A Computational and Evolutionary Approach to Understanding Cryptic Unstable Transcripts in Yeast

By Jessica M. Vera

B.S. University of Wisconsin-Madison, 2007

A thesis submitted to the Faculty of the Graduate School in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Molecular, Cellular, and Developmental Biology

2015

This thesis entitled:

A Computational and Evolutionary Approach to Understanding Cryptic Unstable Transcripts in Yeast

written by Jessica M. Vera

has been approved for the Department of Molecular, Cellular, and Developmental Biology

_____

Tom Blumenthal

_____

Robin Dowell

*Date*_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline

Vera, Jessica M. (Ph.D., Molecular, Cellular and Developmental Biology)

A Computational and Evolutionary Approach to Understanding Cryptic Unstable Transcripts in Yeast

Thesis Directed by Robin Dowell

Cryptic unstable transcripts (CUTs) are a largely unexplored class of nuclear exosome degraded, non-coding RNAs in budding yeast. It is highly debated whether CUT transcription has a functional role in the cell or whether CUTs represent noise in the yeast transcriptome. I sought to ascertain the extent of conserved CUT expression across a variety of *Saccharomyces* yeast strains to further understand and characterize the nature of CUT expression. To this end I designed a Hidden Markov Model (HMM) to analyze strand-specific RNA sequencing data from nuclear exosome *rrp6*Δ mutants to identify and compare CUTs in four different yeast strains: S288c, Σ1278b, JAY291 (*S.cerevisiae*) and N17 (*S.paradoxus*). My RNA-seq based method has greatly expanded upon previous CUT annotations in *S.cerevisiae,* underscoring the extensive and pervasive nature of unstable transcription. Utilizing a four-way genomic alignment I identified a large population of CUTs with conserved syntenic expression across all four strains. Furthermore I observed that certain configurations of gene-CUT pairs, where CUT expression originates from a gene 5' or 3' nucleosome free region, correlate with distinct expression trends for the associated gene. Bidirectional gene-CUT pairs correlate with higher expression of genes, and antisense gene-CUT pairs correlate with reduced gene expression. Interestingly these effects on gene expression are most prevalent in the presence of conserved CUT expression. Additionally I have shown that CUTs lack a well-defined 3' nucleosome free region that is commonly observed at protein-coding genes, and suggests that 3' NFRs are not characteristic of Sen1-dependent terminated transcripts.

Table of Contents

Tables

Figures

Chapter I - Introduction

Cryptic unstable transcripts, or CUTs, are one of a plethora of unstable non-coding RNAs (ncRNA) identified in yeast in the past decade. At the time that CUTs were first globally identified in nuclear exosome *rrp6*Δ mutants (Wyers et al. 2005; Davis and Ares 2006), a complete understanding of the molecular components and pathways leading to CUT degradation were lacking. However there were preliminary indications that CUTs engaged many of the same complexes involved in small nuclear and small nucleolar RNA production (Allmang et al. 1999; Kadaba, Wang, and Robinson 2006). Yet how these components worked together or how and why they acted on CUTs and other RNAs remained unclear. We now know that an alternative RNA polymerase (RNAP) II transcription termination and 3' end processing pathway recognizes CUTs and other RNAPII transcripts, marking these RNAs for degradation or processing into mature transcripts by the nuclear exosome (Eric J. Steinmetz et al. 2001; Wyers et al. 2005; Arigo, Eyler, et al. 2006; LaCava et al. 2005; Kim et al. 2006; Thiebaut et al. 2006).

For what purpose are CUTs transcribed if only to be rapidly degraded by the nuclear exosome? I, like other researchers in the field, have hypothesized that CUT transcription is able to regulate the expression of nearby and overlapping genes. The central focus of the my thesis work has been to assess the extent of conserved syntenic CUT expression across a variety of yeast strains to elucidate potentially important, conserved functional roles for CUT transcription in yeast. To best appreciate and contextualize the findings discussed in this thesis I will first introduce the reader to the various ncRNAs thus far identified in the budding yeast *Saccharomyces cerevisiae* and the basics of RNAPII transcription, including transcription termination and 3' end processing pathways. Then I will discuss in detail the molecular complexes and pathways resulting in CUT degradation by the nuclear exosome and the implications of such pathways for the yeast transcriptome.

Yeast ncRNAs

Numerous transcriptome studies have shown that the yeast genome is highly expressed, revealing pervasive transcription of intergenic and unannotated, non-protein coding regions (David et al. 2006; Nagalakshmi et al. 2008; Churchman and Weissman 2011). RNAPII is responsible for this pervasive transcription, producing the various intergenic and intragenic ncRNAs observed in these studies. Historically the term 'non-coding RNA' simply refers to any non-protein coding RNA, including the well characterized functional ncRNAs such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), etc. However in recent years 'non-coding RNA' has expanded to include non-protein coding RNAs, transcribed by RNAPII, whose functions are often unknown. A great deal of effort has been put into determining the functions of RNAPII ncRNAs. Well documented examples include ncRNAs that regulate IME4, SER3, and GAL10 expression (Martens, Laprade, and Winston 2004; Hongay et al. 2006; Houseley et al. 2008; Gelfand et al. 2011; Thebault et al. 2011)

In yeast recently discovered ncRNAs are commonly distinguished from one another by their stability or persistence in the cell after transcription termination. One group of stable ncRNAs are the so-called stable unannotated transcripts (SUTs) (Xu et al. 2009), though many stable ncRNAs exist beyond those identified as SUTs. Unstable RNAs are not readily detectable in the steady-state RNA population and are generally defined by the molecular complexes or enzymes involved in rapidly removing these RNAs from the cell. Yeast unstable ncRNAs include cryptic unstable transcripts (CUTs; nuclear exosome, Rrp6p) (Xu et al. 2009; Wyers et al. 2005; Davis and Ares 2006), Xrn1-sensitive unstable transcripts (XUTs; Xrn1p) (van Dijk et al. 2011), Nrd1-dependent unterminated transcripts (NUTs; Nrd1p) (Schulz et al. 2013), meiotic unstable transcripts (MUTs, essentially CUTs specifically found during meiosis) (Frenk, Oxley, and Houseley 2014), and dis3Δ transcripts (nuclear exosome, Dis3p) (Gudipati et al. 2012). Similar to CUTs, unstable ncRNAs called promoter upstream transcripts (PROMPTs)

have been identified in human cells by transient knock down of nuclear exosome components (Preker et al. 2008; Preker et al. 2011). This menagerie of unstable transcripts not only greatly contributes to the complexity of the yeast transcriptome, but also raises many questions regarding the nature of such extensive, but short lived, ncRNAs. Research investigating CUTs, and other unstable ncRNAs, has raised many questions regarding the specificity and regulation of RNA polymerase activity and has also brought to light a complex network of RNA processing and degradation pathways in yeast.

The Basics of Transcription

Nuclear transcription in yeast is carried out by three multi-subunit RNAP complexes: I, II, and III which catalyze the polymerization of ribonucleoside triphosphates from a DNA template; each RNAP is responsible for the production of a distinct population, or populations, of RNA in the cell. RNAPI transcribes the 35S primary rRNA transcript that is processed into the 25S, 18S, and 5.8S mature rRNAs. In addition to the aforementioned ncRNAs described in the previous section (**Yeast ncRNAs**) RNAPII transcribes all nuclear, protein-coding RNAs (mRNAs), all small nucleolar RNAs (snoRNAs) except snR52, and the U1, U2, U4, and U5 small nuclear RNAs (snRNAs). RNAPIII transcribes the 5S rRNA, all nuclear tRNAs, the U6 snRNA, and the snR52 snoRNA. The structure and function of yeast RNAPI, II, and III are highly conserved throughout eukaryotes (Sentenac 1985; reviewed in Werner and Grohmann 2011; Huang 2001; Cramer et al. 2008) making yeast an excellent model to study eukaryotic transcription.

RNAPII is the most widely studied polymerase complex of the three, no doubt due to its role in mRNA production and its unique C-terminal domain (CTD) structure. Transcription by RNAPII can be broken into three phases: initiation, elongation and termination. These three steps of transcription are accompanied by several co-transcriptional RNA processing events

including capping, splicing, and 3' end formation that are coordinated by post-translational modifications of RNAP II CTD (reviewed in Eick and Geyer 2013). During initiation, RNAPII binds the promoter, located upstream of the gene, and begins transcription of a short RNA product. Once transcription begins a guanosine triphosphate cap is added to the 5' end of the nascent RNA, protecting it from 5'→3' degradation and aiding in translation efficiency. Elongation marks the start of processive polymerase activity where RNAPII moves along the DNA adding nucleotides to the 3' end of the growing nascent RNA. During elongation introns are removed from the pre-mRNA. Finally, in termination, RNAPII is either released or removed from the DNA thereby ending polymerization. Termination coincides with 3' end formation and can follow one of two distinct, though not mutually exclusive, pathways: poly(A)-dependent termination or Sen1-dependent termination[1] (Kim et al. 2006). These two pathways play a crucial role in determining the fate (i.e. stability) of RNAP II RNAs.

Alternative RNAP II Transcription Termination Pathways

In yeast, RNAPII-transcribed pre-mRNAs and stable ncRNAs undergo canonical poly(A)-dependent termination utilizing the cleavage and polyadenylation factor (CPF) as well as the cleavage and polyadenylation factor IA (CFIA) complexes. RNAPII-transcribed snRNAs, snoRNAs, and unstable ncRNAs, such as CUTs, undergo non-canonical, Sen1-dependent termination utilizing the Nrd1-Nab3-Sen1 (NNS) complex (Steinmetz et al. 2001; Kim et al. 2006; Thiebaut et al. 2006; reviewed in Bernstein and Toth 2012). Though each pathway utilizes a distinct set of termination and 3' processing complexes, several key termination factors, such as Pcf11p, Ssu72, and Ess1, are important to both pathways (Steinmetz and Brow 2003; Kim et al. 2006; Singh et al. 2009; Krishnamurthy et al. 2009). **Figure 1** highlights the key

[1] Sometimes referred to as poly(A)-independent termination or non-poly(A) termination

components involved in each pathway discussed below. I note that several mRNAs and stable ncRNAs show evidence of occasional Sen1-dependent termination (Schulz et al. 2013; Webb et al. 2014) demonstrating modest overlap or redundancy between these two pathways. It is thought that both sets of termination complexes associate with elongating RNAPII, via phosphorylated CTD serine residues, awaiting transcription of appropriate sequences in the nascent RNA. Both Pcf11p of the CFIA complex and Nrd1p of the NNS complex contain a CTD interacting domain that preferentially binds phosphorylated Serine2 (Ser2-P) and Serine5 (Ser5-P), respectively (Licatalosi et al. 2002; Vasiljeva et al. 2008; Kubicek et al. 2012), while Ctf1 of the CPF complex has a general affinity for phosphorylated forms of the CTD (Dichtl et al. 2002). As shown in **Figure 1** the prevalence of RNAPII CTD Ser5-P and Ser2-P are inversely correlated, with Ser5-P dominating early in transcription and Ser2-P dominating late in transcription (reviewed in Eick and Geyer 2013). The difference in CTD binding between Pcf11p and Nrd1p may provide additional specificity in determining which termination pathway is used. Additionally H3K4 trimethylation by Set1, which is enriched at promoter-proximal regions of highly transcribed genes, aids in Sen1-dependent termination (Terzi et al. 2011). The combined preference for Ser5-P and H3K4 trimethylation may explain why most Sen1-dependent terminated RNAs are relatively short[2].

Poly(A)-dependent termination occurs via recognition of an AU-rich polyadenylation signal[3] by the CPF complex causing a slowing of RNAPII and Ysh1-dependent endonucleolytic cleavage of the nascent RNA. Cleavage is followed by polyadenylation of the upstream cleavage product by Pap1p and degradation of the downstream cleavage product by the 5'→3'

---

[2] because Ser5-P occurs early in transcription and H3K4 is promoter proximal

[3] with the consensus motif: AAUAAA

exonuclease Rat1p/Xrn2p. RNAPII continues to transcribe the downstream cleavage product until it dissociates from the DNA template thereby ending transcription. The mechanisms governing transcription termination remain unclear, but popular models suggest that Rat1p/Xrn2p degradation of the downstream RNA collides with RNAP II (the torpedo model) and/or conformational changes in RNAP II, resulting from the loss of CTD-associated proteins after transcription of the polyadenylation signal, disrupt polymerase activity (the allosteric model) (Kim et al. 2004; Luo, Johnson, and Bentely 2006; Park, Kang, and Kim 2015; reviewed in Rosonina, Kaneko, and Manley 2006).

Non-canonical, Sen1-dependent termination occurs via recognition of short motifs (consensus GUAA|G and UCUU) in the nascent RNA by the RNA binding proteins Nrd1p and Nab3p (Steinmetz et al. 2001; Carroll et al. 2004; Hobor et al. 2011; Creamer et al. 2011). Interaction of the Nrd1-Nab3-Sen1 complex (NNS) with both the RNAPII CTD and the nascent RNA is thought to slow and possibly disrupt transcription. Sen1 helicase activity may further induce transcription termination, possibly by unwinding the DNA:RNA hybrid in the RNAPII active site (Steinmetz et al. 2006; Porrua and Libri 2013; Martin-Tumasz and Brow 2015). Sen1-dependent termination is thought to occur without cleavage of the nascent RNA or Rat1p/Xrn2p degradation (torpedo model) (Kim et al. 2006). Following 3' end formation, association with the NNS complex marks the RNA for polyadenylation by the non-canonical poly(A) polymerase Trf4p/Pap2p of the TRAMP complex (Tudek et al. 2014). It was recently shown that the Nrd1p CTD interacting domain, which binds RNAP II CTD Ser5-P, also binds a CTD mimic in Trf4p, there by facilitating substrate hand over to TRAMP (Tudek et al. 2014).

RNAs that undergo Sen1-dependent termination are polyadenylated by TRAMP and rapidly degraded or processed into mature transcripts by the nuclear exosome. The fact that many of the RNAs terminated by this pathway are rapidly degraded, i.e. CUTs, has led to the suggestion that this pathway acts as a surveillance or quality control mechanism, guarding the

yeast transcriptome from spurious transcription. Indeed, the NNS complex is essential, likely due in part to its role in sn/snoRNA production. Temperature sensitive nrd1 mutants and nuclear depletion of Nrd1p lead to defective NNS termination resulting in extended transcripts (NUTs) that often transcribe into neighboring genes (Schulz et al. 2013). Genes affected by these extended transcripts often showed aberrant expression patterns, consistent with models where Sen1-dependent termination acts to protect the genome from uncheck/spurious transcription by early termination of ncRNAs.

TRAMP (Trf4/Air2/Mtr4 polymerase complex)

The TRAMP complex is an activating cofactor of the nuclear exosome (Callahan and Butler 2010) made up of three proteins: a poly(A) polymerase (Trf4/5p), a zinc knuckle domain protein (Air1/2p), and a DExD/H family RNA helicase (Mtr4). Air1p and Air2p have largely redundant functions, acting to recognize and bind the RNA substrates of TRAMP (LaCava et al. 2005; Wyers et al. 2005; Schmidt et al. 2012). Although over expression of Trf5p can rescue defects in RNA 3' processing in *trf4*Δ, Trf4p and Trf5p primarily have non-redundant functions (LaCava et al. 2005; Egecioglu, Henras, and Chanfreau 2006; Houseley and Tollervey 2006), often leading to the designations TRAMP4 and TRAMP5 to distinguish between these two forms of the complex. Mtr4 RNA helicase is hypothesized to dissociate or remodel stable ribonucleoprotein structures both as a component of TRAMP and in TRAMP-independent processes (Holub et al. 2012).

The nuclear exosome is thought to be relatively inactive in the absence of TRAMP activity (Mitchell et al. 1997), likely to protect the cell from inappropriate or nonspecific RNA degradation. Both trf4Δ and *rrp6*Δ mutants show accumulation of CUTs and 3' extended sn/snoRNAs and rRNAs, with *trf4*Δ/*rrp6*Δ double mutants showing that Trf4 is epistatic to Rrp6 (Wyers et al. 2005; LaCava et al. 2005; Callahan and Butler 2010; Kadaba, Wang, and

Robinson 2006). In addition to Sen1-dependent terminated transcripts, TRAMP polyadenylates introns, rRNAs, and aberrant tRNAs targeting them for 3' end processing and maturation or complete degradation by the nuclear exosome (Wyers et al. 2005; Egecioglu, Henras, and Chanfreau 2006).

Nuclear Exosome

The exosome is a conserved, RNA-degrading complex composed of 3'→5' exoribonucleases that functions in RNA surveillance, turnover, and processing. Two forms of the exosome are present in yeast: cytoplasmic and nuclear. The cytoplasmic exosome is involved in general mRNA turnover, no-go decay, and non-stop decay while the nuclear exosome is involved in RNA 3' end processing, maturation, and degradation (Anderson and Parker 1998; Allmang et al. 1999; van Hoof, Lennertz, and Parker 2000; Kadaba, Wang, and Robinson 2006; reviewed in Butler and Mitchell 2010; Parker 2012). The nuclear exosome is also responsible for degrading various populations of unstable ncRNAs such as CUTs (Wyers et al. 2005; Davis and Ares 2006; Xu et al. 2009). Both exosome forms share a core of nine catalytically inactive subunits and one active subunit, Rrp44p/Dis3p[4], which contains both endo and exoribonuclease domains (Mitchell et al. 1997; Lebreton et al. 2008). The nuclear exosome additionally contains the active exonuclease Rrp6p, its binding partner Rrp47p, and the accessory protein Mpp6p. Mtr4p also directly interacts with the nuclear exosome, via Rrp6p-Rrp47p, in either a TRAMP dependent or independent manner (Schuch et al. 2014; Thoms et al. 2015). Mtr4p helicase activity likely serves to unfold highly structured RNAs to aid in processing or degradation by the nuclear exosome. A recent publication has also demonstrated

[4] Henceforth simply referred to as Dis3

that Nab3p can directly recruit Rrp6p to ncRNAs for processing or degradation independent of Nrdp1, providing combinatorial flexibility in RNA processing (Fasken, Laribee, and Corbett 2015).

Structure studies of the nuclear exosome reveal a tunnel-like structure composed of the nine core subunits flanked by Dis3p and Rrp6p at opposite ends of the tunnel (Aloy et al. 2002; Aloy et al. 2004; Liu, Greimann, and Lima 2006; Bonneau et al. 2009; Makino et al. 2015). A recent report demonstrates how Dis3p and Rrp6p active sites can be accessed by different paths depending on the nature of the RNA substrate (Makino et al. 2015). As shown in **Figure 2**, RNA substrates first bind Rrp6p-Rrp47p and stochastically reach the active site of Rrp6p depending on the conformation of Rrp6p Tyrosine361. Alternatively, the RNA substrate can be funneled along the central channel to the Dis3p active site leaving Rrp6p to stabilize RNA binding. It is hypothesized that in the absence of RNA binding by Rrp6p, Dis3p can be an open conformation, seen left in **Figure 2**, allowing for direct access to the Dis3p active site. This observation that Dis3p can be accessed independently of Rrp6p may explain why partially redundant populations of ncRNAs are stabilized in *rrp6*Δ and *dis3*Δ cells (Gudipati et al. 2012).

Rrp6 and its accessory factors Mpp6 and Rrp47 are the only viable exosome deletion mutants (Giaever et al. 2002). However loss of Rrp6 greatly abrogates nuclear exosome function, resulting in a slow growth phenotype, increased sensitivity to a number of stresses, and RNA accumulation (Briggs, Burkard, and Butler 1998; Giaever et al. 2002; Hieronymus, Yu, and Silver 2004; Wyers et al. 2005; Davis and Ares 2006; Stead et al. 2007). It's important to note that *rrp6*Δ does not completely inhibit exosome function, as the exosome is an essential

complex, suggesting that Dis3p can to some extent compensate for Rrp6[5]. Conversely, complete loss of Dis3 is inviable (Giaever et al. 2002). Point mutations, temperature sensitive mutants, and partial deletions have been used to characterize Dis3p functions (Dziembowski et al. 2007; Gudipati et al. 2012). While structural and biochemical studies are helping to understand substrate specificity of Rrp6p and Dis3p, we can also gain important insights by studying nuclear exosome targeted RNAs themselves.

Evolutionary Context of Sen1-dependent Termination

Sen1-dependent termination is an alternative transcription and 3' end processing pathway in the budding yeast *Saccharomyces cerevisiae.* Sen1-dependent termination may be ancestrally linked to an analogous termination pathway recently characterized in *E.coli* that utilizes the Rho, NusG, and H-NS proteins (Peters et al. 2012). Intriguingly, Nrd1p has been shown to regulate NRD1 expression by means of transcriptional attenuation, a strategy that is commonly utilized in bacteria but rarely observed in eukaryotes (Arigo et al. 2006). Though Sen1-dependent termination may represent an ancestral termination pathway, Nrd1, Nab3, and Sen1 are only moderately conserved in higher eukaryotes. In humans SCAF8/4 are putative homologs of Nrd1, Sentaxin is the homolog of Sen1p, and only very recently was RALY identified as a possible homolog of Nab3 (Yuryev et al. 1996; Chen et al. 2006; Fasken, Laribee, and Corbett 2015). However, these putative homologs have not been isolated in a complex, nor do SCAF8 or Sentaxin appear to have a role in termination or processing of snRNAs in humans (O'Reilly et al. 2014). It is likely that these putative Nrd1, Nab3, and Sen1 homologs have evolved to carry out other functions in humans, though Sentaxin is likely

[5] Compensation by Dis3 in *rrp6*Δ is largely dependent on the exonucleolytic activity of Dis3, while *rrp6*Δ/*dis3endo-* only show synergistic slow growth phenotype (Lebreton et al. 2008)

conserved for NNS-independent Sen1p functions such as coordinating transcription during DNA replication, aiding in transcription-coupled DNA repair, and resolving R loops to protect genome integrity (Ursic et al. 2004; Mischo et al. 2011; Chan et al. 2014).

Likewise, the TRAMP complex has not yet been isolated in humans though the human genome does contain homologous sequences to all three TRAMP components (Walowsky et al. 1999; Chen et al. 2001; Houseley and Tollervey 2008). In bacteria polyadenylation of RNAs stimulates exonuclease degradation much in the same way that polyadenylation by TRAMP stimulates nuclear exosome activity. LaCava et al. 2005 speculate that the ancestral role of polyadenylation was to stimulate RNA degradation and that TRAMP has retained this function within the eukaryotic nucleus, leaving polyadenylation outside of the nucleus to evolve separate functions. Given the seemingly strong functional conservation of Sen1-dependent termination and TRAMP activity to pathways in bacteria, I speculate that the NNS and TRAMP complexes have directly evolved from a common ancestor shared between bacteria and yeast. As eukaryotes evolved, NNS and TRAMP functions became less important and took on different functions in the cell. In this way, yeast may serve as an intermediary model for understanding the evolution of RNA biology in eukaryotes, thus making it particularly important to study those RNAs that are affiliated with the NNS and TRAMP complexes.

Cryptic Unstable Transcripts (CUTs)

Historically CUTs are defined as unannotated, ncRNAs that are stabilized upon loss of Rrp6 (Wyers et al. 2005). Current estimates put the total number of CUTs between 900 and 1500 (Wyers et al. 2005; Neil et al. 2009), though my work outlined in Chapter II suggests that CUT expression is far more extensive than previously observed. As Sen1-dependent transcripts, CUTs are relatively short, averaging 200-700bp in length (Wyers et al. 2005; Davis and Ares 2006; Xu et al. 2009). CUT expression is widely distributed throughout the genome,

with many being intergenic; however the vast majority of CUT expression is strongly associated with protein-coding genes (i.e. antisense, intragenic, and promoter proximal). While many propose that CUTs are the result of spurious transcriptional activity, and therefore rapidly degraded as a quality control mechanism (Wyers et al. 2005; Thiebaut et al. 2006), others have argued for possible functional roles for CUTs or CUT transcription in regulating the expression of nearby or overlapping genes (Xu et al. 2009; Thiebaut et al. 2008). It is increasingly clear that the act of transcription greatly influences the local chromatin environment through histone modifications, nucleosome repositioning, and transcriptional interference (Carrozza et al. 2005; Shearwin, Callen, and Egan 2005; Hainer et al. 2011; Thebault et al. 2011). In this way CUT transcription may also be able to affect and regulate gene expression. Regrettably, examples of CUT-based gene regulation are lacking and limited to but a few published studies[6].

Conclusion

CUTs undergo Sen1-dependent termination, are polyadenylated by TRAMP, and rapidly degraded by the nuclear exosome. This chapter has provided a broad understanding of these RNA processing complexes and pathways involved in CUT biogenesis and degradation. Additionally I have also discussed roles for Sen1-depedent termination, TRAMP polyadenylation, and nuclear exosome degradation beyond that of CUTs, touching on the broader implications for these complexes within yeast. In Chapter II I will discuss my work to identify and compare CUTs in four different strains of yeast from both S.cerevisiae and its closest relative, S.paradoxus. By assessing the extent of conserved syntenic CUT expression within these four strains of yeast I have gained several insights into possible functional roles for

[6] Two of these studies are discussed in detail in Chapter IV

CUT expression. In the course of this work I have also observed that CUTs lack a well-defined 3' nucleosome free region, in contrast to what is commonly seen at protein-coding genes, a distinction that has been largely overlooked within the field. A discussion on the implications of distinct 3' nucleosome occupancy profiles between CUTs and protein-coding is also included in Chapter II. Chapter III will discuss how I have leveraged strain-unique CUT expression, as identified in Chapter II, to inform on the role that promoter nucleosome free regions and sequence variation may play in regulating CUT expression. Lastly, in Chapter IV I will describe a method for direct quantitation of nascent CUT expression by qPCR which I adapted from NET-seq (Churchman and Weissman 2011) and have termed NET-qPCR. This tool makes it possible to study CUTs without the use of stabilizing mutant backgrounds which could prove useful to future studies investigating examples of CUT-based gene regulation.

Figures

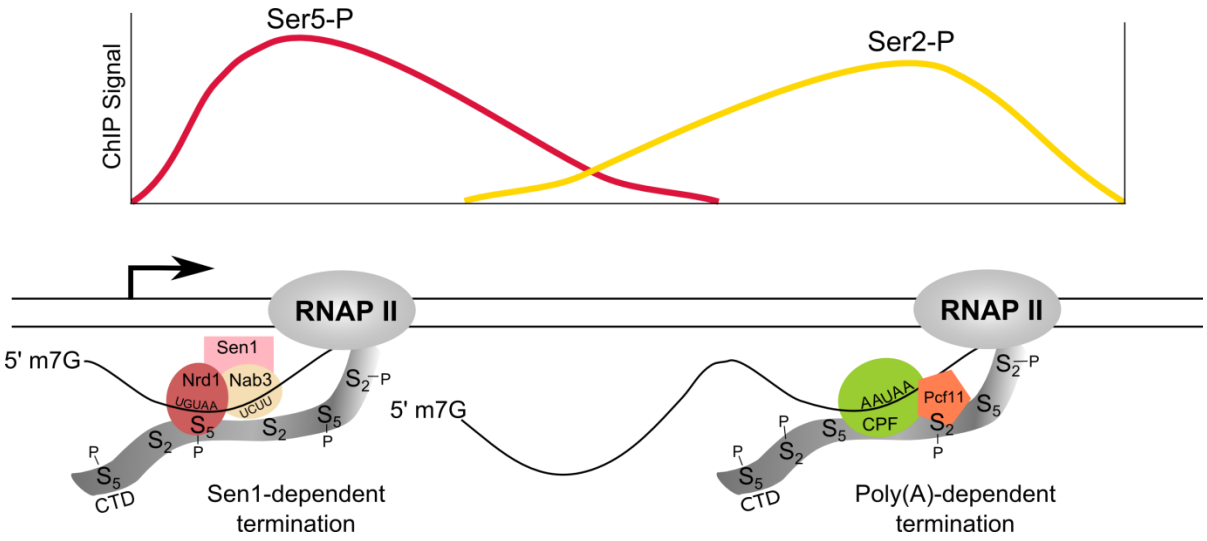


Figure 1 - Alternative Transcription Termination Pathways in Yeast (adapted from Eick and Geyer 2013). A schematic highlighting the key factors involved in Sen1-dependent termination (left) and poly(A)-dependent termination (right). The top of the figure shows hypothetical chromatin immunoprecipitation (ChIP) sequencing results for RNAPII Ser5-P (red), which peaks early in transcription, and Ser2-P (yellow), which peaks late in transcription. This inverse CTD phosphorylation pattern correlates with the binding preferences of Nrd1p (part of the NNS complex) and Pcf11p (part of the CFIA complex) helping to confer termination pathway specificity.

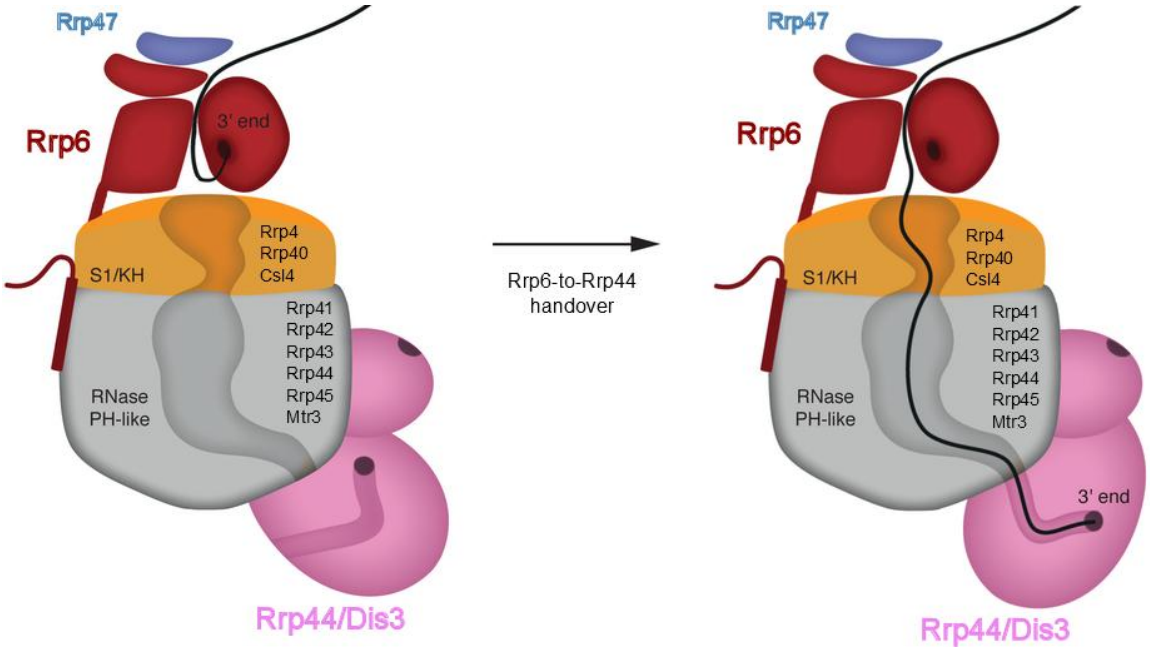


Figure 2 - RNA Degradation Paths in the Yeast Nuclear Exosome modified from Makino et al. 2015; http://www.nature.com/nature/journal/v524/n7563/fig_tab/nature14865_SF4.html).
Schematic drawing of the sequence of events leading to RNA degradation by the nuclear exosome (drawn in a longitudinal section, showing the inner central channel). RNA substrates bind at the top of Rrp6p–Rrp47p and stochastically reach the active site of Rrp6p (left panel). Alternatively, the RNA can be funneled through the central channel to the active site of Dis3p (right panel) leaving Rrp6p to stabilize substrate binding. The general location of all nuclear exosome subunits, except Mpp6, is labeled in the schematic.

Chapter II - Survey of Cryptic Unstable Transcripts in Yeast

This chapter was written for publication and is currently under review at BMC Genomics.

Authors' Contributions

The authors are Jessica M. Vera and Robin D. Dowell. JMV and RDD designed the research; JMV performed the experiments; JMV and RDD analyzed data and wrote the manuscript.

Introduction

Numerous transcriptome studies have shown the eukaryotic genome to be highly expressed, revealing pervasive transcription of intergenic and unannotated, non-protein coding regions (Nagalakshmi et al. 2008; Sultan et al. 2008; Core, Waterfall, and Lis 2008; Churchman and Weissman 2011). The discovery of unstable transcripts further adds to the complexity of the eukaryotic transcriptome. Cryptic unstable transcripts (CUTs) comprise a fraction of the unstable RNA population in yeast. These unstable, non-coding RNAs (ncRNAs) are RNA polymerase II transcribed and capped, but are terminated and polyadenylated by a non-canonical pathway involving the RNA binding proteins Nrd1, Nab3, and the poly(A) polymerase Trf4 of the TRAMP complex (Wyers et al. 2005; Thiebaut et al. 2006; Arigo, Eyler, et al. 2006; Davis and Ares 2006). Following transcription termination, CUTs are rapidly degraded by the nuclear exosome (Wyers et al. 2005) thereby rendering them virtually undetectable in wild type cells by traditional methodologies. Disrupting any step in this pathway will lead to CUT stabilization. However CUTs are customarily defined by dependency on Rrp6p nuclear exosome activity, and disrupting upstream steps, such as Nrd1p depletion or TRF4 deletion, result in extended or non-polyadenylated transcripts respectively (Davis and Ares 2006; Schulz et al. 2013; Wyers et al. 2005), that do not accurately reflect CUTs as they would be in wildtype

(WT) cells. Similar unstable ncRNAs have been identified in human cells by transient knock down of nuclear exosome components (Preker et al. 2008). While many propose that CUTs are the result of spurious transcriptional activity and therefore rapidly degraded as a quality control mechanism (Wyers et al. 2005; Thiebaut et al. 2006), others have argued for possible functional roles for CUTs or CUT expression in regulating gene expression (Xu et al. 2009; Thiebaut et al. 2008).

Historically regulation of gene expression has been attributed to sequence-specific DNA binding factors (transcription factors), transcription start site availability (via nucleosome positioning), and large co-activator complexes (mediator). However it is increasingly clear that the act of transcription greatly influences the local chromatin environment through histone modifications and nucleosome repositioning (Carrozza et al. 2005; Hainer et al. 2011; Thebault et al. 2011). Given the pervasive nature of CUT transcription and prevalent association with protein-coding genes, this transcriptional activity holds great potential to regulate gene expression. Although documented cases exist in which transcription of a CUT regulates the expression of a gene (Thiebaut et al. 2008; Arigo, Carroll, et al. 2006), the functional basis of CUT expression remains highly debated and largely unexplored.

To date CUTs have only been identified in a single species of yeast, *Saccharomyces cerevisiae*, using the reference laboratory strain S288c (Wyers et al. 2005; Xu et al. 2009; Neil et al. 2009). We have utilized a hidden Markov model (HMM) to annotate CUTs from RNA-seq data in a variety of strains from *S.cerevisiae* and *S.paradoxus* thereby allowing us to identify conserved syntenic expression of CUTs between these two species which are predicted to have diverged 2-5 million years ago (Tsankov et al. 2010; Tirosh et al. 2009). It is well documented that important cellular functions are evolutionarily conserved, and we sought to identify the population of CUTs with conserved syntenic expression to gain insights into possible functional

roles for CUT expression in yeast. Likewise, we can leverage CUT expression in other species of yeast to inform on the mechanisms underlying CUT expression.

Results and Discussion

Explicit duration HMM identifies CUTs de novo from RNA-seq data

To assess the extent of conserved CUT expression we utilized three strains of *S.cerevisiae*: S288c, Σ1278b, and JAY291, and a single strain of *S.paradoxus*: N17. In each strain background, strand-specific RNA-seq libraries were prepared for wildtype (WT) and nuclear exosome mutant *rrp6*Δ backgrounds using the Illumina RNA ligation library protocol (Levin et al. 2010). Reads were mapped to each strain's respective genome assembly (Argueso et al. 2009; Dowell et al. 2010; Engel et al. 2014)(see Methods) and CUTs were identified by an explicit duration HMM (**Figure 3A**) utilizing per nucleotide fold change values calculated from *rrp6*Δ and WT RNA-seq data (GEO accession GSE74028). Following previously established methods (Neil et al. 2009; Xu et al. 2009) our HMM was parameterized to identify CUTs as regions of the transcriptome with elevated RNA-seq coverage in *rrp6*Δ approximately ≥ 2 fold over WT. Using the HMM we derived an initial set of raw CUT annotations that are subsequently filtered to remove specific nuclear exosome targeted transcripts such as snRNAs, snoRNAs, and rRNAs (Xu et al. 2009; Wyers et al. 2005), as well as expected hits resulting from genotypic differences in *rrp6*Δ strains relative to WT. Adjacent CUTs were merged based on an RT-PCR informed strategy (**Figure 4**). Lastly we removed regions with low average *rrp6*Δ read coverage, to reduce potential false positives, as well as any remaining regions less than 100bp in length, in keeping with previously reported methods (Wyers et al. 2005; Xu et al. 2009).

To benchmark and inform our HMM parameters we leveraged previous S288c *rrp6*Δ CUT annotations based on tiling arrays from Xu et al. 2009. In S288c we have identified 687 of

885 possible Xu et al. 2009 CUTs (**Figure 3B**), where a positive hit requires that our CUT annotation overlaps ≥ 25% the length of the Xu et al. 2009 annotation or vice versa (example in **Figure 3C**), though overlap results were largely independent of extent of overlap between features (**Figure 5A**). In each case the number of positive hits is far greater than would be expected by chance (**Figure 5B**). Those Xu et al. 2009 CUTs missed by our HMM do not appear to be stabilized by disruption of nuclear exosome activity resulting by the loss of Rrp6p, though they do appear to be expressed in WT cells at levels equivalent to those CUTs we do identify and thus are not undetected due to low signal (**Figure 3D**, **Figure 5C**). Furthermore, our results are in high agreement with the 622 Xu et al. 2009 CUTs found upregulated in *rrp6*Δ by Fox et al. 2015 (Fox et al. 2015). Additionally our HMM identified a large number of novel CUTs relative to previous Xu et al. 2009 annotations (**Figure 3C**).

To further support our method of de novo CUT identification, we compared our CUTs to the *dis3*Δ transcripts from Gudipati et al. 2012 (Gudipati et al. 2012). It was recently shown that the nuclear exosome subunit Dis3p/Rrp44p, which along with Rrp6p are the major catalytic components of the nuclear exosome, plays an active role in CUT degradation, showing a synergistic cooperation with Rrp6p (Gudipati et al. 2012). While Gudipati et al. largely excluded the *rrp6*Δ Xu et al. 2009 CUTs from their *dis3*Δ annotations, producing little overlap between those two data sets (**Figure 3B**), we note that a large number, 640 of a possible 1972 *dis3*Δ transcripts (**Figure 3B**), are detected by our HMM in an *rrp6*Δ background, far more than we would expect by chance (**Figure 5E**). This demonstrates greater cooperation between the Dis3p and Rrp6p nuclear exosome subunits in the degradation of CUTs than was previously appreciated. **Figure 3E** and **Figure 5F** shows that the *dis3*Δ transcripts identified in our study have an overall lower *rrp6*Δ read coverage than the *dis3*Δ transcripts as a whole, suggesting that these transcripts are lowly expressed and may have been missed previously due to the sensitivity limitations of hybridization-based assays (Wyers et al. 2005; Xu et al. 2009). In

contrast, the *dis3*Δ transcripts not identified by our study have an overall lower fold change in *rrp6*Δ relative to WT and are more likely to comprise a Dis3p-specific subset of nuclear exosome targets. These results underscore the need for high sensitivity methods for the detection of low abundance transcripts.

CUTs lack a defined 3' nucleosome free region

To further assess the accuracy of our annotations, we compared our CUT 5' and 3' ends, as called by our HMM, to transcription start site (TSS) and transcription termination site (TTS) annotations obtained by TSS sequencing and 3' SAGE sequencing (Malabat et al. 2015; Neil et al. 2009) (**Figure 6A,B**) performed in assorted *rrp6*Δ mutants. As many as 51% and 23% of our S288c HMM CUT annotation start and stop sites were found within 50bp of these TSS and TTS annotations, respectively. It has been previously established that CUTs, like other transcripts, have a 5' nucleosome free region (NFR) upstream of the TSS (Xu et al. 2009). **Figure 6C** shows a metagene plot of 5' nucleosome occupancy (Field et al. 2008) in S288c comparing genes with a 5' UTR annotation (Nagalakshmi et al. 2008), CUTs identified in this study, and CUT TSS clusters (Malabat et al. 2015). It is clear that our CUTs have the characteristic nucleosome depletion 5' upstream of the TSS. However, when we compare the 3' end of our CUT calls to both genes with a 3' UTR annotation (Nagalakshmi et al. 2008) and CUT TTS annotations (Neil et al. 2009), it is clear there is no distinct nucleosome depletion at the 3' end (**Figure 6D**) of CUTs. We observe a similar lack of 3' nucleosome depletion for CUTs in Σ1278b and *S.paradoxus* (N17) (**Figure 8A**). The previously identified Xu et al. 2009 CUTs, however, showed a mild 3' NFR, but we found that this signal was dominated by the set of CUTS that we failed to detect in our study (**Figure 8B**).

While it is clear that chromatin remodelers, DNA binding proteins, and A/T rich sequences are driving NFRs throughout the genome (Kaplan et al. 2009; Field et al. 2008; Yuan

et al. 2005; Whitehouse et al. 2007), and that 5' NFRs are regulating transcription initiation, the role of 3' NFRs is poorly understood. In humans 3' nucleosome depletion is hypothesized to regulate polyadenylation site selection and therefore subsequent 3' end processing (Huang et al. 2013). Transcription termination, 3' end processing, and maturation of mRNAs is dependent on the cleavage and polyadenylation factor complex and comprises a pathway distinct from that of CUTs. Along with snRNAs, snoRNAs, and to some degree rRNAs, CUT transcription termination and 3' end processing is dependent on an alternative, non-canonical pathway that is dependent on the Nrd1-Nab3-Sen1 (NNS) complex (Arigo, Eyler, et al. 2006). As they utilize distinct termination and 3' end processing pathways it is therefore not surprising to see distinct 3' nucleosome structures between mRNAs and CUTs, though it has been largely overlooked within the field.

In humans a similar difference between coding and non-coding gene 3' nucleosome structure has also been observed (Huang, Liu, and Sun 2013). Though we see moderate nucleosome depletion for yeast stable ncRNAs (**Figure 7**), there is not a well-defined NFR. While it is presumed that stable ncRNAs predominately utilize the same pathways as protein-coding genes for transcription termination and polyadenylation, it has been shown in several studies that these ncRNAs accumulate in NNS and nuclear exosome mutants (Xu et al. 2009; Yassour et al. 2010; Fox et al. 2015; Schulz et al. 2013) demonstrating that these transcripts utilize the NNS pathway to some extent. That yeast stable ncRNAs lack of a well-defined 3' NFR, as we have observed for CUTs, may indicate greater utilization of the NNS pathway than was previously appreciated. These results hint at a possible role for mRNA-specific termination sequences and factors in 3' NFR production and maintenance as we only see a 3' NFR at protein-coding genes. Admittedly the varied and heterogeneous nature of CUT 3' ends makes it difficult to precisely annotate CUT TTSs, however given that we see the same nucleosome

structure for CUT TTSs from Neil et al. 2009 (Neil et al. 2009) we believe this lack of a 3' NFR (**Figure 6D**) is an inherent property of the NNS termination pathway.

A large set of CUTs show conserved expression between S.cerevisiae and S.paradoxus

Having demonstrated that our HMM successfully annotates CUTs in S288c we then applied it to the remaining three strains: Σ1278b, JAY291, and N17 (**Figure 9A**). Median CUT length in all four samples is approximately 400nt, consistent with previous findings (**Figure 9A,B**). As it remains largely unknown, we first sought to assess the extent of conserved CUT expression, here defined as detectable CUT expression within a syntenic genomic location. We used Pecan (Paten et al. 2008; Paten et al. 2009) to perform a whole genome, multiple sequence alignment of the S288c, Σ1278b, JAY291, and N17 (*S.paradoxus*) genomes. The Pecan alignment generated a universal genomic coordinate system to which all CUT annotations were converted, allowing us to identify regions where detected CUTs overlapped across the strains. In order to be confident in identification of conserved expression, CUTs with no or poor 4-way alignment (see methods) were excluded from subsequent analyses regarding CUT conservation, roughly excluding 20% of all CUT annotations. In total 64% of S288c CUTs are conserved out to *S.paradoxus* (N17) (**Figure 9C**). Alternatively we grouped all *S.cerevisiae* CUTs, 2663 in total, and found that just about half are conserved out to *S.paradoxus* which corresponds to 62% of all *S.paradoxus* CUTs (**Figure 9C**). From our identified CUTs, 855 showed conserved syntenic expression across all four strains (labeled 4x in **Figure 9D**). Our set of 4x conserved CUTs include many well-known CUTs that are expressed at NRD1, IMD3, URA2, URA8, ADE12, and LEU4 (Thiebaut et al. 2008; Arigo, Carroll, et al. 2006; Davis and Ares 2006). We selected three 4x conserved CUTs, occurring at the SIF2/YBR103W, YKU80/YMR106C, and YKL151C loci, for validation by strand-specific quantitative PCR (RT-qPCR) (**Figure 9E**). In each case strand-specificity was necessary for validation as the

candidate CUTs are antisense to an expressed mRNA. To confirm the strand-specificity of our RT reactions, we measured signal from both strands of the amplicon, (i.e. both the CUT and the mRNA) which also allowed us to measure any changes in mRNA expression. In the case of both SIF2/YBR103W and YKU80/YMR106C the fold change from *rrp6*Δ to WT for the mRNA is relatively static (log2 fold ~ 0) while the CUT is elevated in *rrp6*Δ relative to WT. In the case of YKL151C, while again we see that the CUT is elevated in *rrp6*Δ, the YKL151C mRNA shows a moderate decrease in expression in both the S288c and N17 strains, though it remains unchanged in Σ1278b and JAY291.

In addition to 4x conserved CUTs, we identified CUT expression unique to each strain (**Figure 9D**) and expression in intermediate patterns (in either 3 of 4 strains or 2 of 4 strains). We note that the N17 (*S.paradoxus*) unique CUTs contain a combination of both strain and species unique CUTs whereas for the *S.cerevisiae* strains unique CUTs are predominantly strain specific, hence the greater number of unique CUTs for N17. We selected a small number of CUTs predicted in three of the four strains for validation by RT-qPCR in order to assess our false negative rate. Doing so, we failed to confirm the absence of the CUT in the fourth strain, implying that our method may have an appreciable false negative rate (**Figure 10**). We note that many of these candidates pushed the lower bounds of qPCR detection, and we suspect that the fourth, unannotated CUT was likely missed by the HMM for similarly low abundance in our RNA-seq libraries. These results exemplify the difficulty in distinguishing between noise and true signal of low abundance RNAs even with the use of RNA-seq for their detection. Given these results, we suspect our assessment of conserved CUT expression to be conservative however it is quite clear that a large, and potentially larger, subset of CUTs have conserved expression between these two species of yeast.

Using our 4-way genome alignment we sought to examine to what extent sequence conservation parallels conserved CUT expression patterns across the strains. **Figure 11A**

shows the distribution of average percent identity for 4x conserved CUTs compared to a random set of regions demonstrating that the sequence conservation of 4x conserved CUTs is no more or less than what can be expected by chance. CUT proximal promoters (300 or 50bp upstream) have higher sequence conservation than corresponding regions of our randomized annotations. We note that the CUT and CUT promoter sequence conservation distributions are statistically distinct (p-value by two-sided KS test) possibly demonstrating distinct pressures for sequence conservation of these regions. Unique CUTs show a greater, but nonsignificant, variation in sequence conservation relative to 4x conserved CUTs, particularly in the promoter regions which may reflect sequence differences related to unique CUT expression (**Figure 11B**). Admittedly, given that our four strains are closely related, the differences we see in sequence conservation are modest. Future studies at greater evolutionary depth are required to better elucidate the relationship between conserved CUT expression and sequence conservation.

Distinct trends of gene expression correlate with CUT expression in specific architectures with genes

It has been suggested that spurious transcription at open chromatin leads to CUT expression (Wyers et al. 2005), and indeed it has been shown that a large fraction of CUTs originate from the 5' or 3' NFR of protein-coding genes (Xu et al. 2009; Neil et al. 2009). In total 1060 S288c CUTs identified by our HMM (52%) originate within either the 5' or 3' NFR of a gene (**Figure 12A**). These CUTs show greater average depletion in 5' nucleosome occupancy than CUTs that do not originate from a gene NFR (**Figure 12A**). Interestingly the 4x conserved set of CUTs are over-enriched for CUTs that originate from a gene NFR ($p=8.13 \times 10^{-25}$ by hypergeometric test) (**Figure 12B**) and this enrichment is apparent as a moderate enrichment in 5' nucleosome depletion of 4x conserved CUTs relative to all CUTs in S288c (**Figure 12B**). We see a similar trend for increased 5' nucleosome depletion for 4x conserved CUTs over all CUTs in both Σ1278b and *S.paradoxus* (N17) (**Figure 13**).

We propose that CUTs that originate from or share a gene NFR are in a strong position to influence expression of the associated gene in *cis*. CUTs originating from the 3' NFR of a gene could reduce gene expression via transcriptional inference (Shearwin, Callen, and Egan 2005) whereas CUTs originating from shared 5' NFR regions may contribute to maintaining an open chromatin conformation (Xu et al. 2009) to aid gene expression. To test for possible CUT-based regulation of these genes genome-wide, we subdivided gene and CUT NFR sharing into two general configurations: convergent, overlapping gene-CUT pairs where the CUT 5' NFR overlaps the gene 3' NFR (subsequently referred to as antisense) and divergent, non-overlapping gene-CUT pairs that share a 5' NFR (subsequently referred to simply as divergent) (**Figure 12C,D**). We note that the remaining configurations, in which CUT transcription is same sense and overlapping with a gene, not only occur less frequently but are also more difficult to analyze as we cannot distinguish read coverage between the two features (CUT and gene) and therefore cannot accurately assess transcript levels for either.

Antisense CUT expression shows evidence of transcriptional interference on sense strand

First we examined antisense gene-CUT pairs, identifying 483 such pairs in S288c (**Figure 12C**). We compared expression of the genes in these gene-CUT pairs to all expressed genes, excluding those with a same sense overlapping CUT over ≥ 50% the length of the gene CDS. Overall the genes associated with antisense CUTs showed generally decreased expression compared to all expressed genes, a trend that is more pronounced when considering only the 4x conserved CUTs (**Figure 12C**). This trend is lost, however, when we examine nascent transcription by NET-seq (Churchman and Weissman 2011) (**Figure 12C**, bottom right). This pattern is consistent with a model where CUTs impact the overlapping gene through transcriptional interference (Shearwin, Callen, and Egan 2005) as NET-seq would still be able to detect nascent pre-mRNA transcripts before transcription is terminated whereas

these aborted pre-mRNA transcripts would be missed by our RNA-seq protocol which selects for mature, polyadenylated RNAs.

Several studies report anti-correlation between stable sense-antisense transcript expression (Xu et al. 2009; Neil et al. 2009; Xu et al. 2011) however we did not observe a (anti)correlation between CUT and gene RNA-seq or NET-seq expression levels, nor did observe a (anti)correlation between CUT expression and gene repression levels, where repression was measured as the difference in gene NET-seq signal and WT RNA-seq signal. Conversely, we observed an increased trend for reduced gene expression in *rrp6*Δ over WT similar to previous reports (Yassour et al. 2010; Xu et al. 2011) regarding stable sense-antisense pairs. While mechanisms of transcriptional interference do not require a stable interfering transcript (Shearwin, Callen, and Egan 2005), we speculate that stabilization of the interfering transcript upon loss of Rrp6p may result in increased gene repression through increased DNA:RNA hybrid formation (Wahba et al. 2011; Chan et al. 2014).

We leveraged our sequence alignment to examine all antisense pairs containing 4x conserved CUTs in our remaining strains. We observed the same general trend of reduced expression of the genes in these gene-CUT pairs, but this shift is not statistically significant by the two-sided KS test (**Figure 14**). It is possible that this lack of statistical significance results from fewer total gene-CUT pairs in the remaining strains. In some cases we simply lack an annotation for the corresponding gene; in other cases the gene is not expressed and was thus removed from the analysis.

We have observed a trend for reduced expression of the genes found in antisense gene-CUT pairs similar to what is observed for stable sense-antisense pairs (Xu et al. 2011). Antisense transcription is often found to elicit a negative effect on sense transcription via transcriptional interference, and has been widely studied in yeast (Xu et al. 2011; Houseley et al. 2008; Hongay et al. 2006; Yassour et al. 2010), but almost exclusively in the context of

stable ncRNAs. Our results demonstrate that antisense CUTs elicit a negative effect on sense gene transcription in a manner consistent with stable ncRNAs and thus establish CUTs as possible sources of transcriptional interference.

Divergent CUT expression correlates with higher gene expression

Next we examined divergent gene-CUT pairs, identifying 698 in S288c (**Figure 12D**). We find that genes in this configuration have increased expression relative to all expressed genes and that this trend is more pronounced when looking only at those gene-CUT pairs with 4x conserved CUTs. We observed moderate gene ontology enrichment for various metabolic processes for genes found in divergent gene-CUT pairs, but this enrichment is lost when we only look at 4x conserved CUT pairs (**Table 1**). Notably this trend of higher gene expression appears to originate at the level of transcription as it is observed in both nascent (Churchman and Weissman 2011) and steady state RNA levels. This trend is consistent across all strains (**Figure 15**). Additionally we did not observe a correlation between CUT expression and gene expression levels in S288c in any sequencing data set (data not shown). These results are consistent with a model where divergent expression of a CUT may help to maintain an open chromatin confirmation (Xu et al. 2009).

Next we wondered if increased gene expression is a general phenomenon of divergent transcripts or if this effect is specific to gene-CUT pairs. To address this we examined divergent gene-gene pairs, identifying 398 pairs, far fewer than gene-CUT divergent pairs despite a far greater number of protein coding genes overall suggesting a bias for CUTs in divergent transcript pairs with protein coding genes. When we compared the expression of divergent gene pairs to all expressed genes (**Figure 16**) we did not find a significant difference in the expression distribution suggesting the effect seen in **Figure 12D** is specific to CUTs.

Many have characterized bidirectional transcription, looking at both CUTs and stable ncRNAs (Xu et al. 2009; Neil et al. 2009) but have failed to report on any observed effects on the expression of the associate genes. We hypothesized that divergent CUT expression from a shared NFR may help maintain the NFR thereby allowing for rapid and efficient expression of the associated gene and most likely benefitting higher expressed genes. Others have reported that long and deep NFRs commonly correlate to constitutive and highly expressed growth genes (Tsankov et al. 2010). That genes found in divergent gene-CUT pairs are enriched for various metabolic processes is consistent with these previous findings. While we cannot rule out that CUT expression is an incidental result of higher expression at these genes, we note that we do not see divergent CUT expression at all highly expressed, or even the highest expressed genes. Additionally we see little correlation between CUT and gene expression levels further suggesting that CUT expression not a spurious result of leaky promoters of highly expressed genes. Strikingly divergent gene-gene pairs did not elicit the same expression trends observed in gene-CUT pairs in the same configuration. This further supports a role for divergent CUT expression in regulating the expression of associated genes and hints to the possibility of CUT-specific factors in mediating this trend.

Conclusion

In this study, we used an explicit duration HMM to annotate CUTs from RNA-seq in an *rrp6*Δ background for a variety of yeast strains from the species *S.cerevisiae* and *S.paradoxus*. This allowed us make the first assessment of conserved intra- and interspecies CUT expression. Though our estimates appear conservative, we find that CUT expression is highly conserved within and between these two species of yeast despite the presence of sequence variation within upstream promoter regions. These finding warrant additional studies assessing CUT expression in other, more distantly related yeast species to better understand the

relationships between DNA sequence and CUT expression. Our work has additionally demonstrated that CUTs lack 3' nucleosome depletion as commonly observed for protein-coding genes and that CUT expression is not only highly associated with protein-coding genes but may also be regulating these genes in a manner consistent with the orientation of CUT transcription relative to these genes.

Methods

Strain construction

Σ1278b WT and S288c (BY4741) WT were provided by the Fink lab. Σ1278b *rrp6*Δ and S288c *rrp6*Δ were provided by the Boone lab (Dowell et al. 2010). JAY291 WT was provided by Lucas Argueso (Argueso et al. 2009). We transformed JAY291 WT with the KanMX cassette from S288c *rrp6*Δ to delete RRP6 in JAY291. N17 WT was provided by the Fay lab, and transformed with a NatMX cassette to delete RRP6 in N17. See Appendix A for complete strain genotypes.

Genome sequences and annotations

S288c genome and annotations are from the Saccharomyces Genome Database (SGD) S288c genome version 64 (Engel et al. 2014). Σ1278b genome and annotations are available from Dowell et al. 2010 (Dowell et al. 2010). JAY291 genome and annotations are from the Duke 2009 (Argueso et al. 2009) release, downloaded from SGD. We used a modified version of the JAY291 Duke 2009 assembly, where the reverse compliment of several contig sequences were used so as to match the orientation of homologous S288c sequences. N17 genome and annotations were downloaded from the Sanger Welcome Trust FTP site as part of the Saccharomyces Genome Sequencing project (Kellis et al. 2003).

RNA-sequencing libraries

Cells were grown in YPD to an OD of 0.6. Total RNA was isolated via hot acid phenol method and DNAse treated with Promega DNAse RQ1 to remove contaminating DNA. Poly(A) RNA was isolated using either a single round of Qiagen oligotex mRNA isolation kit or two rounds of Dyna bead mRNA isolation kit. Strand specific RNA-seq libraries were constructed from 500ng of poly(A) RNA using the Illumina RNA ligation library protocol from (Levin et al. 2010). We sequenced, by Illumina HiSeq, biological duplicates of each sample. To remove rRNA reads, we first used Bowtie v0.12.7 (Langmead et al. 2009) to map reads to a single repeat of the rDNA locus allowing two mismatches. The remaining reads were mapped uniquely to the genome sequence of each respective strain allowing up to two mismatches. Per nucleotide read coverage was obtained using BEDTools (Quinlan and Hall 2010), corrected for read first nucleotide biases and read mappability, and then normalized by the tens of millions of mapped reads per sample. Per nucleotide coverage was averaged across replicates. Fold change from *rrp6*Δ to WT was calculated for every nucleotide in the genome using bias corrected coverage values. A Laplace prior (+1) was added to all coverage values to avoid division by zero when calculating the per nucleotide fold change.

Explicit duration hidden Markov model

We developed an explicit duration hidden Markov model (HMM) to analyze per nucleotide *rrp6*Δ/WT RNA-seq fold change signal (**Figure 3A**) using the Matlab HMM toolkit (MATLAB 2012b, The MathWorks Inc., Natick, MA, 2012). The HMM consists of two main states, one parameterized to non-elevated regions of the transcriptome (i.e. not CUTs) and one for elevated (approximately ≥ 2 fold) regions of the transcriptome (i.e. CUTs). Specifically we expanded the CUT state into nine identical sub-states with unidirectional movement through the model (**Figure 18**) thereby setting the minimum length of a CUT to nine nucleotides and

producing a 10-State model that approximates a hidden semi-Markov model (Datta, Hu, and Ray 2008). This allowed us to deviate from the exponential duration modelling of traditional HMMs and produce CUT annotations with a length distribution that better approximated previous studies (Wyers et al. 2005; Xu et al. 2009). We note that when the model is used to generate representative sequences, the CUT state of the model produced sequences that are generally long (> 34,000bp) reflecting our bias to identify long regions of relatively consistent elevated coverage. Per nucleotide fold change values were converted to discrete values for analysis by our HMM as necessitated by the Matlab toolkit (**Table 2**). Transition and emission probabilities are available in **Table 3** and **Table 4**.

CUT identification

From the HMM we derived an initial set of raw CUT annotations. These raw annotations were filtered to remove snRNAs, snoRNAs, and rRNAs as well as expected hits resulting from genotypic differences in *rrp6*Δ strains relative to WT. Any remaining regions within 450bp were merged together into a single annotation. Regions with average *rrp6*Δ read coverage less than 33% the coverage of all nonzero coverage bases for that sample and any regions less than 100nt in length also were removed.

Annotation overlap and significance test

We used IntersectBed (Quinlan and Hall 2010) to quantify the extent of overlap between our HMM S288c CUT annotations and other data sets (**Figure 3B**) requiring overlap of ≥25% the length of either annotation. Because we removed raw HMM CUT annotations that overlapped snRNAs, snoRNAs, and rRNAs, we likewise removed any annotations from Xu et al. 2009 and Gudipati et al. 2012 that overlapped the removed raw HMM CUTs in S288c to properly reflect the extent of overlap between these data sets and our S288c CUTs. Hence only 885 of a total 925 Xu et al. 2009 CUTs and 1972 of a total 2032 Gudipati et al. 2012 *dis3*Δ

transcripts were used in subsequent overlap analyses. To determine statistical significant we randomly sampled genomic regions with the same length distribution as S288c identified CUTs. After 200 iterations, overlap of these randomly sampled regions and previously annotated CUTs or *dis3*Δ transcripts approximate a normal distribution (**Figure 5B,E**). We use two standard deviations from the mean to assess significance within our CUT annotations.

Nucleosome occupancy and metagene analysis

For S288c nucleosome occupancy we used summarized nucleosome occupancy from Field et al. 2008 (Field et al. 2008) data available from the SGD website. For Σ1278b and N17 we mapped the raw reads from Tsankov et al. 2010 (Tsankov et al. 2010) according to their methods with the exception that we used the N17 *S.paradoxus* genome instead of NRRLY-17217 used in their study. Metagene plots were constructed by averaging the nucleosome occupancy for each base pair in a 500bp window for all annotations in the analyzed data sets.

CUT transcription start site comparisons

The Malabat et al. 2015 study identified TSS clusters in various mutant backgrounds including *rrp6*Δ. TSS clusters were sorted and grouped according to their relative positions to annotated features. Since clusters assigned to CUTs required overlap with previous CUT annotations, we included all antisense, same sense, and intergenic (i.e. A, B, and I) clusters with an *rrp6*Δ/WT fold change ≥ 1.5 as calculated in their study.

Pecan whole genome alignment

We used Pecan version 0.9 (Paten et al. 2009; Paten et al. 2008) to generate a four-way whole genome multiple sequence alignment of the S288c, Σ1278b, JAY291, and N17 genomes. As the JAY291 genome is currently only available in a contig assembly (Argueso et al. 2009),

we first used BLAT to find the single best hit for each contig to the S288c genome in order to produce a pseudo-genome assembly as required by Pecan.

Conserved CUT expression

First we converted all CUT annotations from strain-specific coordinates to the 4-way alignment coordinate system. Then we calculated a histogram of CUT annotations along the 4-way alignment and all continuous regions ≥ 1 in the histogram were selected. The total histogram signal over these selected regions was averaged and used to determine the total number of CUTs overlapping that region Regions with an average histogram signal > 4 denoted 4x conserved CUT expression. We identified 208 regions where the CUT annotations were incongruent across the four strains and applied hand edits to resolve these incongruences where possible. Additionally, we examined those CUTs in 3 of the 4 strains and if the CUT is missed in the fourth strain by our filtering procedure (i.e. the fourth strain has a CUT in the raw HMM output) we brought back the filtered CUT annotation and considered these to be 4X conserved CUTs. The resulting changes in CUT annotations are reflected in summaries reported in **Figure 9A**. After removing those CUTs with indels (relative to the four-way alignment) for more than 25% the length of the CUT, we derived the conserved expression results reported in **Figure 9C,D**. In the case of unique CUTs (**Figure 9D**) we only reported those CUTs that did not overlap a raw (but removed) annotation in either of other strains. To determine the significance of our CUT conservation analysis we randomized CUT annotations in all four strains to assess the chance of CUT conservation simply by chance. With 200 iterations, little to no random 4x conserved CUTs were found (**Figure 17**).

CUT expression validation by RT-qPCR

Primer sequences can be found in Appendix B. We selected candidate CUTs that were novel to our study relative to Xu et al. 2009 however in some cases candidates were also

identified by Gudipati et al. 2012 as *dis3*Δ transcripts. To validate CUTs identified by the HMM we performed strand specific RT-qPCR using a 5' tagged gene-specific RT primer (Plaskon, Adelman, and Myles 2009) for cDNA synthesis of DNAsed, total RNA. In many cases strand specificity was necessary to distinguish CUT transcripts in the presence of overlapping, antisense mRNAs. Tagged RT primer distinguishes primer-specific cDNA from false primed cDNA that frequently occurs between overlapping, antisense transcripts. Subsequent PCR reactions used a universal forward primer complimentary to the RT tag and a gene specific reverse primer. Primer sequences can be found in Additional File 15. In some cases it was necessary to use the tagged RT primer as the forward primer during qPCR to avoid primer dimers between the universal forward primer and the gene-specific reverse primer. ACT1 was used as a normalizing endogenous control and was also measured strand specifically. A few candidates did not require strand-specific RT-qPCR (see Appendix B). These samples instead used random hexamer RT primers and gene-specific qPCR primers. Fermentas Maxima Reverse Transcriptase was used for all RT reactions. Three biological replicates were grown to O.D. 0.6 in YDP and total RNA was isolated by hot acid phenol method and DNAse treated with Promega DNAse RQ1.

NFR sharing between CUTs and protein-coding genes

Metagene plots in Figure 2C,D show the general location of the 5' NFR ranging from -200 to 0bp from the transcription start site and the 3' NFR ranging from +100 to -100 from the transcription termination site. We annotated these regions for each gene where corresponding untranslated region annotations were available (Nagalakshmi et al. 2008). We annotated CUT 5' NFRs in the same fashion. We considered potential instances of NFR sharing when the CUT 5' NFR annotation overlapped ≥50% (minimum 100bp) the length of a gene 5' or 3' NFR.

Figures

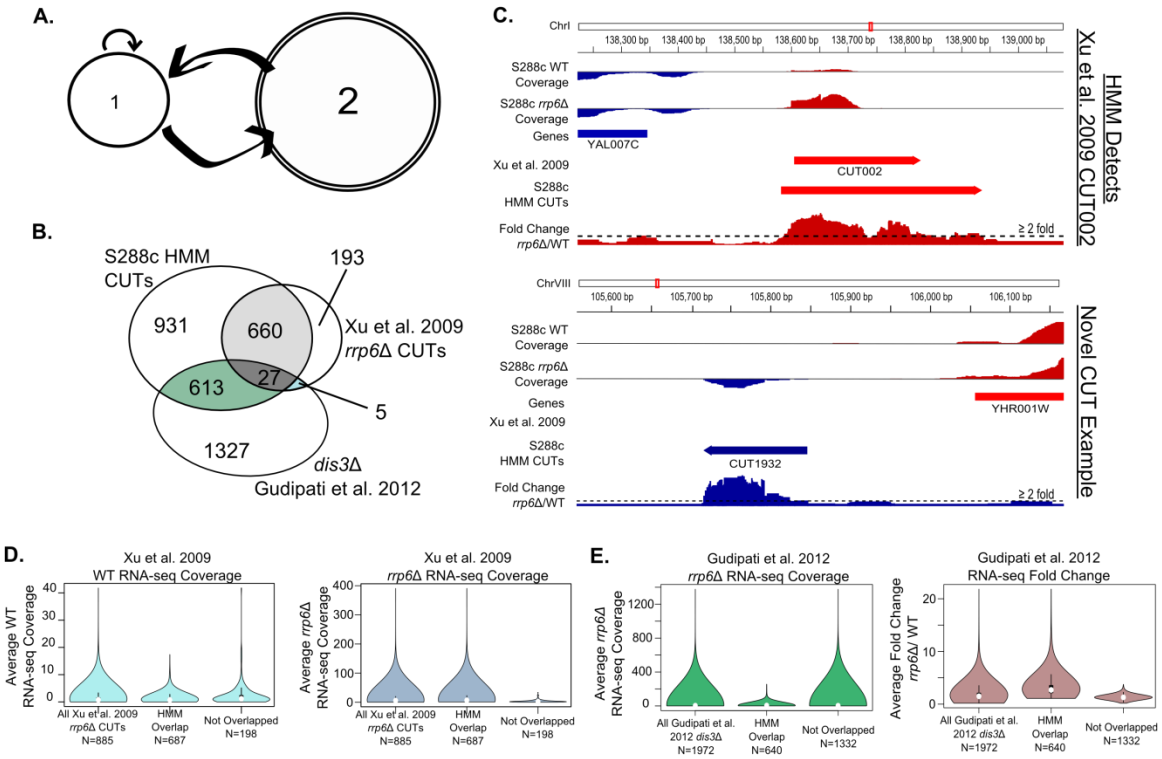


Figure 3 - 10-state HMM Identifies CUTs de novo from RNA-seq

A) A state diagram of our explicit duration HMM. State 1 describes regions which have little to no fold change, while state 2 captures CUT regions of elevated fold change in *rrp6*Δ relative to WT. **B)** Venn diagram showing overlap between S288c CUTs annotations as determined by our HMM, *rrp6*Δ CUTs from Xu et al. 2009, and *dis3*Δ transcripts from Gudipati et al. 2012. Minimum overlap of 25% the length of the annotations in one set or the other is required for positive matches. **C)** IGV[47, 48] snapshots showing two examples of CUTs detected in S288c by our HMM. Top example shows previously identified Xu et al. CUT002 which is also identified by our HMM. Bottom example shows a novel CUT identified in this study. For each example, tracks are S288c WT RNA-seq coverage, S288c *rrp6*Δ RNA-seq coverage, annotated genes, Xu et al. 2009 annotated CUTs, CUTs called by our HMM, and *rrp6*Δ/WT fold change within the region. Strand-specific data is color coded with Watson/plus strand in red and Crick/minus strand in blue. **D)** Violin plots comparing the average S288c RNA-seq WT coverage and *rrp6*Δ/WT fold change for all 885 possible Xu et al. 2009 CUTs, the 687 CUTs overlapped by CUTs detected by our HMM, and the 198 remaining CUTs not overlapped by CUTs detected by our HMM. The Xu et al. 2009 CUTs not identified in this study are presumably missed due to lack of stabilization in *rrp6*Δ. **E)** Violin plots comparing the average S288c RNA-seq *rrp6*Δ and *rrp6*Δ /WT fold change for all 1972 possible Gudipati et al. 2012, the 640 transcripts overlapped by CUTs detected by our HMM, and the 1332 remaining transcripts not overlapped by CUTs detected by our HMM. The *dis3*Δ transcripts missed in previous *rrp6*Δ only, tiling array studies are presumably missed to limitations in tiling array sensitivity.

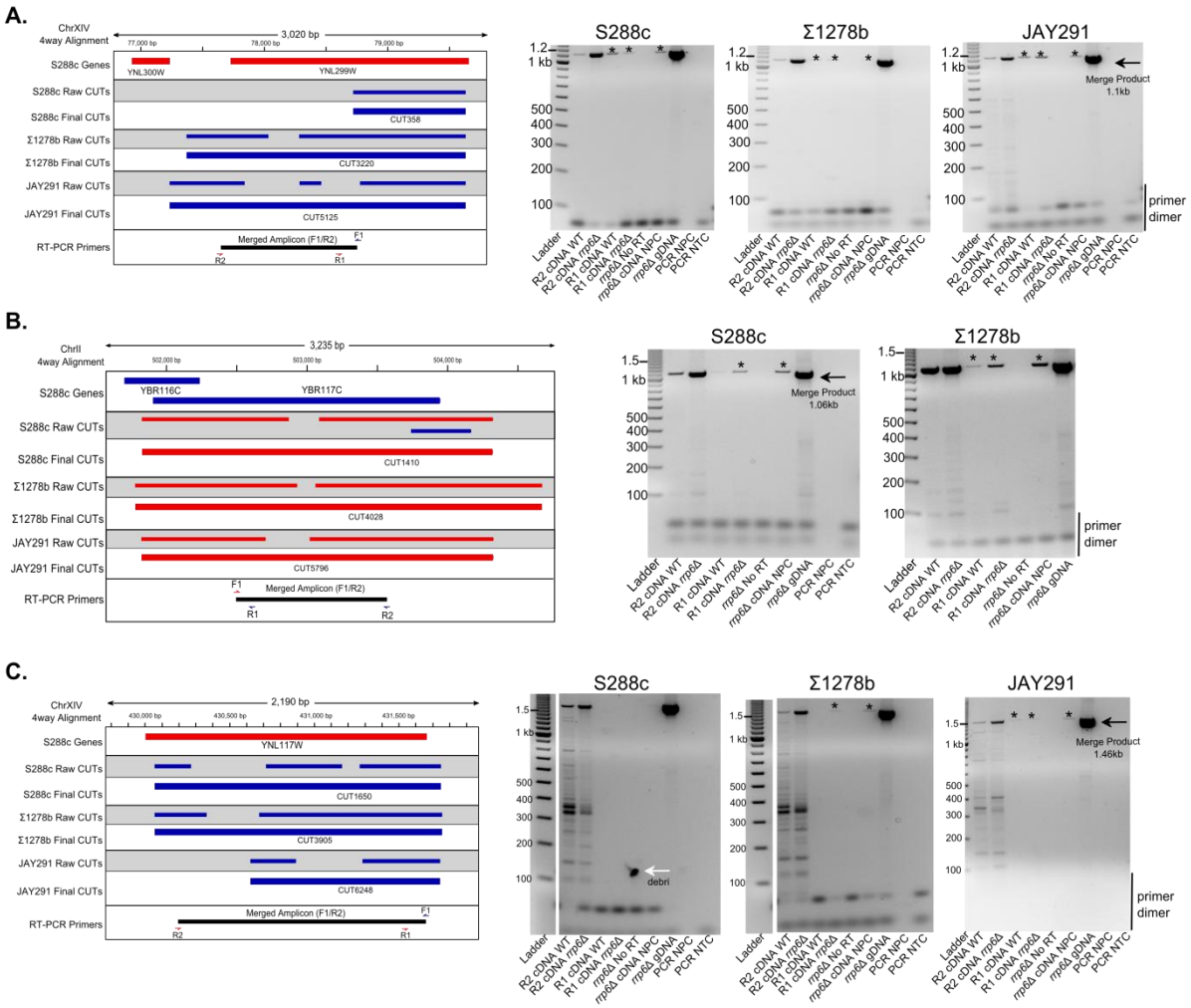


Figure 4 - RT-PCR validation of raw CUT annotations merging strategy

Three candidate regions selected to determine whether adjacent CUT regions, supported by calls in multiple strains, should be merged in post processing. Candidates tested are located at the A) YNL299W/TRF5 locus B) YBR117C/TKL2 locus and C) YNL117W/MLS1 locus. In each case strand-specific RT primers were used to generate cDNA and PCR was performed to produce an amplicon that spans the gap in the raw annotations. Left: An IGV[47, 48] snapshot with tracks showing the gene, our raw CUT, and our final CUT annotations for the strains S288c, Σ1278b, and JAY291 after conversion to the 4-way Pecan alignment (see Methods). Additionally we show the location of each primer used and the resultant amplicon of a positive merge result. Strand-specific data is color coded with Watson/plus strand in red and Crick/minus strand in blue. Right: 2% agarose gel showing RT-PCR results. For each candidate we designed two primer pairs with each pair located on either side of the gap between raw CUT annotations as identified by our HMM. We generated strand-specific cDNA from both WT and *rrp6*Δ total RNA samples with each reverse primer and performed PCR on these cDNA with F1/R2 primer pair. F1/R2 primers should produce a merge amplicon product only if the candidate CUT is a single transcript spanning the gap in raw CUT annotations. Amplification in R1 primed cDNA served as a negative control, as amplification should only occur in R1 primed cDNA; this also helped to confirm strand-specificity. We included genomic positive control, a no primer control (NPC) RT sample to distinguish false-primed cDNAs (denoted with *), and a no template control (NTC) to distinguish primer dimers.

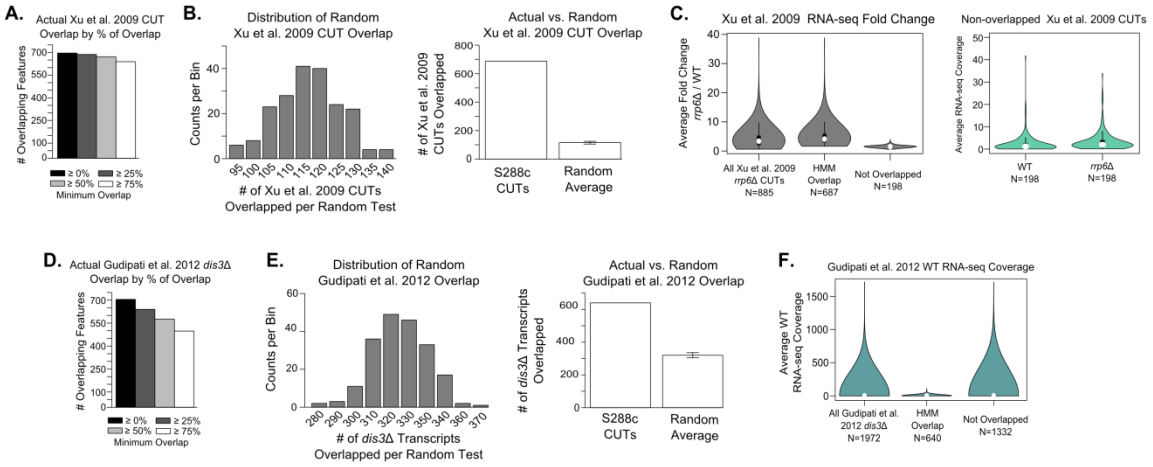


Figure 5 - S288c HMM CUT comparison to Xu et al. 2009 and Gudipati et al. 2012 annotations

Comparisons of S288c CUTs identified by our HMM and Xu et al. 2009 CUTs or Gudipati et al. 2012 *dis3*Δ transcripts. Extent to which minimum overlap influences number of features concordant between HMM detected CUTs and **A)** Xu et al. 2009 CUTs. **B)** Overlap is more than would be expected by chance. S288c CUT annotations were randomized (see Methods) and the number of features overlapped in each data set was collected over 200 iterations and plotted as a histogram. The average number of features overlapped after 200 iterations, with error bars denoting standard deviation, is plotted for comparison to actual S288c overlap results. Actual S288c CUTs overlap is greater than 2 standard deviations from random trials. **C)** Violin plots as seen in Fig1.D showing average RNA-seq fold change for all Xu et al. 2009 CUTs, Xu et al. 2009 CUTs overlapped by CUT identified by our HMM, and Xu et al. 2009 CUTs missed by our study where we observe equivalent expression in WT and *rrp6*Δ backgrounds. **(D-F)** Similar comparison for Gudipati et al. 2012 ***dis3Δ*** transcripts.

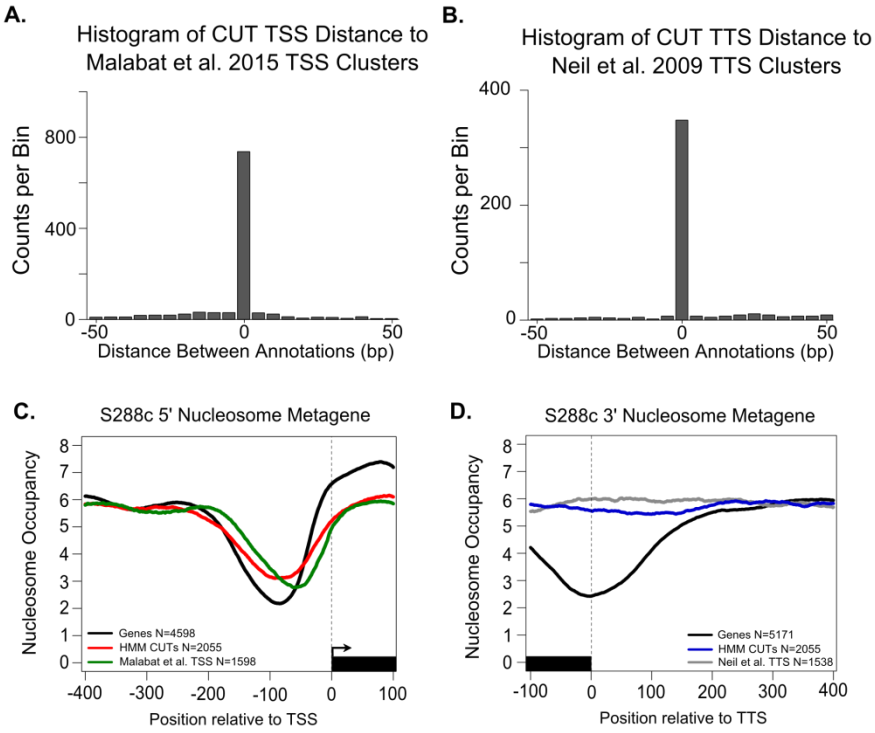


Figure 6 - CUT Start and Stop Sites Concurrent with Previous Data and Show Distinct 3' Nucleosome Structure

A) Histogram showing the distribution of the distance between S288c CUT TSSs relative to Malabat et al. 2015 CUT, intergenic, same sense, and antisense TSS clusters (see Methods). Histogram is only reporting distances for S288c CUTs that are within 50bps of a TSS cluster. Bin widths are 5bp. **B)** Histogram showing the distribution of the distance between S288c CUT TSSs relative to Neil et al. 2009 TTS clusters. Histogram is only reporting distances for S288c CUTs that are within 50bps of a TTS cluster. Bin widths are 5bp. **C)** Metagene plot showing the average S288c nucleosome occupancy of a 500bp window around the TSS for all genes with a 5' UTR annotation (black), our HMM identified CUTs (red), and Malabat et al. 2015 CUT TSS clusters (green). **D)** Metagene plot showing the average S288c nucleosome occupancy of a 500bp window around the TTS of all genes with a 3' UTR annotation (black), our HMM identified CUTs (blue), and Neil et al. 2015 TTS clusters (grey).

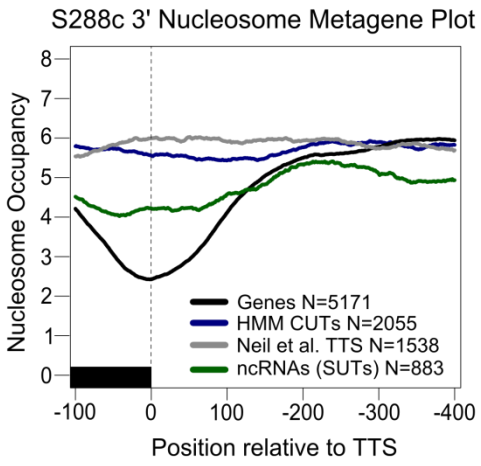


Figure 8- ncRNAs have moderate 3' nucleosome depletion
Metagene plot showing the average S288c nucleosome occupancy of a 500bp window around the TTS of all genes with a 3' UTR annotation (black), our HMM identified CUTs (blue), Neil et al. 2015 TTS clusters (grey), and ncRNAs (green) also known as stable unannotated transcripts (SUTs) from Xu et al. 2009. ncRNAs show moderate 3' nucleosome depletion within the same 200bp region where genes have a strong 3' NFR producing a nucleosome occupancy pattern that is distinct from both CUTs and genes.

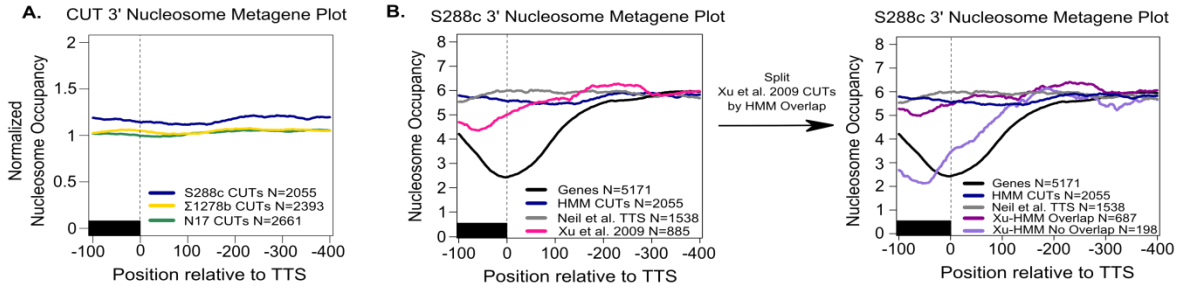


Figure 7 - CUTs lack a 3' NFR
A) Metagene plot showing the average nucleosome occupancy of a 500bp window around the TTS of all S288c(blue), Σ1278b(yellow), and *S.paradoxus*$_{N17}$(teal) CUTs identified by our HMM. For comparison across strains, nucleosome occupancy was normalized by the average nucleosome occupancy per base pair in each strain. Like S288c CUTs, we see do not see 3' nucleosome depletion in our other strains for which nucleosome occupancy data is available. **B)** Left: Metagene plot showing the average S288c nucleosome occupancy of a 500bp window around the TTS of all genes with a 3' UTR annotation (black), our HMM identified CUTs (blue), Neil et al. 2015 TTS clusters (grey), and Xu et al. 2009 CUTs (pink). Moderate 3' nucleosome depletion can be seen for Xu et al. CUTs 2009. Right: When we split the Xu et al. 2009 CUT annotations into two groups, those overlapped by an S288c CUT identified by our HMM (maroon), and those that are not (lilac), we see distinct nucleosome occupancy patterns for the two groups. Those Xu et al. 2009 CUTs that overlap an S288c CUT identified by our HMM also appear to lack a 3'NFR and the moderate depletion previously seen in the left graph is largely restricted to those Xu et al. 2009 CUTS that we failed to detect and which also appear to be stable, albeit lowly expressed RNAs (see Figure 1D and Figure S2C).

A.

Summary of HMM Results

Strain	Total CUTs	Median Length
S288c	2055	399
Σ1278b	2393	437
JAY291	2268	400
S.paradoxus$_{(N17)}$	2661	480

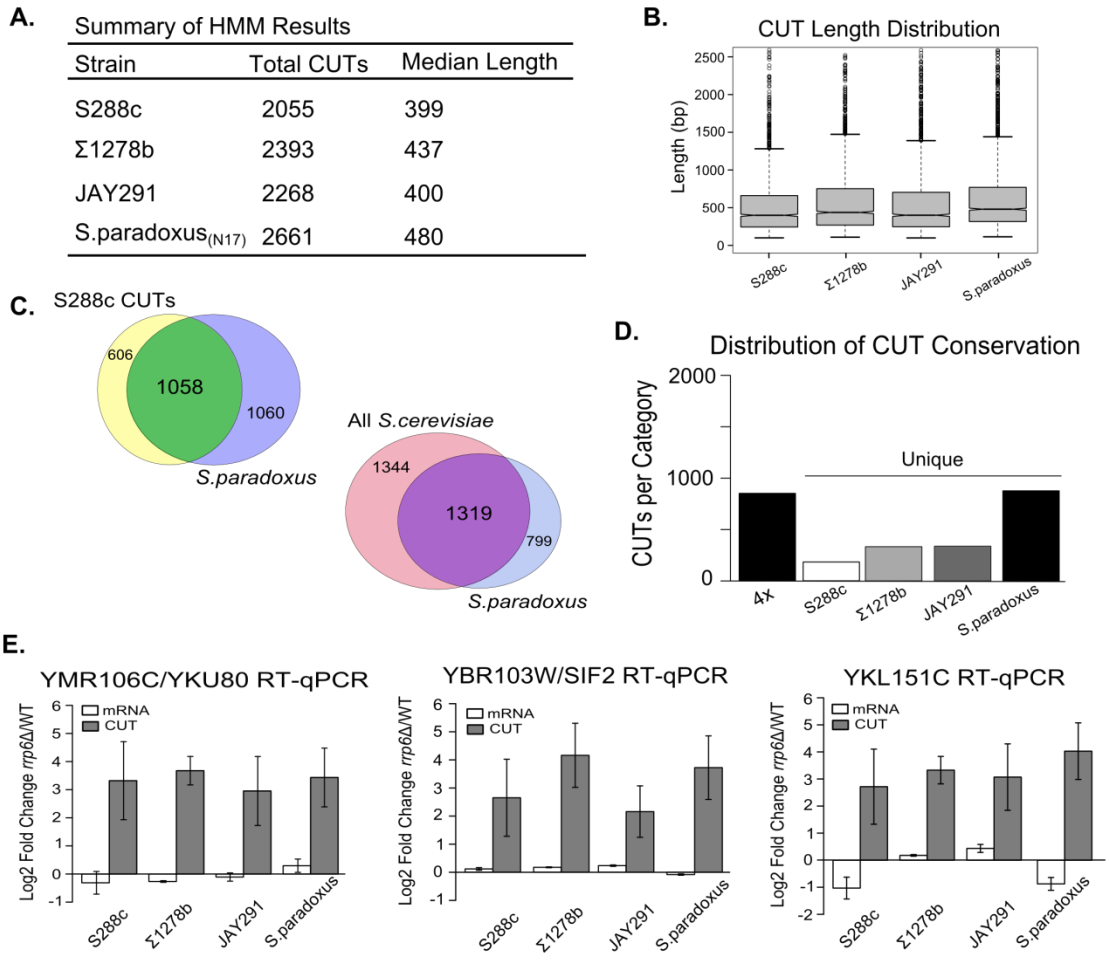


Figure 9 - Assessment and Validation of Conserved CUT expression

A) A summary of the HMM identified CUTs in each strain. **B)** Box and whiskers plot showing CUT length distribution for each strain. We note that the y-axis range was limited to a maximum length of 2.5kb for better comparison of the distributions across the strains. **C)** Venn diagrams showing conserved CUT expression between the *S.cerevisiae* strain S288c and *S.paradoxus* (N17) and the conserved CUT expression between all *S.cerevisiae* strains (S288c, Σ1278b, and JAY291) and *S.paradoxus* (N17). **D)** Distribution of CUTs with conserved syntenic expression across all four strains (4x) profiled or present in only one strain (unique). **E)** RT-qPCR validation of three 4x conserved CUTs. In each case the candidate CUT is expressed antisense to an annotated gene and qPCR was performed strand-specifically with the same amplicon to distinguish between signal from the mRNA and the antisense CUT. Log2 fold change of *rrp6*Δ/WT was calculated after normalization to ACT1 (also acquired strand-specifically). In each case the CUT-specific strand shows a significant increase in transcript abundance in *rrp6*Δ relative to WT while the mRNA-specific strand shows little to no change, except with YKL151C mRNA. All qPCR was performed with biological triplicates and error bars denote standard deviation of fold change by coefficient of variation calculations.

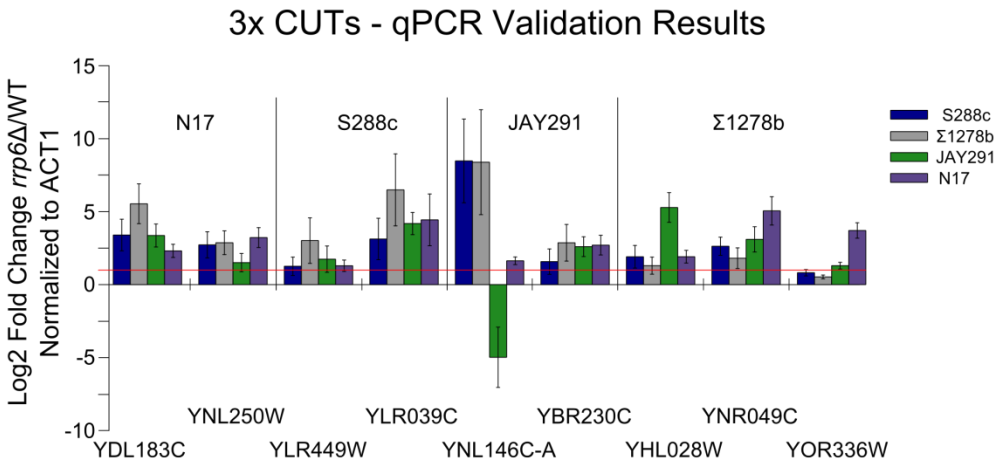


Figure 10 - Assessment of HMM false negative rate by RT-qPCR

RT-qPCR of CUTs expressed in three out of four strains (3x CUTs). For simplification candidates are named based on closest or overlapping protein-coding gene annotations (x-axis). Candidates are grouped and labeled (above the bar plot) according to the strain that lacks the corresponding CUT annotation. RT-qPCR was performed either strand-specifically or non-strand specifically depending on the presence of overlapping antisense gene annotations (see Methods, Table S7). Log2 fold change of *rrp6*Δ/WT was calculated after normalization to ACT1. The red dashed line marks two-fold cutoff. In all but one instance, JAY291 YNL146C-A, the "missing" CUT shows elevated expression, as seen in the remaining strains. All qPCR was performed with biological triplicates and error bars denote standard deviation of fold change by coefficient of variation calculations.

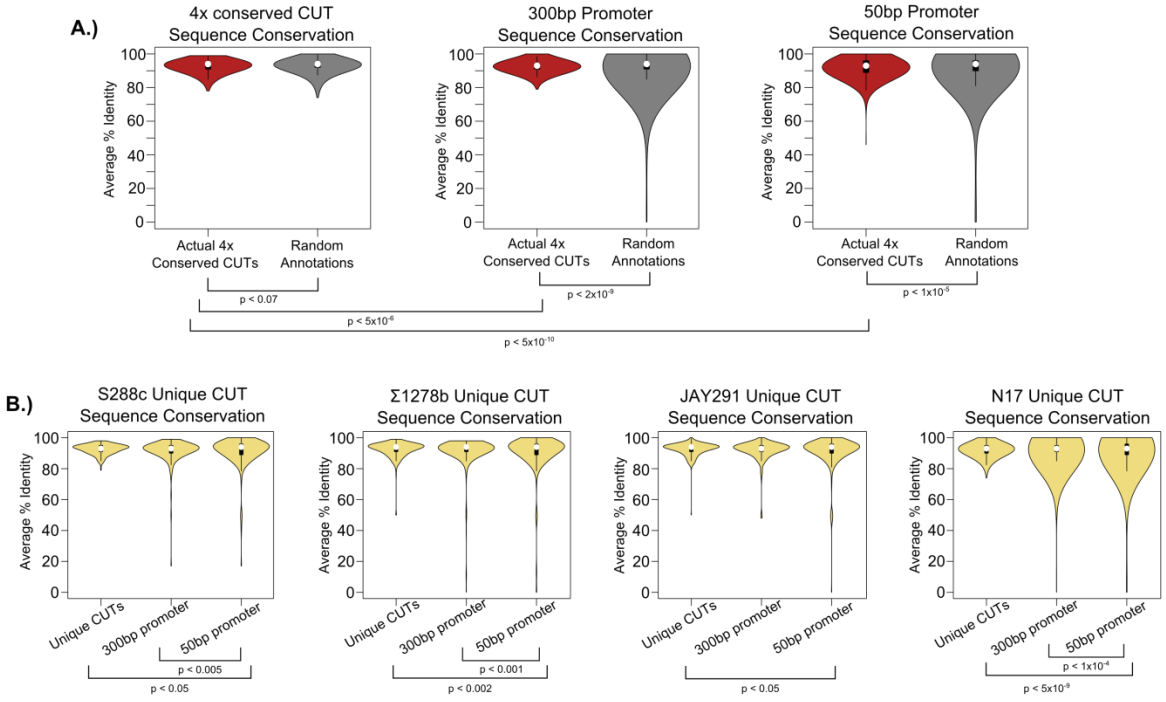


Figure 11 - Sequence conservation of CUTs

A) Violin plots showing the average sequence conservation, calculated from our 4-way genome alignment, of all 4x conserved CUTs, 300bp upstream and 50bp upstream promoters (red), and compared to the average percent identity of a randomized set of annotations (grey) that recapitulates the 4x conserved CUTs in length and frequency. We used the S288c start coordinate and the longest stop coordinate as the start and stop coordinates for the 4x conserved CUTs when calculating average percent identity. Included are all p-values < 0.1 obtained by the two-sided KS test. **B)** Violin plots showing the average sequence conservation, calculated from our 4-way genome alignment, of the CUTs unique to each strain and the 300bp upstream and 50bp upstream promoters. Included are all p-values < 0.1 obtained by the two-sided KS test.

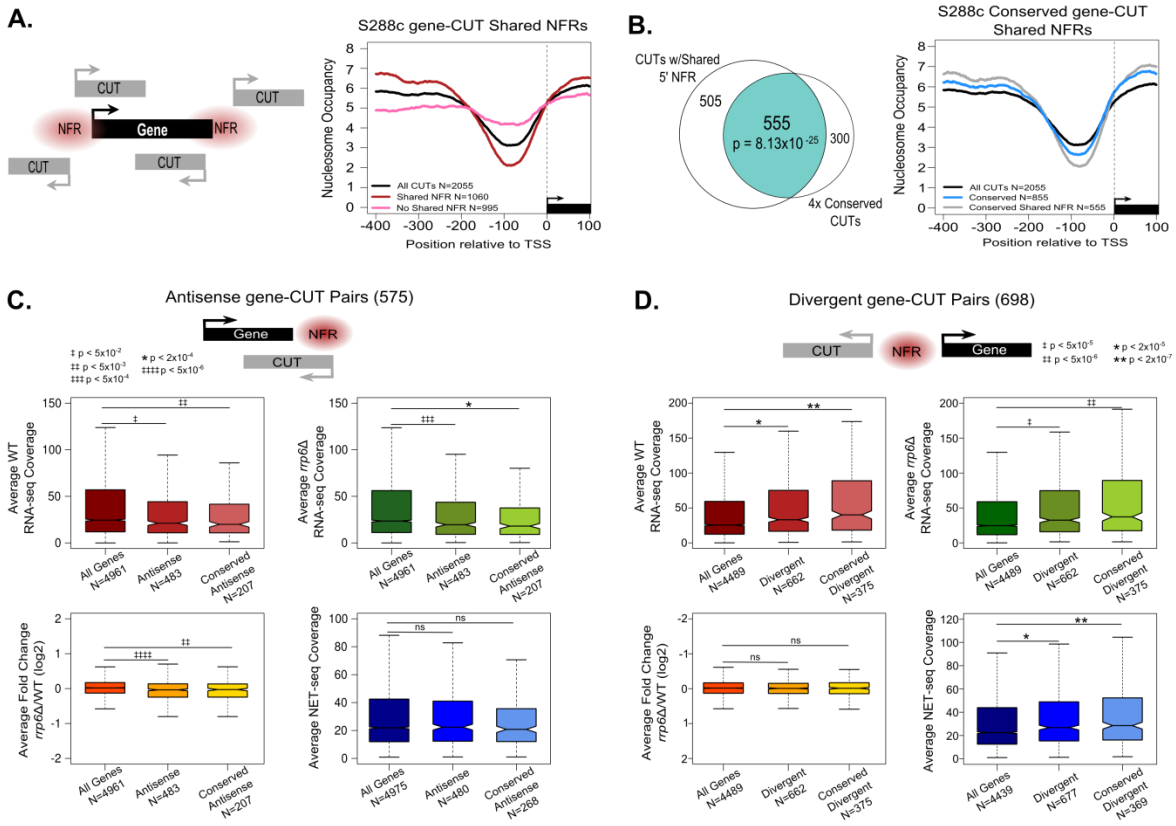


Figure 12 - Distinct trends of gene expression correlate with CUT expression in specific architectures with genes

A) Left: Schematic demonstrating the configurations in which a CUT can originate from a gene NFR. Right: Metagene plot of S288c nucleosome occupancy for a 500bp window around the TSS of all S288c CUTs identified by our HMM, the subset of CUTs found to originate from a gene NFR (red), and the remaining CUTs that do not originate from a gene NFR (pink). **B)** Left: Venn diagram of the overlap of CUTs that originate from a gene NFR and 4x conserved CUTs. Right: Metagene plot of S288c nucleosome occupancy for a 500bp around the TSS of all S288c CUTs identified by our HMM (black), all of 4x conserved CUTs from S288c (blue), and the subset of 4x conserved CUTs that originate from a gene NFR (grey). **C)** Examination of antisense gene-CUT pairs in S288c. Box and whisker plots shows the distribution the average WT RNA-seq coverage (red), *rrp6*Δ RNA-seq coverage (green), log2 *rrp6*Δ/WT RNA-seq fold change (yellow), and NET-seq coverage from Churchman and Weissman 2011 (blue) for all expressed genes with a 3' UTR annotation, those genes in antisense gene-CUT pairs, and the genes from the subset of Antisense gene-CUT pairs with a 4x conserved CUT. All points outside the whiskers (outliers) are not displayed. All p-values are derived from the two-sided KS test. **D)** Examination of divergent gene-CUT pairs in S288c. Box and whisker plots shows the distribution the average WT RNA-seq coverage (red), *rrp6*Δ RNA-seq coverage (green), log2 *rrp6*Δ/WT RNA-seq fold change (orange), and NET-seq coverage from Churchman and Weissman 2011 (blue) for all expressed genes with a 5' UTR annotation, those genes in divergent gene-CUT pairs, and the genes from the subset of divergent gene-CUT pairs with a 4x conserved CUT. All points outside the whiskers (outliers) are not displayed. All p-values are derived from the two-sided KS test.

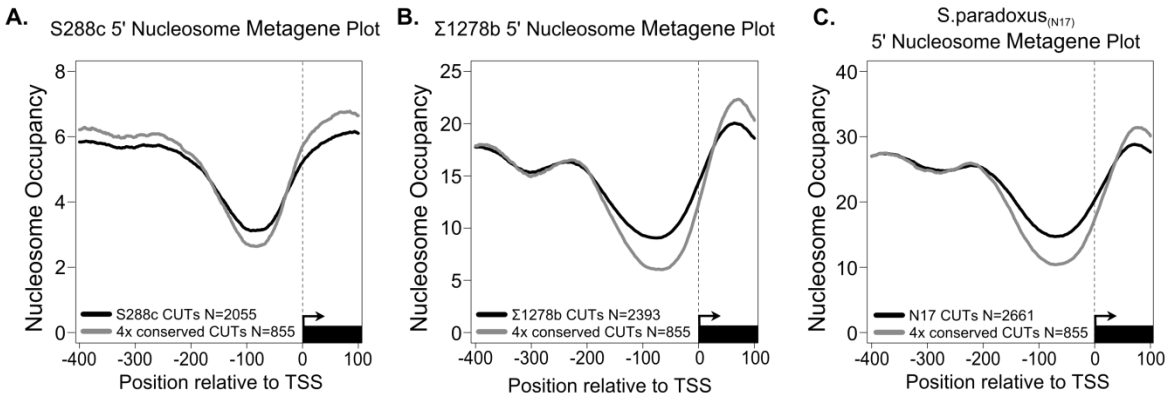


Figure 13 - 4x conserved CUTs show increased 5' nucleosome depletion relative to all CUTs
Metagene plot showing the average nucleosome occupancy in **A)** S288c, **B)** Σ1278b, and **C)** N17 of a 500bp window around the TSS for all CUTs identified by our HMM in the respective strain (black) and all 4x conserved CUTs as annotated in each respective strain (grey).

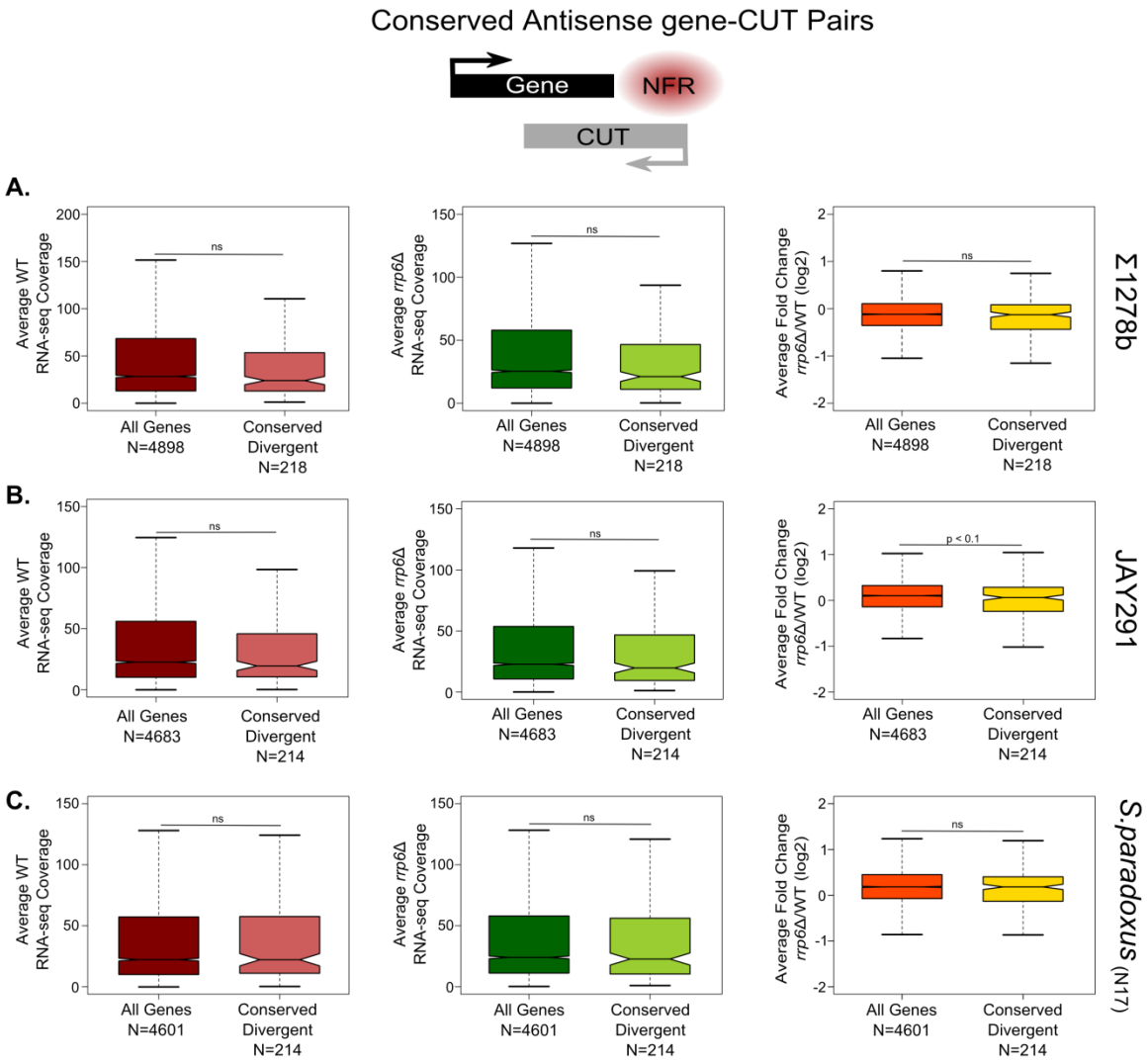


Figure 14 - Conserved antisense gene-CUT pairs in Σ1278b, JAY291, and *S.paradoxus*

Examination of antisense gene-CUT pairs containing a 4x conserved CUT. Box and whisker plots shows the distribution the average WT RNA-seq coverage (red), *rrp6*Δ RNA-seq coverage (green), log2 *rrp6*Δ/WT RNA-seq fold change (orange) for all expressed genes with a 3' UTR annotation and the subset of genes from antisense gene-CUT pairs with a 4x conserved CUT in A)Σ1278b, B)JAY291, and C)*S.paradoxus*. All points outside the whiskers (outliers) are not displayed. All p-values are derived from the two-sided KS test. Nonsignificant (ns) p-value $\geq$ 0.1.

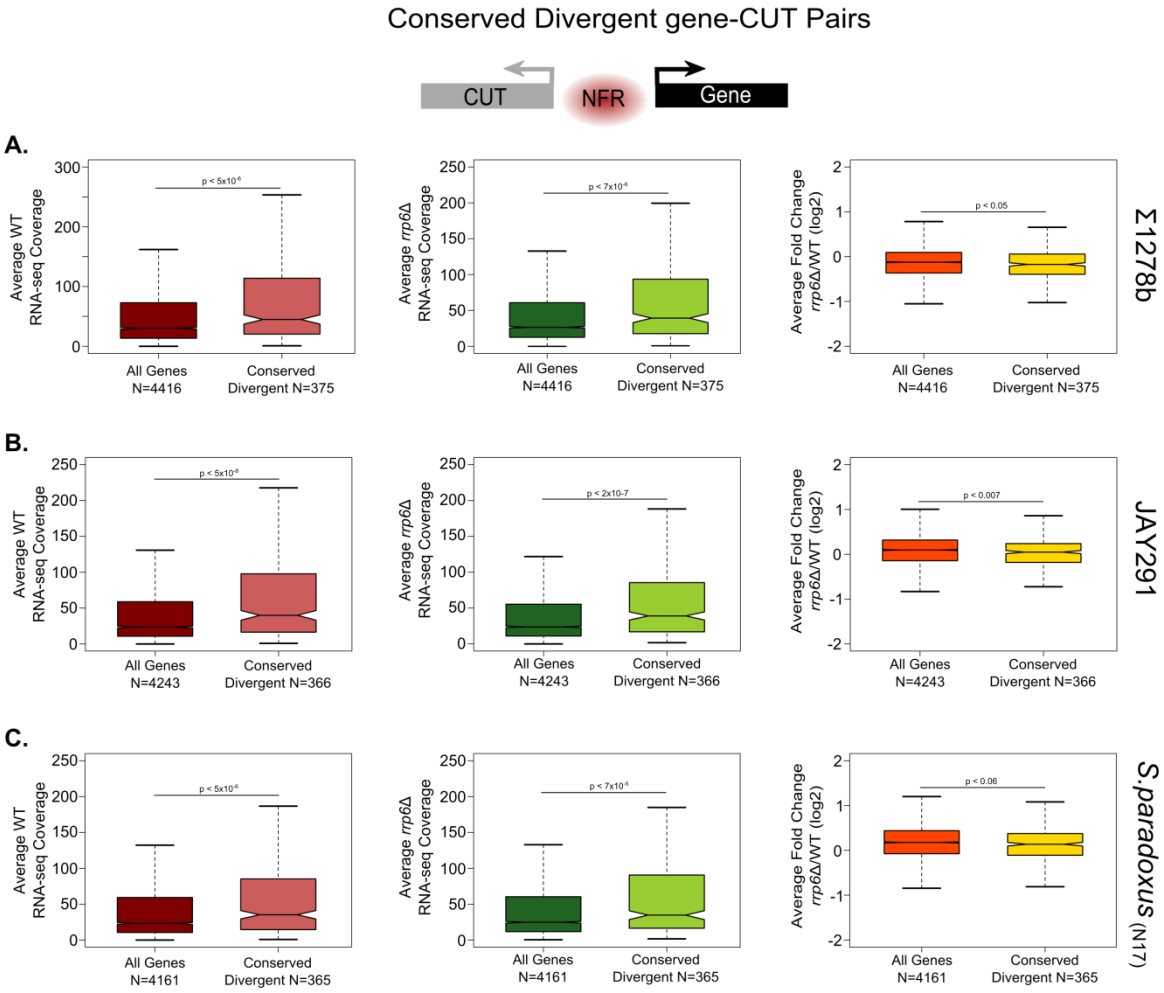


Figure 15 - Conserved divergent gene-CUT pairs in Σ1278b, JAY291, and *S.paradoxus*

Examination of divergent gene-CUT pairs containing a 4x conserved CUT. Box and whisker plots shows the distribution the average WT RNA-seq coverage (red), *rrp6*Δ RNA-seq coverage (green), log2 *rrp6*Δ/WT RNA-seq fold change (orange) for all expressed genes with a 5' UTR annotation and the subset of genes from antisense gene-CUT pairs with a 4x conserved CUT in **A)** Σ1278b, **B)** JAY291, and **C)** *S.paradoxus* (N17). All points outside the whiskers (outliers) are not displayed. All p-values are derived from the two-sided KS test.

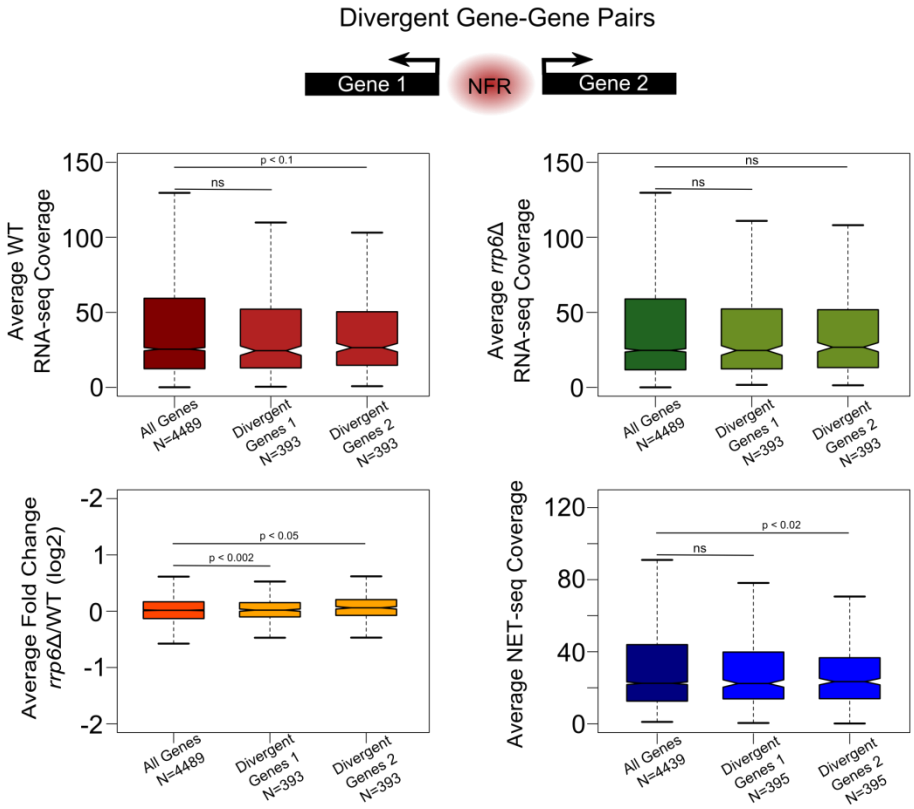


Figure 16 - Divergent gene-gene pairs in S288c
Examination of divergent gene-gene pairs in S288c. Box and whisker plots shows the distribution the average WT RNA-seq coverage (red), *rrp6*Δ RNA-seq coverage (green), log2 *rrp6*Δ/WT RNA-seq fold change (orange) for all expressed genes with a 5' UTR annotation and the subset of genes from gene-gene pairs. Gene set 1 and gene set 2 are derived from the two separate genes from each gene-gene pair where gene 1 is also on the crick strand as shown in the schematic. All points outside the whiskers (outliers) are not displayed. All p-values are derived from the two-sided KS test. Nonsignificant (ns) p-value ≥ 0.1.

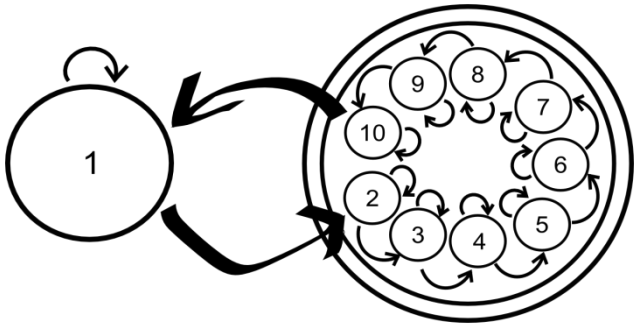


Figure 18 - 10-state explicit duration HMM
A state diagram of our explicit duration HMM showing expansion of state 2 into nine equivalent sub-states. The first state is parameterized to non-elevated regions of the transcriptome (i.e. not CUTs) and the remaining states are parameterized for elevated (approximately ≥ 2 fold) regions of the transcriptome (i.e. CUTs). We expanded the CUT state into nine identical sub-states with unidirectional movement through the model thereby setting the minimum length of a CUT to nine nucleotides and producing a 10-State model that approximates a hidden semi-Markov model (Datta et al. 2008).

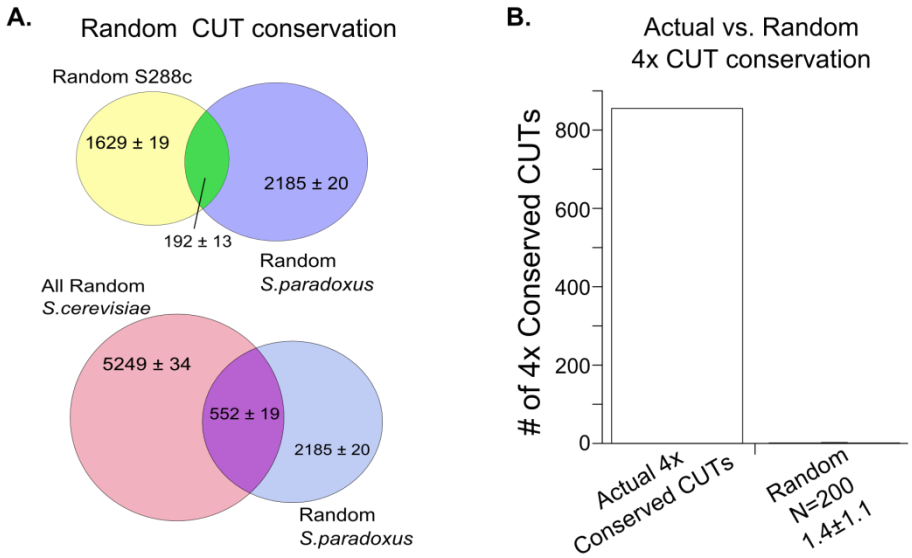


Figure 17 - Results of Randomized CUT Conservation Analysis
To determine the significance of our CUT conservation analysis we randomized CUT annotations in all four strains to assess the chance of CUT conservation simply by chance.
A.) Venn diagrams as seen in Figure 3C showing the average and standard deviation of conserved CUT expression between the S.cerevisiae strain S288c and S.paradoxus (N17) and the conserved CUT expression between all S.cerevisiae strains (S288c, Σ1278b, and JAY291) and S.paradoxus (N17) after 200 randomized trials. B.) Bar graph showing the actual total number of 4x conserved CUTs as found by our study and the average and standard deviation of the total number of 4x conserved CUTs after 200 randomized trials.

Tables

Table 1 - Divergent gene-CUT pairs enriched for metabolic process genes

Gene Ontology (GO)	p-value	# Genes
nucleobase-containing small molecule metabolic process [GO:0055086]	0.011139	81
phosphorus metabolic process [GO:0006793]	0.011852	125
phosphate-containing compound metabolic process [GO:0006796]	0.038567	120
organophosphate metabolic process [GO:0019637]	0.045294	77
carbohydrate derivative metabolic process [GO:1901135]	0.046665	70
alcohol biosynthetic process [GO:0046165]	0.049517	24
small molecule metabolic process [GO:0044281]	0.05444	133
ribose phosphate metabolic process [GO:0019693]	0.060834	39

A total of 698 divergent gene-CUT pairs were identified in S288c. The subset of genes in these gene-CUT pairs are enriched for various metabolic processing gene ontologies (GO). P values are based on the hypergeometric test after Holm-Bonferroni correction using the default background from YeastMine http://yeastmine.yeastgenome.org. The total number of genes in each GO category is listed far right.

Table 2 - Fold Change Conversion to Discrete Values

Bin	1	2	3	4	5	6	7	8	9
Fold Change	≤ 0.5	(0.5,1]	(1,1.25]	(1.25,1.75]	(1.75,3]	(3,4.5]	(4.5,6]	(6,10]	> 10

The Matlab HMM Toolkit only accepts discrete emission values. Per nucleotide *rrp6*Δ/WT fold change values were converted to a discrete value according to the table above.

Table 3 - HMM Emission Probabilities

		Fold Change Discrete Value								
Strain	**State**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
S288c	1	0.148289049	0.535045612	0.154105087	0.060321476	0.057662431	0.025783771	0.006114807	0.007648488	0.005029279
	2-10	0.005050152	0.006520196	0.015120454	0.017490525	0.235147054	0.217736532	0.188185646	0.16401492	0.150734522
Σ1278b	1	0.233759421	0.510798823	0.147520694	0.062861323	0.061962822	0.029996997	0.007887819	0.010687626	0.008718621
	2-10	0.007928039	0.006198992	0.014414585	0.018151635	0.251639689	0.252269069	0.241747529	0.228239262	0.260229245
JAY291	1	0.204011192	0.493874904	0.126978689	0.063365041	0.07728963	0.045661738	0.012141639	0.014895602	0.011005159
	2-10	0.006919117	0.005993605	0.012407379	0.018297087	0.313883999	0.384006578	0.372119472	0.318102571	0.328476752
N17	1	0.229862062	0.490867392	0.096278902	0.061062461	0.090558207	0.051315066	0.013296108	0.017747487	0.015367037
	2-10	0.007795859	0.005957106	0.009407632	0.017632202	0.367769548	0.431549995	0.407501888	0.379005913	0.458668008

The HMM emission probability for each discrete *rrp6*Δ/WT RNA-seq fold change value (see Table S3) for states 1-10. Because states 2-10 have the same emission probabilities we only show a single iteration of these emission probabilities for simplification.

Table 4 - **HMM Transition Probabilities**

State	1	2	3	4	5	6	7	8	9	10
1	0.999999637	3.63241E-07	0	0	0	0	0	0	0	0
2	0	0.9999999999991	9E-13	0	0	0	0	0	0	0
3	0	0	0.9999999999991	9E-13	0	0	0	0	0	0
4	0	0	0	0.9999999999991	9E-13	0	0	0	0	0
5	0	0	0	0	0.9999999999991	9E-13	0	0	0	0
6	0	0	0	0	0	0.9999999999991	9E-13	0	0	0
7	0	0	0	0	0	0	0.9999999999991	9E-13	0	0
8	0	0	0	0	0	0	0	0.9999999999991	9E-13	0
9	0	0	0	0	0	0	0	0	0.9999999999991	9E-13
10	9E-13	0	0	0	0	0	0	0	0	0.9999999999991

The HMM transition probabilities for states 1-10. Movement through the HM is unidirectional and only two transition probabilities exist for each state.

Chapter III - What Mechanisms Govern CUT Expression?

Introduction

The work discussed in this chapter pertains to analyses related to the work in Chapter II, but which were excluded from the publication on account of being preliminary findings. That CUTs are so rapidly degraded following transcription termination has led many to believe that CUTs are simply transcriptional noise, resulting from spurious, unchecked transcriptional activity at open chromatin (Wyers et al. 2005; Schulz et al. 2013). Supporters of this hypothesis cite high coincidence of CUT transcription originating from gene 5' and 3' NFRs as further motivation for this hypothesis. Likewise many attribute CUT transcription to a seemingly inherent bidirectional properties of yeast promoters (Xu et al. 2009; Neil et al. 2009). While it is clear that DNA must be accessible to RNA polymerase complexes for transcription to occur, there is inconclusive evidence of whether CUT expression is the indirect result of NFRs or is simply contributing the open chromatin formation. It is nearly impossible to distinguish between these two scenarios, but if we are to understand the function or role of CUT transcription in yeast we must also understand the mechanisms governing CUT expression. To this end I have leveraged strain unique CUTs, identified by my comparative analysis, to gain insights into these questions. These instances of unique CUT expression provide excellent opportunities to inform on the role of nucleosome positioning and sequence variation in regulating CUT expression by making cross strain comparisons and looking for variations that do or do not correlate with unique CUT expression.

As discussed previously (page 22) I have found that my RNA-seq/HMM method of CUT identification may have an appreciable false negative rate, and the work discussed in this chapter should be interpreted with this fact in mind. To minimize potential effects from falsely categorized unique CUTs in the analyses of this chapter, I have only considered those unique

CUTs that do not have corresponding syntenic raw HMM CUT calls in any of the other remaining strains[7]. In this way I have provided the best estimate of unique CUT expression in these four strains given the available data.

Results

I showed in Chapter II that CUTs have an NFR centered approximately 200bp upstream of the transcription start site (TSS) (**Figure 6**). Using publically available nucleosome occupancy data for S288c, Σ1278b, and N17 I assessed the 5' nucleosome occupancy profile of strain unique CUTs to determine if there are any inherent differences in 5' nucleosome positioning specific to unique CUTs. While unique CUTs also appear to have a 5' NFR, they show overall less nucleosome depletion within the NFR relative to the population average (**Figure 19**). This is in contrast to 4x conserved CUTs which show strong 5' nucleosome depletion within the NFR relative to the population average (**Figure 12**, **Figure 13**). Less nucleosome depletion in strain unique CUTs is consistent seen across S288c, Σ1278b, and N17 with Σ1278b showing the least amount of nucleosome depletion and the most deviation from the population average (**Figure 19**). While it is possible that strain unique CUTs have higher overall nucleosome occupancy within their promoters, I think we are instead seeing the effects of sampling error as the strain unique CUT populations within each strain are small in numbers.

If open chromatin is a primary driver of CUT expression then one could expect to see changes in nucleosome positioning that corresponds with the gain or loss of CUT expression. To determine if changes in nucleosome positioning correlate with strain unique CUTs I used my

---

[7] see *CUT identification* on page 30 for details regarding HMM post-processing steps

Pecan 4-way genome alignment[8] to compare 5' nucleosome occupancy in each strain background along the promoter region for a single set of strain unique CUTs. Remarkably there is strong conservation in the nucleosome occupancy profiles across all three strains (**Figure 20**) despite the CUT being present in only one strain. This is not surprising given that previous investigations comparing whole genome nucleosome profiles across 12 Hemiascomycota fungi (Tsankov et al. 2010) have also found nucleosome positioning to be conserved across deep evolutionary distances. But it does suggest that open chromatin is not a major factor governing CUT expression as we cannot attribute obvious changes in nucleosome occupancy to the gain or loss of CUT expression. To confirm that this result is not sensitive to the TSS annotations, which may be less accurate in the non-S288c strains, I compared 5' nucleosome occupancy of 41 S288c unique CUTs with accurate TSSs[9]. I again see conserved nucleosome depletion in both Σ1278b and N17 (**Figure 21**) further suggesting that open chromatin is not sufficient to initiate CUT transcription.

An important caveat to this comparison of nucleosome occupancy across strains is that the metagene plots used thus far are population averages, which may be obscuring more minute variations in CUT promoter nucleosome occupancy. To gain a better resolution of promoter nucleosome occupancy patterns, I clustered strain unique nucleosome occupancy profiles based upon their similarity to one another. By clustering CUTs with similar nucleosome occupancy profiles, distinct patterns emerge that were obscured within the population average (**Figure 22**). Some clusters, such as S288c cluster 1, Σ1278b cluster 7, and N17 cluster 1,

---

[8] See Chapter II for details regard the 4-way sequence alignment of the S288c, Σ1278b, JAY291, and N17 genomes

[9] Where the HMM 5' end call is within 50bp of a Malabat et al. 2015 TSS

show far greater variation in nucleosome occupancy within unique CUT promoters across strains (**Figure 23, Figure 24, Figure 25**). I would like to note that Σ1278b and N17 nucleosome occupancy profiles often show better correlation to one another than to S288c; this is most likely artefactual, due to the fact that the Σ1278b and N17 nucleosome occupancy data were collected in the same study but separately from S288c. Unfortunately differences in nucleosomal DNA isolation and data collection are likely introducing biases that are reflected in my correlation studies. While it is possible that the observed variations in nucleosome occupancy across strains, subtle though they may appear, are effecting CUT expression at these loci across my strain, I think my results suggest that nucleosome positioning is not a primary factor dictating CUT expression. If nucleosome positioning alone cannot explain strain unique CUT expression, then it may be that promoter sequence variation is major contributor to unique CUT expression.

As discussed in Chapter II, I assessed sequence conservation for 300bp and 50bp promoter regions for 4x conserved and unique CUTs. Unique CUTs appear to have a greater, but statistically nonsignificant, variation in promoter sequence conservation relative to 4x conserved CUTs (**Figure 11**) implicating sequence variation as a probable route to varied CUT expression. I used my unique CUT promoter nucleosome occupancy clusters to look for correlations in sequence variation relative to nucleosome positioning. For each set of strain unique CUTs I collected a histogram of all single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) unique to that particular strain relative to the three remaining strains. This histogram of SNPs and indels was plotted atop clustered nucleosome occupancy profiles revealing a subtle pattern of increased or localized strain-specific sequence variation within regions of low nucleosome occupancy (**Figure 23, Figure 24, Figure 25**). That strain-specific sequence variation is seen within regions of open chromatin hints to probable changes in transcription factor binding motif and suggests that the gain or loss of these motifs is a

contributing factor of unique CUT expression. Because these findings lack statistical significance, to confirm this finding I would need to be experimentally validated.

It may be possible to determine if any particular TF consensus motifs are disrupted by these strain-specific SNPs and indels, however the small number of sites will likely provide insufficient statistical power.

Unfortunately I have yet to determine is any of these strain-specific SNPs or indels have actually caused changes in TF consensus motifs. My preliminary findings suggest that CUT promoter sequence may be a primary factor dictating CUT expression and that small sequences changes such as SNPs and indels are driving the gain or loss of CUT expression across the four strains used in my study.

Discussion

What is the link between open chromatin and CUT expression? Many attribute CUT transcription to inherently bidirectional promoters and adventitious transcription at open chromatin (NFRs) due to a high coincidence of CUT transcription originating from gene 5' and 3' NFRs. I likewise observe that the majority of S288c CUTs identified by my HMM appear to originate from, or share their 5' NFR with, a gene 5' or 3' NFR. However many other S288c CUTs do not appear to originate from gene NFRs suggesting that CUTs can arise as independent transcriptional processes not linked to other transcriptional activity. Furthermore I did not observe any obvious changes in nucleosome positioning that could explain the gain or loss of strain unique CUT expression. While it is clear that CUT expression is strongly associated with protein-coding genes, these findings suggest that CUT expression does not simply occur in the presence of open chromatin and highlight our lack of understanding and appreciation for the mechanisms governing CUT expression.

Methods

Unique CUT Expression

See *Pecan whole genome alignment* on page 32 for details regarding the identification of unique and conserved syntenic CUT expression.

Nucleosome Occupancy and Metagene Analysis

See page 32 for details regarding acquisition and analysis of nucleosome occupancy data sets and design of metagene plots. For cross strain comparisons nucleosome occupancy data was normalized within each strain by dividing the occupancy at each nucleotide by the genomic occupancy mean. Normalized occupancy data was then converted from strain-specific genomic coordinates to 4-way Pecan alignment genomic coordinates.

Promoter Nucleosome Occupancy Profile Clustering

Using R and the DTW time series clustering package (Giorgino 2009) I performed hierarchical clustering of individual CUT 5' nucleosome occupancy profiles. The number of clusters produced for each set of strain unique CUTs was chosen subjectively based on cluster dendrograms. To ensure that clustering was performed based on relative nucleosome positioning, and not overall occupancy values, I normalized nucleosome occupancy for each individual CUT before clustering by setting the highest single nucleotide occupancy value to 1 scaling occupancy values the remaining occupancy values accordingly.

Figures

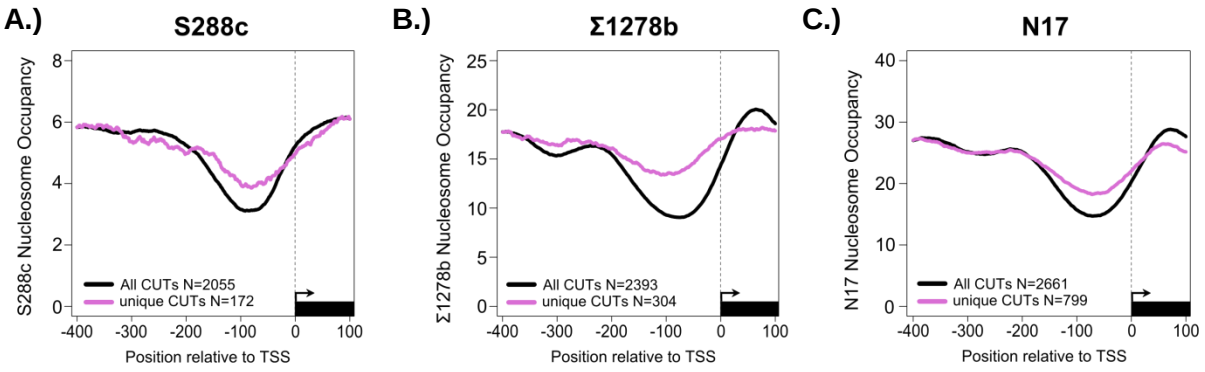
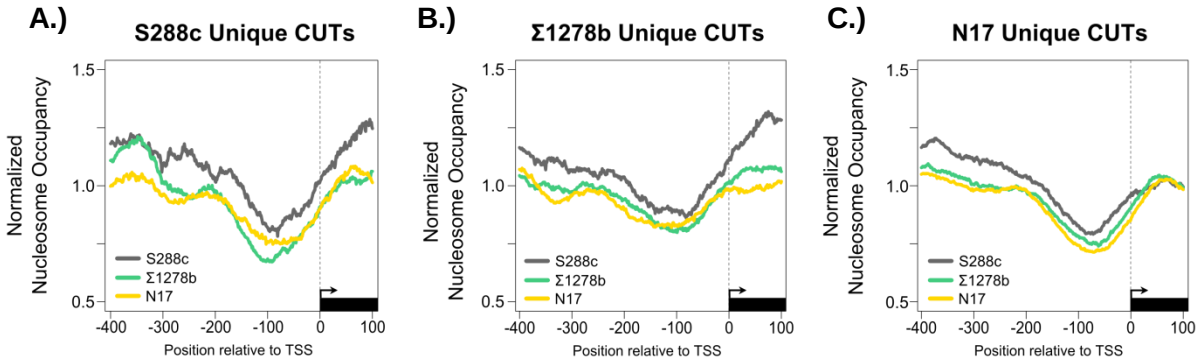


Figure 19 – Strain Unique CUTs Show Increased 5' Nucleosome Occupancy Relative to All CUTs
Metagene plot showing the average nucleosome occupancy in **A)** S288c, **B)** Σ1278b, and **C)** N17 of a 500bp window around the TSS for all CUTs identified by my HMM in the respective strain (black) and all CUTs unique to that strain (pink).



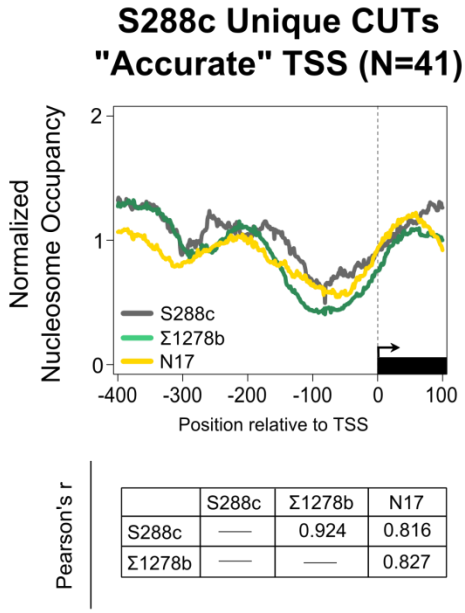
Pearson's r

	S288c	Σ1278b	N17
S288c	—	0.916	0.962
Σ1278b	—	—	0.926

	S288c	Σ1278b	N17
S288c	—	0.956	0.848
Σ1278b	—	—	0.905

	S288c	Σ1278b	N17
S288c	—	0.936	0.926
Σ1278b	—	—	0.992

Figure 20 – Cross Strain Nucleosome Occupancy is Highly Conserved Within the Promoters of Strain Unique CUTs
Metagene plot showing the average, normalized S288c (grey), Σ1278b (teal), and N17 (yellow) nucleosome occupancy within a 500bp window around the TSS for **A.)** S288c unqiue CUTs, **B.)** Σ1278b unique CUTs, and **C.)** N17 unique CUTs. Below each metagene plot are tables showing the pearson's r correlation coefficient calculated for all cross strain comparisons of the average, normalized nucleosome occupancy data in each metagene plot.



	S288c	Σ1278b	N17
S288c	——	0.924	0.816
Σ1278b	——	——	0.827

Figure 21 – Cross Strain Conservation of Unique CUT 5' Nucleosome Occupancy Is Not an Artefact of Inaccurate TSS Annotations
Metagene plot showing the average, normalized S288c (grey), Σ1278b (teal), and N17 (yellow) nucleosome occupancy within a 500bp window around the TSS only for S288c unique CUTs with an HMM TSS annotation that is within 50bp of a Malabat et al. 2015 TSS. Below the metagene plot is tables showing the pearson's r correlation coefficient calculated for all cross strain comparisons of the average, normalized nucleosome occupancy data shown in the metagene plot.

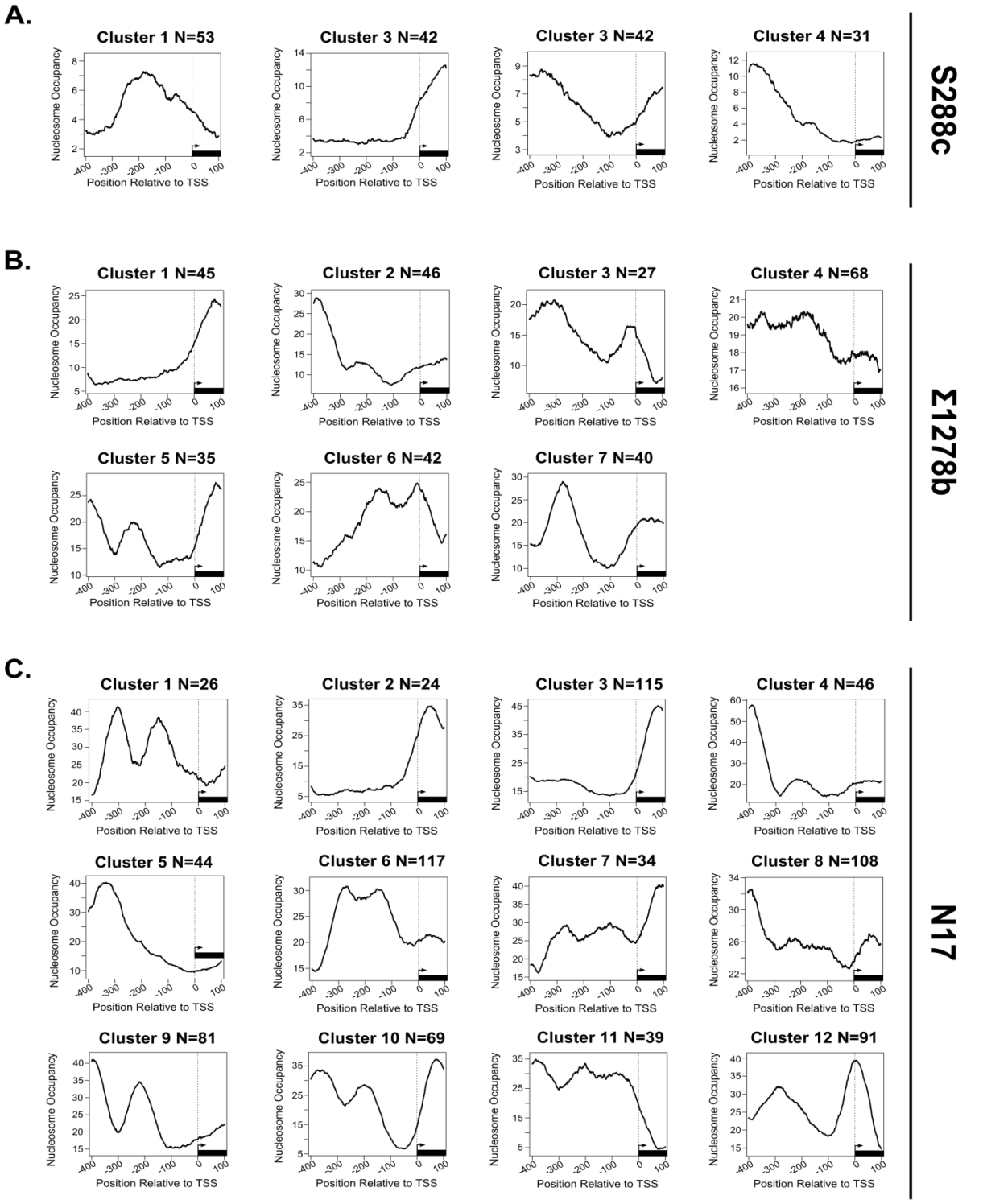
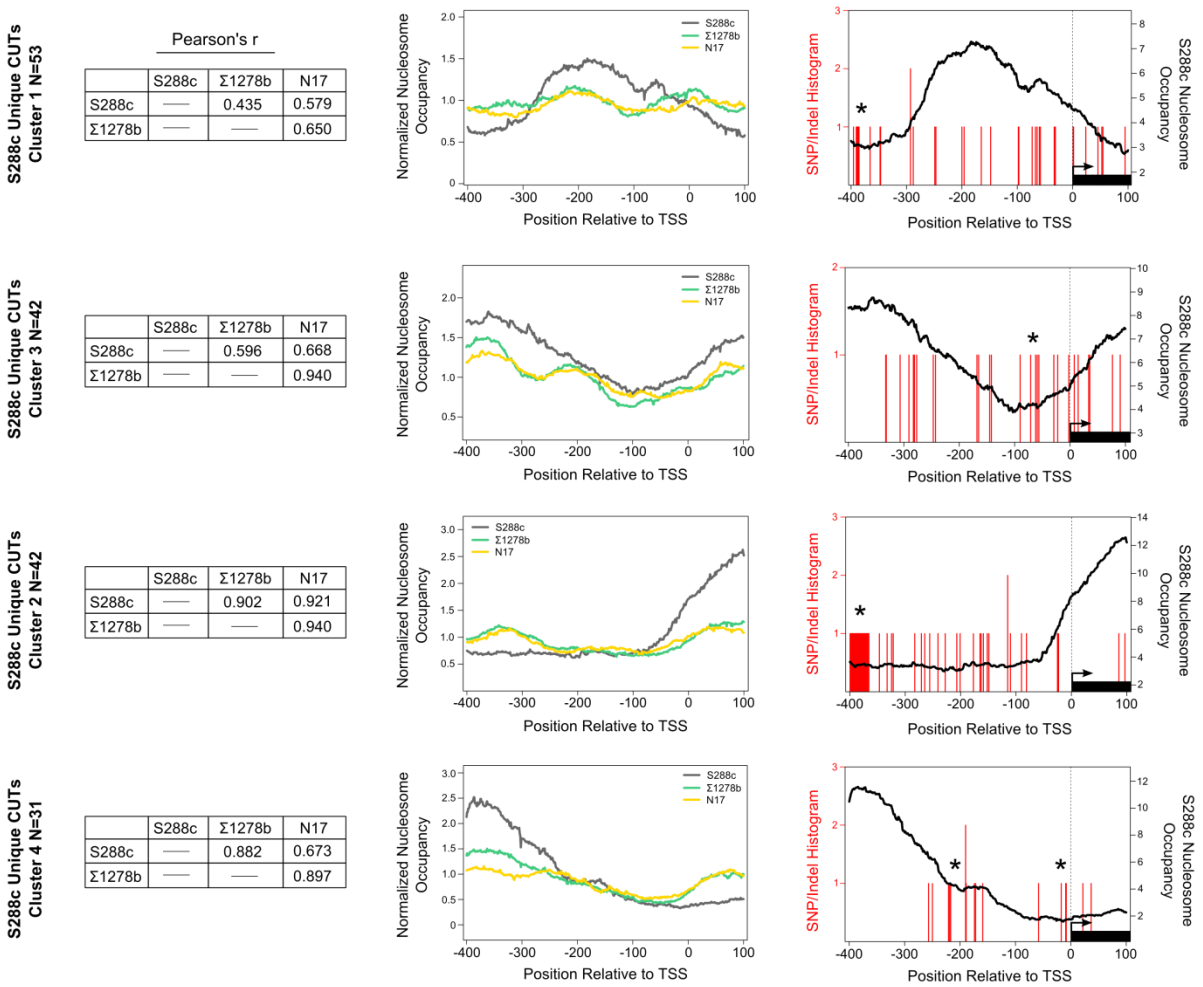


Figure 22 –Unique CUT 5' Nucleosome Occupancy Clusters
The individual 5' nucleosome occupancy profiles of **A.)** S288c, **B.)** Σ1278b, and **C.)** N17 unique CUTs were clustered based upon their similarity to one another revealing distinct patterns of promoter nucleosome occupancy that were obscured within the population average.



	S288c	Σ1278b	N17
S288c	—	0.435	0.579
Σ1278b	—	—	0.650

	S288c	Σ1278b	N17
S288c	—	0.596	0.668
Σ1278b	—	—	0.940

	S288c	Σ1278b	N17
S288c	—	0.902	0.921
Σ1278b	—	—	0.940

	S288c	Σ1278b	N17
S288c	—	0.882	0.673
Σ1278b	—	—	0.897

Figure 23 – S288c Unique CUT 5' Nucleosome Occupancy Clusters Show Greater Cross Strain Nucleosome Occupancy Variation

Middle Column: Metagene plots showing the average, normalized S288c (grey), Σ1278b (teal), and N17 (yellow) nucleosome occupancy within a 500bp window around the TSS for each S288c unique CUT 5' nucleosome occupancy cluster**. Left Column:** A table showing the pearson's r correlation coefficient calculated for all cross strain comparisons of the average, normalized nucleosome occupancy data within each S288c unique CUT 5' nucleosome occupancy cluster. **Right Column:** Metagene plots showing S288c nucleosome occupancy (black) within a 500bp window around the TSS for each S288c unique CUT 5' nucleosome occupancy cluster overlaid with a histogram of S288c-specific SNPs and indels.

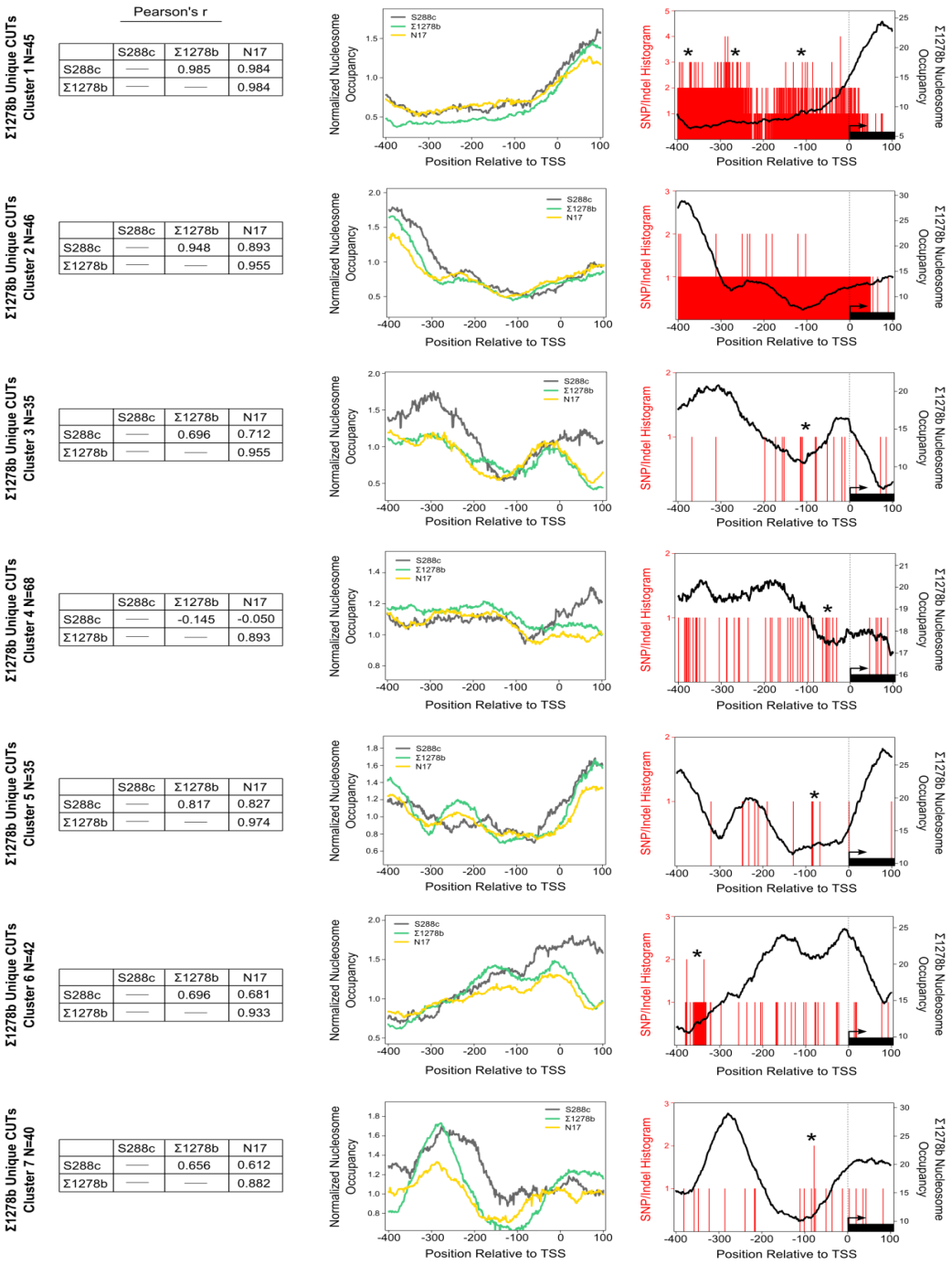


Figure 24 – Σ1278b Unique CUT 5' Nucleosome Occupancy Clusters Show Greater Cross Strain Nucleosome Occupancy Variation

Middle Column: Metagene plots showing the average, normalized S288c (grey), Σ1278b (teal), and N17 (yellow) nucleosome occupancy within a 500bp window around the TSS for each S288c unique CUT 5' nucleosome occupancy cluster. **Left Column:** A table showing the Pearson's r correlation coefficient calculated for all cross strain comparisons of the average, normalized nucleosome occupancy data within each Σ1278b unique CUT 5' nucleosome occupancy cluster. **Right Column:** Metagene plots showing Σ1278b nucleosome occupancy (black) within a 500bp window around the TSS for each Σ1278b unique CUT 5' nucleosome occupancy cluster overlaid with a histogram of Σ1278b-specific SNPs and indels.

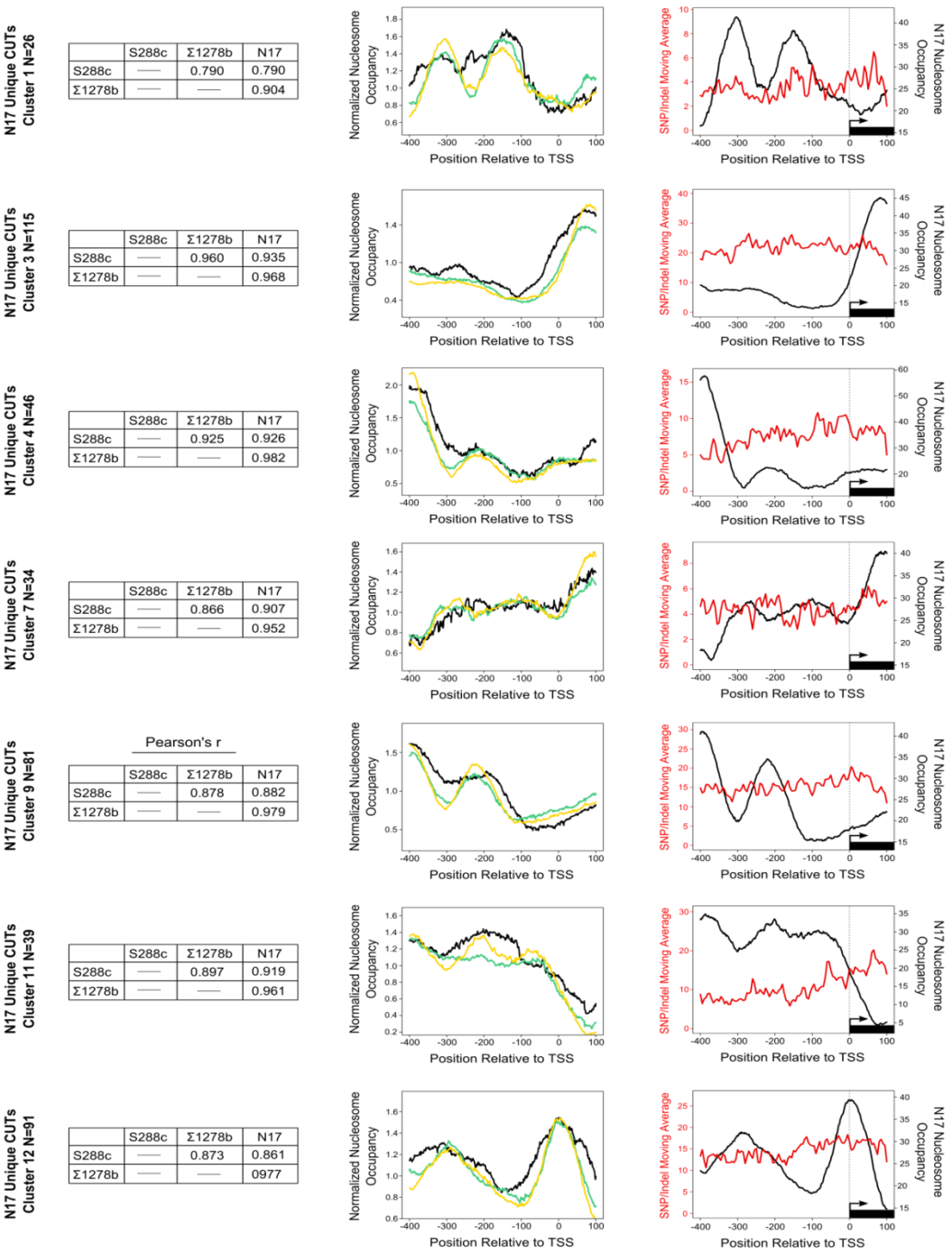


Figure 25 – N17 Unique CUT 5' Nucleosome Occupancy Clusters Show Greater Cross Strain Nucleosome Occupancy Variation

Middle Column: Metagene plots showing the average, normalized S288c (grey), Σ1278b (teal), and N17 (yellow) nucleosome occupancy within a 500bp window around the TSS for each N17 unique CUT 5' nucleosome occupancy cluster**. Left Column:** A table showing the pearson's r correlation coefficient calculated for all cross strain comparisons of the average, normalized nucleosome occupancy data within each N17 unique CUT 5' nucleosome occupancy cluster. **Right Column:** Metagene plots showing N17 nucleosome occupancy (black) within a 500bp window around the TSS for each N17 unique CUT 5' nucleosome occupancy cluster overlaid with a moving average of N17-specific SNPs and indels histogram. A moving average was used in place of a histogram to more clearly show changes in SNP/Indel frequency. Only six of the 12 N17 clusters are shown for simplification.

Chapter IV - Assessment of Nascent CUT Expression by NET-qPCR

Introduction

Global and evolutionary studies into CUTs, such as described in Chapter II, produce broad reaching general insights into CUT expression and its role within the yeast transcriptome. However more in-depth case studies focused on individual CUTs are needed if we are to confirm and exemplify CUT-based regulation of gene expression. While there is growing support in the literature documenting regulatory functions for CUT expression, the functional basis of CUTs remains highly debated and largely unexplored. Two examples of CUT-based gene regulation are well documented in the literature: NRD1 and URA2. The NRD1 gene is self-regulating through a negative feedback loop that is dependent on the Nrd1-Nab3-Sen1 (NNS) termination complex. When Nrd1p level are high, NRD1 transcription terminates early, via Sen1-dependent termination, creating a CUT. However when Nrd1p levels are low there is increased read through of Nrd1 and Nab3 terminator motifs and full length NRD1 mRNAs are produced (Arigo et al. 2006). These full length mRNAs are translated into functional Nrd1p increasing the Nrd1p concentration in the cell and increasing early termination of NRD1 until Nrd1p levels drop again and full length NRD1 mRNAs are needed. NRD1 autoregulation is not only a unique example of CUT-based gene regulation but also an example of transcriptional attenuation, which is a common regulatory strategy in bacteria but is not commonly observed in eukaryotes. PCF11 and RPB10 were recently identified as possible targets of Nrd1p attenuation (Creamer et al. 2011) but follow-up studies are needed to verify these findings.

URA2 is another example of CUT-based gene regulation. URA2 encodes a pyrimidine biosynthesis enzyme and its expression is regulated by cellular concentrations of uracil, such that URA2 is lowly expressed in the presence of uracil and upregulated when uracil is low or absent. URA2 expression is not dependent on the Ppr1p transcriptional activator, a factor that

controls other genes in the same pathway, nor have other transcription factors been implicated in URA2 expression. However a CUT, usURA2[10], is transcribed upstream of and overlapping with the URA2 5' UTR, initiating approximately 90bp upstream of the URA2 transcription start site (TSS_{URA2}) (**Figure 26**) (Thiebaut et al. 2008). Similar promoter architectures are also seen at URA8, IMD2, and ADE12 suggesting that these genes may be regulated by CUT-based mechanisms similar to URA2. Mutating the usURA2 transcription start site (TSS_{usURA2}) inhibits usURA2 transcription and results in constitutive upregulated expression of URA2, suggesting that usURA2 transcription represses URA2 expression in non-inducing, uracil-replete conditions (Thiebaut et al. 2008). The authors hypothesized that usURA2 expression would decrease to allow for upregulation of URA2 when uracil is low, however surprisingly usURA2 transcription was not observed to decrease in during URA2 upregulation; Instead, the authors concluded that repression of URA2 by usURA2 depends on two factors: 1.) a "weaker" TSS_{usURA2} more proximal to a TATA-box shared by a "stronger" distal TSS_{URA2} and 2.) an AT-rich region between the TSS_{usURA2} and TSS_{URA2} dubbed the R-box (**Figure 26**) that contains Sen1-dependent terminator sequences that ensure early termination and degradation of usURA transcripts. Upregulation of URA2 is thought to occur via unidentified factors or mechanisms that promote read-through of TSS_{usURA2} allowing productive transcription starting at TSS_{URA2}. Alternatively the R-box may act as an internal promoter to direct preinitiation complex assembly to selectively enhance transcription of URA2 in low uracil growth conditions.

Work on URA2/usURA2 demonstrates how CUT transcription may be involved in regulating gene expression, and poignantly illustrates a lack in understanding of the various mechanisms involved in CUT-based gene regulation. Likewise the work of Thiebaut et al. 2008

---

[10] upstream sense URA2

also highlights potential methodological flaws in the study of CUTs. The observations of usURA2 expression in both uracil+ and uracil- conditions were made using northern blots of steady-state RNA in both WT and *trf4*Δ backgrounds. When analyzing steady-state RNA, transcript quantification is dependent on both expression and degradation rates and it is not possible, without the appropriate controls, to attribute transcript changes specifically to one process over the other. Assessing changes in CUT expression is particularly difficult given that mutant backgrounds used for their detection disrupt RNA steady-state dynamics and artificially cause CUTs to accumulate by inhibiting degradation. It is not known how long stabilized CUTs persist in the cell thus potentially rendering any decreases in CUT transcription undetectable in the steady-state RNA pool. I hypothesized that usURA2 transcript levels were not observed to change (decrease) in uracil- conditions relative to uracil+ conditions because the northern blot assay used by Thiebaut et al. 2008 only assess steady state RNA levels, which are disrupted by *trf4*Δ. Instead the usURA2 signal observed in the northern blot may pertain to transcripts made in uracil+ conditions that failed to be degraded and persisted in uracil- conditions. What the authors had actually hypothesized was a change in nascent usURA2 transcription, but did not assess nascent transcription with their northern blot assay. To avoid complications arising from mutant, CUT stabilizing backgrounds and to direction assess changes in nascent expression of URA2 and usURA2 I propose to adapt a new method in yeast, called nascent elongating transcript sequencing (NET-seq) (Churchman and Weissman 2011), for downstream analysis by qPCR (NET-qPCR).

Results

First I established the growth conditions for low, basal URA2 expression and for upregulated URA2 expression. To ensure ample uracil availability synthetic complete (SC) medium with twice the standard amount of uracil (at 40ug/mL) was used as a control for low,

basal URA2 expression. SC without uracil (SC-ura) medium was used to activate URA2 upregulation. Once the cell begins to produce its own uracil URA2 expression is downregulated, but URA2 upregulation can be sustained with the addition of 6-azauracil (6AU). 6AU is a competitive uracil antagonist that inhibits Ura3p, an enzyme that functions downstream of URA2 in the uracil biosynthesis pathway. **Figure 28** shows URA2 expression by RT-qPCR under control and inducing growth conditions, where URA2 expression upregulated 1.8 fold in SC-ura over SC control and is upregulated 9.5 fold in SC-ura with the addition of 6AU at 10mg/mL. These expression trends are in keeping with published results under similar growth conditions (Potier et al. 1990).

Using cryofixation, NET-seq isolates nascent RNA by immunoprecipitation elongating RNA Pol II, via 3x-Flag tagged Rpb3p, and the associated, nascent RNA. I isolated both total RNA by hot acid phenol and nascent RNA in yJV001 (see **Appendix A** for strain table) grown in SC and SC-ura/6AU from biological duplicates. I only selected SC and SC-ura/6AU because of the robust up regulation of URA2 in the later growth condition. Nascent RNA preps showed an approximately 5-fold reduction in 18S rRNA compared to total RNA preps suggesting that my preps had good enrichment of nascent, RNAP II RNA (**Figure 27**). Nascent URA2 and usURA2 expression was assessed by qPCR (i.e. NET-qPCR) (**Figure 29**). While I observed a 6.6 fold increase in nascent URA2 expression in SC-ura/6AU relative to control, I did not observe a change in usURA2 expression. Though I did not directly compare nascent usURA2 expression to steady-state expression in WT or *rrp6*Δ backgrounds, I can say that usURA2 NET-qPCR consistently yielded cycle thresholds ~ 5 cycles earlier than usURA2 RT-qPCR in either steady-state RNA sample, confirming an enrichment of CUTs in nascent RNA over steady-state RNA even in *rrp6*Δ.

Discussion

NET-qPCR is able to directly quantitate nascent expression in yeast making it possible to assess nascent CUT expression without the use of mutant, CUT stabilizing backgrounds. With NET-qPCR I am able to detect upregulated nascent URA2 expression in SC-ura growth conditions however I did not observe a change in usURA2 nascent expression. Though usURA2 expression appears to be inhibiting URA2 expression, in this instance, the simplest explanation for URA2 upregulation is not via downregulation of usURA2. It is possible that mutating the TSS_{usURA2} disrupted motifs necessary for URA2 repression in SC+uracil conditions, independent of inhibiting usURA2 expression. This would negate a role for the usURA2 CUT in regulating URA2 expression. However given the thorough investigation of Thiebaut et al. 2008, it is more likely that URA2 regulation does indeed involve usURA2 and is more complex than a simple on-off switch. Although I did not observed the expected expression trends in usURA2, NET-qPCR will be useful to future studies of CUT-based gene regulation, even if just to confirm results found in mutant backgrounds such as trf4Δ or *rrp6*Δ. Alternatively, NET-qPCR may be advantageous in cases where the slow growth, temperature-sensitive, or nonviable[11] phenotypes of CUT stabilizing mutants make some experiments difficult to perform.

Materials and Methods

Strains

All work in this chapter used strain yJV001. This strain is a derivative of the original NET-seq, RPB3-3xFLAG-NAT1 strain that was kindly provided by Dr. Stirling Churchman. This

[11] Ndr1 is essential requiring the use of temperature sensitive mutants or depletion techniques like anchor-way

strain was transformed with wildtype URA3 to allow for growth in SC-uracil medium. See **Appendix A** for complete strain details.

Total RNA Isolation

Cells were grown in synthetic complete medium (SC) with 40mg/mL uracil, SC without uracil, or SC without uracil with 6-azauracil at 10mg/mL to an OD of 0.6. 6-azauracil was not added until cells reach OD 0.3 because inhibits growth in these condition. Total RNA was isolated via hot acid phenol method and DNAse treated for two hours with Promega DNAse RQ1 to remove contaminating DNA.

Nascent RNA Isolation via NET-seq Method

The yeast NET-seq protocol was kindly provided by Dr. Stirling Churchman. A complete, detailed version of the NET-seq protocol is available in the literature (Churchman and Weissman 2012). Briefly, 1L of cells were grown in either synthetic complete medium (SC) with 40mg/mL uracil or SC without uracil with 6-azauracil at 10mg/mL to an OD of 0.6. 6-azauracil was not added until cells reach OD 0.3 because inhibits growth in these condition. Cells were pelleted and cryofixed in liquid nitrogen. Frozen pellets were cryoground using a Retsch Mortar Grinder RM100. One gram of ground cells was suspended in lysate buffer, clarified by centrifugation, and immunoprecipitated with Sigma Anti-Flag M2 Affinity Gel. RNAP II complexes were eluted from the anti-flag beads with 2mg/mL 3×-Flag peptide. Nascent RNA was isolated from the eluate using the Qiagen miRNeasy kit. Before qPCR, RNA was DNAse treated for two hours with Promega DNAse RQ1 to remove contaminating DNA.

cDNA Synthesis and qPCR

For all cDNA samples, 1ug of RNA was reverse transcribed with Fermentas Maxima RT using random hexamers. qPCR was performed with SYBR green fluorescence reporter dye. ACT1 served as an endogenous control.

Primer Sequences

Primer sequences can be found in Appendix B. Because usURA2 is transcribed into URA2, leaving only ~90bp of sequence (Rbox) unique to usURA2, extra care was taken to design primers that amply within this unique stretch of usURA2.

Figures

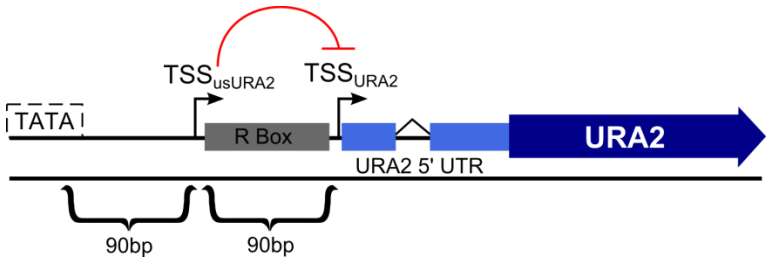


Figure 26 - **URA2 Promoter Architecture**
usURA2 is transcribed from a TSS ~90bp upstream of TSSURA2 and terminates within the URA2 5'UTR intron (~500bp). Both transcripts depend on a TATA box ~90bp upstream of TSSusURA2. The T-rich DNA region between the two TSSs is the R box. usURA2 expression in thought to inhibit URA2 expression. A similar promoter architecture is observed at other nucleotide biosynthesis genes (URA8, ADE12, IMD2, and IMD3). Image not drawn to scale.

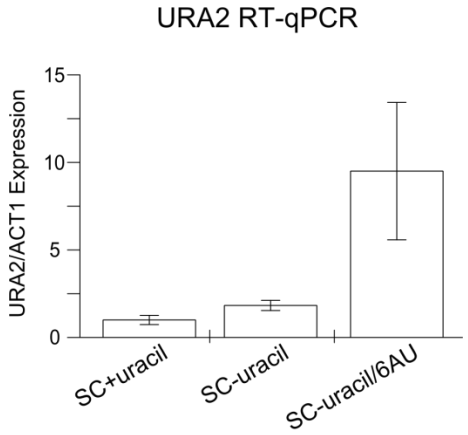


Figure 28 – URA2 expression is upregulated in the absence of uracil
yJV001 yeast were grown in synthetic complete (SC) with 40mg/mL uracil (SC+uracil), SC without uracil (SC-uracil), and SC-uracil with 10mg/mL 6AU (SC-uracil/6AU) in biological triplicates. Total RNA was isolated and URA2 expression was assessed by RT-qPCR. URA2 expression was normalized to ACT1 and expression levels are set relative to URA2 expression in SC+uracil. Error bars denote standard deviation of URA2 expression by coefficient of variation calculations.

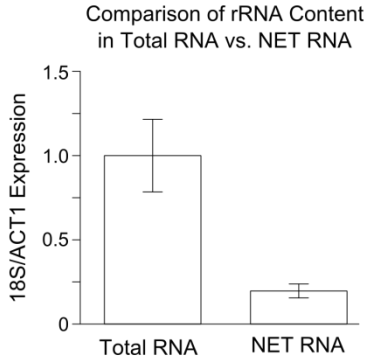


Figure 27 – Nascent RNA Preps are depleted of rRNA
Both total and nascent RNA were isolated from cells grown in synthetic complete (SC) with 40mg/mL uracil and SC-uracil with 10mg/mL 6AU (SC-uracil/6AU) in biological duplicates. 18S rRNA was assessed by RT-qPCR. Results from total RNA and nascent RNA were grouped respectively and show a consistent ~5 fold depletion of 18S rRNA in nascent RNA relative to total RNA preps. 18S rRNA levels were normalized to ACT1 and set relative to total RNA signal. Error bars denote standard deviation of URA2 expression by coefficient of variation calculations.

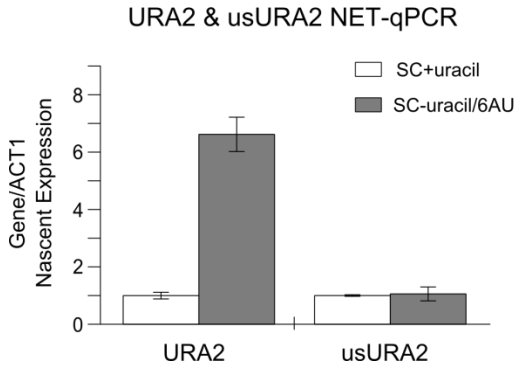


Figure 29 – NET-qPCR of the CUT usURA2 Upon Activation of URA2 Expression
Nascent RNA was isolated from cells grown in synthetic complete (SC) with 40mg/mL uracil and SC-uracil with 10mg/mL 6AU (SC-uracil/6AU) in biological duplicates. Nascent expression of URA2 (white bars) and the CUT usURA2 (grey bars) was assessed by NET-qPCR. URA2 expression is upregulated in SC-uracil/6AU relative to SC+uracil but usURA2 nascent expression remains unchanged. Both URA2 and usURA2 expression levels were normalized to nascent ACT1 expression and set relative to expression in SC+uracil. Error bars denote standard deviation by coefficient of variation calculations.

Chapter V - Conclusion

Major Conclusions from This Work

The work presented in this thesis has outlined two methods, NET-qPCR and HMM analysis of RNA-seq, for the detection and identification of cryptic unstable transcripts. NET-qPCR allows for direct assessment of nascent CUT expression and obviates the need for mutant, CUT stabilizing backgrounds that may otherwise confound some analyses. My HMM allows for global identification of CUTs with RNA-seq data from wild-type and *rrp6*Δ RNA samples. With my HMM/RNA-seq method I have greatly expanded upon previous CUT annotations in the *S.cerevisiae* reference strain S288c, and though I suspect my HMM of generating a conservative estimate of CUT expression, my findings still show that CUT expression is far more extensive than previous estimates. I have also generated the first set of CUT annotations outside of the strain S288c by applying my HMM method to other strains of yeast. In doing so, I have provided the first evolutionary comparison of CUT expression in yeast. Using conservative estimates of CUT expression I have shown that syntenic CUT expression is well conserved between the species *S.cerevisiae* and *S.paradoxus*.

By identifying populations of both conserved and unique syntenic expression in my four strains I have had a unique opportunity to inform on the mechanisms underlying CUT expression. Not surprisingly, it appears that CUT promoter sequence conservation may be a primary factor dictating CUT expression. While 4x conserved CUTs do not show preferential sequence conservation within the body of the CUT transcript, I did observe preferential sequence conservation of 4x conserved CUT promoters (**Figure 11**). One interpretation of this observation is that the CUT sequence itself is less important than the promoter sequence for 4x conserved CUTs, consistent with a model where CUT expression, not the CUT itself, is functionally important. These observations coincide with increased sequence variation within

the promoters of strain unique CUTs (**Figure 11**). Looking nucleosome occupancy within the promoters of strain unique CUTs I found that the gain or loss of CUT expression cannot be explained by obvious changes in nucleosome positioning. Instead I observed a high frequency of SNPs and indels within regions of low nucleosome occupancy, suggesting that sequence changes in *cis*, possibly affecting TF binding, are a major factor influencing CUT expression.

Using the CUTs identified by my HMM I have gained insights into possible functional roles for CUT expression in yeast. My work has demonstrated that antisense CUT expression, originating from the 3' NFR of the associated gene, can elicit a negative effect on sense gene transcription (**Figure 12C**) in a manner consistent with traditional antisense transcripts. Furthermore I have shown that bidirectional gene-CUT expression correlates with higher levels of gene expression (**Figure 12D**) demonstrating that bidirectional CUT expression may act to aid or promote gene expression, possibly by helping to maintain an open promoter conformation. Remarkably this trend of higher gene expression is not observed for bidirectional gene-gene pairs, demonstrating that this effect of higher gene expression is specific to instances of bidirectional transcription involving CUTs (**Figure 16**). These findings have implications for how unstable RNAs, not just necessarily CUTs, may contribute to regulating gene expression, further demonstrating that the plethora of unstable RNAs found in both yeast and humans should not be overlooked despite their unstable nature.

Using publically available nucleosome occupancy data for S288c, Σ1278b, and N17 I have shown that Nrd1-Nab3-Sen1 terminated CUTs lack a 3' nucleosome free region (NFR) that is commonly observed at the 3' end of transcripts that undergo poly(A)-dependent termination (**Figure 6**, **Figure 7**). As one might expect based on my findings, stable ncRNAs, which show evidence of occasional Sen1-dependent termination in addition to poly(A)-dependent termination, only show moderate 3' nucleosome depletion (**Figure 8**), producing a 3' nucleosome occupancy profile that looks almost like a hybrid between CUTs and protein-coding

genes. This distinct pattern of 3' nucleosome occupancy is an additional distinguishing characteristic of Sen1-dependent termination. Despite the extensive work on Sen1-dependent termination in recent years this distinct nucleosome occupancy pattern has surprisingly been overlooked in the field. That we only see a 3' NFR at protein-coding genes hints at a possible role for mRNA-specific termination sequences and factors in 3' NFR production and maintenance.

While the study of unstable transcripts remains a burgeoning field, overall I believe my work will prove instrumental to future studies into CUTs, providing a framework to better understand the role of unstable RNAs in the yeast transcriptome.

Limitations of This Work

Of the four strains used in my comparative analysis, both JAY291 and N17 have incomplete, low coverage genome assemblies. Lacking complete genomic sequences for these strains reduced the amount of 4-way aligned sequence in the Pecan alignment, limiting my searchable space for syntenic CUT expression. As my assessment of syntenic CUT expression was limited to only 4-way aligned regions my estimates of conserved CUT expression were made even more conservative. While the reference strain S288c is thoroughly curated and has numerous publically available data sets and annotations, the same cannot be said of my remaining strains: Σ1278b, JAY291, and N17. For instance there is no publically available JAY291 nucleosome occupancy data. Therefore I had to exclude JAY291 from analyses involving nucleosome occupancy. Likewise JAY291 and N17 have 500+ fewer protein coding gene annotations compared to S288c and Σ1278b. These missing gene annotations limit my ability to detect all instances of antisense and bidirectional CUT expression and possibly limit the statistical power of the analyses shown in **Figure 14**, **Figure 15**. Furthermore, a complete lack of transcription start site and transcription termination site annotations in Σ1278b, JAY291,

and N17 prevented me from independently identifying antisense and bidirectional gene-CUT pairs in these strains. Instead my analysis was limited to antisense and bidirectional gene-CUT pairs identified in S288c which were extrapolated to the remaining strains via conserved CUT expression.

At the onset of this project there was little indication regarding the evolutionary divergence of CUT expression in closely related species of yeast. While I have shown that CUT expression is well conserved between S.cerevisiae and S.paradoxus, I have been unable to make assertions regarding the relationship between sequence conservation and CUT expression given the limited evolutionary depth of my comparative analysis. Though it appears that CUT promoter sequence may be a primary factor dictating CUT expression, it is clear now that I cannot rule out the possibility that conserved CUT expression is an indirect consequence of sequence conservation due to the limited amount of sequence divergence between *S.cerevisiae* and *S.paradoxus.*

How This Work Relates To the Work of Others

CUTs were first globally identified in 2005 by Wyers et al. who were seeking to identify new targets of the nuclear exosome. At the time it was known that the nuclear exosome was responsible for processing and maturation of rRNA and sn/snoRNAs and turnover of introns and aberrant pre-mRNAs (Petfalski et al. 1998; Allmang et al. 1999). With the advent of microarrays it was possible to take an unbiased approach to find new targets of the nuclear exosome upon deletion of RRP6. While the work of Wyers et al. 2005 provided the first global assessment of CUT expression, but was just a snapshot of CUT expression, limited to the probes of the ORFeome microarray used in their experiment. Xu et al. 2009 expanded upon the work of Wyers et al. 2005 by using a whole genome tiling array providing CUT annotations that are still commonly used today. However as discussed in Chapter II, microarrays have a limited

detection range, often struggling to detect low abundance (**Figure 3**, **Figure 5**) and/or AT-rich transcripts. By utilizing traditional RNA sequencing to detect CUT expression I have not only greatly expanded upon previous CUT annotations in S288c, but I have also shown CUT expression to be far more pervasive than previously imagined. Furthermore I have applied my RNA-seq/HMM method to not one, but four strains of yeast, providing the most extensive CUTs annotations currently available. My work is a natural progression in the study of CUTs, advancing the field by applying modern next-generation sequencing technology.

Early studies of CUT expression focused on characterizing the genomic organization of CUTs and the extensive association of CUTs with protein-coding genes. When I began my project it was well known that CUTs largely originate from gene 5' and 3' NFRs (Xu et al. 2009; Neil et al. 2009). My work corroborates the earlier findings of others but also makes it clear that not all CUTs originate from a gene 5' or 3' NFR nor is the presence of an NFRs sufficient for CUT expression. Little work has actually been done to understand why CUTs are transcribed, not in the sense of whether they are functional or not, but in the sense of what cryptic DNA elements lead to CUT expression. What sequences are common among cryptic promoters? Why do some gene promoters generate bidirectional transcription involving CUTs? What can CUTs tell us about the specificity and regulation of RNAP II? My work has only begun to address some of these questions. In terms of understanding whether CUTs or CUT expression are functional, there are but a handful of examples demonstrating CUT-based regulation of gene expression[12] (Arigo et al. 2006; Thiebaut et al. 2008) while previous global CUT studies have only speculated on possible modes of CUT-based gene expression. My work has gone a step further to demonstrate possible activating and inhibiting functions for CUT expression. My

[12] See Chapter IV for details on these examples

work is the first global evidence of CUT-based regulation of gene expression. While it is well documented that antisense transcription can reduce or inhibit sense transcription both globally (Xu et al. 2011) and in well studied instances (e.g. IME4 Hongay et al. 2006), CUTs have been largely ignored as possible sources of transcriptional interference. While most instances of ncRNA-based gene regulation are believed to repress the corresponding gene, there is growing evidence in the literature for an activating role of ncRNAs. A recent publication studying bidirectional transcription in mice observed a decrease in mRNA levels upon transient knockdown of corresponding promoter-associated divergent ncRNAs (Uesaka et al. 2014). This result is consistent with the idea that bidirectional expression is promoting or activating gene expression. Why this trend of activating bidirectional transcription appears specific to coding-noncoding transcript pairs remains unclear. Though most studies of ncRNA-based gene regulation have focused on stable ncRNAs, my work suggests that an RNA need not be stable to confer an effect on the expression of corresponding genes.

Future Directions

These findings warrant further investigation of CUT expression at greater evolutionary depth to better determine at what rate CUT expression is gained and lost. Likewise it would be important to know if the effects of antisense and bidirectional CUT expression on the expression of associated genes hold up at greater evolutionary distances. Expanding this comparative analysis to include additional yeast strains and species would greatly improve our understanding of the relationship between sequence conservation and conserved CUT expression. If CUT expression is found to be conserved even in the presence of significant promoter sequence variation that would suggest that CUT expression was actively maintained despite underlying sequence changes. This would be an important example demonstrating that conserved CUT expression is not simply an unintentional consequence of sequence

conservation and would go a long way in supporting a functional role for CUT expression in yeast.

As discussed in Chapter III, instances of unique CUT expression provide excellent opportunities for cross strain comparisons, helping to inform on the role of nucleosome positioning and sequence variation in regulating CUT expression. Continued analyses utilizing examples of strain unique CUT expression should first set out to validate a number of strain unique CUTs by RT-qPCR. Not only would validation of strain unique CUT expression better inform on the false negative rate of our RNA-seq/HMM method of CUT identification but it would also provide suitable candidates for more in depth case studies. So far no one particular motif or transcription factor has been implicated in regulating gene expression. I suspect that any activating TF could induce CUT expression, thus making it difficult to identify motifs with statistically significant enrichment within CUT promoters. Instead, with validated examples of unique CUT expression one could conduct promoter bashing experiments to identify the sequences important for expression of the candidate CUTs. Furthermore, it would be intriguing to swap out unique CUT promoters between strains to determine if these promoters are sufficient for inducing CUT expression in the other strains, or if strain-specific *trans* factors are also needed to elicit CUT expression.

Although I have demonstrated potential regulatory functions for CUT expression, I have not determined if CUT expression is sufficient to induce the observed effects on gene expression. It may be that CUT expression needs to surpass some threshold in order to elicit an effect (Xu et al. 2011). To address these questions future analyses regarding the effects of antisense and bidirectional CUT expression should incorporate gene and CUT expression levels. Future analyses should also include all instances of antisense CUT expression, not just where antisense CUT expression originates from gene 3' NFRs, to determine if there is a minimum amount of sense-antisense overlap required for sense inhibition or if transcription

through the 3' NFR is important for eliciting transcription interference. Furthermore, instances of transcriptional interference by antisense CUT expression would make excellent candidates for in depth studies of CUT-based regulation of gene expression. Unlike bidirectional gene-CUT pairs, antisense gene-CUT pairs do not share a promoter; therefore CUT expression can be attenuated with promoter sequence alterations while largely avoiding changes to the promoter sequence of the sense gene.

A number of follow-up experiments are needed to further demonstrate the connection between Sen1-dependent termination and high 3' nucleosome occupancy. Yeast stable ncRNAs would be useful resources for future experiment as at least some of these ncRNAs already show evidence of both Sen1-dependent and poly(A)-dependent termination. These stable ncRNAs should be grouped based on sensitivity to *rrp6*Δ [13] to identify the population that most utilizes Sen1-dependent termination in addition to poly(A)-dependent termination. By comparing the 3' nucleosome occupancy profiles of *rrp6*Δ-sensitive and *rrp6*Δ-insensitive populations of stable ncRNAs I would expect to see a correlation in 3' nucleosome occupancy related to utilization of Sen1-dependent termination. If the *rrp6*Δ-sensitive stable ncRNAs showed greater 3' nucleosome occupancy than rrp6Δ-insensitive ncRNAs that would provide further evidence that 3' nucleosome occupancy is largely dictated by termination pathways. Furthermore it would be interesting to add or remove Sen1-dependent terminator sequences (i.e. the Nrd1 and Nab3 motifs) within stable ncRNAs to see if it were possible to produce changes in 3' nucleosome occupancy by driving increased or decreased usage of Sen1-dependent termination.

---

[13] i.e. increased RNA-seq coverage in *rrp6*Δ relative to WT

References

Allmang, C., Joanna Kufel, G.F. Chanfreau, Philip Mitchell, Elisabeth Petfalski, and David Tollervey. 1999. “Functions of the Exosome in rRNA, snoRNA and snRNA Synthesis.” *The EMBO Journal* 18 (19): 5399–5410. doi:10.1093/emboj/18.19.5399.

Aloy, P., Bettina Böttcher, H. Ceulemans, C Leutwein, C. Mellwig, S. Fischer, AC Gavin, et al. 2004. “Structure-Based Assembly of Protein Complexes in Yeast.” *Science* 303 (5666): 2026–29. doi:10.1126/science.1092645.

Aloy, P., F.D. Ciccarelli, C. Leutwein, AC Gavin, Giulio Superti-Furga, Peer Bork, Bettina Böttcher, and R.B. Russel. 2002. “A Complex Prediction: Three-Dimensional Model of the Yeast Exosome.” *EMBO Reports* 3 (7): 628–35. doi:10.1093/embo-reports/kvf135.

Anderson, J. S.J., and R. Parker. 1998. “The 3’ to 5’ Degradation of Yeast mRNAs Is a General Mechanism for mRNA Turnover That Requires the SKI2 DEVH Box Protein and 3’ to 5’ Exonucleases of the Exosome Complex.” *The EMBO Journal* 17 (5): 1497–1506. doi:10.1093/emboj/17.5.1497.

Argueso, J. L., M. F. Carazzolle, P. A. Mieczkowski, F. M. Duarte, O. V.C. Netto, S. K. Missawa, F. Galzerani, et al. 2009. “Genome Structure of a Saccharomyces Cerevisiae Strain Widely Used in Bioethanol Production.” *Genome Research* 19 (12): 2258–70. doi:10.1101/gr.091777.109.

Arigo, John T., Kristina L. Carroll, Jessica M. Ames, and Jeffry L. Corden. 2006. “Regulation of Yeast NRD1 Expression by Premature Transcription Termination.” *Molecular Cell* 21 (5): 641–51. doi:10.1016/j.molcel.2006.02.005.

Arigo, John T., Daniel E. Eyler, Kristina L. Carroll, and Jeffry L. Corden. 2006. “Termination of Cryptic Unstable Transcripts Is Directed by Yeast RNA-Binding Proteins Nrd1 and Nab3.” *Molecular Cell* 23 (6): 841–51. doi:10.1016/j.molcel.2006.07.024.

Bonneau, Fabien, Jérôme Basquin, Judith Ebert, Esben Lorentzen, and Elena Conti. 2009. “The Yeast Exosome Functions as a Macromolecular Cage to Channel RNA Substrates for Degradation.” *Cell* 139 (3): 547–59. doi:10.1016/j.cell.2009.08.042.

Briggs, M. W., K. T. D. Burkard, and J. S. Butler. 1998. “Rrp6p, the Yeast Homologue of the Human PM-Scl 100-kDa Autoantigen, Is Essential for Efficient 5.8 S rRNA 3’ End Formation.” *Journal of Biological Chemistry* 273 (21): 13255–63. doi:10.1074/jbc.273.21.13255.

Butler, J. Scott, and Phil Mitchell. 2010. “Rrp6, Rrp47 and Cofactors of the Nuclear Exosome.” In *RNA Exosome*, edited by Torben Heick Jensen, 702:91–104. New York, NY: Springer US. http://link.springer.com/10.1007/978-1-4419-7841-7_8.

Callahan, K. P., and J. S. Butler. 2010. “TRAMP Complex Enhances RNA Degradation by the Nuclear Exosome Component Rrp6.” *Journal of Biological Chemistry* 285 (6): 3540–47. doi:10.1074/jbc.M109.058396.

Carroll, K. L., D. A. Pradhan, J. A. Granek, N. D. Clarke, and J. L. Corden. 2004. “Identification of Cis Elements Directing Termination of Yeast Nonpolyadenylated snoRNA Transcripts.” *Molecular and Cellular Biology* 24 (14): 6241–52. doi:10.1128/MCB.24.14.6241-6252.2004.

Carrozza, Michael J., Bing Li, Laurence Florens, Tamaki Suganuma, Selene K. Swanson, Kenneth K. Lee, Wei-Jong Shia, et al. 2005. “Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription.” *Cell* 123 (4): 581–92. doi:10.1016/j.cell.2005.10.023.

Chan, Yujia A., Maria J. Aristizabal, Phoebe Y. T. Lu, Zongli Luo, Akil Hamza, Michael S. Kobor, Peter C. Stirling, and Philip Hieter. 2014. “Genome-Wide Profiling of Yeast DNA:RNA Hybrid Prone Sites with DRIP-Chip.” Edited by Michael Snyder. *PLoS Genetics* 10 (4): e1004288. doi:10.1371/journal.pgen.1004288.

Chen, Ching-Yi, Roberto Gherzi, Shao-En Ong, Edward L. Chan, Reinout Raijmakers, Ger J.M. Pruijn, Georg Stoecklin, Christoph Moroni, Matthias Mann, and Michael Karin. 2001. “AU Binding Proteins Recruit the Exosome to Degrade ARE-Containing mRNAs.” *Cell* 107 (4): 451–64. doi:10.1016/S0092-8674(01)00578-5.

Chen, Ying-Zhang, Sayed H. Hashemi, Susan K. Anderson, Yongzhao Huang, Maria-Ceu Moreira, David R. Lynch, Ian A. Glass, Phillip F. Chance, and Craig L. Bennett. 2006. “Senataxin, the Yeast Sen1p Orthologue: Characterization of a Unique Protein in Which Recessive Mutations Cause Ataxia and Dominant Mutations Cause Motor Neuron Disease.” *Neurobiology of Disease* 23 (1): 97–108. doi:10.1016/j.nbd.2006.02.007.

Churchman, L. Stirling, and Jonathan S. Weissman. 2011. “Nascent Transcript Sequencing Visualizes Transcription at Nucleotide Resolution.” *Nature* 469 (7330): 368–73. doi:10.1038/nature09652.

2012. “Native Elongating Transcript Sequencing (NET-Seq).” In *Current Protocols in Molecular Biology*, edited by Frederick M. Ausubel, Roger Brent, Robert E. Kingston, David D. Moore, J.G. Seidman, John A. Smith, and Kevin Struhl. Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.wiley.com/10.1002/0471142727.mb0414s98.

Core, L. J., J. J. Waterfall, and J. T. Lis. 2008. “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.” *Science* 322 (5909): 1845–48. doi:10.1126/science.1162228.

Cramer, P., K.-J. Armache, S. Baumli, S. Benkert, F. Brueckner, C. Buchen, G.E. Damsma, et al. 2008. “Structure of Eukaryotic RNA Polymerases.” *Annual Review of Biophysics* 37 (1): 337–52. doi:10.1146/annurev.biophys.37.032807.130008.

Creamer, Tyler J., Miranda M. Darby, Nuttara Jamonnak, Paul Schaughency, Haiping Hao, Sarah J. Wheelan, and Jeffry L. Corden. 2011. “Transcriptome-Wide Binding Sites for Components of the Saccharomyces Cerevisiae Non-Poly(A) Termination Pathway: Nrd1, Nab3, and Sen1.” Edited by Nick J. Proudfoot. *PLoS Genetics* 7 (10): e1002329. doi:10.1371/journal.pgen.1002329.

Datta, Ritendra, Jianying Hu, and Bonnie Ray. 2008. “On Efficient Viterbi Decoding for Hidden Semi-Markov Models.” In , 1–4. IEEE. doi:10.1109/ICPR.2008.4761926.

David, Lior, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J. Palm, Lee Bofkin, Ted Jones, Ronald W. Davis, and Lars M. Steinmetz. 2006. “A High-Resolution Map of Transcription in the Yeast Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (14): 5320–25. doi:10.1073/pnas.0601091103.

Davis, C. A., and M. Ares. 2006. “Accumulation of Unstable Promoter-Associated Transcripts upon Loss of the Nuclear Exosome Subunit Rrp6p in Saccharomyces Cerevisiae.” *Proceedings of the National Academy of Sciences* 103 (9): 3262–67. doi:10.1073/pnas.0507783103.

Dichtl, B., D. Blank, M. Sadowski, W. Hubner, S. Weiser, and W. Keller. 2002. “Yhh1p/Cft1p Directly Links poly(A) Site Recognition and RNA Polymerase II Transcription Termination.” *The EMBO Journal* 21 (15): 4125–35. doi:10.1093/emboj/cdf390.

Dowell, R. D., O. Ryan, A. Jansen, D. Cheung, S. Agarwala, T. Danford, D. A. Bernstein, et al. 2010. “Genotype to Phenotype: A Complex Problem.” *Science* 328 (5977): 469–469. doi:10.1126/science.1189015.

Dziembowski, Andrzej, Esben Lorentzen, Elena Conti, and Bertrand Séraphin. 2007. “A Single Subunit, Dis3, Is Essentially Responsible for Yeast Exosome Core Activity.” *Nature Structural & Molecular Biology* 14 (1): 15–22. doi:10.1038/nsmb1184.

Egecioglu, D. E., A.K. Henras, and G.F. Chanfreau. 2006. “Contributions of Trf4p- and Trf5p-Dependent Polyadenylation to the Processing and Degradative Functions of the Yeast Nuclear Exosome.” *RNA* 12 (1): 26–32. doi:10.1261/rna.2207206.

Eick, Dirk, and Matthias Geyer. 2013. “The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code.” *Chemical Reviews* 113 (11): 8456–90. doi:10.1021/cr400071f.

Engel, S. R., F. S. Dietrich, D. G. Fisk, G. Binkley, R. Balakrishnan, M. C. Costanzo, S. S. Dwight, et al. 2014. “The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now.” *G3: Genes|Genomes|Genetics* 4 (3): 389–98. doi:10.1534/g3.113.008995.

Fasken, Milo B., R. Nicholas Laribee, and Anita H. Corbett. 2015. “Nab3 Facilitates the Function of the TRAMP Complex in RNA Processing via Recruitment of Rrp6 Independent of Nrd1.” Edited by J. Scott Butler. *PLOS Genetics* 11 (3): e1005044. doi:10.1371/journal.pgen.1005044.

Field, Yair, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K. Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. 2008. “Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals.” Edited by Uwe Ohler. *PLoS Computational Biology* 4 (11): e1000216. doi:10.1371/journal.pcbi.1000216.

Fox, Melanie J., Hongyu Gao, Whitney R. Smith-Kinnaman, Yunlong Liu, and Amber L. Mosley. 2015. “The Exosome Component Rrp6 Is Required for RNA Polymerase II Termination at Specific Targets of the Nrd1-Nab3 Pathway.” Edited by Jeffry Corden. *PLOS Genetics* 10 (2): e1004999. doi:10.1371/journal.pgen.1004999.

Frenk, Stephen, David Oxley, and Jonathan Houseley. 2014. “The Nuclear Exosome Is Active and Important during Budding Yeast Meiosis.” Edited by Jürg Bähler. *PLoS ONE* 9 (9): e107648. doi:10.1371/journal.pone.0107648.

Gelfand, B., J. Mead, A. Bruning, N. Apostolopoulos, V. Tadigotla, V. Nagaraj, A. M. Sengupta, and A. K. Vershon. 2011. “Regulated Antisense Transcription Controls Expression of Cell-Type-Specific Genes in Yeast.” *Molecular and Cellular Biology* 31 (8): 1701–9. doi:10.1128/MCB.01071-10.

Giaever, Guri, Angela M. Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, et al. 2002. “Functional Profiling of the Saccharomyces cerevisiae Genome.” *Nature* 418 (6896): 387–91. doi:10.1038/nature00935.

Giorgino, Toni. 2009. “Computing and Visualizing Dynamic Time Warping Alignments in *R* : The **dtw** Package.” *Journal of Statistical Software* 31 (7). doi:10.18637/jss.v031.i07.

Gudipati, Rajani Kanth, Zhenyu Xu, Alice Lebreton, Bertrand Séraphin, Lars M. Steinmetz, Alain Jacquier, and Domenico Libri. 2012. “Extensive Degradation of RNA Precursors by the Exosome in Wild-Type Cells.” *Molecular Cell* 48 (3): 409–21. doi:10.1016/j.molcel.2012.08.018.

Hainer, S. J., J. A. Pruneski, R. D. Mitchell, R. M. Monteverde, and J. A. Martens. 2011. “Intergenic Transcription Causes Repression by Directing Nucleosome Assembly.” *Genes & Development* 25 (1): 29–40. doi:10.1101/gad.1975011.

Hieronymus, H., M.C. Yu, and P.A. Silver. 2004. “Genome-Wide mRNA Surveillance Is Coupled to mRNA Export.” *Genes & Development* 18 (21): 2652–62. doi:10.1101/gad.1241204.

Hobor, F., R. Pergoli, K. Kubicek, D. Hrossova, V. Bacikova, M. Zimmermann, J. Pasulka, C. Hofr, S. Vanacova, and R. Stefl. 2011. “Recognition of Transcription Termination Signal by the Nuclear Polyadenylated RNA-Binding (NAB) 3 Protein.” *Journal of Biological Chemistry* 286 (5): 3645–57. doi:10.1074/jbc.M110.158774.

Holub, P., J. Lalakova, H. Cerna, J. Pasulka, M. Sarazova, K. Hrazdilova, M. S. Arce, F. Hobor, R. Stefl, and S. Vanacova. 2012. “Air2p Is Critical for the Assembly and RNA-Binding of the TRAMP Complex and the KOW Domain of Mtr4p Is Crucial for Exosome Activation.” *Nucleic Acids Research* 40 (12): 5679–93. doi:10.1093/nar/gks223.

Hongay, Cintia F., Paula L. Grisafi, Timothy Galitski, and Gerald R. Fink. 2006. “Antisense Transcription Controls Cell Fate in Saccharomyces Cerevisiae.” *Cell* 127 (4): 735–45. doi:10.1016/j.cell.2006.09.038.

Houseley, Jonathan, Liudmilla Rubbi, Michael Grunstein, David Tollervey, and Maria Vogelauer. 2008. “A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster.” *Molecular Cell* 32 (5): 685–95. doi:10.1016/j.molcel.2008.09.027.

Houseley, Jonathan, and David Tollervey. 2006. “Yeast Trf5p Is a Nuclear poly(A) Polymerase.” *EMBO Reports* 7 (2): 205–11. doi:10.1038/sj.embor.7400612.

2008. “The Nuclear RNA Surveillance Machinery: The Link between ncRNAs and Genome Structure in Budding Yeast?” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1779 (4): 239–46. doi:10.1016/j.bbagrm.2007.12.008.

Huang, Huan, Jiao Chen, Hongde Liu, and Xiao Sun. 2013. “The Nucleosome Regulates the Usage of Polyadenylation Sites in the Human Genome.” *BMC Genomics* 14 (1): 912. doi:10.1186/1471-2164-14-912.

Huang, Huan, Hongde Liu, and Xiao Sun. 2013. “Nucleosome Distribution near the 3′ Ends of Genes in the Human Genome.” *Bioscience, Biotechnology and Biochemistry* 77 (10): 2051–55. doi:10.1271/bbb.130399.

Huang, Y. 2001. “Comparison of the RNA Polymerase III Transcription Machinery in Schizosaccharomyces Pombe, Saccharomyces Cerevisiae and Human.” *Nucleic Acids Research* 29 (13): 2675–90. doi:10.1093/nar/29.13.2675.

Jade Bernstein, and Eric A Toth. 2012. “Yeast Nuclear RNA Processing.” *World J Biol Chem* 3 (1): 7–26.

Kadaba, S., Xuying Wang, and James T Robinson. 2006. “Nuclear RNA Surveillance in Saccharomyces Cerevisiae: Trf4p-Dependent Polyadenylation of Nascent Hypomethylated tRNA and an Aberrant Form of 5S rRNA.” *RNA* 12 (3): 508–21. doi:10.1261/rna.2305406.

Kaplan, Noam, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, et al. 2009. “The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome.” *Nature* 458 (7236): 362–66. doi:10.1038/nature07667.

Kellis, Manolis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. 2003. “Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements.” *Nature* 423 (6937): 241–54. doi:10.1038/nature01644.

Kim, Minkyu, Nevan J. Krogan, Lidia Vasiljeva, Oliver J. Rando, Eduard Nedea, Jack F. Greenblatt, and Stephen Buratowski. 2004. “The Yeast Rat1 Exonuclease Promotes Transcription Termination by RNA Polymerase II.” *Nature* 432 (7016): 517–22. doi:10.1038/nature03041.

Kim, Minkyu, Lidia Vasiljeva, Oliver J. Rando, Alexander Zhelkovsky, Claire Moore, and Stephen Buratowski. 2006. “Distinct Pathways for snoRNA and mRNA Termination.” *Molecular Cell* 24 (5): 723–34. doi:10.1016/j.molcel.2006.11.011.

Krishnamurthy, S., M. A. Ghazy, C. Moore, and M. Hampsey. 2009. “Functional Interaction of the Ess1 Prolyl Isomerase with Components of the RNA Polymerase II Initiation and Termination Machineries.” *Molecular and Cellular Biology* 29 (11): 2925–34. doi:10.1128/MCB.01655-08.

Kubicek, K., H. Cerna, P. Holub, J. Pasulka, D. Hrossova, F. Loehr, C. Hofr, S. Vanacova, and R. Stefl. 2012. “Serine Phosphorylation and Proline Isomerization in RNAP II CTD Control Recruitment of Nrd1.” *Genes & Development* 26 (17): 1891–96. doi:10.1101/gad.192781.112.

LaCava, John, Jonathan Houseley, Cosmin Saveanu, Elisabeth Petfalski, Elizabeth Thompson, Alain Jacquier, and David Tollervey. 2005. “RNA Degradation by the Exosome Is Promoted by a Nuclear Polyadenylation Complex.” *Cell* 121 (5): 713–24. doi:10.1016/j.cell.2005.04.029.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Lebreton, Alice, Rafal Tomecki, Andrzej Dziembowski, and Bertrand Séraphin. 2008. “Endonucleolytic RNA Cleavage by a Eukaryotic Exosome.” *Nature* 456 (7224): 993–96. doi:10.1038/nature07480.

Levin, Joshua Z, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. 2010. “Comprehensive Comparative Analysis of Strand-Specific RNA Sequencing Methods.” *Nat Meth* 7 (9): 709–15. doi:10.1038/nmeth.1491.

Licatalosi, Donny D, Gabrielle Geiger, Michelle Minet, Stephanie Schroeder, Kate Cilli, J.Bryan McNeil, and David L Bentley. 2002. “Functional Interaction of Yeast Pre-mRNA 3′ End Processing Factors with RNA Polymerase II.” *Molecular Cell* 9 (5): 1101–11. doi:10.1016/S1097-2765(02)00518-X.

Liu, Quansheng, Jaclyn C. Greimann, and Christopher D. Lima. 2006. “Reconstitution, Activities, and Structure of the Eukaryotic RNA Exosome.” *Cell* 127 (6): 1223–37. doi:10.1016/j.cell.2006.10.037.

Luo, W., A.W. Johnson, and D.L. Bentely. 2006. “The Role of Rat1 in Coupling mRNA 3’-End Processing to Transcription Termination: Implications for a Unified Allosteric-Torpedo Model.” *Genes & Development* 20 (8): 954–65. doi:10.1101/gad.1409106.

Makino, Debora Lika, Benjamin Schuch, Elisabeth Stegmann, Marc Baumgärtner, Claire Basquin, and Elena Conti. 2015. “RNA Degradation Paths in a 12-Subunit Nuclear Exosome Complex.” *Nature* 524 (7563): 54–58. doi:10.1038/nature14865.

Malabat, Christophe, Frank Feuerbach, Laurence Ma, Cosmin Saveanu, and Alain Jacquier. 2015. “Quality Control of Transcription Start Site Selection by Nonsense-Mediated-mRNA Decay.” *eLife* 4 (April). doi:10.7554/eLife.06722.

Martens, Joseph A., Lisa Laprade, and Fred Winston. 2004. “Intergenic Transcription Is Required to Repress the Saccharomyces Cerevisiae SER3 Gene.” *Nature* 429 (6991): 571–74. doi:10.1038/nature02538.

Mischo, Hannah E., Belén Gómez-González, Pawel Grzechnik, Ana G. Rondón, Wu Wei, Lars Steinmetz, Andrés Aguilera, and Nick J. Proudfoot. 2011. “Yeast Sen1 Helicase Protects the Genome from Transcription-Associated Instability.” *Molecular Cell* 41 (1): 21–32. doi:10.1016/j.molcel.2010.12.007.

Mitchell, Philip, Elisabeth Petfalski, Andrej Shevchenko, Matthias Mann, and David Tollervey. 1997. “The Exosome: A Conserved Eukaryotic RNA Processing Complex Containing Multiple 3′→5′ Exoribonucleases.” *Cell* 91 (4): 457–66. doi:10.1016/S0092-8674(00)80432-8.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.” *Science* 320 (5881): 1344–49. doi:10.1126/science.1158441.

Neil, Helen, Christophe Malabat, Yves d/’Aubenton-Carafa, Zhenyu Xu, Lars M. Steinmetz, and Alain Jacquier. 2009. “Widespread Bidirectional Promoters Are the Major Source of Cryptic Transcripts in Yeast.” *Nature* 457 (7232): 1038–42. doi:10.1038/nature07747.

O’Reilly, D., O. V. Kuznetsova, C. Laitem, J. Zaborowska, M. Dienstbier, and S. Murphy. 2014. “Human snRNA Genes Use Polyadenylation Factors to Promote Efficient Transcription Termination.” *Nucleic Acids Research* 42 (1): 264–75. doi:10.1093/nar/gkt892.

Parker, R. 2012. “RNA Degradation in Saccharomyces Cerevisae.” *Genetics* 191 (3): 671–702. doi:10.1534/genetics.111.137265.

Park, J., M. Kang, and M. Kim. 2015. “Unraveling the Mechanistic Features of RNA Polymerase II Termination by the 5’-3’ Exoribonuclease Rat1.” *Nucleic Acids Research* 43 (5): 2625–37. doi:10.1093/nar/gkv133.

Paten, B., J. Herrero, K. Beal, and E. Birney. 2009. “Sequence Progressive Alignment, a Framework for Practical Large-Scale Probabilistic Consistency Alignment.” *Bioinformatics* 25 (3): 295–301. doi:10.1093/bioinformatics/btn630.

Paten, B., J. Herrero, K. Beal, S. Fitzgerald, and E. Birney. 2008. “Enredo and Pecan: Genome-Wide Mammalian Consistency-Based Multiple Alignment with Paralogs.” *Genome Research* 18 (11): 1814–28. doi:10.1101/gr.076554.108.

Peters, J. M., R. A. Mooney, J. A. Grass, E. D. Jessen, F. Tran, and R. Landick. 2012. “Rho and NusG Suppress Pervasive Antisense Transcription in Escherichia Coli.” *Genes & Development* 26 (23): 2621–33. doi:10.1101/gad.196741.112.

Petfalski, Elisabeth, Thomas Dandekar, Yves Henry, and David Tollervey. 1998. “Processing of the Precursors to Small Nucleolar RNAs and rRNAs Requires Common Components.” *Molecular and Cellular Biology* 18 (3): 1181–89. doi:10.1128/MCB.18.3.1181.

Plaskon, Nicole E., Zach N. Adelman, and Kevin M. Myles. 2009. “Accurate Strand-Specific Quantification of Viral RNA.” Edited by Peter Sommer. *PLoS ONE* 4 (10): e7468. doi:10.1371/journal.pone.0007468.

Porrua, Odil, and Domenico Libri. 2013. “A Bacterial-like Mechanism for Transcription Termination by the Sen1p Helicase in Budding Yeast.” *Nature Structural & Molecular Biology* 20 (7): 884–91. doi:10.1038/nsmb.2592.

Potier, S., F. Lacroute, J.C. Hubert, and J.L. Souciet. 1990. “Studies on Transcription of the Yeast URA2 Gene.” *FEMS Microbiology Letters* 72 (1-2): 215–19. doi:10.1111/j.1574-6968.1990.tb03891.x.

Preker, P., K. Almvig, M. S. Christensen, E. Valen, C. K. Mapendano, A. Sandelin, and T. H. Jensen. 2011. “PROMoter uPstream Transcripts Share Characteristics with mRNAs and Are Produced Upstream of All Three Major Types of Mammalian Promoters.” *Nucleic Acids Research* 39 (16): 7179–93. doi:10.1093/nar/gkr370.

Preker, P., J. Nielsen, S. Kammler, S. Lykke-Andersen, M. S. Christensen, C. K. Mapendano, M. H. Schierup, and T. H. Jensen. 2008. “RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters.” *Science* 322 (5909): 1851–54. doi:10.1126/science.1164096.

Quinlan, A. R., and I. M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.

Rosonina, E., S Kaneko, and J.L. Manley. 2006. “Terminating the Transcript: Breaking up Is Hard to Do.” *Genes & Development* 20 (9): 1050–56. doi:10.1101/gad.1431606.

Schmidt, K., Z. Xu, D. H. Mathews, and J. S. Butler. 2012. “Air Proteins Control Differential TRAMP Substrate Specificity for Nuclear RNA Surveillance.” *RNA* 18 (10): 1934–45. doi:10.1261/rna.033431.112.

Schuch, B., M. Feigenbutz, D. L. Makino, S. Falk, C. Basquin, P. Mitchell, and E. Conti. 2014. “The Exosome-Binding Factors Rrp6 and Rrp47 Form a Composite Surface for Recruiting the Mtr4 Helicase.” *The EMBO Journal* 33 (23): 2829–46. doi:10.15252/embj.201488757.

Schulz, Daniel, Bjoern Schwalb, Anja Kiesel, Carlo Baejen, Phillipp Torkler, Julien Gagneur, Johannes Soeding, and Patrick Cramer. 2013. “Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis.” *Cell* 155 (5): 1075–87. doi:10.1016/j.cell.2013.10.024.

Sentenac, Andre. 1985. “Eukaryotic RNA Polymerase.” *Critical Reviews in Biochemistry and Molecular Biology* 18 (1): 31–90. doi:10.3109/10409238509082539.

Shearwin, K, B Callen, and J Egan. 2005. “Transcriptional Interference – a Crash Course.” *Trends in Genetics* 21 (6): 339–45. doi:10.1016/j.tig.2005.04.009.

Singh, Navjot, Zhuo Ma, Trent Gemmill, Xiaoyun Wu, Holland DeFiglio, Anne Rossettini, Christina Rabeler, et al. 2009. “The Ess1 Prolyl Isomerase Is Required for Transcription Termination of Small Noncoding RNAs via the Nrd1 Pathway.” *Molecular Cell* 36 (2): 255–66. doi:10.1016/j.molcel.2009.08.018.

Stead, J. A., J. L. Costello, M. J. Livingstone, and P. Mitchell. 2007. “The PMC2NT Domain of the Catalytic Exosome Subunit Rrp6p Provides the Interface for Binding with Its Cofactor Rrp47p, a Nucleic Acid-Binding Protein.” *Nucleic Acids Research* 35 (16): 5556–67. doi:10.1093/nar/gkm614.

Steinmetz, E. J., and D. A. Brow. 2003. “Ssu72 Protein Mediates Both Poly(A)-Coupled and Poly(A)-Independent Termination of RNA Polymerase II Transcription.” *Molecular and Cellular Biology* 23 (18): 6339–49. doi:10.1128/MCB.23.18.6339-6349.2003.

Steinmetz, Eric J., Nicholas K. Conrad, David A. Brow, and Jeffry L. Corden. 2001. “RNA-Binding Protein Nrd1 Directs poly(A)-Independent 3′-End Formation of RNA Polymerase II Transcripts.” *Nature* 413 (6853): 327–31. doi:10.1038/35095090.

Steinmetz, Eric J., Christopher L. Warren, Jason N. Kuehner, Bahman Panbehi, Aseem Z. Ansari, and David A. Brow. 2006. “Genome-Wide Distribution of Yeast RNA Polymerase II and Its Control by Sen1 Helicase.” *Molecular Cell* 24 (5): 735–46. doi:10.1016/j.molcel.2006.10.023.

Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, et al. 2008. “A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome.” *Science* 321 (5891): 956–60. doi:10.1126/science.1160342.

Terzi, N., L. S. Churchman, L. Vasiljeva, J. Weissman, and S. Buratowski. 2011. “H3K4 Trimethylation by Set1 Promotes Efficient Termination by the Nrd1-Nab3-Sen1 Pathway.” *Molecular and Cellular Biology* 31 (17): 3569–83. doi:10.1128/MCB.05590-11.

Thebault, P., G. Boutin, W. Bhat, A. Rufiange, J. Martens, and A. Nourani. 2011. “Transcription Regulation by the Noncoding RNA SRG1 Requires Spt2-Dependent Chromatin Deposition in the Wake of RNA Polymerase II.” *Molecular and Cellular Biology* 31 (6): 1288–1300. doi:10.1128/MCB.01083-10.

Thiebaut, Marilyne, Jessie Colin, Helen Neil, Alain Jacquier, Bertrand Séraphin, François Lacroute, and Domenico Libri. 2008. “Futile Cycle of Transcription Initiation and Termination Modulates the Response to Nucleotide Shortage in S. Cerevisiae.” *Molecular Cell* 31 (5): 671–82. doi:10.1016/j.molcel.2008.08.010.

Thiebaut, Marilyne, Elena Kisseleva-Romanova, Mathieu Rougemaille, Jocelyne Boulay, and Domenico Libri. 2006. “Transcription Termination and Nuclear Degradation of Cryptic Unstable Transcripts: A Role for the Nrd1-Nab3 Pathway in Genome Surveillance.” *Molecular Cell* 23 (6): 853–64. doi:http://dx.doi.org/10.1016/j.molcel.2006.07.029.

Thoms, Matthias, Emma Thomson, Jochen Baßler, Marén Gnädig, Sabine Griesel, and Ed Hurt. 2015. “The Exosome Is Recruited to RNA Substrates through Specific Adaptor Proteins.” *Cell* 162 (5): 1029–38. doi:10.1016/j.cell.2015.07.060.

Tirosh, I., S. Reikhav, A. A. Levy, and N. Barkai. 2009. “A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation.” *Science* 324 (5927): 659–62. doi:10.1126/science.1169766.

Tsankov, Alexander M., Dawn Anne Thompson, Amanda Socha, Aviv Regev, and Oliver J. Rando. 2010. “The Role of Nucleosome Positioning in the Evolution of Gene Regulation.” Edited by Peter B. Becker. *PLoS Biology* 8 (7): e1000414. doi:10.1371/journal.pbio.1000414.

Tudek, Agnieszka, Odil Porrua, Tomasz Kabzinski, Michael Lidschreiber, Karel Kubicek, Andrea Fortova, François Lacroute, et al. 2014. “Molecular Basis for Coordinating Transcription Termination with Noncoding RNA Degradation.” *Molecular Cell* 55 (3): 467–81. doi:10.1016/j.molcel.2014.05.031.

Uesaka, Masahiro, Osamu Nishimura, Yasuhiro Go, Kinichi Nakashima, Kiyokazu Agata, and Takuya Imamura. 2014. “Bidirectional Promoters Are the Major Source of Gene Activation-Associated Non-Coding RNAs in Mammals.” *BMC Genomics* 15 (1): 35. doi:10.1186/1471-2164-15-35.

Ursic, D., K Chinchilla, J.S. Finkel, and M.R. Culbertson. 2004. “Multiple Protein/protein and protein/RNA Interactions Suggest Roles for Yeast DNA/RNA Helicase Sen1p in Transcription, Transcription-Coupled DNA Repair and RNA Processing.” *Nucleic Acids Research* 32 (8): 2441–52. doi:10.1093/nar/gkh561.

van Dijk, E. L., C. L. Chen, Y. d’Aubenton-Carafa, S. Gourvennec, M. Kwapisz, V. Roche, C. Bertrand, et al. 2011. “XUTs Are a Class of Xrn1-Sensitive Antisense Regulatory Non-Coding RNA in Yeast.” *Nature* 475 (7354): 114–17. doi:10.1038/nature10118.

van Hoof, A., P. Lennertz, and R. Parker. 2000. “Yeast Exosome Mutants Accumulate 3’-Extended Polyadenylated Forms of U4 Small Nuclear RNA and Small Nucleolar RNAs.” *Molecular and Cellular Biology* 20 (2): 441–52. doi:10.1128/MCB.20.2.441-452.2000.

Vasiljeva, Lidia, Minkyu Kim, Hannes Mutschler, Stephen Buratowski, and Anton Meinhart. 2008. “The Nrd1–Nab3–Sen1 Termination Complex Interacts with the Ser5-Phosphorylated RNA Polymerase II C-Terminal Domain.” *Nature Structural & Molecular Biology* 15 (8): 795–804. doi:10.1038/nsmb.1468.

Wahba, Lamia, Jeremy D. Amon, Douglas Koshland, and Milena Vuica-Ross. 2011. “RNase H and Multiple RNA Biogenesis Factors Cooperate to Prevent RNA:DNA Hybrids from Generating Genome Instability.” *Molecular Cell* 44 (6): 978–88. doi:10.1016/j.molcel.2011.10.017.

Walowsky, C., D. J. Fitzhugh, I. B. Castano, J. Y. Ju, N. A. Levin, and M. F. Christman. 1999. “The Topoisomerase-Related Function Gene TRF4 Affects Cellular Sensitivity to the

Antitumor Agent Camptothecin." *Journal of Biological Chemistry* 274 (11): 7302–8. doi:10.1074/jbc.274.11.7302.

Webb, Shaun, Ralph D Hector, Grzegorz Kudla, and Sander Granneman. 2014. "PAR-CLIP Data Indicate That Nrd1-Nab3-Dependent Transcription Termination Regulates Expression of Hundreds of Protein Coding Genes in Yeast." *Genome Biology* 15 (1): R8. doi:10.1186/gb-2014-15-1-r8.

Werner, Finn, and Dina Grohmann. 2011. "Evolution of Multisubunit RNA Polymerases in the Three Domains of Life." *Nature Reviews Microbiology* 9 (2): 85–98. doi:10.1038/nrmicro2507.

Whitehouse, Iestyn, Oliver J. Rando, Jeff Delrow, and Toshio Tsukiyama. 2007. "Chromatin Remodelling at Promoters Suppresses Antisense Transcription." *Nature* 450 (7172): 1031–35. doi:10.1038/nature06391.

Wyers, Françoise, Mathieu Rougemaille, Gwenaël Badis, Jean-Claude Rousselle, Marie-Elisabeth Dufour, Jocelyne Boulay, Béatrice Régnault, et al. 2005. "Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase." *Cell* 121 (5): 725–37. doi:10.1016/j.cell.2005.04.030.

Xu, Zhenyu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Munster, Jurgi Camblong, Elisa Guffanti, Francoise Stutz, Wolfgang Huber, and Lars M. Steinmetz. 2009. "Bidirectional Promoters Generate Pervasive Transcription in Yeast." *Nature* 457 (7232): 1033–37. doi:10.1038/nature07728.

Xu, Z., W. Wei, J. Gagneur, S. Clauder-Munster, M. Smolik, W. Huber, and L. M. Steinmetz. 2011. "Antisense Expression Increases Gene Expression Variability and Locus Interdependency." *Molecular Systems Biology* 7 (1): 468–468. doi:10.1038/msb.2011.1.

Yassour, Moran, Jenna Pfiffner, Joshua Z Levin, Xian Adiconis, Andreas Gnirke, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman, and Aviv Regev. 2010. "Strand-Specific RNA Sequencing Reveals Extensive Regulated Long Antisense Transcripts That Are Conserved across Yeast Species." *Genome Biology* 11 (8): R87. doi:10.1186/gb-2010-11-8-r87.

Yuan, G.-C., Liu, Y-J, Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Oliver J. Rando, O.J. 2005. "Genome-Scale Identification of Nucleosome Positions in S. cerevisiae." *Science* 309 (5734): 626–30. doi:10.1126/science.1112178.

Yuryev, A., M. Patturajan, Y. Litingtung, R. V. Joshi, C. Gentile, M. Gebara, and J. L. Corden. 1996. "The C-Terminal Domain of the Largest Subunit of RNA Polymerase II Interacts with a Novel Set of Serine/arginine-Rich Proteins." *Proceedings of the National Academy of Sciences* 93 (14): 6975–80. doi:10.1073/pnas.93.14.6975.

Appendix A – Strains Used in This Study

Appendix A - Strains Used In This Study

Name	Strain	Genotype	Mating Type	Species
BY4742	S288c WT	his3Δ1, lys2Δ0, leu2Δ0, ura3Δ0	MATα	S.cere
yRD004	S288c *rrp6*Δ	rrp6::KanMX, his3Δ1, leu2Δ0, ura3Δ0, met15Δ0	MATa	S.cere
L6441	S1278b WT	parent: 10512-3C, ura3-52, leu2::hisG, his3::hisG	MATα	S.cere
yRD003	Σ1278b *rrp6*Δ	rrp6::KanMX, his3Δ1, leu2Δ0, ura3Δ0	MATa	S.cere
JAY291	JAY291 WT	WT	MATa	S.cere
yJV003	JAY291 *rrp6*Δ	rrp6::KanMX	MATa	S.cere
N17	N17 WT	HO::lox-Kan-lox	MATα	S.para
yJV009	N17 *rrp6*Δ	HO::lox-Kan-lox, rrp6::NatMX	MATα	S.para
yJV001	S288c	his3Δ1, leu2Δ0, met15Δ0, RPB3-3xFLAG-NAT1	MATa	S.cere

Appendix B – Oligos/Primers

RT-qPCR and NET-qPCR primer names and sequences. Primers are specific to either *S.cerevisiae* (S.cere) or *S.paradoxus* (S.para), but in some cases could be used in either species background. All primers were named for the nearest or overlapping gene annotation. Those primers labeled "–T" denote the presence of the unique 5' tagged used for strand-specificity in RT reactions (for more details regarding strand-specific cDNA see *CUT expression validation by RT-qPCR* on page 33). An asterisk in the final column denotes candidates not requiring strand-specific RT-qPCR; for these candidates qPCR was performed on random hexamer primed cDNA.

Species	Name	Sequence	6mer cDNA
N/A	Universal Fwd	GGC AGT ATC GTG AAT TCG ATG C	
S.cere	SIF2 F1-T	GGC AGT ATC GTG AAT TCG ATG CAG ACG TTT ACC TGC CCA TCC	
S.cere	YKU80 F1-T	GGC AGT ATC GTG AAT TCG ATG CGT GTC GGC GGT AAT GAA GGA	
S.cere	YKL151C F2-T	GGC AGT ATC GTG AAT TCG ATG CTT GGC CTC CTA CCC TCT TGT	
S.cere	YKL151C F1-T	GGC AGT ATC GTG AAT TCG ATG CAA TGA CCG TAC CAG CGT TGT	
S.cere	YKL151C R1-T	GGC AGT ATC GTG AAT TCG ATG CAC AGG GGC ACC GTA TTT CAG	
S.cere	ACT1 R1-T	GGC AGT ATC GTG AAT TCG ATG CAC CGG CAG ATT CCA AAC CCA	
S.cere	ACT1 R1	ACG TGA GTA ACA CCA TCA CCG G	
S.cere	ACT1 F1	ACG TCG CCT TGG ACT TCG AAC A	
S.cere	SIF2 R1-T	GGC AGT ATC GTG AAT TCG ATG CTC AAT CGT GGA TGG TGT CCC	
S.cere	YKU80 R1-T	GGC AGT ATC GTG AAT TCG ATG CGC TAC CGT CCG TTC TAG TCG	
S.cere	YKL151C R2-T	GGC AGT ATC GTG AAT TCG ATG CGC TTG ATC GCC CAG GAA TTG	
S.cere	YKL151C F1	AAT GAC CGT ACC AGC GTT GT	
S.cere	YKL151C R1	ACA GGG GCA CCG TAT TTC AG	
S.cere	YKL151C F2	TTG GCC TCC TAC CCT CTT GT	
S.cere	YKL151C R2	GCT TGA TCG CCC AGG AAT TG	
S.cere	YKU80 F1	GTG TCG GCG GTA ATG AAG GA	
S.cere	YKU80 R1	GCT ACC GTC CGT TCT AGT CG	
S.cere	SIF2 F1	AGA CGT TTA CCT GCC CAT CC	
S.cere	SIF2 R1	TCA ATC GTG GAT GGT GTC CC	
S.cere	TKL2 F1	GGC AAT AGC GCA GGC CAA CTT T	
S.cere	TKL2 R1	TGC TGC AGG AGC CGT TAG GTT A	
S.cere	TKL2 R2	TTG GTG CGT TGG ACC ATC CTC A	
S.cere	TRF5 R2	GGT TAA GCT GGT TCG TTT CAC TAG C	

S.cere	TRF5 R1	ACG AAC GGG TTA GAG GCT GCA A	
S.cere	TRF5 F1	AAC CTC CCA ATC CTC CTG TGT GC	
S.cere	MLS1 R2	AGA CTC GGG CTC CTA TCA TCT GG	
S.cere	MLS1 F1	TTG CTC AAA TCA GTG GGC GTC G	
S.cere	MLS1 R1	AAT TCG CGC TGG CCG CTA AGT A	
S.para	YKL151C F3	TGG TCT GCA TTG CAC GTC CCT T	
S.para	YKL151C R3	TGA GTT GTT ACG CAG GCT GCA C	
S.para	YKL151C F3-T	GGC AGT ATC GTG AAT TCG ATG CTG GTC TGC ATT GCA CGT CCC TT	
S.para	YKL151C R3-T	GGC AGT ATC GTG AAT TCG ATG CTG AGT TGT TAC GCA GGC TGC AC	
S.para	YKL151C F4	TTG ACC GCC CAC TCT CTT GTT G	
S.para	YKL151C R4	GGC AGC TTG ATC GCA CAA GAA C	
S.para	YKL151C F4-T	GGC AGT ATC GTG AAT TCG ATG CTT GAC CGC CCA CTC TCT TGT TG	
S.para	YKL151C R4-T	GGC AGT ATC GTG AAT TCG ATG CGG CAG CTT GAT CGC ACA AGA AC	
S.para	YKU80 F4	ACA ACC AAG TCT TGT ATC TGC GGT	
S.para	YKU80 R4	TGT GCT ACC GTC CAT TCT AGT CG	
S.para	YKU80 F4-T	GGC AGT ATC GTG AAT TCG ATG CAC AAC CAA GTC TTG TAT CTG CGG T	
S.para	YKU80 R4-T	GGC AGT ATC GTG AAT TCG ATG CTG TGC TAC CGT CCA TTC TAG TCG	
S.para	SIF2 F2	AGC GAA CGG AGC CAT CCA TC	
S.para	SIF2 R2	TGC CTC GGA CGA TGG TAC TCT	
S.para	SIF2 F2-T	GGC AGT ATC GTG AAT TCG ATG CAG CGA ACG GAG CCA TCC ATC	
S.para	SIF2 R2-T	GGC AGT ATC GTG AAT TCG ATG CTG CCT CGG ACG ATG GTA CTC T	
S.para	ACT1 F2	TGA GAG TTG CCC CAG AAG AGC A	
S.para	ACT1 R2	ACG TAG AAG GCT GGG ACG TTG A	
S.para	ACT1 R2-T	GGC AGT ATC GTG AAT TCG ATG CAC GTA GAA GGC TGG GAC GTT GA	
S.cere/S.para	YBR230C_F1	CCG TCC AGC GTC AAA AGA CCC A	
S.para	YBR230C_R1	ACC ACA ACG TGA GAT TCT TGA AGG G	
S.para	YBR230C_R1-T	GGC AGT ATC GTG AAT TCG ATG CAC CAC AAC GTG AGA TTC TTG AAG GG	
S.cere	YBR230C_R1	CCA CAA CGC CAG ATT CTT GAA GGG G	
S.cere	YBR230C_R1-T	GGC AGT ATC GTG AAT TCG ATG CCC ACA ACG CCA GAT TCT TGA AGG GG	
S.cere	YDL183C_F2	ACT GAA TCT CAA GCG CAC GCA GT	*
S.para	YDL183C_F1	TGG ATC TCA AAT GTA CGC TGC ACA C	*
S.para	YDL183C_R1	AGG ATG GTG CTC GTG GCT AAG T	*
S.cere	YDL183C_R2	TTG ATC TCT CCA AGG TTA GCC GCC	*
S.cere	YDR234W_F3	CCA GGC CCT AGA ACT GTG GAA CGA	
S.para	YDR234W_F3	AGG TTC TGA CCC CTC AAC CAG GTC	
S.cere	YDR234W_R3	AGC TAG TTT TGC GCT GCC TCT T	
S.para	YDR234W_R3	GCT AGT TTC GCA CTA CCT CTT ACG G	
S.para	YDR234W_R3-T	GGC AGT ATC GTG AAT TCG ATG CGC TAG TTT CGC ACT ACC TCT TAC GG	
S.cere	YDR234W_R3-T	GGC AGT ATC GTG AAT TCG ATG CAG CTA GTT TTG CGC TGC CTC TT	
S.cere	YDR518W_F1	GAC ACT TCA GAA TCC TTG GCC TGG T	
S.cere/(Σ1278b)	YDR518W_F1.2	GAC ACT TCA GAA TCC TTA GCC TGG T	
S.para	YDR518W_F1	CAT GCA CGG TGT CCT CTG TAG AC	

S.para	YDR518W_R1	AGG TAC GCA TCA TAT TGA CGG CCA	
S.cere	YDR518W_R1	GGT ACA CAT CAC ATT GAC GGC CAG	
S.para	YDR518W_R1	GGC AGT ATC GTG AAT TCG ATG CAG GTA CGC ATC ATA TTG ACG GCC A	
S.cere	YDR518W_R1-T	GGC AGT ATC GTG AAT TCG ATG CGG TAC ACA TCA CAT TGA CGG CCA G	
S.cere	YHL028W_F1	TGG GTC TAG CCT CTG ATC CAC CA	
S.para	YHL028W_F1_N17	TCC ACC AGT ACC GTG TCT TCG C	
S.cere	YHL028W_R1	ACG TTG GCA AGC CCA TTT CAC GA	
S.para	YHL028W_R1	GCA AGC CCA TTT CAC GAC CCT	
S.para	YHL028W_R1-T	GGC AGT ATC GTG AAT TCG ATG CGC AAG CCC ATT TCA CGA CCC T	
S.cere	YHL028W_R1-T	GGC AGT ATC GTG AAT TCG ATG CAC GTT GGC AAG CCC ATT TCA CGA	
S.cere	YLR039C_F1	AGA GCT TTG CTT TCC TCA GTC CCT	
S.para	YLR039C_F1	GCC TTT TGG ACC AAT TGT GTG TTG T	
S.para	YLR039C_R1	TGA CAA GTG CTG CGA ACG CT	
S.para	YLR039C_R1-T	GGC AGT ATC GTG AAT TCG ATG CTG ACA AGT GCT GCG AAC GCT	
S.cere	YLR039C_R1	TGA GGC TAA AAT GCT TGG CGT ACT T	
S.cere	YLR039C_R1-T	GGC AGT ATC GTG AAT TCG ATG CTG AGG CTA AAA TGC TTG GCG TAC TT	
S.cere	YLR449W_F1	GCA ATG GCT CGC TAG GAC AGC A	
S.para	YLR449W_F1	ACC GCA AGT ATG CGA TTG CAG C	
S.para	YLR449W_R1	AGG GTT CAG GAT TAG GCC GTC ACC	
S.para	YLR449W_R1-T	GGC AGT ATC GTG AAT TCG ATG CAG GGT TCA GGA TTA GGC CGT CAC C	
S.cere	YLR449W_R1	ATC AGG TCA TCG CCG TTC TGG G	
S.cere	YLR449W_R1-T	GGC AGT ATC GTG AAT TCG ATG CAT CAG GTC ATC GCC GTT CTG GG	
S.cere	YNL146C-A_F1	GGC TTT AGA CTT CAA ATC GCG GTG	
S.para	YNL146C-A_F1	ACA TCA GAA CCT TCG GCG GAA CT	
S.para	YNL146C-A_R1	CAC TGA ACT GAT CTC CAA AAA CGC A	
S.para	YNL146C-A_R1-T	GGC AGT ATC GTG AAT TCG ATG CCA CTG AAC TGA TCT CCA AAA ACG CA	
S.cere	YNL146C-A_R1	AAG TTC CGC CGA AGG TTC TGA	
S.cere	YNL146C-A_R1-T	GGC AGT ATC GTG AAT TCG ATG CAA GTT CCG CCG AAG GTT CTG A	
S.cere	YNL250W_F1	TTG CCT TGT CTC GTG CGC TAG T	*
S.para	YNL250W_F1	ATA GCT TTT CCG CAC CCC GTG T	*
S.para	YNL250W_R1	AAC CGC GAG ATG AAG CCA TTT CT	*
S.cere	YNL250W_R1	AGG GAC AAG ATG AAA ACC GGA ACC T	*
S.cere	YNR049C_F1	TCC TGG GAA CCC CTG GAA AGG A	
S.para	YNR049C_F1	TCC CCT AAC CAA CCT GGG AAC C	
S.cere	YNR049C_R1	TTC CCA AGA AGG ATC CGG GCG A	
S.para	YNR049C_R1	TCA GGA AGG GTC TGG GCG ATT	
S.para	YNR049C_R1-T	GGC AGT ATC GTG AAT TCG ATG CTC AGG AAG GGT CTG GGC GAT T	
S.cere	YNR049C_R1-T	GGC AGT ATC GTG AAT TCG ATG CTT CCC AAG AAG GAT CCG GGC GA	
S.cere	YOR336W_F2	GCT CAG CAC GCT CTG TCT TAC G	
S.para	YOR336W_F1	ACC GAA GTG CTT GCT GTT ATC GT	*
S.para	YOR336W_R1	ACG TAG CAC AGA AGG CGC TGA A	*
S.cere	YOR336W_R2	ACA GAG GCA CTC ACA CTG ATA CGT C	

S.cere	YOR336W_R2-T	GGC AGT ATC GTG AAT TCG ATG CAC AGA GGC ACT CAC ACT GAT ACG TC	
S.cere	URA2_F1	ACC AGG CGC CAA AGG AAA ATG C	*
S.cere	URA2_R1	TGC ATC CTC CGC GGC ATC TAA A	*
S.cere	18s_F1	TCA CTA CCT CCC TGA ATT AGG ATT G	*
S.cere	18s_R1	AGA AAC GGC TAC CAC ATC CAA	*
S.cere	usURA2_F1	ATT CAC CAG CGA CGG ATT TCT CAG	*
S.cere	usURA2_R1	TGC TTT CGT CAT CGT CAA CGC CA	*