3)</td>
</tr>
<tr>
<td></td>
<td>CAAGAA</td>
<td>TTCTTG</td>
<td>18.87</td>
<td>10.54</td>
<td>8.33</td>
<td>contains RC(Nab3)</td>
</tr>
<tr>
<td></td>
<td>TTTGAA</td>
<td>TTCAAA</td>
<td>18.18</td>
<td>12.90</td>
<td>5.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TGGTG</td>
<td>CACA</td>
<td>17.70</td>
<td>13.66</td>
<td>4.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTAC</td>
<td>GTAAA</td>
<td>14.53</td>
<td>10.79</td>
<td>3.73</td>
<td>contains RC(Nrd1)</td>
</tr>
<tr>
<td></td>
<td>TTGGG</td>
<td>CAAA</td>
<td>14.48</td>
<td>6.75</td>
<td>7.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TGCCA</td>
<td>TGGCA</td>
<td>12.38</td>
<td>11.64</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAAGAA</td>
<td>TTCTTG</td>
<td>12.08</td>
<td>10.55</td>
<td>1.54</td>
<td>contains RC(Nab3)</td>
</tr>
<tr>
<td></td>
<td>TGCCA</td>
<td>TGGCA</td>
<td>12.03</td>
<td>8.16</td>
<td>3.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAGTG</td>
<td>CACTTA</td>
<td>10.31</td>
<td>6.83</td>
<td>3.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTACCA</td>
<td>TCTAAA</td>
<td>8.62</td>
<td>6.97</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCACCA</td>
<td>TGGTGG</td>
<td>8.52</td>
<td>6.23</td>
<td>2.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTACCA</td>
<td>TTGTAAA</td>
<td>8.42</td>
<td>5.07</td>
<td>3.36</td>
<td>contains RC(Nrd1)</td>
</tr>
<tr>
<td></td>
<td>CAAAAGA</td>
<td>TTCTTTTG</td>
<td>8.39</td>
<td>6.92</td>
<td>1.47</td>
<td>contains RC(Nab3)</td>
</tr>
<tr>
<td></td>
<td>GCAGCAGC</td>
<td>GCTGCTGC</td>
<td>8.30</td>
<td>14.22</td>
<td>-5.92</td>
<td>CAG repeats</td>
</tr>
<tr>
<td></td>
<td>AGGGGC</td>
<td>GCCCTT</td>
<td>8.27</td>
<td>4.29</td>
<td>3.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAAGCG</td>
<td>GCTTA</td>
<td>8.21</td>
<td>5.62</td>
<td>2.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAAGTGA</td>
<td>CACTTTTC</td>
<td>8.19</td>
<td>5.66</td>
<td>2.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCCATTG</td>
<td>CAATGCG</td>
<td>7.95</td>
<td>5.39</td>
<td>2.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAGCAGCA</td>
<td>TGCTGCTG</td>
<td>7.79</td>
<td>15.26</td>
<td>-7.46</td>
<td>CAG repeats</td>
</tr>
<tr>
<td></td>
<td>GAAAAGGA</td>
<td>TTCTTTTC</td>
<td>7.27</td>
<td>5.11</td>
<td>2.16</td>
<td>contains RC(Nab3)</td>
</tr>
<tr>
<td></td>
<td>TGCCA</td>
<td>TGACA</td>
<td>7.14</td>
<td>7.13</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCACTTTT</td>
<td>AAAGTTGG</td>
<td>7.11</td>
<td>3.44</td>
<td>3.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAAATCA</td>
<td>TGATTTTC</td>
<td>7.03</td>
<td>5.43</td>
<td>1.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGCAGCAGC</td>
<td>CTGCTGCT</td>
<td>6.97</td>
<td>25.19</td>
<td>-18.22</td>
<td>CAG repeats</td>
</tr>
<tr>
<td></td>
<td>TTGGAAGA</td>
<td>TCTTCAAA</td>
<td>6.88</td>
<td>6.53</td>
<td>0.36</td>
<td>contains RC(Nab3)</td>
</tr>
<tr>
<td></td>
<td>TTGGAAAA</td>
<td>TTCTCAAA</td>
<td>6.87</td>
<td>5.62</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTACTG</td>
<td>CAGTAG</td>
<td>6.79</td>
<td>5.87</td>
<td>0.92</td>
<td>contains RC(Nrd1)</td>
</tr>
<tr>
<td></td>
<td>CAAATAT</td>
<td>ATATTTTG</td>
<td>6.74</td>
<td>4.78</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CG</td>
<td>CG</td>
<td>-37.35</td>
<td>-30.06</td>
<td>7.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTCAG</td>
<td>CTGAC</td>
<td>-13.99</td>
<td>-7.22</td>
<td>6.76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td>TCGG</td>
<td>-13.16</td>
<td>-12.62</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAAAAAA</td>
<td>TTTTTTTT</td>
<td>-12.72</td>
<td>-12.19</td>
<td>0.52</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>CGCG</td>
<td>CGCG</td>
<td>-12.55</td>
<td>-5.77</td>
<td>6.78</td>
<td>Rsc3</td>
</tr>
<tr>
<td></td>
<td>AAAAAAAA</td>
<td>TTTTTTTTT</td>
<td>-11.73</td>
<td>-6.34</td>
<td>5.39</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>ACATA</td>
<td>TATG</td>
<td>-9.52</td>
<td>-8.44</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTAAA</td>
<td>TTTTTTA</td>
<td>-8.74</td>
<td>-4.45</td>
<td>4.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATCC</td>
<td>GGGAT</td>
<td>-8.62</td>
<td>-5.59</td>
<td>3.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTTAA</td>
<td>TTAAG</td>
<td>-8.43</td>
<td>-4.99</td>
<td>3.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTTGAC</td>
<td>GTGAC</td>
<td>-8.36</td>
<td>-5.58</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAATTA</td>
<td>TTATTA</td>
<td>-7.86</td>
<td>-7.65</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGGGG</td>
<td>CCCCG</td>
<td>-7.22</td>
<td>-3.63</td>
<td>3.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTCCG</td>
<td>CGGAC</td>
<td>-7.19</td>
<td>-3.88</td>
<td>3.31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCCCCC</td>
<td>GGGGG</td>
<td>-7.11</td>
<td>-5.23</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATAGCT</td>
<td>AGCTAT</td>
<td>-6.88</td>
<td>-4.97</td>
<td>1.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTAATA</td>
<td>TATTTAA</td>
<td>-6.86</td>
<td>-6.76</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCGAC</td>
<td>GTGCG</td>
<td>-6.58</td>
<td>-3.55</td>
<td>3.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCGAG</td>
<td>CTGCG</td>
<td>-6.52</td>
<td>-3.09</td>
<td>3.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAAAAACG</td>
<td>CGTTTTTC</td>
<td>-6.51</td>
<td>-4.46</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTAACG</td>
<td>CGTTAG</td>
<td>-6.50</td>
<td>-5.76</td>
<td>0.74</td>
<td></td>
</tr>
</tbody>
</table>
because both AAA and TTT encode amino acids with only one other option, and the overall usage of each is similar between the two amino acids; lysine is encoded by both AAA and AAG, and overall 58% of lysines are AAA, while phenylalanine is encoded by TTT and TTC, and a similar fraction (59%) are encoded by TTT. Although both A- and U-rich sequences play roles in cleavage and polyadenylation [355], it is possible that the relative difference in the levels of depletion of poly-A and poly-U sequences in ORFs reflects differences in their ability to initiate transcript cleavage.

**Enriched/depleted k-mers in ORFs show a functional enrichment**

I next wanted to ask if, in general, there tended to be a functional enrichment of GO terms in ORFs containing these motifs. To address this, I compared how well the 50 most enriched/depleted k-mers could enrich/deplete GO terms, compared with all other k-mers (of length 2 to 8). I first counted the instances of each k-mer in all ORFs and performed a rank sum test for how well the number of motifs divided each GO slim term. This gave me p-values and ROC scores indicating the significance and relative enrichment of each GO slim term, for each k-mer. I next compared the distributions of ROC scores and p-values between the 50 most enriched/depleted k-mers and all possible k-mers by making a histogram illustrating the frequencies of p-values and ROC scores within each group (Figure 3.6). To ensure that differences in functional enrichment were not due to differences in distribution of the k-mers within each group (e.g. there are 17 5-mers and 7 8-mers in the top 50, when overall there are 64 times more 8-mers than 5-mers), I scaled the frequencies within each value of k such that the contribution to the overall histogram was identical between the two groups (e.g. for the background, the frequencies for 8-mers are scaled down while 5-mers are scaled up).

Overall, the enriched and depleted k-mers are more significantly associated with GO terms than randomly selected k-mers (Figure 3.6). For instance, the dinucleotide “CG” is the most significantly depleted k-mer in ORFs and is especially depleted in genes
Table 3.7: Association of GO slim terms with k-mer occurrence within ORFs for the 50 most significant associations. ROCs < 0.5 indicate a negative association between the k-mer and GO term. p-values are Bonferroni-adjusted.

<table>
<thead>
<tr>
<th>k-mer</th>
<th>GO ID</th>
<th>GO term</th>
<th>ROC</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG</td>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>0.161</td>
<td>1.52e-18</td>
</tr>
<tr>
<td>TGTTG</td>
<td>GO:0005886</td>
<td>plasma membrane</td>
<td>0.681</td>
<td>3.02e-27</td>
</tr>
<tr>
<td>TGTTG</td>
<td>GO:0003674</td>
<td>molecular function</td>
<td>0.387</td>
<td>1.88e-26</td>
</tr>
<tr>
<td>TGTTG</td>
<td>GO:0022857</td>
<td>transmembrane transporter activity</td>
<td>0.696</td>
<td>3.72e-25</td>
</tr>
<tr>
<td>TTTGG</td>
<td>GO:0005886</td>
<td>plasma membrane</td>
<td>0.670</td>
<td>2.52e-23</td>
</tr>
<tr>
<td>TTTGG</td>
<td>GO:0016020</td>
<td>membrane</td>
<td>0.598</td>
<td>2.43e-22</td>
</tr>
<tr>
<td>TTTGG</td>
<td>GO:0003674</td>
<td>molecular function</td>
<td>0.397</td>
<td>2.64e-21</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.696</td>
<td>2.19e-19</td>
</tr>
<tr>
<td>TTATAC</td>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>0.241</td>
<td>2.67e-19</td>
</tr>
<tr>
<td>TTTGG</td>
<td>GO:0022857</td>
<td>transmembrane transporter activity</td>
<td>0.675</td>
<td>3.21e-19</td>
</tr>
<tr>
<td>CAGA</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.692</td>
<td>7.58e-19</td>
</tr>
<tr>
<td>CCGA</td>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>0.247</td>
<td>3.98e-18</td>
</tr>
<tr>
<td>CAGCAGCA</td>
<td>GO:0006366</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.563</td>
<td>5.32e-18</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0003674</td>
<td>molecular function</td>
<td>0.408</td>
<td>2.55e-16</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0030427</td>
<td>site of polarized growth</td>
<td>0.683</td>
<td>6.13e-16</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0005740</td>
<td>mitochondrial envelope</td>
<td>0.347</td>
<td>6.63e-16</td>
</tr>
<tr>
<td>AAAAAGA</td>
<td>GO:0003634</td>
<td>nucleus</td>
<td>0.566</td>
<td>5.58e-15</td>
</tr>
<tr>
<td>TTATAC</td>
<td>GO:0016020</td>
<td>membrane</td>
<td>0.581</td>
<td>5.70e-15</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0005886</td>
<td>plasma membrane</td>
<td>0.640</td>
<td>1.29e-14</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0005840</td>
<td>ribosome</td>
<td>0.337</td>
<td>4.00e-14</td>
</tr>
<tr>
<td>TGCCA</td>
<td>GO:0003674</td>
<td>molecular function</td>
<td>0.415</td>
<td>4.31e-14</td>
</tr>
<tr>
<td>TGTTG</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.667</td>
<td>4.97e-14</td>
</tr>
<tr>
<td>TTTGAA</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.665</td>
<td>6.57e-14</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0000705</td>
<td>mitochondrial organization</td>
<td>0.354</td>
<td>1.30e-13</td>
</tr>
<tr>
<td>TGGTG</td>
<td>GO:0016020</td>
<td>membrane</td>
<td>0.578</td>
<td>1.57e-13</td>
</tr>
<tr>
<td>CAAGA</td>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>0.645</td>
<td>4.41e-13</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0005933</td>
<td>structural molecule activity</td>
<td>0.346</td>
<td>1.94e-12</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0003674</td>
<td>molecular function</td>
<td>0.421</td>
<td>3.21e-12</td>
</tr>
<tr>
<td>GAAGAA</td>
<td>GO:0005730</td>
<td>nucleolus</td>
<td>0.609</td>
<td>6.58e-12</td>
</tr>
<tr>
<td>ACATA</td>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>0.295</td>
<td>6.63e-12</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0005933</td>
<td>cellular bud</td>
<td>0.675</td>
<td>8.91e-12</td>
</tr>
<tr>
<td>CAAAGA</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.651</td>
<td>1.41e-11</td>
</tr>
<tr>
<td>TAAATCAA</td>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>0.625</td>
<td>2.89e-11</td>
</tr>
<tr>
<td>CAAGA</td>
<td>GO:0004386</td>
<td>helicase activity</td>
<td>0.776</td>
<td>3.08e-11</td>
</tr>
<tr>
<td>CAAGAA</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.651</td>
<td>3.58e-11</td>
</tr>
<tr>
<td>CAAGA</td>
<td>GO:0008150</td>
<td>biological process</td>
<td>0.374</td>
<td>4.37e-11</td>
</tr>
<tr>
<td>CAGCAGC</td>
<td>GO:0006366</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.542</td>
<td>5.73e-11</td>
</tr>
<tr>
<td>ACATA</td>
<td>GO:0006468</td>
<td>protein phosphorylation</td>
<td>0.660</td>
<td>6.27e-11</td>
</tr>
<tr>
<td>ACATA</td>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>0.629</td>
<td>1.12e-10</td>
</tr>
<tr>
<td>TTTGG</td>
<td>GO:0006811</td>
<td>ion transport</td>
<td>0.654</td>
<td>1.40e-10</td>
</tr>
<tr>
<td>TAAATCAA</td>
<td>GO:0001649</td>
<td>oxidoreductase activity</td>
<td>0.372</td>
<td>2.56e-10</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>0.631</td>
<td>4.32e-10</td>
</tr>
<tr>
<td>TGCCA</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.645</td>
<td>4.40e-10</td>
</tr>
<tr>
<td>TTATAC</td>
<td>GO:0016887</td>
<td>ATPase activity</td>
<td>0.645</td>
<td>4.40e-10</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0004386</td>
<td>helicase activity</td>
<td>0.764</td>
<td>5.31e-10</td>
</tr>
<tr>
<td>TGTTG</td>
<td>GO:0006811</td>
<td>ion transport</td>
<td>0.630</td>
<td>5.57e-10</td>
</tr>
<tr>
<td>CG</td>
<td>GO:0002181</td>
<td>cytoplasmic translation</td>
<td>0.233</td>
<td>7.83e-10</td>
</tr>
<tr>
<td>TAAATCAA</td>
<td>GO:0005634</td>
<td>nucleus</td>
<td>0.559</td>
<td>8.56e-10</td>
</tr>
<tr>
<td>TAAATCAA</td>
<td>GO:0006366</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.598</td>
<td>1.05e-09</td>
</tr>
<tr>
<td>CCGA</td>
<td>GO:0016301</td>
<td>kinase activity</td>
<td>0.656</td>
<td>1.23e-09</td>
</tr>
</tbody>
</table>
Figure 3.6: The most enriched/depleted k-mers within ORFs show functional enrichment. (A) Histogram comparing the frequencies of ROC values for GO-term enrichment for all k-mers vs. the top 50 most enriched/depleted k-mers. (B) Same as A, only comparing p-values for the GO-term enrichment.

involved in translation (Table 3.7); the term “structural constituent of ribosome” is very significantly separated by the number of CG dinucleotides within the ORF (ROC=0.161, \( P < 10^{-33} \)). Together this indicates that, as well as functioning generally, some of these sequence elements may function to regulate specific processes. There are several ways in which this could be modulated; for instance, by regulating the efficiency of transcript elongation (e.g. via chromatin modification), or controlling the longevity of the mRNA or translation efficiency via RNA-binding proteins.

### 3.7 3’ UTRs

Consistent with an earlier report [167], within the bodies of 3’ UTRs, A/T-content appears overall higher than surrounding regions, with T-content being especially high. After about 100 bp following the cleavage site, the overall base content is approximately equal to the genomic average (Figure 3.1D). The different stop codons are visible at the start
of the 3' UTR (Figure 3.1G), while the cleavage site motif is visible at the 3' end and is surrounded by a relatively high T content (Figure 3.1J). These biases towards A and T-rich DNA in the 3' UTR potentially reflect the presence of the A and U-rich elements that initiate transcript cleavage [355].

3.7.1 Over/underrepresented k-mers in 3' UTRs

I next analyzed over and underrepresented k-mers within 3' UTRs (Table 3.8). Consistent with previous analyses [501, 167], many of the most enriched k-mers correspond to binding sites for Hrp1, and variants thereof [408]. Many of the depleted sequences included the dinucleotide "TA", but contained other nucleotides that would prevent it from being part of the larger Hrp1 consensus of TAYRTA. I also found many enriched poly-A and poly-T sequences, which may represent binding sites for the other cleavage factors [355]. In general, the sequences corresponding to the known cleavage site cis-elements had a high K-S statistic, indicating a strong positional preference relative to the cleavage site (Table 3.8).

TTCTT/AAGAA motifs in 3' UTRs are recognized by Nab3

Similar to 5' UTRs, I also found an enrichment for TTCTT/AAGAA nucleotides, consistent with a previous report [167]. I hypothesized that, potentially as a result of the NFRs present at many 3' ends [550, 344], the TTCTT/AAGAA motifs might act as Nab3 binding sites in the 3' ends of genes as well. I performed the same analysis here as on these motifs in 5' UTRs, comparing motif occurrences in the 3' UTR to Nab3 occupancy [221] and CUT expression [433], and found that, as in 5' UTRs, there is a positive correlation between the number of Nab3 motif matches (TCTT) and both Nab3 occupancy and CUT expression on the sense (Spearman R = 0.149 and P<10^{-25}, and R=0.161 and P<10^{-29}, respectively) and antisense (Spearman R = 0.115 and P<10^{-15}, and R=0.0423 and P<0.003, respectively) strands. This suggests that these motifs could play a similar
### Table 3.8: Top 50 motifs enriched and depleted in 3’ UTRs, with K-Ss calculated for the 150 bp surrounding CPA sites.

<table>
<thead>
<tr>
<th>Direction</th>
<th>k-mer</th>
<th>RC(k-mer)</th>
<th>Sub-Z</th>
<th>Super-Z</th>
<th>Z-difference</th>
<th>K-S</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enriched</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTT</td>
<td>AAA</td>
<td>49.34</td>
<td>37.94</td>
<td>11.40</td>
<td>15.93</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>ATA</td>
<td>TAT</td>
<td>41.83</td>
<td>28.34</td>
<td>13.48</td>
<td>21.87</td>
<td>Hrp1 submotif</td>
</tr>
<tr>
<td></td>
<td>ATATA</td>
<td>TATAT</td>
<td>40.30</td>
<td>36.02</td>
<td>4.28</td>
<td>15.70</td>
<td>Hrp1 submotif</td>
</tr>
<tr>
<td></td>
<td>TTTT</td>
<td>AAAA</td>
<td>38.95</td>
<td>33.41</td>
<td>5.54</td>
<td>15.38</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>TATATA</td>
<td>TATATA</td>
<td>38.59</td>
<td>32.00</td>
<td>6.59</td>
<td>20.65</td>
<td>Hrp1</td>
</tr>
<tr>
<td></td>
<td>AAA</td>
<td>TTTTT</td>
<td>36.05</td>
<td>34.02</td>
<td>2.03</td>
<td>23.64</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>AAAAA</td>
<td>TTTTTT</td>
<td>34.66</td>
<td>33.70</td>
<td>0.96</td>
<td>21.67</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>AAAAAA</td>
<td>TTTTTT</td>
<td>34.20</td>
<td>32.99</td>
<td>1.21</td>
<td>19.04</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>TTTTT</td>
<td>AAAA</td>
<td>34.13</td>
<td>32.03</td>
<td>2.10</td>
<td>13.64</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>AAAAAAA</td>
<td>TTTTTTT</td>
<td>33.39</td>
<td>29.70</td>
<td>3.69</td>
<td>17.16</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>TTTTTT</td>
<td>AAAAAA</td>
<td>32.56</td>
<td>31.97</td>
<td>0.60</td>
<td>12.52</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>ATATA</td>
<td>TATATAT</td>
<td>32.42</td>
<td>30.26</td>
<td>2.16</td>
<td>15.97</td>
<td>contains Hrp1</td>
</tr>
<tr>
<td></td>
<td>TTTTTT</td>
<td>TTTTTTTT</td>
<td>32.40</td>
<td>28.42</td>
<td>3.98</td>
<td>11.76</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>ATATAT</td>
<td>ATATATAT</td>
<td>30.61</td>
<td>32.96</td>
<td>-2.35</td>
<td>13.41</td>
<td>contains Hrp1</td>
</tr>
<tr>
<td></td>
<td>AAAAAAA</td>
<td>TTTTTTTT</td>
<td>30.03</td>
<td>27.96</td>
<td>2.07</td>
<td>15.27</td>
<td>poly-A</td>
</tr>
<tr>
<td></td>
<td>TATATA</td>
<td>TATATA</td>
<td>29.18</td>
<td>33.95</td>
<td>-4.78</td>
<td>13.44</td>
<td>contains Hrp1</td>
</tr>
<tr>
<td></td>
<td>TTTTTTT</td>
<td>TTTTTTTTT</td>
<td>28.77</td>
<td>26.63</td>
<td>2.14</td>
<td>10.76</td>
<td>poly-T</td>
</tr>
<tr>
<td></td>
<td>GC</td>
<td>GC</td>
<td>27.79</td>
<td>11.31</td>
<td>16.49</td>
<td>5.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TGTA</td>
<td>TACA</td>
<td>18.33</td>
<td>11.23</td>
<td>7.10</td>
<td>15.36</td>
<td>Hrp1 submotif</td>
</tr>
<tr>
<td></td>
<td>TACATA</td>
<td>TATGTA</td>
<td>18.04</td>
<td>9.32</td>
<td>8.72</td>
<td>11.93</td>
<td>Hrp1</td>
</tr>
<tr>
<td></td>
<td>TATGTA</td>
<td>TACATA</td>
<td>17.89</td>
<td>8.09</td>
<td>9.80</td>
<td>12.46</td>
<td>Hrp1</td>
</tr>
<tr>
<td></td>
<td>AAATA</td>
<td>TATT</td>
<td>14.90</td>
<td>6.33</td>
<td>8.58</td>
<td>3.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TGTATA</td>
<td>TATA</td>
<td>14.69</td>
<td>9.32</td>
<td>5.37</td>
<td>10.98</td>
<td>Hrp1</td>
</tr>
<tr>
<td></td>
<td>CACAC</td>
<td>GTGTGTG</td>
<td>14.53</td>
<td>10.52</td>
<td>4.02</td>
<td>3.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TATTT</td>
<td>AAATA</td>
<td>14.10</td>
<td>5.08</td>
<td>9.02</td>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTAT</td>
<td>ATAAA</td>
<td>14.05</td>
<td>6.94</td>
<td>7.12</td>
<td>3.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATACA</td>
<td>TGTAT</td>
<td>13.79</td>
<td>10.75</td>
<td>3.04</td>
<td>9.14</td>
<td>Hrp1 submotif</td>
</tr>
<tr>
<td></td>
<td>TTCCTT</td>
<td>AAGAA</td>
<td>13.23</td>
<td>13.12</td>
<td>0.11</td>
<td>5.79</td>
<td>TTCCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>AAGAA</td>
<td>TTCCTT</td>
<td>13.02</td>
<td>11.12</td>
<td>1.90</td>
<td>5.46</td>
<td>TTCCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>TTCCTT</td>
<td>AAGAA</td>
<td>12.51</td>
<td>11.36</td>
<td>1.15</td>
<td>4.10</td>
<td>TTCCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td>TGC</td>
<td>11.91</td>
<td>3.84</td>
<td>8.07</td>
<td>2.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTCTT</td>
<td>AAGAAA</td>
<td>11.89</td>
<td>11.60</td>
<td>0.29</td>
<td>4.07</td>
<td>TTCCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>ATATA</td>
<td>TATAT</td>
<td>11.36</td>
<td>9.46</td>
<td>2.01</td>
<td>1.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAGAAA</td>
<td>TTTCTT</td>
<td>11.31</td>
<td>10.73</td>
<td>0.59</td>
<td>4.81</td>
<td>TTCCTT/AAGAA</td>
</tr>
<tr>
<td></td>
<td>TTTATTT</td>
<td>AATAA</td>
<td>11.15</td>
<td>4.81</td>
<td>6.34</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTCTTT</td>
<td>AAGAAA</td>
<td>11.00</td>
<td>9.32</td>
<td>1.68</td>
<td>3.00</td>
<td>TTCCTT/AAGAA</td>
</tr>
<tr>
<td><strong>Depleted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTA</td>
<td>TAA</td>
<td>-33.56</td>
<td>-17.11</td>
<td>16.44</td>
<td>2.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAA</td>
<td>TTA</td>
<td>-26.95</td>
<td>-20.64</td>
<td>6.31</td>
<td>5.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAT</td>
<td>ATT</td>
<td>-25.80</td>
<td>-17.26</td>
<td>8.54</td>
<td>6.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATT</td>
<td>AAT</td>
<td>-25.59</td>
<td>-17.67</td>
<td>7.91</td>
<td>6.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTA</td>
<td>TAG</td>
<td>-22.63</td>
<td>-10.50</td>
<td>12.13</td>
<td>3.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>GAT</td>
<td>-21.85</td>
<td>-9.52</td>
<td>12.33</td>
<td>3.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAT</td>
<td>ATC</td>
<td>-19.38</td>
<td>-8.88</td>
<td>10.50</td>
<td>2.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AATT</td>
<td>AATT</td>
<td>-16.66</td>
<td>-6.59</td>
<td>10.07</td>
<td>5.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTAA</td>
<td>TTAA</td>
<td>-16.53</td>
<td>-4.81</td>
<td>11.73</td>
<td>4.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGT</td>
<td>ACT</td>
<td>-15.84</td>
<td>-9.69</td>
<td>6.16</td>
<td>2.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TAG</td>
<td>CTA</td>
<td>-15.18</td>
<td>-14.14</td>
<td>1.04</td>
<td>8.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACT</td>
<td>AGT</td>
<td>-13.96</td>
<td>-9.62</td>
<td>4.34</td>
<td>2.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTC</td>
<td>GAC</td>
<td>-13.60</td>
<td>-7.46</td>
<td>6.14</td>
<td>1.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTAG</td>
<td>CTA</td>
<td>-11.93</td>
<td>-4.77</td>
<td>7.16</td>
<td>1.42</td>
<td></td>
</tr>
</tbody>
</table>
Clustered HRp1 binding sites

I noted that in addition to the optimal HRp1 binding site (TATAA), enriched k-mers also included TA-repeats of length >6. Strikingly, HRp1 binding sites in terminator regions are often flanked by additional TA repeats that tend to remain in phase (Figure 3.7).

Figure 3.7: HRp1 binding sites are clustered and in phase in terminator regions. Base content surrounding the 3457 optimal non-overlapping HRp1 binding sites in terminator regions (=150 bp upstream of CPA site). Colours indicate the base at the corresponding position, from 25 bp upstream to 25 bp downstream of the motif match.

I wanted to determine if HRp1 binding site clustering could be explained by cooperative binding, and so I looked at occurrences of the HRp1 binding motif TAYRTA, where Y=pyrimidine and R=purine [211, 408]. Since it had been previously shown that two HRp1 molecules are capable of simultaneously binding (UA)$_6$ but not (UA)$_5$ in vitro [387], I compared the occurrence of motifs that could be bound by two HRp1 molecules simultaneously (i.e. TAYRTATAYRTA) to those that could be bound by only one HRp1.

role in CUT recognition at both the 5’ and 3’ ends of genes.
molecule at a time (i.e. TAYRTAYRTA), ensuring each motif instance represented only two potential binding sites (e.g. TATATATAYRTA has three overlapping binding sites, and so was excluded). In terminator regions, I found 598 10 bp motifs representing two overlapping binding sites and 38 12 bp motifs representing two non-overlapping Hrp1 binding sites. By correcting for the two bases of extra information (TA) encoded in non-overlapping binding sites, I obtained the expected number of adjacent binding sites (57), which is greater than the number observed. Further, I also analyzed the number of (TA)$_6$ repeats (to which two Hrp1 molecules could simultaneously bind) and found these to be approximately the number expected given the abundance of (TA)$_5$ and (TA)$_7$, which were 447, 278 and 181 for 5, 6 and 7 AU repeats, respectively. I conclude that the overall pattern of Hrp1 motif occurrence in terminator regions is inconsistent with cooperative binding and hypothesize that this pattern may simply reflect the fact that each additional TA dinucleotide creates an additional potential Hrp1 binding site.

**Variation in Hrp1 binding site number has multiple explanations**

The diversity in the number of potential Hrp1 binding sites in terminator regions was surprising to me because I expected mRNA 3' end processing to function similarly across different genes, resulting in the same number of minimal elements present in each terminator. The fact that genes have diverse numbers of potential binding sites could mean that different transcripts require different amounts of Hrp1 binding for efficient 3' end processing or that 3' end processing is differentially regulated. I considered the possibility that 3' end processing could be differentially regulated for different cellular processes. However, when I compared the numbers of Hrp1 binding sites in terminator regions for genes with/without each GO slim term, none had a significant association ($P>0.05$ by rank sum, Bonferroni corrected). I found that expression level is significantly correlated with Hrp1 binding site number, where more highly expressed genes [314] have more Hrp1 binding sites (Spearman $R=0.123$, $P<5 \times 10^{-18}$). This association could be
a result of increased selective pressure on correct 3’ end formation for highly expressed genes, but could also be explained if genes with more Hrp1 binding sites are processed more efficiently, resulting in mRNAs with a longer half life. I next considered different CTD phosphorylation marks in terminator regions [33] and found that some showed a significant association with Hrp1 binding site number (maximum Spearman R=-0.167, P<2×10^{-17} for Ser7P). However, this too could be a cause or a result of the different numbers of binding sites. Genes in the convergent orientation have significantly more potential Hrp1 binding sites (ROC=0.531, P<0.0001), which could be a result of overlapping cleavage sites (see Chapter 4). Together, this shows that there is probably no single cause for the diversity in the numbers of potential Hrp1 binding sites in terminator regions.

Are Hrp1 binding sites also TATA boxes?

I noted that the optimal Hrp1 motif (TATATA [387, 408]) is extremely similar to the optimal TATA box motif (TATAAWWR [32]). This led me to ask whether the Hrp1 motifs were frequently occupied by TBP in terminator regions and, if so, whether these could initiate transcription (Figure 3.8A). To address this, I first identified perfect instances of the optimal Hrp1 motif (TATATA) in terminator regions and then measured the relative TBP and TFIIIB occupancy surrounding each motif using available ChIP-exo data [412]. To provide a basis for comparison, I collected the same data for true TATA boxes (instances of the TATA box motif in the region from -80 to -50 relative the TSS). These data show that Hrp1 binding sites and true TATA boxes tend to be occupied by TBP to a similar extent (Figure 3.8B and C). I next wanted to evaluate whether these motifs could nucleate preinitiation complex (PIC) formation, for which I used TFIIIB occupancy as a proxy. In promoters, this protein shows a clear bias towards binding downstream of the TATA box and a lack of binding at the TATA box motif (Figure 3.8C). For Hrp1 binding sites in terminator regions, the motifs are similarly depleted in binding by TFIIIB, but
generally have much lower levels of surrounding TFIIB and have no clear directional bias (Figure 3.8D and E), indicating that, although Hrp1 motifs appear similarly bound by TBP, PIC formation is much less efficient on these elements.

Figure 3.8: Hrp1 binding sites are bound by the TATA-binding protein. (A) Diagram of the region analyzed. The blue arrow represents the ends of ORFs, while the thick green arrow represents the sense mRNA containing an Hrp1 binding site (the alternating bright red and green). I analyzed the region to see if the corresponding DNA could be bound by TBP and nucleate PIC assembly. I analyzed the region surrounding the Hrp1 binding site for (C) TBP binding and (E) TFIIB binding. (B,C) The average number of TBP ChIP-exo [412] reads per base across (B) 453 TATA boxes in promoters, and (C) 3455 perfect instances of the Hrp1 motif. The surrounding lighter lines represent the 90th and 10th percentile estimates of the mean, sampling 100 times from the data, with replacement. (D,E) TFIIB ChIP-exo [412] reads surrounding (D) 453 TATA boxes, and (E) 3455 potential Hrp1 binding sites in 3’ UTRs.
Next, I asked if these PIC-like structures in terminators could explain the prevalence of antisense transcription [544]. For this, I compared the amount of TFIIB surrounding the Hrp1 binding site to the expression level of the region corresponding to the beginning of the antisense transcript, assuming the Hrp1 binding site is used as a TATA box. To gauge TFIIB occupancy, I summed the ChIP-exo read counts upstream of the Hrp1 binding site (-50 to the Hrp1 site), since TFIIB tends to be occupied more downstream of the TATA box in true promoters. When more than one perfect Hrp1 binding site was present in terminator regions, I only considered the most highly occupied site. I compared TFIIB occupancy to the average number of RNA-Seq reads [299] upstream of the Hrp1 binding site (-180:-80 relative to the Hrp1 binding site, on the antisense strand, Figure 3.8A), which would encompass the first part of antisense transcripts initiated using the Hrp1 binding site as a TATA box. I observed a weak, but significant correlation between antisense expression and TFIIB occupancy (Spearman $R=0.114$, $P<2 \times 10^{-8}$), which was on the same order as the agreement in the expression level between different RNA-Seq datasets [299, 314] for the same region (Spearman $R=0.107$, $P<3 \times 10^{-10}$). Importantly, if I compare the same TFIIB occupancy to transcription immediately surrounding the Hrp1 binding site (-50 to +50, relative to the Hrp1 site), which could not have been initiated using the Hrp1 motif as a TATA box, the two are no longer positively correlated. In fact, the correlation is significantly negative (Spearman $R=-0.070$, $P<6 \times 10^{-4}$), indicating that antisense transcription over the Hrp1 binding site may inhibit TFIIB binding and/or PIC formation.

3.8 Significance

I found my method of analyzing over and under represented sequence elements to be extremely sensitive and informative. Because of the comparative ease of my approach and the ability to compare to a desired background sequence, I can emphasize the differences
between two sets of sequences and control for known biases, such as coding sequence. The sensitivity was demonstrated by the recovery of many of the known functional elements in the regions analyzed, including a robust identification of all the known splicing elements, including the less well-described poly-U sequences.

In 5' UTRs, I find an enrichment of poly-A and poly-T elements which are thought to play roles in inhibiting nucleosome formation [215]. In 3' UTRs I find these same elements, which, in this location, correspond to binding sites for cleavage and polyadenylation factors [355], although these sequences are probably at least partly responsible for the low nucleosome occupancy observed at cleavage sites [462, 242]. I also find an enrichment for TA-repeats in 5' UTRs, which I show may act as alternate TATA boxes, while similar TA-repeat elements in 3' UTRs likely encode binding sites for the cleavage factor, Hrp1 [408], but can also act as TATA boxes that initiate antisense transcription. These Hrp1 binding sites frequently form arrays of TA-repeats, but their occurrence is inconsistent with combinatorial binding and likely reflects the presence of additional binding sites. Hrp1 binding site number correlates with several features, including expression level and PolIII CTD marks, but it is unclear whether this is a cause or a consequence of these binding sites. Many of the sequence elements are enriched in both 5' and 3' UTRs. This could explain why some genes are transcribed on both strands [544]; some terminator sequences are recognized as promoters.

I showed that the base bias observed within yeast transcripts can almost entirely be explained by the coding sequence and overall yeast G/C-content. In addition, I was able to find many enriched and depleted sequences within ORFs relative to a shuffled ORF background that maintains the same codon bias. I identified an over representation of CAG-repeats, which generally encode for poly-glutamine, and an underrepresentation of certain sequences that may represent TFBSs or binding sites for the yeast termination factors. The occurrence of some of the over and underrepresented sequences in ORFs can significantly segregate GO terms, which implies that they can have process-specific
functions, potentially regulating gene expression post-transcriptionally. Comparing the occurrences of motifs within ORFs involved in different processes might be fruitful for identifying more of these types of sequences.

I also find that binding sites for the Nrd1 complex are enriched antisense to ORFs as well as in both orientations in 5' and 3' UTRs, and demonstrate that these sites appear to be bound and functional \textit{in vivo}. These could play roles in preventing antisense transcription [544] and degrading the CUTs that are associated with promoters and cleavage sites [367, 538]; for instance, the sense CUTs that overlap 5' UTRs [367]. Alternatively, some of these sites could play a regulatory role (e.g. [357]). The prevalence of these motifs across transcriptional units implies a more global role for the Nrd1 complex than previously appreciated.

In addition to the sequence elements to which we were able to assign a role, I identified many sequence elements with unknown roles in transcription. Given that many of the elements to which we assigned functions correspond to the binding sites for known RBPs, it seems likely that many of the unknown motifs also correspond to binding sites for such factors. Of the RNA-binding proteins in yeast, few have known specificities [408], and so we may be able to assign additional functions to these sequence elements as more specificities become known.
Chapter 4

A unified model for yeast transcript definition

A modified version of this chapter is encompassed in the following publication, which is available under the terms of the Creative Commons Attribution License:


In its current form, I performed all the described work, except as noted here. Annie Bang (from Mount Sinai Hospital) performed the flow cytometry. The Donnelly Sequencing Centre performed the sequencing. I also had technical help and advice from Mariana Kekis, Debashish Ray, Kyle Tsui, Christoph Kurat, and Harm van Bakel. Debashish Ray, Kate Cook, Hong Zheng, and Ying Wu performed the Hrp1 RNA-compete experiment. Tim and I wrote the text, and were provided useful commentary from Harm van Bakel, Corey Nislow, Jack Greenblatt, Quaid Morris, and Alan Moses. I am also grateful to Alan Moses, Jack Greenblatt, and especially Quaid Morris for useful discussions regarding the nature of my findings, and modelling approaches.
4.1 Summary

Identifying genes in the genomic context is central to a cell’s ability to interpret the genome. Yet, in general, the signals used to define eukaryotic genes are poorly described. Here, I derived simple classifiers that identify where transcription will initiate and terminate using nucleic acid sequence features detectable by the yeast cell, which I integrate into a Unified Model (UM) that models transcription as a whole. The cis-elements that denote where transcription initiates function primarily through nucleosome depletion, and, using a synthetic promoter system, I show that most of these elements are sufficient to initiate transcription in vivo. Hrp1 binding sites are the major characteristic of terminators and can terminate transcription bidirectionally. The UM predicts global transcript structure by modelling the entire process of transcription using a hidden Markov model whose emissions are the outputs of the initiation and termination classifiers. I validated the novel predictions of the UM with available RNA-Seq data, and I test it further by directly comparing the transcript structure predicted by the model to the transcription generated by the cell for synthetic DNA segments of random design. I show that the UM identifies transcription start sites more accurately than the initiation classifier alone, indicating that the relative arrangement of promoter and terminator elements influences their function. My model presents a concrete description of how the cell defines transcription units, explains the existence of non-genic transcripts, and provides insight into genome evolution.

4.2 Motivation/introduction

Identification of the cellular mechanisms that define gene structure has been a long-standing problem in molecular biology: it is of interest not only for the study of transcription and its regulation, but also in understanding how new genes arise [71], creating synthetic regulatory elements [42], and explaining the many types of “cryptic” transcripts
[538, 544]. While much is known about individual elements that function in eukaryotic transcription initiation and termination (see below), to my knowledge it has not been systematically explored which elements are critical on a genome-wide basis, what proportion of all transcription these elements can account for, and how they work together across entire chromosomes. For example, sequences that can function as promoters in some contexts may be disabled by transcription originating from an upstream or antisense promoter via repressive changes in the chromatin environment [241, 203, 55] or transcriptional interference [348]. Interactions in cis are also clearly important for transcriptional terminators, since a sequence can only be used to terminate a transcript if it is first transcribed. Moreover, the existence of physical connections between the 5' and 3' ends of genes that depend on proper 3' end formation [478, 12] suggests the existence of a feedback mechanism between terminators and promoters.

In eukaryotes, protein-coding genes are transcribed by RNA Polymerase II (PolII), which is loaded onto the promoter region via the general transcription factors (GTFs) [233]. However, the GTFs generally lack sequence specificity. In yeast, the only exception is the TATA-binding protein (TBP), which recognizes the TATA box. Although there appear to be TATA-like elements bound by TBP in most promoters [412], only a minority have a canonical TATA box [32], and TATA boxes and TATA-like sequences also occur elsewhere in the genome. Motif matches for many yeast sequence-specific transcription factors (TFs) are enriched in promoters [123, 294], but it is not clear that they are sufficient to define transcription start sites (TSSs), as their activity is often dependent on other features of the promoter [215, 445]. Moreover, most yeast promoters have multiple TSSs, varying within a range of about 26 bp [386]. The exact start site appears to be controlled by the sequence surrounding the site [81, 178, 365] and has a consensus of YR (where Y=pyrimidine, and R=purine) [500], which is likely recognized by TFIIB and/or PolII [60, 304, 394].

In many eukaryotes, including yeast, a distinguishing feature of promoters is the pres-
ence of a nucleosome-depleted region (NDR) immediately upstream of the TSS [294, 550]. Yeast promoters often contain a high A/T content and poly-dA:dT tracts, which inherently inhibit nucleosome formation [215, 550, 242, 488]. Many promoters also contain binding sites for General Regulatory Factors (GRFs), including Rap1, Reb1, and Abf1 [185], which cause nucleosome depletion \textit{in vivo} where bound [294, 20]. However, none of these features are present in all promoters, and some occur outside of known promoters. Moreover, promoter definition is complicated by the fact that transcript initiation occurs bidirectionally in many promoter regions, with cryptic unstable transcripts (CUTs) produced by transcription in the reverse orientation [538, 367]. Promoters can further be regulated in \textit{cis} by transcription over the promoter region, which results in transcriptional interference or histone modification [241, 348].

Termination of coding transcripts in yeast (and other eukaryotes) occurs when a cleavage and polyadenylation (CPA) site is recognized, mRNA cleavage occurs, and a polyadenine tail is added, producing a mature mRNA. Cleavage tends to occur over a range of about 36 bases and prefers to occur at (C/G)AA motifs [386], which may be recognized by the 3' endonuclease Ysh1 [337, 148]. Following cleavage, the nascent mRNA no longer has a 5' m7G cap and becomes a substrate for a 5'-3' exonuclease that degrades the nascent RNA, leading to destabilization of the PolIII/RNA association and termination of transcription [94, 257, 523]. CPA sites have been previously described as comprising several relatively simple motifs that are bound and recognized by the sequence specific RNA-binding protein (RBP) components of the cleavage machinery. In yeast, these include the AU-rich efficiency element, bound by Hrp1 [80, 252], the A-rich positioning element, bound by Rna15 [172], and several U-rich elements surrounding the cleavage site (likely bound by the Yth1 and/or Yhh1 [470, 26, 106]). In addition to mRNAs, PolIII also synthesizes several types of non-coding transcripts, including snRNAs, snoRNAs, and CUTs, which have independent termination mechanisms [219]. For example, the RBPs Nab3 and Nrd1 recognize sequences present in RNAs and are important
for the maturation of snoRNAs [219] and the labelling of CUTs as TRAMP/exosome substrates [15, 536]. The CUT termination pathway provides a non-productive termination mechanism for PolII transcripts, as the resulting transcripts are immediately degraded.

Despite this extensive literature describing the sequences and factors involved, there has been no global examination of which features are critical to yeast gene identity, and therefore the problem of how yeast (or any other eukaryote) delineates transcription units remains an open question. Many programs to identify genes in genomic DNA exist (e.g. [458, 6, 58, 334]), but these generally rely on sequence features unlikely to be used in the process of transcription, including sequence conservation and open reading frames (ORFs). Attempts to identify promoters [399, 352] and CPA sites [168, 83] using mechanistic features, such as TF and RBP specificities, have met with varying success. To my knowledge no study has taken an integrated approach to model the entire process of transcript definition or has tested the predictions of the model in vivo. Several groups have used synthetic promoter systems to assay the expression levels of pools of constructs (e.g. [445, 407]). However, these studies aimed to measure how TF binding sites (TFBSs) or nucleosome positioning sequences affect expression level, and so they used a basal promoter sequence into which individual sequence elements were inserted or deleted. To my knowledge, no study has directly tested what sequences are necessary to generate a functional promoter in vivo.

Here, I have created a computer model that can explain genome-wide yeast transcript structure and is supported by multiple lines of experimental validation. The model is underpinned by classifiers that mimic the choices the cell makes when initiating and terminating transcripts. These classifiers indicate that the transcription of most yeast genes can be explained by relatively few features and reveal which trans-acting factors are most influential and which cis-elements help to define individual genes. Despite the fact that I did not incorporate CUTs into the training procedure, my model predicts bidirectional transcription from unidirectional promoters, indicating that the same sequence features
generally drive transcript initiation for both mRNAs and CUTs and that CUTs are an inherent, and possibly unavoidable, feature of yeast promoters. My model indicates that yeast CPA sites are also generally bidirectional, suggesting that convergent genes use the same termination elements. I combined these classifiers into a unified model that can predict where transcription will initiate more accurately than the initiation classifier alone, indicating that the relative arrangement of promoter and terminator elements in cis is a likely mechanism for orienting promoters. I go on to verify the predictions of the model using both existing data and experiments of my own design. This is the first model that describes how a eukaryotic cell defines transcript structure genome wide.

### 4.3 Models of transcript initiation, elongation, and termination

I first sought to ask how well I could explain the specific recognition of yeast promoters, transcripts, and CPA sites by cellular factors, using the known sequence features associated with these elements. This question can be framed as a computational classification problem, in which algorithms seek to classify input sequences as positives (e.g. promoter) or negatives (e.g. non-promoter) on the basis of features in the sequences (e.g. TF motif scores at specific positions). I refer to the resulting classifiers as the “initiation”, “elongation”, and “termination” classifiers because they identify the regions where transcripts initiate (promoters), elongate (transcript bodies), and terminate (CPA sites).

Because I wanted the models to mimic cellular mechanisms, I restricted the input features to those that can be realistically sensed by nuclear factors, including TF binding sites (TFBSs), nucleosome-excluding sequences, and DNA structural features, as well as the binding sites for RBPs and nucleotide content for genomic regions encompassing transcripts (since NTP concentrations can affect PolII elongation rates [342]) (Figure 4.1A). I subdivided promoter and CPA site sequences into bins to reflect known or potential
Figure 4.1: Design, refinement, and performance of the classifiers. (A) Classifier pipeline. Training and test examples were generated by calculating the relevant features (rounded boxes) using the DNA sequence of the example. The features were calculated over the bins shown in the coloured boxes. At the bottom of the coloured boxes, the components of the minimal feature sets are shown. Feature colours represent the feature type, including: transcription factors (TF), general transcription factors (GTF), base content (%NT), or RNA-binding proteins (RBP). (B) ROC curves representing initiation, elongation, and termination classifiers with either all features tested or the minimal feature sets, derived from the test data. The line $y=x$ represents the curve expected by random classification. (C) Same as B, only displaying precision-recall curves, with the legend as in B and separate “Random” curves for each classifier.
location preferences for specific features relative to the TSS or CPA sites (see Table 4.1), and calculated a single score (e.g. a TF motif score) for each feature within each bin. The elongation model had only a single bin. This binning procedure limits the resolution of the classifiers, since they cannot identify the exact locations of features within bins, but it has two major advantages. First, it accounts for the fact that most yeast promoters and terminators use a range of initiation and cleavage sites [386] (see Figure 4.3A and B). Second, it greatly reduces the number of features considered. A full list of the features initially included in each classifier is available in Supplementary Tables 1 and 2.

4.3.1 Creation of classifiers

I created the classifiers using Random Forests [48], an approach that is capable of capturing non-linear relationships between the features and classes, such as cooperative interactions among TFs. To produce positive examples for the two classifiers, I first identified TSSs and CPA sites for protein coding genes. I defined TSSs using those observed in Lipson et al., 2009 [314] and CPA sites using RNA-Seq reads containing poly-A sequences [364]. This yielded 5,010 genes with both TSS and CPA sites. I scored the initiation classifier features in six bins, as shown in Figure 4.1A, for 600 bp surrounding the TSS (from -500 to +100, relative to the TSS). I obtained negative examples by dividing non-promoter portions of the genome into overlapping 600 bp windows, yielding 72,276 negative examples. For the termination model, I scored features in three 50 base bins encompassing 150 bases surrounding the CPA site (from -75 to +75, relative to the CPA site; see Figure 4.1A). I derived negative examples by dividing sense ORFs into overlapping 150 bp segments, yielding 155,093 negative examples. A complete description of the sequence selection rules is given below. I split the genome into two halves (chromosomes 1-8 and 9-16), using half to train the classifiers and leaving the other half of the genome for testing the model. Within each training half, I created eight random forests: each training on seven chromosomes and withholding one chromosome for model refinement.
The predictions from the eight forests were then averaged to produce a classification score for the test data.

**Transcript annotations**

I created a new map of yeast transcripts based on initial manual annotations, and refined automatically based on RNA-Seq data. Manual identification of transcript boundaries was based on a combination of gene annotations, RNA-Seq, and tiling array expression data [99, 364, 499], and was originally performed by Harm van Bakel, Kyle Tsui, and Desiree Tilloy. I automatically searched for the best representative TSS within 100 bp of the manually annotated start using RNA-Seq data in which reads cluster at the TSS [314]. Using these data, I identified the site of the greatest increase in log(read-density) as the best representative TSS, where the number of \( \log_{10}(\text{reads}) \) had to be at least 1.5.

I defined CPA sites as the position with the highest read density of poly-A reads from [364] within 300 bp of the manually-curated transcript end, and I required that this site follow the stop codon. This procedure produced transcript TSS and TTS annotations for 5,010/5,772 of the non-dubious ORFs.

**Positive and negative classifier examples**

I took the promoter region as the sequence from -500 to +100 relative to the TSS, as defined above (see Table 4.1 and below for justification of this definition). The “positive” promoter set consisted of the 5,010 promoters from my transcript annotations. To generate “negative” promoter examples, I first identified all TSS regions, ORF starts, rRNAs, transposons, tRNAs, ARSs, telomeres, and ncRNAs. I then identified regions which contained none of these features and, among those greater than 601 bp, I divided the region up into a maximal number of negative examples, allowing 400 bp of overlap, and spaced out these examples within the region. For instance, a 1,004 bp region would generate three negative examples (1-602, 202-803, 403-1004). In all cases, I chose neg-
ative examples in a strand-specific manner, such that a positive example on one strand would not preclude a negative example at the same position on the opposite strand. This yielded 72,276 negative promoter examples.

Table 4.1: Known features of promoters and their occurrence in the initiation classifier bins.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Reference</th>
<th>-500:-200</th>
<th>-200:-150</th>
<th>-150:-80</th>
<th>-80:-50</th>
<th>-50:TSS</th>
<th>TSS:+100</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATA-box</td>
<td>[123]</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most TFs (e.g. Reb1, Cbf1, Abf1)</td>
<td>[123, 294]</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some TFs (e.g. Pbf1/2, Mcm1)</td>
<td>[123]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1 Nucleosome</td>
<td>[294]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>NFR (most genes)</td>
<td>[294]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>NFR (TATA genes)</td>
<td>[123]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>G/C-content/differences</td>
<td>[294]</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

The positive examples for the elongation classifier were derived by dividing transcripts and ORFs into a maximal number of non-overlapping 50 bp sections, while avoiding the last 100 bp of the transcript/ORF to avoid potential terminators. Negative examples were derived by dividing the rest of the genome up into non-overlapping 50 bp examples (excluding sense ORFs and transcripts). This yielded 165,109 positive and 257,338 negative examples with which to train the elongation classifier.

I defined the CPA region as -75 to +75 relative to the annotated CPA sites of transcripts, and I calculated features over three 50 bp bins as shown in Figure 4.1A. I chose these bins to capture the known positional preferences of known cleavage elements [175, 378, 107]. The “positive” examples included the CPA sites from my annotated transcripts (as described above), and I obtained “negative” examples by dividing ORFs (excluding 100 bp near the 3’ end) into a maximal number of (non-CPA) examples that could overlap by no more than 100 bp. This yielded 5,010 “positive” and 155,093 “negative” examples for the termination classifier.
Initiation classifier features

I calculated feature values independently over the six sequence windows (bins) shown in Figure 4.1A. I determined the bin locations manually on the basis of the characteristic architecture of yeast promoters (see Table 4.1).

Features can generally be grouped into four broad categories: motifs for TFs, motifs for RBPs, DNA structural features and base content, and nucleosome excluding sequences. These features are fully described in Supplementary Table 1. For the initiation classifier, I only included RBPs and strand-specific base content for the 100 bp following the TSS, since these are the only bases that are transcribed. I included all other features in all bins. For both RBPs and TFs, I calculated features using position frequency matrices (PFMs) representing the factor’s specificity. The “score” for a sequence bin is the probability of the factor binding anywhere in the sequence given the PFM, which I calculated using the method described previously [82]. For RBPs, I performed all motif scans strand-specifically, whereas I scanned both strands for TFs. Two exceptions to this are the TATA-box (including only the forward orientation) and Rap1 (including forward, reverse, and strand-unspecific orientations), since these factors had been shown to have asymmetrical binding effects [454, 35]. Initial analyses were complicated by the fact that there are many motifs of varying quality for each yeast TF; this motivated us to develop a database of expert curated yeast TF motifs [101], from which I used the Expert-Curated set of TF motifs (version 1.00). Other members of the Hughes lab also generated new motifs for Hrp1, Npl3, and Yra1 using the RNAcompete method [408] to expand my existing catalogue of RNA features [355, 380, 261, 260, 303, 104, 471, 470, 252].

I calculated DNA structural features as the average value across the sequence bin for each mono-, di-, or trinucleotide, which I mapped to the corresponding structural value [324, 430], as previously described [294]. The mappings are available on the accompanying website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/transcription_model/). For nucleosome excluding sequences [294], I simply counted the number of occurrences of
the hexanucleotide within the sequence bin. I calculated the poly-A feature as the total length of all poly-A tracts in the bin, only considering those with a length of at least five.

I calculated these features across the six bins for every example as input to the classifier algorithms. This yielded a matrix of 77,286 examples and 1,698 features which I then used to create the initiation classifier.

**Elongation and termination classifier features**

The elongation and termination classifier features included base content and RBP motifs. I calculated the RBP features in a strand-specific manner, as described above for the RBPs included in the initiation classifier. Base content included strand-specific counts of A, T, G, and C, as well as G/C-content, and the ratio of As to Ts and Gs to Cs. I calculated each feature over each of the three bins for both classifiers, yielding a matrix of 160,103 examples by 147 features that I used to create the termination classifier, and 422,447 examples by 49 features for the elongation classifier.

**Random forest classifiers**

I divided each chromosome into positive and negative examples for each classifier, where the initiation classifier compared examples of promoters to non-promoter sequences, the elongation classifier compared transcript sequences to non-transcript sequences, and the termination classifier compared the CPA sites of genes to gene bodies. I created the classifiers using the randomForests (version 4.6-6) R module [307]. I trained the classifiers on examples from half of the yeast chromosomes, and within these eight chromosomes, I left one chromosome out for evaluation of performance and optimization of the models (see below). For each chromosome left out, I made four replicate forests, each with 50 trees (facilitating distributed computation). Thus, each classifier consisted of 1,600 decision trees and was trained on half the genome. I used regression, sampling with replacement, and a minimum node size of five to make the forests. I used the resulting
classifiers to classify new data and they provided a score between zero and one for how promoter/transcript/CPA site-like the sequence is. Figure 4.1B and C shows the performance of the classifiers in segregating the positives from negatives in the test data. When scanning the test data (e.g. chromosomes 1-8), I averaged the predictions from all forests trained on the training data (e.g. trained on chromosomes 9-16, for each of 9-16 held out).

Model refinement

To minimize overfitting of the models, remove redundant features, and make the models easier to interpret, I sought to reduce the number of included features. I did this by iteratively rebuilding the model, adding one feature at a time, with features sorted by decreasing "importance" (the "mean decrease in node impurity" provided by Random Forests), and retaining only those that appreciably improved the performance of the model measured on the "refinement" data. I iteratively rebuilt the classifiers including only the top N most important features for increasing N. Each time I added another feature, I calculated the change in the error rate (ER) based on the held-out chromosome. The ER is defined as the probability that a randomly selected negative example has a score greater than a randomly selected positive example, and corresponds to one minus the area under the ROC curve (1-AUROC). I only retained features whose inclusion reduced the ER by at least 3% relative to the last ER. For the termination classifier, the ER was minimal after inclusion of the fourth feature, and so no further features were retained. This procedure allowed us to gauge how beneficial the addition of each additional feature is to the classification ability of each model, and allowed us to remove redundant features, since these would not provide an additional decrease in the error rate. Because this procedure was computationally intensive, I only considered the most important features when identifying the critical features (top 150 for initiation and 100 for termination and all 49 for elongation). However, no features were retained after
the 23rd and 4th most important features for the initiation and termination classifiers, respectively, indicating that few, if any, additional features beyond 150 and 100 would have been included had I continued the procedure with all features. Several different feature selection criteria yielded similar results. For instance, simply ranking by the Random Forests importance measure (same order as in Figure 4.2) achieves a similar feature set, but does not remove redundancies.

This feature reduction procedure only slightly decreased the performance of each classifier on the “test” data, while greatly reducing the numbers of features (Figure 4.1B and C). A striking result of this procedure is that the two classifiers are dominated by a small number of critical feature types: I retained a total of 15 features for the initiation classifier, two for the elongation classifier, and four features for the termination classifier (Figure 4.1A and Figure 4.2). I also found that, using these reduced feature sets, a linear model (logistic regression) performed only marginally worse than the Random Forests model (AUROC of 0.901 vs. 0.908 for initiation, and 0.946 vs. 0.952 for termination), indicating that there are few important non-linear feature interactions.

4.3.2 Critical features

The initiation classifier was largely dependent (14/15 features) on six main types of features: binding sites for the GRFs Reb1, Abf1, and Rap1; binding sites for Rsc3 (a putative GRF and component of the RSC complex [20]); DNA structural features and poly-A tracts (which correlate with G/C-content, presumably serving to deplete nucleosomes, and so are hereafter described collectively as “G/C-content” features); and TATA boxes (Figure 4.1A). A classifier with this reduced feature set has an AUROC only 2.6% lower than one containing all 1,698 features (AUROC = 0.908 vs. 0.934, Figure 4.1B), suggesting that these signals are responsible for establishing the identity of the vast majority of yeast promoters.

The elongation classifier retained only two features: strand specific T content,
Figure 4.2: Feature selection and feature importance for the initiation, elongation, and termination classifiers. (A-C) The change in error rate as each additional feature is included in the classifier, where the dashed lines indicate the feature inclusion threshold (3%) for each of the (A) initiation, (B) elongation, and (C) termination classifiers. Retained features are shown in bold. For each, only the top 40 features are shown as no further features attained the 3% threshold. Brackets indicate motif IDs when multiple motifs for the same factor are used and refer to Supplementary Tables 1 & 2. *The Stb2 motif appears to be a variant of the Reb1 Motif. **For the termination classifier, the minimal error rate is achieved upon inclusion of the fourth feature, so no more features were considered. (D, E) The \( \Delta \text{AUROC} \) for models created without each feature type, relative to the minimal (reduced feature set) for the (D) initiation and (E) termination classifiers.
the ratio of As to Ts in the transcript. Together, these yield a AUROC of 0.658 which, although not as good as the other classifiers, is still much better than expected by chance. However, as shown in Section 3.6.1 these features may simply reflect a bias from the coding sequence.

The termination classifier was reduced to only two feature types (Figure 4.1A): G/C-content and Hrp1 binding sites. Together, these confer an AUROC of 0.952 on the test data, only slightly lower than that obtained when all 147 features are included (AUROC = 0.968). I initially included A- and U-rich motifs to capture the canonical sequence specificities of the other CPA factors [172, 470, 26, 106], but these were not selected by the above procedure. This result might be explained by several possibilities: the sequence preferences of these factors may be better captured by base content than by the motifs; the motifs may not accurately capture the specificities of the factors; or the factors may not strongly impact cleavage site selection. G/C-content can also impact nucleosome occupancy, which appears low at terminators, although the significance of this observation is unknown [128, 88, 7, 49].

4.3.3 Resolution of the models

I wanted to compare the resolution of the initiation and termination classifiers to that achieved by the cell. To do this, I analyzed RNA-Seq data specific to TSSs/CPA sites [386], which I normalized within promoter regions (-300 to +300, relative to the annotated TSS) to get the % usage of each base as a TSS/CPA site. First, I wanted to gauge the TSS/CPA site resolution of the cell, so I aligned the % usage around the annotated TSS/CPA site (within 100 bp) to the most frequently used site (Figure 4.3A and B). When there were ties, I went with the closest site to the annotated site. I next aligned the classifier scores (within 100 bp of the annotated TSS/CPA site) and analyzed the TSS/CPA usage around the aligned scores (Figure 4.3C and D). Finally, to ask if I could further narrow down the exact site of initiation/cleavage, I identified motif matches for
the initiator and cleavage site motifs (which have the consensus sequences CA and SAAT, respectively) closest to the centre of the classifier score peaks. Here, I first searched for consensus matches to the motif within the peak and, if none were found, I searched for more and more degenerate motifs until one was found. I then aligned the % usage data to these peak-specific motif matches (Figure 4.3E and F).

I next asked how well the classifiers identify the TSS and CPA sites on a per-base level. Despite the low intrinsic resolution of the predictors (70 nt for the initiation model, and 50 nt for the termination model, due to the bin sizes), the majority of bases used as TSSs/CPA sites in promoter/terminator regions [386] lie near the centre of these peaks (Figure 4.3C and D). Moreover, searching for the initiator (CA) and the cleavage site ((G/C)AA) motifs within the peaks often identifies the exact TSS or CPA site (Figure 4.3E and F), consistent with earlier observations that the exact transcript start and end bases are determined by local sequence cues, following identification of the general regions for transcript initiation [178, 365] and termination [425].

4.3.4 Promoters and terminators are bidirectional in yeast

Strikingly, the key features for both the classifiers contain little strand specificity. The most important features of the initiation classifier (GRF sites, and G/C-content/poly-A sequences) are thought to function through their role in NDR formation [294, 20], and so should function bidirectionally, since NDRs have no orientation. Indeed, for the initiation classifier, the predictions for the top and bottom strands in the test set are highly correlated (Pearson R≈0.5), and the correlation is maximal when the DNA strands are offset by -208 bp, relative to the top strand. Thus, the initiation classifier predicts that many promoters initiate transcripts in either orientation. Indeed, the average distance between the predicted sense and antisense TSSs (212 bp) is consistent with bidirectional initiation observed in vivo [367].

Both the Hrp1 motif and G/C-content are symmetric, and correspondingly, the ter-
Figure 4.3: Resolution of the classifiers and of cellular transcription. Each panel shows a heatmap illustrating the % usage of each base as a TSS/CPA site across the aligned region for each gene, sorted by average TSS/CPA site position (weighted by the percent usage). Overlaid, is the average % usage across the region for all genes (i.e. the average of the data in the heatmap). (A, B) The % usage of each base as a (A) TSS and (B) CPA site, aligned to the most frequently used base. (C, D) The % usage of each base as a (C) TSS and (D) CPA site, aligned to the centre of the initiation and termination classifier peaks, respectively. (E, F) The % usage of each base as a (E) TSS and (F) CPA site, aligned to initiator and cleavage site motifs found within initiation and termination classifier peaks, respectively.
Figure 4.4: Terminators are bidirectional. Alignment of RNA-Seq reads corresponding to poly-A sites [364] on both DNA strands for convergent intergenic regions (Conv.) and elsewhere in the genome (Else.). Data are aligned to sense poly-A sites and include all poly-A sites in the genome with at least two reads. Data is smoothed over a 5 bp window.
mination classifier’s predictions are also highly correlated between the top and bottom DNA strands (Pearson $R \approx 0.8$ when offset by 90 bp), suggesting that the same cleavage signals can operate in either orientation. In fact, 42% of convergent genes have only one optimal Hrp1 site between them. Further, RNA-Seq reads containing poly-A sequences [364] indicate that a substantial fraction of CPA sites terminate transcripts in both orientations, that this is especially common for CPA sites between convergent genes, and that the distance between cleavage sites on either strand is consistent with the same cleavage signals being used to terminate both transcripts (Figure 4.4).

4.4 Testing the initiation classifier

To my knowledge, my classifiers represent the first rigorous demonstration that the few features highlighted can account for the identity of most yeast gene structures. I next sought experimental evidence that the features are necessary and sufficient in vivo.

Trans-acting factors affect transcription

I began by examining published expression data sets corresponding to mutations in the key trans-acting factors: Abf1, Reb1, Rap1, Rsc3, and Spt15 (TBP) [20, 11]. For a given DNA sequence, my model provides a score for how "promoter-like" a sequence is (from 0-1). To categorize each gene as “controlled” or “not controlled” by a factor, I calculated the difference between the initiation score for the full model and when the corresponding feature is set to the median value for non-promoters, effectively simulating that feature’s absence. If the score decreased by at least 0.1, I considered the model to predict that the gene’s promoter is controlled by the corresponding factor (“predicted-controlled”). I then compared these to the expression changes in the corresponding trans-factor mutants [20, 11] and gauged the significance of the association using the rank sum test (Figure 4.5). To compare this to simply using predicted binding sites, I used the predicted binding of
each TF given its motif (the same values as the inputs to the initiation classifier). For binding sites, I considered the highest ranking binding events to be targets and used the same number of targets for each TF as was used for the classifier scores.

In general, the expression of genes whose promoters are predicted to depend on a factor changes significantly more in the corresponding mutant (Figure 4.5). This finding demonstrates that the initiation classifier can identify the set of genes controlled by the predicted promoter-defining factors. Consistent with the notion that there is little interaction between promoter-defining features, motif scores for the individual TFs perturbed in each experiment predict the genes affected nearly as well as the initiation classifier (Figure 4.5).

4.4.1 Promoter-defining elements drive transcription in vivo

I next tested whether these cis-elements are sufficient to initiate gene expression in vivo using a combinatorial library of promoter constructs driving GFP, embedded in a context that otherwise has no promoter-like properties. These constructs encompassed the parts of the promoter most critical to the model (from -150 to +80 relative to the TSS). I designed the sequences in three different sections (Figure 4.6A) encompassing four of the bins used in the model (one of the sections encompassed two bins). The designed sequences were selected computationally from a large excess of randomly-generated sequences of varying G/C-content, with or without sites for Abf1, Reb1, Rap1, Rsc3, or TBP randomly placed within the regions in which they are relevant to the model. Some of the sequences are predicted to form functional promoters by virtue of their G/C-content alone. For each fragment containing a TFBS, I synthesized a fragment that is identical except that the TFBS was disrupted at several key bases. I used conventional oligonucleotide synthesis followed by pooled ligations to create a library of promoters from these promoter fragments that theoretically contained 86,688 distinct sequences. I assayed the relative activity of each promoter in yeast by using cell sorting and sequencing to esti-
Figure 4.5: Initiation classifier can predict the genes that will be affected by trans-acting factor mutants (sources: (A) [11] (B) [20]). This plot shows the ROC curve for how genes that are predicted to be controlled by a factor are stratified from those that are not by the expression changes in the corresponding trans-acting factor mutant. Thus, data above the line $Y=X$ indicates that the “predicted-controlled” genes have a higher expression level in the mutant, while data below this line indicates these genes have a generally lower expression in the mutant. AUROC and rank sum P-values are as indicated.
mate the expression level of each promoter, similar to a procedure used previously [445], yielding sufficient data to estimate the expression level of 48,928 different promoters.

I designed several hundred promoter segments encompassing the most important feature bins of the initiation model. I generated random DNA sequences, embedded TFBSs within each, and scored them by the initiation classifier in every possible combination. I selected those sequences that had relatively high initiation scores in multiple contexts for synthesis, and designed corresponding control sequences which were identical, but with the added TFBS perturbed. I also designed high and low G/C-content promoter segments that either score highly or lowly, but into which I did not specifically add TFBSs. TFBSs were generated by sampling from the position frequency matrix of the corresponding factor using the frequencies as weights. The segments had 43, 48, and 41 different sequences in the -150:-80, -80:TSS, and TSS:+80 positions, respectively, and these had complementary overhangs at either end that ensured the segments would combine in the desired order.

I ordered the promoter parts as single stranded oligonucleotides from Sigma-Aldrich. To generate double-stranded promoter parts, I phosphorylated and annealed the top and bottom strands for each sequence. I selectively phosphorylated (to prevent unwanted ligation products), and annealed the segments by denaturing at 95°C for 5 minutes and cooling slowly to 4°C (1°C/min). I then pooled the resulting double stranded promoter fragments, ligated them together at room temperature for 24 hours, and purified the full length promoters by gel extraction. I cloned the resulting promoters into a modified GFP expression vector [237] (pTH7638) using BamHI and NheI sites, and transformed them into E. coli by electroporation. I generated approximately 6 million transformants and, by pooling these, I isolated the plasmid library and transformed yeast en masse using the lithium acetate method. Approximately 2 million yeast transformants were generated and I pooled these to yield the final yeast promoter library.

I grew the pooled yeast library in SC-Leu overnight and isolated the cells, washed
them in water, and resuspended them in 1xTE to an OD of 0.68. Together with Annie Bang from Mount Sinai Hospital, we then sorted the cells on a FACSARia flow cytometer into six bins of GFP fluorescence and used gates to select for single cells. After sorting, I isolated the cells, resuspended them in SC-Leu, grew them overnight, and isolated the plasmids. I then barcoded the promoters for multiplexing using PCR, and the barcoded promoter libraries were sequenced in one lane on the Illumina HiSeq platform at the Donnelly Centre Sequencing Facility, using paired end sequencing, reading 111 and 114 bases from the ends to ensure I could uniquely identify the source promoters. I mapped reads to the promoter sequences using Bowtie [288] and only considered promoters that had > 50 reads (summing over all bins) for further analysis. I normalized the read count per promoter by the number of reads per bin to correct for differing numbers of reads per bin and multiplied this normalized count by the proportion of cells ending up in each bin. From this, I calculated the expression level by weighting the proportion of each promoter in each bin by the average fluorescence of the bin.

I obtained a strong and significant correlation between initiation score (a number between 0 and 1) and GFP expression level (Spearman R=0.64, i.e. 41% (R^2) of variance in ranks explained; P≈0) (Figure 4.6B), indicating that the features of the promoter model are sufficient for promoter function. At an optimal threshold (promoter score = 0.37, GFP expression level = 2.69), the promoter classifier achieves a true positive rate of 73% with a false positive rate of only 21%. Promoters containing Rap1, Reb1, and Abf1 binding sites, to which the model assigns high scores, are nearly always expressed (expression>2.69) (Figure 4.6C-E), and the expression level is consistently higher than promoters that are identical except for disruption of the binding site (e.g. Figure 4.6H). The impact of Rsc3 sites and the TATA box appear weaker; many promoters containing these elements are neither expressed nor predicted to be functional promoters (Figure 4.6F, G). Nonetheless, comparison of GFP levels from the promoters that are identical except for the presence and absence of Rsc3 or TATA elements showed that these elements have
Figure 4.6: Construction and analysis of the combinatorial promoter library. (A) Synthetic double stranded promoter fragments with complementary overhangs were ligated together to yield full-length promoters, which were then cloned into a GFP expression vector. I used flow cytometry and sequencing to measure the expression level of each promoter. (B) Point-density scatter plot showing the correlation between the initiation score and the expression level (log-scale). Darkness corresponds to point density. Horizontal and vertical lines indicate the expression level and initiation score thresholds for considering sequences “expressed” and a “predicted promoter”, respectively. (C-G) Identical to (B), but divided into promoters containing (C) Reb1, (D) Abf1, (E) Rap1, and (F) Rsc3 binding sites in the -150:-80 bin, and (G) the TATA box in the -80:-50 bin. (H-J) Point-density scatter plots showing the expression level of promoters that are identical except for the presence or absence of functional (H) Rap1, (I) Rsc3, or (J) TBP binding sites. The line y=x marks the point at which expression is identical between the two promoters, regardless of the binding site’s presence. The other GRFs (Abf1 and Reb1) are similar to Rap1 (H).
a positive impact on promoter function (Figure 4.6I, J). Overall, the model explains a considerable amount of the variation in GFP expression; the additional diversity in expression may be due to experimental variability as well as the presence of additional TFBSs not considered in the analysis.

4.5 A unified model improves computational gene identification

I next created a “Unified Model” (UM) aimed at describing the full process of transcription. The UM uses a hidden Markov model (HMM), which is ideal for this purpose because it can probabilistically model the states of PolIII as it initiates, transcribes, and terminates transcription in a similar way to how it occurs in the cell.

4.5.1 Creation of the UM

In essence, an HMM takes as inputs a model structure (as in Figure 4.7A) and a set of “observations” (i.e. the classifier scores for each base of a chromosome), and outputs the probability of the model being in each state at each base (i.e. chromosome-wide transcript structures). The model structure includes parameters describing the probability distributions of the observations in each state (i.e. the distributions of the classifier scores within each of the states shown in Figure 4.7A), as well as the probability distribution of transitions between states (e.g. the probability of going from intergenic to TSS+). Since there are eight different states and four different observations (two classifiers for both DNA strands) (Figure 4.7A), the model requires 32 (8×4) means and 32 (8×4) variances for observations, as well as 64 (8×8) different transition probabilities (40 of which are zero, since not all state transitions are allowed, e.g. Gene+ directly to Gene-). To ensure that the model is symmetric (a chromosome’s strand labels are arbitrarily assigned), I used the same means and variances for equivalent states between the + and - strands
Figure 4.7: A genome-scale yeast transcript model. (A) The structure of the Unified Model HMM. Circles represent states and arrows represent inter-state transitions. Inside state circles, the number of bases the model expects to remain in each state is shown in parentheses. Transition probabilities, as a percent of outgoing transitions, are shown on transition arrows. Very infrequent transitions (probability < 1%) are not shown. (B) Genome browser display illustrating the predictions of the models at the GAL1-10 locus of chromosome 2. The tracks on the top half represent data for the top strand of DNA, with the bottom strand on the lower half. From centre: blue bars represent genes, with thinner bars representing UTRs, and the grey bar represents a Ty element. Black tracks represent RNA-Seq read density on a log scale [299]. The Unified Model’s predictions are shown with dark green, blue, and red on a single track representing the probability of being in each of the states, where the probabilities are shown stacked. The light green and red tracks on the outer edge represent the scores for the initiation and termination classifiers, respectively. Initiation peaks corresponding to the true TSS and other potential TSSs for the CHS3 gene are as indicated, and some examples of predicted non-genic transcripts that are supported by RNA-Seq are shown boxed.
and ensured that the transition probabilities were symmetrical, resulting in a total of 56 non-zero parameters (see Tables 4.2 and 4.3).

Table 4.2: UM transition probabilities, representing the probability of transitioning from source states (first column) to destination states (other columns).

| Source | IG    | Gene+ | TSS+   | CPA+   | CPA-   | Gene- | TSS-   | CPA+/-
|--------|-------|-------|--------|--------|--------|-------|--------|---------
| IG     | 0.99777 | 0     | 0.001246 | 0.000986 | 0     | 0     | 1.19E-06 |
| Gene+  | 0     | 0.999 | 0       | 0.000865 | 0     | 0     | 0.00013796 |
| TSS+   | 0     | 0.047471 | 0.9525 | 0       | 0     | 0     | 2.42E-05 |
| CPA+   | 0.025805 | 0     | 0.002395 | 0.96983 | 5.35E-05 | 0     | 0.0019219 |
| CPA-   | 0     | 0     | 0       | 0.99983 | 5.35E-05 | 0     | 0       |
| Gene-  | 0     | 0     | 0       | 0       | 0.99983 | 0.001003 | 0     |
| TSS-   | 0.04424 | 0     | 0       | 0.003249 | 0     | 0.9525 | 6.05E-06 |
| CPA+/-.| 6.39E-05 | 0     | 9.13E-06 | 0       | 0.003934 | 0.009849 | 3.65E-05 | 0.98611 |

The UM states correspond to TSSs, transcript bodies, and CPA sites on the top and bottom strands, as well as a state corresponding to bidirectional transcript termination and a state for intergenic DNA (Figure 4.7A). Because TSS and CPA sites typically cover a range of bases [386], I represented TSS and CPA sites as states that can span multiple bases. As a result, I could not use my original transcript map (the one used to train the initiation and termination classifiers) to derive the UM parameters because it has only a single base for each TSS and CPA site, and so does not capture the actual range of bases used for these sites or the transitions between these and the other model states. Further, these annotations excluded many cellular RNAs, including ORF-containing genes for which I did not have TSS/CPA site estimates.

Table 4.3: Final UM observation distributions. These represent means and variances of BoxCox-transformed data (using \( \lambda \) of 0.050942 for initiation and -0.006859 for termination), which are normally distributed \( (N(\mu,\sigma^2)) \) for each state.

<table>
<thead>
<tr>
<th>State</th>
<th>initiation+</th>
<th>termination+</th>
<th>initiation-</th>
<th>termination-</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSS+</td>
<td>N(-0.4001, 2.0736)</td>
<td>N(-4.5387, 5.0823)</td>
<td>N(-4.4165, 2.5278)</td>
<td>N(-6.0073, 4.6229)</td>
</tr>
<tr>
<td>CPA+</td>
<td>N(-2.5527, 2.4549)</td>
<td>N(-0.63571, 7.4256)</td>
<td>N(-2.6374, 2.6348)</td>
<td>N(-1.9809, 6.6425)</td>
</tr>
<tr>
<td>CPA-</td>
<td>N(-2.6374, 2.6348)</td>
<td>N(-1.9809, 6.6425)</td>
<td>N(-2.5527, 2.4549)</td>
<td>N(-0.63571, 7.4256)</td>
</tr>
<tr>
<td>TSS-</td>
<td>N(-4.4165, 2.5278)</td>
<td>N(-6.0073, 4.6229)</td>
<td>N(-4.001, 2.0736)</td>
<td>N(-5.5387, 6.0823)</td>
</tr>
<tr>
<td>CPA+/-.</td>
<td>N(-3.0298, 2.1931)</td>
<td>N(-0.084816, 4.8119)</td>
<td>N(-3.0298, 2.1931)</td>
<td>N(-0.084816, 4.8119)</td>
</tr>
</tbody>
</table>
To overcome these limitations, I defined a new transcript map that could be used in deriving the HMM parameters. In my initial analyses, the model outputs (i.e. the predicted transcript structures) were surprisingly sensitive to the choice of parameters and the transcript maps from which they were derived. Ultimately, I derived the parameters using a transcript map based on multiple types of RNA-Seq data [314, 299].

Defining a new transcript map to train the UM

Overview. The HMM parameters included the mean and variance of the “observations” (the two classifier scores for both strands) in each state (the nodes shown in Figure 4.7), as well as the probabilities of transitioning between the states shown in Figure 4.7. To obtain these parameters, I needed a map of the occurrence of the HMM states to use as training data. For reasons previously described, I could not use the original transcript maps employed for the initiation and termination classifiers. I therefore used RNA-Seq data [299, 378] to create a state label map. The state labels needed to have the same states as the UM. I therefore used a procedure similar to that used to derive the initiation and termination classifiers, and the UM, but instead based on RNA-Seq data. I first developed classifiers that identify the range of TSS/CPA sites, which I then combined into a HMM capable of predicting transcript structure. Finally, I used this HMM to label the bases of the genome with the corresponding UM states, based entirely on RNA-Seq data.

Classifiers. I created classifiers that would identify TSSs and CPA sites given RNA-Seq data [299, 378] as features (designated TSS-R/CPA-R, where “R” stands for “RNA-Seq”). I made these classifiers with random forests and designed them to distinguish TSSs and CPA sites from non-TSS and non-CPA sites, respectively, using raw read counts and the change in counts over different windows (since read counts go up at the TSS and down at the CPA site), as well as CPA-specific reads (i.e. poly-A reads). These classifiers provided a score reflecting how much each base in the genome resembles a
TSS/CPA site based on RNA-Seq data surrounding that base (see Figure 4.8B for an example of the predictions).

Hidden Markov Model. I then used the TSS-R and CPA-R scores, as well as the raw RNA-Seq data, to train a HMM (designated HMM-R) using the Hidden Markov Model Toolbox for Matlab (http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html). The HMM outputs states corresponding to those in Figure 4.8A. In order to train the HMM-R, I required a set of state labels for every base in the genome representing the TSS, transcript, CPA site, and intergenic states (treating strands independently, see Figure 4.8A). I derived these labels by identifying peaks in the TSS-R score upstream of ORFs, and peaks in the CPA-R score downstream of ORFs (both within 1 kb of the ORF start/end) and labelling these as the TSS and CPA states, respectively. I labelled everything between these peaks as the transcript state and I labelled everything else intergenic. I then used these state labels to train the RNA-Seq based HMM-R.

The observations of the HMM-R included the TSS-R and CPA-R classifier scores, as well as log read counts for RNA-Seq data [299, 314] for every base. I assumed the observation distributions are Gaussian within each state, with each state having a mean and variance for each observation. I calculated the maximum likelihood parameters given the aforementioned state labels based on CPA-R/TSS-R peaks surrounding ORFs for half the chromosomes (1-8, or 9-16) and then used the resulting HMM-R to generate a transcript state map for the other half, swapping the two chromosome sets to get maps for both halves. I used the Viterbi path, corresponding to the single most likely path through all the HMM states given the observations, to generate two state label maps (one per strand) that correspond to the UM states. Since these are derived from RNA-Seq data, they include both ORFs and non-ORF stable transcripts. I removed state labels corresponding to very small transcripts (<150 bp) since these may represent artifacts in the RNA-Seq data. Next, I combined the single stranded state labels such that CPA sites were allowed to overlap (in the bidirectional terminator case), but transcripts and TSSs
Figure 4.8: Structure and example predictions of RNA-Seq-based transcript identifiers. (A) The structure of the HMM-R HMM which uses RNA-Seq to identify transcript structure. (B) The predictions of HMM-R, TSS-R, and CPA-R are depicted on a part of chromosome 1. Here, TSS-R scores for the top and bottom strands are shown in light green, with CPA-R predictions immediately below. The two HMM-R tracks depict merged transcript state maps, where, when both agree, both sets include the transcript and where they conflict, as in the case of FUN14 and ERP2, the gene on the top strand is included in “top” and the bottom-strand gene in “bottom”. In the centre, gene annotations are shown in dark blue, with thinner bars representing the UTRs.
could not. In the case of two convergent transcripts where the ends of transcripts overlap, I treated the overlapping bases as bidirectional terminators up to an overlap of 130 bp, above which the transcripts were considered to conflict. When the state labels conflicted, I kept the two strands separate, resulting in two final state label maps designated “top” and “bottom”, which are identical except where transcripts conflict, in which case they contain the top and bottom strand predictions, respectively (Figure 4.8B). This yielded two state label maps for the entire genome, which could then be used for training the UM. When calculating the maximum likelihood (ML) estimates for the UM parameters, I used both “top” and “bottom” state labels, essentially weighting the conflicting data half as much in the ML estimate.

Building the unified model

I built the unified model (UM) using the Hidden Markov Model Toolbox for Matlab (http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html). For the initial parameters, I used the ML estimates given the RNA-Seq-based transcript state label map (derived as described above) and the corresponding classifier scores. This yielded a matrix of transition probabilities representing the probability of moving from every state to every other state, as well as a matrix of means and a matrix of variances representing the expected observation distributions for each classifier score in each state (see Tables 4.2 and 4.3). I assumed that the observations were approximately normally distributed within each state and calculated the observation probabilities (probability of observing a given combination of classifier scores in a given state) using a Gaussian probability density function with diagonal covariance (one mean and one variance for each classifier score, for each strand). The classifier scores were not initially normally distributed, so I first transformed them to be approximately normal by Box-Cox power transformation.

I noted that the performance of the UM was initially worse than expected and hypothesized that this was because the ML parameter estimates did not accurately represent
the parameters used by the cell. To address this, I optimized the observation means to maximize the Pearson correlation of the transcript predictions to observed RNA-Seq read counts, on a per-base level [299], for the training chromosomes, using a hill-climbing algorithm. To reduce overfitting of the parameters, I first scaled all means together, then I scaled the initiation and termination means together, then I scaled each mean individually, each time optimizing until convergence. I tuned the means iteratively such that each was changed by at most 50% per round. Each time no additional changes to the parameters improve the correlation, I halved the amount by which the means are changed (50%, 25%, 12.5%, etc.) and repeated the process until the parameters converge. The algorithm stops when the amount by which the parameters are changed is 3.125%. This procedure did not add any additional parameters to the model, since I altered the values of existing parameters, and I found it selected a stable set of parameters and improved the quality of the model’s predictions (as measured by the AUROC predicting TSSs, CPA sites, and transcribed bases). The final observation distribution parameters are shown in Table 4.3. I speculate that this process may compensate for errors or omissions in the initial transcript map. It is also possible that some aspects of transcription are not captured well by the binary state annotations in my transcript map.

Observations that are extremely unlikely in any state can tend to dominate the predictions of a HMM. For example, if an observation probability is nearly zero in all states but is many fold higher in one state compared with the others, the most probable state will be strongly favoured. For this reason, I also added a constant (0.001) to the observation probabilities of every state, which I then re-normalized to sum to one over all states, preventing these extremely improbable observations from dominating the prediction. I used the forward-backward algorithm to generate the predictions of the model, which yields the probability of being in each state at every base, given the classifier scores for the entire chromosome.

Here I present results for the UM trained on chromosomes 9-16 and tested on chro-
mosomes 1-8 (similar model parameters and performance were attained by swapping the training and test data).

4.5.2 Evaluation of the model

I could now use the UM to scan the test data to get transcript predictions across that half of the genome. I used these data to evaluate the quality of the model’s predictions relative to my annotated transcripts and expression data.

Model performs better without elongation

Given the classifier scores for an entire chromosome, the UM outputs the probability of being in each state at every base. The probability corresponds to the posterior marginals of the HMM states. I first wanted to quantify the accuracy of the predictions of the UM. I tried several different scoring metrics, but settled on ROC/precision-recall analysis for how well the model is able to differentiate the positive and negative examples of the classifiers for TSSs and CPA sites, as well as for transcribed vs. non-transcribed bases (where transcribed bases include those that are within a transcript in the transcript map, or within an open reading frame). This allowed me to compare how different variants of the UM perform in the different aspects of identifying transcriptional units.

The first comparison I made was between a UM including the elongation model and a UM without it. Surprisingly, the UM performs better in every respect when it does not include the elongation model scores as evidence (Figure 4.9A and B). I believe this is because the elongation model is so noisy that the evidence it provides is of extremely low confidence, resulting in it often misleading the predictions. By manual inspection of the predictions of the UM including the elongation classifier, this seems to be the case; the predicted transcribed strand often switches back and forth within a gene to whichever strand is preferred by the elongation model. Regardless, because the model including the elongation classifier performed worse than the model without, all subsequent analyses
were performed on the UM including only the initiation and termination classifiers.

An example of the UM predictions at the GAL1-10 locus (and surrounding region), located on Chromosome 2, is shown in Figure 4.7B. There is generally a good correspondence between gene annotations and the transcript predictions (and the exceptions are informative; see below). The UM robustly predicts TSSs, CPAs, and transcripts of known genes (Figure 4.9A and B). I next calculated the overlap of the model’s predictions with transcript predictions on a base-by-base basis. The positive set of transcripts included non-dubious ORF-containing genes and, when available, the boundaries represent the TSSs and CPA sites from the transcript annotations used to train the classifiers, but when unavailable, the boundaries represent the ORF start and end. I considered a base a "predicted transcript" if the UM probability of being part of a transcript was over 0.5 as this value precludes there also being a transcript predicted on the opposite strand. The overall sensitivity (recall) of detecting transcripts on a base-by-base level is 76.6%, and the precision is 76.9% (taken as predicted transcribed bases that are within 100 bp of a known transcript). For comparison, random guessing with a probability of 42.1% (the proportion of the genome encompassed by a transcript on a given strand) would yield precision and recall values of 42.1%. TSSs and CPAs are much more sparse than transcripts, yet the ROC and Precision-Recall analyses indicate that the UM is also adept at identifying these elements (Figure 4.9A and C).

**UM identifies TSSs better than the initiation classifier**

The UM also allowed us to ask whether the interaction of sequence features along the chromosome is important to their function *in vivo*. If so, the UM would be expected to be better at identifying TSSs than the initiation classifier alone. Indeed, the UM is much better than the initiation classifier at distinguishing true TSSs from bases that are within ORFs (AUROC=0.925 vs. 0.863, Figure 4.9B), indicating that the relative arrangement of promoter and CPA sites on the chromosome has an important influence on the usage of
the individual elements, and that this context-dependence is being modelled by the UM.

Manual examination of the UM and the classifier outputs confirm that the HMM often
identifies the correct TSS for a given gene even when the initiation classifier identifies
other, often stronger, potential TSSs nearby, suggesting that this context-dependence is
being incorporated in the UM. CHS3 in Figure 4.7B is one such example (indicated with
arrows).

4.6 Predictions of the UM

Now that I had a model capable of predicting where transcripts would be generated, I
wanted to see whether there was evidence to support the existence of the novel predicted
transcript species, and I wanted to ask how frequently transcripts would be generated in
random DNA sequence.

4.6.1 Predicted transcript species are transcribed

I next investigated the non-genic transcripts predicted by the model. Transposons and
sn/snoRNAs, although not considered as known transcripts in the analyses above, are
also generated by PolII [297, 413]. These elements encompass 2.95% of predicted tran-
scribed intergenic bases, and overall 71.3% of the bases of these elements are predicted
to be transcribed, despite these elements being absent from the classifier training data.
Remainder inconsistencies between the model and the annotated transcripts are split ap-
proximately evenly between predicting transcription on the wrong strand (15.3% of pre-
dicted transcribed bases), and predicting transcripts in intergenic regions (>100 bases
from ORF-containing transcripts; 11.7% of predicted transcribed bases). Figure 4.7B
contains several such instances. In addition to capturing most of the known genes, the
model identifies a known antisense transcript of GAL10 [203], transcripts antisense to
ETR1 and YBR016W, and a short intergenic transcript between ETR1 and OLA1. No-
Figure 4.9: Performance of the UM. (A) ROC curves illustrating how well the UM predicts TSSs, transcripts, and CPA sites, when classifying the positive and negative examples for the initiation and termination classifiers, as well as ORFs/transcripts and non-transcript bases, for both the UM with all three classifiers, and the final UM, lacking the elongation classifier. AUC values as indicated. (B) Same as A, but with precision-recall curves. (C) ROC curve comparing the ability of both the UM and initiation classifier to distinguish between TSSs and bases that are part of non-dubious ORFs. The line y=x represents the curve expected by random classification.
tably, all are also observed in the RNA-Seq data.

In order to determine more globally whether there was evidence to support the existence of transcripts that do not correspond to known features, I compared the UM’s predictions to available RNA-Seq data. For calculating percent overlaps, for instance, with antisense transcripts, I found the number of bases that were both transcribed (>10 reads) and predicted to be transcribed, and compared this to the overlap expected by chance (\(\%\text{transcribed} \times \%\text{predicted}\)). Among the transcripts predicted antisense to known genes, 23.5% of bases are supported by strand-specific RNA-Seq reads [314] (compared with 8.5% expected by chance) and 39.0% of intergenic bases that are predicted to be transcribed are supported by the same RNA-Seq data (compared with 25.6% expected by chance).

Applying the same 0.5 probability threshold to TSS and CPA predictions genome-wide, I find that 25.7% of predicted TSSs and 10.8% of predicted CPA sites are supported by TSS- and CPA-specific RNA-Seq reads [386] (compared with 0.0125% and 0.0438% expected by chance). If I consider only bases outside ORF-containing transcripts (>100 bp away), 5.58% of bases predicted to be part of a TSS and 2.52% of bases predicted to be part of a CPA site are supported by these same data (vs. 0.000615% and 0.0131%, respectively, expected by chance). This is a very positive result given that low abundance transcripts may not be adequately represented in the data, not all bases are used to initiate/terminate transcripts within promoter/terminator regions (see Figure 4.3A and B), and passage through the TSS and CPA states can be very transient and so frequently do not get over 50% within a promoter or terminator region. Further, bidirectional terminators (which are common in convergently-transcribed genes) have their own state that encompasses the terminator regions of both transcripts, whereas in reality both convergent genes undergo cleavage downstream of the cleavage sites of their partner, resulting in an overestimation of the number of bases encompassing the terminator of each individual gene. Regardless, these data overlap with the model’s predictions many
times more than expected by chance.

Figure 4.10: Novel predicted transcripts are often transcribed. (A) dUTP RNA-Seq data from [299], aligned across 1,286 transcripts predicted on chromosomes 1-8 that do not correspond to known transcripts or other genomic features. (B) Average across the predicted transcripts for both the RNA-Seq data shown in (A), and NET-Seq data from [89], smoothed using a 10 bp sliding window. Surrounding bases (upstream and downstream of predicted transcripts) are frequently known genes and tend to be more highly expressed than the predicted transcripts, and so have a greater read count and PolII occupancy.

Finally, I examined complete transcript predictions that do not overlap known features. I first defined transcriptional units predicted by the model. I treated each strand independently and, for the top strand, I identified potential TSSs by, starting from the first base of the chromosome, identifying regions with an increase in the probability of being transcribed of at least 0.08 over 120 bp. I wanted to avoid excluding potential
transcripts, and so intentionally set this threshold relatively low. I then identified CPA sites that followed potential TSSs using the same criteria, only instead using the decrease in the probability of being a transcript. Where suitable TSS and CPA predictions were both found, I considered the position of the maximum increase and maximum decrease in the probability of being transcribed as the TSS and CPA sites of the predicted transcriptional units. I treated the bottom DNA strand identically, only with the direction reversed. I then removed any transcripts that overlap on the same strand with ORFs, annotated transcripts, transposons, sn/snoRNAs, or other known features, leaving us with a set of predicted transcripts that do not correspond to known features. I scaled the data in Figure 4.10 such that the starts and ends of the predicted transcripts line up and averaged and scaled these so that the minimum and maximum between the two data sets displayed in Figure 4.10B are comparable. I find that available RNA-Seq [299] and NET-Seq [89] data show that the predicted TSSs and CPA sites of non-genic transcripts appear in aggregate to be largely correct (Figure 4.10). These comparisons demonstrate that the model is capturing many real non-genic transcripts, and suggest that these transcripts are produced by the same mechanisms that produce conventional PolII transcripts.

4.6.2 Stable transcripts are produced by random sequence

An intriguing feature of my classifiers and UM is that the critical sequence features are relatively simple and would therefore be expected to arise in random sequence at a relatively high frequency. For example, the binding sites for Reb1, Abf1, and Rap1 have only 7, 7, and 9 critical bases, respectively, and tolerate some degeneracy; thus, binding sites for at least one of these factors should appear approximately every two kb in randomly-generated sequence. Hrp1 binding sites, with a six-base optimal binding site, appear even more frequently - roughly once every kb in the A/T-rich yeast genome. Thus, relatively long transcripts should be produced with some frequency even from randomly generated DNA sequence.
Prevalence of transcript-like structures in random DNA sequence

In order to estimate how often promoter-like sequences will originate in DNA with randomly-generated sequences, I created ten independent 10 Mb DNA sequences in which the base frequency matches that of the yeast genome (A=T=0.31). I then scanned these sequences with the initiation classifier. I identified peaks meeting a score threshold corresponding to the median score of true promoters (0.335), and called predicted promoter peaks (which were required to be at least 100 bp apart and selected in a greedy fashion with the highest scoring peaks selected first). The average number of bases between predicted promoters was in good agreement between the ten sequences (940 bp, SD=6 bp among the ten 10 Mb sequences).

To estimate the fraction of bases that are predicted to be transcribed, I scanned these same 10 Mb sequences with the termination classifier and used the initiation and termination classifier scores to predict gene structure using the UM. I then counted the fraction of bases that were predicted to be transcribed (probability >0.5) on each strand. The percent transcribed for each strand ranged from 32.7% to 33.4% for the ten sequences. An overall average of 33.1% of each strand is predicted to be transcribed, and so 66.2% of the sequences are predicted to be transcribed on one of the two strands, not allowing overlapping transcripts.

Generation and analysis of 6 kb randomly-generated DNA

To confirm that such transcripts do arise and are consistent with the predictions of the UM, I assayed transcription in vivo from synthetic DNA fragments that were integrated into the yeast genome. I tested four 6 kb fragments, each composed of two tandem 3 kb DNA fragments, denoted A1B1, A1B2, A2B1, and A2B2. These fragments contained randomly-generated sequences with a G/C-content similar to that of the yeast genome (38%), into which I randomly inserted binding sites for Rap1, Abf1, Rsc3, Reb1, and TBP, to reduce the amount of DNA to be synthesized. The 3 kb fragments I synthesized
contained an average of 3.75 TBP, 2.25 Abf1, 0.75 Reb1, 0.25 Rap1, and 3.25 Rsc3 consensus binding sites each, which is only slightly more than the number expected in completely random sequences of the same length (2.56, 0.99, 0.24, 0.24, and 1.2 binding sites, respectively). I did not intentionally add CPA sites because they appear very frequently in randomly generated DNA sequence. I integrated these constructs into the genome and assayed expression using a custom tiling array.

I designed the four 3 kb fragments that comprised the four 6 kb randomly-designed loci by randomly selecting bases following the \textit{S. cerevisiae} base content (A=T=0.31) and occasionally inserting sites for Abf1, Reb1, Rap1, Rsc3, and TBP. The motifs added were sampled from the position frequency matrix (PFM) for these factors. For instance, if the factor’s PFM includes an A at the first position 99% of the time then the inserted motifs included an A at the corresponding position 99% of the time. I chose four fragments for gene-synthesis that when combined were predicted to form several “transcripts”. I assembled the four fragments into the four 6 kb combinations and integrated these into the yeast genome with a KanMX selectable marker. I confirmed the integration by PCR across the integration junctions. I grew these four yeast strains in YPD to an OD of 1.0, at which point I collected mRNA using the hot acid phenol method. I treated the total RNA with DNase I, purified it using an RNeasy mini kit (Qiagen), and isolated the mRNA using a NucleoTrap mRNA mini kit (Clontech). I isolated genomic DNA (gDNA) separately using a YeaStar Genomic DNA kit (Zymo Research) and subsequently sheared it by sonication to an average size of approximately 250 bp. I labelled RNA and DNA samples with G-coupled Cy3 and Cy5 dyes, respectively (Createch). Following dye-coupling, I partly hydrolyzed mRNA samples to an average size of about 250 bp using a NEBNext Magnesium RNA Fragmentation Module (NEB). I performed the microarray hybridization as described elsewhere [557], but with the hydrolyzed mRNA in place of cDNA.

The probes on the tiling array were 60 bp long and spanned the region in 1 bp
increments, with each probe present on the array at least four times. I normalized the tiling array data using a background model [206] which relates, for each probe \((k)\), the signal intensity \((y_k)\) to the background signal \((B_k)\) plus the amount of nucleic acid in the sample \((x_k)\) times the proportionality factor \((a_k)\), or \(y_k = B_k + x_k a_k\). Here, I wanted to determine the amount of nucleic acid in the sample \((x_k)\). I modified this normalization protocol to take advantage of the fact that the target sequences of many probes were absent from some samples (e.g. the A1B1 probe target sequences are absent from the A2B2 strain). This allowed me to estimated separate mRNA and gDNA background signals \((B_k)\) for each probe by taking the median signal intensity of the probe in strains lacking the target DNA sequences. I estimated the proportionality factor for each probe \((a_k)\), corresponding to how probe intensity changes with the amount of nucleic acid, by comparing the background-normalized gDNA signal at each probe to the actual amount of gDNA present in the sample, where the target DNA was present in the sample. I estimated the amount of mRNA at each probe using this model, and scaled the amounts to the median intensity across the KanMX gene to make mRNA levels comparable across constructs.

The expression data shown in Figure 4.11 illustrate that all four fragments produce a diverse set of transcripts whose expression levels span three orders of magnitude and have a similar range of expression levels as real genes (the 3’ end of SSB1, the 70th most highly expressed gene [314], can be seen on the left hand sides of the panels in Figure 4.11). The UM predicts the majority of transcript species correctly. There are several instances of promoters that initiate transcripts bidirectionally \((A1-a/A1-b, B1-a/B1-b, B2-b/B2-c)\), as well as numerous instances of convergent transcripts that appear to terminate at the same bidirectional CPA site \((B1-b/B1-c, B2-a/B2-b, A2-c/B1-a, B2-c/kan-R)\). Of the 12 easily distinguished transcripts, nine are robustly predicted by the UM \((A1-a, A1-b, A2-b, B1-a, B1-b, B2-a, B2-b, B2-c, kan-R)\). There are several examples where transcripts are generated on both DNA strands. These cases are difficult for the UM to
Figure 4.11: Predicted transcript structure and measured expression of the four randomly-generated 6 kb fragments, including (A) A1B2, (B) A2B1, (C) A1B1, and (D) A2B2. Tracks as in Figure 4.7, except “Expression” was measured using custom Agilent tiling arrays.
predict because it has no states representing transcripts on both strands; nonetheless, in both instances of overlapping transcripts ($A2-a/A2-b$ and $kan$ and transcripts generated antisense to $kan$), the model is uncertain and partially predicts transcripts on both strands. In other cases, it is uncertain which strand is transcribed when the data support the transcription of only one strand under the condition tested ($A1-c$, $B1-a$, $B2-c$). Of the easily distinguished transcripts, my model achieves a precision and a recall of 0.69, on a transcript-by-transcript basis, allowing a 200-base offset for TSS and CPA sites. Because the model is incapable of simultaneously predicting both of the overlapping transcripts $A2-a$ and $A2-b$, the theoretical maximum is a precision and recall of 0.92. By simulating random guessing for the positions and strands of transcripts, but assuming the number and sizes of transcripts are known, the upper bound of what is expected by chance is a precision and recall of about 0.029. Altogether, these results demonstrate that transcripts can arise from pseudo-random sequence at a predictable frequency.

### 4.6.3 Prevalence of bidirectional CPA sites

Because the termination model predicted that terminator elements are frequently bidirectional, I wanted to ask how frequently convergent transcripts use the same terminator. I analyzed ORFs on chromosomes 1-8 that are arranged in the convergent orientation and considered only those that are separated by at most 750 bp between their stop codons (592/647 convergently arranged ORFs). Of these, I then counted the number of convergent pairs that the UM predicts have at least one base in the bidirectional CPA state ($P(\text{CPA+/−})>50\%$), of which there were 240 (40.54%). Of these, the UM predicts that the majority (68.33%) overlap by at least 100 bp.
4.7 Discussion

The modelling and experimental validation presented here indicate that the majority of mRNA transcript definition in \textit{S. cerevisiae} is relatively simple, and that transcription occurs as a probabilistic process that can be faithfully captured in a hidden Markov model. Both the models and the experimental data described here indicate that promoters are generally defined by sequences that inhibit nucleosome formation, either inherently - through G/C-content and poly-A sequences - or using binding sites for a small number of chromatin modifying TFs, mainly GRFs. Most of the features originally tested in the initiation model correspond to motifs for other TFs, and so the slight loss in the initiation classifier’s performance when reducing the number of features (Figure 4.1B and C) may be explained by less frequent utilization of other mechanisms, such as the ability of arrays of TFBSs to exclude nucleosomes and promote transcription [2, 354, 389]. However, the fact that a linear model performed as well as Random Forests suggests that the major promoter-defining elements do not work cooperatively.

4.7.1 Interactions between features in \textit{cis}

Importantly, my model captures the interactions between these features in \textit{cis}: the fact that the UM is more accurate at identifying promoters than the initiation classifier alone indicates that promoter selection can be influenced by cleavage site usage and the relative arrangement of these elements on the chromosome (Figure 4.9B). One possible mechanism for this influence is through gene looping, which is a physical connection between the 5’ and 3’ ends of genes dependent on proper 3’ end formation [478, 12]. Following pioneering rounds of transcription from all promoter-like sequences, successful CPA events may reinforce the “correct” transcript choice (Figure 4.12). Previous studies have also established that transcription can influence the function of nearby elements by transcriptional interference [348] and repressive changes in the chromatin environment of the
downstream promoter [241]. These mechanisms would force promoters to compete with one another in cis, with factors such as initiation frequency, epigenetic state, and gene loop formation determining which promoter becomes dominant (Figure 4.12). Given this model, I would expect that, under conditions where an upstream promoter is inactivated, downstream promoters could become active as repressive transcription from the upstream promoter stops. Indeed, I can find many potential examples of this phenomenon (Figure 4.13). Thus, while few TFBSs contributed to the initiation classifier, they could help determine which among competing promoters is dominant, since the relative activation levels of promoters near each other can influence which transcripts are produced in the neighbourhood. Such a mechanism may also explain why it is beneficial to incorporate expression levels in training the UM parameters: their inclusion indirectly informs the model about the global transcriptional state of the cell.

### 4.7.2 Promoter and terminator directionality

The initiation classifier explains the presence of many non-coding transcripts. Previously reported cryptic transcripts, including CUTs, SUTs, and antisense transcripts [538, 544], are all, on average, associated with promoter-like sequences at the 5’ end (Figure 4.14). Many of these transcripts initiate from bidirectional promoters [538, 367]. In my analysis, the correlation between the initiation classifier’s predictions for the top and bottom DNA strands is only about 0.5, suggesting that there is some asymmetry incorporated into the promoter itself; indeed, the distribution of the critical promoter-defining features is only partly symmetric (Figure 4.1A). It is possible that either the arrangement or orientation of binding sites for some of the other ~200 TFs in *S. cerevisiae* help to make initiation more efficient in one direction. The base content of the nascent transcript could also be a contributor to asymmetrical transcript initiation from bidirectional promoters. I found that the ratio of As to Ts is important in the first 100 transcribed bases, which could help to control the efficiency of PolIII promoter release, since elongation rate can be affected
Figure 4.12: Gene definition model. (A) In the absence of transcription, the DNA forms nucleosomes except where prevented by bound TFs (such as the GRFs) or by the DNA structure. (B) Transcription begins indiscriminately from nucleosome free regions in proportion to the efficiency of pre-initiation complex (PIC) formation. Promoters compete with one another in cis through the act of transcription. (C) An equilibrium is reached where some promoters are active and others are repressed. Successful cleavage and polyadenylation reinforces the promoter choice. (D) If a nucleosome free region is destroyed (for instance, through loss of GRF binding), it is no longer competent for initiating transcription. Downstream promoters are then de-repressed, become active, and a new equilibrium is reached.
Figure 4.13: Context dependence of promoter identity. Several examples of promoters occurring within gene bodies that appear to be repressed by normal transcription of the gene and de-repressed when the gene is no longer expressed in a trans-acting factor mutant, for (A) Rsc3, (B) Abf1, and (C) Rap1 mutants [20]. In each case, three tracks are shown for both strands, ORFs in dark green, expression level, shown relative to wild type, in black, and the initiation score in green. The blue, red, and purple bars in the centre represent the binding sites for Rsc3, Abf1, and Rap1 (respectively) in the promoter regions of each gene. De-repressed transcripts are boxed.
Chapter 4. A unified model for yeast transcript definition

by nucleotide content [342].

![Graph A: CUTs](image)

![Graph B: SUTs](image)

![Graph C: Antisense Transcripts](image)

Figure 4.14: Predictions of models across non-traditional transcripts. Each graph shows the two classifier scores, as well as the UM predictions averaged across the aligned transcripts for (A) CUTs and (B) SUTs [538], and (C) antisense transcripts [544].

The termination classifier is both remarkably accurate (Figure 4.1B) and very simple, depending on only base content and Hrp1 sites. Historically, the Hrp1 binding site was thought to control the efficiency of downstream cleavage, while selection of the site of cleavage was determined by the positioning element [175]. My model is partly consistent
with this view, and further suggests that the general location of cleavage is determined primarily by Hrp1-binding, with a minor but significant contribution by A/T-content to cleavage site identity. The palindromic nature of the features favoured by the termination classifier strongly supports the bidirectional nature of yeast cleavage sites. It has previously been shown that some yeast terminators can stimulate cleavage and polyadenylation in either orientation [13, 119, 388] and that the 3' ends of convergent yeast genes frequently overlap [364, 378]. I provide evidence that the same cis elements are generally used to stimulate transcript termination in either orientation. In fact, my UM predicts that over 40% of convergent genes have overlapping terminator regions. Such a mechanism could have a role in preventing transcriptional interference between adjacent genes by minimizing overlapping transcription, and could also contribute to genome compaction.

4.7.3 Synthetic regulatory elements and the evolution of genes

Yeast is a prevalent system in synthetic biology (e.g. [254, 519, 524]). However, most of the promoter and terminator sequences currently in use are based on native sequences, and so have the potential to recombine [42]. I have demonstrated that these elements are relatively simple in nature, and so functional elements can easily be designed de novo. This simplicity also means that it is relatively straightforward for new promoters to originate in the yeast genome: fundamentally, all that is required is the gradual expansion of A/T-rich tracts and/or addition of binding sites for the five TFs that dominate the model. Indeed, promoter-like sequences will occur in randomly-generated DNA sequence at a rate of about 1/kb. Since a given promoter-sized sequence is likely to contain multiple TFBSs purely by chance (there are, on average, 27 perfect TFBSs per kilobase of random DNA drawn from the yeast base composition, using the YeTFaSCo expert-curated motifs [101]), then a newly-emerged promoter will likely already be regulated in some way. The fact that functional variants can arise frequently at random could explain
both the relatively high evolutionary rates in cis-regulatory sequence [249, 521] and de novo gene birth. Creating new genes from essentially random DNA was thought to be very unlikely [218], but several recent studies have shown that it may be more common than previously thought [71, 63]. Consistent with this, my UM predicts that 66% of bases in randomly-generated DNA sequence will be transcribed on one strand or another.

One model for the origin of new genes involves first generating a stable transcript, then acquiring coding potential, and finally acquiring a function [63]. However, until now, it was unclear how simple it is to generate stable transcriptional units from essentially random intergenic DNA.
Chapter 5

Conclusions

5.1 Summary

For this thesis, I describe the sequence features associated with transcriptional units, identify the factors which recognize them, describe how these affect transcription, and test the resulting models.

In Chapter 2, I described my work collecting the specificities for the complement of yeast TFs. I evaluated these motifs with available data to see which motifs best match the available evidence. On the YeTFaSCo website, I provide many tools and resources to aid in the study of transcriptional networks. These include genome-wide motif matches, custom sequence scanning with the motifs in the database, querying motifs in the database, and a tool to find potential regulators of gene sets. For me, the most important result of this work was the Expert Curated set of TF specificities. With this, I had a comprehensive and non-redundant motif set with which I could identify potential roles for each transcription factor. This set also reduced the feature space of computational models while still incorporating all potentially relevant motifs.

In Chapter 3, I described my work identifying the sequence elements that are associated with yeast transcripts de novo. This naive approach allows the discovery of
novel sequence elements, but has the disadvantage that not all sequence elements identified correspond to known features, which makes testing their function more difficult. I demonstrated that the approach I used is extremely effective at identifying sequence elements that are specific to the region of interest and that, using appropriate backgrounds, I can restrict the identified motifs to only those that are specific to the phenomenon of interest. For instance, I showed that many of the sequence biases associated with yeast transcripts can be accounted for given the overall G/C-content of the yeast genome and constraints of the protein coding sequence. I also identified a widespread enrichment of Nrd1 complex motifs and showed that functional TATA boxes can occur in unconventional places. Further, I found that the sequence elements located in promoter and terminator regions are strikingly similar, with many of the same sequence elements enriched in both.

In Chapter 4, I attempted to identify the *cis*-elements and corresponding *trans*-acting factors that function together to define transcript structure. I created classifiers that find transcription start sites and cleavage and polyadenylation sites using various cellular features, such as the TF specificities collected in the YeTFaSco database. The major features of the initiation classifier included binding sites for the GRFs and G/C-content, both of which are predicted to function by preventing the binding of nucleosomes, which generally do not bind promoters in the cell. The termination classifier relies on binding sites for the termination factor, Hrp1, and G/C-content, which I believe acts as a proxy for the binding sites of the other termination factors. I found that, for both classifiers, many of the defining sequence features are symmetrical, and, correspondingly, both promoters and terminators are predicted to function in either orientation. I confirmed that the *cis*-elements predicted to be sufficient to create a promoter are actually sufficient to initiate transcription by creating a synthetic promoter library and measuring the resulting expression levels. I also provided evidence that yeast terminators can function in either orientation. To model the entire process of transcription, I created a Unified Model that uses a hidden Markov model to predict gene structure with the classifier scores as input.
I found that the UM was robust at identifying transcriptional units and could identify promoters better than the initiation classifier alone. Finally, I integrated synthetic DNA sequences that were predicted to produce several transcripts into the yeast genome and confirmed that the transcripts that are generated correspond closely to those that are predicted.

5.2 Implications

In this section, I describe the conclusions of my work as they relate to the existing literature.

5.2.1 Gene definition

Overall, the picture painted by my modelling efforts of cellular gene definition is relatively simple: a nucleosome free region marks the promoter, the gene body is characterized by not containing terminators, and the terminator is characterized by low G/C-content and Hrp1 binding sites. Together, these elements are arranged along the chromosome in such a way as to define transcriptional units.

The promoter

The initiation classifier’s major features include GRF binding sites and G/C-content, both of which are believed to function through nucleosome depletion, either via GRF-mediated nucleosome depletion or by inherently excluding nucleosomes, respectively. There has been a large amount of work done previously characterizing the roles of the general transcription factors in yeast. The canonical model is that specific factors, such as the TATA binding protein, recognize sequences in the promoter and initiate transcription in a single direction accordingly. My model suggests that PICs will form in nucleosome free regions, more or less indiscriminately, and will form in either orientation. I believe
that these predictions can be reconciled with this past work by the fact that the TATA binding protein will bind to TATA-like sequences when no TATA box is available [412]; the yeast genome is so AT-rich that there are TATA-like sequences found throughout. The discordance between the traditional view that promoters initiate directionally and the bidirectionality now known to be commonplace [538] might be explained by the fact that many of the promoters originally studied in vitro contained TATA boxes, which can load TBP in a preferred orientation [454]. I think that the first stage of PIC assembly is binding of the GTFs to the most energetically favourable position within the NFR.

There are many genes in *S. cerevisiae* that lack a 5' NFR, but these are frequently not expressed under general conditions [292]. It has been demonstrated previously that these regions become nucleosome free upon activation of the gene [43, 272, 10], but it was unclear if the promoter becomes nucleosome free because it is transcribed, or because of the activation. My model suggests that, upon gene activation, transcription factors first evict nucleosomes from the promoter region, and this allows the PIC to form.

In general, I think that the 5' NFR is a result of continual competition among TFs, the GTFs, and nucleosomes. When a nucleosome is evicted, TFs bind and prevent nucleosomes from re-forming. For low G/C-content promoters, nucleosome eviction occurs because of the increased rate of nucleosome dissociation resulting from energetically unfavourable histone-DNA interactions. A PIC can then form within the NFR, and this further helps to exclude the nucleosome. At this point, TFs and GTF-DNA interactions can act to stabilize the PIC, increasing the PIC-promoter residency time. This is consistent with recent experiments showing that TATA box-containing promoters have relatively high burst sizes but low burst frequency [554, 202]; these promoters generally do not inherently exclude nucleosomes, but when a PIC can form, it is more stable.

An open question remains as to why yeast employ two different mechanisms (TF binding sites and nucleosome-refractory DNA-sequence) to achieve the same result (nucleosome depletion). An obvious advantage of TF-mediated nucleosome depletion is
that the target genes can be coordinately regulated, whereas inherently nucleosome free promoters are constitutively active. Given this, we would expect that certain cellular processes are enriched among the different GRFs. Indeed, using GO terms and predicted binding sites, Abf1 enriches for “RNA binding”, Rap1 enriches for “ribosome”, Rsc3 enriches for “plasma membrane”, and Reb1 enriches for “cytoskeleton organization” [101]. This may have the advantage that, in times of stress or a sudden shift in environment, large groups of genes can quickly be shut off without the need for a general downregulation of the transcriptional machinery. In this way, the cell can more quickly bounce back upon a shift to a more favourable environment.

The gene body

Although the elongation model failed to aid in the identification of gene structures using the HMM, it was generally successful in distinguishing sequences that are part of transcripts from those that are not. However, the primary features of this classifier relied on strand-specific base content. As I show in Chapter 3, this is primarily a result of the presence of coding sequences, and so the primary features of the elongation classifier reflected the coding sequence, not the transcript structure.

Another reason the elongation classifier failed was because the negative examples were so ill-defined. Positive examples included anything encompassing a stable transcript. Identifying regions that are not part of a transcript is much more difficult because it is unclear whether the sequence is not efficiently transcribed, lacks a promoter preceding it, or lacks a terminator following it. In the latter two scenarios, the negative example could just as easily be a positive example if it happened to be transcribed.

The terminator

My termination model was both very accurate and remarkably simple. It relied solely upon Hrp1 binding sites followed by an area of low-G/C-content. This more or less cor-
responds to what was already known about yeast terminators, except that there were supposed to be U-rich and A-rich regions bound by the other termination factors. Although I included motifs to represent this in the model, the model preferred G/C-content. I think that more accurate representations of the specificities of the other termination factors would have resulted in a better model, but these specificities are not currently available.

I found that Hrp1 binding sites, because of their repetitive nature, often occur in arrays, but are unlikely to bind Hrp1 cooperatively. These sites also appear to double as TATA boxes in some cases and can therefore result in the initiation of antisense transcripts. These repetitive motifs also mean that Hrp1 binding sites are present in either orientation. My model predicts, and I showed, both with specific examples and genome-wide studies, that these can result in terminators that stimulate cleavage and polyadenylation in either orientation.

5.2.2 Antisense transcription and gene looping

I found that many of the same sequence elements are found at both the beginnings and the ends of transcripts, including poly-As, poly-Ts, and AT-repeats. Indeed, the 5’ and 3’ gene ends are similar enough that the UM often gets the boundaries of a transcript correct, but predicts transcription on the wrong strand. AT-repeats are assumed to be used as Hrp1 binding sites at the 3’ end (and the enrichment of the alternate Hrp1 binding sites in this region supports this), while on the 5’ end these can correspond to TATA boxes. However, in certain cases, I showed that the TA-repeats in 3’ UTRs can be used as TATA boxes to drive antisense transcription. It is possible that the similarity in the sequences at the 5’ and 3’ ends facilitates the formation of gene loops, perhaps by having similar factors bind both NFRs. Gene loop maintenance could require a 3’ NFR, for instance, if gene-looping factors bind the DNA directly, but, to my knowledge, this has not been tested.
Although not requiring gene looping, antisense transcription could be facilitated by the gene loop structure. With 5' and 3' NFRs held together in space, it would not be surprising if, occasionally, PICs formed on the wrong NFR. This could potentially result in switch-like behaviour, oscillating between sense and antisense transcripts depending on which NFR the PIC formed on. This could have the potential evolutionary advantage of priming a gene for activation and having the gene loop structure already set up before the gene is activated.

Since I demonstrated that Hrp1 binding sites could be used as TATA boxes, it would be interesting to ask if TATA boxes could also be used as Hrp1 binding sites. If true, antisense transcripts could use the sense terminator to initiate and the sense promoter to terminate. This could have the advantage of facilitating the formation of sense-antisense switches by reusing the same cis-elements for both transcripts. TATA boxes acting as functional CPA sites could also act to prevent transcription across promoters by inducing cleavage and polyadenylation of upstream transcripts.

### 5.2.3 The role of the Nrd1 complex

Given that many of the enriched motifs within ORFs and at 5' and 3' UTRs correspond to binding sites for the Nrd1 complex, I believe these factors are likely to be of widespread importance in gene definition. Although these factors were included in the elongation classifier and were not found to significantly aid in transcript identification, it does not follow that these factors are not involved in the process; indeed, their known role in CUT degradation, as well as the distribution and strong enrichment of their binding sites, suggests they are. Both Nrd1 and Nab3 are essential proteins [161], which indicates that they are important to cellular function. However, this complex plays multiple roles in the cell, including regulating the expression level of mRNAs [357], affecting snoRNA maturation [464], and causing the degradation of CUTs [15], and it is unclear which of these roles are essential.
5.3 Future directions

Many avenues for future work are here laid out. The modelling of gene definition could be improved and extended by incorporating additional known features of transcription, but a major current limitation is the lack of available RBP specificities. Ultimately, some of the hypotheses of the model must be tested, and the modes of action of the factors involved must be elucidated.

5.3.1 GRF function

My initiation classifier suggests an important role for the GRFs. Although the GRFs are very well studied, the mechanism by which they evict nucleosomes is elusive. Several studies have shown physical interactions between the GRFs and chromatin remodelling complexes (e.g. RSC [151, 152], SWI/SNF [489]), and Rsc3 is itself a member of the RSC chromatin remodelling complex. However, they also interact with histone modifiers (e.g. histone acetyltransferases [310, 231, 265]), the histone proteins themselves [151, 276, 162, 152, 78], and numerous other proteins. It is possible that nucleosomes are evicted by virtue of the GRF out-competing the nucleosome for binding the DNA. Because GRFs tend to be highly abundant proteins [158], and overexpression of a TF can help it compete with nucleosomes [548], this may be the primary method by which GRFs function. By studying nucleosome occupancy across ascomycota, one group was able to show that the GRFs that are used by the cell change over evolutionary time [494]. If an increase in abundance is all that is needed to convert a TF into a GRF, then we would expect them to turn over more frequently than if they have to evolve interactions with chromatin remodellers.
5.3.2 Terminator-promoter feedback

The UM suggested that knowledge of where a transcript could terminate could feed back into promoter selection. It is not clear how this could be accomplished, but I proposed that it was due to gene looping, where the 5’ and 3’ ends of a gene are physically adjacent in 3D space. Although this phenomenon has been well-studied for specific examples, and several roles for it have been proposed (including promoter directionality [478, 338], and transcriptional memory [286]), it is unclear how prevalent it is across the yeast genome. My model suggests this phenomenon is more widespread than previously appreciated and may play a role in promoter identity. If this is the case, we would expect that there might be widespread changes in where transcripts initiate in mutants causing a disruption of gene loops. Indeed, TFIIB mutations cause both gene looping [477] and TSS selection defects [394, 383, 127], although the latter is not thought to result from the former. The common sequence elements I found at the 5′ and 3′ ends could also play a role here. Perhaps similar structures are built at the 5′ and 3′ ends of genes, these structures are physically tethered, and the transcription machinery must then choose which of the two possible NFRs to initiate from, as well as which of the two possible directions to take. As experimental data becomes available, it may soon be possible to answer these questions.

5.3.3 Promoter directionality

How the transcription machinery chooses a direction to transcribe is another unexplained phenomenon. My initiation classifier was only partly symmetric, with the majority of the features dedicated to NFR formation and located symmetrically in the -150:-80 bin. The TATA box is known to be an asymmetric feature, where the motif preferentially occurs in one direction and the orientation of TBP binding determines the direction of transcription [454]. However, most promoters lack a canonical TATA box and, instead, TBP appears to bind whatever TATA-like sequence it can find in the promoter region [412]. Therefore, there must be other sequence signals polarizing transcription in promoter re-
regions. As described above, the presence of a terminator could help, but I think this is more likely (at least upon initial activation) an equilibrium problem, where the combinations, positions, activities, and orientations of the transcription factors and GTFs present in the promoter region push PIC formation in one orientation or the other. Nonetheless, this remains a largely unexplored problem and, with current high-throughput promoter assaying technology (e.g. [445]), could soon become a solved problem.

5.3.4 Determinants of nucleosome binding

Another open question, and one that has been extensively studied, is the sequence determinants of nucleosome binding in the absence of other factors (i.e. the inherent specificity of nucleosomes). Many models have been proposed (e.g. [242, 488, 131]), but none fully explain the observed nucleosome binding \textit{in vitro}. Further, the mechanisms for nucleosome exclusion by certain elements remain unresolved. For instance, poly-dA:dT tracts theoretically exclude nucleosomes by virtue of their intrinsically rigid structures [440]; however, this does not explain why ATATA [242] and AATAA (Figure 3.4) are similarly depleted. As DNA synthesis and sequencing become cheaper, and sequencing reads become longer, it will soon be feasible to test vast numbers of synthetic DNA sequences for nucleosome binding to produce more precise models of nucleosome-DNA interactions. A solid understanding of the determinants of nucleosome binding is fundamental to our interpretation of the genome.

5.3.5 Factor specificities

As of version 1.02 of the YeTFaSCo database, there are still five transcription factors with known DNA-binding domains that are likely to be sequence specific, but have as yet no known motifs available. There are a further 37 low-confidence motifs for 34 TFs that may not represent the true specificities of these factors. In addition, the representation of the motifs currently included in the YeTFaSCo database may not faithfully represent the true
specificities of the factors. As new methods to derive specificities are applied to yeast TFs, YeTFaSCo will need to be updated. Currently, the gold standard representation of a motif is a PWM; however, other methods are available. If a new specificity representation becomes popular, it may become necessary to update the database to include these new representations. The evaluation of the motifs may also have to be changed, since they presently rely on PWM representations.

There are also many yeast RNA-binding proteins with unknown specificities. In a recent publication, other members of the Hughes lab attempted to characterize the binding preferences of several yeast RBPs [408]. Of the fourteen proteins tested, only two yielded a reliable specificity (Hrp1 and Vts1). In addition to these, there are many known and putative RBPs with uncharacterized or under-characterized specificities. Having a complete library of RBP specificities in yeast would allow a more complete modelling of transcriptional processes.

Particularly important examples of under-characterized specificities are those of the Nrd1 complex; the exact sequences bound by these proteins is still an open question [69, 397, 527, 97, 199], and the specificities we do have for these factors reflect their indirect inference from \textit{in vivo} data. Having a motif that accurately reflects the binding specificity of Nrd1 and Nab3 (both separately and in complex) would be extremely beneficial to modelling RNA-recognition by this complex.

### 5.3.6 Unexplained motifs

In Chapter 3, I found many enriched sequences with both known and unknown function. Although these are unlikely to be functional, I was able to explain why they were so enriched. I found many more motifs to which I was unable to assign a function. For instance, when analyzing introns, I found many motifs that were associated with genes located within \textit{Y'} elements or ribosomal genes. It is possible that a portion of the other unexplained motifs are artifacts similar to the \textit{Y'} element and ribosomal motifs, but I
was also unable to explain them as such. Although the yeast transcription factors are generally well-characterized, the RNA-binding proteins are much less so. Given that many of the motifs I found could be explained as binding sites for RNA-binding proteins, it seems likely that some of the other unexplained motifs are also binding sites for RBPs.

In particular, I was only partly able to explain the function of the TTCTT/AAGAA motifs located in the 5’ and 3’ UTRs. As described (Sections 3.5.1 and 3.7.1), these motifs can partly be explained as binding sites for Nab3. However, the motifs found antisense to the coding sequence more closely resembled the known specificity of Nab3 and also included binding sites for its partner, Nrd1 (Section 3.6.2). Why were the alternative Nab3 binding sites not also found in the UTRs? Why were Nrd1 binding sites not more prevalent there? I see two possible explanations: either the TTCTT motif is multifunctional - serving as both a Nab3 binding site and as a binding site for some other factor or else Nab3 plays different roles in UTRs and coding sequences.

5.3.7 Improvements to the UM

The biggest shortcoming of the UM, in my opinion, is the Markov property of the HMM underpinning it. The “Markov” (or memoryless) property is that the model only knows what state it is currently in and does not remember where it came from or how long it has been in that state. The fact that the PolIII CTD has distinct states across the body of a gene [487, 347, 268] suggests that transcription is not a Markov process. Further, it is known that the different termination pathways depend on the phosphorylation state of the CTD, where Nrd1-dependent termination preferentially occurs with a Ser5-P CTD [279], whereas CPA-dependent termination occurs with a Ser2-P CTD [308, 3]. This suggests that the probability of transitioning from the elongation state to the termination state changes as more DNA is transcribed, which could be captured in a hidden semi-Markov model. Each of the seven amino acids of the CTD repeat consensus have at least two different states [191], resulting more than $2^7 = 128$ possible CTD states for a single
repeat; accounting for combinations of repeats would yield many more, meaning that PolIII potentially has a very complex built in memory.

In training the parameters of the UM, I used binary state annotations to derive the transition and observation probabilities and used a range of bases for TSSs and CPA sites. In reality, the cell produces many transcript species with different probabilities (e.g. [386]), and each transcript uses exactly one TSS and CPA site. I believe that the model could be substantially improved if the parameters were derived in a probabilistic, rather than an absolute, fashion. Similarly to how I included conflicting transcripts for different strands (Section 4.5.1), parameters could be derived by repeatedly encompassing the same segment of DNA into the parameter estimation, each time with a different isoform, weighted with the probability of that isoform occurring. This would also have the desirable effect of producing a greater precision for TSS/CPA site identification. The initiator and cleavage site motifs could be included as an additional input for the HMM to further increase precision.

A more probabilistic model would also benefit from a prediction of expression level (or, more precisely, initiation frequency). Because transcripts can interfere with the transcription of one another (e.g. [203, 55]), the expression levels of the competing promoters are likely to come into play in determining which promoter becomes dominant. In the UM, expression level is not accounted for at all, and is more or less treated as binary. In order to incorporate expression level into such a model, it would probably be easiest to modify the initiation classifier to, instead of predicting whether or not a sequence is a promoter, predict the frequency with which a given sequence will generate transcripts, where non-promoters would have an initiation frequency of about zero. The data necessary to produce such a model is not currently available, but high-throughput technologies for assaying expression levels, such as the work described here (Section 4.4.1) and elsewhere [445], will soon provide the data required for such an endeavour. However, incorporation of expression level necessarily requires an incorporation of the growth environment, since
the former depends on the latter. This may be more challenging, as it either requires a separate model for every growth condition, or a more mechanistic model that can take into account transcription factor activities and concentrations. Ultimately, I think the latter approach will be required, since there are essentially infinitely many possible growth conditions.

I think the initiation model would also benefit from incorporation of a feature capturing the propensity for arrays of different transcription factor binding sites to work together to exclude nucleosomes. It is known that multiple TFs competing with the same nucleosome can have a cooperative effect on binding, without any direct interaction [389, 505, 2]. If our model seeks to identify promoters, promoters are defined by the presence of an NFR, and arrays of different TFs can generate an NFR, then our model must necessarily take this into account, even ignoring expression levels. However, this too requires a knowledge of nuclear TF concentrations, an environment specific variable with only limited data available (e.g. [369]).

The fact that my elongation classifier failed does not mean that such a model would not be useful in identifying transcript structure. Indeed, there are many reasons to believe that it would be useful. The existence of CUTs, for instance, was one of the primary motivations for including this model. However, it may be more useful, instead, to build a CUT termination classifier, which distinguishes transcripts from CUT (Nrd1-dependent) terminators. In this case, we could also include a CUT-termination state in the HMM. With the emergence of numerous RNA-Seq datasets (e.g. [433]), such a classifier may soon be possible.

Another potentially useful “elongation” classifier would predict polymerase pausing. We already have excellent datasets identifying pause sites across transcripts [89, 96] and, by normalizing to the total amount of transcription, we can get a % pausing at every base across different transcripts. Using this data, we can create a model that predicts pausing based on DNA sequence. This could help to determine the efficiency of Nrd1
or CPA-dependent termination, since the elongation rate appears to affect things like splicing and termination [102, 395].

Although my termination classifier was extremely accurate at identifying cleavage sites, I believe that it could be substantially improved. With current technology, it would be feasible to test a vast library of millions of cleavage sites. For instance, by cloning the sites being tested downstream of a common promoter, growing the library in bulk, and isolating and sequencing the poly-A+ RNA as well as the plasmid library, we would get a plethora of examples of mRNA cleavage sites and their termination efficiencies. From this, a more mechanistic termination model could be made that outputs a probability of mRNA cleavage for every base. Here too, however, more accurate representations of the RBPs involved in termination are required.

5.4 Summary

Although my work has contributed significantly to our knowledge of gene regulation in yeast, much work lies ahead. We can explain many of the sequence elements associated with yeast transcripts, but many still go unexplained. The UM can predict the structure of many genes, but does not achieve the same resolution as the cell and could potentially be improved by incorporating transcriptional mechanisms known to be employed by the cell, such as CTD modification. Having a more complete inventory of nucleic acid binding proteins and their specificities will aid in these endeavours, first, by facilitating modelling efforts, and, second, by providing potential trans-acting factors to test when new sequence elements are found. Finally, the exact mechanisms of action of the factors involved need to be characterized.
Bibliography


[351] S. Medler, N. Al Husini, S. Raghunayakula, B. Mukundan, A. Aldea, and A. Ansari. Evidence for a complex of transcription factor IIB with poly(A) polymerase and


[398] P. Preker, K. Alnvig, M. S. Christensen, E. Valen, C. K. Mapendano, A. Sandelin,


[516] T. Watanabe, K. Hayashi, A. Tanaka, T. Furumoto, F. Hanaoka, and Y. Ohkuma. The carboxy terminus of the small subunit of TFIIE regulates the transition from


