Title
Coupling of spliceosome complexity to intron diversity

Author
Sales-Lee, Jade

Publication Date
2021

Coupling of spliceosome complexity to intron diversity

by
Jade Sales-Lee

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Hiten D. Madhani
76DDD3CF67EA4D0...

Hiten D. Madhani
Chair

DocuSigned by:
Geeta Narlikar
DocuSigned by:

Geeta Narlikar

DocuSigned by:
Wallace Marshall
43941FCFA7C0447...

Wallace Marshall

Committee Members

Acknowledgments

This body of work could not have been produced with out the support of a villages worth of people. While it would be impossible to fit everyone who has assisted my journey to this point in a few pages, I'd like to especially thank the following:

Hiten D. Madhani as my academic advisor was instrumental to the completion of this work. His support and guidance through my time at UCSF will be forever with me.

The rest of my committee, Geeta Narlikar and Wallace Marshall, I thank for discussion and guidance strengthening the scientific rigor of this work.

The Madhani lab as a whole I thank for years of discussion and support as I navigated the beautiful world of splicing. I would like to especially thank Jordan Burke and Irene Beusch for their mentorship and being amazing bench buddies.

I can't express enough gratitude to the members of the Guthrie lab for unending discussion of the spliceosome and experimental methods. Without your support and knowledge this work could not have come about.

Finally, I would like to acknowledge the support and love of my family both by blood and by choice. Christina, Steven, Brooke, Quin, Kristina, Alexander, Geoff, Racheal, Nick, Matt. This list could go on for the length of this dissertation. You have sustained me with your presence, care, and love. I can not express how valuable you all have been over this time.

Thank you.

Contributions

Chapter 1: The text of this dissertation chapter is a reprint of the material as it appears in Nucleic Acids Research. J. Sales-Lee performed the ribosome profiling experiments in the JEC21 (serotype D) strain of Cryptococcus neoformans.

Chapter 2: The text of this dissertation chapter is a reprint of the material as it appears in G3. J. Sales-Lee performed insertional mutagenesis of the transposon reporter strain and identified the initial set of mutants defective in transposon suppression.

Chapter 3: The text of this dissertation chapter is a reprint of the material as it appears in Cell. J. Sales-Lee contributed to the design and interpretation of the spliceosome profiling method.

Chapter 4: The text of this dissertation chapter is a preprint of the material before journal submission. J. Sales-Lee designed and performed all of the described experiments and performed analysis of the RNA-seq and mass spectrometry data.

Coupling of spliceosome complexity to intron diversity

Jade Sales-Lee

Abstract

We determined that over 60 spliceosomal proteins are conserved between many fungal species and humans but were lost during the evolution of *S. cerevisiae*, an intron-poor yeast with unusually rigid splicing signals. We analyzed null mutations in a subset of these factors, most of which had not been investigated previously, in the intron-rich yeast *Cryptococcus neoformans.* We found they govern splicing efficiency of introns with divergent spacing between intron elements. Importantly, most of these factors also suppress usage of weak nearby cryptic/alternative splice sites. Among these, orthologs of GPATCH1 and the helicase DHX35 display correlated functional signatures and copurify with each other as well as components of catalytically active spliceosomes, identifying a conserved G-patch/helicase pair that promotes splicing fidelity. We propose that a significant fraction of spliceosomal proteins in humans and most eukaryotes are involved in limiting splicing errors, potentially through kinetic proofreading mechanisms, thereby enabling greater intron diversity.

Table of Contents

List of Figures

List of Figures - cont.

List of Figures - cont.

List of Tables

*Tables available in dryad

List of Tables - cont.

*Tables available in dryad

Chapter 1

Quantitative global studies reveal differential translational control by start codon context across the fungal kingdom.

Edward Wallace 2*Ṫ, Corinne Maufrais 1,3Ṫ, Jade Sales-Lee 4, Laura Tuck 2, Luciana de Oliveira 1, Frank Feuerbach 6, Frédérique Moyrand 1, Prashanthi Natarajan 4, Hiten D. Madhani 4,5*, Guilhem Janbon 1*

1. Institut Pasteur, Unité Biologie des ARN des Pathogènes Fongiques, Département de Mycologie, F-75015, Paris, France

2. Institute for Cell Biology and SynthSys, School of Biological Sciences, University of Edinburgh, UK

3. Institut Pasteur, HUB Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, F-75015, Paris, France

4. Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94158, USA

5. Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA

6 Institut Pasteur, Unité Génétique des Interactions Macromoléculaire, Département Génome et Génétique, F-75015, Paris, France

* To whom correspondence should be addressed. Emails: Edward.Wallace@ed.ac.uk, Hiten.Madhani@ucsf.edu, Guilhem.Janbon@pasteur.fr

Ṫ The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

Abstract

Eukaryotic protein synthesis generally initiates at a start codon defined by an AUG and its surrounding Kozak sequence context, but the quantitative importance of this context in different species is unclear. We tested this concept in two pathogenic *Cryptococcus* yeast species by genome-wide mapping of translation and of mRNA 5' and 3' ends. We observed thousands of AUG-initiated upstream open reading frames (uORFs) that are a major contributor to translation repression. uORF use depends on the Kozak sequence context of its start codon, and uORFs with strong contexts promote nonsense-mediated mRNA decay. Transcript leaders in *Cryptococcus* and other fungi are substantially longer and more AUG-dense than in *Saccharomyces*. Numerous *Cryptococcus* mRNAs encode predicted dual-localized proteins, including many aminoacyl-tRNA synthetases, in which a leaky AUG start codon is followed by a strong Kozak context in-frame AUG, separated by mitochondrial-targeting sequence. Analysis of other fungal species shows that such dual-localization is also predicted to be common in the ascomycete mould, *Neurospora crassa*. Kozak-controlled regulation is correlated with insertions in translational initiation factors in fidelity-determining regions that contact the initiator tRNA. Thus, start codon context is a signal that quantitatively programs both the expression and the structures of proteins in diverse fungi.

Introduction

Fungi are important in the fields of ecology, medicine, and biotechnology. With roughly 3 million predicted fungal species, this kingdom is the most diverse of the domain Eukarya (1). Recent initiatives such as the 1000 Fungal Genomes Project at the Joint Genome Institute, or the Global Catalogue of Microorganisms, which aims to produce 2500 complete fungal genomes in the next 5 years, will result in a deluge of genome sequence data (2,3). Comparative analysis of coding sequences enables the generation of hypotheses on genome biology and evolution (4-7). However, these analyses intrinsically depend on the quality of the coding gene identification and annotation, which have limitations. First, they depend on automatic sequence comparisons, which limit the identification of clade-specific genes. Second, fungal genes generally contain introns whose positions are difficult to predict based on the genome sequence alone (8). An uncertain intron annotation results in a poor annotation of the coding region extremities, which are generally less evolutionary conserved (9). Third, annotation pipelines only predict plausible open reading frames (ORFs), initially for yeast a contiguous stretch of at least 100 codons starting with an AUG codon and ending with a stop codon (10). These approaches do not reveal which ORFs are translated to protein, and are biased against short ORFs (11).

The rule of thumb that first AUG of an ORF is used as the start codon can be wrong in both directions: poor-context AUGs may be skipped, and non-AUGs may be used. The rules for selection of start codons in eukaryotes were discovered by Kozak: the most 5' AUG is generally selected if it is far (>20nt) from the transcription start site, and in metazoans efficient selection is associated with a sequence context gccRccAUGG, where the R indicates a purine at the -3 position (12,13). In *S. cerevisiae*, translation likewise generally starts at the 5' AUG in the mRNA sequence, and although the preferred aaaAUG sequence context weakly affects endogenous protein output (14), AUG sequence context also modulates protein output from reporter mRNAs

(15,16). Usage of non-AUG start codons has been observed in diverse eukaryotes, including the fungi *S. cerevisiae, C. albicans, S. pombe,* and *N. crassa* (17-22), and in *S. cerevisiae*, sequence context has a large effect on the usage of non-AUG start codons (23).

Weak or inefficient start codons near the 5'end of mRNA can give rise to translational regulation, explained by the scanning model of eukaryotic translation initiation. Translation starts by the pre-initiation complex binding mRNA at the 5' cap and then scanning the transcript leader (TL) sequence in a 3' direction until it identifies a start codon, at which translation initiates (24). Here we call the 5' regulatory region of mRNA the TL rather than the 5' UTR, because short "upstream" ORFs in this region can be translated (25). The pre-initiation complex sediments at 43S, and comprises the small ribosomal subunit, methionyl initiator tRNA, and numerous eukaryotic translation initiation factors (eIFs). Biochemical, genetic, and structural data indicate that eIF1 and eIF1A associate with the 43S pre-initiation complex (26,27). Recognition of the start codon involves direct interactions of eIF1 and eIF1A with the start codon context and initiator tRNA within a larger 48S pre-initiation complex. Start codon selection occurs when eIF1 is replaced by eIF5's N-terminus (28), then eIF2 is released, the large ribosomal subunit joins catalyzed by eIF5B, and translation begins (27). This work has been largely driven by studies in *S. cerevisiae* and metazoans. Although the core protein and RNA machinery of eukaryotic translation initiation is highly conserved, it is not understood how fungi quantitatively vary in the sequence, structure, and function of their translation initiation machinery.

Cryptococcus are basidiomycete yeasts with a high density of introns in their coding genes (29). These introns influence gene expression and genome stability (30-32). The current genome annotation of pathogenic *C. neoformans* and *C. deneoformans* reference strains are based on both automatic and manual curations of gene structures using RNA-Seq data (33,34). Although the high degree of interspecies conservation of intron numbers and positions within coding sequences suggest that these annotations are reliable (34), the regulatory regions

(transcript leader and 3' UTRs) at transcript extremities are less well identified. In fact, most fungal genomes lack complete transcript annotations, thus we do not know how regulatory structure varies across fungi.

In this paper, we experimentally determine the beginning and the end of both coding regions and of transcripts in two *Cryptococcus* species, providing an important genomic resource for the field. Furthermore, our joint analysis of TL sequences and translation identifies a Kozak sequence context that regulates start codon selection, affecting upstream ORF regulation and also alternative protein targeting to mitochondria. Comparison with other fungal genomes revealed that these types of regulation are common in this kingdom: the first AUG of an mRNA or an ORF is not always the major start codon in fungi. These studies demonstrate that start codon sequence context is an important gene regulatory signal that programs both the abundance and the structures of proteins across the fungal kingdom.

Results

Delineation of transcript ends in *C. neoformans* and *C. deneoformans*

To annotate the extremities of the coding genes in *C. neoformans* and *C. deneoformans*, we mapped the 5' ends (Transcription Start Sites; TSS) with TSS-Seq (38), the 3' ends (Polyadenylation sites; PAS) with QuantSeq 3'mRNA-Seq, and sequenced the same samples with stranded mRNA-Seq. These experiments were done in biological triplicate from cells growing at two temperatures (30°C and 37°C) and two stages of growth (exponential and early stationary phases) with external normalization with spike-in controls.

We identified 4.7 x 106 unique TSSs and 6.3 x 104 unique PASs in *C. neoformans*. Clustering of these positions revealed between 27,339 and 42,720 TSS clusters and between 9,217 and 16,697 PAS clusters depending on the growth conditions (Table 1.1). We used the clusters associated with the coding genes to produce an initial annotation, using the most distal TSS and PAS clusters for each gene. The predicted positions which changed the extremities of the genes by more than 100 bp were manually curated (n=1131 and n=286 for the TSS and PAS, respectively). We then selected the most prominent clusters that represented at least 10% of the normalized reads count per coding gene in at least one condition (i.e., sum across three normalized replicate samples), for wild-type strains. Finally, the most distal of these TL-TSS and 3'UTR-PA clusters were labeled as the 5'and 3' ends of the coding genes for our final annotation (Table 1.1). For the genes for which no TL-TSS cluster or no 3'UTR-PAS cluster could be identified, we maintained the previous annotation. We used the same strategies for *C. deneoformans* and obtained similar results (Table 1.1).

As expected, most of the TSS clusters (62%) were associated with the TL whereas most of the PAS clusters (82%) were associated with the 3'UTR of the coding genes (Table 1.1). We analyzed the 3'UTR sequences, confirming the ATGHAH motif associated with the PAS (33). In

addition, as previously observed in other systems (68) a (C/T)(A/G)-rich motif was associated with the maxima of these transcription start site clusters. Overall, 89% of the coding genes have both their TL and 3'UTR sequences supported by identified TSS and PAS clusters, respectively.

The analysis leads to a scheme of a stereotypical *C. neoformans* coding gene (Figure 1.1A). In average, it is 2,305 bp long (median 2,008 bp) and contains 5.6 short introns (median 5) in its sequence. As previously reported (29), these introns are short (63.4 nt in average) and associated with conserved consensus motifs. The *C. neoformans* TL and 3'UTR have median lengths of 105 nt and 127 nt, respectively (177 nt and 186 nt, mean; Figure 1.1B,C). Only 887 and 429 genes contain one or more introns in their TL and 3'UTR sequence, respectively; these introns are usually larger (118.3 nt) than those that interrupt the CDS. This gene structure is similar in *C. deneoformans* (Table 1.1) and there are good correlations between the 3'UTR and TL sizes of the orthologous genes in the two species (Figure 1.1D,E).

More than a third of genes have upstream AUGs that affect translation

The analysis of the TL sequences in *C. neoformans* revealed the presence of 10,286 AUG triplets upstream (uAUG) of the annotated translation start codon (aAUG). We include uAUGs that are either out-of-frame from the start codon, or in-frame but with an intervening stop codon, which are very unlikely to encode a continuous polypeptide. Strikingly, 2,942 genes possess at least one uAUG, representing 43% of the genes with an annotated TL in *C. neoformans* (Figure 1.1F). A similar result was obtained in *C. deneoformans*, in which we found 10,254 uAUGs in 3,057 genes, and uAUG counts are correlated between orthologous mRNAs in the two species (Fig 1G). This is consistent with previous findings of conserved uAUG-initiated ORFs in *Cryptococcus* species (69).

Translation initiation at uAUGs results in the translation of uORFs, which can regulate translation of the main ORF (41,70). To evaluate the functionality of the uAUGs in *Cryptococcus*,

we generated riboprofiling data in both species and compared densities of ribosome-protected fragments with those of sample-matched poly(A)+ RNA. Our riboprofiling data passes quality metrics of 3-nucleotide periodicity of reads on ORFs indicating active translation by ribosomes, and appropriate read lengths of 26-30nt (Figure 1.10).

Most genes have ribosome occupancy close to that predicted by their RNA abundance, and restricted to the main ORF, for example the most highly translated gene, translation elongation factor eEF1α/CNAG_06125 (Figure 1.2A,B). However, we observed dramatic examples of translation repression associated with uORFs in CNAG_06246, CNAG_03140 in *C. neoformans* (Figure 1.2A,C,D). These patterns are conserved in their homologs in *C. deneoformans* (Figure 1.11). Other spectacularly translationally repressed genes, CNAG_07813 and CNAG_07695 and their *C. deneoformans* homologs (Figure 1.2A, 1.11A) contain conserved uORFs in addition to 5' introns with alternative splicing or intronically expressed non-coding RNAs (Fig 1.12). In all these cases, high ribosomal occupancy on one or more uORFs is associated with low occupancy of the main ORF.

The uncharacterized gene CNAG_06246 has two AUG-encoded uORFs that are occupied by ribosomes, and a predicted C-terminal bZIP DNA-binding domain. This gene structure is reminiscent of the multi-uORF-regulated amino-acid responsive transcription factors Gcn4/Atf4 (70), or the *S. pombe* analog Fil1 (71). The sugar transporter homolog CNAG_03140 has six uAUGs, with substantial ribosome occupancy only at the first. Interestingly, *N. crassa* has a sugar transporter in the same major facilitator superfamily regulated by a uORF (rco-3/ sor-4, (72)), and sugar-responsive translational repression via uORFs has been observed in plants (73).

Since these translationally repressed genes have multiple uAUGs, we investigated the relationship between uAUGs and translation efficiency genome-wide. We observed a clear

negative relationship between the number of uAUGs and translation efficiency (Figure 1.2E, 1.11E), suggesting an uAUG-associated negative regulation of translation in both species.

Position relative to the TSS affects uAUG translation.

Although some uAUGs are recognized and efficiently used as translation start sites, some others are used poorly or not at all, and allow translation of the main ORF. We thus analyzed *Cryptococcus* uAUG position and sequence context to see how translation start codons are specified in these fungi.

We compared the translation efficiency of genes containing only uAUGs close to the TSS to those with uAUGs far from the TSS. In *C. neoformans*, 1,627 of the 10,286 uAUGs are positioned within the first 20 nt of the TL, and 816 uAUG-containing genes have no uAUG after this position. The presence of one or several uAUGs close to the TSS (<20 nt) has nearly no effect on translation efficiency, whereas genes containing uAUGs far from the TSS are less efficiently translated (Figure 1.2F), and similarly in *C. deneoformans* (Figure 1.11F).

A Kozak sequence context determines AUG translation initiation.

To analyze the importance of AUG sequence context for translation initiation in *C. neoformans*, we used the 5% most translated genes (hiTrans n = 330) to construct a consensus sequence surrounding their annotated translation start codon (Figure 1.3A). The context contains a purine at the -3 position, a hallmark of the Kozak consensus sequence (24). However, there is very little enrichment for the +4 position, in contrast with the mammalian Kozak context in which a G is present in +4 ((A/G)CCAUGG) (24). Because of the limited sequence bias downstream of the AUG, and its confounding with signals of N-terminal amino acids and codon usage, we do not consider it further. However, we found a slight sequence bias in the positions -10 to -7 that is outside the metazoan Kozak context.

We thus calculated “Kozak scores” for all uAUGs against the position weight matrix (pwm) of the Kozak context from -10 from AUG through to AUG (Figure 1.3A). We compared the Kozak scores of the annotated AUGs (aAUGs) with those of the 5% most highly translated genes, the first upstream AUG (uAUGs) and the first downstream AUG (d1AUG). Highly translated aAUGs have a higher score than typical aAUGs, and aAUGs have usually a higher score than the uAUGs and d1AUGs (Figure 1.3B). On a given transcript, the uAUG score is usually lower than the aAUG score (Figure 1.3C).

We next asked if the sequence context of uAUGs affected their ability to repress translation of the annotated ORF, focusing on transcripts with only a single uAUG. Surprisingly, there is a weak and insignificant correlation between uAUG Kozak score and the translation efficiency of the aORF, whether the uAUG is close to or far from the TSS (Figure 1.3D). However, the most striking examples of translational repression in Figure 1.2 tend to have multiple high-score uAUGs (scores CNAG_06246, u1AUG 0.93, u2AUG 0.86; CNAG_03140, u1AUG 0.85, u2AUG 0.76; CNAG_07813, u1AUG 0.79; CNAG_07695, u1AUG 0.97, u2AUG 0.90). This is consistent with direct biochemical evidence that AUG context determines translation repression by uORFs in *N. crassa* and *S. cerevisiae* (74).

We also asked if the AUG score affects the AUG usage transcriptome-wide, by comparing the difference in u1AUG and aAUG scores with the ratio in A-site ribosome occupancy in a 10-codon neighbourhood downstream of the u1AUG and aAUG. We considered the relative occupancy to control for transcript-specific differences in abundance and cap-dependent initiation-complex recruitment. We restrict our analysis to a short neighborhood to control for start-codon specific biases in ribosome occupancy caused by addition of cycloheximide prior to cell lysis (57,60). A higher score difference is associated with higher relative ribosome occupancy, while the control comparison with RNA-Seq coverage shows a

smaller effect (Figure 1.3E). We find the same patterns of AUG consensus, scores, and occupancy in *C. deneoformans* (Figure 1.13).

Nonsense-mediated decay acts on uORF-containing genes.

An mRNA molecule translated using an uAUG can be recognized as a premature stop codon bearing molecule and may be as such degraded by the nonsense-mediated mRNA decay (NMD) (75). In *S. cerevisiae*, uORFs are associated with NMD genome-wide (76). To test this concept in *Cryptococcus*, we first sequenced RNA from *C. deneoformans* strains with the conserved NMD factor Upf1 deleted (34), finding 370 genes with increased mRNA abundance and 270 with decreased (Figure 1.4A, Table 1.2; 2-fold difference in levels at 1% FDR).

We next compared the fold-change in abundance of uORF-containing or uORF-free mRNAs. Two genes with extreme increases in upf1Δ are also extremely translationally repressed uORF-containing genes we identified above (Figures 1.2, 1.11, 1.12): CNF00330 (CNAG_07695 homolog, 11-fold) and CNG04240 (CNAG_03140 homolog, 8-fold). Another extreme is the carbamoyl-phosphate synthase CND01050 (5-fold up in upf1Δ), a homolog of *S. cerevisiae* CPA1 and *N. crassa* arg-2. These orthologs are regulated by a conserved uORF encoding a arginine attenuator peptide that have all been verified to repress reporter gene synthesis in a *N. crassa* cell-free translation system (77); both *S. cerevisiae* and *N. crassa* orthologs are NMD substrates, which for ScCPA1 depends on the uORF (78,79). Consistent with this model, in both *C. neoformans* and *C. deneoformans* the native uORF shows strong ribosome occupancy while the aORF is translationally repressed (CnTE = 0.47, CdTE = 0.38; Figure 1.14).

In general, uORF-containing genes are more likely to be upregulated in the upf1Δ mutant than uAUG-free genes (Figure 1.4B), suggesting that uORFs negatively regulate mRNA abundance in *Cryptococcus*, in addition to repressing translation of the main ORF. Similarly,

uORF-containing genes are enriched for NMD-sensitivity only when the uAUG is more than 20 nt from the TSS (Figure 1.4C), suggesting that TSS-proximal uAUGs (< 20 nt) are skipped, and generally not used as translation start codons in *Cryptococcus*.

Next, we asked if uAUG Kozak score affects mRNA decay via the NMD pathway. Restricting our analysis to genes with a single uAUG (n=1,421), we binned genes according to their Kozak score. We find that mRNAs that contain higher Kozak-score uAUG are more likely to increase in abundance in the upf1Δ mutant (Figure 1.4D). Indeed, the abundance increase is monotonically correlated with the mean of the score bins. This could explain the weak effect size of uAUG score on translation efficiency (Figure 1.3D), as higher-scoring uAUGs repress the RNA abundance (denominator of TE) in addition to repressing translation of the main ORF (numerator).

In conclusion, in *Cryptococcus*, the position and the sequence context of uAUGs determines their usage as translation start codons, and the effect of the uORF on stimulating nonsense-mediated decay of the mRNA.

Start codon sequence context and uORF regulation in other fungi

We then examined sequences associated with translation start codons in other fungi, for which both RNA-Seq and Riboprofiling data were available, and for which the annotation was sufficiently complete (i.e. *S. cerevisiae; Neurospora crassa, Candida albicans* and *Schizosacchomyces pombe*). We analyzed the Kozak context associated with aAUG of all annotated coding genes, of the 5% most translated genes (hiTrans), and for mRNAs encoding cytoplasmic ribosomal proteins (CytoRibo), as a model group of highly expressed and co-regulated genes defined by homology (Table 1.3). Cytoplasmic ribosomal proteins have informative Kozak contexts, with a strong A-enrichment at the positions -1 to-3 and weak sequence enrichment after the AUG in all these species (Figure 1.5A). The total information

content of the Kozak sequence is higher for CytoRibo genes than HiTrans, and higher for HiTrans than all annotated genes, in all these fungi (Figure 1.5B). Nevertheless, these contexts have also some species specificity: Kozak sequences for HiTrans and CytoRibo are more informative in *Cryptococcus* and *N. crassa* than in *S. pombe, C. albicans,* and *S. cerevisiae*. In particular, the C-enrichment at positions -1, -2 and -5 in *Cryptococcus* and *N. crassa* is absent in *S. cerevisiae*, and we observed no sequence enrichment upstream of the -4 position for *S. pombe* and very little for *S. cerevisiae*. In contrast, a -8 C enrichment, similar to the *Cryptococcus* and mammalian pattern, was observed in *N. crassa*, confirming previous results (80). The -10:-6 A/T rich region for *C. albicans* is likely to reflect an overall A/T-richness of the TLs in *C. albicans*.

The analysis of the TL sequences from these fungi, excluding *C. albicans* for which no TL annotation is available, also shows species specificity. The average TL length in *S. cerevisiae* (84 nt) is less than half that in *Cryptococcus* (Figure 1.5C). In sharp contrast with *Cryptococcus*, only 985 uAUGs are present in 504 genes, which correspond to 18% of the genes with an annotated TL in *S. cerevisiae*. Moreover, the density of the uAUGs is very low and uAUGs have no global effect on TE in this yeast (Figures 1.5D,E, 1.15). The short uAUG-depleted TLs observed in the SGD annotations of *S. cerevisiae* are conserved in a recent annotation of other *Saccharomyces* species (81) (Figure 1.16).

More broadly, short TLs with very low uAUG density are more the exception than the rule in the fungal kingdom (Figure 1.5C). However fungi vary in how much these uAUGs globally down regulate gene translation (Figure 1.15). Our analysis shows that fungi quantitatively vary in the sequence context of the AUGs that they use, and in the distribution of AUGs in their TLs. Thus, distinct fungi may differ in how much they use AUG sequence context to regulate gene expression at the post transcriptional level.

Kozak context programs leaky scanning in Cryptococcus

We earlier calculated the Kozak score of the first downstream AUG (d1AUG) within each CDS: these d1AUG scores are mostly lower than the score of the aAUGs (Figure 1.3B), consistent with most annotations correctly identifying a good-context AUG as the start codon. Yet, we identified a number of d1AUGs with a high score (n=1109 above 0.826, the median Kozak score for aAUGs; n= 131 above 0.926, the median for hiTrans), which could be efficiently used as a translation start codon. The scanning model of translation initiation predicts that the d1AUG will be used as the start codon only by pre-initiation complexes that leak past the aAUG, which is unlikely if the aAUG has a strong sequence context.

To identify probable leaky translation initiation events, we thus compared the aAUG and d1AUG scores within each of the 50% most abundant mRNAs (Figure 1.6A). For above-median aAUG score genes, the score of the d1AUGs can be very high or very low. By contrast, for the genes with a low aAUG score, there is a bias toward higher d1AUG score, suggesting that for these genes the strong d1AUG could be used as alternative translation start site (Figure 1.6A).

To test whether AUG score affects translation initiation, we calculated the ratio of ribosome protected fragment density and RNA-Seq density around each aAUG and d1AUG on the same mRNA, and the difference in score between these two AUGs (Figure 1.6B), using the same 10-codon neighborhood as for our earlier uAUG-aAUG comparison. We found a weak positive correlation between the difference in scores of the two AUGs and RNA-Seq density at these specific loci, raising the possibility that transcription start sites sometimes occur downstream of a weak aAUG. The relative ribosome density is equal on average when the d1AUG score is less than the aAUG score. However, a nonlinear generalized additive model shows that the relative density sharply increases at d1AUGs when their score increases above that of the aAUG. This suggests that for these genes, both AUGs can be used as translation

start codon, because a subset of scanning ribosomes leak past a lower-score aAUG and then initiate at the higher-score d1AUG.

Kozak context-controlled scanning specifies alternative N-termini in *Cryptococcus* and *Neurospora*

We next determined which groups of genes could be affected by potential alternative start codon usage. We focused our analysis on the 50% most abundant RNAs for which the difference in score between the aAUG and d1AUG was the highest (difference in wide score d1AUG – aAUG > 0.1, n = 167 for *C. neoformans*) (Table 1.4). Strikingly, for 66% of these genes (110/167) the d1AUG is in frame with the corresponding aAUG, with a median of 69 nt (mean 79 nt) between the two AUGs. Thus, alternative usage of in-frame AUGs would result in proteins with different N-terminal ends. Supporting this hypothesis, 37% of these proteins (41/110) possess a predicted mitochondrial targeting sequence located between the two AUGs, far exceeding the 8% genome-wide (560/6788). This suggests that the usage of the annotated start codon would target the isoform to mitochondria, whereas the usage of the d1AUG would produce a protein specific to the cytoplasm or another organelle. Examples of alternative localization driven by alternative N-termini have been observed across eukaryotes (82,83)

The pattern of predicted dual-localization, i.e. enrichment of high-score d1AUGs in-frame with predicted mitochondrial localization signal on the longer N-terminal, is conserved in some fungi but not others (Figure 1.6C). In a null model where coding sequences have random nucleotide content, we would expect roughly 1/3 of d1AUGs to be in frame. In 6 fungal species we examined, for d1AUGs whose score is comparable to or less than the aAUG they follow, the proportion in frame is close to (*Cryptococcus, N. crassa*) or less than 1/3. These proportions are similar when we considered high abundance (top 50%) or low abundance (bottom 50%)

mRNAs. The pattern differs for mRNAs with a d1AUG whose score is high relative to the aAUG they follow (d1AUG score > aAUG score + 0.1). In *Cryptococcus* and *N. crassa*, most high abundance mRNAs are in-frame and over 1/3 of these in-frame high-score d1AUGs have predicted mitochondrial localization. In *S. cerevisiae* and *C. albicans*, we observe only a slight relative enrichment for high-scoring d1AUGs to be in-frame and to follow a mitochondrial targeting sequence. By contrast, in *S. pombe* we see depletion in the in-frame/out-of-frame ratio, even in these proteins with high-scoring d1AUGs.

These results suggest that the extent to which alternate translation start codons regulate proteome diversity is variable in fungi. Accordingly, we identified a number of *Cryptococcus* proteins with potential alternative start codons and N-terminal targeting sequences, whose two homologs in S. cerevisiae are known to be necessary in two compartments of the cells. For instance, CnPUS1/CNAG_06353 is an homolog of both the mitochondrial and cytoplasmic tRNA:pseudouridine synthases encoded by the PUS1 and PUS2 paralogs in *S. cerevisiae*. In *C. neoformans*, ribosome occupancy at both the aAUG and d1AUG of CNAG_06353, and the presence of transcription start sites both sides of the aAUG (Figure 1.6D, 1.17A), argues that both AUGs are used as start codons, and transcription and translation regulation could co-operate to set isoform levels. Similarly, CnGLO1/CNAG_04219 encodes both the cytoplasmic and nuclear isoforms of the glyoxalase I depending on the alternate AUG usage (Figure 1.17B). The next enzyme in this pathway, Glyoxalase II, is likewise encoded by CnGLO2/CNAG_01128, which is a homolog of both cytoplasmic (Glo2) and mitochondrial (Glo4) enzymes in *S. cerevisiae*. CNAG_01128 has a weak aAUG, strong d1AUG, and N-terminal predicted mitochondrial targeting sequence (Figure 1.17C). Finally, we observed that nine members of the amino-acyl tRNA synthetase gene family have predicted alternate localization from alternate AUG start codons (Figure 1.7A/B).

Amino-acyl tRNA synthetases (aaRSs) are frequently single-copy and dual-localized in Cryptococcus

The tRNA charging activity of aaRSs is essential in both cytosol and mitochondria to support translation in each compartment, and examples of alternative localization of two aaRS isoforms of a single gene have been observed in fungi, plants, and animals (84-86). This implies that a eukaryote with a single genomic homolog of an aaRS activity is likely to make distinct localized isoforms from that locus. Thus, we examined predicted aaRS localization in fungi. We assembled gene lists of aaRSs in diverse fungi from homology databases OrthoDB (64) and PANTHERdb (65), adding a mitochondrial SerRS (CNAG_06763/CNB00380) to the list of *Cryptococcus* aaRSs analysed by Datt and Sharma (87).

In *C. neoformans* and *C. deneoformans*, eleven aaRSs are each expressed from a single genomic locus, including the homologs of all five *S. cerevisiae* aaRSs whose dual-localization has been verified (Table 1.5). Nine of these *Cryptococcus* aaRSs have the same structure of a poor-context annotated AUG followed by a predicted mitochondrial targeting sequence and a strong-context d1AUG (Figure 1.7A/B; AlaRS, CysRS, GlyRS, HisRS, ValRS, LysRS, ProRS, ThrRS, TrpRS). The similar annotated AUG contexts, sharing an unfavourable -3U, suggests that the same mechanism could lead to leaky translation initiation at most of these (Table 1.6). At the downstream AUGs, the strong Kozak context is consistent with efficient translation initiation of the cytoplasmic isoform from this start codon (Table 1.6).

The two remaining single-copy aaRSs have near-AUG translation initiation sites upstream of predicted mitochondrial targeting sequences. Translation of ArgRS starts at an AUU codon with otherwise strong context (cccaccAUU) conserved in both *Cryptococcus* species. This N-terminal extension includes a predicted mitochondrial targeting sequence (mitofates $p > 0.95$ for both species). Translation of LeuRS starts at adjacent ACG and AUU codons which collectively provide strong initiation context (gccaccACGAUU in *C. neoformans*, gccACGAUU in

C. deneoformans). This N-terminal extension also includes a predicted mitochondrial targeting sequence (mitofates p≈0.7 for both species).

In *Cryptococcus*, alternative aaRS isoforms appear to be mostly generated by alternative translation from a single transcript, and sometimes by alternative transcription start sites. On all the predicted dual-localized aaRSs, we observe ribosomal occupancy starting at the earliest start codon (Figure 1.7C/D, 1.18). LysRS/CNAG_04179 contains only a single cluster of transcription start sites, upstream of the aAUG (Figure 1.7C). ProRS/CNAG_04082 contains a wider bimodal cluster of TSSs, both upstream of the aAUG. Similarly, most transcription initiation is well upstream of the aAUG in CysRS/CNAG_06713, LeuRS/CNAG_06123, ThrRS/CNAG_06755, and ValRS/CNAG_07473. However, for GlyRS/CNAG_05900, and HisRS/CNAG_01544, we observe alternative transcription start sites closely upstream of the annotated start codon, that are likely to affect the efficiency of start codon usage. In ArgRS/CNAG_03457 there is also an alternative transcription start site, close to the near-AUG start codon for the mitochondrial form. In AlaRS/CNAG_05722 and TrpRS/CNAG_04604 we detect some transcription start sites between the alternative start codons, and TrpRS also has an uORF in the transcript leader that is likely to affect translation. These observations suggest that dual-localization of the single-copy aaRSs in *Cryptococcus* is regulated largely by start codon choice. For some genes this regulation is backed up by alternative TSS usage.

Some dual-localized genes use an upstream near cognate codon (DualNCC) in all these fungi, but the NCC-initiated aaRS are not the same from one fungus to the other. For instance, both *Cryptococcus* and *N. crassa* AlaRS use DualAUG whereas in *S. pombe*, *S. cerevisiae* and *C. albicans* a DualNCC is used. On the other hand, *S. pombe* GlyRS is regulated by DualNCC whereas the other ones use a DualAUG regulation. Substitution between weak AUG codons and near-cognate codons seems thus to have taken place multiple times in the fungal kingdom.

Amino-acyl tRNA synthetases as an evolutionary case study

To understand patterns of dual-localization, we next examined the evolution of aaRSs. The ancestral eukaryote is thought to have had two complete sets of aaRS, one mitochondrial and one cytoplasmic, but all mitochondrial aaRSs have been captured by the nuclear genome and many have been lost (88). Thus we examined aaRS phylogenetic trees in more detail. For some amino acids (Asn, Asp, Glu, Iso, Met, Phe, Ser, Tyr), reference fungi have distinct cytoplasmic and mitochondrial aaRSs that cluster in separate trees (89). We also do not consider Gln, because organellar Gln-tRNA charging in some species is achieved by an indirect pathway (90).

Dual-localized AlaRS, CysRS, and HisRS in the 6 fungi we focus on are each monophyletic (89). Even these aaRS can be encoded by two genes in some other fungi: AlaRS is duplicated to one exclusively mitochondrial and another exclusively cytoplasmic gene in the Saccharomycete yeast *Vanderwaltozyma polyspora* (91). For CysRS, *Aspergillus versicolor* (ASPVEDRAFT_141527 and ASPVEDRAFT_46520) and *Coprinus cinerea* (CC1G_03242 and CC1G_14214) have two copies, one of which has a predicted mitochondrial targeting sequence. For HisRS, *Rhizopus delemar* (RO3G_01784 and RO3G_16958) and *Phycomyces blakesleeanus* (PHYBL_135135 and PHYBL_138952) likewise contain gene duplications. Similarly, *S. cerevisiae* has two ArgRS genes that arose from the whole-genome duplication: RRS1/YDR341C is essential, abundant, and inferred to be cytoplasmic (92) while MSR1/ YHR091C has a mitochondrial localization sequence and MSR1 deletions have a petite phenotype (93), although both have been detected in mitochondria suggesting some residual dual-localization of the cytoplasmic enzyme (94). The second *S. cerevisiae* stress-responsive cytoplasmic copy of GlyRS also arose from the whole-genome duplication (95). *S. pombe* cytoplasmic ValRS is monophyletic with dual-localized ValRS in other fungi, and

Schizosaccharomyces also has a paralogous but diverged mitochondrial ValRS that appears to be descended from an early eukaryotic ValRS of mitochondrial origin (96).

LysRS appears to have been duplicated in an ancestor of ascomycetes: ascomycete mitochondrial homologs cluster together, and ascomycete cytoplasmic homologs cluster together, while the single basidiomycete homolog clusters close to the base of this split from other opisthokonts (89). By contrast, LeuRS, ProRS, and TrpRS are each represented by two distinct proteins in ascomycetes, one cytoplasmic and one mitochondrial and of independent descent, but the mitochondrial homolog has been lost in *Cryptococcus* species. In basidiomycetes *Ustilago* and *Puccinia*, homologs of mitochondrial LeuRS and ProRS are not present, but there is a homolog of mitochondrial TrpRS; all these have a single homolog of the cytoplasmic TrpRS (89). Our independent phylogenetic analysis of LysRS and ProRS agrees with the conclusions from PANTHERdb (Figures 1.7E/F). These analyses show that aaRSs have undergone multiple incidences of at least two processes during fungal evolution: losses associated with the dual-localization of the remaining gene, and duplications followed by specialization.

Evolutionary conservation of gene-specific feedback regulation by alternate AUG usage

We also observed striking examples of gene-specific regulation by start codon context in *Cryptococcus*, in translation factors affecting start codon selection, supporting previously proposed models of feedback regulation (97,98).

Translation initiation factor eIF1, which enforces the accurate selection of start codons, is encoded by an mRNA with poor start codon context in diverse eukaryotes, driving an

autoregulatory program (97,99). In *C. neoformans*, eIF1 (SUI1/CNAG_04054) also initiates from a poor-context cuuaguugaAUG start (score 0.75), and ribosome profiling reads are spread across the annotated ORF (Figure 1.8A). Intriguingly, the next AUG is out-of frame and has strong context cuccaaaaAUG (score 0.98), with a same-frame stop codon 35 codons later, suggesting that this could represent a downstream short ORF that captures ribosomes that have leaked past the poor-context start. To test this hypothesis, we examined the 5' ends of riboprofiling reads, which report on the translation frame of the ribosomes (41). Riboprofiling reads from the 5' and 3' of the eIF1 annotated ORF are roughly 77% in frame 0, 10% in +1, and 13% in +2, as are reads on two other highly expressed genes, eEF1α and HSP90. By contrast, in the hypothesized downstream ORF, reads are only 57% in frame 0, 32% in frame +1, and 11% in frame +2, consistent with translation occurring in both frame 0 and +1. The gene structure is conserved in *C. deneoformans* eIF1 (CNB05380), with a weak aAUG (score 0.76), a strong d1AUG (score 0.98) in the +1 frame, followed by an enrichment in +1-frame riboprofiling reads (Figure 1.19A,B). We observe small increases in eIF1 mRNA levels in the upf1Δ strain of *C. deneoformans* at both 30°C (1.16x, p=0.04) and 37°C (1.09x), so NMD could regulate this transcript. Overall, our data support the hypothesis that the downstream ORF of eIF1 is translated after leaky scanning past the annotated AUG, and that the downstream ORF contributes to translation regulation of the annotated ORF.

Translation initiation factor eIF5 reduces the stringency of start codon selection, and is encoded by an mRNA with a repressive uORF initiated from a poor-context uAUG in diverse eukaryotes (98). In *C. neoformans*, eIF5 (TIF5/CNAG_01709) also contains a uAUG with the poor sequence context aaagaguucAUG (score 0.72), while the main ORF of eIF5 is initiated by a strong context cccgcaaaAUG (score 0.94). We detect ribosomal density on the uORF of TIF5 comparable to that on the main ORF (Figure 1.8C), suggesting substantial translation initiation at the uAUG, while there is also clear translation initiation at a further upstream CUG codon. The gene structure is conserved in *C. deneoformans* eIF5 (CNC02150), with the same pattern

of riboprofiles at upstream poor-context AUG and near-cognate codons (Figure 1.19C). Further, the *C. deneoformans* homolog transcript abundance increases substantially in the upf1Δ strain (2.6x, p < 10-50). In *N. crassa*, eIF5 has two uORFs and direct analysis of mRNA stability indicated that its transcript is a NMD target (79). The present data support the model that eIF5 translation in *Cryptococcus* is also repressed by upstream reading frames initiated from poor start codons, leading to nonsense-mediated decay of the transcript.

Variable inserts in eTIFs correlate with variation in translation initiation determinants

The conserved proteins eIF1, eIF5, and eIF1A play pivotal roles in start codon selection, and specific mutations in these factors give rise to suppressor of upstream initiation codon (Sui-) phenotypes and their suppressors (Ssu-) (99). To ask if between-species variability in start codon preference is linked to these initiation factors, we generated multiple sequence alignments of their homologs in fungi.

Translation initiation factor eIF1 shows striking sequence variation across fungi, notably at multiple *Cryptococcus*-specific sequence insertions that result in a 159-aa protein substantially larger than the 108-aa *S. cerevisiae* homolog (Figure 1.9A). Variation in eIF1 occurs at and around positions known to modulate start codon selection in *S. cerevisiae* (99). For instance, a T15A substitution increases fidelity in SceIF1 (99), and an analogous T15A substitution is present in eIF1s from *Neurospora* and other filamentous fungi, while both *Cryptococcus* homologs have the T15V substitution. The three fungi that tend not to use alternative AUG start codons in the regulation of proteome diversity, *S. cerevisiae*, *C. albicans*, and *S. pombe*, all have a threonine residue at position 15. Variation in fungal eIF1 extends far beyond this N-terminal region: similar patterns of sequence diversity occur at the positions E48, L51, D61 that have been shown to increase fidelity in SceIF1 (99). By contrast, positions K56,

K59, D83, Q84, at which mutations have been shown to reduce fidelity in SceIF1 (99), are highly conserved in fungi.

We next tested how the translation pre-initiation complex could be affected by the insertions in *Cryptococcus* eIF1 using published structures of the *S. cerevisiae/K. lactis* “Pin” complex engaged in the act of AUG selection (100). We found that the insertions in eIF1 are facing either the methionine initiator tRNA (tRNAi) or the solvent-exposed side (Figure 1.9B). The N-terminal insertion is not visible in the structure, but could be close to the acceptor arm of tRNAi. The N-proximal loop insertion of CneIF1 extends from the SceIF1 sequence (18-DETATSNY-25) that contacts the acceptor arm of tRNAi. The CneIF1 insertion in loop 2 extends the SceIF1 loop 2 (70-KDPEMGE-76) that contacts the D-loop of tRNAi; substitutions D71A/R and M74A/R increase the charge of SceIF1 loop 2 and increase initiation at UUG codons and weak AUG codons (101). CneIF1 loop 2 has substitutions at both these functionally important sites, and is extended by a further 14 hydrophobic and negative residues. The last insertion in CneIF1 extends a loop facing the solvent-exposed surface of SceIF1. Collectively, this shows that there are likely major differences in the eIF1-tRNAi interaction surface in *Cryptococcus* relative to other fungi, an interaction critical for start codon selection (101).

The N-terminal domain of eIF5 (eIF5-NTD) replaces eIF1 upon start codon recognition, and we found between-species variation in CneIF5 at tRNAi interaction surfaces corresponding to variability in CneIF1 (Figure 1.9C,1. S11A). SceIF5 Lys71 and Arg73 in loop 2 make more favourable contacts with the tRNAi than the corresponding residues of SceIF1, so that the shorter loop 2 of SceIF5 may allow the tRNAi to tilt more towards the 40S subunit (28). Although Arg73 is conserved across fungi, Lys71 is absent in CneIF5 loop 2 (67-SMAN-70), which is two amino acids shorter than SceIF5 loop 2 (66-SISVDK-71). Collectively, the longer loop 2 of CneIF1 and the shorter loop 2 of CneIF5 suggest that the conformational changes

accompanying start codon recognition may be more exaggerated in *Cryptococcus*, providing a mechanistic hypothesis for stronger genomic patterns of start codon recognition.

Fungal eIF1A homologs also diverge from SceIF1A at regions that modulate translation initiation fidelity (Figure 1.20B), for example the N-terminal element DSDGP (99). The *Cryptococcus* eIF1A C-terminus is diverged from all other fungi at SceIF1A positions 110-120, and along with other basidiomycetes lacks a loop at SceIF1A positions 135-149. This C-terminal region of SceIF1A contributes to pre-initiation complex assembly and binds eIF5B (102) and eIF5 (103), and domain deletions or local alanine substitutions reduce fidelity of translation start site selection (99,102,104).

Thus, although structural analysis of the Cryptococcal initiation complex will be required for a detailed mechanistic understanding, our initial analysis suggests that sequence variability in fungal eIFs could plausibly account for differences in start codon selection between different species.

Discussion

Our annotation of transcript structure and translation in two pathogenic *Cryptococcus* species and our analysis of published data from other species show that start codon context has a major effect on protein production, regulation, diversity, and localization in diverse fungi. As such this work represents a useful resource for the field. While the genome-wide effect of start codon context is weak in *S. cerevisiae* (14), we find that other fungi, from *Neurospora* to *Cryptococcus*, use start codon context to regulate translation initiation to a far greater extent. These fungi have long and AUG-rich TLs, and more information-rich and functionally important Kozak sequences. Further, *Cryptococcus* and *Neurospora* display extensive evidence of leaky scanning of weak AUG codons that is used for regulation by upstream ORFs and to generate alternate N-terminal isoforms with different subcellular localization.

Widespread leaky scanning controlled by start codon context in *C. neoformans*

Translation initiation regulation can be enabled by start codons that are imperfectly used, so that scanning pre-initiation complexes can leak past them. According to the scanning model of translation initiation, a “perfect” strong start codon would prevent this by capturing all the scanning PICs, and leave few for downstream initiation. For example, the downstream out-of-frame ORF of *Cryptococcus* eIF1 is likely to be translated only by PICs that leak past the annotated AUG. The alternative second in-frame AUG of dual-localized proteins is also initiated only by PICs that have leaked past the initial AUG. Our data show this leakiness-driven dual-localization is common in *Cryptococcus*, in addition to being conserved across eukaryotes in gene classes such as tRNA synthetases. Our data also argue that AUGs that are proximal to the 5’ cap, or that have poor sequence context, are commonly leaked past in *Cryptococcus*, as shown previously in studies of yeast (105) and mammals (12,106). We note that leakiness-driven translation regulation is not the only mechanism regulating alternative translation from a

single mRNA and is distinct from those that depend on either blocking scanning, or on recycling of post-termination ribosomes such as in the case of *S. cerevisiae* GCN4 (70).

Functional role of start codon context varies across the fungal kingdom

Cryptococcus and *Neurospora* have long TLs that are AUG-rich, and extended start codon context sequences that suggest a higher ability to discriminate against poor-context AUGs. Several lines of evidence argue that the efficiency with which upstream AUGs capture initiation complexes is determined by the AUG sequence context, notably in vitro translation studies in *N. crassa* and *S. cerevisiae* from the Hinnebusch and Sachs labs (74). The most spectacular examples of uORF-associated translation repression in *Cryptococcus* are associated with good-context uAUGs with high ribosome occupancy. However, such strong-context high-occupancy uAUGs are rare. In *Cryptococcus* and *Neurospora*, the leakiness of potential AUG translation start sites is also extensively used to diversify the proteome by alternative N-terminal formation.

In comparison, *S. cerevisiae, S. pombe* and *C. albicans* appear to be less efficient in discriminating AUGs based on their sequence context. *S. cerevisiae* has minimized the possibility of regulation of translation by uORFs: it has unusually short TLs, these TLs are unusually AUG-poor, uAUGs tend to have poor context, and there is no statistical association between uAUG score and translation efficiency of the main ORF. Reporter gene studies (15,16) and classic examples such as GCN4 show that uAUGs can repress translation in *S. cerevisiae*, but genome-wide analysis show that this is rare during exponential growth in rich media (Fig S5.1). Recent work on meiosis (107) and stress (108) shows that 5'-extended transcript leaders that contain repressive uAUGs ("long undecoded transcript isoforms") are more common during alternative growth conditions for this yeast. Moreover, in *S. cerevisiae*, near-cognate codons appear to be more common starts for alternative N-terminal formation (109). This suggests that leaky scanning from near-cognate codons, more than from AUGs, might be an important mode

of regulation in *S. cerevisiae*. The situation is different in *S. pombe*, which has long AUG-rich TLs but is depleted for downstream in-frame AUGs. Consequently, uAUGs globally repress aORF translation, but do not appear to regulate alternative protein production through alternative AUG start codons. We speculate that the comparatively uninformative Kozak context in *S. pombe* might be variable enough to regulate translation initiation rate but not proteome diversity.

We found that multiple near-cognate start codons are used for leaky initiation in *Cryptococcus*: ACG for the mitochondrial isoform of LeuRS, AUU for the mitochondrial isoform of ArgRS, and the upstream CUG in eIF5. Further work will be needed to quantify the extent of near-cognate start codon usage in *Cryptococcus* in different growth conditions and to compare it to other organisms (20,110).

Leaky scanning through weak AUGs could regulate the mitochondrial proteome

We computationally predicted dozens of dual-localized proteins with alternative start codons that confer an N-terminal mitochondrial targeting sequence in their longest isoform. We did not identify enrichment of proteins with predicted dual-localization in the cytoplasm and in the nucleus, or with a signal peptide followed by an alternative start codon (data not shown). Thus, increasing the efficiency of weak-context to strong-context translation initiation would predominantly upregulate a regulon consisting of the mitochondrial isoforms of dozens of proteins.

Mechanisms to control initiation efficiency of a mitochondrial-localized regulon could include intracellular magnesium concentration (111), variations in availability or modification status of shared initiation factors, variations of the ratio of mitochondrial volume to intracellular volume (112), or specialized factors to promote initiation specifically of mitochondrial isoforms

with their specialized start codon context. Nakagawa et al (113) previously suggested that distinct Kozak contexts might be recognized by different molecular mechanisms.

One candidate mechanism involves the translation initiation factor 3 complex, which has a role in regulating the translation initiation of mitochondrial-localized proteins across eukaryotes. In *S. pombe*, subunits eIF3d/e promote the synthesis of mitochondrial electron transfer chain proteins through a TL-mediated mechanism (114). In *S. cerevisiae* and *Dictyostelium discoideum*, the conserved eIF3-associated Clu1/CluA protein affects mitochondrial morphology (115), and the mammalian homolog CLUH binds and regulates mRNAs of nuclear-encoded mitochondrial proteins (116,117). Metazoans have 12 stably-associated subunits of eIF3, which are conserved in most fungi including *N. crassa* (118), *Cryptococcus*, and the *Saccharomycetale* yeast *Yarrowia lipolytica* (Table 1.7). Interestingly, species that tend not to use alternate AUG codons for dual-localization have lost eIF3 subunits: eIF3d/e/k/l/m are lost in *C. albicans*, and additionally eIF3f/h in the related *S. cerevisiae*; *S. pombe* has independently lost eIF3k/l (Table 1.7; (89)). Further work will be needed to investigate the role of eIF3 in regulating mitochondrial- and dual-localized proteins in the fungal kingdom.

How could evolutionary plasticity of translational initiation in the fungal kingdom have arisen?

Selection on genome compaction in unicellular yeasts, which has independently led to gene loss and high gene density in multiple lineages of yeast, could lead to shorter TLs. However, *Saccharomyces*, *Schizosaccharomyces*, and *Cryptococcus* have all independently evolved yeast lifestyles with compact genomes, yet their average TL lengths differ three-fold. Mutations in gene expression machinery, such as the variation in eIF1 noted above, would alter selective pressure on start codon context, and thus uAUG density. Cells have multiple redundant quality control mechanisms, and flexible protein production through leaky scanning

could be buffered by such mechanisms enabling their evolution. Key control mechanisms acting on mRNA, such as RNAi and polyuridylation, have been lost in fungal lineages such as *Saccharomyces*, which might explain their more 'hard-wired' mechanism of translation initiation.

Unexpectedly, highly conserved core translation initiation factors, such as eIF1, have distinctive sequence inserts in *Cryptococcus* that are not shared even by basidiomycetes such as Puccinia and Ustilago. One possibility is genetic conflict, as genetic parasites hijack the gene expression machinery (119). Thus, the unique aspects of the *Cryptococcus* translation initiation machinery could have arisen from a past genetic conflict in which rapid evolution of initiation factors in an ancestor enabled evasion of a genomic parasite (e.g. a mycovirus) that would otherwise hijack initiation.

Materials and Methods

DNA and RNA purification, sequencing library preparation

C. neoformans strain H99 and *C. deneoformans* strain JEC21 were grown in 100 mL YPD at 30°C or 37°C under agitation up to exponential or early stationary phase as previously described (33). Briefly, early stationary phase was obtained after 18 h of growth (final 0D600= 15) starting from at OD600 =0.5. *C. deneoformans* strain NE579 (upf1Δ) (34) was grown in YPD at 30°C under agitation in exponential phase. Each *Cryptococcus* cell preparation was spiked in with one tenth (OD/OD) of *S. cerevisiae* strain FY834 (35) cells grown in YPD at 30°C in stationary phase. Cells were washed, snap frozen and used to prepare RNA and total DNA samples as previously described (36,37). Briefly, total DNA was extracted by bead-beating and phenol:chloroform extraction, and RNA was extracted from lyophilized cells using Trizol. Each condition was used to prepare biological triplicate samples.

For RNA-Seq, strand-specific, paired-end cDNA libraries were prepared from 10 µg of total RNA by polyA selection using the TruSeq Stranded mRNA kit (Illumina) according to

manufacturer's instructions. cDNA fragments of ~400 bp were purified from each library and confirmed for quality by Bioanalyzer (Agilent). DNA-Seq libraries were prepared using the TruSeq DNA PCR-Free kit (Illumina). Then, 100 bases were sequenced from both ends using an Illumina HiSeq2500 instrument according to the manufacturer's instructions (Illumina).

TSS-Seq libraries preparations were performed starting with 75 µg of total RNA as previously described (38) replacing the TAP enzyme by the Cap-clip Pyrophosphatase Acid (TebuBio). For each *Cryptococcus* species we also constructed a control "no decap" library.

Briefly, for these control libraries, poly A RNAs were purified from 75 µg of RNA from *Cryptococcus* and 75 µg of RNA from *S. cerevisiae* before being dephosphorylated using Antarctic phosphatase. Then, *S. cerevisiae* RNAs and one half of the RNAs extracted from *Cryptococcus* were treated with Cap-clip Pyrophosphatase Acid enzyme. The second half of *Cryptococcus* RNAs was mock treated. Each half of Cap-clip Pyrophosphatase Acid *Cryptococcus* RNA samples was mixed with the same quantity of *S. cerevisiae* Cap-clip Pyrophosphatase Acid treated RNAs. The subsequent steps of the library preparation were identical to the published protocol (38). 50 base single end reads were obtained using an an Illumina HiSeq2500 instrument according to the manufacturer's instructions (Illumina).

For QuantSeq 3'mRNA-Seq preparation we followed the manufacturer's instructions for the QuantSeq fwd kit (Lexogen GmbH, Austria). 100 base single end reads were obtained using an Illumina HiSeq2000 instrument according to the manufacturer's instructions (Illumina).

Sequencing data analyses

For TSS analysis we kept only the reads containing both the oligo 3665 (AGATCGGAAGAGCACACGTCTGAAC) and the 11NCGCCGCGNNN tag (38). These sequences were removed and the trimmed reads were mapped to the *Cryptococcus* genome and *S. cerevisiae* genomes using Bowtie2 and Tophat2 (39). Their 5' extremities were

considered as potential TSSs. For each condition we kept only the positions that were present in all three replicates. Their coverage was normalized using the normalization factor used for spiked in RNA-Seq. TSS positions were then clustered per condition. As most of the observed TSS sites appeared as clusters, we grouped them into clusters by allowing an optimal maximum intra-cluster distance (at 50 nt) between sites as previously used (38). We then removed the false TSS clusters using the “no-cap” data keeping the clusters i for which

$$R = \frac{Weight_{cluster_i}}{\sum Weight_{cluster}} / \frac{Weight_{cluster_{nodecap_i}}}{\sum Weight_{cluster_{nodecap}}} > 1$$

Similarly, QuantSeq 3’mRNA-Seq reads containing both the Sequencing and indexing primers (Lexogen) were sorted. The reads were then cleaned using cutadapt/1.18 (40) and trimmed for polyA sequence in their 3’end. PolyA untrimmed and trimmed reads were mapped to the adapted *Cryptococcus* and to the *S. cerevisiae* genomes with Tophat2 (39) with the same setting as for RNA-Seq. To eliminate the polyadenylated reads corresponding to genomic polyA stretches, we considered only the reads that aligned to the genomes after polyA trimming but not before the trimming. The 3’end position of these reads were considered as potential PAS. As for the TSS, for each condition we kept only the positions that were present in all three replicates. Similarly, the PAS dataset was normalized using the spike in normalization factor and the PAS positions were clustered using the same strategies.

Ribosome profiling and matched mRNA-seq

Ribosome profiling was performed on both *C. neoformans* H99 and *C. deneoformans* JEC21, two biological replicates of WT-H99 and one replicate each of H99 ago1Δ and H99 gwo1Δ strains from (31), and one replicate each of WT-JEC21 and JEC21 ago1Δ. We detected

negligible differential abundance between these deletions and their background strains, so in our analyses we treat the deletion strains as biological replicates.

Cells were grown to exponential phase in 750 mL of YPAD with shaking at 30°C. 100 ug/ml cycloheximide (Sigma) (dissolved in 100% ethanol) was added to the culture and incubated for 2 minutes. 50 mL of the culture was withdrawn for performing RNA-Seq in parallel. Cells were then pelleted, resuspended in 5ml of lysis buffer (50mM Tris-HCl pH. 7.5, 150mM NaCl, 10mM MgCl2, 5mM DTT, 0.5% Triton and 100ug/mL cycloheximide) and snap frozen. Lysis, clarification, RNaseI digestion, sucrose gradient separation and monosome isolation was performed as previously described (41).

Ribosome protected fragments were isolated from the monosome fraction using hot phenol. 150ug of the total RNA extracted from the 50 ml of culture in parallel was polyA selected using the Dynabeads mRNA purification kit (Thermo Fisher Scientific) and digested using freshly made fragmentation buffer (100mM NaCO3 pH. 9.2 and 2mM EDTA) for exactly 20 mins.

RNA was resolved on a 15% TBE-Urea gel. A gel slab corresponding to 28-34 nt was excised for footprint samples and around 50 nt for mRNA samples, then eluted and precipitated. Sequencing libraries were generated from the RNA fragments as described in Dunn et al. (42) with the following modifications. cDNA was synthesized using primer oCJ11 (Table 1.8). Two rounds of subtractive hybridization for rRNA removal was done using oligos asDNA1-8 (Table 1.8). After circularization Illumina adaptors were added through 9 cycles of PCR. Libraries were sequenced on a HiSeq 2500 (Illumina).

Ribosome profiling data analysis

Ribosome profiling and matched RNA-seq reads were demultiplexed on BaseSpace (Illumina) and then analyzed essentially with the RiboViz pipeline v.1.1.0 (43). In brief, sequencing adapters were removed with cutadapt (40), and then reads aligned to rRNA were

removed by alignment with hisat2 (44). Cleaned non-rRNA reads were aligned to (spliced) transcripts with hisat2 (44), sorted and indexed with samtools (45), and then quantified on annotated ORFs with bedtools (46), followed by calculation of transcripts per million (TPM) and quality control with R (https://www.R-project.org/, 47) scripts included in RiboViz. The cleaned non-rRNA reads were also aligned to the genome with hisat2, and processed analogously, then used to generate figures of genome alignments using ggplot2 (https://ggplot2.tidyverse.org, 48) in R (47).

Data analysis and visualization

Data analysis and visualization were scripted in R (47), making extensive use of dplyr (https://dplyr.tidyverse.org/, 49), ggplot2 (48), and cowplot (https://CRAN.R-project.org/package=cowplot, 50). Sequence logos were prepared in ggseqlogo (https://CRAN.R-project.org/package=ggseqlogo, 51). Analysis of differential mRNA abundance for upf1Δ data was performed in DeSeq2 (52). Some figures were assembled and annotated in Inkscape v0.92 (https://inkscape.org).

Protein sequences were aligned using muscle (53), with default parameters for protein sequences and 100 iterations. Phylogenetic trees were constructed using ClustalW2 tool v2.1 (54) by using the neighbor-joining method with 1000 bootstrap trial replications.

Structural figures were prepared in the PyMOL Molecular Graphics System (Schrödinger).

External datasets

N. crassa (strain OR74A) ribosome profiling data from ((20), GEO:GSE97717) was used to generate highly-translated genes, and ribosome profiling and RNA-seq data from ((55), GEO: GSE71032) used to estimated TE. In both cases, we estimated TPMs using the RiboViz

pipeline as above, using the NC12 genome annotation downloaded from EnsemblGenomes (56). TL sequences were also obtained from NC12.

S. pombe (strain 972h) ribosome profiling and RNA-seq data are from (57), and the authors provided us with a table of RPKMs for all replicates as described. Genome sequence and annotation ASM294v2, including TL annotation, were downloaded from EnsemblGenomes (56).

C. albicans (strain SC5314) ribosome profiling and RNA-seq data are from (58), GEO:GSE52236), processed with the RiboViz pipeline as above using the assembly 22 of the strain SC5414 genome annotation from CGD (59).

S. cerevisiae (strain S288C/BY4741) highly-translated mRNAs use the RPKM table from ((60), GEO:GSE59573), and highly-abundant mRNAs use (61). For TE estimates, we used matched ribosome profiling and RNA-seq estimates from (62), although we did not use this for the list of highly translated genes because near-duplicate paralogous ribosomal protein genes were not present in the dataset, which thus omits a substantial fraction of highly-translated genes. TL sequences were downloaded from SGD (63).

Protein homolog lists were assembled with OrthoDB (64) and PANTHERdb (65), with reference to FungiDB (66). The list of cytoplasmic ribosomal proteins was assembled in *S. cerevisiae* based on (67) with help from SGD (63), extended to other fungi with PANTHERdb (47), and manually curated.

Availability and Accession Numbers

Raw and summarized sequencing data are available on GEO under accession numbers GSE133695 (RNA-seq, TSS-seq, PAS-seq, DNA-seq) and GSE133125 (ribosome profiling and

matched RNA-seq). Custom analysis code in R, and intermediate data files are available at https://github.com/ewallace/CryptoTranscriptome2018.

Data

Data are available at NAR online.

Acknowledgements

We thank members of the Wallace, Janbon, and Madhani labs for helpful discussions and comments on the manuscript. We thank Juan Mata for sharing intermediate data related to (Duncan and Mata 2017). We are grateful to J. Weissman (UCSF) for advice on ribosome profiling. We thank 3 anonymous reviewers for their helpful comments.

Funding

This work in the Madhani lab was supported by grants from the US National Institutes of Health [R01AI120464, R01GM71801 to H.D.M.]. H.D.M. is an Investigator of the Chan-Zuckerberg Biohub. E.W.J.W. is a Sir Henry Dale Fellow, supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 208779/Z/17/Z). L.T. is supported by a Wellcome-University of Edinburgh ISSF3 award. This work in the Janbon lab was supported by an Infect-ERA grant (project Cryptoview).

Figures

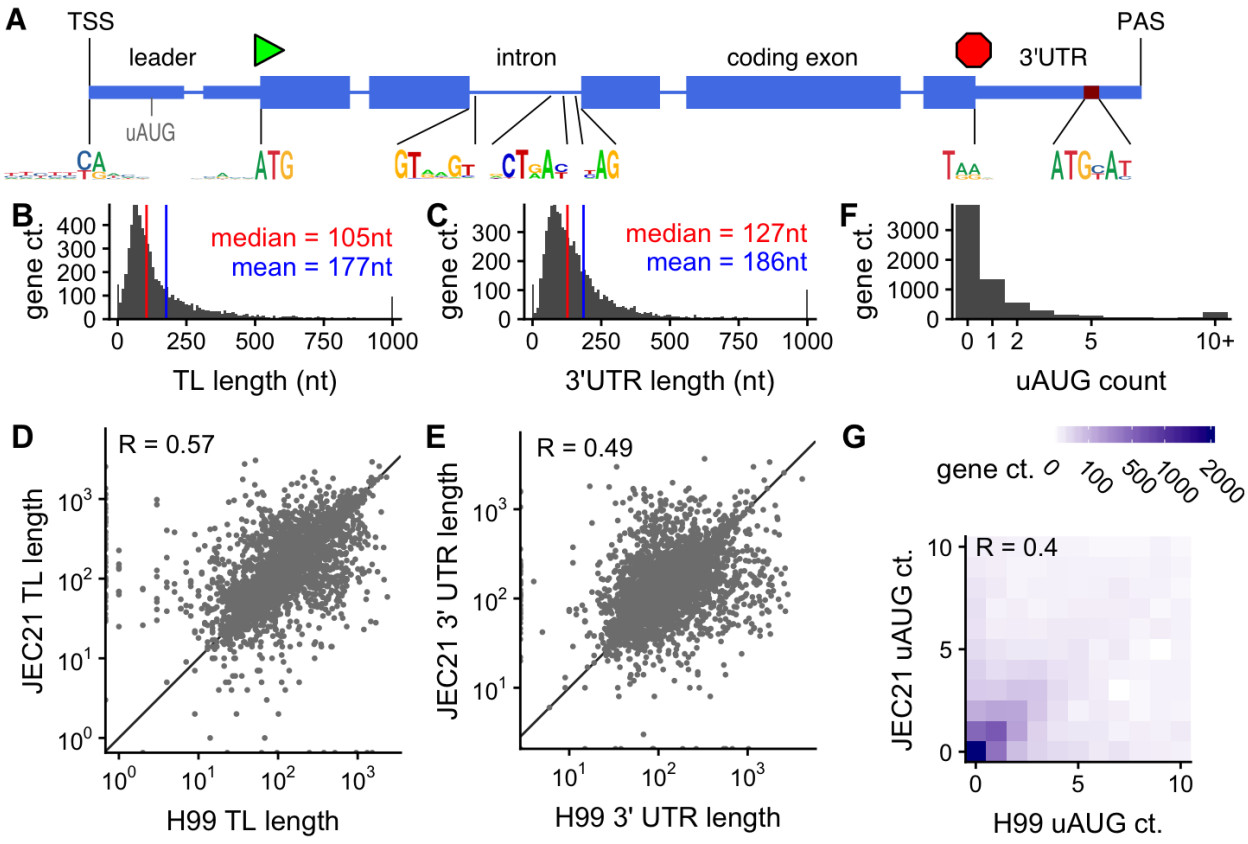


Figure 1.1: Mapping the coding transcriptome of *Cryptococcus neoformans*.

A, Representation of a stereotypical gene of *C. neoformans* H99, showing the sequence logos for the transcription start site (TSS), AUG start codon, intron splicing, stop codon, and polyadenylation site (PAS). B, Distribution of transcript leader (TL) lengths over *C. neoformans* genes, for yeast cells growing exponentially in YPD at 30°C. C, Distribution of 3' untranslated region (3'UTR) lengths over *C. neoformans* genes. D,E : Comparisons of TL and 3'UTR lengths between orthologous genes in *C. neoformans* H99 and *C. deneoformans* JEC21 growing exponentially in YPD at 30°C. F, Distribution of upstream AUG (uAUG) counts over *C. neoformans* genes, and G, comparison of uAUG counts with *C. deneoformans*.

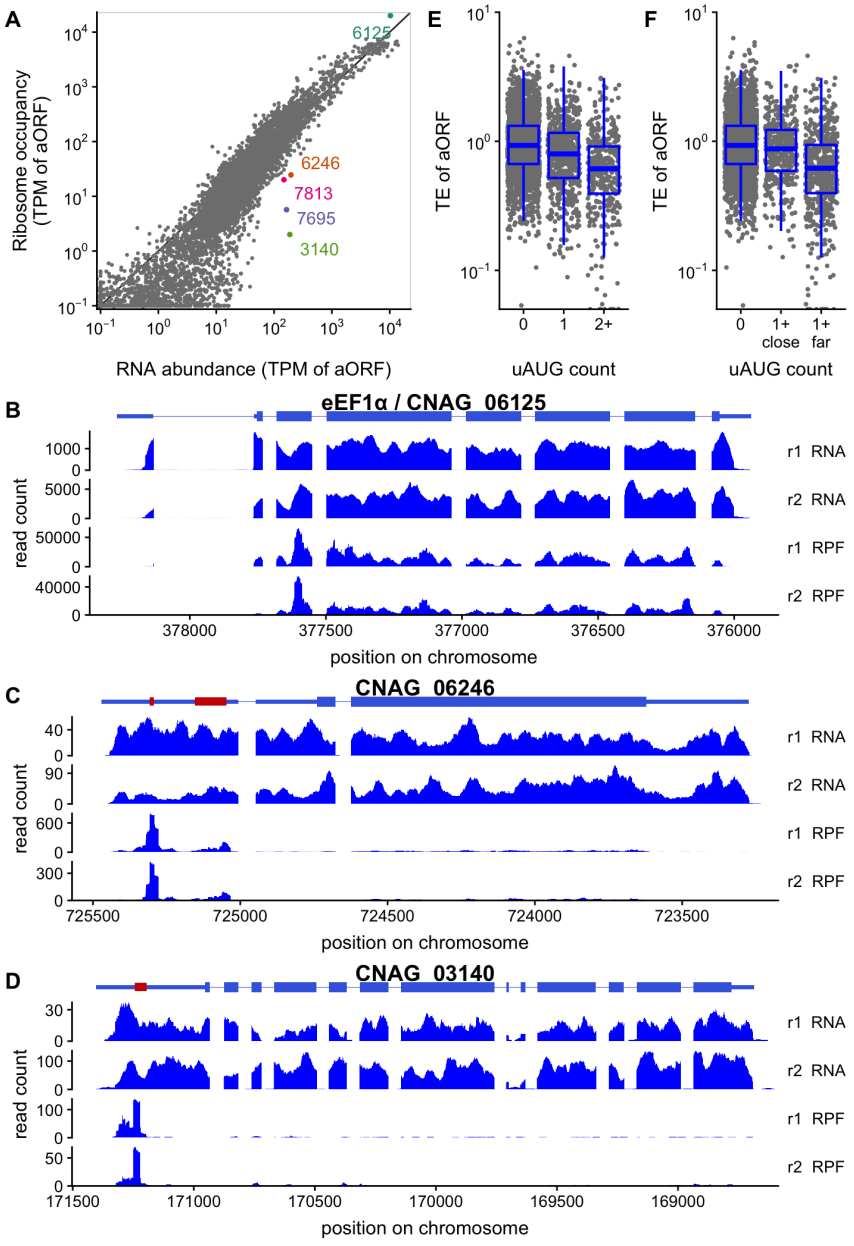


Figure 1.2: Upstream AUGs repress translation in *C. neoformans*.

A, translation regulation of annotated ORFs (aORFs) in *C. neoformans* H99 growing exponentially in YPD at 30°C (equivalent data for *C. deneoformans* shown in Figure 1.11). Ribosome occupancy is plotted against the RNA abundance, both calculated in transcripts per million (TPM) on the aORF. Select genes discussed in the text are highlighted in color. B, Translation elongation factor eEF1α/CNAG_06125 has high ribosome occupancy in the annotated ORF. Translationally repressed mRNAs CNAG_06246 (C) and CNAG_03140 (D) have high ribosome occupancy in uORFs in the transcript leader (red), and low ribosome occupancy in the aORF. Only the first of 5 uORFs in CNAG_03140 is shown. Other genes highlighted in panel A are shown in Figure 1.12. Homologous genes in *C. deneoformans* have

similar structure and regulation (Figure 1.11, 1.12). E, uAUGs are associated with lower translation efficiency (TE) of annotated ORFs, measured as the ratio of ribosome occupancy to RNA-seq reads. F, only uAUGs far from the transcription start site are associated with low TE. A gene is in the “1+ far” category if it has at least one uAUG more than 20nt from the TSS, “1+ close” if all uAUGs are within 20nt of the TSS.

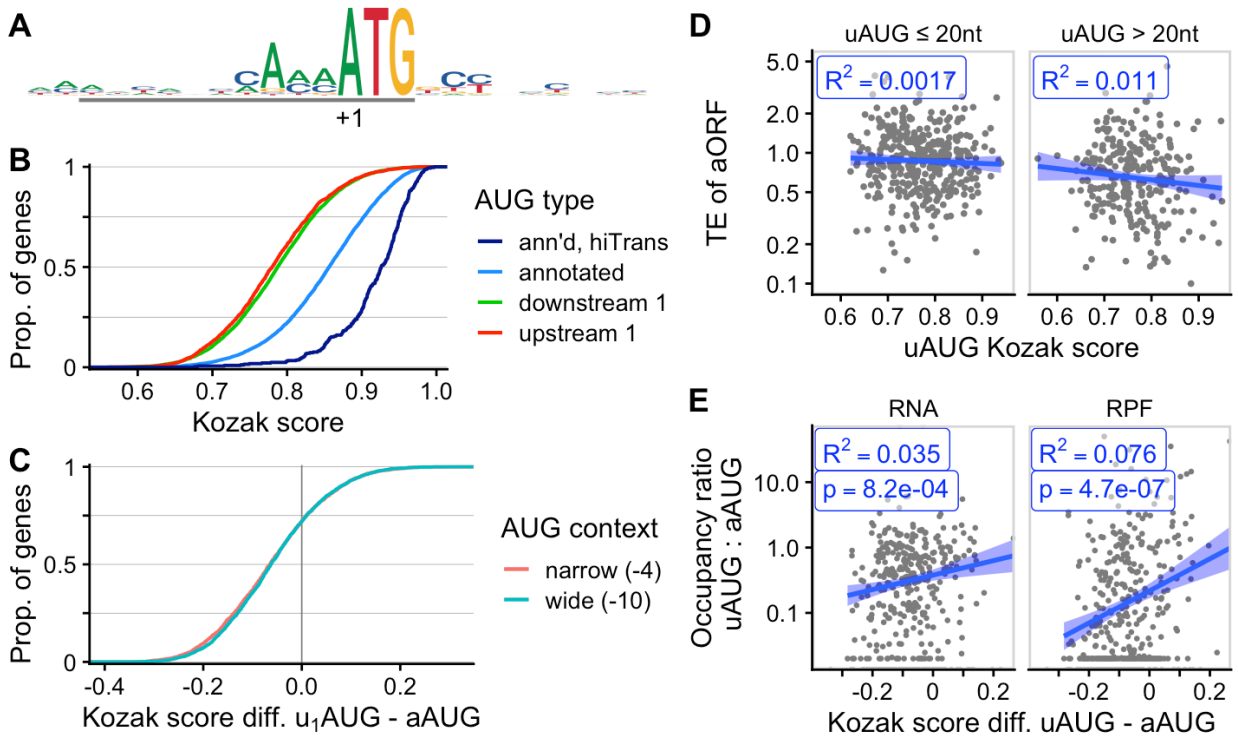


Figure 1.3: An AUG sequence context is associated with translation in C. neoformans.

A, Kozak-like sequence context of AUGs, from -12 to +12, for highest-translated 5% of genes (hiTrans). This sequence context is used to create “Kozak scores” of other AUG sequences by their similarity to the consensus from -10 onwards. B, Cumulative density plot of Kozak scores from various categories of AUG, showing that high scores are associated with annotated AUGs of highly translated genes (hiTrans), somewhat with annotated AUGs, and not with the most 5’ downstream AUG (downstream 1) or 5’ most upstream AUG (upstream 1) in a transcript. C, Cumulative density plot of differences in scores between most 5’ upstream (u1AUG) and annotated AUG, showing that for 75% of genes the upstream AUG score is less than the annotated AUG, whether we take a wide (-10:AUG) or a narrow (-4:AUG) window to calculate the score. D, High upstream AUG score is weakly and not significantly associated with translation repression of the annotated ORF. E, The relative occupancy of ribosomes (RPF) at the upstream AUG and annotated AUG depends on the difference in scores, even when compared to RNA-seq reads; linear model trend fit shown (blue) with R2, and p-value of associated t-test. Panels D and E show data only for genes in the top 50% by RNA abundance, and with only a single upstream AUG. Figure 1.13 shows homologous data for C. deneoformans.

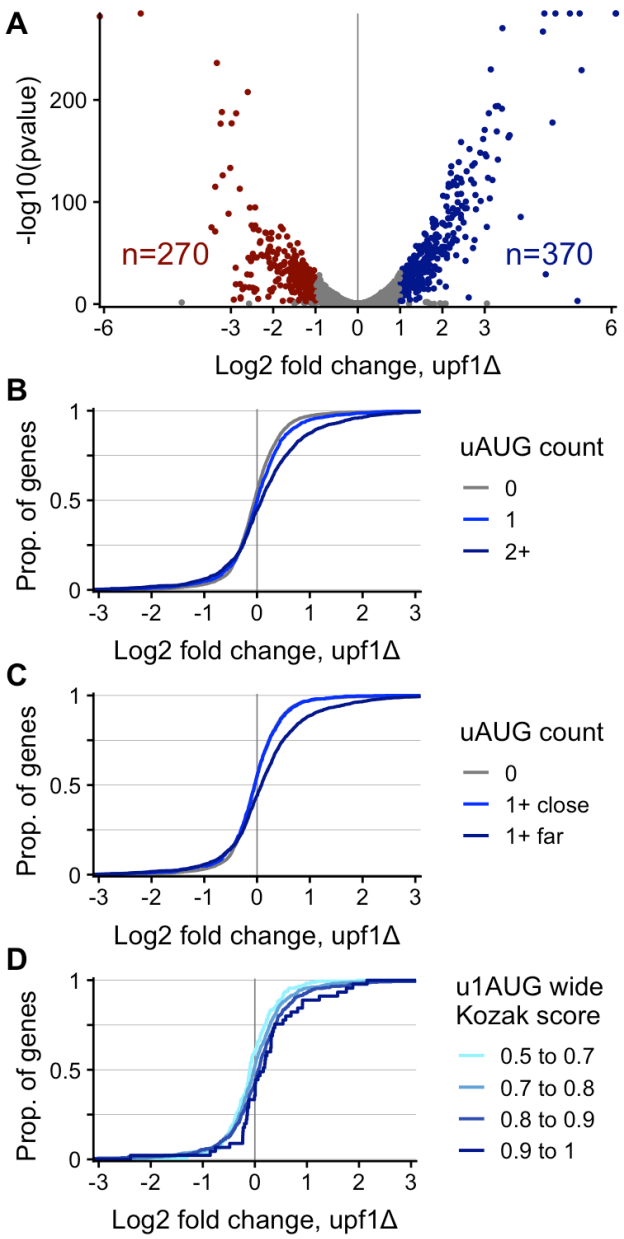


Figure 1.4: Nonsense-mediated decay (NMD) acts on upstream-AUG-containing mRNAs in C. deneoformans.

A, Differential expression results from RNA-Seq in C. deneoformans JEC21, comparing expression in wild-type cells with a mutant deleted for NMD factor UPF1/CNC02960, and using DeSeq2 to identify genes upregulated in the upf1Δ mutant. B, uAUG containing genes are enriched for NMD-sensitivity. A one-sided Kolmogorov-Smirnov test shows that these

differences are significant comparing 1 uAUG to 0 (p=5.7 x 10-5) and 2 or more AUGs to 1 (p=7.3 x 10-11). C, uAUG-containing genes are enriched for NMD-sensitivity only when the uAUG is more than 20nts from the TSS (1+far; $p < 2.2 \times 10^{-16}$), but not when the uAUG is less than 20nts (1+close; p=0.73). D, Start codon sequence context affects NMD sensitivity of genes containing a single upstream AUG: RNAs starting with higher Kozak-score uAUG are more likely to increase in abundance in the upf1Δ mutant ($p = 1.1 \times 10^{-4}$, comparing score < 0.8 with score > 0.8).

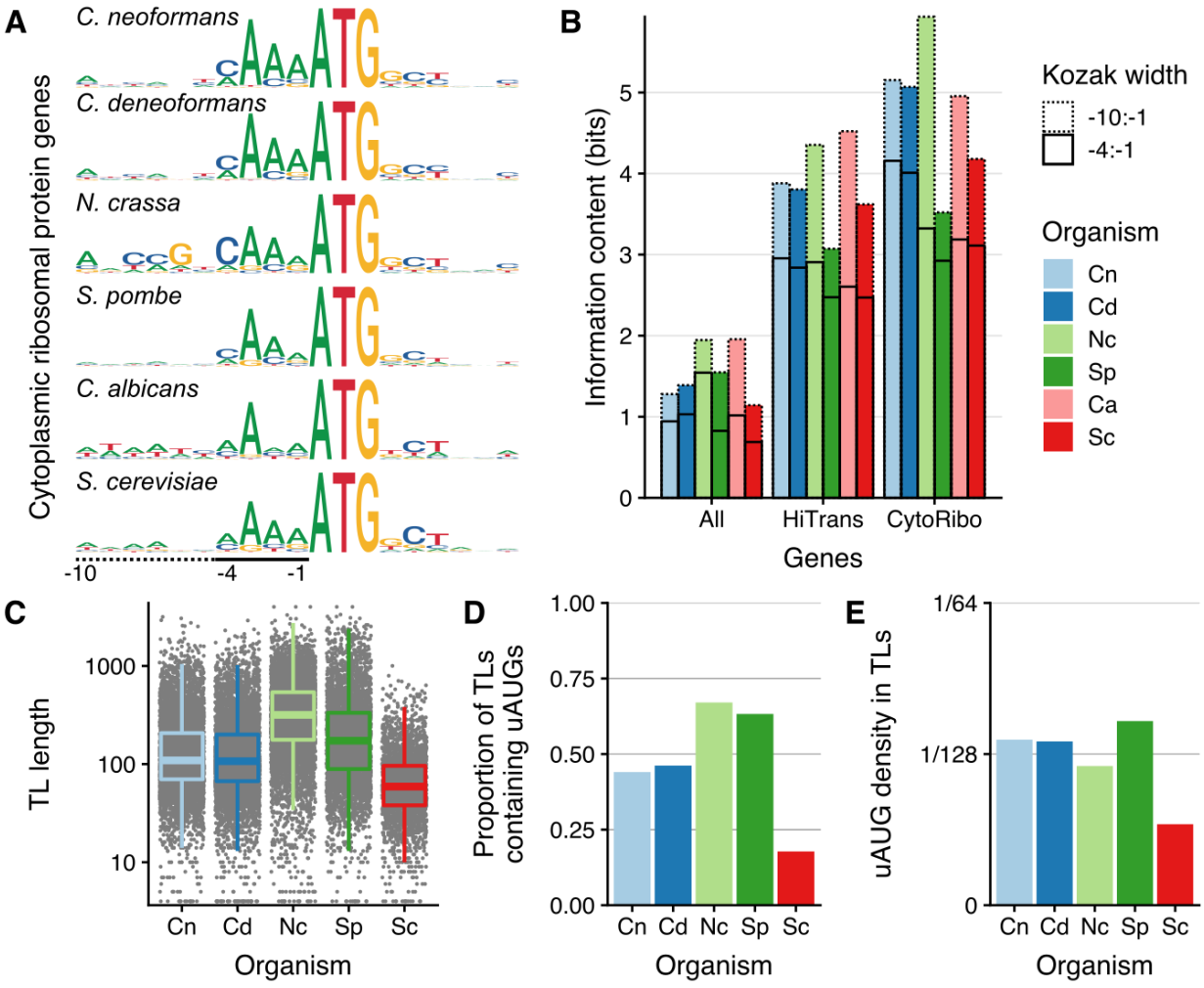
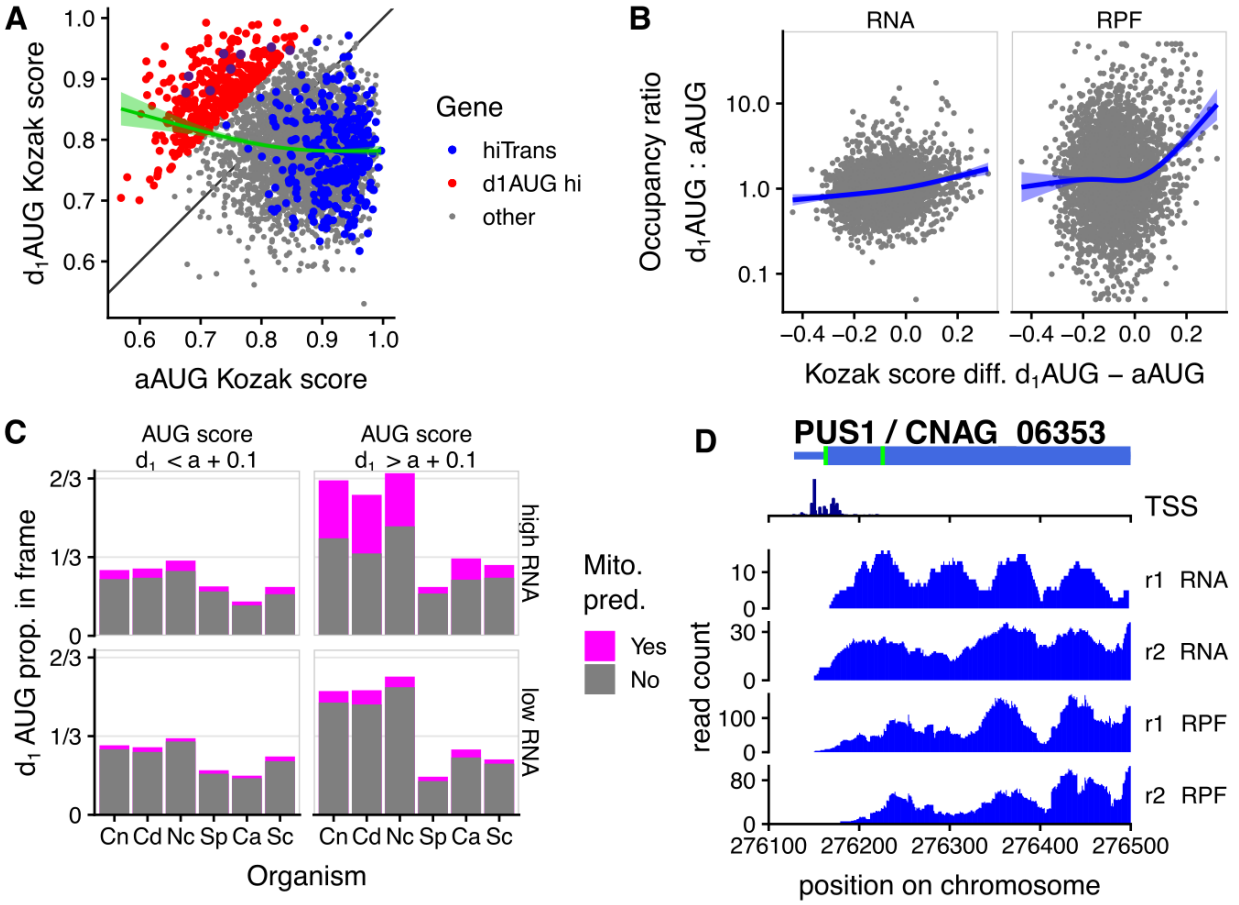


Figure 1.5: Sequences specifying start codon selection are quantitatively different in different fungi.

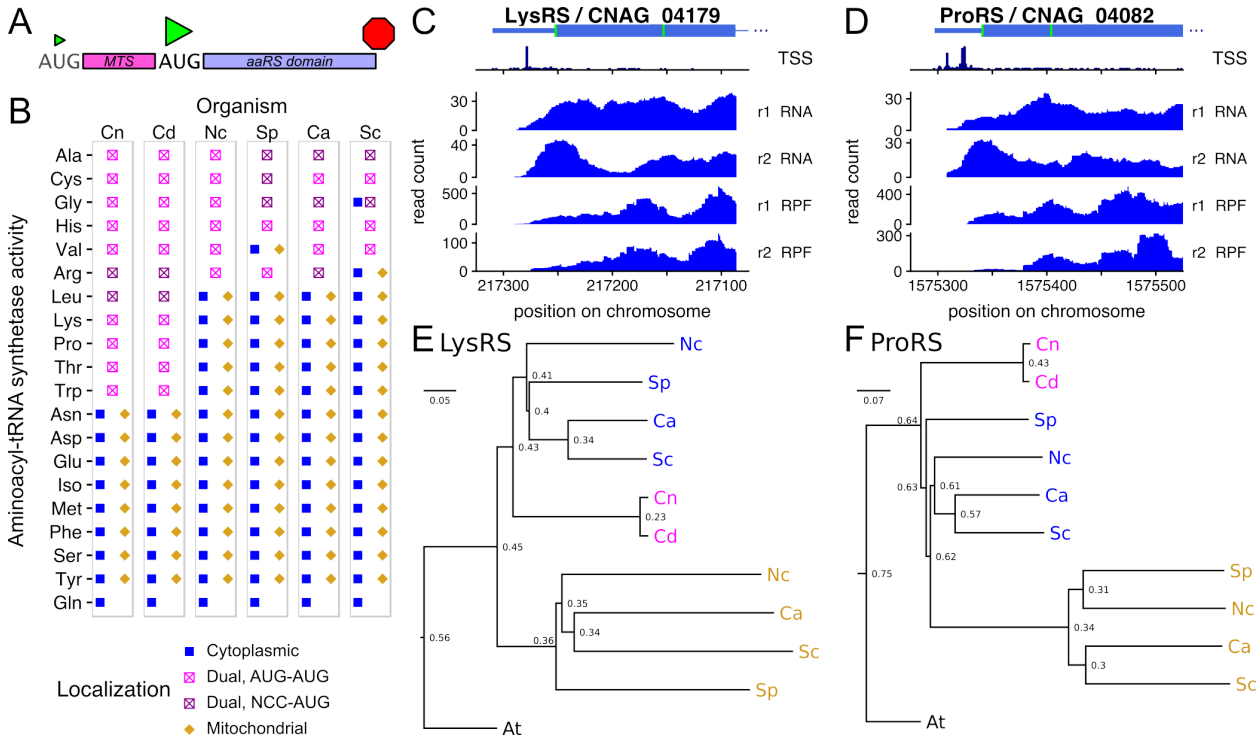
A, Kozak consensus sequence logo for annotated start codons of cytoplasmic ribosomal protein genes from 6 fungal species. The height of each letter represents the Shannon information content in bits, so that the anchor ATG sequence has height 2 bits. B, Information content at annotated start codons in bits per base (i.e. summed height of stacked letters in sequence logo) for 3 groups of genes, in the 6 fungi from panel A. Solid line indicates information from -1 to -4 of ATG, and dotted line additionally to -10 (see bottom of panel A). Gene groups are all annotated ORFs, highly translated ORFs (HiTrans) and cytoplasmic ribosomal proteins (CytoRibo, as panel A). HiTrans used the highest-translated 5% of genes, or the highest 400 genes for fungi with more than 8000 annotated genes (C. albicans and N. crassa; see methods). C-E, For 5 fungi for which transcript leader (TL) annotations were available, TL length (C), proportion of



annotated TL containing an upstream AUG (D), and proportions of AUGs per nucleotide in the TL (E; a uniform random model would have density 1/64).

Figure 1.6: High-scoring downstream AUGs specify alternative N-terminal isoforms in C. neoformans.

A, Most genes with high RNA abundance (top 50% by RNA abundance shown), especially very highly-translated genes (blue, top 5%), have lower Kozak score at the 1st downstream AUG than at the annotated AUG. However there are exceptions (red, d1AUG hi: d1AUG score > annotated AUG score + 0.1), and there is a trend for genes with low aAUG score to have a higher d1AUG score (green, generalized additive model fit). B, Higher d1AUG score than aAUG score drives higher ribosome protected fragment (RPF) occupancy at the d1AUG compared to the aAUG, but much smaller differences in RNA-seq density. Blue line indicates generalized additive model fit. C, Downstream AUGs with high Kozak scores (d1AUG score > annotated AUG score + 0.1) and high RNA abundance (top 50%) are likely to be in-frame and enriched for N-terminal mitochondrial localization signals in C. neoformans, C. deneoformans, and N. crassa, but not in S. pombe, C. albicans, or S. cerevisiae. D, The pseudouridine synthase CnPus1 is a candidate alternate-localized protein with a low-score aAUG and high-score



d1AUG, and transcription start sites on both sides of the aAUG. RNA-Seq and RPF reads on the first exon are shown, and the full length of the gene shown in Figure 1.17.

Figure 1.7: Aminoacyl-tRNA synthetases (aaRSs) are commonly alternatively localized to cytoplasm and mitochondria by use of alternative start codons in fungi.

A, Schematic of the structure of a dual-localized aaRS with alternate AUG start codons. B, Predicted localization of all aaRS enzymes in the fungi C. neoformans (Cn), C. deneoformans (Cd), N. crassa (Nc), S. pombe (Sp), C. albicans (Ca), S. cerevisiae (Sc). C/D, Transcription start site reads, RNA-seq, and ribosome profiles of 5'-ends of CnLysRS (C) and CnProRS (D) show that most transcription starts upstream of both AUG start codons (green), and both AUG codons are used for translation initiation. E/F Simplified neighbour-joining phylogenetic trees show that LysRS (E) and ProRS (F) genes were duplicated in ascomycete fungi, and Cryptococcus retained a single dual-localized homolog. Arabidopsis thaliana (At) was used as an out-group. The scale bar represents the number of amino acid substitutions per residue, and the numbers at nodes are the proportion of substitutions between that node and its parent. See table 1.14, for details of identifiers for genes (GeneID).

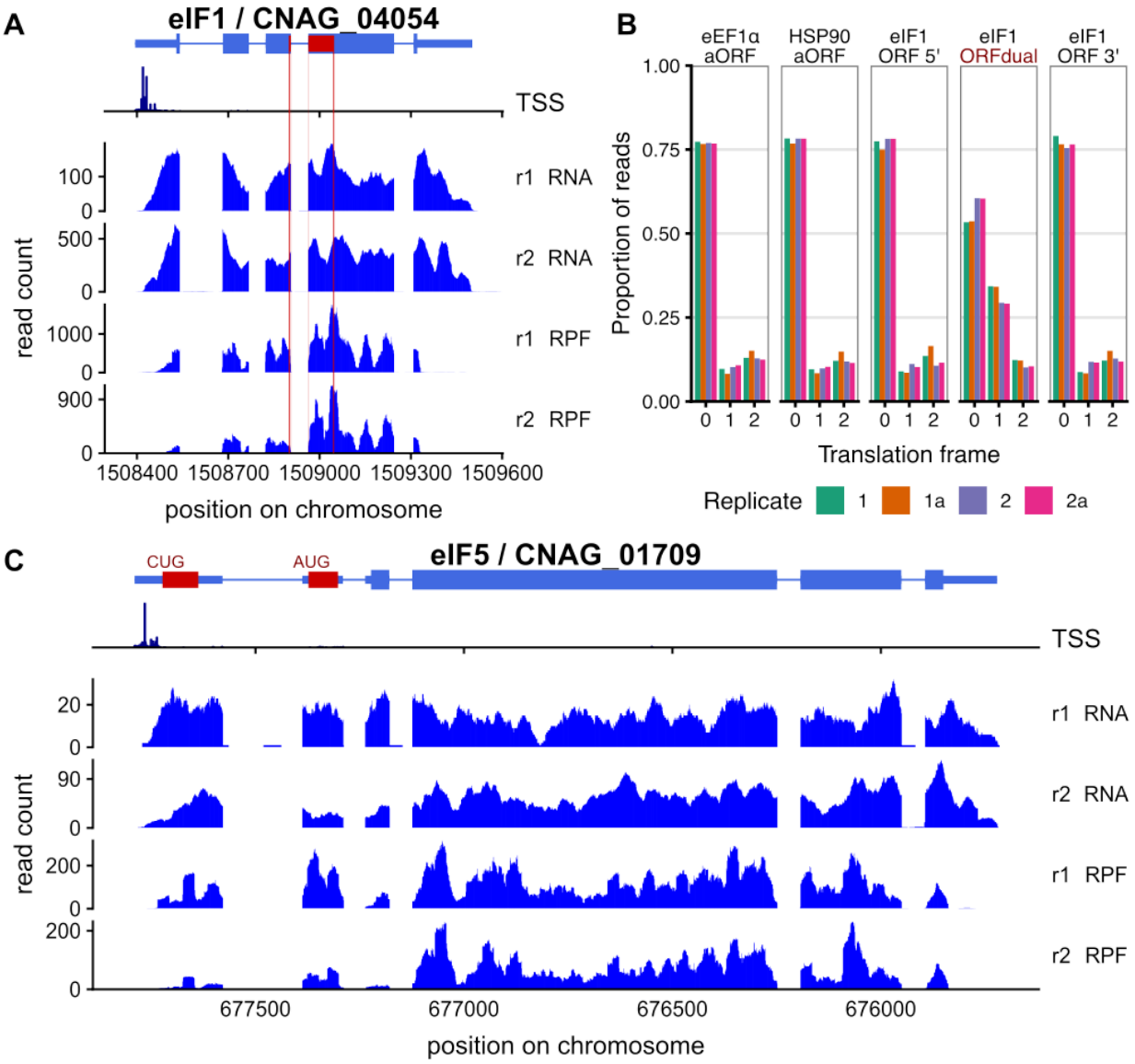


Figure 1.8. Translation initiation factors eIF1 and eIF5 are regulated by alternate start codon usage in C. neoformans.

A, Reads on CneIF1/CNAG_04054, showing frame +1 “downstream ORF” in dark red, breaking for an intron. B, The downstream ORF of CneIF1 is dual-translated in two frames. Most ribosome profiling read 5’ ends are in a consistent frame, including in control genes eEF1α/ CNAG_06125 and HSP90/CNAG_06125, and in the 5’ and 3’ ends of the CneIF1 ORF, but there is 2x enrichment of reads in frame+1 in the dual-decoded ORF. C, Reads on CneIF5/

CNAG_01709 showing substantial ribosomal occupancy over upstream ORFs. The first upstream ORF shown is translated from a CUG start codon and the second from an AUG codon, and other uORFs potentially initiated from near-cognate codons are not shown. C. deneoformans homologs have the same structure and regulation (Figure 1.19).

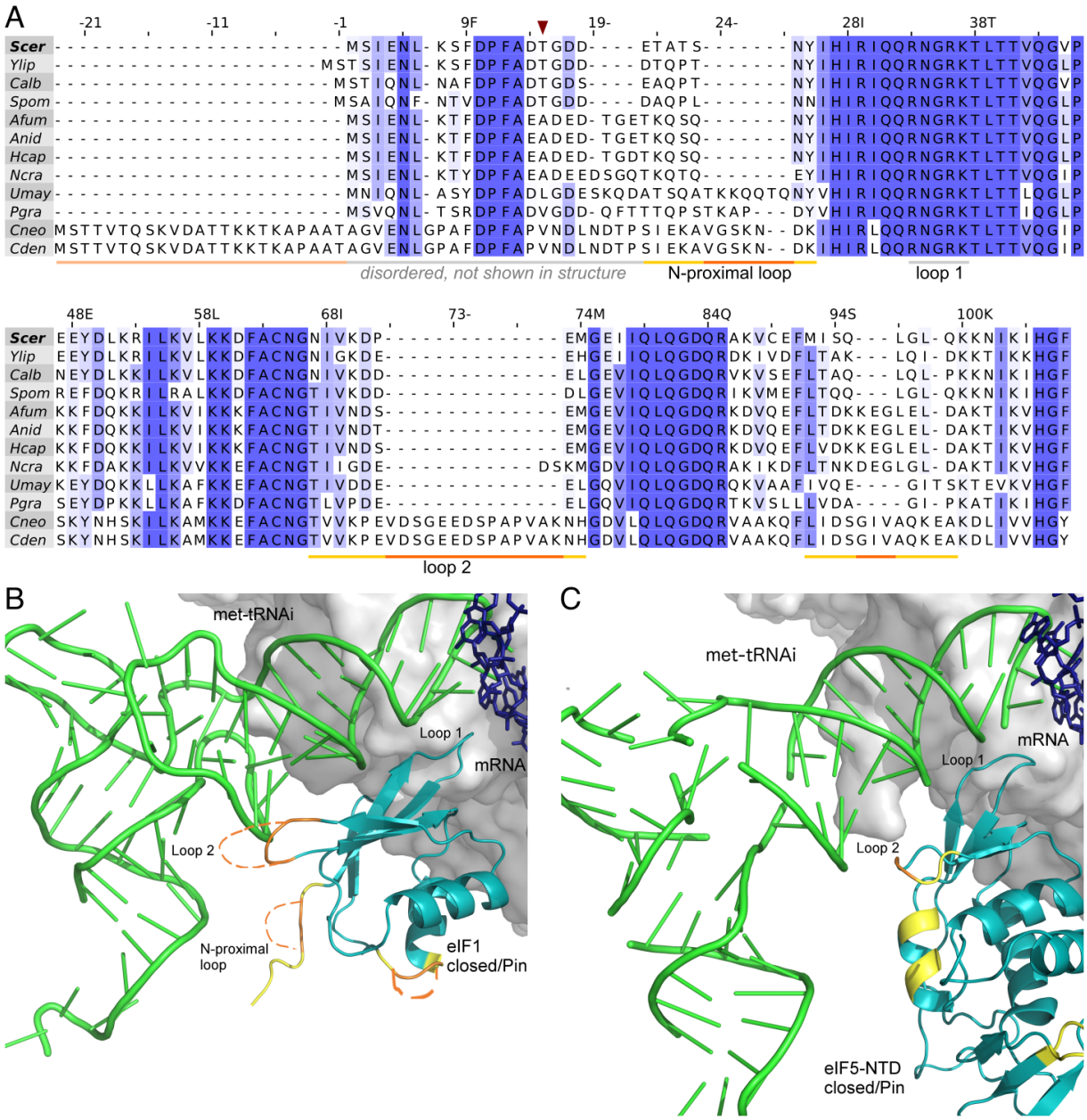


Figure 1.9. Eukaryotic translation initiation factor 1 is highly variable across fungi.

A, Multiple sequence alignment of translation initiation factor eIF1 from 12 fungi, numbered as S. cerevisiae (Scer, top line). Cryptococcus insertions are indicated in orange, and surrounding variable residues in yellow. The N-terminal extension in Cryptococcus eIF1, that is predicted disorded, is shown in pale orange, and T15 residue with dark red arrow. B, Structural predictions of insertions (orange) and non-conserved neighborhoods (yellow) in Cryptococcus

eIF1 mapped onto the closed pre-initiation complex of S. cerevisiae/K.lactis (PDB:3J81, Hussein 2015). eIF1 (teal) and Met-tRNAi (green) in closed conformation, shown with synthetic mRNA sequence (pink), and eIF2 (pale pink) and ribosomal subunit surface (greys) in background. Approximate ribosomal contacts are shown as grey background surface and eIF2-alpha subunit is shown as pale pink sticks. C, Structural predictions of variations in Cryptococcus eIF5 mapped on to S. cerevisiae PIC (PDB:6FYX, (27)). Multiple sequence alignment of eIF5 is shown in figure 1.20A.

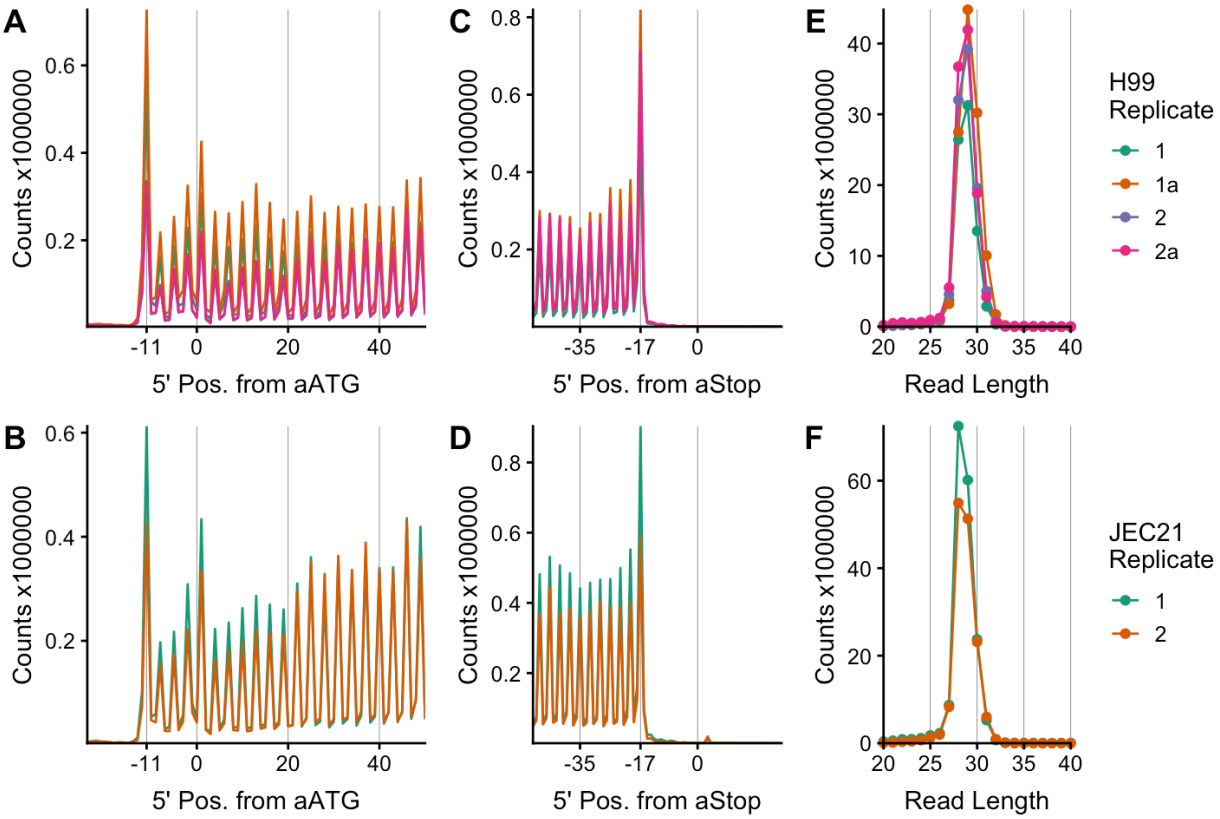


Figure 1.10: Ribosome profiling data passes quality control metrics.

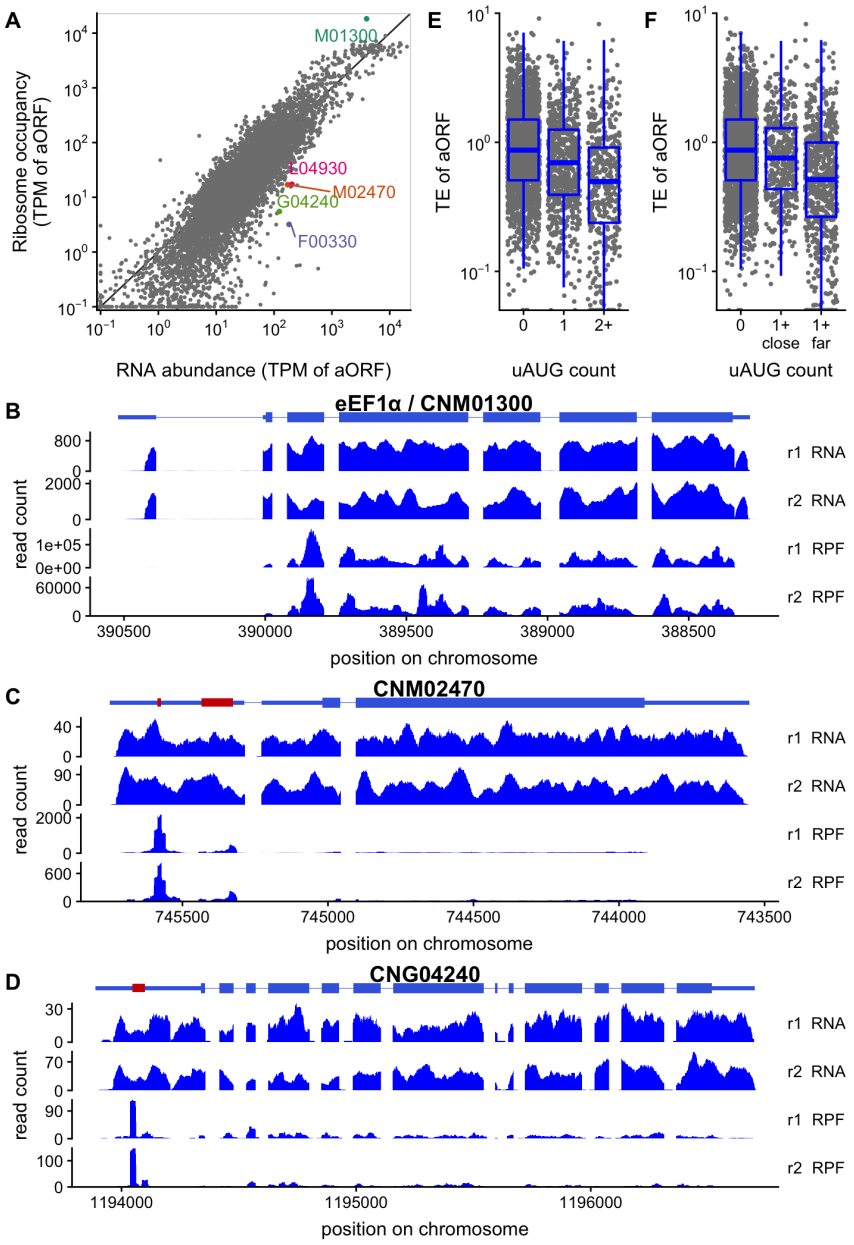


Figure 1.11: Upstream AUGs repress translation in C. deneoformans.

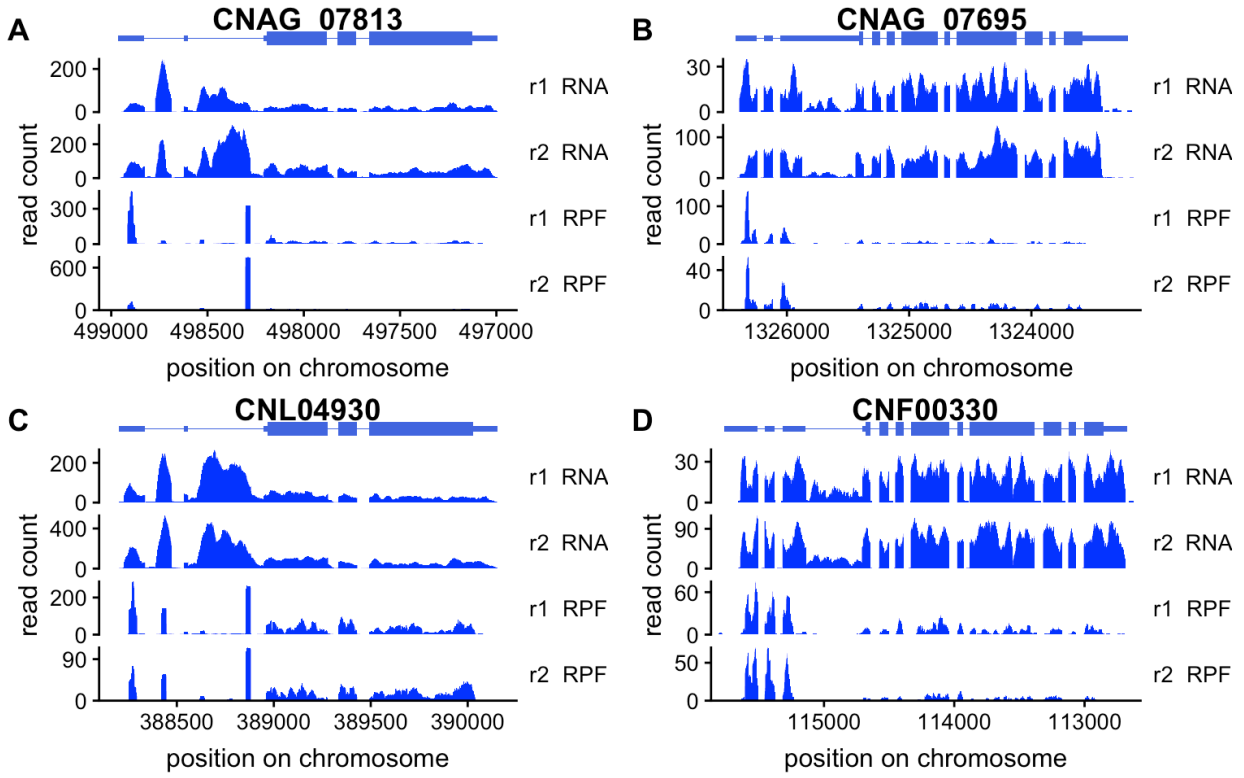


Figure 1.12: Further examples of upstream AUG and 5'-end regulation in C. neoformans and C. deneoformans.

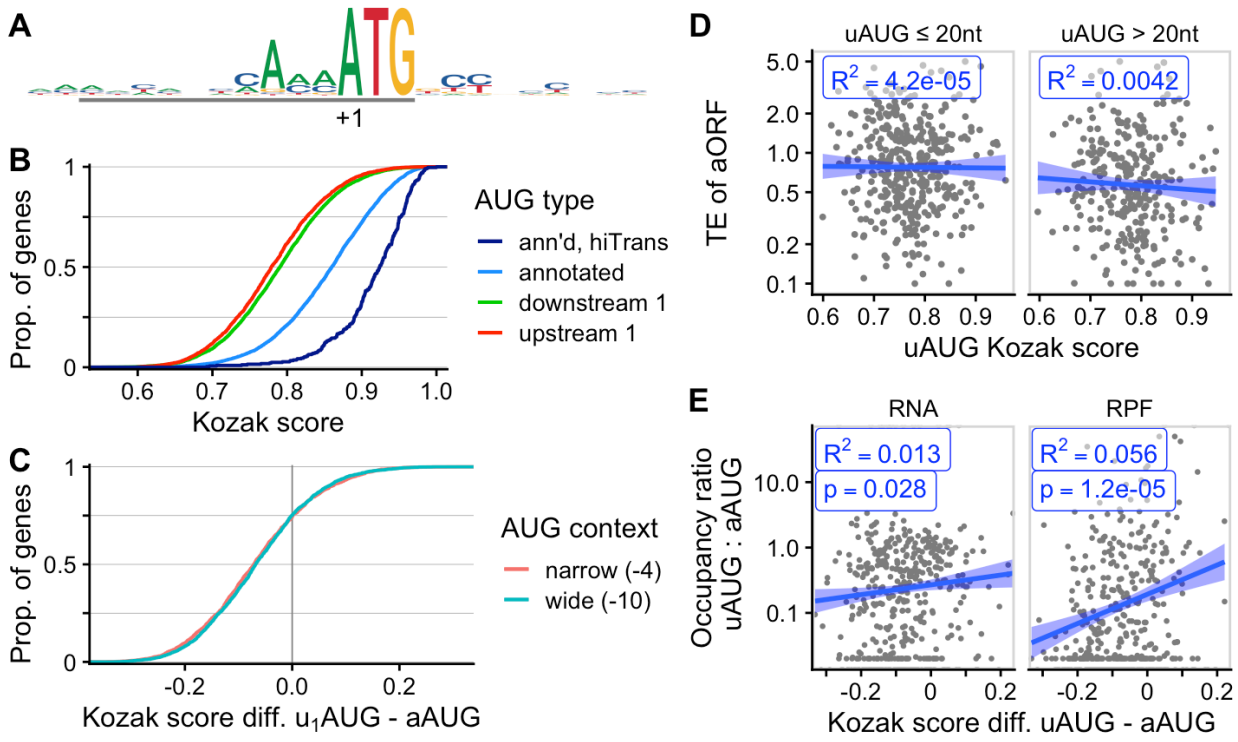


Figure 1.13: AUG sequence context is associated with translation in C. deneoformans.

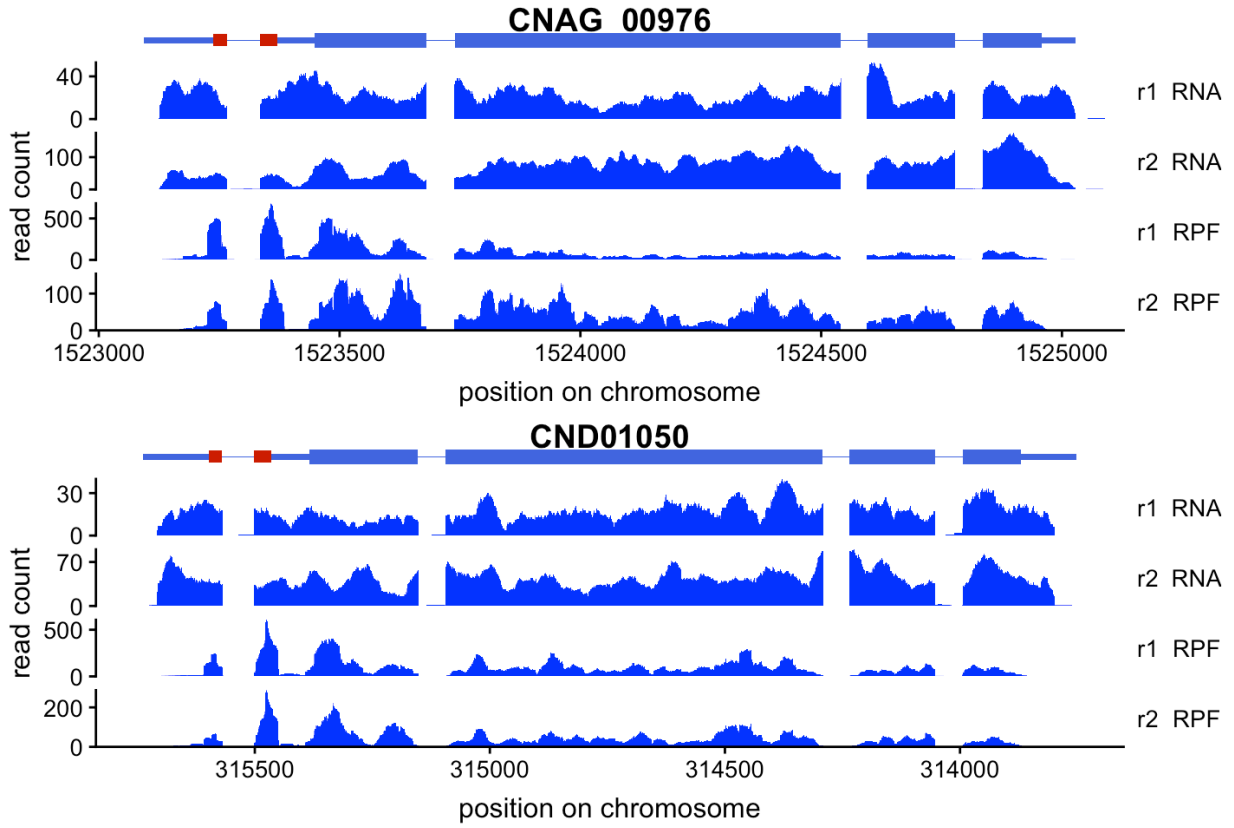


Figure 1.14: Carbamoyl-phosphate synthase CPA1 homologs have a conserved uORF that is occupied by ribosomes in C. neoformans (CNAG_00976, top) and C. deneoformans (CND01050, bottom).

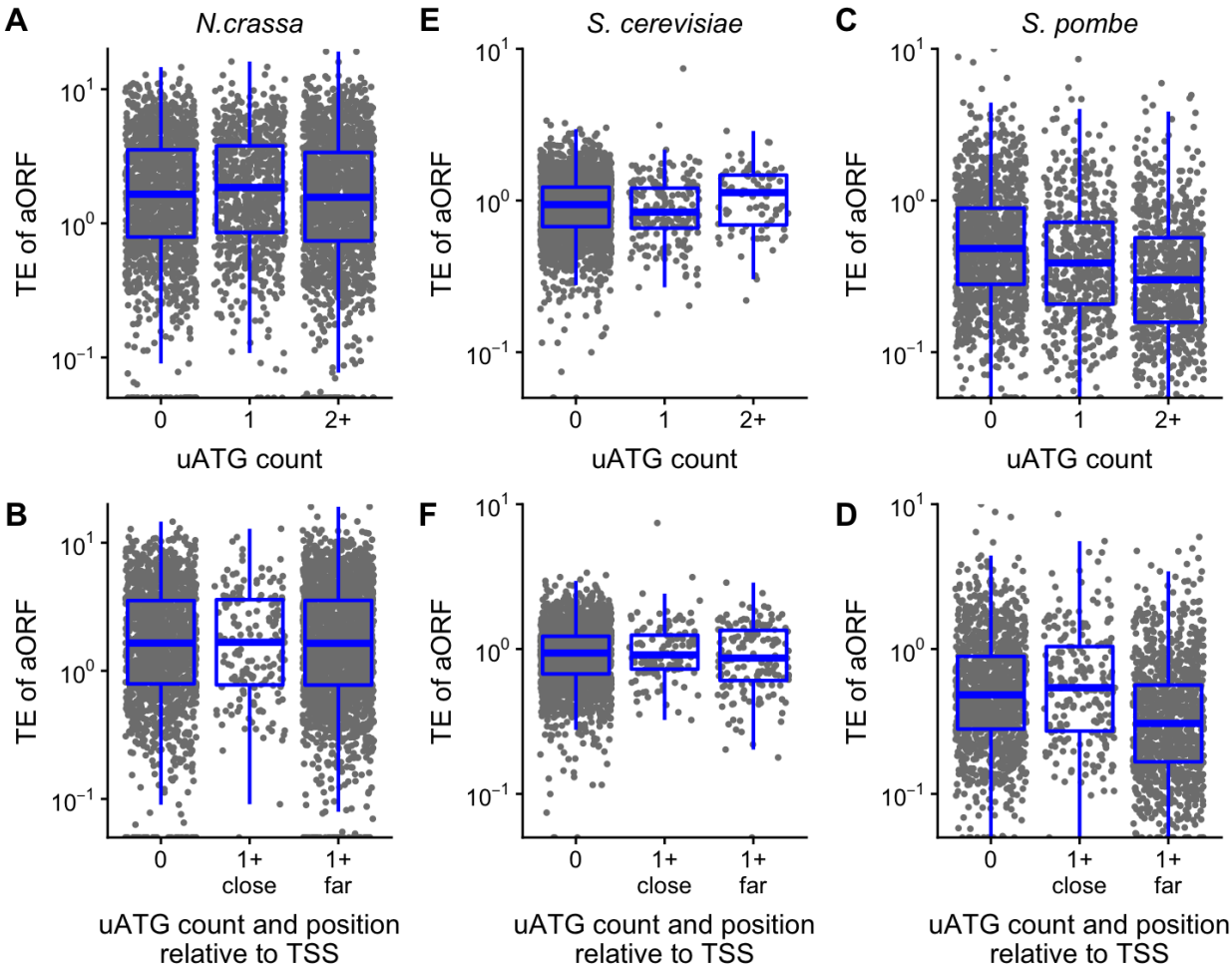


Figure 1.15: Effect of uATGs on translational efficiency in N. crassa, S. pombe, and S. cerevisiae.

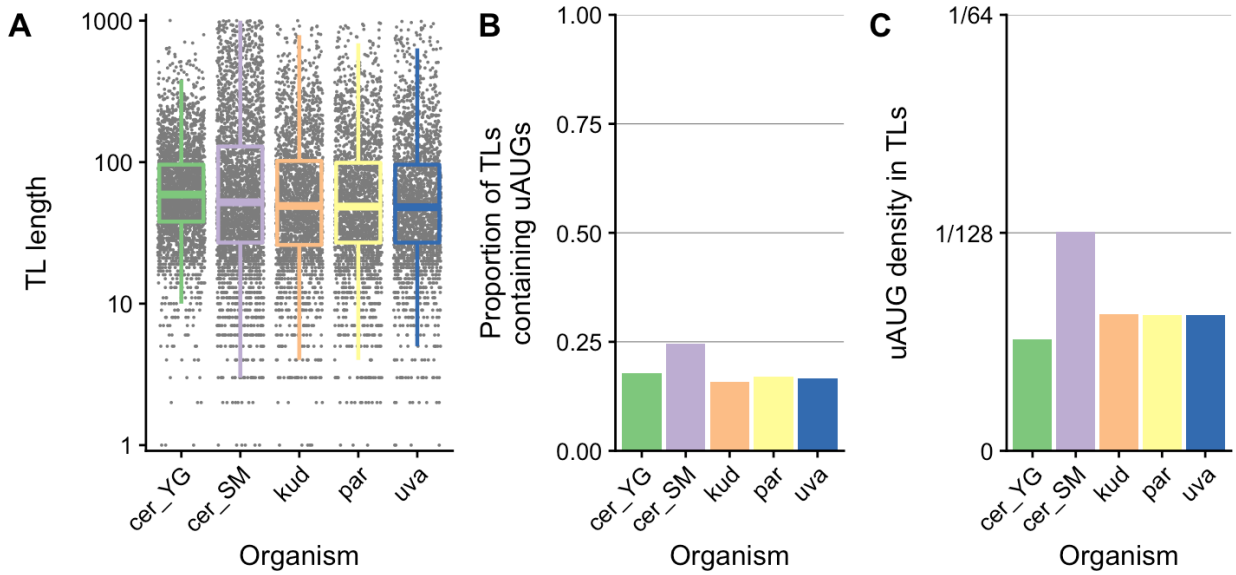


Figure 1.16: Saccharomyces sensu stricto species have short AUG-poor transcript leaders.

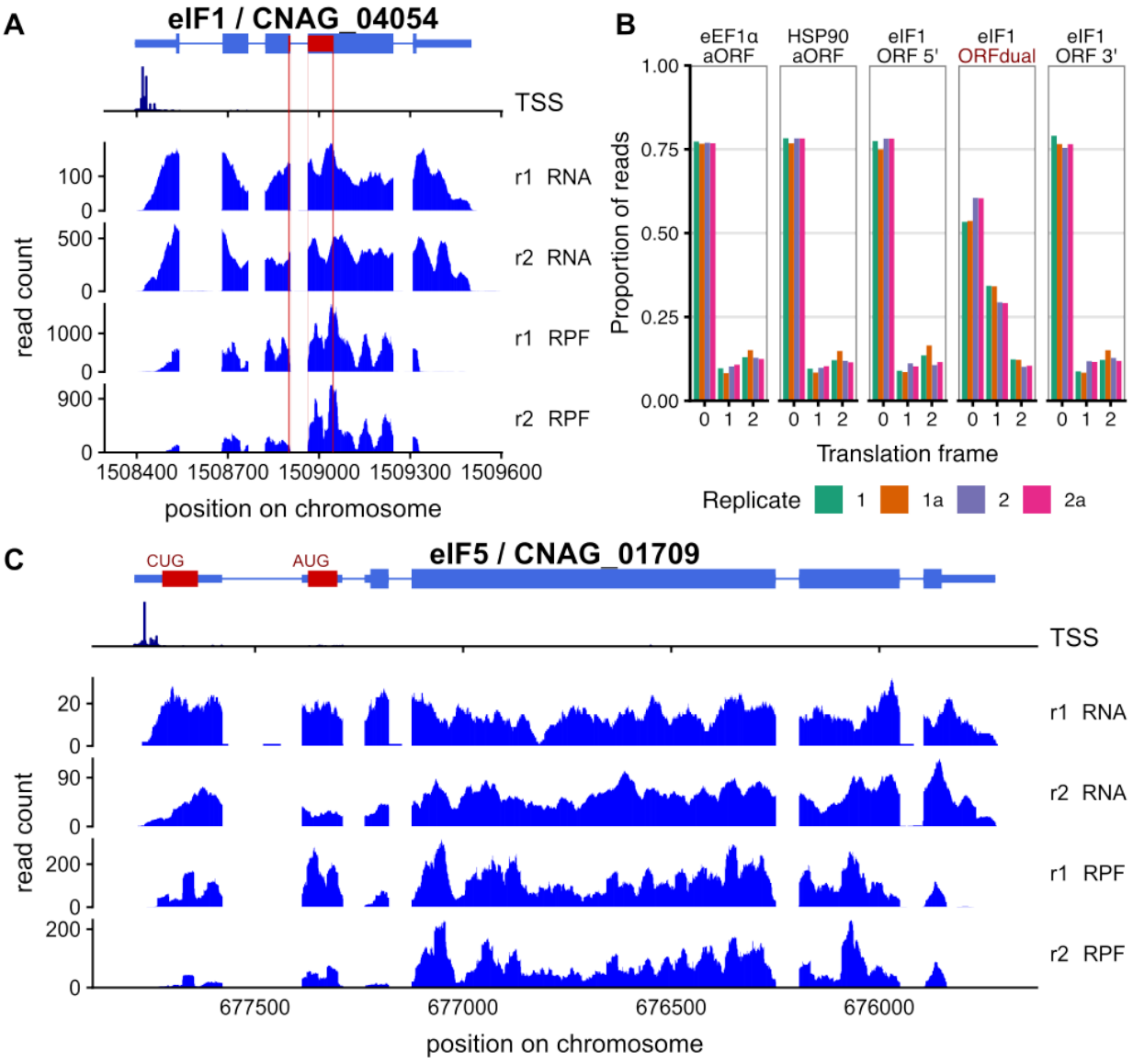


Figure 1.17: Ribosome profiles of some C. neoformans genes with predicted dual-localization specified by alternative N-termini.

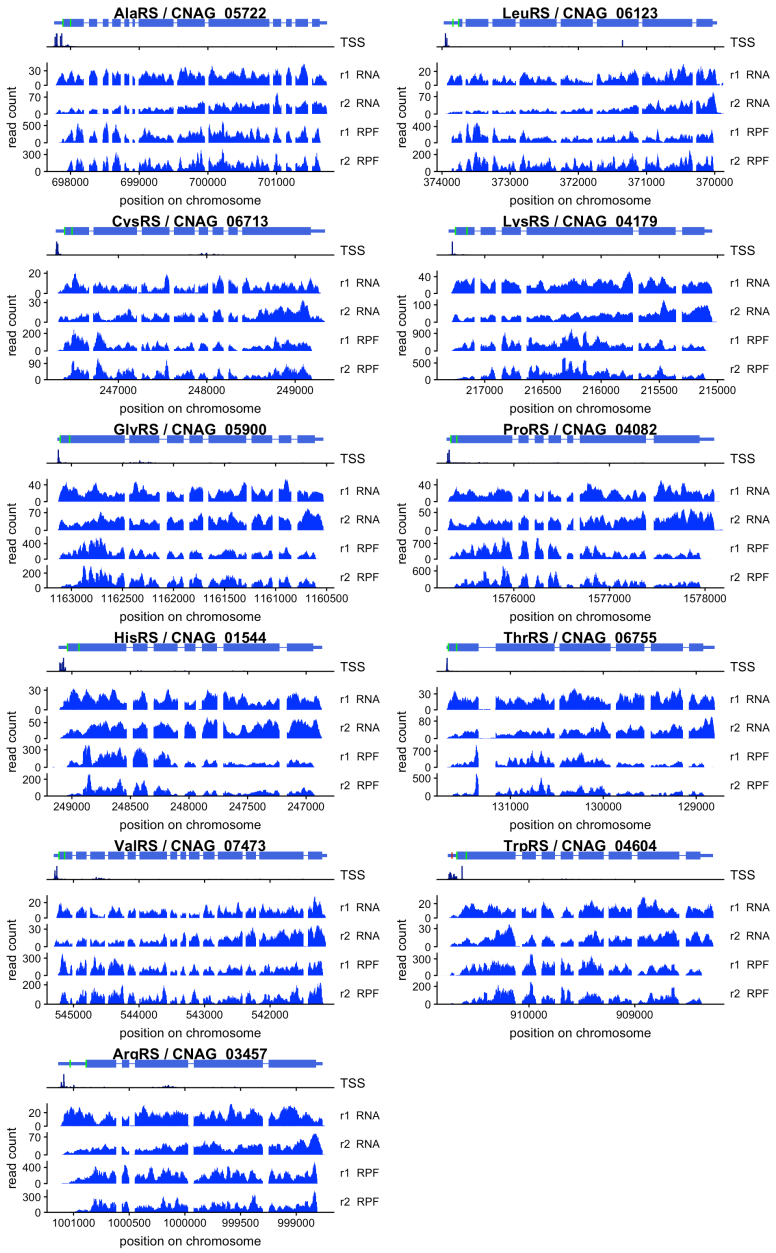


Figure 1.18: Ribosome profiles along the 11 C. neoformans aaRS genes with predicted dual-localization.

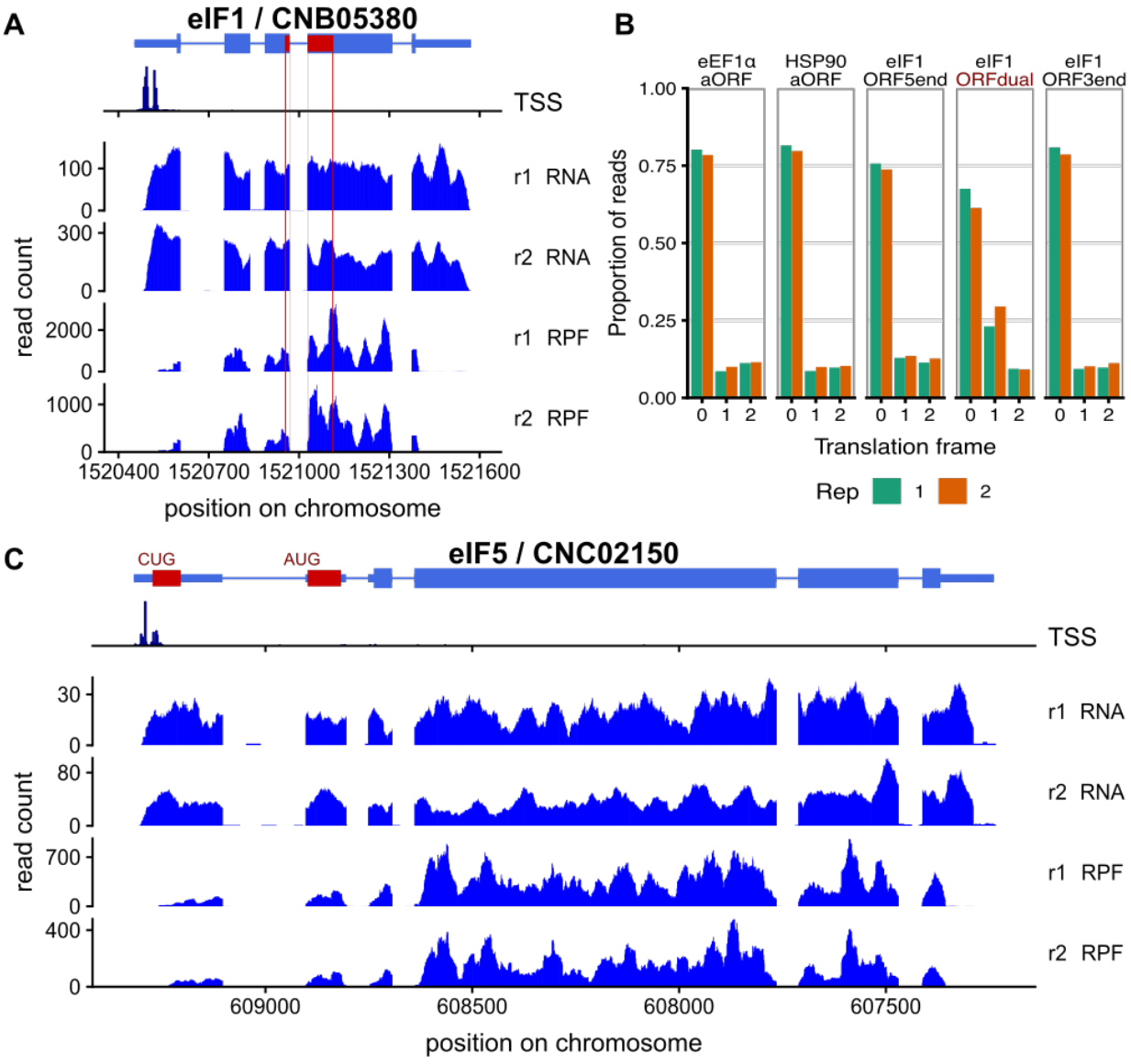


Figure 1.19: Translation initiation factors eIF1 and eIF5 are regulated by alternate start codon usage in C. deneoformans.

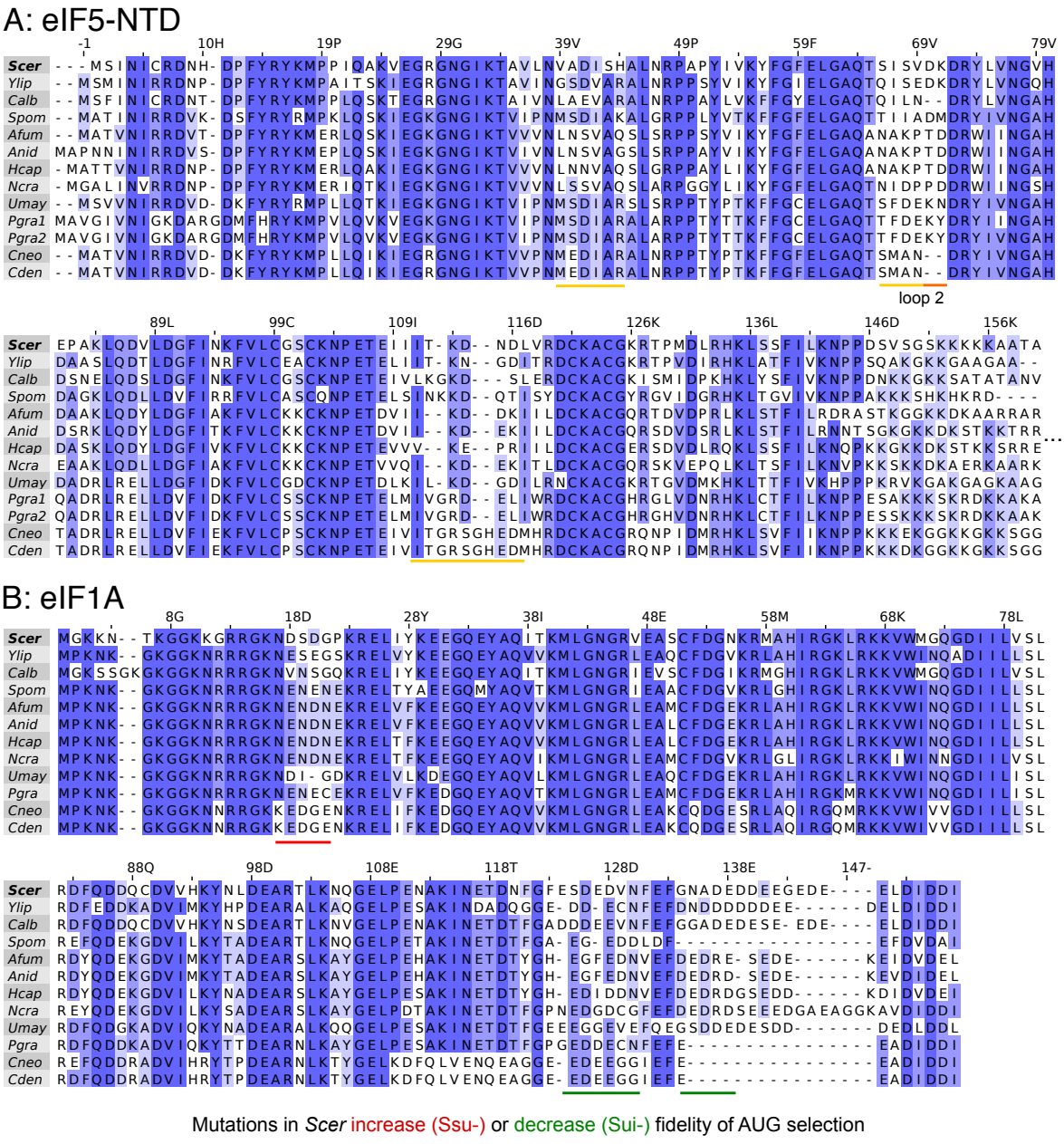


Figure 1.20: multiple sequence alignments of eIF5-NTD and eIF1A from select fungi.

Tables

Table 1.1: Sequencing and annotation numbers.

DOI https://doi.org/10.7272/Q6T72FP1

Table 1.2: Differential expression results in Cryptococcus deneoformans upf1Δ.

DOI https://doi.org/10.7272/Q6T72FP1

Table 1.3: Cytoplasmic ribosomal proteins in 6 fungal species.

DOI https://doi.org/10.7272/Q6T72FP1

Table 1.4: Genes with score d1AUG – aAUG > 0.1, n = 167.

DOI https://doi.org/10.7272/Q6T72FP1

Table 1.5: List of aaRS in Cryptococcus and select fungi.

DOI https://doi.org/10.7272/Q6T72FP1

Table 1.6: Initiation contexts of annotated and downstream AUGs in 9 Cryptococcus aaRSs

H99	JEC21	aATG.context.H99	aATG.context.JEC21	d1.context.H99	d1.context.JEC21
CNAG_04179	CNI03160	CAGATTGTGCATG	CAGATTGTGCATG	AACGCTCACAATG	AACGTTCATAATG
CNAG_06713	CNB00900	CAGCATCTGCATG	CAGCATCTGCATG	CTCATACACAATG	CTCGAACACAATG
CNAG_05900	CNF00730	TCCTACATCTATG	TCCCACATCTATG	ATCATCAAAAATG	ATCATCAAAAATG
CNAG_05722	CNF02520	CCGATATTGTATG	ACGACATTGTATG	CTTAAAGAAAATG	ATCAAAGAAAATG
CNAG_04082	CNB05640	CGACAGATTCATG	CGACAGATTCATG	TATCCACATCATG	TATCCACACCATG
CNAG_01544	CNC06400	CAGCAGCTTTATG	CAGCAGCTTTATG	AGTCGCAAAAATG	TGTAGCAAAAATG
CNAG_04604	CNJ00430	CCACCCCTGCATG	CCACCCCTGCATG	GAAACCCACAATG	GAAACCCACAATG
CNAG_07473	CNB01880	ATAACGGTGTATG	ATAACGGTGTATG	GTACGCAACAATG	GTACGCAGCAATG

H99	JEC21	aATG.context.H99	aATG.context.JEC21	d1.context.H99	d1.context.JEC21
CNAG_06755	CNB00480	GGTATATCGTATG	GGTATATCGTATG	AGTACGCACTATG	AGCACGTACTATG

Table 1.7: Initiation factor 3 components in 12 fungal species.

DOI https://doi.org/10.7272/Q6T72FP1

Table 1.8: Oligonucleotides used in ribosome profiling.

Name	Sequence	Purpose
oCJ11	GATCGTCGGACTGTAGAACTCTGAACCTGTCG	cDNA synthesis
asDNA1	GTTTTTTTACTTATTCAATGAAGCGGAGCT	rRNA subtraction
asDNA2	GCGCCGGTGAAATACCACTACCTCCA	rRNA subtraction
asDNA3	TGTCGCATACACTGGTTGGGACTGAGGAATG	rRNA subtraction
asDNA4	ATCCGCAAGGAGCACCTTCGACCGATCCGG	rRNA subtraction
asDNA5	CCGAAGGGCATGCCTGTTTGAGAGTCATGA	rRNA subtraction
asDNA6	ATTACCCAATCCCGACACGGGGAGGTAGTGA	rRNA subtraction
asDNA7	TATGGTGAATCATAATAACTTCTCGAATCGCAT	rRNA subtraction
asDNA8	AGATTAAGCCATGCATGTCTAAGTATAAACGA	rRNA subtraction

References

1. Hawksworth, D.L. and Lücking, R. (2017) Fungal Diversity Revisited: 2.2 to 3.8 Million Species. *Microbiology Spectrum*, **5**.
2. Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F. *et al.* (2014) MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.*, **42**, D699-D704.
3. Fan, G., Sun, Q., Li, W., Shi, W., Li, X., Wu, L., Ma, J., Kim, C.Y., Lee, J.-S., Zhou, Y. *et al.* (2018) The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. *GigaScience*, **7**.
4. Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T. *et al.* (2018) Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell*, **175**, 1533-1545.e1520.
5. Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A.S., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L. *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature*, **459**, 657.
6. Dujon, B., Sherman, D., Fisher, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35-44.
7. Stajich, J.E. (2017) Fungal Genomes and Insights into the Evolution of the Kingdom. *Microbiology spectrum*, **5**, 10.1128/microbiolspec.FUNK-0055-2016.
8. Stajich, J.E., Dietrich, F.S. and Roy, S.W. (2007) Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol*, **8**, R223.
9. Coletta, A., Pinney, J.W., Solís, D.Y.W., Marsh, J., Pettifer, S.R. and Attwood, T.K. (2010) Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst. Biol.*, **4**, 43-43.

10. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. *et al.* (1996) Life with 6000 Genes. *Science*, **274**, 546.

11. Haas, B.J., Zeng, Q., Pearson, M.D., Cuomo, C.A. and Wortman, J.R. (2011) Approaches to Fungal Genome Annotation. *Mycology*, **2**, 118-141.

12. Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283-292.

13. Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125-8148.

14. Dever, T.E., Kinzy, T.G. and Pavitt, G.D. (2016) Mechanism and Regulation of Protein Synthesis in Saccharomyces cerevisiae. *Genetics*, **203**, 65-107.

15. Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L.B., Weinberger, A. and Segal, E. (2013) Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci USA*, **110**, E2792-E2801.

16. Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S. and Seelig, G. (2017) Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res*, **27**, 2015-2024.

17. Fervers, P., Fervers, F., Makałowski, W. and Jąkalski, M. (2018) Life cycle adapted upstream open reading frames (uORFs) in Trypanosoma congolense: A post-transcriptional approach to accurate gene regulation. *PLOS ONE*, **13**, e0201461.

18. Duncan, C.D.S., Rodríguez-López, M., Ruis, P., Bähler, J. and Mata, J. (2018) General amino acid control in fission yeast is regulated by a nonconserved transcription factor, with functions analogous to Gcn4/Atf4. *Proc Natl Acad Sci USA*, **115**, E1829-E1838.

19. Sundaram, A. and Grant, C.M. (2014) A single inhibitory upstream open reading frame (uORF) is sufficient to regulate Candida albicans GCN4 translation in response to amino acid starvation conditions. *RNA (New York, N.Y.)*, **20**, 559-567.

20. Ivanov, I.P., Wei, J., Caster, S.Z., Smith, K.M., Michel, A.M., Zhang, Y., Firth, A.E., Freitag, M., Dunlap, J.C., Bell-Pedersen, D. *et al.* (2017) Translation Initiation from Conserved Non-AUG Codons Provides Additional Layers of Regulation and Coding Capacity. *mBio*, **8**, e00844-00817.

21. von Arnim, A.G., Jia, Q. and Vaughn, J.N. (2014) Regulation of plant translation by upstream open reading frames. *Plant Sci.*, **214**, 1-12.

22. Barbosa, C., Peixeiro, I. and Romão, L. (2013) Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet*, **9**, e1003529-e1003529.

23. Chen, S.-J., Lin, G., Chang, K.-J., Yeh, L.-S. and Wang, C.-C. (2008) Translational Efficiency of a Non-AUG Initiation Codon Is Significantly Affected by Its Sequence Context in Yeast. *J Biol Chem*, **283**, 3173-3180.

24. Hinnebusch, A.G., Ivanov, I.P. and Sonenberg, N. (2016) Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science*, **352**, 1413-1416.

25. Wethmar, K. (2014) The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdisciplinary Reviews: RNA*, **5**, 765-768.

26. Llácer, J.L., Hussain, T., Marler, L., Aitken, C.E., Thakur, A., Lorsch, J.R., Hinnebusch, A.G. and Ramakrishnan, V. (2015) Conformational Differences between Open and Closed States of the Eukaryotic Translation Initiation Complex. *Mol Cell*, **59**, 399-412.

27. Hinnebusch, A.G. (2017) Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. *Trends Biochem. Sci.*, **42**, 589-611.

28. Llácer, J.L., Hussain, T., Saini, A.K., Nanda, J.S., Kaur, S., Gordiyenko, Y., Kumar, R., Hinnebusch, A.G., Lorsch, J.R. and Ramakrishnan, V. (2018) Translational initiation factor eIF5 replaces eIF1 on the 40S ribosomal subunit to promote start-codon recognition. *eLife*, **7**, e39273.

29. Janbon, G. (2018) Introns in Cryptococcus. *Mem. Inst. Oswaldo Cruz*, **113**, e170519-e170519.

30. Goebels, C., Thonn, A., Gonzalez-Hilarion, S., Rolland, O., Moyrand, F., Beilharz, T.H. and Janbon, G. (2013) Introns regulate gene expression in *Cryptococcus neoformans* in a Pab2p dependent pathway. *PLoS Genet*, **9**, e1003686.

31. Dumesic, P.A., Natarajan, P., Chen, C., Drinnenberg, I.A., Schiller, B.J., Thompson, J.D., Moresco, J.J., Yates Iii, J.R., Bartel, D.P. and Madhani, H.D. (2013) Stalled spliceosomes are a signal for RNAi-mediated genome defense. *Cell*, **152**, 957-968.

32. Bonnet, A., Grosso, A.R., Elkaoutari, A., Coleno, E., Presle, A., Sridhara, S.C., Janbon, G., Géli, V., de Almeida, S.F. and Palancade, B. (2017) Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. *Mol Cell*, **67**, 608-621.e606.

33. Janbon, G., Ormerod, K.L., Paulet, D., Byrnes III, E.J., Chatterjee, G., Yadav, V., Mullapudi, N., Hon, C.C., Billmyre, R.B., Brunel, F. *et al.* (2014) Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. *grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genet*, **10**, e1004261.

34. Gonzalez-Hilarion, S., Paulet, D., Lee, K.-T., Hon, C.-C., Lechat, P., Mogensen, E., Moyrand, F., Proux, C., Barboux, R., Bussotti, G. *et al.* (2016) Intron retention-dependent gene regulation in Cryptococcus neoformans. *Sci. Rep.*, **6**, 32252.

35. Winston, F., Dollard, C. and Ricupero-Hovasse, S.L. (1995) Construction of a set of convenient saccharomyces cerevisiae strains that are isogenic to S288C. *Yeast*, **11**, 53-55.

36. Lee, N. and Janbon, G. (2006) In Kavanagh, K. (ed.), *Med Mycol*, pp. 275-304.

37. Moyrand, F., Lafontaine, I., Fontaine, T. and Janbon, G. (2008) *UGE1* and *UGE2* regulate the UDP-glucose/UDP-galactose equilibrium in *Cryptococcus neoformans*. *Eukaryot Cell*, **7**, 2069-2077.

38. Malabat, C., Feuerbach, F., Ma, L., Saveanu, C. and Jacquier, A. (2015) Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *Elife*, **4**, e06722.

39. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14**, R36-R36.

40. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*.

41. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, **324**, 218.

42. Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R. and Weissman, J.S. (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife*, **2**, e01179-e01179.

43. Carja, O., Xing, T., Wallace, E.W.J., Plotkin, J.B. and Shah, P. (2017) riboviz: analysis and visualization of ribosome profiling datasets. *BMC bioinformatics*, **18**, 461-461.

44. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. and Salzberg, S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, **11**, 1650.

45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078-2079.

46. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, **26**, 841-842.

47. R Core Team. (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

48. Wickham, H. (ed.) (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

49. Wickham, H., François, R., Henry, L. and Müller, K. (2018) dplyr: A Grammar of Data Manipulation. R package version 0.7.8.

50. Wilke, C.O. (2018) cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 0.9.3.

51. Wagih, O. (2017) ggseqlogo: A 'ggplot2' Extension for Drawing Publication-Ready Sequence Logos. R package version 0.1.

52. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

53. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792-1797.

54. Wilm, A., Higgins, D.G., Valentin, F., Blackshields, G., McWilliam, H., Wallace, I.M., Thompson, J.D., Larkin, M.A., Brown, N.P., McGettigan, P.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.

55. Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S. and Liu, Y. (2015) Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol Cell*, **59**, 744-754.

56. Kersey, P.J., Allen, J.E., Allot, A., Barba, M., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C. *et al.* (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.*, **46**, D802-D808.

57. Duncan, C.D.S. and Mata, J. (2017) Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. *Sci. Rep.*, **7**, 10331.

58. Muzzey, D., Sherlock, G. and Weissman, J.S. (2014) Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans. *Genome Res*, **24**, 963-973.

59. Skrzypek, M.S., Binkley, J., Binkley, G., Miyasato, S.R., Simison, M. and Sherlock, G. (2017) The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.*, **45**, D592-D596.

60. Gerashchenko, M.V. and Gladyshev, V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.*, **42**, e134-e134.

61. Csárdi, G., Franks, A., Choi, D.S., Airoldi, E.M. and Drummond, D.A. (2015) Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *PLoS Genet*, **11**, e1005206.

62. Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B. and Bartel, D.P. (2016) Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.*, **14**, 1787-1799.

63. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700-D705.

64. Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A. and Zdobnov, E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807-D811.

65. Mi, H., Muruganujan, A. and Thomas, P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377-D386.

66. Basenko, Y.E., Pulman, A.J., Shanmugasundram, A., Harb, S.O., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecoechea, C., Stoeckert, J.C., Kissinger, C.J. *et al.* (2018) FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *J. Fungi*, **4**.

67. Ban, N., Beckmann, R., Cate, J.H.D., Dinman, J.D., Dragon, F., Ellis, S.R., Lafontaine, D.L.J., Lindahl, L., Liljas, A., Lipton, J.M. *et al.* (2014) A new system for naming ribosomal proteins. *Current opinion in structural biology*, **24**, 165-169.

68. Li, H., Hou, J., Bai, L., Hu, C., Tong, P., Kang, Y., Zhao, X. and Shao, Z. (2015) Genome-wide analysis of core promoter structures in Schizosaccharomyces pombe with DeepCAGE. *RNA Biol.*, **12**, 525-537.

69. Neafsey, D.E. and Galagan, J.E. (2007) Dual Modes of Natural Selection on Upstream Open Reading Frames. *Mol Biol Evol*, **24**, 1744-1751.

70. Hinnebusch, A.G. (2005) Translational regulation of *GCN4* and the general amino acid control of yeast. *Ann Rev Microbiol*, **59**, 407-450.

71. Duncan, C.D.S., Rodríguez-López, M., Ruis, P., Bähler, J. and Mata, J. (2018) General amino acid control in fission yeast is regulated by a nonconserved transcription factor, with functions analogous to Gcn4/Atf4. *Proc Natl Acad Sci USA*, **115**, E1829.

72. Madi, L., McBride, S.A., Bailey, L.A. and Ebbole, D.J. (1997) rco-3, a gene involved in glucose transport and conidiation in Neurospora crassa. *Genetics*, **146**, 499-508.

73. Wiese, A., Elzinga, N., Wobbes, B. and Smeekens, S. (2005) Sucrose-induced translational repression of plant bZIP-type transcription factors. *Biochemical Society Transactions*, **33**, 272.

74. Gaba, A., Wang, Z., Krishnamoorthy, T., Hinnebusch, A.G. and Sachs, M.S. (2001) Physical evidence for distinct mechanisms of translational control by upstream open reading frames. *The EMBO journal*, **20**, 6453-6463.

75. Kervestin, S. and Jacobson, A. (2012) NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol*, **13**, 703-712.

76. Arribere, J.A. and Gilbert, W.V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res*, **23**, 977-987.

77. Hood, H.M., Spevak, C.C. and Sachs, M.S. (2007) Evolutionary changes in the fungal carbamoyl-phosphate synthetase small subunit gene and its associated upstream open reading frame. *Fungal Genet Biol*, **44**, 93-104.

78. Gaba, A., Jacobson, A. and Sachs, M.S. (2005) Ribosome Occupancy of the Yeast CPA1 Upstream Open Reading Frame Termination Codon Modulates Nonsense-Mediated mRNA Decay. *Mol Cell*, **20**, 449-460.

79. Zhang, Y. and Sachs, M.S. (2015) Control of mRNA Stability in Fungi by NMD, EJC and CBC Factors Through 3′UTR Introns. *Genetics*, **200**, 1133.

80. Wei, J., Zhang, Y., Ivanov, I.P. and Sachs, M.S. (2013) The stringency of start codon selection in the filamentous fungus Neurospora crassa. *The Journal of biological chemistry*, **288**, 9549-9562.

81. Spealman, P., Naik, A.W., May, G.E., Kuersten, S., Freeberg, L., Murphy, R.F. and McManus, J. (2018) Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res*, **28**, 214-222.

82. Danpure, C.J. (1995) How can the products of a single gene be localized to more than one intracellular compartment? *Trends Cell Biol.*, **5**, 230-238.

83. Silva-Filho, M.C. (2003) One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. *Current Opinion in Plant Biology*, **6**, 589-595.

84. Mireau, H., Lancelin, D. and Small, I.D. (1996) The same Arabidopsis gene encodes both cytosolic and mitochondrial alanyl-tRNA synthetases. *Plant Cell*, **8**, 1027-1039.

85. Mudge, S.J., Williams, J.H., Eyre, H.J., Sutherland, G.R., Cowan, P.J. and Power, D.A. (1998) Complex organisation of the 5′-end of the human glycine tRNA synthetase gene. *Gene*, **209**, 45-50.

86. Natsoulis, G., Hilger, F. and Fink, G.R. (1986) The HTS1 gene encodes both the cytoplasmic and mitochondrial histidine tRNA synthetases of S. cerevisiae. *Cell*, **46**, 235-243.

87. Datt, M. and Sharma, A. (2014) Novel and unique domains in aminoacyl-tRNA synthetases from human fungal pathogens Aspergillus niger, Candida albicans and Cryptococcus neoformans. *BMC Genomics*, **15**, 1069.

88. Duchêne, A.-M., Pujol, C. and Maréchal-Drouard, L. (2009) Import of tRNAs and aminoacyl-tRNA synthetases into mitochondria. *Curr Genet*, **55**, 1-18.

89. Muruganujan, A., Ebert, D., Mi, H., Thomas, P.D. and Huang, X. (2018) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419-D426.

90. Frechin, M., Duchêne, A.-M. and Becker, H.D. (2009) Translating organellar glutamine codons : A case by case scenario? *RNA Biol.*, **6**, 31-34.

91. Chang, C.-P., Tseng, Y.-K., Ko, C.-Y. and Wang, C.-C. (2012) Alanyl-tRNA synthetase genes of Vanderwaltozyma polyspora arose from duplication of a dual-functional predecessor of mitochondrial origin. *Nucleic Acids Res.*, **40**, 314-322.

92. Geslain, R., Martin, F., Delagoutte, B., Cavarelli, J., Gangloff, J. and Eriani, G. (2000) In vivo selection of lethal mutations reveals two functional domains in arginyl-tRNA synthetase. *RNA (New York, N.Y.)*, **6**, 434-448.

93. Merz, S. and Westermann, B. (2009) Genome-wide deletion mutant analysis reveals genes required for respiratory growth, mitochondrial genome maintenance and mitochondrial protein synthesis in Saccharomyces cerevisiae. *Genome Biol*, **10**, R95.

94. Sickmann, A., Reinders, J., Wagner, Y., Joppich, C., Zahedi, R., Meyer, H.E., Schönfisch, B., Perschil, I., Chacinska, A., Guiard, B. *et al.* (2003) The proteome of Saccharomyces cerevisiae mitochondria. *Proc Natl Acad Sci USA*, **100**, 13207-13212.

95. Chen, S.-J., Wu, Y.-H., Huang, H.-Y. and Wang, C.-C. (2012) Saccharomyces cerevisiae Possesses a Stress-Inducible Glycyl-tRNA Synthetase Gene. *PLOS ONE*, **7**, e33363.

96. Chiu, W.-C., Chang, C.-P., Wen, W.-L., Wang, S.-W. and Wang, C.-C. (2010) Schizosaccharomyces pombe Possesses Two Paralogous Valyl-tRNA Synthetase Genes of Mitochondrial Origin. *Mol Biol Evol*, **27**, 1415-1424.

97. Ivanov, I.P., Loughran, G., Sachs, M.S. and Atkins, J.F. (2010) Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc Natl Acad Sci USA*, **107**, 18056-18060.

98. Loughran, G., Sachs, M.S., Atkins, J.F. and Ivanov, I.P. (2012) Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. *Nucleic Acids Res.*, **40**, 2898-2906.

99. Martin-Marcos, P., Cheung, Y.-N. and Hinnebusch, A.G. (2011) Functional elements in initiation factors 1, 1A, and 2β discriminate against poor AUG context and non-AUG start codons. *Mol Cell Biol*, **31**, 4814-4831.

100. Hussain, T., Llácer, J.L., Fernández, I.S., Munoz, A., Martin-Marcos, P., Savva, C.G., Lorsch, J.R., Hinnebusch, A.G. and Ramakrishnan, V. (2014) Structural changes enable start codon recognition by the eukaryotic translation initiation complex. *Cell*, **159**, 597-607.

101. Thakur, A. and Hinnebusch, A.G. (2018) eIF1 Loop 2 interactions with Met-tRNA(i) control the accuracy of start codon selection by the scanning preinitiation complex. *Proc Natl Acad Sci USA*, **115**, E4159-E4168.

102. Olsen, D.S., Savner, E.M., Mathew, A., Zhang, F., Krishnamoorthy, T., Phan, L. and Hinnebusch, A.G. (2003) Domains of eIF1A that mediate binding to eIF2, eIF3 and eIF5B and promote ternary complex recruitment in vivo. *EMBO J.*, **22**, 193-204.

103. Luna, R.E., Arthanari, H., Hiraishi, H., Akabayov, B., Tang, L., Cox, C., Markus, M.A., Luna, L.E., Ikeda, Y., Watanabe, R. *et al.* (2013) The interaction between eukaryotic initiation factor 1A and eIF5 retains eIF1 within scanning preinitiation complexes. *Biochemistry*, **52**, 9510-9518.

104. Fekete, C.A., Applefield, D.J., Blakely, S.A., Shirokikh, N., Pestova, T., Lorsch, J.R. and Hinnebusch, A.G. (2005) The eIF1A C-terminal domain promotes initiation complex assembly, scanning and AUG selection in vivo. *EMBO J.*, **24**, 3588-3601.

105. Slusher, L.B., Gillman, E.C., Martin, N.C. and Hopper, A.K. (1991) mRNA leader length and initiation codon context determine alternative AUG selection for the yeast gene MOD5. *Proc Natl Acad Sci USA*, **88**, 9789.

106. Calvo, S.E., Pagliarini, D.J. and Mootha, V.K. (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA*, **106**, 7507-7512.

107. Cheng, Z., Otto, G.M., Powers, E.N., Keskin, A., Mertins, P., Carr, S.A., Jovanovic, M. and Brar, G.A. (2018) Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell*, **172**, 910-923.e916.

108. Van Dalfsen, K.M., Hodapp, S., Keskin, A., Otto, G.M., Berdan, C.A., Higdon, A., Cheunkarndee, T., Nomura, D.K., Jovanovic, M. and Brar, G.A. (2018) Global Proteome Remodeling during ER Stress Involves Hac1-Driven Expression of Long Undecoded Transcript Isoforms. *Dev Cell*, **46**, 219-235.e218.

109. Monteuuis, G., Miścicka, A., Świrski, M., Zenad, L., Niemitalo, O., Wrobel, L., Alam, J., Chacinska, A., Kastaniotis, A.J., Kufel, J. *et al.* (2019) Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. *Nucleic Acids Res.*, in press.

110. Brar, G.A. (2016) Beyond the Triplet Code: Context Cues Transform Translation. *Cell*, **167**, 1681-1692.

111. Feeney, K.A., Hansen, L.L., Putker, M., Olivares-Yañez, C., Day, J., Eades, L.J., Larrondo, L.F., Hoyle, N.P., O'Neill, J.S. and van Ooijen, G. (2016) Daily magnesium fluxes regulate cellular timekeeping and energy balance. *Nature*, **532**, 375-379.

112. Tsuboi, T., Viana, M.P., Xu, F., Yu, J., Chanchani, R., Arceo, X.G., Tutucci, E., Choi, J., Chen, Y.S., Singer, R.H. *et al.* (2019) Mitochondrial volume fraction controls translation of nuclear-encoded mitochondrial proteins. *bioRxiv*, 529289.

113. Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H. and Miura, K.-i. (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.*, **36**, 861-871.

114. Shah, M., Su, D., Scheliga, J.S., Pluskal, T., Boronat, S., Motamedchaboki, K., Campos, A.R., Qi, F., Hidalgo, E., Yanagida, M. *et al.* (2016) A Transcript-Specific eIF3 Complex Mediates Global Translational Control of Energy Metabolism. *Cell Rep.*, **16**, 1891-1902.

115. Fields, S.D., Conrad, M.N. and Clarke, M. (1998) The S. cerevisiae CLU1 and D. discoideum cluA genes are functional homologues that influence mitochondrial morphology and distribution. *J. Cell Sci.*, **111**, 1717.

116. Gao, J., Schatton, D., Martinelli, P., Hansen, H., Pla-Martin, D., Barth, E., Becker, C., Altmueller, J., Frommolt, P., Sardiello, M. *et al.* (2014) CLUH regulates mitochondrial biogenesis by binding mRNAs of nuclear-encoded mitochondrial proteins. *J. Cell Biol*, **207**, 213-223.

117. Schatton, D., Pla-Martin, D., Marx, M.-C., Hansen, H., Mourier, A., Nemazanyy, I., Pessia, A., Zentis, P., Corona, T., Kondylis, V. *et al.* (2017) CLUH regulates mitochondrial

metabolism by controlling translation and decay of target mRNAs. *J. Cell Biol*, **216**, 675-693.

118. Smith, M.D., Gu, Y., Querol-Audí, J., Vogan, J.M., Nitido, A. and Cate, J.H.D. (2013) Human-Like Eukaryotic Translation Initiation Factor 3 from Neurospora crassa. *PLOS ONE*, **8**, e78715.

119. Madhani, H.D. (2013) The frustrated gene: origins of eukaryotic gene expression. *Cell*, **155**, 744-749.

Chapter 2

A non-Dicer RNase III and four other novel factors required for RNAi-mediated transposon suppression in the human pathogenic yeast Cryptococcus neoformans

Jordan E. Burke*, Adam D. Longhurst*, Prashanthi Natarajan*, Beiduo Rao*, S. John Liu*, Jade Sales-Lee*, Yasaman Mortensen*, James J. Moresco†, Jolene K. Diedrich†, John R. Yates III‡, Hiten D. Madhani§

Author affiliations:

*Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158, USA.

†Salk Institute for Biological Studies, La Jolla, CA 92037, USA.

‡Department of Molecular Medicine, the Scripps Research Institute, La Jolla, CA, USA.

§Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158, USA; Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA.

Abstract

The human pathogenic yeast *Cryptococcus neoformans* silences transposable elements using endo-siRNAs and an Argonaute, Ago1. Endo-siRNAs production requires the RNA-dependent RNA polymerase, Rdp1, and two partially redundant Dicer enzymes, Dcr1 and Dcr2, but is independent of histone H3 lysine 9 methylation. We describe here an insertional mutagenesis screen for factors required to suppress the mobilization of the *C. neoformans HARBINGER* family DNA transposon *HAR1*. Validation experiments uncovered five novel genes (*RDE1-5*) required for *HAR1* suppression and global production of suppressive endo-siRNAs. The *RDE* genes do not impact transcript levels, suggesting the endo-siRNAs do not act by impacting target transcript synthesis or turnover. *RDE3* encodes a non-Dicer RNase III related to *S. cerevisiae* Rnt1, *RDE4* encodes a predicted terminal nucleotidyltransferase, while *RDE5* has no strongly predicted encoded domains. Affinity purification-mass spectrometry studies suggest that Rde3 and Rde5 are physically associated. *RDE1* encodes a G-patch protein homologous to the *S. cerevisiae* Sqs1/Pfa1, a nucleolar protein that directly activates the essential helicase Prp43 during rRNA biogenesis. Rde1 copurifies Rde2, another novel protein obtained in the screen, as well as Ago1, a homolog of Prp43, and numerous predicted nucleolar proteins. We also describe the isolation of conditional alleles of *PRP43*, which are defective in RNAi. This work reveals unanticipated requirements for a non-Dicer RNase III and presumptive nucleolar factors for endo-siRNA biogenesis and transposon mobilization suppression in *C. neoformans*.

Introduction

Transposons are ancient mobile genetic elements that have invaded the genomes of nearly all living organisms and can make up substantial proportions of host DNA (Nekrutenko and Li 2001). DNA transposons and retrotransposons that retain their ability to mobilize and proliferate can both mutagenize and disrupt regulation of the host genome (Chuong *et al.* 2017). Host organisms such as fungi have therefore developed a diverse set of mechanisms to defend against transposable elements, including strategies at the levels of DNA and RNA. Transcriptional silencing of transposons is achieved via histone modifications and DNA methylation and transposons are also neutralized by repeat-induced point mutation (RIP) in some systems (Muszewska *et al.* 2017). RNAi based mechanisms for post-transcriptional silencing are found in fungi as distantly related as the ascomycete *Neurospora crassa* and the basidiomycete *Cryptococcus neoformans* (Billmyre *et al.* 2013) and the role of RNAi has shifted in *Schizosaccharomyces pombe* to regulate the formation of heterochromatin (Volpe *et al.* 2002). However, fungi such as *Saccharomyces cerevisiae* and even a subtype of *Cryptococcus gattii* have lost the RNAi machinery (Billmyre *et al.* 2013).

In *Cryptococcus neoformans*, transposable elements are silenced by endogenously produced small RNAs (endo-siRNAs). *C. neoformans* harbors a semi-canonical RNAi pathway including of a single Argonaute (Ago1), two partially redundant Dicers (Dcr1, Dcr2) and an RNA-dependent RNA polymerase (Rdp1) (Janbon *et al.* 2010). Ago1 is found in both a nuclear complex, Spliceosome-coupled and Nuclear RNAi or SCANR complex, and a complex with Gwo1 that localizes to P-bodies, the P-body associated RNA Silencing Complex or PRSC (Dumesic *et al.* 2013). siRNA biogenesis and transposon silencing appear to be particularly important during meiosis (Wang *et al.* 2010); however, transposon mobilization may also occur in vegetative cells and is suppressed by the RNAi machinery (Wang *et al.* 2010; Dumesic *et al.* 2013).

While endo-siRNA systems vary from organism to organism (Claycomb 2014), one common characteristic is the presence of a triggering double stranded RNA (dsRNA) species. In some cases, the dsRNA is produced by transcription of repetitive elements that form inter- and intra-molecular duplexes (Sijen and Plasterk 2003; Slotkin *et al.* 2005). In other organisms, an RNA-dependent RNA polymerase, such Rdp1 in *C. neoformans* and QDE-1 in *N. crassa*, is thought to produce dsRNA (Cogoni and Macino 1999; Lee *et al.* 2010; Janbon *et al.* 2010). In the latter case, the trigger for production of dsRNA is unclear, as is the manner in which the RNA-dependent RNA polymerase is recruited. More generally, how invasive genetic elements are detected by the host genome defense machinery remains unclear.

Genome defense mechanisms can be triggered by introduction of transgenes and repetitive sequences. In fact, several transposon silencing pathways were discovered due to co-suppression of transgenes and cognate endogenous genes (Billmyre *et al.* 2013). In *C. neoformans*, the RNAi pathway was initially characterized by studying the co-suppression of multiple copies of a *SXI2**a*-*URA5* transgene (Wang *et al.* 2010). Upon mating and selection on media that selects against uracil biosynthesis, strains can be recovered that silence all *URA5* loci by RNAi. This suggests that cells sense either the copy number or expression level of genes and targets the RNAi pathway against anomalous transcripts. However, the manner in which such events are detected is unknown.

To pursue this question, we searched for additional components of *C. neoformans* endo-siRNA production pathway, using an active DNA transposon, *HAR1*, a member of the *HARBINGER* family as a reporter for the ability of the cell to silence transposable elements. *HARBINGER* is a cut-and-paste DNA transposon with a DDE nuclease that excises double-stranded DNA directly and reinserts the gene encoding the transposon elsewhere in the host genome (Muszewska *et al.* 2017). We took advantage of an active copy of *HARBINGER* that is silenced by the RNAi machinery in *C. neoformans* (Wang *et al.* 2010) to search for factors

required for transposable element silencing. We report here five new genes required for production of endo-siRNAs and the suppression of *HAR1* mobilization.

Results

A colony-level assay to measure the mobilization of a *HARBINGER* DNA transposon

To screen for previously unreported factors involved in silencing transposable elements, we developed a reporter that links uracil prototrophy to mobilization of the HARBINGER DNA transposon. HARBINGER is suppressed by the RNAi pathway in C. neoformans (Wang et al. 2010). A copy of HARBINGER, HAR1 (CNAG_02711) was inserted into the second intron of URA5 (CNAG_03196), grossly disrupting the intron and resulting in uracil auxotrophy (Figure 2.1A). To estimate the rate of HAR1 transposition, cell populations in which URA5 is still disrupted by HARBINGER were isolated by selection on 5-FOA. The selected cells were then transferred directly to rich media (YPAD) or media lacking uracil, and colony forming units (CFUs) were counted between 2-6 days at 30°C (Figure 2.1B). Upon deletion of genes encoding a core RNAi factor such as Argonaute (AGO1) or the RNA-dependent RNA polymerase (RDP1), the number of CFUs on media lacking uracil increased 1,000 to 10,000-fold in comparison to rich media (Figure 2.1C), consistent with loss of RNAi-based transposition suppression. The ura5::HAR1 system provides a platform for high-throughput screening for factors involved in HARBINGER silencing.

Insertional mutagenesis uncovers previously unidentified RNAi factors

We next performed insertional mutagenesis of the *C. neoformans* genome in the *ura5::HAR1* background. A strain of *A. tumefaciens* carrying a nourseothricin (NAT) resistance cassette bounded by *T-DNA* inverted repeat terminal sequences (McClelland *et al.* 2005) was co-cultured with *C. neoformans* resulting in ~25,000 NAT-resistant transformants (Figure 2.1D). The mutant pool yielded 96 uracil prototroph strains within 2-6 days at 30°C.

T-DNA insertions in all NAT resistant (NATR+) and uracil prototroph (Ura+) strains were identified by performing linear PCR originating in the NATR cassette followed by purification of the biotinylated ssDNA from genomic DNA [Figure 2.1E, adapted from (Schmidt *et al.* 2007)]. The upstream flanking sequence was then amplified by nested PCR and insertion sites were identified by high throughput sequencing and alignment to the *C. neoformans* genome (Figure 2.1F). We detected insertions with at least 10 reads in 1804 genes in the NATR+ pool and 752 genes in the Ura+/NATR+ pool. We further refined the list of enriched insertion sites by determining Z-scores from the $\log_2$ ratio of reads in the Ura+/NATR+ pool to reads in the NATR+ pool (Figure 2.1G, black, Table 2.1). The 97 genes with a Z-score of at least 2.0 included four genes encoding known RNAi factors (*AGO1*, *RDP1*, *GWO1*, and *QIP1*) (Dumesic *et al.* 2013) and the gene neighboring *RDP1* (Figure 2.1G, magenta). Interestingly, an insertion in one of the *HARBINGER* loci, *HAR2* (*CNAG_00903*) exhibited two-fold enrichment in the uracil prototroph pool (Table 2.1), suggesting that decreased expression of *HAR2* might improve the ability of *HAR1* to transpose. The remaining annotated hits primarily occur in factors involved in starvation response, nutrient and small molecule transport, DNA repair and RNA processing (Figure 2.1H).

To validate the initial hits of the screen, we isolated RNA from 52 available coding sequence knockouts from a *C. neoformans* knockout collection being constructed in our laboratory and detected siRNAs against the endo siRNA-producing locus *CNAG_06705* by small RNA northern analysis (data not shown). Five of the knockouts that exhibited partial or complete loss of siRNAs (Figure 2.2A) were selected for further study and named *RDE1-5* (Figure 2.2B; *R*NAi-*DE*fective). Knockouts of these five genes also exhibited decreased levels of siRNAs against all three copies of *HARBINGER* (Figure 2.2C) and dramatically increased transposition of *HAR1* (Figure 2.2D and Table 2.2). Finally, RNA-seq analysis of the knockout strains revealed minimal differential transcript expression and no reduction in the expression of RNAi pathway members (Figure 2.2E and Table 2.3).

Mutants of new RNAi pathway members display defects in siRNA accumulation

To determine the extent and specificity of the RNAi defect in the newly discovered RNAi pathway members, we compared the small RNA populations in the knockouts to wild type and knockouts of canonical pathway members by sequencing the total 15-30 nt RNA population in the cell (small RNA-seq). In wild type, we observe that the majority of reads fall within the range of 21-24 nt (Figure 2.3A-B). Upon loss of a member of the RNAi pathway such as *AGO1*, this peak is no longer discernable (Figure 2.3A, panel 1, purple) and the proportion of reads between 21-24 nt decreases (Figure 2.3B). In contrast, when transcriptional silencing of siRNA targets in heterochromatic regions is lost, as in the case of *clr4Δ* (whose gene encodes the sole H3K9 methylase in *C. neoformans*) the 21-24 nt small RNA population increases (Figure 2.3A, panel 1, yellow). In mutants of each of the factors identified in this screen we observe a different extent of loss of 21-24 nt siRNAs with *rde4Δ* and *rde5Δ* exhibiting the most severe loss and *rde2Δ* exhibiting the least (Figure 2.3A-B).

To address whether the observed decrease in siRNAs was specific to certain targets, we also counted small RNA reads antisense to genes (Figure 2.3C, Table 2.4) or transposable elements (Figure 2.3D, Table 2.4). We observe a similar pattern of siRNA abundance changes in the knockouts of RNAi pathway members and the newly discovered factors (Figure 2.3C-D), which differs primarily in the magnitude of the small RNA loss. Notably, siRNAs that decrease when RNAi pathway members are lost are largely mutually exclusive with siRNAs that increase when heterochromatic silencing is lost upon deletion of *CLR4* (Figure 2.3E). Additionally, deletion of *EZH2,* the histone 3 lysine 27 methyltransferase component of Polycomb complex (Dumesic *et al.* 2015) has little effect on the small RNA population (Figure 2.3C-D). Taken together, these results indicate that the factors identified in our screen are required for the biosynthesis of endo-siRNAs in *C. neoformans*, while heterochromatin formation antagonizes

siRNA production, presumably by limiting expression of the transcriptions that serve as templates for siRNA production.

Nucleolar protein homologs required for endo-siRNA production

To expand our understanding of the function and localization of the newly discovered RNAi factors, we performed tandem affinity purification followed by label-free mass spectrometry analysis on each Rde factor using a C-terminal epitope tag [Calmodulin binding protein (CBP), 2xFLAG]. Mass spectrometry of purified Rde1, which contains a G-patch domain thought to interact with DEXD-box helicases (Figure 2.4A), primarily detected ribosomal proteins, factors involved in translation and nucleolar proteins (Figure 2.4B, Table 2.5). The most abundant protein detected with Rde1 is with the essential DEXD-Box helicase Prp43, which is both a ribosome biogenesis and pre-mRNA splicing factor. The *S. cerevisiae* ortholog of Rde1, Sqs1/Pfa1, also associates with Prp43 and is implicated in ribosome biogenesis (Lebaron *et al.* 2009; Pandit *et al.* 2009; Pertschy *et al.* 2009). Notably, Ago1 and Rde2 are also detected in the Rde1-CBP-2xFLAG purified material. A single nucleolar protein, Nop1 is detected in the Rde2-CBP-2xFLAG purified material; however, Prp43 and Ago1 were not detected (Figure 2.4C).

To determine whether Prp43 function is important for endo-siRNA biogenesis, we generated a randomly mutagenized library of *PRP43* alleles and screened for mutants that increased transposition of *HAR1* using the *ura5::HAR1* system (Figure 2.4D). We identified two alleles of *PRP43*, termed *prp43-ts5* (K277R, T29A, H511R, R620Q) and *prp43-ts12* (F708S), that result in increased transposition of *HARBINGER* as well as a severe growth defect at 37°C (data not shown). Reconstructions of these alleles also exhibited increased transposition of *HAR1* (Figure 2.4E, *prp43-ts12* not shown) and loss of siRNA production (Figure 2.4F). In the presence of *prp43-ts5,* the amount of Ago1 in the Rde1-CBP-2xFLAG purified material increases and the P-body localized RNAi factor, Gwo1 is also detected. Additionally, RT-qPCR

reveals that enrichment for the 18S ribosomal RNA by both Rde1 and Prp43 is reduced in the *prp43-ts5* background (Figure 2.4H).

RNA processing factors link RNAi with RNA surveillance

The remaining RNAi factors contain domains that suggest they are involved in mRNA processing and RNA surveillance (Figure 2.5A). Rde3 contains an RNase III domain, but no PAZ domain, suggesting that it can cleave double stranded RNA (dsRNA) but is not a canonical Dicer enzyme. Rde4 contains a terminal-nucleotidyl transferase domain commonly found in terminal-uridylyl transferases and polyA polymerases, while Rde5 has no strong domain predictions, but may contain a disordered region between amino acids 293-317 (Piovesan *et al.* 2018).

In contrast with Rde1-2, rRNA processing and nucleolar proteins are largely not detected in material purified by Rde3-CBP-2xFLAG, Rde4-CBP-2xFLAG or Rde5-CBP-2xFLAG. . Rde3, the putative RNase III, is detected in Rde5-CBP-2xFLAG purified material and vice versa. A variety of other nuclear proteins (Figure 2.5B) are also present in both the Rde3 and Rde5 data sets. RNA quality control enzymes, such as Rnh1, an RNase H that cleaves RNA-DNA duplexes, as well as Rrp6, a component of the nuclear exosome are detected in Rde3-CBP-2xFLAG purified material. Spt16 and Pob3, members of the FACT complex involved in chromatin remodeling (Lejeune *et al.* 2007), Nop1, a nucleolar protein, and Fkbp4 are all detected in both Rde3-CBP-2xFLAG and Rde5-CBP-2xFLAG purified material. Interestingly, Aga1, which is a known binding partner of Ago1 (Dumesic *et al.* 2013) is also detected in the Rde5-CBP-2xFLAG material. Rde4, the putative terminal nucleotidyltransferase, did not co-purify with any proteins aside from common contaminants (Table 2.5).

Impact on RNAi target transcript levels and siRNA abundance by the nuclear exosome

Finally, to investigate the connection between RNAi and nuclear RNA surveillance and quality control, we performed RNA-seq and small RNA-seq in a strain lacking the nuclear exosome component, Rrp6. Consistent with the known role of Rrp6 in other systems, deletion of *RRP6* results in substantial changes in the mature RNA population (Figure 2.5C). While many transcripts are significantly more abundant in the *rrp6Δ* strain (Figure 2.5C), targets of the RNAi pathway exhibit a greater increase in abundance (Figure 2.5D), suggesting they are turned over at a somewhat higher rate by the nuclear exosome than transcripts on average.

Additionally, a subset of small RNAs increase in abundance in *rrp6Δ* compared with wild type (Figure 2.5E). However, these small RNAs could just be degradation products of suboptimal transcripts that are normally degraded by the nuclear exosome. The size distribution of the small RNAs increased in *rrp6Δ* indicates that these are canonical 21-24 nt siRNAs (Figure 2.5F). The small RNAs in these regions are also dependent on *AGO1,* indicating that they are normally produced by the canonical RNAi pathway (Figure 2.5F). Finally, the mRNA targets of these small RNAs exhibit a modestly significant differential increase in expression in *rrp6Δ* compared with the general mRNA population (Figure 2.5G). Together with the result that the RNase III Rde3 may associate with Rrp6 and other RNA surveillance factors, our findings are consistent with a competition relationship between nuclear RNA surveillance and the production of endo-siRNAs in *C. neoformans.*

Discussion

In this study, we describe a genetic screen aimed at identifying novel factors involved in transposon mobilization in *C. neoformans*. We report the impact of deletion alleles of five genes identified in the screen, *RDE1-5*, on global endo-siRNA levels, global RNA levels, and transposon mobilization. We tagged each gene and performed tandem affinity purification and mass spectrometry experiments to investigate protein-protein interactions. We also describe two conditional alleles of *C. neoformans PRP43* that impact endo-siRNA levels and transposon mobilization.

Two of these factors, Rde1 (a homolog of the *S. cerevisiae* nucleolar protein Sqs1/Pfa1) and Rde2, appear to associate with one another based on mass spectrometry. Other factors detected in Rde1 purified material are factors with *S. cerevisiae* homologs that localize to the nucleolus. Additionally, we find that two different mutants of the nuclear/nucleolar helicase Prp43 display reduced siRNA production, further implicating this rRNA processing machinery in siRNA biogenesis. Given that Prp43 also disassembles stalled and post-catalytic spliceosomes and that stalled spliceosomes can serve to trigger siRNA production in *C. neoformans* (Dumesic et al., 2013), it is possible that Prp43 is released from the nucleolus in these mutants enabling it to disassemble otherwise stalled spliceosomes in the nucleoplasm, thereby inhibiting RNAi. Consistent with this view, we find that the mutant Prp43 alleles display decreased association with 18S RNA and altered protein interactions. Alternatively, the apparently increased association of Rde1 with Ago1 in the *prp43-ts5* strain may point to sequestration of Ago1, perhaps via relocalization in the nucleolus, as an alternative possible mechanism of mutant action. While further work will be required to understand the underlying mechanisms, our results point to a connection between RNAi and the nucleolus.

Rde4, a predicted terminal nucleotidyl transferase that resembles polyA polymerases and terminal-uridylyl transferases (TUTases), is also required for siRNA biogenesis. In some

other systems, TUTases have been reported to dampen the effectiveness of the RNAi pathway by inactivating small RNAs and miRNAs (Pisacane and Halic 2017). However, we do not observe any evidence for this in the *C. neoformans* system. We are unable to detect oligoA/U tails on small RNAs in our data and Rde4 promotes rather than inhibits siRNA biogenesis. Rde4 may be responsible for marking for RNAi suboptimal transcripts detected by RNA surveillance that would otherwise target them for degradation by the nuclear exosome (Lim *et al.* 2014). Additionally, the recent finding that LINE-1 elements are modified by TUT4/TUT7 uridylyltransferases to impede mobilization (Warkocki *et al.* 2018) suggests that these modifications may also be protective against proliferation of transposable elements. The discovery that TUTases are involved in silencing both DNA transposons and retrotransposons supports the proposal that regulation by uridylation occurs through multiple mechanisms and is specific to different cellular compartments (Warkocki *et al.* 2018). Moreover, the finding that both fungi and mouse cells employ this mechanism supports a model in which TUTases are import general regulators of RNA stability and function.

Finally, based on mass spectrometry, we hypothesize that Rde3, an RNase III related to *S. cerevisiae* Rnt1, and Rde5, a protein of unknown function, associate with one another at a low level and that Rde3 interacts with homologs of the RNA surveillance factors Rrp6 and Rnh1. In *S. pombe*, a RNAi system mediates formation of heterochromatin in pericentromeric regions presumably via production of dsRNA against ncRNA transcripts by an RdRP (Volpe *et al.* 2002) and subsequent production of siRNAs that direct Ago1. Formation of heterochromatin in *S. pombe* is dependent on the RNAi pathway as well as other factors including Rrp6 and the RNA polymerase II pausing and termination factor Seb1 (Reyes-Turcu *et al.* 2011; Parsa *et al.* 2018). Disruption of Seb1 does not affect siRNA abundance (Marina *et al.* 2013) and Rrp6 is not required for siRNA production and can cause a major increase in the level of siRNAs (Bühler *et al.* 2007; Chalamcharla *et al.* 2015) indicating that they function in parallel with the RNAi system to maintain heterochromatin. Additionally, transcripts that may typically undetectable by the

RNAi machinery due to high turnover can become targets of RNAi when Rrp6 is disrupted (Yamanaka *et al.* 2013).

Our findings indicate that the *C. neoformans* heterochromatin pathway is not required for RNAi as it is in *S. pombe*. In *C. neoformans*, heterochromatin likely silences transposable element transcription at centromeres which in this species limits the production of transcripts that template endo-siRNA production. This model would explain our finding that global endo-siRNAs increase rather than decrease in abundance in cells lacking H3K9me. As RNAi does not generally impact transcript levels in *C. neoformans*, it likely acts at another level such as nuclear export or translation. This two-level mechanism may enable more stringent transposon silencing. Deletion of *RRP6* does not affect the abundance of siRNAs, suggesting that RNAi and exosome-mediated surveillance act in parallel to inactivate target transcripts in *C. neoformans*, providing a third potential layer of transposon suppression.

Based on our analysis of transposition phenotypes of the five genes described here, our screen may have been biased towards identifying strong effects on transposon mobilization. Moreover, as *HAR1* is present in a non-heterochromatic region, factors selectively involved in silencing of retrotransposons, all of which lie in heterochromatic centromeric regions in *C. neoformans*, would have been missed. Screens with a broader dynamic range and those aimed at centromeric elements are thus likely to reveal additional factors that limit transposon mobilization.

Materials and Methods

Strain construction

The *ura5::HAR1:NEO* transposition screening strain was constructed using a plasmid containing the second intron of *URA5* interrupted by the *HAR1* sequence plus 1 kb of upstream flanking sequence, followed by the 3' end of *URA5* and 500 bp of downstream flanking sequence and the *C. neoformans* G418 resistance cassette by in vivo recombination in S*. cerevisiae* into the pRS416 backbone (Finnigan and Thorner 2015). Plasmids were recovered by preparing DNA from *S. cerevisiae* and electroporation into DH5α *E. coli* followed by miniprep. The plasmid was linearized by restriction digest with PmeI and SbfI (NEB) and incorporated into the genomes of CM018 and Kn99a by biolistic transformation (Chun and Madhani 2010). Incorporation was confirmed by colony PCR and Sanger sequencing. Knockouts were incorporated into the strain by mating on Murashige and Skoog medium as previously described (Xue *et al.* 2007) followed by selection on G418 and nourseothricin (NAT). Isolates with uracil prototrophy were selected against on 5-FOA before characterization of *HARBINGER* transposition.

Proteins of interest were C-terminally tagged with CBP-2xFLAG. The tag was incorporated immediately before the annotated stop codon and followed by a terminator and the G418 resistance cassette. Linearized plasmid was introduced into Kn99a by biolistic transformation and the presence and sequence of the tag was confirmed by colony PCR, Sanger sequencing and Western blotting against the FLAG epitope.

Transposition assay

Transposition of the *HARBINGER* transposon out of the *URA5* intron was assayed by selecting for growth on synthetic complete medium lacking uracil (SC-Ura). Strains were initially recovered from frozen stocks on YPAD, then selected for uracil auxotrophy by patching on 5-

FOA plates. Enough cells to achieve 0.2-0.3 OD were resuspended in YPAD liquid medium and incubated with shaking at 30°C until doubled. 1 ml of cells were then concentrated by centrifugation at 2000xg, resuspended in 0.2 ml of supernatant and spread onto -Ura plates. Colonies were counted after 3-6 days of growth at 30°C.

Agrobacterial insertional mutagenesis

Agrobacterium tumefaciens bearing a the *T-DNA* plasmid with the *C. neoformans* NAT resistance cassette (McClelland *et al.* 2005) were cultured in 120 ml AMM (0.35% potassium phosphate, 2.6 mM sodium chloride, 2 mM magnesium sulfate, 0.45 mM calcium chloride, 10 µM iron sulfate, 0.05% ammonium sulfate and 0.2% glucose) with 10 µg/ml kanamycin for at least 16 hours at 28°C with shaking. Cells were harvested by centrifugation at 4500xg for 15 min at room temperature and enough bacteria resuspended in 20 ml induction medium (40 mM 2-(N-morpholino)ethanesulfonic acid or MES, pH 5.3, 3% sucrose, 0.5% glycerol) with 10 µg/ml kanamycin and 200 µM acetosyringone to achieve an optical density of 0.15 at 600 nm. This culture was then incubated an additional 6 hours at 28°C and then harvested by centrifugation at 4500xg for 15 min. Finally, the cells were resuspended in 10 ml induction medium and adjusted to an OD_{600} of 1.25 to yield approximately 30 ml.

The *ura5::HAR1* strain was also cultured overnight in YPAD at 30°C with shaking. *C. neoformans* cells were diluted to an OD_{600} of 1.0 and grown for an additional 6 hours at 30°C with shaking. The cells were harvested by centrifugation for 5 min at 2000xg, washed twice with sterile ddH_2O and resuspended in 10 ml induction media. The volume was adjusted to achieve an OD_{600} of 5.85 with induction medium.

Equal volumes of *A. tumefaciens* and *C. neoformans* were mixed together to yield 500 µl per plate and spread onto an OSMONICS Nylon membrane on induction medium plates with 200 µM acetosyringone and 0.6% agar. Induction plates were incubated upside down (agar

down) for 72 hours at room temperature. Membranes were then transferred to YPAD plates containing 75 µg/ml carbenicillin,100 µg/ml NAT and 200 µM cefotaxime and incubated at 30°C for 48 hours. Once colonies appeared, they were replica plated using sterile velvets onto new plates composed of the same medium and incubated at 30°C for 24 hours. Following this final selection step, the colonies were replica plated onto synthetic complete medium lacking uracil and incubated at 30°C for up to 6 days, checking each day for the appearance of new colonies. The YPAD/NAT/Cefotaxime plates from the previous steps were retained to determine the background level of insertions.

Determination of insertion sites

Genomic DNA was prepared from the NATR+ and NATR+/Ura+ pools as previously described (Chun and Madhani 2010). 20 µg of genomic DNA was sonicated for 10 min (30 sec on, 1 min rest) at 4°C. DNA was extracted with phenol/chloroform/isoamyl alcohol (25:24:1, Sigma), washed with chloroform and precipitated with 3 volumes ethanol. The extent of the fragmentation was determined by separation on a 0.8% agarose gel.

Linear PCR originating in the *T-DNA*:NAT insertion was performed with Accuprime Taq High Fidelity polymerase (Thermo Fisher) using the JEBPN-Biotin2 primer (Table 2.6). First strand DNA was then purified over M280 Dynabeads (Thermo Fisher) according to the manufacturer's instructions. Purified linearized DNA was ligated to the JEBPN-DNA linker (Table 2.6) using Circligase II (Epicentre) according to the manufacturer's instructions. Finally, DNA was amplified by nested PCR with JEBPN-SA-II and JEBPN_index-SA-I (Table 2.6) with various indexes using Accuprime Taq High Fidelity polymerase (Thermo Fisher). Libraries were size selected by non-denaturing PAGE (8% Novex TBE, Thermo Fisher) and extracted from the gel by crushing and elution into 0.3 M sodium acetate overnight at 4°C. The libraries were precipitated in isopropanol and sample quality and quantity was assessed by the Agilent Bioanalyzer High Sensitivity DNA assay and qPCR. Libraries were mixed with a PhiX sample to

improve sequence diversity and sequenced on a HiSeq2500 with the JEBPN-SP3 primer (Table 2.6).

Data were pre-processed for alignment using a custom script as follows: First, reads were filtered for those beginning with at least 6 nt of the *T-DNA* sequence ("TTGTCTAAGCGTCAATTTGTTTACACCACAATATATC"). Second, the adaptor was removed from the 3' end of the reads ("GTATGCCGTCTTCTGCTTG"). Trimmed reads that were at least 18 nt long were retained for alignment to the genome with bowtie1 (additional parameters: -v2 and -m1) (Langmead *et al.* 2009). Samtools (Li *et al.* 2009) was used to convert, sort and index bam files. Bedgraph files were generated with BEDTools (Quinlan and Hall 2010) and visualized in the Integrative Genomics Viewer (IGV) (Robinson *et al.* 2011). Reads in annotated genes were counted using HTseq-count (Anders *et al.* 2015).

Targeted mutagenic screening of PRP43

A library of *PRP43* alleles was generated by mutagenic PCR (Cadwell and Joyce 2006) of the *PRP43* open reading frame. The mutagenized PRP43 amplicons were combined with the G418 resistance cassette in the pRS316 backbone by in vivo recombination in *S. cerevisiae* (Finnigan and Thorner 2015). Repaired plasmids were retrieved from *S. cerevisiae* by DNA extraction and electroporated in DH10β cells. All *E. coli* transformants were pooled and the plasmid library was prepared using a Qiagen Maxi Prep kit. The mutagenic diversity of the library as assessed by retransformation into *E. coli* followed by miniprep and Sanger sequencing. The plasmid library was then linearized with HindIII (NEB) and introduced into the *ura5*::*HAR1* screening strain by biolistic transformation.

Strains resistant to both G418 and NAT were then further screened for growth defects at 25°C, 30°C, 34°C and 37°C as well as growth on SC-Ura medium indicative of *HARBINGER* mobilization. The identified alleles were then reconstructed by amplification of the *PRP43*

coding sequence from genomic DNA and incorporation into the same plasmid construct in the manner described above. The HindIII linearized plasmid was introduced by biolistic transformation into *C. neoformans* and the allele was confirmed by colony PCR and Sanger sequencing.

RNA preparation and siRNA Northern blotting

RNA samples were prepared as previously described (Dumesic *et al.* 2013) from log phase YPAD cultures. 30 µg of total RNA were desiccated, re-dissolved in formamide loading dye and separated on a 15% TBE-Urea gel (Novagen) in 1X TBE at 180 V for 80 min. RNA was transferred to Hybond-NX membrane (Amersham) 1X TBE in a Invitrogen Xcell II blot module for 90 min at 20 V. RNA was crosslinked to the membrane in 0.16 M N-(3-Dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride (Sigma) in 0.13 M 1-methylimidazole (Sigma), pH 8 for 1 hour at 60°C. The membrane was equilibrated in 10 ml Roche Easy Hyb solution at 25°C. Riboprobe against *CNAG_06705* antisense RNA was prepared using the Roche Dig Northern Starter Kit according to the manufacturer's instructions. 10 µl of the riboprobe was hydrolyzed in 120 µl 0.1 M sodium carbonate + 180 µl 0.1M sodium bicarbonate for 30 min at 60°C then added directly to the Easy Hyb solution. 10 ng/ml of Dig-anti-U6 probe (Table 2.6) was also added to the hybridization mixture. The probe was hybridized to the membrane overnight at 25°C. The membrane was then washed twice with 6X SSC, 0.1% SDS twice at 37°C for 10 min then once with 2X SSC, 0.1% SDS at 25°C for 10 min. Detection of digoxin on the membrane was performed using the Roche Dig Northern Starter Kit according to the manufacturer's instructions. Chemiluminescence was detected using an Azure c600.

RNA-seq and siRNA-seq library preparation

For RNA-seq (all except *rrp6Δ*), 2.5 µg of RNA were treated with DNase as previously described (Zhang *et al.* 2012). 3' end RNA sequencing libraries were prepared with the Lexogen

QuantSeq 3' mRNA-Seq Library Prep Kit FWD. For RNA from the *rrp6Δ* strain and a corresponding wild type sample, 50 µg of RNA was first selected using the Qiagen Oligotex mRNA mini kit following the manufacturer's instructions and DNase treated as for QuantSeq samples. Libraries were then prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina following the manufacturer's instructions. For small RNA-seq, 20 µg of RNA were treated with DNase as previously described (Zhang *et al.* 2012). 3' end RNA sequencing libraries were prepared with the Lexogen Small RNA-Seq Library Prep Kit.

Analysis of RNA-seq data

Reads were prepared for alignment by trimming adaptor sequences with Cutadapt (A_{10} for QuantSeq and "TGGAATTCTC" for small RNA-seq) (Martin 2011). Trimmed reads were aligned to the *C. neoformans* genome with either STAR for QuantSeq data (Dobin *et al.* 2013) or Bowtie for small RNA-seq data (Langmead *et al.* 2009). Split reads were ignored when aligning with STAR (--alignIntronMax 1). For small RNA-seq, 2 mismatches were allowed (-v2) and reads aligning to more than one locus were randomly assigned (-M1 --best). Reads aligning to genes and transposable elements were counted using custom scripts using the Python library Pysam and differential expression of mRNAs and siRNAs was determined using DESeq2 (Love *et al.* 2014). The location of transposable elements was determined based on homology with the consensus sequences determined from *C. neoformans var neoformans* (Janbon *et al.* 2010) with custom scripts using BLAST (McGinnis and Madden 2004).

Mass spectrometry

Tandem immunoprecipitation using C-terminal 2xFLAG and calmodulin binding peptide epitopes was performed as previously described (Dumesic *et al.* 2013) from 2 L of *C. neoformans* cultured in YPAD to OD 2.

Reagent and data availability

Strains are available upon request and are listed in Table 2.6. Mass spectrometry and processed sequencing data are available as supplemental tables (Tables 2.1-2.5). Raw and processed sequencing data are available from GEO with accession code GSE128009.

Acknowledgments

We would like to thank all members of the Madhani lab for helpful discussions. We thank Dr. Sandra Catania for construction of the CM1926 strain and Nguyen Nguyen for technical support. We also thank Eric Chow and the UCSF Center for Advanced Technology for assistance with sequencing and sample analysis. We thank anonymous reviewers for critical comments on the manuscript. Supported by NIH grants R01 GM71801 to H.D.M., P41 GM103533 to J.R., and R01 GM120507 to J.J.L. J.E.B was supported by postdoctoral fellowship 127531-PF-15-050-01-R from the American Cancer Society. H.D.M. is a Chan-Zuckerberg Biohub Investigator.

Figures

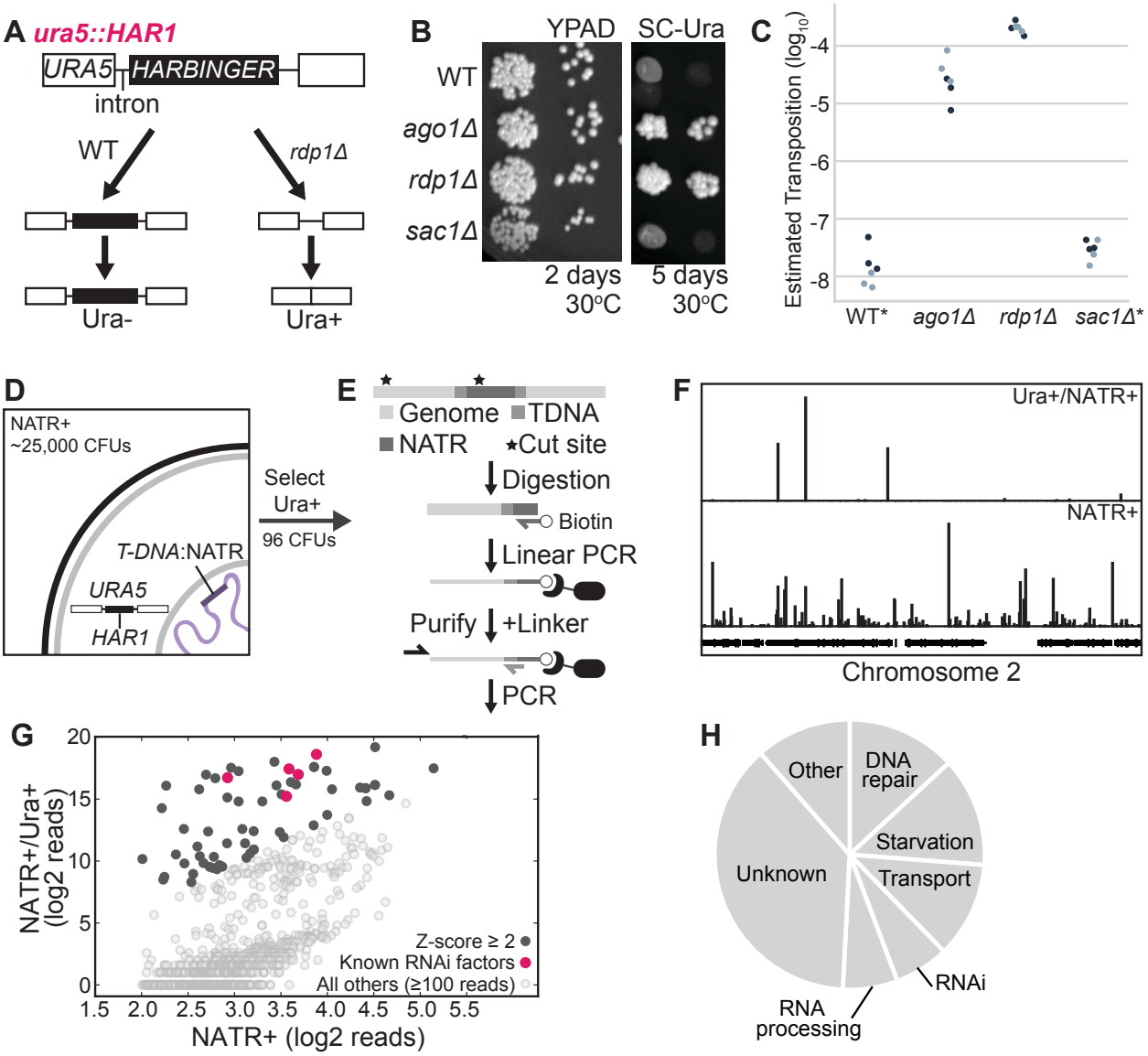


Figure 2.1

A. A copy of the HARBINGER transposon (HAR1) was inserted into the second intron of URA5, resulting in failure to splice and uracil auxotrophy when the RNAi pathway is functional. Upon loss of RNAi (rdp1Δ), transposition may occur and some cells are able to synthesize uracil. B. Mobilization of HARBINGER in the presence of RNAi pathway knockouts. 1/10 dilutions of log phase cultures were spotted on YPAD or SC-Ura and incubated at 30°C for 2 or 5 days. C.

Quantitation of HARBINGER mobilization. CFUs were counted after 2 days (YPAD) or 6 days (SC-Ura) at 30°C. Ura+ CFUs were then normalized to YPAD CFUs. *No colonies in WT and sac1 genotypes were detected on SC-Ura by the end 6 days, so a maximum estimated transposition rate is indicated. D. Schematic of the insertional mutagenesis strategy used for screening. Cells co-cultured with Agrobacteria carrying a T-DNA:NATR transposable element were selected for resistance to NAT, then for the ability to grow on media lacking uracil. E. Sequencing strategy for identifying T-DNA insertions. C. neoformans genomic DNA was fragmented by sonication and then ssDNA against the insertion site was generated by linear amplification. The biotinylated ssDNA product was purified, a DNA linker of known sequence was added to the 5' end and the genomic flank was amplified by nested PCR. F. Mapping of insertion sites to the C. neoformans genome in the NAT resistant uracil prototrophic pool (NATR+/Ura+) versus NAT resistant pool (NATR+). G. Quantitative comparison of the number of reads spanning the T-DNA boundaries for insertions within annotated genes for NATR+ and NATR+/Ura+ pools. Z-scores were determined from the distribution of the log2 ratio of reads from the NATR+/Ura+ pool over reads from the NATR+ pool. H. Functional classification of genes with enriched insertions in the HARBINGER mobilization screen based on FungiDB and hand-curated annotations.

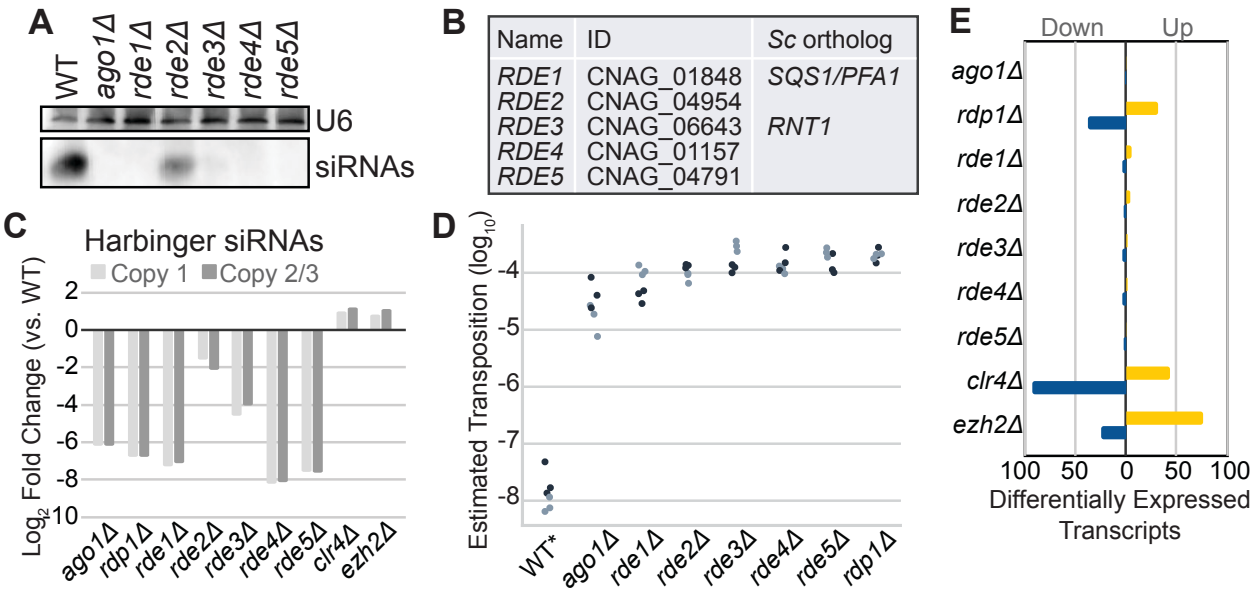


Figure 2.2

A. siRNA Northern analysis of screen hits. Shown are those with a qualitatively apparent loss of siRNAs. Total RNA from each strain was separated by denaturing PAGE, transferred, chemically crosslinked and probed for siRNAs against CNAG_6705 with a Dig-labeled riboprobe. The loading control, U6 snRNA, was detected with a Dig-labeled DNA oligo against the C. neoformans U6 sequence. B. Names, gene identifiers and orthologs from S. cerevisiae (Sc) as determined by PSI blast (Altschul et al. 1997). C. Quantitation of the fold change in siRNAs against each HARBINGER locus (Copy 1: CNAG_00903, Copy 2/3: CNAG_02711 and CNAG_00549) as determined by small RNA-seq. Two copies of HARBINGER are virtually identical (Copy 2/3) and thus reads map to both with roughly equivalent frequency in our alignment strategy (see methods). D. Estimated transposition rate of HARBINGER in knockouts of each of the newly discovered factors. Assay conducted as described in Figure 2.1 legend. E. Differential expression of mRNA in C. neoformans (QuantSeq). Transcripts were determined to be significantly differentially expressed if they exhibit at least a 2-fold increase (yellow) or decrease (blue) in expression with an adjusted p-value of at most 0.01 (as determined by DESeq2, 2 replicates).

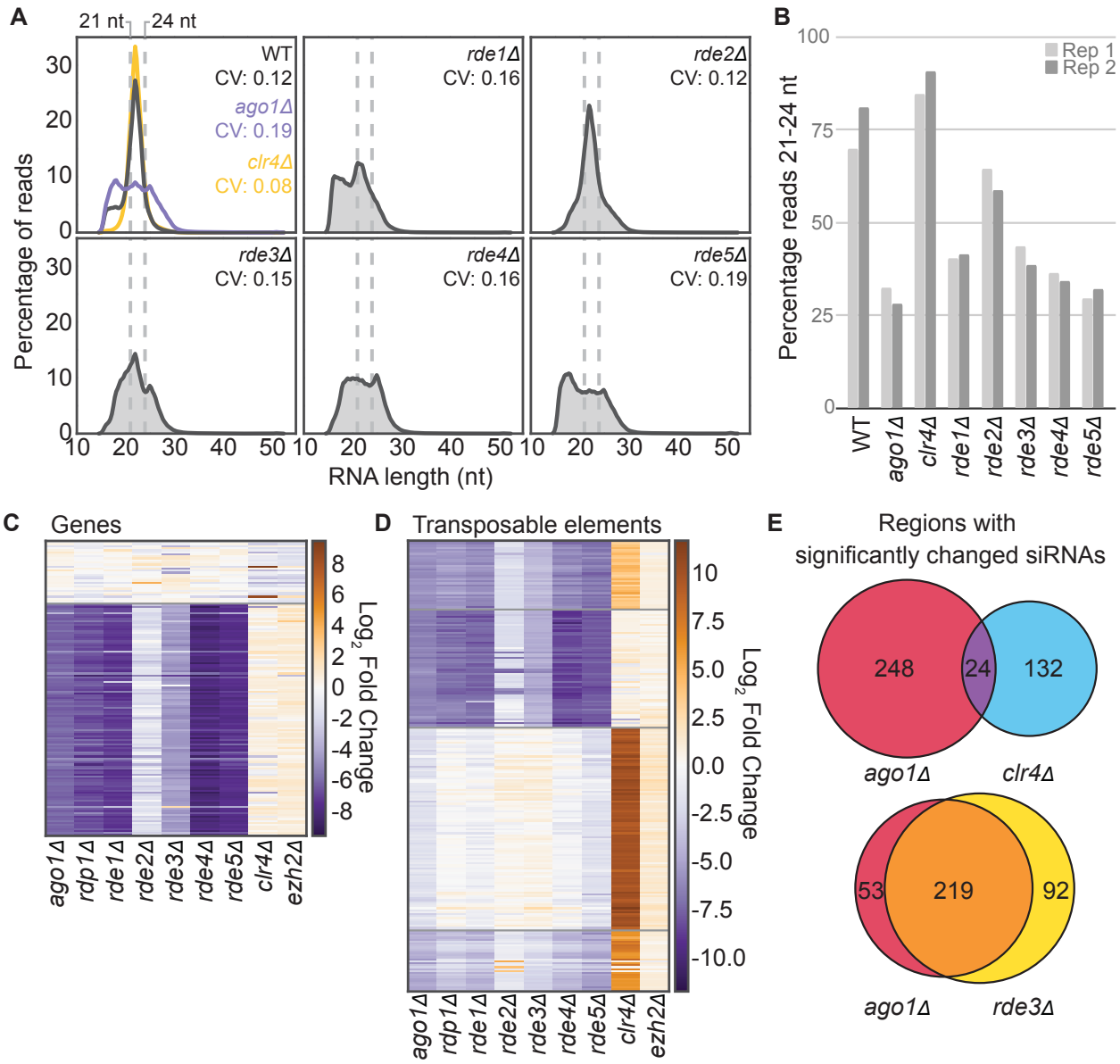


Figure 2.3

A. Distribution of small RNA-seq read sizes and coefficient of variation (CV) in each RDE knockout as well as ago1Δ and a partial deletion of clr4Δ (single replicate shown). B. Percentage of small RNA-seq reads between 21-24 nt in length for each knockout compared to wild type. C. Fold change of siRNA abundance for annotated genes with at least 100 small RNA-seq reads in both of the wild-type libraries (determined by DESeq2, 2 replicates). Order determined by K-means clustering. D. Fold change of siRNA abundance (determined by DESeq2, 2 replicates) against transposable elements and transposable element remnants (see methods for annotation strategy) compared to wild type. E. Overlap of transposable elements and genes with significantly changed siRNA populations.

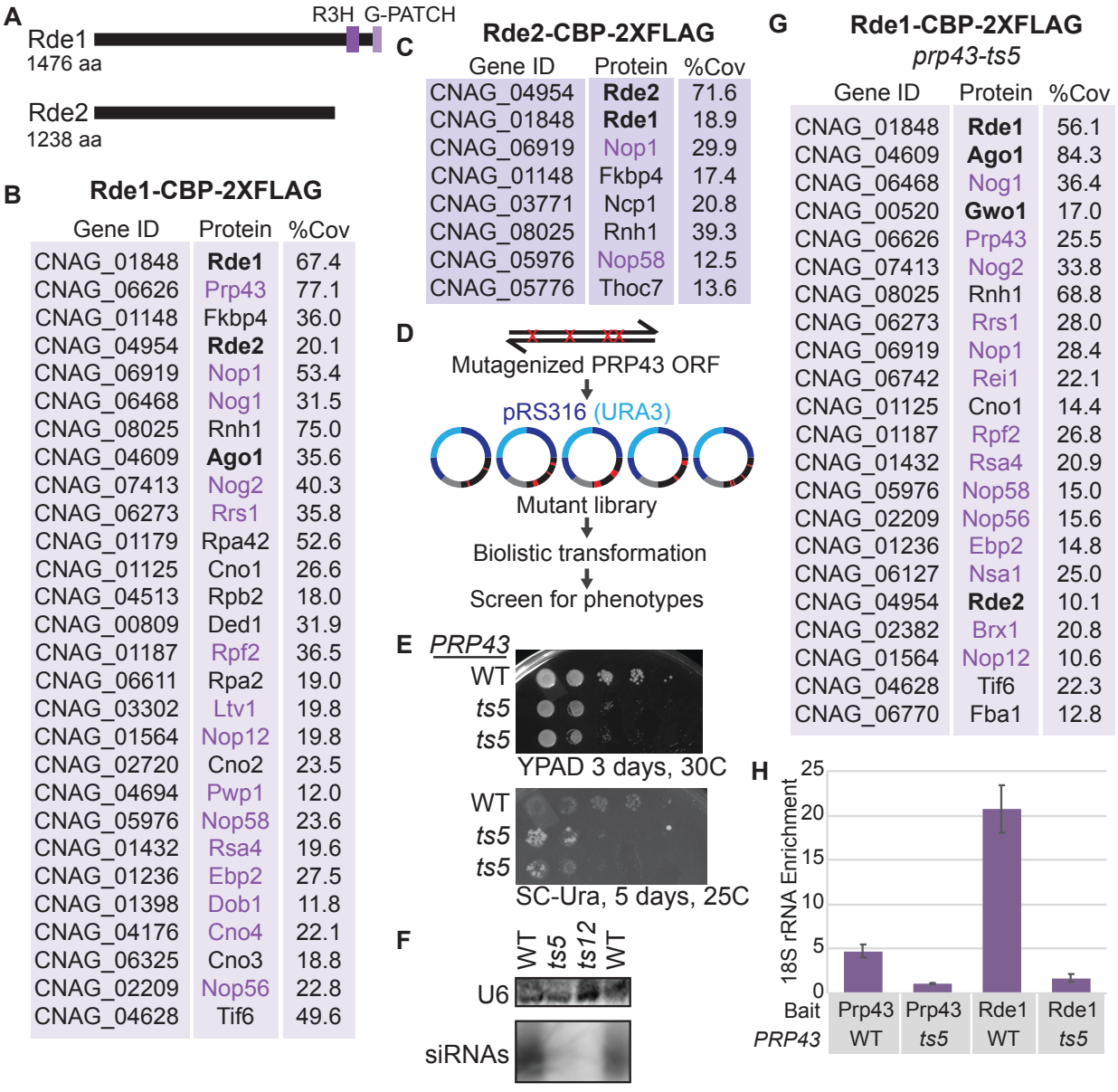


Figure 2.4

A. Predicted domain structures of Rde1 and Rde2 determined by PSI-BLAST with an e-value of at most 1X10-5. B. Proteins detected in Rde1-CBP-2xFLAG purified material as determined by tandem IP-MS, named based on current C. neoformans annotation (FungiDB) or S. cerevisiae homolog. Three factors had no clear homolog, so we refer to them as putative C. neoformans NucleOlar protein 1-3 (Cno1, Cno2 and Cno3). Proteins in bold were identified in the screen. Proteins in purple are predicted nucleolar proteins, typically involved in rRNA processing and

ribosome biogenesis. Percent coverage is the average between two replicates. Common contaminants and proteins with less than 10% coverage are excluded and proteins are in order of ascending total spectral counts. C. Proteins detected in Rde2-CBP-2xFLAG purified material (see B). D. Mutagenic library strategy for PRP43. E. Growth phenotype of wild type and prp43-ts5 C. neoformans strains bearing the ura5::HAR1 insertion on rich media and media lacking uracil. F. Loss of siRNAs against CNAG_06705 as determined by Northern analysis (see Fig. 2A). G. RT-qPCR of 18S rRNA associated with Rde1 (native FLAG affinity purification). Data are from three technical replicates and two biological replicates.

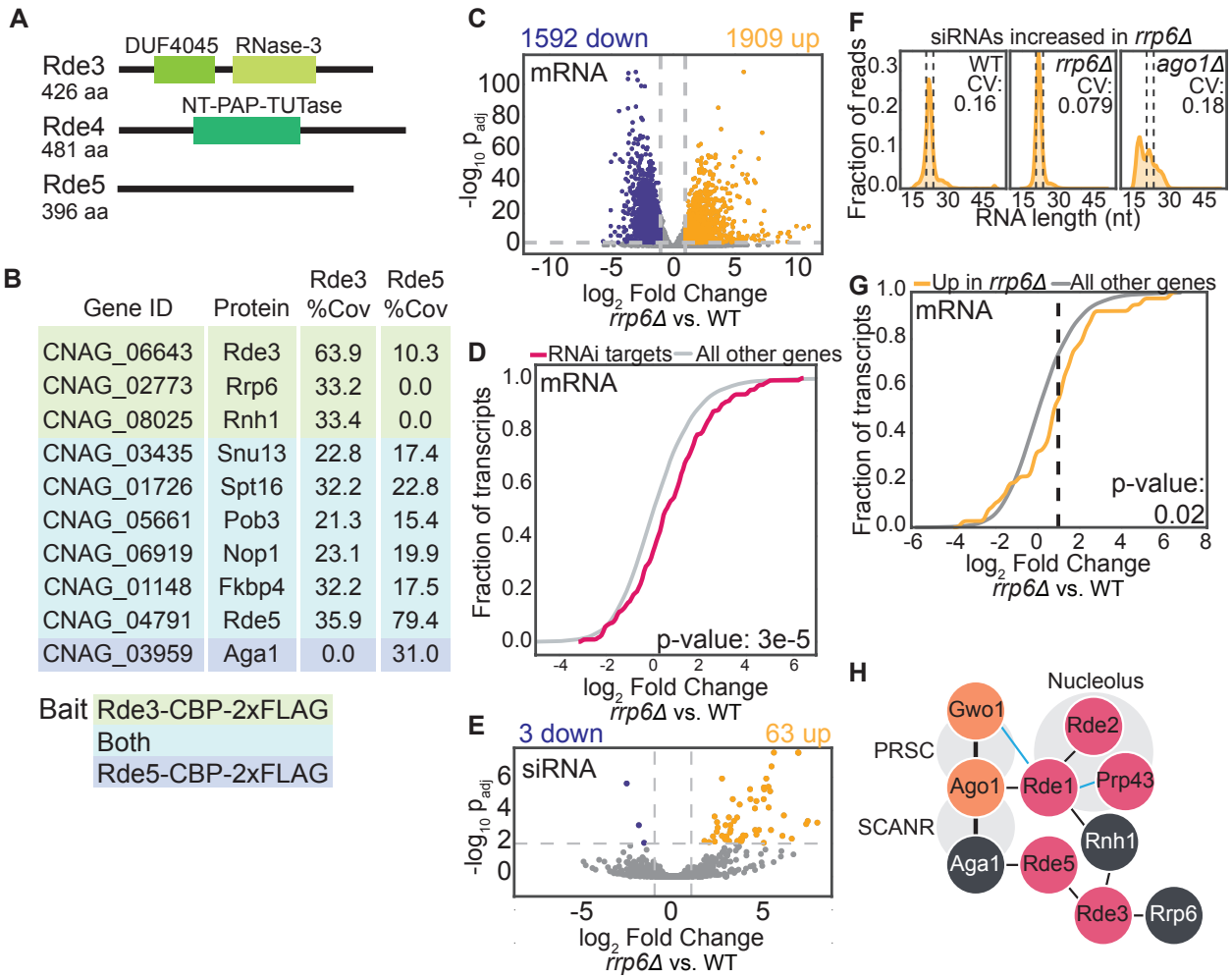


Figure 2.5

A. Predicted domain structures of Rde3, Rde4 and Rde5 determined by PSI-BLAST (Altschul et al. 1997) with an e-value of at most 1X10-5. B. Proteins detected in Rde3-CBP-2xFLAG and Rde5-CBP-2xFLAG purified material as determined by tandem affinity purification and mass spectrometry. Proteins highlighted in green were found with Rde3; those highlighted in teal were detected with both Rde3 and Rde5 and Aga1 (in blue) was only detected with Rde5. Percent coverage is the average between two replicates. Common contaminants and proteins with less than 10% coverage are excluded and proteins are in order of ascending total spectral counts. C. Differential expression of polyA mRNA in the rrp6Δ strain. Colored points indicate transcripts with at least a 2-fold increase (yellow) or decrease (blue) in expression with an adjusted p-value of at most 0.01 (DE-Seq2). D. Fold change of transcripts that template endo-siRNA production (Dumesic et al. 2013) compared with the rest of the transcriptome. P-value determined by Mann-Whitney U test. E. Differential abundance of small RNAs in the rrp6Δ strain. See panel C for color-code. F. Small RNA size distribution and CV of the population significantly increased in the rrp6Δ strain. Dashed lines indicate 21-24 nt region. G. Fold change of transcripts targeted by siRNAs that are differentially increased in rrp6Δ. P-value determined by Mann-Whitney U test. H. Physical associations of new RNAi factors described in this study (pink) with known RNAi

factors (orange) and other RNA processing factors (white). Thin lines are interactions from IP-MS data only. Thick lines indicate interactions confirmed by co-IP and/or yeast 2-hybrid (Dumesic et al. 2013). Blue lines indicate interactions that are affected by the prp43-ts5 allele.

Tables

Table 2.1

Counts of reads spanning T-DNA insertion sites for all insertions within genes with Z-score analysis. Related to Figure 2.1.

DOI https://doi.org/10.7272/Q6T72FP1

Table 2.2

Transposition assay CFU counts and estimated transposition rates for wild type and mutant *C. neoformans ura5::HAR1* strains. Related to Figure 2.2.

DOI https://doi.org/10.7272/Q6T72FP1

Table 2.3

Read counts from QuantSeq, expression changes and significance as determined by DESeq2. Related to Figure 2.2.

DOI https://doi.org/10.7272/Q6T72FP1

Table 2.4

Read counts from siRNA sequencing, expression changes and significance as determined by DESeq2 and read size distributions. Related to Figure 2.3.

DOI https://doi.org/10.7272/Q6T72FP1

Table 2.5

Mass spectrometry results from Rde1-5 CBP-2xFLAG affinity purifications. Related to Figure 2.4-2.5.

DOI https://doi.org/10.7272/Q6T72FP1

Table 2.6

DNA oligomers and *C. neoformans* strains used in this study.

Name	Sequence
JEBPN-Biotin2	5’-Biosg/GAAGGGCAATCAGCTGTTGC
JEBPN-DNA-linker	5’Phos/ TCGTATGCCGTCTTCTGCTTGACTCAGTAGTTGTGCGATGGATTGATG/ddC-3’
JEBPN-SA-II	5’-CAAGCAGAAGACGGCATACGA-3’

Name	Sequence
JEBPN-index-SA-I	5'-AATGATACGGCGACCACCGAGATCT GATCGGAAGAGCACACGTCTGAACTCCAGTCAC TCAAGT ACGCTCGTCCGATCT CGTCCGCAATGTGTTATTAAG-3'
JEBPN-SP3	ACGCTCGTCCGATCT CGTCCGCAATGTGTTATTAAG
Dig-anti-U6	5’-/5DigN/TCCTCTCTGCTCGAGTTTGTC-3’
18S rRNA qPCR forward	GTCCAGACATAGTGAGGATTGACAG
18S rRNA qPCR reverse	GACAGTCCCTCTAAGAAGTCATACG

References

1. Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang et al., 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–402.

2. Anders, S., P. T. Pyl, and W. Huber, 2015 HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–9.

3. Billmyre, R. B., S. Calo, M. Feretzaki, X. Wang, and J. Heitman, 2013 RNAi function, diversity, and loss in the fungal kingdom. Chromosome Res. 21: 561–72.

4. Bühler, M., W. Haas, S. P. Gygi, and D. Moazed, 2007 RNAi-Dependent and -Independent RNA Turnover Mechanisms Contribute to Heterochromatic Gene Silencing. Cell 129: 707–721.

5. Cadwell, R. C., and G. F. Joyce, 2006 Mutagenic PCR. CSH Protoc. 2006: pdb.prot4143.

6. Chalamcharla, V. R., H. D. Folco, J. Dhakshnamoorthy, and S. I. S. Grewal, 2015 Conserved factor Dhp1/Rat1/Xrn2 triggers premature transcription termination and nucleates heterochromatin to promote gene silencing. Proc. Natl. Acad. Sci. U. S. A. 112: 15548–55.

7. Chun, C. D., and H. D. Madhani, 2010 Applying genetics and molecular biology to the study of the human pathogen cryptococcus neoformans. Methods Enzymol. 470: 797–831.

8. Chuong, E. B., N. C. Elde, and C. Feschotte, 2017 Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18: 71–86

9. Claycomb, J. M., 2014 Ancient Endo-siRNA Pathways Reveal New Tricks. Curr. Biol. 24: R703–R715.

10. Cogoni, C., and G. Macino, 1999 Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. Nature 399: 166–169.

11. Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski et al., 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

12. Dumesic, P. A., C. M. Homer, J. J. Moresco, L. R. Pack, E. K. Shanle et al., 2015 Product Binding Enforces the Genomic Specificity of a Yeast Polycomb Repressive Complex. Cell 160: 204–218.

13. Dumesic, P. a, P. Natarajan, C. Chen, I. a Drinnenberg, B. J. Schiller et al., 2013 Stalled spliceosomes are a signal for RNAi-mediated genome defense. Cell 152: 957–68.

14. Finnigan, G. C., and J. Thorner, 2015 Complex in vivo Ligation Using Homologous Recombination and High-efficiency Plasmid Rescue from Saccharomyces cerevisiae. Bio-protocol 5: e1521.

15. Janbon, G., S. Maeng, D.-H. Yang, Y.-J. Ko, K.-W. Jung et al., 2010 Characterizing the role of RNA silencing components in Cryptococcus neoformans. Fungal Genet. Biol. 47: 1070–80.

16. Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

17. Lebaron, S., C. Papin, R. Capeyrou, Y.-L. Chen, C. Froment et al., 2009 The ATPase and helicase activities of Prp43p are stimulated by the G-patch protein Pfa1p during yeast ribosome biogenesis. EMBO J. 28: 3808–3819.

18. Lee, H.-C., A. P. Aalto, Q. Yang, S.-S. Chang, G. Huang et al., 2010 The DNA/RNA-Dependent RNA Polymerase QDE-1 Generates Aberrant RNA and dsRNA for RNAi in a

Process Requiring Replication Protein A and a DNA Helicase (M. Egli, Ed.). PLoS Biol. 8: e1000496.

19. Lejeune, E., M. Bortfeld, S. A. White, A. L. Pidoux, K. Ekwall et al., 2007 The Chromatin-Remodeling Factor FACT Contributes to Centromeric Heterochromatin Independently of RNAi. Curr. Biol. 17: 1219–1224.

20. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

21. Lim, J., M. Ha, H. Chang, S. C. Kwon, D. K. Simanshu et al., 2014 Uridylation by TUT4 and TUT7 Marks mRNA for Degradation. Cell 159: 1365–1376.

22. Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

23. Marina, D. B., S. Shankar, P. Natarajan, K. J. Finn, and H. D. Madhani, 2013 A conserved ncRNA-binding protein recruits silencing factors to heterochromatin through an RNAi-independent mechanism. Genes Dev. 27: 1851–1856.

24. Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12.

25. McClelland, C. M., Y. C. Chang, and K. J. Kwon-Chung, 2005 High frequency transformation of Cryptococcus neoformans and Cryptococcus gattii by Agrobacterium tumefaciens. Fungal Genet. Biol. 42: 904–913.

26. McGinnis, S., and T. L. Madden, 2004 BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32: W20–W25.

27. Muszewska, A., K. Steczkiewicz, M. Stepniewska-Dziubinska, and K. Ginalski, 2017 Cut-and-Paste Transposons in Fungi with Diverse Lifestyles. Genome Biol. Evol. 9: 3463–77.

28. Nekrutenko, A., and W. H. Li, 2001 Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 17: 619–21.

29. Pandit, S., S. Paul, L. Zhang, M. Chen, N. Durbin et al., 2009 Spp382p interacts with multiple yeast splicing factors, including possible regulators of Prp43 DExD/H-Box protein function. Genetics 183: 195–206.

30. Parsa, J.-Y., S. Boudoukha, J. Burke, C. Homer, and H. D. Madhani, 2018 Polymerase pausing induced by sequence-specific RNA-binding protein drives heterochromatin assembly. Genes Dev. 32: 953–964.

31. Pertschy, B., C. Schneider, M. Gnädig, T. Schäfer, D. Tollervey et al., 2009 RNA helicase Prp43 and its co-factor Pfa1 promote 20 to 18 S rRNA processing catalyzed by the endonuclease Nob1. J. Biol. Chem. 284: 35079–91.

32. Piovesan, D., F. Tabaro, L. Paladin, M. Necci, I. Mičetić et al., 2018 MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res. 46: D471–D476.

33. Pisacane, P., and M. Halic, 2017 Tailing and degradation of Argonaute-bound small RNAs protect the genome from uncontrolled RNAi. Nat. Commun. 8: 15332.

34. Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

35. Reyes-Turcu, F. E., K. Zhang, M. Zofall, E. Chen, and S. I. S. Grewal, 2011 Defects in RNA quality control factors reveal RNAi-independent nucleation of heterochromatin. Nat. Struct. Mol. Biol. 18: 1132–8.

36. Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander et al., 2011 Integrative genomics viewer. Nat. Biotechnol. 29: 24–26.

37. Schmidt, M., K. Schwarzwaelder, C. Bartholomae, K. Zaoui, C. Ball et al., 2007 High-resolution insertion-site analysis by linear amplification–mediated PCR (LAM-PCR). Nat. Methods 4: 1051–1057.

38. Sijen, T., and R. H. A. Plasterk, 2003 Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature 426: 310–314.

39. Slotkin, R. K., M. Freeling, and D. Lisch, 2005 Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nat. Genet. 37: 641–644.

40. Volpe, T. A., C. Kidner, I. M. Hall, G. Teng, S. I. S. Grewal et al., 2002 Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 297: 1833 – 7.

41. Wang, X., Y.-P. Hsueh, W. Li, A. Floyd, R. Skalsky et al., 2010 Sex-induced silencing defends the genome of Cryptococcus neoformans via RNAi. Genes Dev. 24: 2566–82.

42. Warkocki, Z., P. S. Krawczyk, D. Adamska, K. Bijata, J. L. Garcia-Perez et al., 2018 Uridylation by TUT4/7 Restricts Retrotransposition of Human LINE-1s. Cell 174: 1537–1548.e29.

43. Xue, C., Y. Tada, X. Dong, and J. Heitman, 2007 The Human Fungal Pathogen Cryptococcus Can Complete Its Sexual Cycle during a Pathogenic Association with Plants. Cell Host Microbe 1: 263–273.

44. Yamanaka, S., S. Mehta, F. E. Reyes-Turcu, F. Zhuang, R. T. Fuchs et al., 2013 RNAi triggered by specialized machinery silences developmental genes and retrotransposons. Nature 493: 557–60.

45. Zhang, Z., W. E. Theurkauf, Z. Weng, and P. D. Zamore, 2012 Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. Silence 3: 9.

Chapter 3

Spliceosome profiling visualizes the operations of a dynamic RNP in vivo at nucleotide resolution

Jordan E. Burke 1, Adam D. Longhurst 1, Daria Merkurjev 2, Jade Sales-Lee 1, Beiduo Rao 1, James Moresco 3, John Yates III 3, Jingyi Jessica Li 2, Hiten D. Madhani 1,4,5

1 Dept. of Biochemistry and Biophysics, University of California, San Francisco, CA 94158

2 Dept. of Statistics, University of California, Los Angeles, CA

3 Dept. of Molecular Medicine The Scripps Research Institute La Jolla, CA

4 Chan-Zuckerberg Biohub San Francisco, CA 94158

5 Lead Contact hitenmadhani@gmail.com

Abstract

Tools to understand how the spliceosome functions *in vivo* have lagged behind advances in its structural biology. We describe methods to globally profile spliceosome-bound precursor, intermediates and products at nucleotide resolution. We apply these tools to three divergent yeast species that span 600 million years of evolution. The sensitivity of the approach enables detection of novel cases of non-canonical catalysis including interrupted, recursive and nested splicing. Employing statistical modeling to understand the quantitative relationships between RNA features and the data, we uncover independent roles for intron size, position and number in substrate progression through the two catalytic stages. These include species-specific inputs suggestive of spliceosome-transcriptome coevolution. Further investigations reveal ATP-dependent discard of numerous endogenous substrates at both the precursor and lariat-intermediate stages and connect discard to intron retention, a form of splicing regulation. Spliceosome profiling is a quantitative, generalizable global technology to investigate an RNP central to eukaryotic gene expression.

Introduction

A signature feature of eukaryotic gene expression is the splicing of primary transcripts by the spliceosome. Rather than serving as a mere impediment to gene expression, introns are central to gene regulation at many levels. In addition to controlling proteome diversity (Blencowe, 2017), splicing also impacts RNA stability, for example through the inclusion of exons with premature termination codons subject to nonsense-mediated decay (NMD) (Zhou et al., 2000; Luo et al., 2001; Wiegand et al., 2003; Maquat, 2004; Amrani et al., 2006; Brogna and Wen, 2009). Splicing also plays critical roles in RNA export and translation efficiency (Zhou et al., 2000; Luo et al., 2001; Wiegand et al., 2003; Maquat, 2004). Biogenesis of noncoding RNA often requires splicing, including many snoRNAs and miRNAs that are encoded within introns rather than exons (Hesselberth, 2013). Finally, splicing is a major player in human disease; a large fraction of single nucleotide polymorphisms associated with human disease impact splicing (Li et al., 2016), and many human cancers harbor driver mutations in components of the spliceosome itself (Bejar, 2016).

Our understanding of the assembly dynamics of the spliceosome comes almost entirely from decades of studies of model intron substrates in splicing-competent cell extracts from *S. cerevisiae* and HeLa cells (Wahl et al., 2009), and has been substantially further illuminated by recent high-resolution structural work (Fica and Nagai, 2017; Shi, 2017). Activation of the spliceosome after its assembly involves major ATP-dependent rearrangements that expel two RNA-protein complexes, the U1 and U4 snRNPs, and several proteins (Wahl et al., 2009). At the same time, the Prp19 complex (NTC) and NTC related proteins (NTR) join the complex to produce the B^{act} complex. These events trigger assembly of the RNA catalytic core of the spliceosome. Splicing then proceeds via two transesterification reactions. In step 1, nucleophilic attack of the 2' hydroxyl of an adenosine at the branchpoint (BP) on the phosphate at the 5' splice site (ss) produces a cleaved 5' exon and branched species harboring a 2'-5'

phosphodiester bond called the lariat intermediate (Figure 3.1A). The second transesterification is achieved through attack of the 3' hydroxyl of the cleaved 5' exon on the phosphate at the 3'ss, producing the ligated exons and excising the intron as a branched RNA, referred to as the lariat intron. A handful of additional factors are required to bind and then leave the spliceosome during these two steps. These include the DEXD-box ATPases Prp16 and Prp22, which are thought to remodel/inactivate the active site after the first and second chemical steps, respectively, enabling intermediates and products to proceed to the next step in mRNA biogenesis (Schwer, 2008; Tseng et al., 2011; Krishnan et al., 2013; Ohrt et al., 2013; Semlow et al., 2016). Following mRNA release, the excised lariat intron and associated snRNPs are dissociated by the helicase Prp43 (Martin et al., 2002; Tsai et al., 2005). The 2'-5' phosphodiester bond at the BP of the lariat intron is hydrolyzed by the debranching enzyme, Dbr1 (Ruskin and Green, 1985; Chapman and Boeke, 1991).

The spliceosome is thus both remarkably dynamic and complex. These properties are thought to promote the fidelity of splicing through kinetic proofreading schemes, while permitting flexibility and regulation. Studies using *S. cerevisiae* splicing-competent extracts have shown that mutant pre-mRNA substrates that assemble into spliceosomes but are kinetically slow at either chemical step trigger spliceosome disassembly prior to completion of the reaction, a process termed "discard." Together with prior suppressor genetics and associated biochemical work, this work has led to a model for spliceosomal fidelity in which the failure to perform catalysis prior to ATP hydrolysis by Prp16 (step 1) or Prp22 (step 2) produces a spliceosome "marked" for discard, which signals disassembly by Prp43 (Mayas et al., 2006; Koodathingal et al., 2010; Mayas et al., 2010; Semlow et al., 2016).

While these studies have been important for our understanding of the mechanism of fidelity, they are based on studies of mutant RNAs in a single species, *S. cerevisiae*, that harbors few introns and these display unusually strong matches to 5'ss, BP and 3'ss consensus

signals (Kupfer et al., 2004; Irimia and Roy, 2008). The match of each signal to consensus positively impacts progression of an intron through the steps of splicing (Lesser and Guthrie, 1993). Fungi from other subphyla, such as *S. pombe* and *C. neoformans*, are considerably more intron-rich and display a much broader range of splicing signals, akin to those of plants and animals (Kupfer et al., 2004; Irimia and Roy, 2008; Suzuki et al., 2013; Duff et al., 2015; Sibley et al., 2015). Because of these considerations, it remains unclear to what extent inefficiency triggers discard in native introns *in vivo* and whether discard contributes significantly to the regulation of the RNA population as a whole, particularly in intron populations with diverse sequences at the 5'ss and BP regions.

RNA-seq, the high-throughput sequencing of cDNAs from fragmented mRNA, has enabled genome-wide identification of splicing events *in vivo.* Although the sensitivity of RNA-seq is limited by RNA stability, high-depth experiments can identify rare sequencing reads corresponding to splicing precursors and products. Even non-canonical splicing events such as recursive splicing involving exons of size zero have been identified in mammals and Drosophila (Suzuki et al., 2013; Duff et al., 2015; Sibley et al., 2015). Mutation of factors involved in nuclear or cytoplasmic RNA turnover or surveillance mechanisms can enable detection of splicing events that produce highly unstable RNAs, but these mutations are confounding as they necessarily alter the RNA population. Likewise, excised lariats can be stabilized in cells containing a knockout of *DBR1* (Awan et al., 2013; Gould et al., 2016; Mayerle et al., 2017), but this approach has the same drawback. Sequencing of chromatin-associated RNA has been used to enrich for shorter-lived products of splicing (Bhatt et al., 2012; Mayer et al., 2015), but this method requires such products to co-fractionate with the genome. Lariat-introns have been enriched by immunoprecipitation of *S. cerevisiae* extracts with polyclonal antibodies against Prp16 (Qin et al., 2016); this qualitative method enriches for a subpopulation of spliceosomes transiently bound by this ATPase. Despite these advances, none of these approaches directly

interrogates the full population of transcripts bound to active spliceosomes, nor can they be used to quantify the progression of substrates through the spliceosome cycle.

We describe here a strategy for purification of the full complement of endogenous active spliceosomes using a cross-organism tagging and purification strategy followed by high throughput sequencing library preparation schemes to identify total spliceosome bound RNA, step 1 and step 2 cleavage sites as well as methods to quantify relative precursor and intermediate levels. We apply it to three highly divergent yeasts, the brewing yeasts *S. cerevisiae* and *S. pombe* and the human pathogenic yeast *C. neoformans*, which diverged from a common ancestor approximately 600 million years ago (Lucking et al., 2009; Prieto and Wedin, 2013). Below we describe these methods, which we term the spliceosome profiling suite, their utility for investigating the spliceosome *in vivo*, and the insights gained.

Results

Prp19 is a species-independent handle for isolating active spliceosomes

The spliceosome is not a single entity, but rather a series of dynamic complexes (Figure 3.1A). Based on *in vitro* assembly studies, the NTC component Prp19 joins the spliceosome during the process of activation and is associated primarily with *S. cerevisiae* spliceosomes that are activated and bound to mRNA precursor (B^{act}), intermediate (C), or product (ILS) (Figure 3.1A) (Fabrizio et al., 2009). Affinity purification of Prp19-TAP from *S. pombe* also yields components diagnostic of these complexes (Ren et al., 2011). To extend these studies to *C. neoformans*, we affinity-purified Prp19-FLAG associated spliceosomes from cell extracts prepared at a range of salt (KCl) concentrations (Figure 3.8A) and performed tandem mass tag mass spectrometry (TMT-MS) to quantify the purified proteins (Figure 3.1B). We find that while orthologs of most annotated splicing factors copurify with Prp19 in lysates prepared in buffers containing up to 500 mM KCl (Figure 3.1B and Table 3.1), 55 splicing components exhibit quantitatively decreased abundance (based on reporter ion intensities) at KCl concentrations at and above 300 mM. Because copurification of common contaminants is also decreased at KCl concentrations above 300 mM KCl (Figure 3.1B, compare lane 6 to lane 2 and Table 3.1), we chose the 300 mM KCl condition for *C. neoformans*. For *S. cerevisiae* and *S. pombe*, we found enrichment comparable to that observed using *C. neoformans* could not be achieved unless the KCl concentration was lowered to 200 mM KCl (data not shown).

Of splicing factor orthologs that co-purify with Prp19 in *C. neoformans,* 49/72 are orthologs of proteins that are part of the B^{act}, C and ILS complexes of *S. cerevisiae* (Figure 3.1C and Table 3.1), indicating that the Prp19 purification effectively enriches for actively splicing and post-catalytic spliceosomes. We also examined the levels of snRNAs in Prp19 associated RNA and found that U2, U5 and U6 are highly enriched over U1 and U4 (Table 3.1), further confirming that the spliceosomes associated with Prp19 are primarily the B^{act}/B*, C complexes

and ILS and not earlier assembly intermediates. This is also true in *S. pombe* and *S. cerevisiae* (Table 3.1). An additional 24 factors associated with *C. neoformans* Prp19 have no orthologs in *S. cerevisiae* and all but three have human orthologs annotated as spliceosome-associated factors (Cvitkovic and Jurica, 2013). These include GPATCH1, HNRNPUL1, DHX35 and members of the exon junction complex (Figure 3.8C and Table 3.1), which are associated with human B^{act} and C complex (Bessonov et al., 2010).

Profiling spliceosome-associated RNA

We extracted RNA from affinity-purified spliceosomes from *S. cerevisiae*, *S. pombe* and *C. neoformans* using endogenously tagged Prp19 alleles (TAP in *S. cerevisiae* and *S. pombe* and FLAG in *C. neoformans*) and a single-step purification protocol (Figure 3.8A). We also isolated total RNA from whole-cell extracts and selected polyadenylated (polyA) RNA to measure the abundance of spliceosome-bound RNA and mature RNA. Comparing the abundance of the transcript associated with the spliceosome to the abundance of the polyA mRNA, we observed that, in all three yeasts, Prp19-associated RNA is enriched for transcripts having at least one annotated intron (Figure 3.1D). Furthermore, we observed significantly higher read coverage in introns relative to exons in libraries prepared from Prp19-associated RNA compared to those prepared from polyA RNA or RNA extracted from purifications that used cells lacking a tagged Prp19 allele (Figure 3.8B). As detailed below, RNA associated with Prp19-spliceosomes was processed, sequenced and analyzed in a number of different ways to illuminate the action of the spliceosome *in vivo* (Figure 3.1E).

3’ end profiling of transcripts reveals intermediates and products of splicing

Genome-wide methods to analyze the rate or efficiency of splicing, such as splicing microarrays (Clark et al., 2002) and high-throughput sequencing (Katz et al., 2010; Brooks et

al., 2011; Shen et al., 2014) typically rely on measuring precursor and spliced RNA. In contrast, using RNA from spliceosomes, we could directly observe the spliceosome-bound precursors, intermediates and products of the splicing reaction globally. To detect the intermediates and products, we labeled free 3' hydroxyl groups using an approach adapted from NET-seq (Figure 3.2A) (Churchman and Weissman, 2011). In this protocol, a 5'-adenylated DNA linker is ligated to available RNA 3' OH groups, allowing cDNA to be subsequently synthesized from the 3' end of the RNA via priming from the DNA linker. The cDNA is then circularized, PCR-amplified, and subject to sequencing. Importantly, we include a 6 nt random barcode at the end of the DNA linker, allowing the computational detection and collapse of sequencing reads corresponding to a single ligation event to prevent over-counting of PCR amplified DNA species. In addition to profiling 3' ends, we also prepared spliceosome-bound RNA and polyA RNA from whole cell extract for standard RNA-seq analysis (Figure 3.2A). For these samples, we first hydrolyzed the RNA by heating under alkaline conditions and then repaired the 3' phosphate with T4 polynucleotide kinase as previously described for ribosome profiling (Ingolia et al., 2012). We then reverse-transcribed the RNA, circularized and amplified the resulting cDNA to produce sequencing libraries using the same steps as 3' end profiling.

In each of the yeasts, we observed pile-ups of reads beginning at annotated 5' splice sites that correspond to 3'OH of the cleaved 5' exon, an intermediate of the splicing reaction (Figure 3.2B). This species is also sometimes detectable in NET-seq data (Mayer et al., 2015; Nojima et al., 2015), consistent with the occurrence of cotranscriptional splicing. We also observe pile-ups of reads at annotated 3'ss whose ends correspond the 3' OH of the excised lariat intron, a product of the reaction (Figure 3.2B). Comparing the RNA-seq data from spliceosome-bound RNA to polyA mRNA, we observe substantially more reads in introns, consistent with enrichment for unspliced transcripts and excised intron. 3' end intensities at annotated splice sites and read density across spliceosome-bound and polyA RNA are

reproducible between biological replicates (Figure 3.9A-F shows data for the intron-rich species, *S. pombe* and *C. neoformans*).

To automatically detect peaks, we developed a pipeline using the PELT (Pruned Exact Linear Time) method (Killick et al., 2012) in the R package changepoint (Killick and Eckley, 2014). We applied this approach (see Methods) to two replicate experimental datasets obtained using a tagged Prp19 allele and data from an untagged control experiment. Peaks that were detected in both experimental datasets but not in the control were retained for analysis. The goal of the analysis was to identify high-confidence peaks suitable for downstream quantitative analysis.

The number of annotated introns detected with high confidence using this approach is dependent on the number of introns in the organism and the expression level of intron-containing transcripts under the culture conditions used. Further, we observed that particular introns in highly expressed genes frequently do not exhibit read pile-ups, presumably due to low intermediate levels produced by rapid second step kinetics. 9%, 34% and 50% of peaks were reproducibly detected in the tagged strains and absent from untagged in *S. cerevisiae*, *S. pombe* and *C. neoformans*, respectively. However, 69%, 89% and 88% of peaks at annotated 5' splice sites and 71%, 92% and 79% of peaks at annotated 3' splice sites were reproducible and absent from untagged. This means that the peaks corresponding to splicing events are reproducible. In *S. cerevisiae*, our computational method reproducibly picked peaks at 34% of annotated 5' splice sites and 58% of annotated 3' splice sites. In *S. pombe*, our approach picked peaks at 62% of annotated 5' splice sites and 21% of 3' splice sites. In *C. neoformans*, our method picked peaks at 16% of annotated 5' splice sites and 10% of annotated 3' splice sites. Overall, we detected peaks at splice sites in 53%, 78% and 31% of transcripts with annotated introns in *S. cerevisiae*, *S. pombe* and *C. neoformans* respectively.

We observed peaks corresponding to annotated splice sites, as well as peaks throughout transcripts (Figure 3.2C). While the sequences neighboring peaks at annotated sites mirror the consensus sequence for each organism as expected, peaks corresponding to unannotated sites did not follow splicing consensus sequences (Figure 3.2D, right panel). The majority of these may correspond to non-adenylated/uridylated transcript 3' ends produced by pausing or termination of RNA polymerase II or endonucleolytic cleavage. While we observed a large number of unpredicted peaks (Figure 3.2D), only a minority of reads at peaks are at unpredicted sites (6.5% in *S. cerevisiae*, 7.2% in *S. pombe* and 22% in *C. neoformans*).

Junction and branch profiling confirm identities of 3' end peaks

To identify the subset of unpredicted peaks that correspond to *bona fide* cleaved 5' exons at unannotated 5' splice sites and or excised intron lariats at unannotated 3' splice sites, we took advantage of existing tools for detecting the complementary species produced by the splicing reaction, namely the branch formed after the first step of splicing or the exon-exon junction formed after the second step (Figure 3.3A). To observe branching events, we enriched for lariat-intermediate and lariat-intron species by treating spliceosome-bound RNA with the 3'-5' exonuclease, RNase R (Suzuki et al., 2006). We then synthesized cDNA under conditions that enable read-through of lariats (see Methods), constructed sequencing libraries and performed 100 bp paired-end sequencing. After aligning reads to the genome with Tophat (Trapnell et al., 2009), we searched those that did not align to the genome for reads that begin downstream of a 5'ss or unpredicted peak and end downstream inside the same intron (Figure 3.3A). To identify exon-exon junctions, we prepared cDNA suitable for sequencing using random priming and performed 100 bp paired-end sequencing on spliceosome-bound RNA, using TopHat to detect exon-exon junctions. Of junctions detected in RNA-seq analysis of spliceosome-bound RNA in *C. neoformans*, 26% were observed only in spliceosome-bound RNA and not in polyA RNA

(Figure 3.3B). In contrast, only 3% of junctions observed by analysis of polyA RNA were not observed in spliceosome-bound RNA (Figure 3.3B). Transcripts with junctions only apparent on the spliceosome display slightly but significantly lower expression levels than those without; however, they also exhibit significantly higher levels in the spliceosome-bound RNA-seq data (Figure 3.3C). Additionally, they show significantly higher ratios of spliceosome-bound RNA to polyA RNA (Figure 3.3C, right panel), suggesting that they are either subject to turnover and/or accumulate on the spliceosome. Thus, analysis of junctions in spliceosome-associated RNA enables the detection of thousands of events that do not lead to the accumulation of polyadenylated RNA, presumably because such products are subject to decay mechanisms.

To select high-confidence splicing events, we developed a pipeline that compares branches and junctions to peaks detected by 3' end profiling (see Figure 3.3D for examples, algorithm design in Figure 3.10A). Using this approach, we identified thousands of peaks that correspond to annotated and unannotated complementary splicing events in *S. pombe* and *C. neoformans*. Figure 3.3E shows examples of alternative 5'ss choice in *S. pombe* and *C. neoformans*. Note that low frequency alternative splicing events are not expected to contribute significantly to the mature polyA mRNA population and may represent splicing errors subject to downstream RNA turnover mechanisms.

We observe 89 unannotated 5' or 3' splice sites in *S. pombe* and 446 in *C. neoformans* that have a detectable peak by 3' end profiling and either a junction or branch. Within these high-confidence sets, dinucleotides immediately before the peaks are strongly enriched for the GT dinucleotide (5'ss) and the dinucleotides immediately after the peaks are strongly enriched for the AG dinucleotide (3'ss), consistent with the identification of *bona fide* splicing events in both *C. neoformans* (Figure 3.3F, 7% of unpredicted peaks) and *S. pombe* (Figure 3.10B, 9% of unpredicted peaks).

Spliceosome profiling detects non-canonical events

We next tested whether the sensitivity of spliceosome profiling enabled the detection of unusual splicing events. Using junction profiling, we detected numerous alternative splicing events (738 in *S. pombe* and 1367 in *C. neoformans*) that coincide with at least one peak from 3' end profiling. Within these unannotated events, we searched for examples of types of non-canonical splicing. One such type is interrupted splicing, where the spliceosome recognizes a 5'ss and BP and performs the first step but then does not proceed to the second step due to lack of an acceptable splice receptor (Figure 3.4A). In *S. cerevisiae*, interrupted splicing of the *BDF2* transcript (encoding a BET family bromodomain protein) results in degradation of the transcript by the nuclear exosome in a process called spliceosome mediated decay (Volanakis et al., 2013). Indeed, we observed a substantial amount of spliceosome-bound intermediate for the *BDF2* transcript and no peaks that correspond to a 3'ss (Figure 3.4B). In *S. pombe*, the 3' end of the telomerase RNA, *TER1*, is formed by the first step of splicing occurring without the second (Kannan et al., 2013). We observed the intermediate formed by this reaction (Figure 3.4C), but also the exon-exon junction to a nearby 3'ss; however, transcripts with this junction appear to be unstable as only the full length *TER1* transcript was detected in polyA mRNA (Figure 3.4C, grey trace) and the relative read coverage is low compared to that of canonically spliced transcripts (Figure 3.11A). We also uncovered one other transcript subject to interrupted splicing, an unannotated *S. pombe* transcript named SPBC530.07c (Figure 3.4D and 3.11A). This gene encodes a member TENA/THI-4 family of proteins, which includes enzymes involved in thiamine metabolism (Pang et al., 1991; Akiyama and Nakashima, 1996). The *BDF2*, *TER1* and SPBC530.07c transcripts are enriched 30-75 fold on the spliceosome relative to transcripts without annotated introns (Figure 3.11B, compare *BDF2* to *FET5* and Figure 3.11C, compare *TER1* and *TENA* to *ACT1*), consistent with accumulation of intermediates on the spliceosome.

In addition to interrupted splicing, we observed other non-canonical phenomena. This includes recursive splicing, the removal of introns that flank a zero-nucleotide exon, which has recently been reported to be extensive in *D. melanogaster* (Duff et al., 2015) and in mammals (Sibley et al., 2015) based on intron coverage patterns and junctions in high-depth RNA-seq data. We identified a recursive splicing event in *C. neoformans* in a 375 nt intron of the *BOT1*/*CNAG_07884* gene, which encodes an protein essential for *C. neoformans* viability (Ianiri and Idnurm, 2015) related to the *S. pombe* Bot1 mitochondrial ribosomal protein (Wiley et al., 2008). Multiple lines of evidence indicate that recursive splicing effectively splits this intron into two smaller ones of 210 and 165 nt (Figure 3.4E). Specifically, we identified a peak that includes signal from both the lariat intron of the first splicing event and the free 5' exon of the second splicing event. We also identified junctions and branches for both events (Figure 3.4E, magenta and yellow, shorter arches) and the junction that corresponds to the final exon-exon junction (Figure 3.4E, magenta, long arch). We observed two additional large peaks (Figure 3.4E, grey) that do not correspond to splicing events but appear to be the 3' ends of premature transcripts based on spliceosome-bound RNA-seq (Figure 3.4E, blue trace). In another transcript, an alternative 3'ss is chosen that, curiously, results in the precise skipping of the downstream exon (Figure 3.4F). The isoform lacking this exon was not detected in the mature polyA mRNA (data not shown).

Another type of non-canonical splicing is nested splicing where a smaller intron is excised from a larger intron potentially reducing the size of the intron ("intron within an intron"). We observed 13 examples of nested splicing in *C. neoformans* (one example is shown in Figure 3.4G). We also identified many instances (63 in *S. pombe* and 104 in *C. neoformans)* of introns where branch sites unusually distal from the 3'ss are recognized and used for the first step (Figure 3.4H). These "early" or "distal" branching events have also been detected in RNA-seq data from human cells (Taggart et al., 2017). In both *S. pombe* and *C. neoformans*, they are particularly common in introns larger than 150 nt (Figure 3.11D). In *C. neoformans*, 43% of early

branching events result in BP-3'ss distances greater than 50 nt (Figure 3.11E), which may be prohibitively large for the second step of splicing.

Estimating the efficiency of splicing with spliceosome profiling

Assuming steady state kinetics, reproducible measurements of the relative abundance of pre-mRNA and intermediate bound to Prp19-associated spliceosomes (normalized to total spliceosome-bound RNAs) would enable estimates of the ability of endogenous substrates to progress through the first and second steps *in vivo* (including conformational changes that precede the chemical steps), thereby providing insights into how substrates behave once activated spliceosomes have assembled.

To accomplish this, we focused on high-confidence splicing events identified by our pipeline by first defining a transcript set that is above a threshold for the quantification limit of the assay (at least one high-confidence splicing event within each transcript; *S. cerevisiae*: 128, *S. pombe*: 1889, *C. neoformans*: 2121). Within this set, we quantified spliceosome-bound precursor and intermediate levels at each annotated intron and any unpredicted 5' splice sites. We defined the level of bound precursor as the number of unique reads that begin at least 5 nt downstream of the 5'ss and end at least 5 nt upstream (Figure 3.5A) normalized to the total read density of spliceosome-bound RNA across the transcript (Figure 3.5B). Based on the assumption that the steady-state level of precursor bound to the spliceosome would be inversely proportional to the rate of the first step of splicing, we used precursor level to estimate the relative efficiency of the first step of splicing. We defined the level of bound intermediate as the number of unique reads that begin at the 5'ss (corresponding to the cleaved 5' exon, Figure 3.5A) normalized to the total read density of spliceosome-bound RNA across the transcript (Figure 3.5B) and we employed this ratio to estimate the relative efficiency of the second step of splicing. Spliceosome-bound precursor and intermediate level measurements were both

reproducible between biological replicates, with intermediate level measurements displaying higher reproducibility (Figure 3.5B and 3.12A-B). Precursor levels form a narrower distribution in all three organisms than intermediate levels (Figure 3.5B and 3.12A-B, Levene test (log_2 transform) p-values – *S. cerevisiae*: 2 $X10^{-22}$, *S. pombe*: 3 $X10^{-279}$, *C. neoformans*: 5 $X10^{-295}$).

The total amount of transcript associated with Prp19 spliceosomes positively correlates with the abundance of polyA mRNA in the cell as determined by RNA-seq (Figure 3.5C) and is likely related to the rate of transcription. In contrast, mRNA abundance does not positively correlate with spliceosome-bound precursor or intermediate level in any yeast, suggesting a strong role for substrate differences in controlling progression through the spliceosome cycle *in vivo*. Unexpectedly, relative intermediate levels on spliceosomes from *S. cerevisiae* are strongly negatively correlated with mRNA abundances (Figures 3.5C and D). This negative correlation is also apparent when comparing *S. cerevisiae* intermediate levels with amount of spliceosome-bound transcript (Figure 3.5E). Slight negative trends are seen for these metrics in *S. pombe* and *C. neoformans* as well (Figure 3.5E). Additionally, precursor levels for introns on the same transcript exhibit slightly but significantly lower variance than for introns from randomly selected transcripts (Figure 3.12D) in *C. neoformans*. The variance in intermediate levels is not significantly different between introns in the same transcript and randomly selected introns

Bayesian modeling identifies features that predict bound precursor and intermediate levels

Because spliceosome profiling provides the opportunity to assay splicing efficiency across thousands of substrates, we sought to exploit the statistical power of the sample sizes to investigate, in an unbiased fashion, intron features that predict the levels of normalized spliceosome-bound precursor and intermediate. As any such approach is data-driven, we restricted our analysis to the intron-rich species *S. pombe* and *C. neoformans*. We employed

Bayesian Model Averaging, a probabilistic approach that generates large numbers of linear models, averages the best models and determines the probability, magnitude and uncertainty of the contribution of a given feature to the final ensemble of models (Clyde, 2017). The probability of including an intron feature in the final model is expressed as the marginal posterior inclusion probability or PIP (see Methods). A PIP of 0.5 is interpreted as an equal probability (50%) that a feature is significantly predictive of the response variable. In our case, the response variable is the level of normalized spliceosome-bound precursor or intermediate. We tested a range of features based on our current understanding of introns and the spliceosome (Figure 3.13A and Table 3.6). We focus here on three aspects of the results of this analysis: 1) the correlation of match to intron consensus with spliceosome-bound precursor and intermediate, 2) the correlation of distance-related features such as intron size, and 3) roles of intron number and position.

We observed PIPs of 1.0 for a negative correlation of 5'ss and BP scores with spliceosome-bound precursor levels in both *S. pombe* and *C. neoformans*, presumably reflecting the role of these sequences in step 1 catalysis (Figure 3.6A-D and 3.13A-B), such that a stronger 5'ss and BP results in more rapid transition through the first step (or the immediately preceding steps). 5'ss, BP, pyrimidine tract, and 3'ss scores are positively correlated with intermediate levels in *S. pombe* and BP and pyrimidine tract scores in *C. neoformans* (PIPs=1.0, Figure 3.6B-C and 3.13B-C). The predictive roles of these sequences for intermediate levels may reflect higher accumulation of the products of step 1 catalysis (the cleaved 5' exon and lariat-intermediate) when splicing signals are more optimal. They could also reflect reduced rates of spliceosomal remodeling steps required to execute the second chemical step (see Discussion).

Strikingly, a predictive feature of both precursor and intermediate levels (PIPs=1.0) in *C. neoformans* identified by the models is intron size (Figure 3.6D-E and 3.13A). This feature is

not predictive in *S. pombe* (PIP=0.02 and 0.15, for precursor and intermediate). Notably, *C. neoformans* introns display a size distribution that is significantly narrower than that of *S. pombe* (Figure 3.13D). Precursor and intermediate levels are contributed to by the distance between the BP to 3'ss in *S. pombe* and *C. neoformans* (PIPs=1.0; Figure 3.6F), and are also positively predicted by the presence of detectable alternative BP or 3'ss utilization events (PIP=1.0, Figure 3.13A and F). We also identified sizes of upstream and downstream exons as species-specific predictors (Figure 3.13A). Taken together, these data indicate that the distance relationships between substrate signals play a role in the efficiency of splicing after spliceosome assembly *in vivo* in a species that displays a relatively narrow intron length distribution.

Finally, we found that the number of introns in the transcript and the position of the intron in the transcript are also predictive of precursor and intermediate levels (Figure 3.6G-H). Transcripts with more introns predict lower intermediate levels in both *S. pombe* and *C. neoformans* (PIPs=1.0, Figure 3.6G and H, bottom panels). Precursor levels are also negatively predicted by intron number (PIPs=0.9-1.0, Figure 3.6G-H). Introns that are first in the transcript also exhibit higher intermediate level in *C. neoformans*, but lower precursor and intermediate level in *S. pombe* (Figure 3.6G and 3.13A and G).

Intron retention may be triggered by discard after spliceosome assembly

Intron retention is a major form of alternative splicing that is generally thought to be the result of inefficient spliceosome assembly. An untested alternative possibility for the source of such intron-retained species derives from the kinetic proofreading model of splicing fidelity described in the Introduction, in which suboptimal pre-mRNA substrates can be discarded from the spliceosome *after* assembly due to slow catalysis (Koodathingal and Staley, 2013). This hypothesis predicts a correlation between the levels of intron retention in the polyA RNA

population and the normalized levels of spliceosome-bound precursor, as high levels of the latter reflect slower progression through first-step catalysis after assembly.

We measured intron retention in RNA-seq data produced from polyA RNA, defining the level of retention as the read density inside the intron normalized to the total read density across the transcript. In *S. cerevisiae*, we observed no significant correlation between spliceosome-bound precursor level and intron retention (Figure 3.7A, left panel). However in the other two yeasts, which display denser and more diverse intron populations, we observed significant correlations, indicating that precursor molecules that progress slowly through the first chemical step after spliceosome assembly are more likely to be retained in the polyA population (Figure 3.7A). These data are consistent with the model that *S. pombe* and *C. neoformans* execute discard after assembly to produce intron-retained species in the polyA population.

A potential alternative explanation for the correlation between intron retention in polyA RNA and spliceosome-bound precursor levels is that introns adjacent to active spliceosomes that are not directly recognized are contributing to the measurement of precursor level. We addressed this possibility in two ways. We first examined single-intron transcripts. In *S. pombe*, we could measure splicing efficiency at 690 such introns and found that indeed the correlation persists (Figure 3.14A). In *C. neoformans*, single introns are rare. To separate spliceosomes bound to the same transcript, we treated immobilized, affinity purified spliceosomes with RNase I (Figure 3.14B-C). Analysis of RNA extracted these samples revealed that the correlation was maintained (Figure 3.14D).

Unexpectedly, we also observed similar correlations with intermediate level (Figure 3.14E); however, in this case, *S. cerevisiae* and *C. neoformans* exhibited the strongest effects. This correlation may occur because defects in signals that result in slow step 1 kinetics also result in slow step 2 kinetics – this is known to be the case for certain splice site mutations in *S. cerevisiae* (Lesser and Guthrie, 1993; Query and Konarska, 2004; Liu et al., 2007).

Prp43 mediated discard of precursor and intermediate species *in vivo*

Nonetheless, the result that inefficiently spliced introns are more likely to appear as retained introns in the polyadenylated mRNA population generally supports the hypothesis that substrates can be discarded from the spliceosome after its assembly (Mayas et al., 2010). To further test this idea, we employed a mutation in the RNA helicase that disassembles the spliceosome upon successful completion of splicing, Prp43 (Tsai et al., 2005). *In vitro* studies in *S. cerevisiae* have demonstrated that addition of a dominant-negative form of Prp43 that binds the spliceosome but cannot hydrolyze ATP results in accumulation of spliceosome-bound lariat intermediate that would otherwise be released from spliceosomes via its disassembly (Mayas et al., 2010). These studies utilized a substrate mutated at the 3'ss to block the second catalytic step. We hypothesized that expression of this allele would also lead to accumulation of precursor and intermediate species of endogenous introns on spliceosomes *in vivo* if they were normally subjected to discard and Prp43-dependent disassembly (Figure 3.7B). To test this hypothesis, we created a strain of *C. neoformans* that expresses the equivalent dominant negative Prp43 allele, *prp43-Q435E* (hereafter referred to as *prp43DN*) under control of the galactose-inducible *GAL7* promoter inserted at the *URA5* locus. Under repressive glucose media conditions, the *prp43DN* transcript accumulates to roughly the same abundance as the wild type transcript (Figure 3.7C), indicating leaky expression of the *GAL7* promoter, but the strain grows as well as wild type on glucose (Figure 3.7C). Under inducing galactose media conditions, the *prp43DN* transcript accumulates to 30-60 times the level of the wild type transcript and the strain displays a marked growth defect (Figure 3.7C). No splicing factors exhibited decreased expression in either condition (see RNA-seq data in Figure 3.14F and Table 3.7); nonetheless, to minimize potential indirect effects, we performed spliceosome profiling analysis on wild type and the *prp43DN* strain cultivated in glucose-containing media. As with the analyses described above, experiments were performed in duplicate.

To examine changes in the steady-state spliceosome-bound substrate populations in the *prp43DN* strain, we grouped introns based on normalized bound precursor and intermediate levels using K-means clustering (Figure 3.7D, see Methods). Of these clusters, one exhibits a significant increase in spliceosome-bound precursor levels (Cluster 1, Figure 3.7D-E) and another exhibits a significant increase in spliceosome-bound intermediate levels in the *prp43DN* strain (Cluster 2, Figure 3.7D-E). These data support the model that Prp43 disassembles spliceosomes *in vivo* prior to completion of the catalytic steps and that specific sets of introns are subject to this activity.

Introns whose spliceosome-bound precursor and intermediate levels are controlled by Prp43 do not display significantly weaker 5'ss or 3'ss sequences (Figure 3.14G-H). Rather, introns with increased intermediate levels in the *prp43DN* strain display shorter BP to 3'ss distances on average (Figure 3.7F) and detectably stronger 5'ss and BP scores (Figure 3.14G-H, BP scores – Mann-Whitney U p-value: 4 $X10^{-6)}$. Introns in this cluster are also more likely to be the last intron (Figure 3.14I). In contrast, introns with increased spliceosome-bound precursor in the *prp43DN* background are significantly longer (Figure 3.7F). Introns in this cluster are also more likely to be first introns (Figure 3.14I). Together with the studies of intron retention above, these studies provide evidence for widespread discard of natural spliceosome-bound substrates and reveal endogenous predictors of this behavior.

Discussion

The spliceosomal layer of gene expression determines the amount and structure of proteins and ncRNAs produced in the cell. Understanding these mechanisms during normal homeostasis and disease requires tools to interrogate this RNA-protein machine *in vivo*. We have described a suite of spliceosome profiling tools that marry the biochemical purification of endogenous spliceosomes with high-throughput sequencing/analysis methods to identify high-confidence splicing events in spliceosome-bound RNA, quantifying both the precursor and intermediates as well as identifying products that evade detection by RNA-seq. By implementing these methods in three widely divergent yeast species, we demonstrate that the method can be established in multiple systems. Although we used tagged alleles of the essential splicing factor Prp19, which is located at the periphery of all reported high resolution spliceosome structures, in principle, any accessible tag or epitope could be used, making the approach adaptable to any organism for which sufficient numbers of cells can be obtained. Spliceosome profiling has substantial advantages over current methodologies because 1) it quantifies spliceosome-bound precursor and intermediate levels, 2) it identifies trans-esterification events that cannot be easily identified by RNA-seq because of their transient nature or decay after release from the spliceosome, 3) it does not require additional mutations to stabilize RNA species. Consequently, we anticipate its widespread application for the analysis of conditions, drugs, and mutations that impact gene expression via the spliceosome. Below we discuss insights gained from this initial study and prospects for further development.

Integration of data reveals novel spliceosome-catalyzed transesterification events

Our initial analysis of RNA 3' ends associated with spliceosome-bound transcripts revealed cleavages at the boundaries of annotated introns as well as numerous unpredicted ends. Because many of these ends do not appear to be produced by the action of the

spliceosome, we filtered these data using information gained from profiling spliceosome-bound branchpoints and junctions. Through this approach, we obtained a high-confidence set of previously unannotated spliceosome-catalyzed events in *S. pombe* and *C. neoformans*. We detected non-canonical events including interrupted splicing events in *S. cerevisiae* and *S. pombe* including those involved in spliceosome-mediated decay and the biogenesis of telomerase RNA. We also identified a number of nested splicing events ("introns within introns") in *C. neoformans*. Finally, we identified an essential *C. neoformans* gene, *BOT1*, that is subject to recursive splicing in *C. neoformans*. Such events have previously only been described in animal cells for very long introns (Duff et al., 2015; Sibley et al., 2015). We anticipate that at least some of these non-canonical events will be verified as targets for regulation.

Quantification enables investigation of substrate features that influence progression

We proceeded beyond qualitative identification of cleavage and ligation sites by focusing on spliceosome-bound substrates sufficiently abundant to be quantified reproducibly. This enabled searches for the determinants of substrate behavior in fully assembled spliceosomes *in vivo*, which was not possible with prior technologies. The much larger number of quantifiable splicing events in the intron-rich yeast species *S. pombe* and *C. neoformans* allowed for the use of an automated statistical modeling approach to identify predictive substrate features. A clear finding was that stronger 5'ss and BP sequences predict lower bound precursor levels in *S. pombe* and *C. neoformans*, as expected for a role for the intron itself in catalytic efficiency (or immediately preceding conformational changes). This finding is consistent with recent cryo-EM structures that show known and novel intimate interactions between the 5'ss and BP sequences with the catalytic core of the spliceosome (Fica and Nagai, 2017; Shi, 2017). Surprisingly, more optimal intron sequences (5'ss, BP, pyrimidine tract and 3'ss) predicted higher intermediate

levels, particularly in *S. pombe*. As mentioned above, this may reflect the more rapid accumulation of intermediate as a result of the effects of more optimal sequences on the first chemical step.

Modeling also revealed predictive roles for splice site and BP spacing and spliceosome-bound precursor and intermediate levels in *C. neoformans*, but not *S. pombe*. As mentioned above, this is fully consistent with the tighter distribution of intron sizes in *C. neoformans*, which has been shown to be under complex evolutionary selective forces (Hughes et al., 2008). Consistent with prior work, we observe that long introns in *S. pombe* and *C. neoformans* tend to harbor detectably stronger splicing signals (Figure 3.13C), further suggesting evolutionary pressure to compensate for non-optimal intron size. Our observation of 'early branching' in long introns may be indicative of size being measured prior to the first chemical step. The corresponding longer distance between the BP and 3'ss could then impact the second chemical step as this is a well established determinant of step 2 kinetics in *S. cerevisiae* (Luukkonen and Séraphin, 1997; Zhang and Schwer, 1997).

Numerous species display narrow intron length distributions, with the most remarkable being the ciliate *Stentor* in which nearly all introns are precisely 15 nucleotides in length (Slabodnick et al., 2017). The *Drosophila* genome is enriched for introns of 60-65 nucleotides in length (Lim and Burge, 2001) and smaller introns may be selected for during evolution due to improved splicing efficiency (Carvalho and Clark, 1999). Intron length has been shown to impact overall efficiency during *in vitro* splicing in HeLa and in *Drosophila* cell extracts (Guo and Mount, 1995; Fox-Walsh et al., 2005) and *in vivo* (Ulfendahl et al., 1985) but whether this is due to an impact on spliceosome assembly or catalysis has not been assessed; our studies demonstrate a role after assembly *in vivo*. Identifying the yardstick that measures intron size in assembled spliceosomes in species that display such constraints will likely require structural analysis.

Our modeling studies also revealed that higher numbers of introns in a transcript correlate with lower levels of bound intermediate in both *S. pombe* and *C. neoformans*. An intriguing explanation for this correlation would be that spliceosomes bound to nearby introns on a transcript can influence each other's activities. Testing these and other hypotheses inspired by the modeling will require further investigation. While the modeling presented in this study begins to describe the intron and transcript features that contribute to splicing efficiency, the parameters considered here (which originate largely from studies in *S. cerevisiae*) have limited predictive power. This underscores the need to discover and understand additional contributors to splicing efficiency in the context of the cell. Not only does splicing happen in the context of chromatin, which may be able to promote or inhibit splicing (Sorenson et al., 2016), but it also occurs to a large extent in the context of transcription (Oesterreich et al., 2016) and may be influenced by the rate of transcription (Braberg et al., 2013; Aslanzadeh et al., 2017). We anticipate that spliceosome profiling will enable further investigation these poorly-understood connections.

Spliceosome discard of native transcripts

The ability of spliceosomes stalled *in vitro* to be disassembled is well established. However, as these studies require alterations to the substrate that induce stalling, the extent to which such events occur frequently in normal cells and what their function might be is unclear. Spliceosomal discard has been conceptualized as a mechanism for enhancing the fidelity of splicing by limiting the impact of splicing errors. This thinking is based on substantial *in vivo* and *in vitro* analysis of mutant substrates in *S. cerevisiae*. While a reasonable inference, it has not been tested in the context of naturally-occurring transcripts, in part because deviation from consensus is rare in *S. cerevisiae*, but also because tests for discard *in vivo* have been challenging to develop. A notable exception is the finding that Prp43 is important for the

biogenesis of the *TER1* RNA in *S. pombe* via interrupted splicing (Kannan et al., 2013). For this transcript, discard appears to promote a specific biogenesis pathway, rather than a mechanism for the suppression of errors. Our finding that intron retention correlates with high levels of spliceosome-bound precursor suggests another role for discard in the biogenesis of intron-retained RNA, a common form of alternative splicing. Intron retention has recently been shown to be regulated by environmental stresses in *C. neoformans* (Gonzalez-Hilarion et al., 2016) raising the possibility that there exist corresponding signal transduction mechanisms that impinge on fully assembled spliceosomes.

In complementary studies, we identified precursor and intermediate populations that accumulate on spliceosomes in cells expressing ATPase-defective version of Prp43. This finding provides the *in vivo* evidence for Prp43 activity upstream of spliceosome disassembly on a substantial fraction of native transcripts prior to the completion of splicing. Our analysis revealed several predictors of this behavior (intron size for Prp43-dependent precursor discard and BP to 3'ss distance for Prp43-dependent intermediate discard), suggesting that these introns are more likely to undergo discard under normal conditions. Notably, these substrates did not differ in the overall strength of splicing signals. Additional parameters may determine the sensitivity of these transcripts to Prp43 activity.

Prospects

The suite of spliceosome profiling tools described here was implemented in three divergent yeasts, demonstrating the adaptability of the methods. The ability to interrogate mutants in spliceosomal components combined with the data analysis approaches described here offers an attractive approach for understanding the function of spliceosomal components/ protein domains/residues, including many whose three-dimensional location is now known but whose precise function across substrates remains to be dissected. Our proteomic

characterization of spliceosomes in *C. neoformans*, a species that harbors over 40,000 annotated introns, indicates it harbors numerous factors that have human orthologs but are not found in the intron-reduced species *S. cerevisiae.* These include components of the exon junction complex. Analysis of these factors by spliceosome profiling may illuminate mechanisms important for species harboring a higher degree of splicing complexity. Finally, translation of our protocols to mammalian cells should enhance studies of the role of the spliceosome in human disease.

Experimental procedures

Strain construction

C. neoformans strains were constructed by biolistic transformation (Chun and Madhani, 2010). Briefly, the ORF, 3' UTR and downstream 500 bp of *PRP43*-Q435E followed by a selection marker and flanked by 1 kb of homology to the *URA5* locus were cloned into vector pRS316 using homologous recombination in *S. cerevisiae*. 10 µg of the plasmid were linearized and desiccated and then precipitated onto gold beads using spermidine and $CaCl_2$. DNA-coated gold beads were shot into cells using a gene gun (BioRad, Model PDS-1000/He Biolistic-Particle Delivery System). Insertion of the full construct into the genome was confirmed by colony PCR using primers outside the sequence included in the plasmid and failure to grow on SC –Ura medium. The existence of the wild type and mutant copies of *PRP43* were confirmed by Sanger sequencing and detected by RNA-seq. Tags in all strains (including those previously constructed) were confirmed by Western blot.

Prp19 immunoprecipitation (FLAG)

2 L of all strains were grown in Difco YPAD medium starting at a low density (~0.002 OD/ml) overnight at 30°C. 1% glucose was added when the cultures reached 1 OD. Cells were harvested at 2 OD by centrifugation at 6200xg for 6 min at 4°C (Beckman JLA-8.1000 rotor). They were then washed in ice-cold sterile water and harvested again at 9000xg for 6 min at 4°C (Beckman JLA-8.1000 rotor). Cells were resuspended in 25 ml cold H03 buffer (25 mM HEPES-KOH, pH 7.9, 100 µM EDTA, 500 µM EGTA, 2 mM $MgCl_2$, 300 mM KCl, 16% glycerol, 0.1% Tween) with protease inhibitors (Complete-EDTA free, Pierce) and 1 mM DTT. The resuspension was dripped into liquid nitrogen to form popcorn. Cells were lysed by cryo-grinding in a SPEX Sample Prep 6870 for 3 cycles of 2 min at 10 cps with 2 min of cooling between each cycle.

Lysate was thawed quickly with mixing in a 30°C water bath until just liquid. 500 units of RNaseOUT (Invitrogen) were added immediately after thawing. Lysate was cleared by centrifugation at 40000xg for 40 min at 4°C (Beckman JA17.0 rotor). Whole cell extract samples were removed and flash frozen at this point. 400 µl M2-FLAG agarose beads (Sigma) were washed in 10 ml H03 buffer and collected by centrifugation at 1200 rpm (228xg) for 2 min three times. Lysate was incubated with the prepared beads for 3 hours at 4°C. The beads were collected by centrifugation at 1200 rpm (228xg) for 2 min and then resuspended and incubated for 10 min with 15 ml H03 buffer (protease inhibitor in the first wash) at 4°C. This was repeated two additional times (three washes total). Bound material was eluted three times in 150 µl elution buffer (25 mM HEPES-KOH, pH 7.9, 2 mM $MgCl_2$, 300 mM KCl, 20% glycerol) with 1 mg/ml 3X FLAG peptide (Sigma).

Prp19 immunoprecipitation (TAP)

2 L of all strains were grown in the appropriate rich medium (*S. cerevisiae*: Difco YPAD, *S. pombe*: YS + 3% glucose) starting at a low density (~0.002 OD/ml) overnight at 30°C. 1% glucose was added when the cultures reached 1 OD. Cells were harvested at 2 OD by centrifugation at 6200xg for 6 min at 4°C (Beckman JLA-8.1000 rotor). They were then washed in ice-cold sterile water and harvested again at 9000xg for 6 min at 4°C (Beckman JLA-8.1000 rotor). Cells were resuspended in 25 ml cold H03-200 buffer (25 mM HEPES-KOH, pH 7.9, 100 µM EDTA, 500 µM EGTA, 2 mM $MgCl_2$, 200 mM KCl, 16% glycerol, 0.1% Tween) with protease inhibitors (Complete-EDTA free, Pierce) and 1 mM DTT. The resuspension was dripped into liquid nitrogen to form popcorn. Cells were lysed by cryo-grinding in a SPEX Sample Prep 6870 for 3 cycles of 2 min at 10 cps with 2 min of cooling between each cycle.

Lysate was thawed quickly with mixing in a 30°C water bath until just liquid. 500 units of RNaseOUT (Invitrogen) were added immediately after thawing. Lysate was cleared by centrifugation at 40000xg for 40 min at 4°C (Beckman JA-17.0 rotor). Whole cell extract

samples were removed and flash frozen at this point. 400 µl IgG agarose beads (Sigma) were washed in 10 ml H03-200 buffer and collected by centrifugation at 1200 rpm (228xg) for 2 min three times. Lysate was incubated with the prepared beads for 3 hours at 4°C. The beads were pelleted by centrifugation at 1200 rpm (228xg) for 2 min and then washed three times for 10 min with 15 ml H03-200 buffer (protease inhibitor in the first wash). Prp19 was eluted in 500 µl 1X TEV buffer with 50 units AcTEV protease (Invitrogen) and 40 units RNaseOUT for 1 hour at room temperature.

Treatment of purified material with RNase I

Prp19 associated were affinity purified as above (FLAG protocol). However, immediately before elution, beads were resuspended in 500 µl H03 buffer with 1X PhoStop solution (Roche) and 30 U RNase I. The bead-associated spliceosomes were then incubated 1 hour at room temperature with agitation. After 1 hour, 5 µl Superasin (Ambion) was immediately added to the mixture and the slurry was placed on ice. The beads were washed twice for 10 min with 1 ml H03 buffer at 4°C. After careful removal of all wash buffer, Prp19 was eluted from the beads as described above.

The amount of RNase I to use for this assay was determined by adding 0.2-30 U RNase I or 0.2-2 U RNase T1 to the reaction mixture. 7.5 µl of eluate or digest supernatant were mixed with 7.5 µl formamide loading dye, boiled for 2 min, loaded on 10% polyacrylamide gel (8 M Urea, 1XTBE) and separated at 200 V for 90 min. The extent of digestion of RNA originating from *CNAG_07888* was then assayed by Northern blot using the Dig Northern starter kit (Sigma) according to the manufacturer's instructions. The membrane was exposed to film for 2 min. U6 was detected on the same membrane using a complementary ^{32}P end labeled oligomer. The probe was hybridized to the membrane for 6.5 hours at 45°C, then washed 3 times for 5 min in 2X SSC, 0.1% SDS and twice for 15 min in 0.1X SSC, 0.1% SDS at 45°C. Membrane-associated ^{32}P was detected by overnight exposure to a phosphorimager screen.

Mass Spectrometry

Immunoprecipitation of Prp19 was performed as described above, except that protein was eluted three times in 0.1 M glycine HCl, pH 3.5, for 5 min at 4°C. The samples were dissolved in 60 µL of 8 M urea in 100 mM TEAB, pH 8.0. Reduction was performed with at 5 mM TCEP 20 min at room temperature. Alkylation followed at 55 mM 2-chloroacetamide for 20 min in the dark at room temperature. The samples were precipitated in cold acetone (1:6) at −20 °C overnight. Tryptic digestion and Ten-plex TMT labeling (Thermo Scientific) were performed according to the manufacturer's instructions. The ten samples were combined and run. Additionally, the five control (untagged) samples were combined and analyzed as were five experimental (Prp19-2XFLAG) conditions. The TMT labeled samples were analyzed on a Fusion Orbitrap tribrid mass spectrometer (Thermo Fisher Scientific) with multinotch data acquisition (McAlister et al., 2014). Samples were injected directly onto a 30 cm, 75 µm ID column packed with BEH 1.7 µm C18 resin (Waters). Samples were separated at a flow rate of 400 nL/min on a Thermo Easy-nLC 1000 (Thermo Fisher Scientific). Buffer A and B were 0.1% formic acid in water and acetonitrile, respectively. A gradient of 0−10% B over 20 min, 10−45% B over 270 min, an increase to 90% B over another 60 min and held at 90% B for a final 10 min of washing was used for 360 min total run time. Column was re-equilibrated with 20 µL of buffer A prior to the injection of sample. Peptides were eluted directly from the tip of the column and nanosprayed directly into the mass spectrometer by application of 2.5 kV voltage at the back of the column. The Orbitrap Fusion was operated in a data dependent mode. Full MS1 scans were collected in the Orbitrap at 120k resolution. The cycle time was set to 3 s, and within this 3 s the most abundant ions per scan were selected for CID MS/MS in the ion trap. MS3 analysis with multinotch isolation was utilized for detection of TMT reporter ions at 60k resolution.

Monoisotopic precursor selection was enabled and dynamic exclusion was used with exclusion duration of 30 s.

Protein and peptide identification and protein quantitation were done with Integrated Proteomics Pipeline - IP2 (Integrated Proteomics Applications, Inc., San Diego, CA. http://www.integratedproteomics.com/). Tandem mass spectra were extracted from raw files using RawConverter (He et al., 2015) and were searched against a *C. neoformans* protein database (http://www.broadinstitute.org) with reversed sequences using ProLuCID (Peng et al., 2003; Xu et al., 2015). The search space included all half- and fully-tryptic peptide candidates, with static modifications of 57.02146 on cysteine and of 229.1629 on lysine and the N-terminus. Peptide candidates were filtered using DTASelect, with these parameters -p 1 -y 2 --trypstat --pfp .01 --extra --pI -DM 10 --DB --dm -in -t 1 --brief --quiet (Tabb et al., 2002). Quantitation was performed using Census (Park et al., 2014). Common contaminants, including heat shock, ribosomal and chaperone proteins as well as several other proteins we commonly find in mass spectrometry experiments in *C. neoformans* are excluded from Figure 3.8A. All detected proteins can be found in Table 3.1, excluded contaminants are shown in grey. Proteins with at least 15 spectral counts in all tagged samples were considered for further analysis.

RNA preparation

RNA was extracted from affinity purified spliceosomes and whole cell extract by first treating with 120 µg Proteinase K in reverse buffer (10 mM Tris-HCl, pH 7.4, 5 mM EDTA, 1% SDS). RNA was then isolated by phase separation with 1 volume acid Phenol:Chloroform (1:1) and then washed with 1 volume chloroform. RNA was precipitated with 2 µl GlycoBlue in 0.3 M sodium acetate and 50% isopropanol, then recovered by centrifugation for 30 min at 18400xg at 4°C and washed with 70% ethanol. RNA was initially checked by RT-qPCR and later by Bioanalyzer for enrichment of U2, U5 and U6 snRNAs after DNase treatment with 6 units TURBO DNase I (Invitrogen) and purification using the Zymo RNA Clean and Concentrate kit.

RNA from whole cell extract was polyA selected with OligoTex beads (Qiagen) and DNase treated in the same manner as the spliceosome-bound RNA.

3’ end profiling and total spliceosomal RNA-seq

Library preparation

RNA from each IP was split with 4 parts used for the “A” sample (3’ end profiling) and 1 part for the “B” sample (spliceosome-bound RNA). “A” sample RNA was ligated to a pre-adenylated adaptor with a 5’ random hexamer (see Key Resources Table) using T4 RNA ligase 2, truncated, K227Q (NEB) for 2.5 hours at 37°C (25% PEG-8000, 12.5% DMSO, 1X buffer). After precipitation and resuspension, RNA was hydrolyzed in hydrolysis buffer (100 mM sodium carbonate, pH 9.2) for a predetermined amount of time (10-15 min) at 95°C then immediately neutralized with 0.3 M sodium acetate and precipitated. “B” sample RNA was first hydrolyzed (see “A” sample) and precipitated. Hydrolyzed RNA ends were then repaired by treating with 25 units PNK (NEB) for 1 hour at 37°C. After precipitation, “B” sample RNA was ligated to the pre-adenylated adaptor as for the “A” sample. PolyA selected samples were treated the same as the “B” samples. All samples were loaded on a 10% polyacrylamide gel with 8 M urea and 1X TBE in 1X formamide dye with bromophenol blue. The gel was run for 50 min at 200 V then stained for 5 min in 1X Sybr Gold (Invitrogen), 1X TBE. RNA bands were excised between 50 and 400 nt. RNA was extracted from crushed gel in nuclease free water for 10 min at 70°C and then separated from the gel by centrifugation in a Spin-X column (Costar) for 3 min at 16000xg. RNA was precipitated with GlycoBlue and 0.3 M sodium acetate in 50% isopropanol.

All samples were converted to cDNA using the Superscript III kit (Invitrogen) and the oLSC007 reverse transcription primer (see Key Resources Table). cDNA was separated from primer by denaturing PAGE (same conditions as for RNA above). Resuspended cDNA was circularized with the Circligase kit (Epicentre). Circularized cDNA was amplified by PCR with the

Phusion polymerase kit (NEB) using one indexed primer and oNT231 (see Key Resources Table). Libraries were size selected on 8% non-denaturing polyacrylamide gels and confirmed by Bioanalyzer (High Sensitivity DNA kit). Libraries were sequenced on a HiSeq 4000 with single-end 50 bp reads with the oLSC006 custom sequencing primer.

Data analysis

5' adaptor sequence was trimmed from the 3' end of reads using custom scripts. Reads shorter than 24 bp after adaptor trimming were excluded from analysis. Reads were then collapsed using fastx_collapser (http://hannonlab.cshl.edu/fastx_toolkit/). The 6 nt barcode was removed from the beginning of collapsed reads using a custom script. Collapsed, trimmed reads were first aligned to rRNA and snRNA sequences using Bowtie (Langmead et al., 2009) with the following parameters:

bowtie -p10 -v2 -m1 --un CM764no_A_rRNA_un /home/jordan/GENOMES/crypto_rRNA -f CM764no_A_S64_L007_R1_001_debarcoded.fasta --sam CM764no_A_rRNA_al.sam

Unaligned reads were then aligned to the genome with Bowtie using the following parameters:

bowtie -p10 -v2 -M1 --best --max CM763db_A_multi --un CM763db_A_un /home/jordan/GENOMES/Crypto_for_gobs -f CM763db_A_UBC4_un --sam CM763db_A_al.sam

Conversion to BAM format, sorting and indexing was accomplished using Samtools (Li et al., 2009). BAM files were converted to bedgraph files using BEDTools (https://github.com/arq5x/bedtools2). Data were visualized using IGV (Robinson et al., 2011). Total reads in transcripts in RNA-seq style libraries were determined using custom scripts written around HTseq (Anders et al., 2014). Reads spanning 5' splice sites (precursor) and reads beginning at 5' splice sites (intermediate) were counted using custom scripts using the Pysam package (https://github.com/

pysam-developers/pysam). Precursor values are included for sites with at least 20 reads covering the exon-intron junction.

Peaks were identified using the PELT (Pruned Exact Linear Time) algorithm (Killick et al., 2012) in the R package changepoint (Killick and Eckley, 2014). First, 5' ends of aligned reads were converted to the bedgraph format using BedTools. Then, a range of thresholds between 1 and 100 were tested for each sample from each yeast. Peaks were detected after the threshold at which the number of detected peaks dropped most steeply (10 for all samples Figure 3.9G-I) was applied. Peaks called in data from both tagged samples but absent from untagged data were selected for further analysis.

Junction and branch profiling

Library preparation

RNA recovered by affinity purification of Prp19 was treated with 10 units RNase R (Epicentre) or mock treated for 1 hour at 37°C. RNA was then immediately purified using the Zymo RNA Clean and Concentrate kit and DNase treated as above. Libraries were prepared using the NEBNext directional kit (NEB) according to the manufacturer's instructions. To improve read-through of branches in lariat RNA species, we added 1 mM $MnCl_2$ to the reverse transcription reaction (Madhura Raghavan and Jeffrey Pleiss, personal communication). Final size selection was performed using an 8% non-denaturing 1X TBE polyacrylamide gel. Libraries were sequenced on a HiSeq 4000 with paired-end 100 bp reads.

Data analysis

Adaptor was trimmed from either end of reads using Cutadapt (https://github.com/marcelm/cutadapt). Reads shorter than 25 bp after trimming or with a quality score less than 10 were not considered in the analysis.

cutadapt -a AGATCGGAAGA -A AGATCGGAAGA --trim-n -m 25 -q 10 -o CM763-RR_1_trim.fq -p CM763-RR_2_trim.fq ../FASTQ_FILES/CM763-RR_S31_L007_R1_001.fastq.gz ../FASTQ_FILES/CM763-RR_S31_L007_R2_001.fastq.gz

Trimmed reads were aligned to the genome with Tophat

tophat -p 1 -o CM763-RR --library-type rf-firststrand -i 30 -I 400 -G CNA3_FINAL_CALLGENES_1_gobs.gff3 H99-2 CM763-RR_1_trim.fq CM763-RR_2_trim.fq

Data were converted for visualization in IGV as described above. Junction position and read depth were extracted from junction.bed files from Tophat using custom scripts. Only junctions with at least 5 reads were included in the analysis.

Reads that did not align to the genome were searched for branches with custom Python scripts. First, a set of potential 5' splice sites were established based on annotation and cleavage events that occur immediately before a "GC" or "GT" in the genome. Unaligned reads were extracted from the Tophat generated BAM file using BEDTools and searched for the 15 nt sequence immediately downstream of this site. All reads containing one of these sequences were split and realigned to the genome with Bowtie using the following parameters:

bowtie -p2 -v1 -M1 --best /home/jordan/GENOMES/Crypto_for_gobs -f Cn_ann_split.fa --sam Cn_ann_branches.sam

The 5' splice site position was embedded in the read name so it could be recovered later during analysis. Finally, potential branches were filtered based on the presence of an adenosine within 3 nt of the read split and required that the branch be with 1 kb of the 5' splice site. Only branches with at least 5 reads were considered for further analysis.

Statistical approaches

Precursor and intermediate level calculations

Precursor and intermediate levels were determined for high confidence 5' splice sites (as called by our pipeline described above and in Figure 3.10A). Both metrics are normalized to spliceosome-bound RNA-seq coverage for the corresponding transcript (SB). This is defined as the number of reads aligning to the transcript (on the correct strand) divided by the length of the transcript in kb then normalized to the sum of all such densities in all transcripts (T) divided by a million (similar to TPM).

$$SB_i = \frac{\frac{reads_i}{length_i}}{\left(\sum_{j \in T} \frac{reads_j}{length_j}\right) * 1X10^{-6}}$$

Precursor reads for a given splicing event, P, were determined by counting the number of reads that begin between 3 and 50 nt downstream of a 5' ss and end between 3 and 50 nt upstream of the same site (using Pysam). Reads were divided by the window size (0.1 kb) and then normalized to the total number of precursor reads from all splicing events (SE) divided by the window size (0.1 kb) and then a million. Precursor level, PL, was determined by dividing this value by the spliceosome-bound RNA-seq coverage (SB) for the corresponding transcript.

$$PL_k = \frac{\frac{P(k)}{0.1}}{\left(\sum_{l \in SE} \frac{P(l)}{0.1}\right) * 1X10^{-6}} / SB_i$$

Similarly, intermediate reads for a given splicing event, I, were determined by counting the number of reads that begin at the peak called by our pipeline. Intermediate reads were

normalized to the total number of intermediate reads from all splicing events (SE) divided by a million. Intermediate level, IL, was determined by dividing this value by the spliceosome-bound RNA-seq coverage (SB) for the corresponding transcript.

$$IL_k = \frac{I(k)}{\sum_{l \in SE} I(l) * 1X10^{-6}} / SB_i$$

Splice site scoring

Splice site scores of each site were determined using a position-specific scoring matrix (PSSM) method. A matrix is constructed based on all the annotated splice sites in the genome as follows:

$F[i,j] = \frac{n_{ij}}{k}$ where n_{ij} is the number of sequences with base *i* at site *j* and

k is the total number of sequences.

$P[i,j] = \frac{F[i,j]}{f(i)}$ where *f(i)* is the frequency of the base in the genome

Finally, $S[i,j] = log_2\big(P[i,j]\big)$

Each individual splice site in a sample is scored against this matrix. Logos were generated using WebLogo3 (Crooks, 2004) .

Branch determination

If a branch was determined experimentally, the branch with the most reads was used as the major branch site. Otherwise, branches were determined based on sequence and distance from the 3' splice site. First, a list of all experimentally determined branches was assembled for

C. neoformans and *S. pombe*. Each unique branch sequence was ranked based on its usage throughout the genome. Introns without experimentally determined branches were searched for each branch in this list in rank order. If no branch was found, then the adenosine closest to the 3' splice site was chosen as the branch site. This method produced a distribution similar in terms of sequence and branch to 3' splice site distance to experimentally determined branches.

Bayesian model averaging

Bayesian model averaging was performed using the R package BAS (Clyde, 2017).The optimal model was determined by Markov chain Monte Carlo sampling. Bayesian inclusion criteria (BIC) were used as the prior distribution of the regression coefficients and a uniform prior distribution was assigned over all models. Highly similar models were also obtained using traditional linear modeling.

K-means clustering

K-means clustering was performed using the Python package sklearn (Garreta and Moncecchi, 2013). The number of centroids was selected by the "elbow method" (Ketchen et al., 1996) by calculating the difference in the sum of squares for 10 clusters each in the range of 2 to 12 centroids.

Acknowledgments

The *S. cerevisiae* Prp19-TAP strain was a gift from C. Guthrie and the *S. pombe prp19-TAP* strain was provided by K. Gould. We thank M. Mayerle and anonymous reviewers for critical comments on the manuscript, J. Staley and D. Bartel for helpful discussions during early stages of this project, and N. Nguyen for technical support. We also thank members of the Madhani group for helpful discussions and support. Supported by NIH grants R01 GM71801 to H.D.M., P41 GM103533 to J.R., and R01 GM120507 to J.J.L. J.E.B was supported by postdoctoral fellowship 127531-PF-15-050-01-R from the American Cancer Society. H.D.M. is a Chan-Zuckerberg Biohub Investigator.

Figures

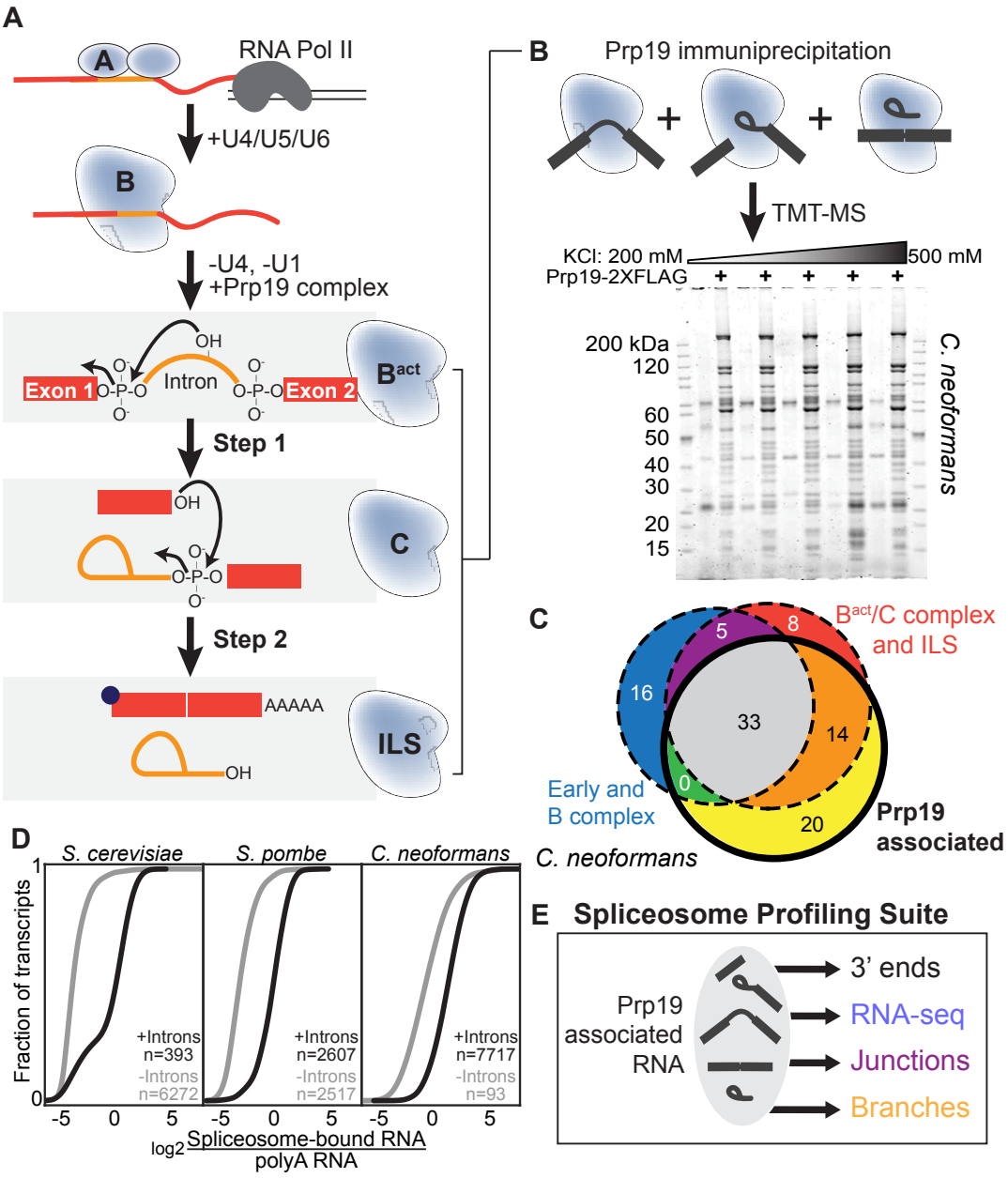


Figure 3.1: Isolating RNA from purified active spliceosomes using Prp19

A. Simplified view of the spliceosome assembly pathway and catalytic cycle indicating relevant substrate transformations and ribonucleoprotein complexes. B. SDS-PAGE (4-12% Acrylamide Tris-Glycine, Novex) analysis of spliceosomes affinity-purified from *C. neoformans* extracts using FLAG-tagged Prp19 at increasing KCl concentrations. Proteins were visualized using SYPRO Ruby. C. TMT-MS reveals that Prp19 associated splicing factors belong primarily to B^{act}, C and ILS complexes. Membership of *S. cerevisiae* orthologs in complexes was used to assign factors. The 20 factors in the yellow portion of the Venn diagram include orthologs of

splicing factors found in *H. sapiens* and/or *S. pombe* but not *S. cerevisiae* (See Figure 3.8C for list). Likely contaminants (ribosomal proteins, metabolic enzymes, chaperones and others) are not shown. D. RNAs associated with Prp19 spliceosomes are enriched for transcripts containing annotated introns (KS test, p-values: *S. cerevisiae*: $3X10^{-154}$, *S. pombe*: $< 2X10^{-308}$, *C. neoformans*: $9X10^{-16}$). E. RNA from Prp19-associated spliceosomes is used for multiple purposes in the spliceosome profiling suite.

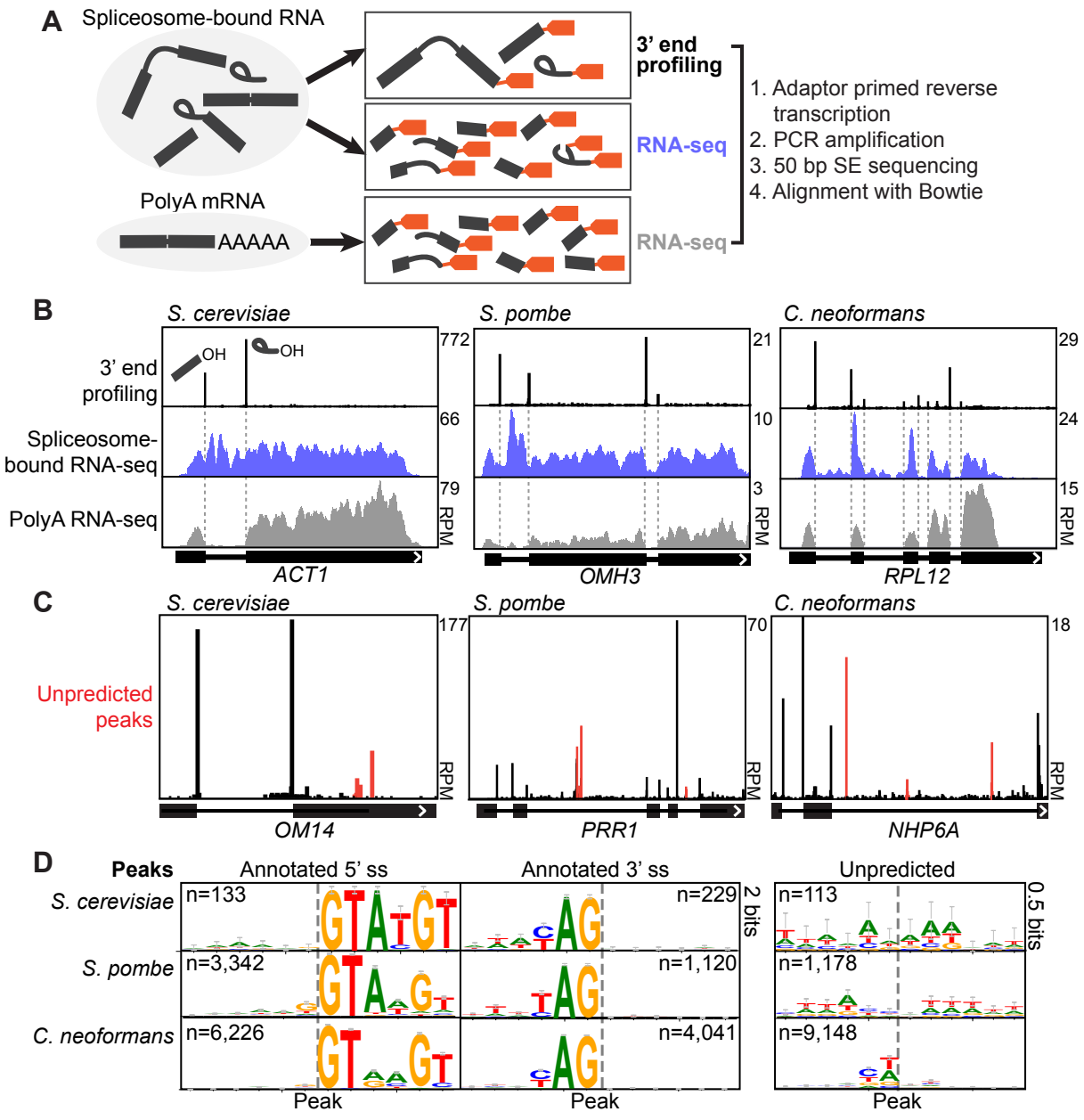


Figure 3.2: 3' end profiling reveals splicing intermediates and products in three yeasts.

A. Strategies for 3' end profiling and RNA-seq of spliceosome-bound and polyA RNA (see Results and Methods for details). B. 3' end profiling (black traces) reveals splicing intermediates (cleaved 5' exon) and products (lariat intron) globally in the three indicated organisms. Spliceosome-bound (blue) and polyA (grey) RNA-seq coverage for each transcript are also shown. C. 3' end profiling also discovers novel 3' ends (red) that may be spliceosomal

cleavages or transcript ends. D. Logos from predicted and unpredicted peaks detected by 3' end profiling in each of the three species.

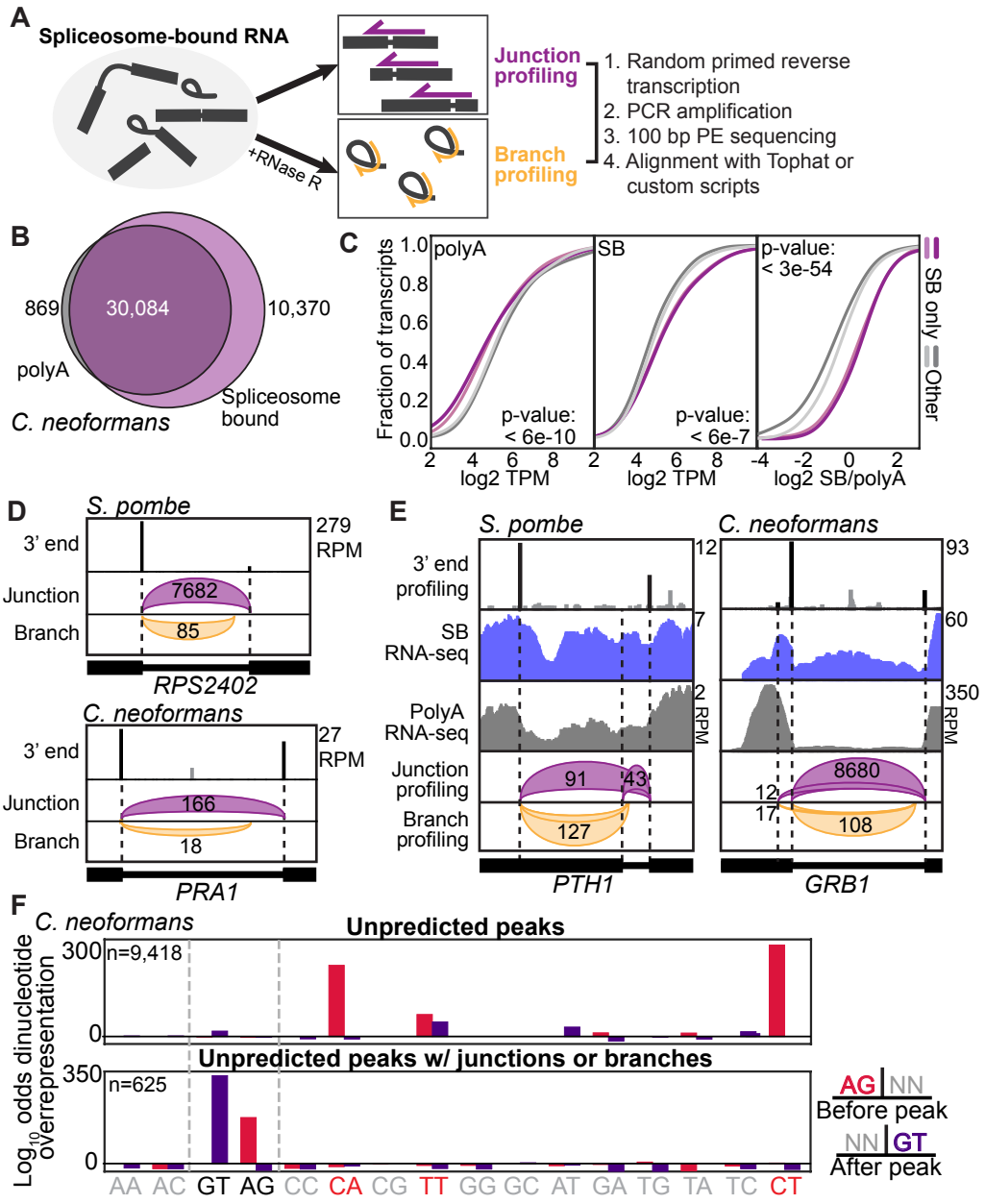


Figure 3.3: Branch and junction profiling confirm splicing events discovered by 3' end profiling.

A. Schematic of junction and branch profiling library preparation. Magenta arrows indicate reads that span exon-exon junctions and curved yellow arrows indicate reads that span branches. B. Overlap of junctions discovered by polyA vs. spliceosome-bound RNA-seq using Tophat reproducible in two replicates. C. Comparison of RNA levels for transcripts with and without junctions detected only in spliceosome-bound RNA. Light and dark traces represent biological replicate data. Log_2 transcript per million (TPM) levels are shown for polyA and spliceosome-

bound RNA-seq. The right panel shows the $\log_2$ ratio of spliceosome-bound to polyA RNA-seq (TPM). D. Example of an exon-exon junction and branch detected at an annotated introns in *S. pombe* and *C. neoformans* (*PRA1*: CNAG_04950). Purple and yellow curves are derived from junction BED files generated by Tophat (junction profiling) or custom scripts (branch profiling) visualized in IGV. Numbers indicate number of reads covering each exon-exon junction or branch (not normalized). E. Examples of junctions and branches that confirmed unannotated alternative 5' splice sites in *S. pombe* and *C. neoformans* (*GRB1*: CNAG_03281). Peaks that align with other splicing events are shown in black. Other 3' end profiling signals are shown in grey. F. The likelihood that a given dinucleotide is overrepresented based on its frequency in the genome (displayed as a log(odds ratio)) immediately upstream and downstream of peaks (determined independently) in *C. neoformans* (see Figure 3.10B for *S. pombe*). Unfiltered unpredicted peaks are enriched for pyrimidine rich dinucleotides (red letters) and splicing signal dinucleotides (black letters) while peaks with a junction or branch are only enriched for splicing signal dinucleotides.

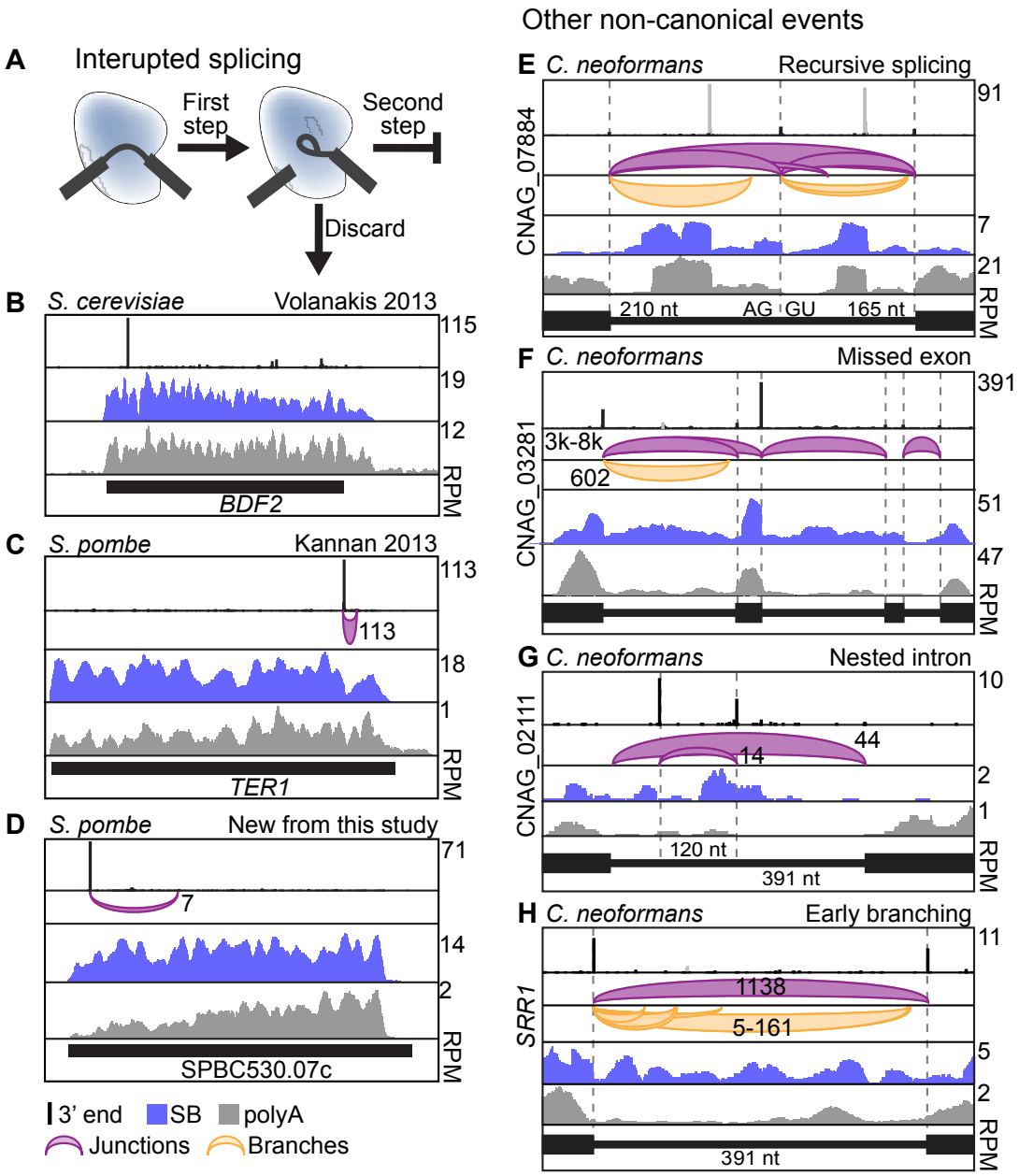


Figure 3.4: Non-canonical splicing events.

A. Splicing can be interrupted between the two steps when discard is triggered by a slow second step (Kannan et al., 2013; Volanakis et al., 2013). B-D. Examples of interrupted splicing. 3' end profiling results are shown in black, spliceosome-bound RNA-seq data are shown in blue, polyA RNA-seq data are shown in grey and junction profiling results are shown in purple. The number of reads corresponding to each junction (not normalized) is indicated. E. An example of recursive splicing in *C. neoformans*, where an annotated intron is actually two introns separated

by a zero nucleotide exon “pivot point.” Traces are colored as above and branch profiling is shown in yellow. Grey peaks correspond to putative 3’ RNA ends based on the profile of the spliceosome-bound RNA-seq data for this transcript. F. An example of a splicing event resulting in exclusion of an exon in *C. neoformans*. Traces colored as above. G. Nested intron inside a relatively large (391 nt) intron. Only the 391 nt intron is detectable by polyA RNA-seq (data not shown). Traces colored as above. H. An example of early branches inside a relatively large (391 nt) intron. In these cases, the branch to 3’ss distances are much larger than the typical range of 15-30 nt. Traces colored as above.

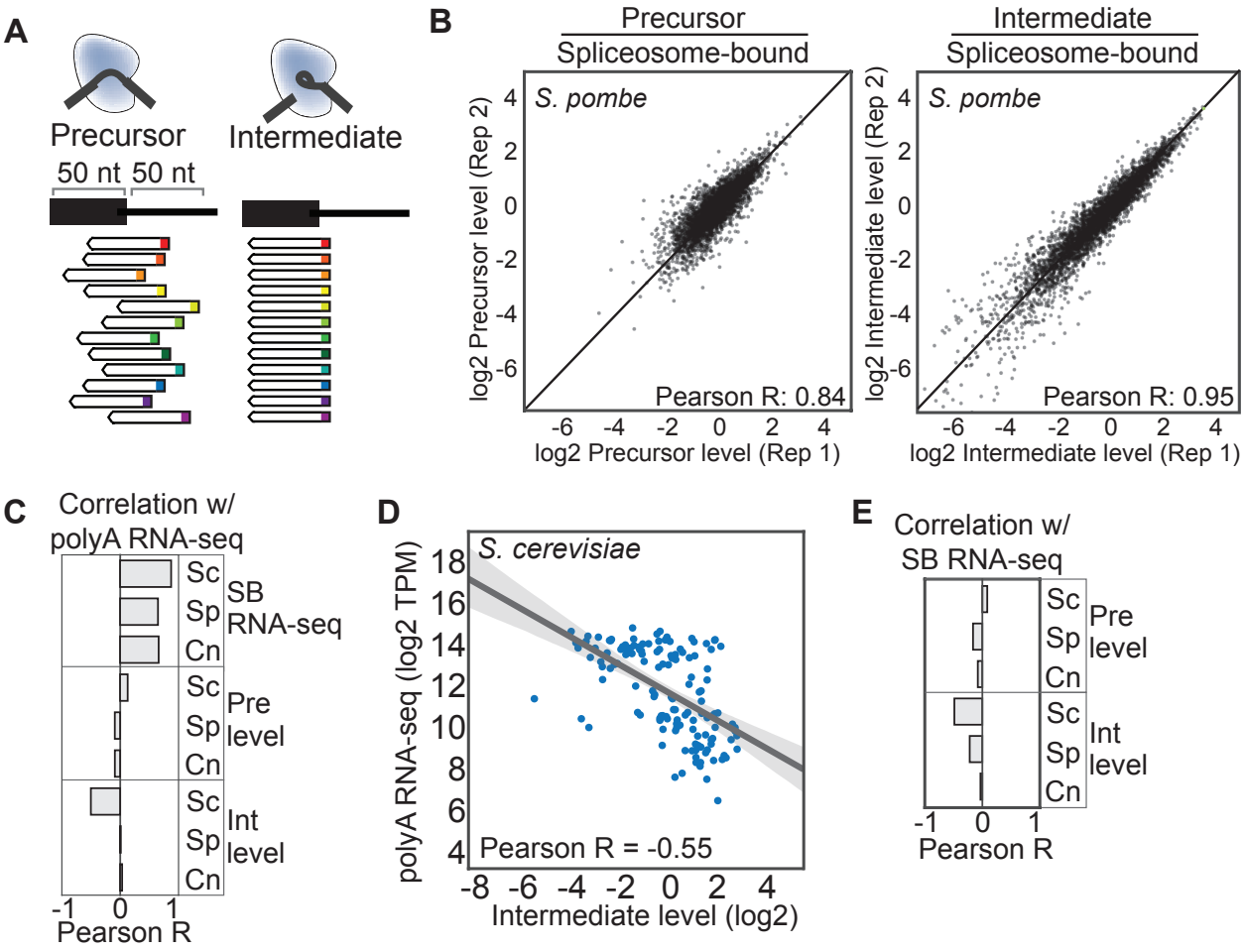


Figure 3.5: Quantitation of spliceosome-bound precursor and intermediate

A. Quantitating the level of precursor and intermediate at each splice site. Precursor is determined by counting reads that start downstream of the splice site and end upstream. Intermediate is determined by counting reads that begin at the splice site. Each read begins with a random hexamer bar code (rainbow colors). B. Reproducibility and distribution of precursor and intermediate levels (precursor and intermediate normalized to the read density of spliceosome-bound transcript) in *S. pombe* (see Figure 3.12A-B for other yeast). Each point represents one splicing event. All annotated introns in any transcript with at least one detected high-confidence splicing event are included to avoid bias against splicing events with low intermediate accumulation. C. Correlation of precursor, intermediate and spliceosome-bound transcript levels with the abundance of the mature transcript as determined by polyA RNA-seq (log_2 transformed to approximate normality for Pearson R analysis). D. Correlation of polyA RNA-seq with intermediate levels in *S. cerevisiae*. Grey area indicates 95% confidence interval for the linear regression model. E. Correlation of precursor and intermediate levels with spliceosome-bound transcript levels (log_2 transforms).

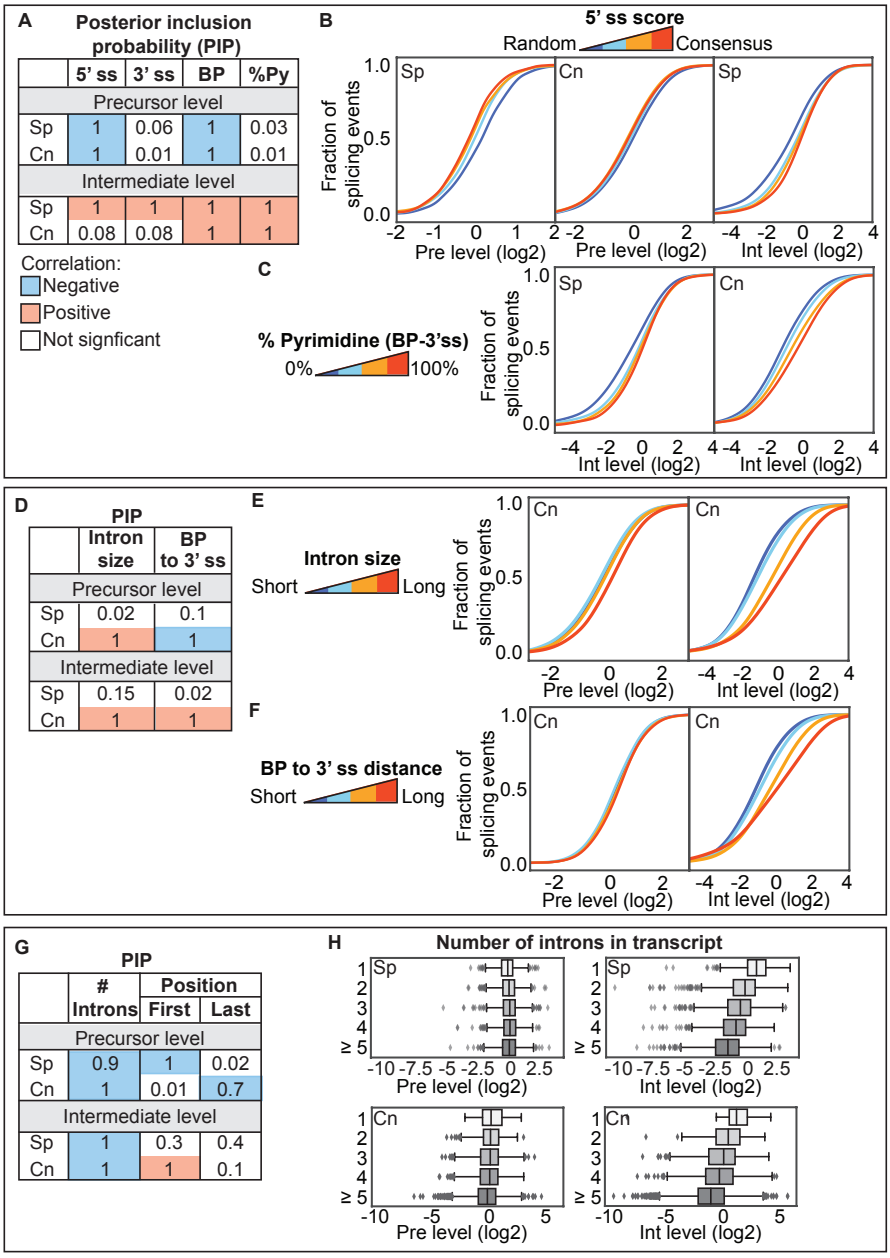


Figure 3.6: The relationship between intron features and splicing efficiency.

A. Posterior inclusion probabilities (PIPs) for Bayesian Model Averaging of the relationship between intron features and precursor and intermediate levels. PIPs for 5'ss, 3'ss and BP scores and % pyrimidine between the BP and 3'ss are shown. Intron features with a PIP higher than 0.5 are highlighted (red = positive correlation, blue = negative correlation). See Figure 3.13A for coefficients and uncertainties. B. Relationship between precursor and intermediate levels and 5'ss scores. Introns were split into quartiles based on 5'ss score (blue=random

sequence, red=consensus) and the distribution of each metric is plotted as a CDF (See Figure 3.13B-C for BP and 3'ss scores). Sp indicates *S. pombe* and Cn indicates *C. neoformans*. C. Same as B but for the percent pyrimidine in the region between the branch point and the 3' ss. D. PIPs for intron size and the distance between the BP and 3'ss. Intron features with a PIP higher than 0.5 are highlighted (red = positive correlation, blue = negative correlation). E. Relationship between precursor and intermediate levels and intron length in *C. neoformans*. Introns were split into quartiles based on length (blue=short, red=long) and the distribution of each metric for each quartile is plotted as a CDF. F. Same as E but for BP to 3'ss distance. G. PIPs for the number of introns in the transcript and the relative position of the intron in the transcript. Intron features with a PIP higher than 0.5 are highlighted (red = positive correlation, blue = negative correlation). H. Difference in precursor and intermediate levels based on the number of introns in the transcript (See Figure 3.13F-G for alternative splicing and position in transcript).

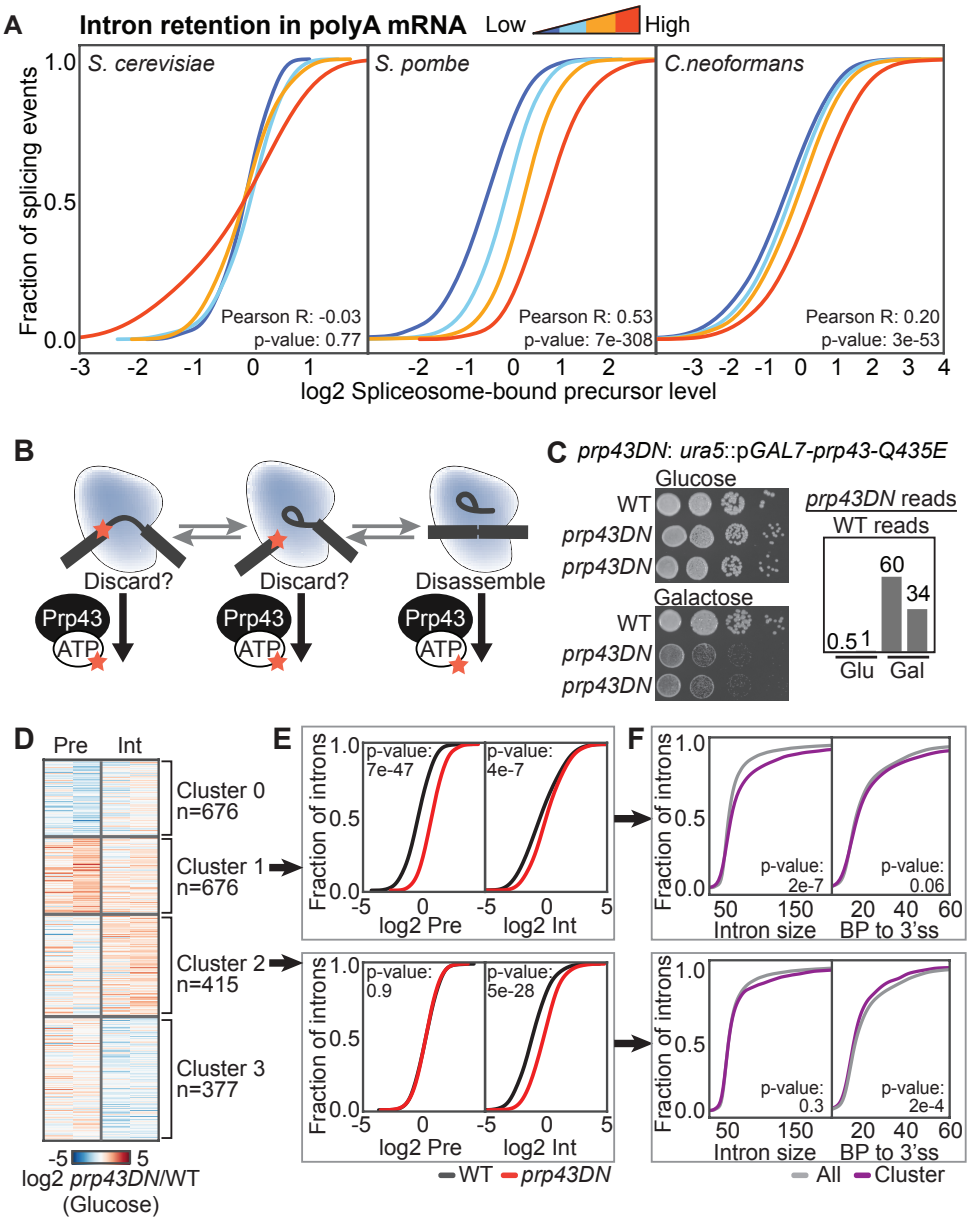


Figure 3.7: Prp43 discards long introns before the first step of splicing.

A. Intron retention correlates with precursor level in *S. pombe* and *C. neoformans* but not *S. cerevisiae*. Introns were split into quartiles based on the level of intron retention calculated in polyA RNA-seq data. The distribution of precursor level in each quartile is plotted as a CDF (See Figure 3.14E for intermediate level). B. Simplified model of Prp43 mediated discard from the spliceosome (adapted from (Koodathingal and Staley, 2013)). This model predicts that

disabling the ATPase activity of Prp43 will cause precursor and intermediates to accumulate on spliceosomes. C. Growth of strains expressing *prp43-Q435E* or *prp43DN* from the *URA5* locus in glucose or galactose media and relative expression of *prp43-DN* (compared to the wild type transcript) in each condition, determined by determining the number of reads with and without the corresponding nucleotide change from polyA RNA-seq data. D. Heat map of the ratio of precursor and intermediate in the *prp43DN* strains vs. wild type grown under repressive conditions. Introns are sorted by K-means clustering with 4 centroids (see Methods). E. CDF plots of precursor and intermediate in *prp43DN* (red) and wild type (black) in glucose for cluster 1 (top panel) and cluster 2 (bottom panel) where intermediate or precursor, respectively, is increased in the *prp43DN* strains. P-values determined by KS test. F. Intron size and BP to 3'ss distance for each cluster compared to all introns for cluster 1 (top panel) and cluster 2 (bottom panel). P-values determined by Mann-Whitney U test.

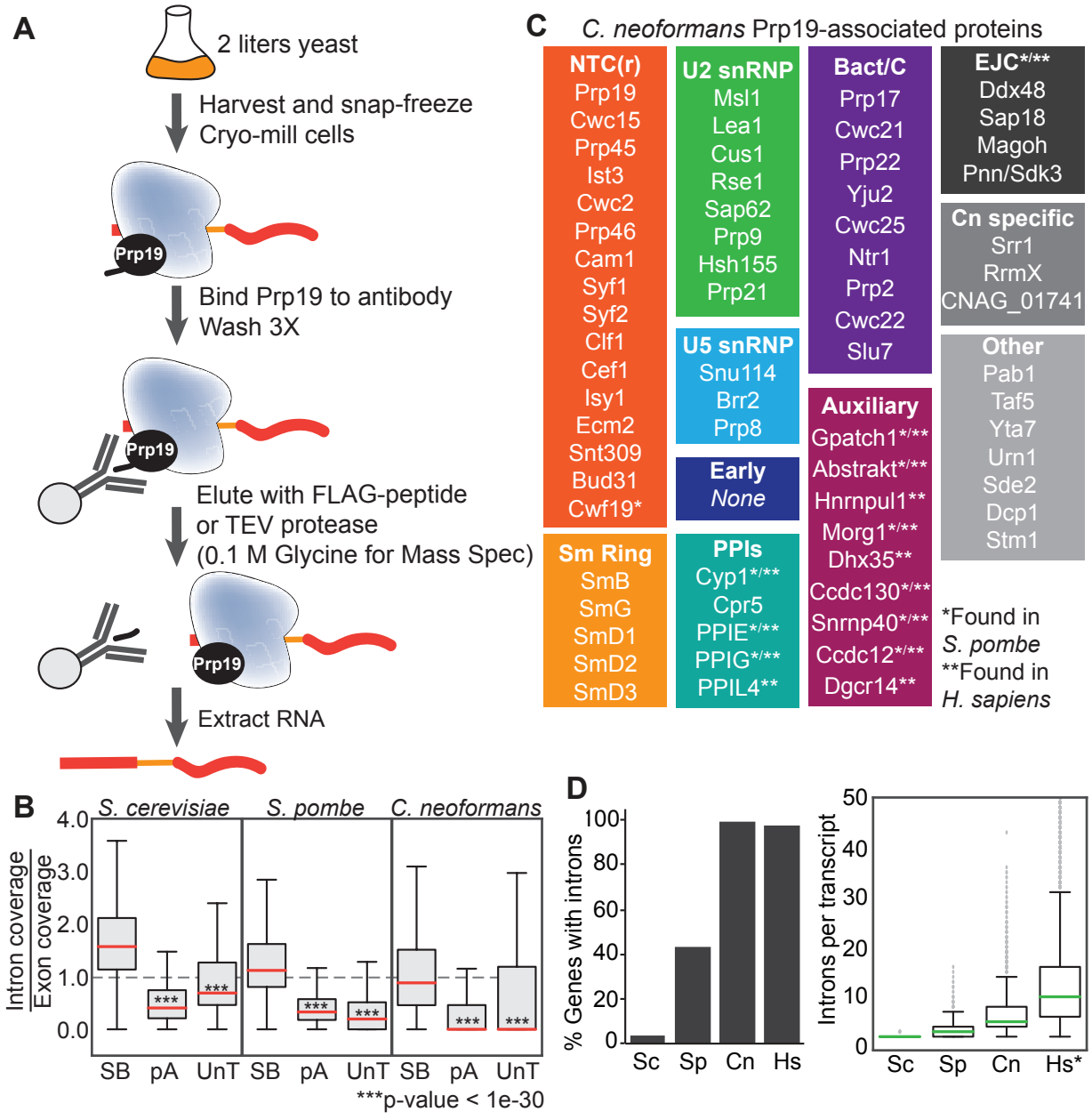


Figure 3.8: Composition of affinity purified spliceosomes, related to Figure 3.1.

A. General scheme for purifying Prp19-associated spliceosomes (see methods for details). B. Comparison of read coverage in introns versus exons. Reads in introns and exons were counted in RNA-seq data sets (SB: spliceosome-bound/Prp19-associated, pA: polyA, UnT: affinity-purified from cells lacking tagged Prp19) then divided by the size of the intron or spliced transcript in kb to calculate coverage. The ratio of intron to exon coverage is plotted for all transcripts with exon coverage of at least 10. P-values were determined by Mann-Whitney U

test. C. Splicing factors detected by TMT-MS in Prp19-IPs organized by complex named based on their *S. cerevisiae* orthologs, unless they are not found in *S. cerevisiae* (“Auxiliary”, “PPIs” and “EJC”) in which case they are named for the *S. pombe* or human ortholog. Category name abbreviations: NTC(r): Prp19 complex and Prp19 complex related; PPIs: peptidyl prolyl isomerases; EJC: exon junction complex; Cn specific: specific to *C. neoformans.* D. Proportion of genes with annotated introns and density of introns per transcript in each organism from this study (Sc: *S. cerevisiae,* Sp: *S. pombe*, Cn: *C. neoformans*) as well as humans (Hs).

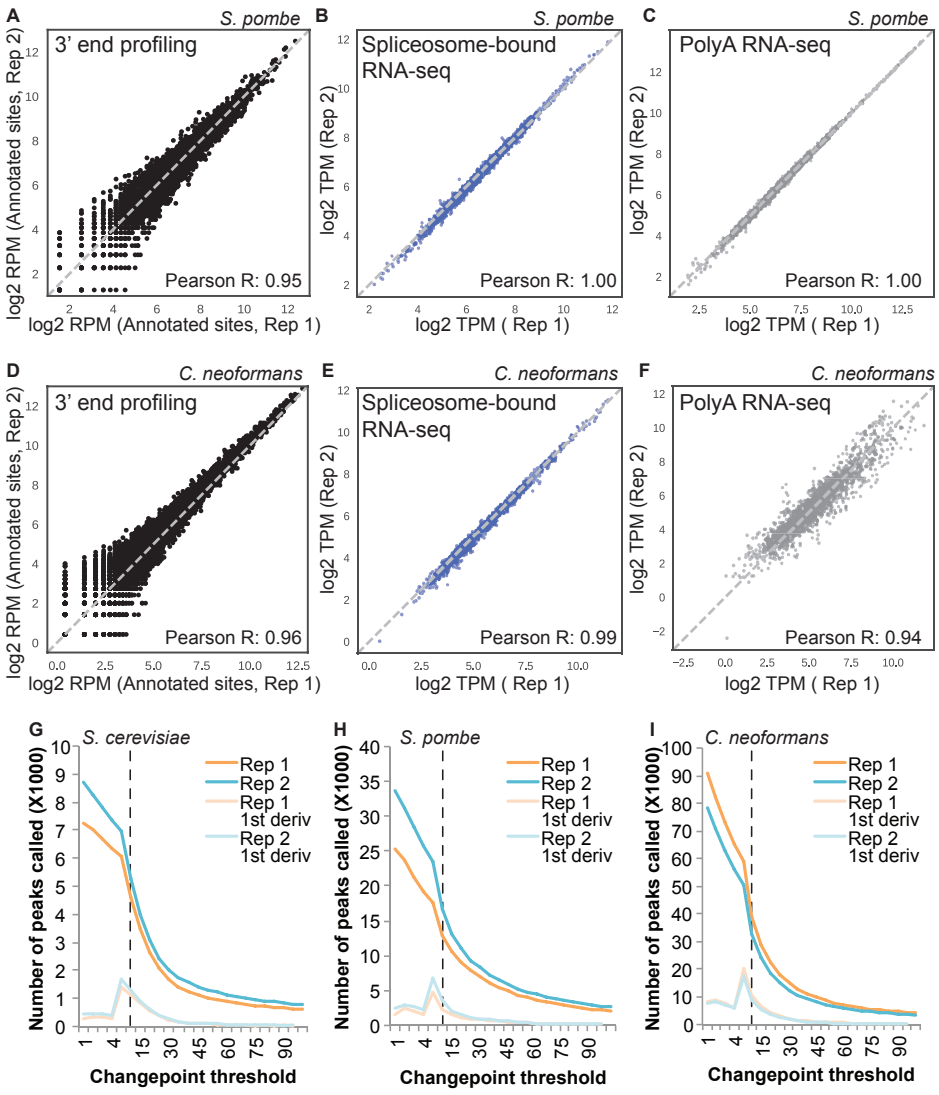


Figure 3.9: Data selection and reproducibility, related to Figure 3.2.

Reproducibility and distributions of 3' end profiling peak heights (A) and RNA-seq of spliceosome-bound RNA (B) and polyA RNA (C) in *S. pombe*. 3' end profiling is measured in reads per million aligned reads. RNA-seq metrics are measured in transcript per million (TPM). D-F. Same as A-C but for *C. neoformans*. G-I. Determination of changepoint thresholds in all three yeast. Peaks were detected using changepoint analysis (see Methods) in each yeast. The number of peaks at each threshold was determined and the threshold that accomplished the

largest drop in the number of peaks (maximum of the 1st derivative traces, light orange and green) was chosen as the threshold.

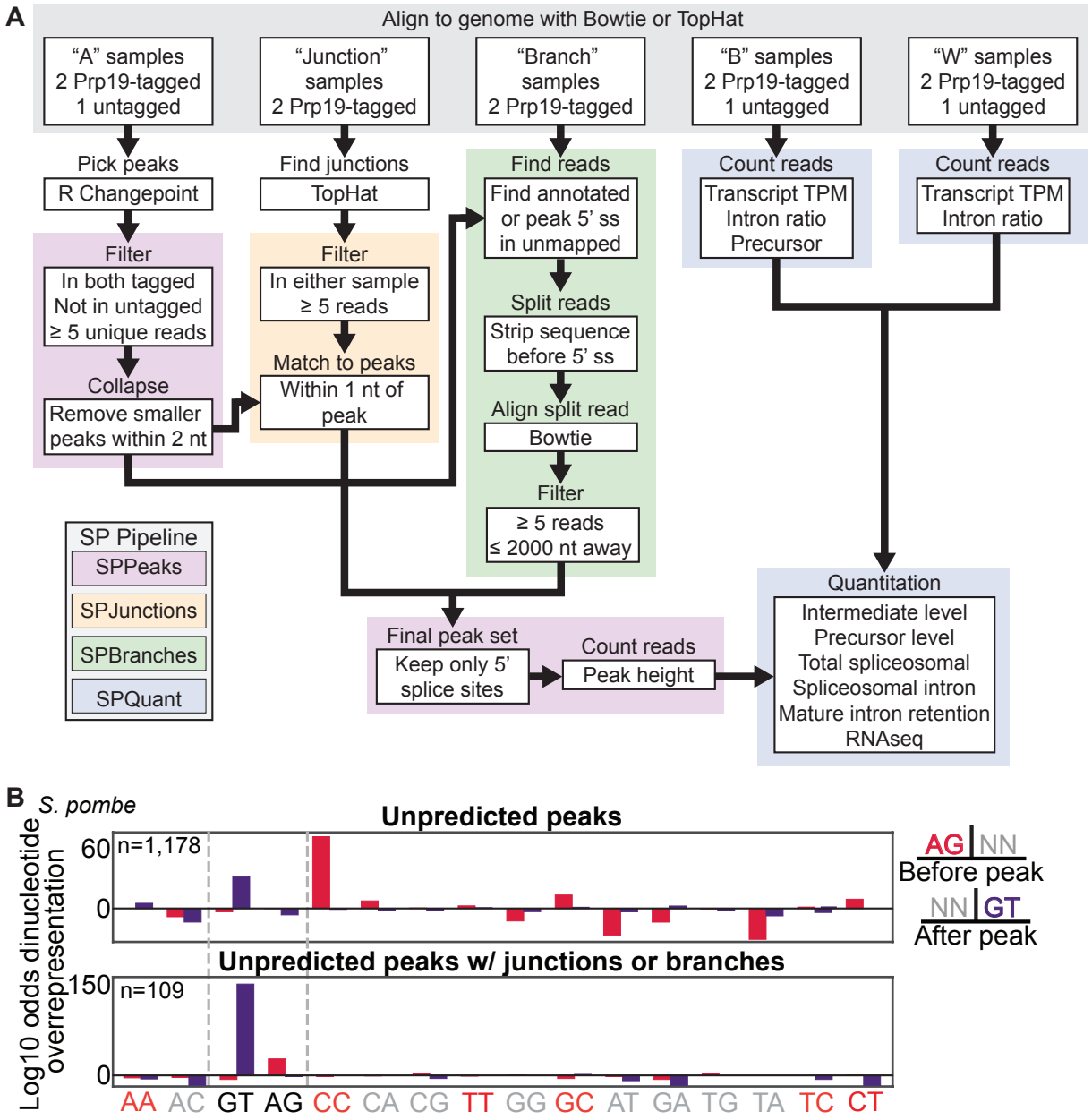


Figure 3.10: Selection of high confidence splicing events, related to Figure 3.3.

A. Workflow and structure of the spliceosome profiling software package (SP Pipeline). See methods for additional details. B. Dinucleotide overrepresentation in unpredicted peaks and unpredicted peaks that overlap with junctions or branches in *S. pombe.*

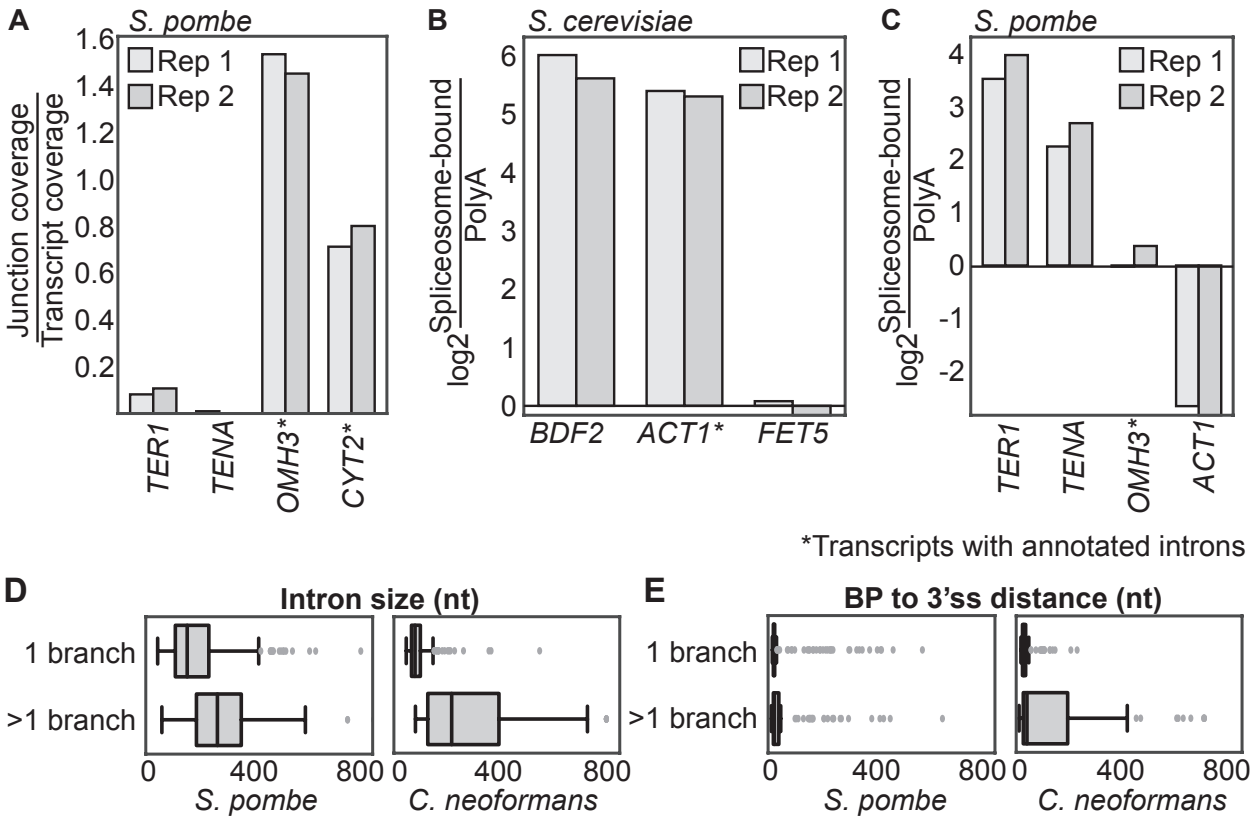


Figure 3.11: Additional information on non-canonical splicing, related to Figure 3.4.

A. Ratio of junction coverage to transcript coverage for exon-exon junction reads originating from either interrupted splicing events (*TER1* and *TENA*/SPBC530.07c) or canonical splicing events (*OMH3* and *CYT2*) in *S. pombe*. B-C. Abundance of spliceosome-bound vs. polyA mRNA for transcripts with interrupted splicing events compared to canonically spliced and unspliced transcripts. D. Size distributions of introns with only one detected branch or more than one detected branch. Branches were detected with the SPBranch module of the SP Pipeline as described in the methods. E. Distance between the branch point and 3'ss in introns with only one detected branch or more than one detected branch.

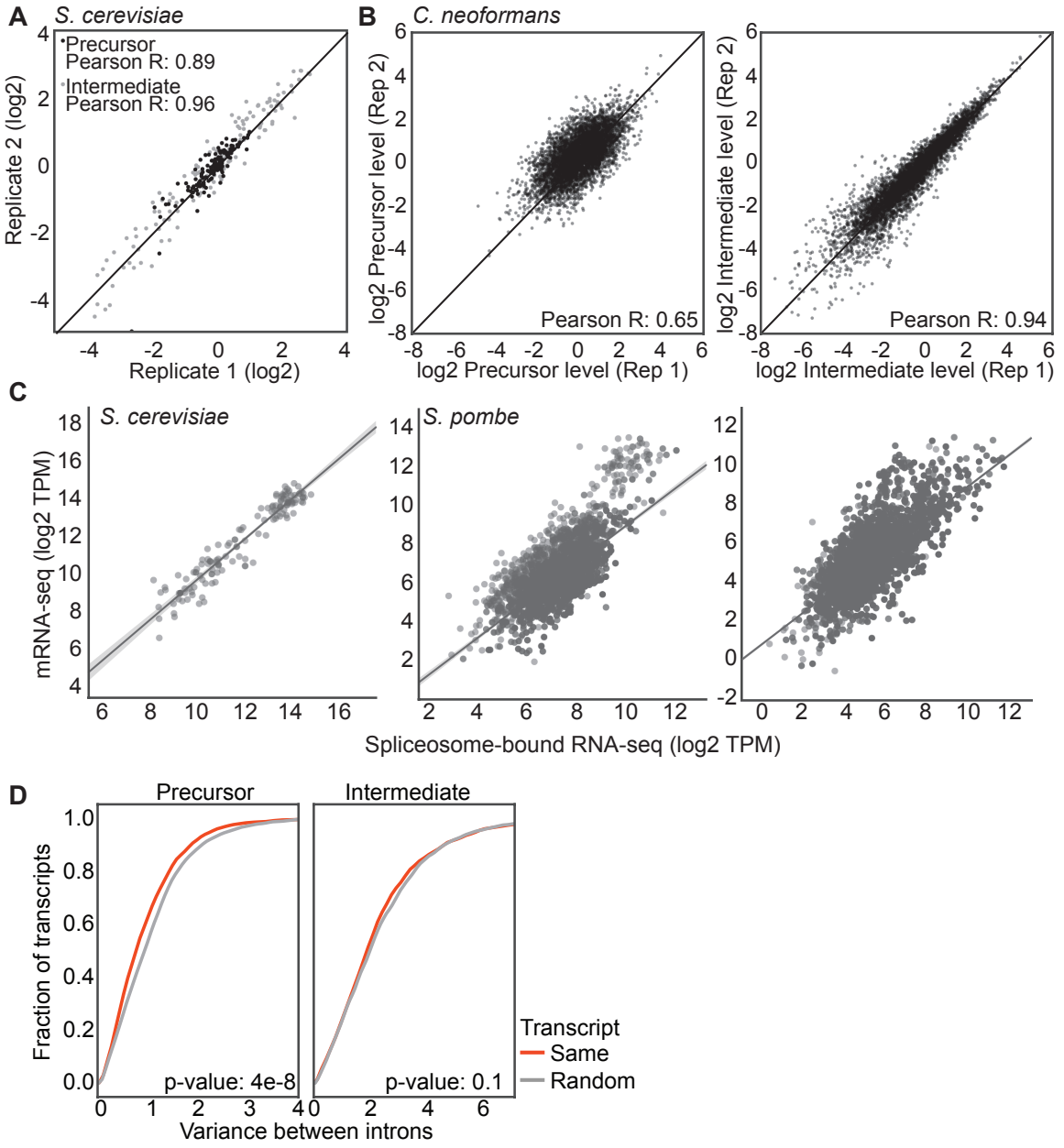


Figure 3.12: Reproducibility and distribution of precursor and intermediate levels, related to Figure 3.5.

Distribution of precursor (dark grey) and intermediate (light grey) levels for *S. cerevisiae* (A) and *C. neoformans* (B). C. Scatter plots of the relationship between polyA RNA-seq and spliceosome-bound RNA-seq for each transcript. D. Variance was calculated for introns in the same transcript (red) in *C. neoformans*. A paired variance was calculated for the same number of introns selected randomly from different transcripts (grey). P-values were calculated using Mann-Whitney U test.

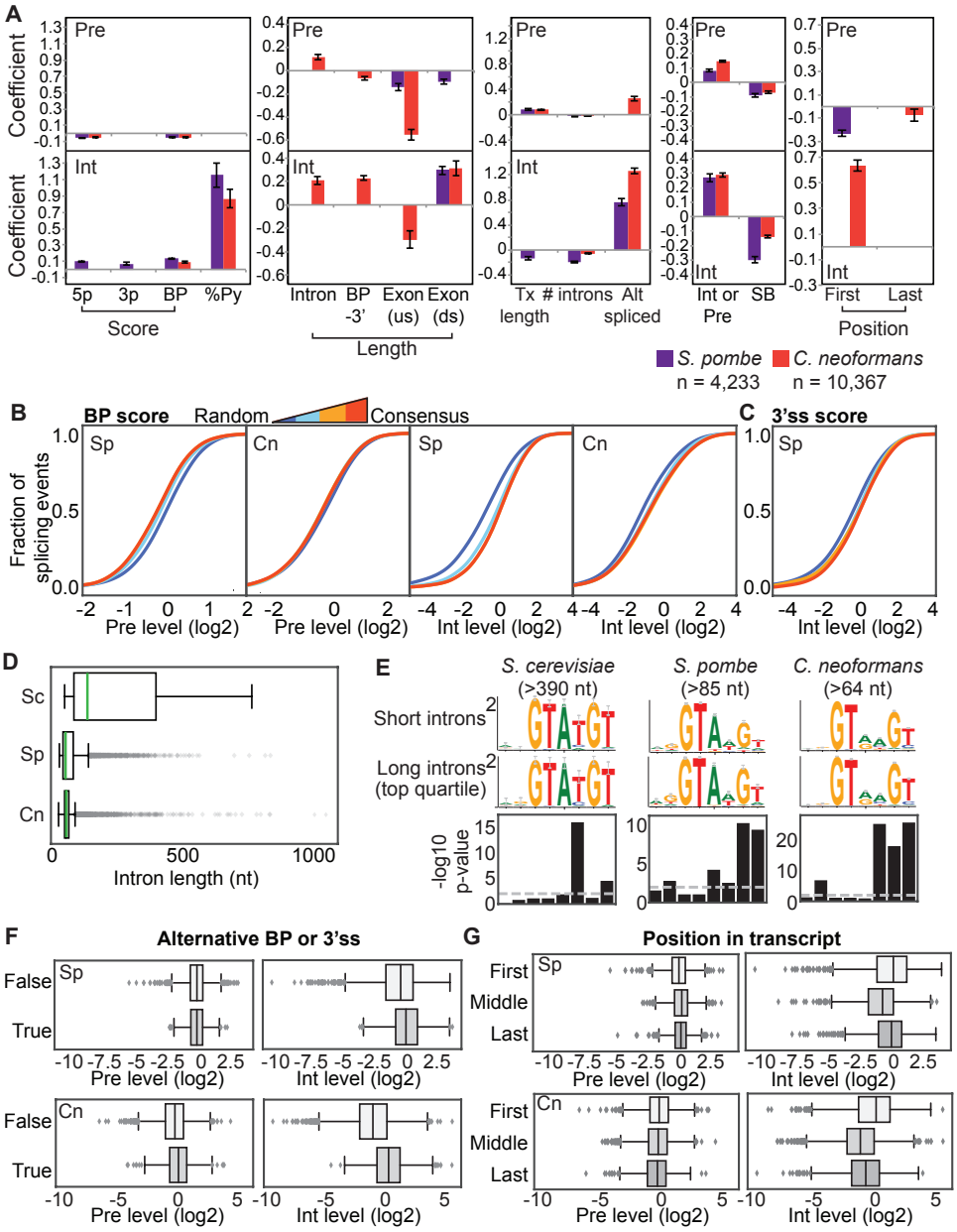


Figure 3.13: Uncovering intron features that predict splicing efficiency, related to Figure 3.6.

A. Bayesian modeling of the correlation between intron features and precursor (Pre) or intermediate (Int) level. Coefficients for explanatory variables with a posterior inclusion probability (PIP) of at least 0.5 are shown. Error bars indicate the uncertainty of each coefficient in the final averaged model. B-C. Relationship between precursor and intermediate levels and BP or 3'ss scores. Introns were split into quartiles based on BP score (blue=random sequence, red=consensus) and the distribution of each metric is plotted as a CDF. D. Differences in intron

length distributions in the three yeasts used in this study. E. Long introns have 5’ splice sites that are closer to consensus. P-values were determined by the chi-squared test for independence and logos were generated with WebLogo3 (Crooks, 2004). F-G. Difference in precursor and intermediate levels based on the whether an alternative BP or 3’ss was detected for an intron and the position of the intron in the transcript.

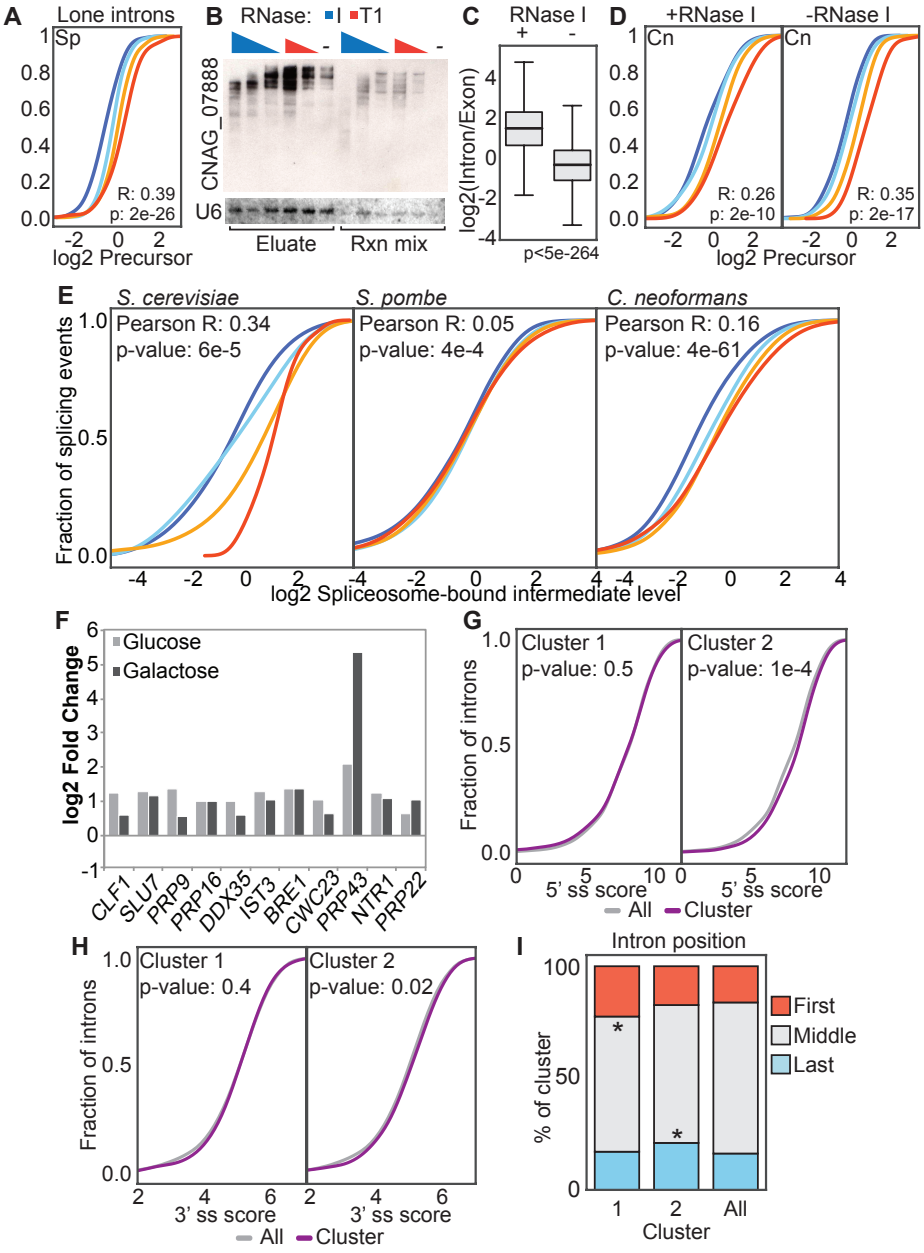


Figure 3.14: Discard of native transcripts, related to Figure 3.7.

A. Relationship between intron retention in polyA mRNA and precursor level in lone introns in *S. pombe* (n=690). Introns were split into quartiles based on the level of intron retention calculated in polyA RNA-seq data. The distribution of precursor level in each quartile is plotted as a CDF. R and p-values were determined by Pearson R test. B. Digestion of spliceosome-bound RNA with RNase. Affinity-purified spliceosomes were treated with 0.2 or 2 units RNase T1 or 0.3, 3 or 30 units RNase I (see Methods). Degradation of the transcript originating from *CNAG_07888* and retention of U6 snRNA were visualized by Northern blot. C. Distribution of the $\log_2$ ratio of read

density in introns to read density in exons after treatment with 30 U RNase I vs mock-treatment. D. Relationship between intron retention in polyA mRNA and precursor level in spliceosomes treated with 30 U RNase I or mock-treated. Introns were split into quartiles based on the level of intron retention calculated in polyA RNA-seq data. The distribution of precursor level in each quartile is plotted as a CDF. R and p-values were determined by Pearson R test. E. Relationship between intron retention in polyA mRNA and intermediate level. Introns were split into quartiles based on the level of intron retention calculated in polyA RNA-seq data. The distribution of intermediate level in each quartile is plotted as a CDF. F. All splicing factors that are significantly differentially expressed in *prp43DN* by polyA RNA-seq are increased (log_2 fold change and significance (adjusted p-value) determined using DESeq2). G-H. 5' and 3' splice site scores as determined by scoring against a PSSM for each cluster (see Methods). P-values determined by Mann-Whitney U test. I. Intron position proportions of each cluster (see Figure 3.7). The proportion of introns that are first, last or in the middle of the transcript are plotted for the clusters with either increased precursor (cluster 1) or increased intermediate (cluster 2) vs. all introns. P-values determined by proportion hypothesis test (*p-value < 0.01).

Tables

Table 3.1

Mass spectrometry results with values and quantitation of snRNAs associated with Prp19, related to Figure 3.1

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.2

Prp19 transcript enrichment for transcripts with and without annotated introns, related to Figure 3.1

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.3

All peaks detected for each organism, related to Figure 3.2

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.4

All detected branches and junctions for *S. pombe* and *C. neoformans*, related to Figure 3.3

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.5

Cross-confirmed peaks for each organism with quantitation, related to Figure 3.3

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.6

Results of Bayesian model averaging, related to Figure 3.6

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.7

Differential expression analysis of polyA RNA-seq for *prp43DN* samples and filtered precursor and intermediate level measurements for *prp43DN*, related to Figure 3.7

DOI https://doi.org/10.7272/Q6T72FP1

Table 3.8

Methods Table

DOI https://doi.org/10.7272/Q6T72FP1

References

1. Akiyama, M., and Nakashima, H. (1996). Molecular cloning of thi-4, a gene necessary for the biosynthesis of thiamine in Neurospora crassa. Curr Genet *30*, 62-67.

2. Amrani, N., Sachs, M.S., and Jacobson, A. (2006). Early nonsense: mRNA decay solves a translational problem. Nat Rev Mol Cell Biol *7*, 415-425.

3. Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data.

4. Aslanzadeh, V., Huang, Y., Sanguinetti, G., and Beggs, J.D. (2017). Transcription Rate Strongly Affects Splicing Fidelity and Co-transcriptionality in Budding Yeast. Genome Res.

5. Awan, A.R., Manfredo, A., and Pleiss, J.A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. Proc Natl Acad Sci U S A *110*, 12762-12767.

6. Bejar, R. (2016). Splicing Factor Mutations in Cancer. Advances in experimental medicine and biology *907*, 215-228.

7. Bessonov, S., Anokhina, M., Krasauskas, A., Golas, M.M., Sander, B., Will, C.L., Urlaub, H., Stark, H., and Luhrmann, R. (2010). Characterization of purified human Bact spliceosomal complexes reveals compositional and morphological changes during spliceosome activation and first step catalysis. RNA *16*, 2384-2403.

8. Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279-290.

9. Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. Trends in biochemical sciences *42*, 407-408.

10. Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F.*, et al.* (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775-788.

11. Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nat Struct Mol Biol *16*, 107-113.

12. Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E., and Graveley, B.R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res *21*, 193-202.

13. Carvalho, A.B., and Clark, A.G. (1999). Intron size and natural selection. Nature *401*, 344.

14. Chapman, K.B., and Boeke, J.D. (1991). Isolation and characterization of the gene encoding yeast debranching enzyme. Cell *65*, 483-492.

15. Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

16. Clark, T.A., Sugnet, C.W., and Ares, M., Jr. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296*, 907-910.

17. Clyde, M. (2017). BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging. R package version 1.4.7.

18. Crooks, G.E. (2004). WebLogo: A Sequence Logo Generator. Genome Res *14*, 1188-1190.

19. Cvitkovic, I., and Jurica, M.S. (2013). Spliceosome database: a tool for tracking components of the spliceosome. Nucleic Acids Res *41*, D132-141.

20. Duff, M.O., Olson, S., Wei, X., Garrett, S.C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S.E., and Graveley, B.R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature *521*, 376-379.

21. Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., and Luhrmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. Mol Cell *36*, 593-608.

22. Fica, S.M., and Nagai, K. (2017). Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. Nat Struct Mol Biol *24*, 791-799.

23. Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.-P., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc Natl Acad Sci U S A *102*, 16176-16181.

24. Garreta, R., and Moncecchi, G. (2013). Learning scikit-learn: Machine Learning in Python (Packt Publishing Ltd).

25. Gonzalez-Hilarion, S., Paulet, D., Lee, K.T., Hon, C.C., Lechat, P., Mogensen, E., Moyrand, F., Proux, C., Barboux, R., Bussotti, G.*, et al.* (2016). Intron retention-dependent gene regulation in Cryptococcus neoformans. Sci Rep *6*, 32252.

26. Gould, G.M., Paggi, J.M., Guo, Y., Phizicky, D.V., Zinshteyn, B., Wang, E.T., Gilbert, W.V., Gifford, D.K., and Burge, C.B. (2016). Identification of new branch points and unconventional introns in Saccharomyces cerevisiae. RNA *22*, 1522-1534.

27. Guo, M., and Mount, S.M. (1995). Localization of sequences required for size-specific splicing of a small Drosophila intron in vitro. J Mol Biol *253*, 426-437.

28. He, L., Diedrich, J., Chu, Y.Y., and Yates, J.R., 3rd (2015). Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. Anal Chem *87*, 11361-11367.

29. Hesselberth, J.R. (2013). Lives that introns lead after splicing. Wiley Interdiscip Rev RNA *4*, 677-691.

30. Hughes, S.S., Buckley, C.O., and Neafsey, D.E. (2008). Complex selection on intron size in Cryptococcus neoformans. Mol Biol Evol *25*, 247-253.

31. Ianiri, G., and Idnurm, A. (2015). Essential gene discovery in the basidiomycete Cryptococcus neoformans for antifungal drug target prioritization. MBio *6*.

32. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc *7*, 1534-1550.

33. Irimia, M., and Roy, S.W. (2008). Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet *4*, e1000148.

34. Kannan, R., Hartnett, S., Voelker, R.B., Berglund, J.A., Staley, J.P., and Baumann, P. (2013). Intronic sequence elements impede exon ligation and trigger a discard pathway that yields functional telomerase RNA in fission yeast. Genes Dev *27*, 627-638.

35. Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods *7*, 1009-1015.

36. Ketchen, D.J., Jr., Ketchen, D.J., Jr., and Shook, C.L. (1996). THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE. Strategic Manage J *17*, 441-458.

37. Killick, R., and Eckley, I.A. (2014). changepoint: AnRPackage for Changepoint Analysis. J Stat Softw *58*.

38. Killick, R., Fearnhead, P., and Eckley, I.A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost, Vol 107.

39. Koodathingal, P., Novak, T., Piccirilli, J.A., and Staley, J.P. (2010). The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5' splice site cleavage during pre-mRNA splicing. Mol Cell *39*, 385-395.

40. Koodathingal, P., and Staley, J.P. (2013). Splicing fidelity: DEAD/H-box ATPases as molecular clocks. RNA Biol *10*, 1073-1079.

41. Krishnan, R., Blanco, M.R., Kahlscheuer, M.L., Abelson, J., Guthrie, C., and Walter, N.G. (2013). Biased Brownian ratcheting leads to pre-mRNA remodeling and capture prior to first-step splicing. Nat Struct Mol Biol *20*, 1450-1457.

42. Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., and Murphy, J.W. (2004). Introns and splicing elements of five diverse fungi. Eukaryot Cell *3*, 1088-1100.

43. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

44. Lesser, C.F., and Guthrie, C. (1993). Mutational analysis of pre-mRNA splicing in Saccharomyces cerevisiae using a sensitive new reporter gene, CUP1. Genetics *133*, 851-863.

45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Data, G.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

46. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600-604.

47. Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A *98*, 11193-11198.

48. Liu, L., Query, C.C., and Konarska, M.M. (2007). Opposing classes of prp8 alleles modulate the transition between the catalytic steps of pre-mRNA splicing. Nat Struct Mol Biol *14*, 519-526.

49. Lucking, R., Huhndorf, S., Pfister, D.H., Plata, E.R., and Lumbsch, H.T. (2009). Fungi evolved right on track. Mycologia *101*, 810-822.

50. Luo, M.L., Zhou, Z., Magni, K., Christoforides, C., Rappsilber, J., Mann, M., and Reed, R. (2001). Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. Nature *413*, 644-647.

51. Luukkonen, B.G., and Séraphin, B. (1997). The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. EMBO J *16*, 779-792.

52. Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol *5*, 89-99.

53. Martin, A., Schneider, S., and Schwer, B. (2002). Prp43 is an essential RNA-dependent ATPase required for release of lariat-intron from the spliceosome. J Biol Chem *277*, 17743-17750.

54. Mayas, R.M., Maita, H., Semlow, D.R., and Staley, J.P. (2010). Spliceosome discards intermediates via the DEAH box ATPase Prp43p. Proc Natl Acad Sci U S A *107*, 10020-10025.

55. Mayas, R.M., Maita, H., and Staley, J.P. (2006). Exon ligation is proofread by the DExD/H-box ATPase Prp22p. Nat Struct Mol Biol *13*, 482-490.

56. Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

57. Mayerle, M., Raghavan, M., Ledoux, S., Price, A., Stepankiw, N., Hadjivassiliou, H., Moehle, E.A., Mendoza, S.D., Pleiss, J.A., Guthrie, C.*, et al.* (2017). Structural toggle in the RNaseH domain of Prp8 helps balance splicing fidelity and catalytic efficiency. Proc Natl Acad Sci U S A *114*, 4739-4744.

58. McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P., Wuhr, M., Huttlin, E.L., Erickson, B.K., Rad, R., Haas, W., and Gygi, S.P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem *86*, 7150-7158.

59. Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526-540.

60. Oesterreich, F.C., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372-381.

61. Ohrt, T., Odenwalder, P., Dannenberg, J., Prior, M., Warkocki, Z., Schmitzova, J., Karaduman, R., Gregor, I., Enderlein, J., Fabrizio, P.*, et al.* (2013). Molecular dissection of step 2 catalysis of yeast pre-mRNA splicing investigated in a purified system. RNA *19*, 902-915.

62. Pang, A.S., Nathoo, S., and Wong, S.L. (1991). Cloning and characterization of a pair of novel genes that regulate production of extracellular enzymes in Bacillus subtilis. J Bacteriol *173*, 46-54.

63. Park, S.K., Aslanian, A., McClatchy, D.B., Han, X., Shah, H., Singh, M., Rauniyar, N., Moresco, J.J., Pinto, A.F., Diedrich, J.K.*, et al.* (2014). Census 2: isobaric labeling data analysis. Bioinformatics *30*, 2208-2209.

64. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res *2*, 43-50.

65. Prieto, M., and Wedin, M. (2013). Dating the diversification of the major lineages of Ascomycota (Fungi). PLoS One *8*, e65576.

66. Qin, D., Huang, L., Wlodaver, A., Andrade, J., and Staley, J.P. (2016). Sequencing of lariat termini in S. cerevisiae reveals 5' splice sites, branch points, and novel splicing events. RNA *22*, 237-253.

67. Query, C.C., and Konarska, M.M. (2004). Suppression of multiple substrate mutations by spliceosomal prp8 alleles suggests functional correlations with ribosomal ambiguity mutants. Mol Cell *14*, 343-354.

68. Ren, L., McLean, J.R., Hazbun, T.R., Fields, S., Vander Kooi, C., Ohi, M.D., and Gould, K.L. (2011). Systematic two-hybrid and comparative proteomic analyses reveal novel yeast pre-mRNA splicing factors connected to Prp19. PLoS One *6*, e16719.

69. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

70. Ruskin, B., and Green, M.R. (1985). An RNA processing activity that debranches RNA lariats. Science *229*, 135-140.

71. Schwer, B. (2008). A conformational rearrangement in the spliceosome sets the stage for Prp22-dependent mRNA release. Mol Cell *30*, 743-754.

72. Semlow, D.R., Blanco, M.R., Walter, N.G., and Staley, J.P. (2016). Spliceosomal DEAH-Box ATPases Remodel Pre-mRNA to Activate Alternative Splice Sites. Cell *164*, 985-998.

73. Shen, S., Park, J.W., Lu, Z.-X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A *111*, E5593-5601.

74. Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat Rev Mol Cell Biol *18*, 655-670.

75. Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J.*, et al.* (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371-375.

76. Slabodnick, M.M., Ruby, J.G., Reiff, S.B., Swart, E.C., Gosai, S., Prabakaran, S., Witkowska, E., Larue, G.E., Fisher, S., Freeman, R.M., Jr.*, et al.* (2017). The Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell. Curr Biol *27*, 569-575.

77. Sorenson, M.R., Jha, D.K., Ucles, S.A., Flood, D.M., Strahl, B.D., Stevens, S.W., and Kress, T.L. (2016). Histone H3K36 methylation regulates pre-mRNA splicing in Saccharomyces cerevisiae. RNA Biol *13*, 412-426.

78. Suzuki, H., Kameyama, T., Ohe, K., Tsukahara, T., and Mayeda, A. (2013). Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. FEBS Lett *587*, 555-561.

79. Suzuki, H., Zuo, Y., Wang, J., Zhang, M.Q., Malhotra, A., and Mayeda, A. (2006). Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. Nucleic Acids Res *34*, e63.

80. Tabb, D.L., McDonald, W.H., and Yates, J.R., 3rd (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res *1*, 21-26.

81. Taggart, A.J., Lin, C.L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. Genome Res *27*, 639-649.

82. Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

83. Tsai, R.T., Fu, R.H., Yeh, F.L., Tseng, C.K., Lin, Y.C., Huang, Y.H., and Cheng, S.C. (2005). Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2. Genes Dev *19*, 2991-3003.

84. Tseng, C.K., Liu, H.L., and Cheng, S.C. (2011). DEAH-box ATPase Prp16 has dual roles in remodeling of the spliceosome in catalytic steps. RNA *17*, 145-154.

85. Ulfendahl, P.J., Pettersson, U., and Akusjarvi, G. (1985). Splicing of the adenovirus-2 E1A 13S mRNA requires a minimal intron length and specific intron signals. Nucleic Acids Res *13*, 6299-6315.

86. Volanakis, A., Passoni, M., Hector, R.D., Shah, S., Kilchert, C., Granneman, S., and Vasiljeva, L. (2013). Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. Genes Dev *27*, 2025-2038.

87. Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

88. Wiegand, H.L., Lu, S., and Cullen, B.R. (2003). Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. Proc Natl Acad Sci U S A *100*, 11327-11332.

89. Wiley, D.J., Catanuto, P., Fontanesi, F., Rios, C., Sanchez, N., Barrientos, A., and Verde, F. (2008). Bot1p is required for mitochondrial translation, respiratory function, and normal cell morphology in the fission yeast Schizosaccharomyces pombe. Eukaryot Cell *7*, 619-629.

90. Xu, T., Park, S.K., Venable, J.D., Wohlschlegel, J.A., Diedrich, J.K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X.*, et al.* (2015). ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J Proteomics *129*, 16-24.

91. Zhang, X., and Schwer, B. (1997). Functional and physical interaction between the yeast splicing factors Slu7 and Prp18. Nucleic Acids Res *25*, 2146-2152.

92. Zhou, Z., Luo, M.J., Straesser, K., Katahira, J., Hurt, E., and Reed, R. (2000). The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature *407*, 401-405.

Chapter 4

Coupling of spliceosome complexity to intron diversity

Jade Sales-Lee[1], Daniela S. Perry[1], Bradley A. Bowser[2], Jolene K. Diedrich[3], Beiduo Rao[1], Irene Beusch[1], John R. Yates III[3], Scott W. Roy[4,6], and Hiten D. Madhani[1,6,7]

[1] Dept. of Biochemistry and Biophysics, University of California – San Francisco, San Francisco, CA 94158

[2] Dept. of Molecular and Cellular Biology, University of California - Merced, Merced, CA 95343

[3] Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA 92037

[4] Dept. of Biology, San Francisco State University, San Francisco, CA 94132

[5] Chan-Zuckerberg Biohub, San Francisco, CA 94158

[6] Corresponding authors: scottroy@gmail.com, hitenmadhani@gmail.com

[7] Lead Contact

Abstract

We determined that over 40 spliceosomal proteins are conserved between many fungal species and humans but were lost during the evolution of *S. cerevisiae*, an intron-poor yeast with unusually rigid splicing signals. We analyzed null mutations in a subset of these factors, most of which had not been investigated previously, in the intron-rich yeast *Cryptococcus neoformans.* We found they govern splicing efficiency of introns with divergent spacing between intron elements. Importantly, most of these factors also suppress usage of weak nearby cryptic/alternative splice sites. Among these, orthologs of GPATCH1 and the helicase DHX35 display correlated functional signatures and copurify with each other as well as components of catalytically active spliceosomes, identifying a conserved G-patch/helicase pair that promotes splicing fidelity. We propose that a significant fraction of spliceosomal proteins in humans and most eukaryotes are involved in limiting splicing errors, potentially through kinetic proofreading mechanisms, thereby enabling greater intron diversity.

Introduction

The splicing of mRNA precursors on the spliceosome is a signature feature of eukaryotic gene expression (Sharp, 1987). Splicing plays numerous critical regulatory roles in organisms as diverse as the unicellular budding yeast *Saccharomyces cerevisiae*, apicomplexan parasites, plants, and humans. Splicing also plays key roles in ncRNA biogenesis (Ruby et al., 2007), RNA export (Zhou et al., 2000), nonsense-mediated decay (Ishigaki et al., 2001; Le Hir et al., 2001), and genome defense (Dumesic et al., 2013). The spliceosome is a complex and dynamic assembly of small nuclear ribonucleoproteins (snRNPs) and proteins that assemble onto the intron substrate and then undergo several large rearrangements to form a catalytically active complex (Wilkinson et al., 2020). Two sequential transesterification steps mediate intron removal. Pre-mRNA splicing by the spliceosome seems surprisingly complex for a process that removes a segment of RNA from a precursor. Splicing requires eight ATP-dependent steps and about 90 proteins in *S. cerevisiae*. Much of our functional understanding of spliceosome components derives from the analysis of conditional and null mutants in *S. cerevisiae* (Wilkinson et al., 2020). Human spliceosomes appear to contain about 60 additional proteins (Wahl and Luhrmann, 2015a, b). The reason for this added complexity is not well-understood.

The atomic structures of the rigid core portion of the spliceosome at various stages of its cycle have been elucidated using cryoEM (Wilkinson et al., 2020). Nearly all structures have been obtained using *in vitro*-assembled spliceosomes using extracts from the budding yeast *S. cerevisiae* or from HeLa cells. While these structures have revealed that the catalytic core of the spliceosome is invariant across divergent species, proteins and structures have been identified in human spliceosomes that are not found in *S. cerevisiae* spliceosomes. While it might be imagined that the higher complexity of human spliceosomes relates to late evolutionary innovations that enabled metazoan complexity, an alternative model is that the common ancestor of *S. cerevisiae* and humans harbored a complex spliceosome, whose

components were lost during the evolution of *S. cerevisiae*. There is anecdotal support for this hypothesis. For example, several orthologs of human splicing factors that do not exist in *S. cerevisiae* have been described in the fission yeast *Schizosaccharomyces pombe* (Chen et al., 2014; Cipakova et al., 2019; McDonald et al., 1999). Prior work indicates that the *Saccharomycotina,* the subphylum to which *S. cerevisiae* belongs, has lost introns that were present in an intron-rich ancestor, such that less than ten percent of genes harbor introns in *S. cerevisiae* (Irimia et al., 2007*)*. As in other lineages, such loss events correlate with intron signals moving towards optimal intron signals. Thus, as introns are lost, intron signals become homogenous and lose diversity. Insofar as certain splicing factors play outsized roles in recognition of introns with divergent splice signals, such homogenization might be expected to be associated with loss of spliceosomal factors and thus overall spliceosomal simplification.

We have examined evolution of human spliceosomal protein orthologs in fungi and find that most fungal lineages, including that of the genetically tractable haploid yeast, *Cryptococcus neoformans*, encode orthologs of most human spliceosomal proteins, indicating a complex ancestral spliceosome and substantial simplification through protein loss during the evolution of *S. cerevisiae*. For many of these proteins, we have obtained viable gene deletion strains of *C. neoformans* lacking an ortholog of one of these factors. Our subsequent functional studies reveal that these factors generally promote the efficiency of splicing of intron subsets that diverge in size, branchpoint-to-3'-splice site distance and intron position from average introns. Most also influence splicing decisions with a bias towards suppressing the utilization of weak/ cryptic nearby 5' and/or 3' splice sites, indicating a role in splicing fidelity. Cryptococcal orthologs of two human proteins, GPATCH1 and DHX35, show highly correlated splicing signatures. Affinity purification-mass spectrometry (AP-MS) identification of GPATCH1-associated proteins revealed strong copurification with DHX35 as well as components characteristic of catalytically active spliceosomes. GPATCH1 and DHX35 suppress the usage of weak/cryptic nearby 5' and 3' splice sites, forming a G-patch/helicase pair governing

spliceosomal accuracy. Orthologs of two other human G-patch-containing proteins, RBM5 and RBM17, also suppress the usage of weak nearby alternative splice sites. Their co-purified proteins indicate that they act at earlier stages of spliceosome assembly to inspect substrates. We propose that the complexity of the spliceosome enables the use of diverse introns while promoting fidelity.

Results

Maintenance of many dozens of human spliceosomal orthologs in fungal lineages

Like the model yeasts, *S. cerevisiae* and *S. pombe*, *C. neoformans* offers a genetically tractable haploid organism in which to investigate fundamental aspects of gene expression. To emphasize the evolutionary diversity of introns in fungi, we highlight the differences between the intron sequences and abundance between *Saccharomyces cerevisiae*, *Cryptococcus neoformans* and *Homo sapiens* in Figure 4.1A. The *S. cerevisiae* genome is estimated to encode only 282 introns (Grate and Ares, 2002) spread over 5410 annotated genes (0.05 introns/gene), while *Cryptococcus neoformans* (H99 strain) has 6941 annotated protein coding genes harboring over 40,000 introns (Janbon et al., 2014; Loftus et al., 2005), comparable to humans (8 introns/gene), with 27,219 annotated genes and over 200,000 introns (Piovesan et al., 2019). Note that the sequences of *C. neoformans* 5' splice sites and branchpoints are considerably more variable than those of *S. cerevisiae*, suggesting its spliceosomes, like those of humans, may be more flexible in terms of substrate utilization (Figure 4.1A).

We asked whether the loss of introns in the *Saccharomycotina*, the subphylum in which *S. cerevisiae* belongs, is accompanied with a loss of spliceosomal protein orthologs. We compiled a list of all spliceosome components reproducibly detected through mass

spectrometry, interaction studies and/or purified and visualized in the spliceosome in structural biology studies (Cvitkovic and Jurica, 2013; Wahl and Luhrmann, 2015a, b). This list includes 157 human proteins (Table 4.1). To identify candidates for fungal orthologs we used a combination of criteria including reciprocal BLASTP searches and the presence of predicted protein domains, followed by the application of additional criteria. Because ortholog identification is not an unambiguous exercise, we generated a confidence score (0-10) for the presence of an ortholog in given species (see Methods). Using this semi-automated process, we analyzed 24 fungal species with at least two representatives from each major clade (Figure 4.1B, left panel). We then plotted the number of proteins for which an ortholog to a human spliceosomal protein could be identified at a given confidence level in each species (Figure 4.1B, right panel). Strikingly, members of the intron-reduced *Saccharomycotina* harbored the fewest strong human spliceosomal protein orthologs. Other species exhibited considerably larger numbers of human spliceosomal orthologs, including *C. neoformans* (Figure 4.1B right panel). Because members of the most early branching groups analyzed harbor the highest number human spliceosomal protein orthologs, clades displaying lower numbers of orthologs have most likely undergone gene loss events, with the *Saccharomycotina* exhibiting the highest degree of loss. This correlates with the reduction in intron number found in species of this group (Irimia and Roy, 2008). For three fungal species of interest (*S. cerevisiae*, *S. pombe* and *C. neoformans*), we performed detailed manual curation of the spliceosome using the literature (including our past studies of purified Cryptococcal spliceosomes – Burke et al, 2018) and an available experimentally curated database (Cvitkovic and Jurica, 2013). Nine proteins in *S. cerevisiae* and one protein in *S. pombe* are included in the curation based on the literature despite the fact they display insufficient sequence identity with the presumptive human ortholog to be detected bioinformatically. This analysis revealed 94 spliceosomal protein orthologs in *S. cerevisiae*, 126 in *S. pombe* and 139 in *C. neoformans* (Figure 4.1C and Table 4.1).

Thus, some 45 genes encoding predicted human spliceosomal orthologs are present in *C. neoformans* but not in *S. cerevisiae.* To investigate these spliceosomal proteins, we searched for viable knockout mutants in these factors in a gene deletion collection for *C. neoformans* constructed in our laboratory and identified strains deleted in 13 of these putative spliceosomal factors. We also identified a strain harboring a deletion of an ortholog of the human spliceosomal protein DHX35, which is found in *S. cerevisiae* (Dhr2) but appears to not be involved in splicing but instead nucleolar ribosomal RNA processing (Colley et al., 2000). We identified the Cryptococcal ortholog of DHX35 previously in purified *C. neoformans* spliceosomes (Burke et al., 2018), and thus included it in this study. The names and confidence scores in fungi of these 14 human spliceosomal proteins are displayed in Figure 4.1D. For readability, we will use the human nomenclature for Cryptococcal spliceosome proteins throughout the manuscript (see Table 4.1 for *C. neoformans* gene locus and name). We also identified a viable gene deletion corresponding to Rrp6, a nuclear exosome subunit, involved in RNA degradation and quality control, whose loss we hypothesized might stabilize RNAs produced by aberrant splicing events compared to wild-type cells.

To examine the impact of these 14 gene deletion mutations on the abundance of pre-mRNA and mRNA along with splice site choice, we cultured these strains, extracted RNA, purified polyadenylated transcripts, and performed RNA-seq. To maximize the quality of the data, the samples were grown in duplicate and paired with duplicate wild-type samples grown on the same day to the same optical density. In addition, paired-end 100 nt reads were obtained at a minimum depth of 12M reads/sample.

Limited impact of spliceosomal protein mutations on global transcript abundance

We first sought to determine whether deletions of putative spliceosomal proteins altered the transcript levels of other spliceosomal proteins. Hence, we subjected RNA-seq reads to mapping and applied DESeq2 to identify changes in transcript levels (Love et al., 2014). Shown in Figure 4.2D is the impact of gene deletions on the levels of spliceosomal protein-encoding transcripts (see Table 4.2 for full results). Among the deletions analyzed only one displayed a significant change (>2 fold change, adjusted p-value<0.01) in the transcript levels of a spliceosome-encoding protein. This strain is deleted for *CNAG_02260,* which encodes the Cryptococcal ortholog of FAM50A. While FAM50A is a spliceosomal protein in humans, it has also been linked to transcription (Kim et al., 2018) suggesting a pleiotropic role. Consistent with this, we observed that many genes display transcript level changes in this mutant, while few global transcript changes were observed for the other gene deletion strains, save for the *rrp6*Δ strain, which increased the levels of ~250 mRNAs, consistent with its predicted role in nuclear RNA turnover (Figure 4.2D). Thus, mutations in putative spliceosomal factors analyzed here do not generally appear to have large effects on the expression of other spliceosomal factors, suggesting that effects on splicing in the corresponding *C. neoformans* mutants likely reflect direct roles. We therefore proceeded to analyze the impact of mutations on splicing.

Altered splicing choice and efficiency in mutants lacking human spliceosomal protein orthologs

Prior RNA-seq analysis of splicing in wild-type *C. neoformans* strains suggests that intron retention is the major form of alternative splicing and can be altered in response to changes in environmental conditions. However, alternative 5’ and 3’ splice sites are also observed in RNA-seq data. The extent to which isoforms have distinct functions is unknown. To

examine splicing changes (Fig 2A-C), we chose to use the Junctional Utilization Method [JUM; (Wang and Rio, 2018)], a high-performing annotation-free approach for measuring splicing changes in RNA-seq data that is also optimized for measuring intron retention (a surrogate for splicing efficiency). Using a stringent read-count and p-value cutoffs (see Methods), we quantified splicing changes in each of the 14 gene deletion mutants described above. Since we did not identify any instances of mutually exclusive exons and only a small handful of cassette exons, we excluded these two categories, along with the 'complex splicing' category, from our downstream analysis (see Methods).

As diagrammed in Figure 4.2B and 4.2C, analysis of intron retention, alternative 5' splice site usage, and alternative 3' splice site usage involves multiple possibilities for a mutant phenotype. For intron retention, the amount of retained intron transcripts (i.e., precursor) can be increased or decreased relative to mRNA. In the metric change to "Percent Splicing In" ($\Delta\psi$) a *positive* value corresponds to an *increase* intron retention in a mutant, while a *decrease* in intron retention produces a *negative* $\Delta\psi$ value (Figure 4.2A-C). For changes in the relative use of a splice site relative to an alternative splice site, we first determined that the site preferred in wild-type cells (>50% usage relative to the alternative site) was always an annotated splice site, while the alternative site was either unannotated or annotated as an alternative site in the current *C. neoformans* H99 strain genome annotation. Whether either was proximal or distal relative to the fixed splice site was not considered. For alternative 5' or 3' splice site changes, a *decrease* of usage of the preferred site in mutant produces a *negative* $\Delta\psi$, while an *increase* in the usage of the wild-type site in a mutant produces a *positive* $\Delta\psi$ value (Figure 4.2B-C).

For each mutant, we quantified the effects across alternative splicing events in the *C. neoformans* genome and tabulated this data across events. The results of this analysis are shown in Figure 4.3A (full dataset available in Table 4.3). Plotted are the number of introns impacted in each gene deletion mutant for each of the three types of splicing changes. The numbers plotted above the line indicate the number of introns whose splicing is altered in a such

a way to produce a positive $\Delta\psi$ value as defined above, while those plotted below the line represent the number of introns impacted for a given splicing type that produce a negative $\Delta\psi$ value as defined above. Note that because of the stringent criteria imposed on the data and limited sensitivity of RNA-seq for rare transcript classes (e.g. unspliced pre-mRNA), these numbers are likely to be substantial underestimates. We observed the largest numbers of affected introns in the intron-retention category, and the fewest in the alternative 5' splice site category.

It appeared that many of the mutants were biased towards a negative $\Delta\psi$ for 3' and 5' splice site choice, indicating a decrease in the use of the canonical splice site in the mutant (and therefore an increase in the use of an alternative site). Likewise, for intron retention, several mutants appeared to be biased towards increasing intron retention, consistent with increased splicing defects (increased pre-mRNA vs. mRNA). To test the statistical significance of these apparent skews, we used the binomial distribution to model the null hypothesis. As can be seen in Figure 4.3B, nine deletion mutants displayed statistically significant bias towards decreased usage of the canonical site (and therefore increased use of an alternative site) for 3' splice site usage. These correspond to strains lacking orthologs of human FAM32A, RBM5, RBM17, GPATCH1, FAM50A, NOSIP, IK, DHX35 and SAP18 (note that in humans RBM5 and RBM10 are paralogs; we refer to the Cryptococcal ortholog as RBM5 for simplicity). Curiously, a mutant for ZNF830, a human spliceosomal protein of unknown function, displayed a bias towards increased use of the canonical 3' splice site. For alternative 5' splice site usage, we observed a similar pattern, with cells lacking orthologs of GPATCH1, NOSIP, and DHX35 displaying a bias towards decreased use of the canonical 5' splice site and increased use of an alternative 5' splice site in the mutant (Figure 4.3C). Again, cells lacking an ortholog of ZNF830 displayed the opposite bias. Finally, five mutants displayed a bias towards an increase in intron retention in the mutant (DHX35, GPATCH1, RBM5, SAP18 and RBM17) suggesting a role in splicing efficiency for a subset of transcripts (Figure 4.3D). Unexpectedly, strains lacking orthologs of

NOSIP, IK, FAM50A, and FRA10AC1, human spliceosomal protein of unknown function, display reduced intron retention in the mutant, indicating that their absence results in increased splicing efficiency for a subset of transcripts, an unexpected phenotype (Figure 4.3D). This initial analysis reveals categories of splicing changes produced by mutations in the factors analyzed and statistically significant biases in the directionality of these changes.

Clustering of gene deletions based on splicing changes suggests some factors act together

To identify candidates for spliceosomal factors that might act together, we calculated correlations of splicing effects for each pair of factors. Specifically, we calculated vectors of $\log_{10}$ corrected P-values (produced by JUM's linear model approach) for each of the ~40,000 *C. neoformans* introns, with nonsignificant p-values corrected to 1; and then calculated correlations for these vectors for each pair of mutants Figure 4.4 displays these correlation matrices for three types of splicing events using a Pearson correlation as the distance metric. Strikingly, we observed that GPATCH1 and DHX35 were nearest neighbors in all three matrices, suggesting they consistently impacted overlapping intron sets. We also noticed that RBM5 and RBM17 were also nearest neighbors in the autocorrelation matrix for the alternative 5' splice site usage data and the intron retention data (Figure 4.4B-C). Human GPATCH1 and DHX35 have both been identified in C complex spliceosomes assembled *in vitro* (Ilagan et al., 2013), while RBM5 and RBM17 have been found in the early A complex that includes U2 snRNP (Hartmuth et al., 2002). Loss of the nuclear exosome factor Rrp6 produced a signature that tended to cluster adjacent to that of strains lacking the ortholog of CTNNBL1 (Figure 4.3A-C), a core component of active human spliceosomes recently visualized by cryoEM (Townsend et al., 2020), indicating an overlap between RNA species normally degraded by Rrp6 and those that accumulate in cells

lacking CTNNBL1. Other mutants also showed some degree of clustering, suggesting functional/biochemical relationships.

GPATCH1 and DHX35 as well as RBM5 and RBM17 associate in spliceosomes

The genetic data above together with existing data on the association of the human orthologs suggests that GPATCH1 and DHX35 might act together. This would require for them to be present in the same spliceosomal complex(es). To test this hypothesis, we generated a FLAG-tagged allele of GPATCH1 and performed immunoprecipitation (IP) of an untagged strain and of the tagged strain under low and high-salt conditions (four IPs total). To quantify the proteins in the coimmunoprecipitated material we performed tandem mass tag (TMT) mass spectrometry analysis (Figure 4.5A). We then ranked proteins based on relative peptide counts/ protein length for all identified proteins. Remarkably, the next most abundant protein in the GPATCH1 IP was DHX35 (Figure 4.5B). Numerous additional spliceosomal proteins were identified, including those characteristic of active C complex spliceosomes (Figure 4.5B), suggesting that, as in human cells, GPATCH1 and DHX35 associate with active spliceosomes in *C. neoformans*. The full mass spectrometry dataset can be found in Table 4.4, which includes raw and annotated data.

We also performed parallel IP experiments with RBM5 and RBM17 (eight additional purifications), as they also harbor a G-patch domain and displayed clustering in their functional signatures. These proteins displayed different associated proteins. RBM5 associated with components of the U2 snRNP including DDX46 (*S. cerevisiae* (*Sc.*) Prp5), SF3A3 (*Sc.* Prp9) and SF3A2 (*Sc.* Prp11) along with SF3B complex proteins (Figure 4.5C), consistent with its association with A complex spliceosomes during *in vitro* splicing reactions (Hartmuth et al., 2002). RBM17 has been found to associate with U2SURP and CHERP in IP-MS studies from

human cells (De Maio et al., 2018). Strikingly, we found that purification of *C. neoformans* RBM17 identified U2SURP as the most abundant coimmunoprecipitating protein (Figure 4.5B), indicating evolutionary conservation of this association. We also identified peptides corresponding to RBM5 (Figure 4.5D), consistent with their clustering in the autocorrelation matrix based on the RNA-seq data described above. The full mass spectrometry datasets for the RBM5 and RBM17 purifications can be found in Tables 4.S5 and 4.S6. Taken together, these data indicate that the clustering of factors based on their impact on splicing choice and efficiency is a useful way to begin to understand their biochemical relationships in the spliceosome.

Identification of intron features that correlate with sensitivity to dependence on specific factors

To investigate why some introns are sensitive to loss of the spliceosomal protein orthologs described above, we tested whether 5' splice site strength, predicted branchpoint strength (see Methods), or 3' splice site strength was distinct for introns affected in each of the mutants studied. These studies identified only weak or marginal effects. Next, we investigated intron geometry. Specifically, we asked whether the intron length distributions of affected versus unaffected introns differed (as determined by a corrected Wilcoxon Rank Sum Test) for a given mutant and splicing type. We performed the same for the number of intronic nucleotides between the predicted branchpoint and the 3' splice site. All mutants that impacted the splicing of introns skewed significantly towards affecting introns with longer lengths (Figure 4.6A; clustered heatmap of corrected p-values is shown on the left panel and top three mutants/ splicing types are shown in cumulative density plots on the right). The impact was strongest for intron retention changes (Figure 4.6A). Differences in branchpoint-to-3' splice site distance [both increases and decreases; denoted by (+) and (-)] were most notable of introns affected for

intron retention for RBM17, CCDC12, FAM32A, and FAM50A. FAM32A has been identified a "metazoan-specific" alternative step 2 factor in human spliceosome cryoEM structures that promotes the splicing *in vitro* of an adenovirus substrate, harboring a relatively short branchpoint-to-3' splice site distance (Fica et al., 2019). The analysis described here suggests it also limits the splicing of longer introns as well as those with nonoptimal branchpoint-3' splice sites *in vivo*. These analyses indicate that intron size and branchpoint-3' splice site distance are significant determinants of sensitivity to loss of the spliceosomal proteins investigated here.

Mutants result in activation of weak alternative 5' and/or 3' splice sites

As many mutants that we examined were found to trigger reduced use of the canonical 5' or 3' splice site and a shift toward an alternative 5' or 3' splice site, we asked whether the corresponding splice site sequence differed between the canonical and alternative sites. To accomplish this, we examined the frequency of each of the four bases at the first six and last six position of each intron for the canonical versus alternative 5' or 3' splice site. We tested whether the nucleotide biases of the canonical versus alternative site were significantly different at a given position for a given gene deletion using a corrected Chi-squared test. Plotted in Figure 4.7A are the results ($-\log_{10}$P value) for the first six nucleotides of the intron for the cases of alternative 5' slice site usage. We observed highly significant differences at many nucleotides depending on the mutant, particularly, positions 4-6 of the 5' splice site, which normally base-pair with U6 snRNA in the spliceosome (Figure 4.7A). For introns displaying alternative 3' splice site usage in the mutants, we observed significant deviation between the canonical and alternative site primarily at position -3, which is typically a pyrimidine. To examine the mutants in greater detailed, we generated sequence logo plots of the canonical and alternative sites. Shown in Figure 4.7C-D are those for the introns displaying decreased use of the canonical site for mutants in the three G-patch proteins analyzed above as well as DHX35. Remarkably, the

alternative splice site is consistently considerably weaker than the canonical and in many cases lacking conservation at key intronic positions (e.g. positions 5 and 6 of the 5' splice site or -3 of the 3' splice site). We consistently observed similar patterns in mutant of the other factors. We conclude the spliceosomal proteins investigated here display a functional bias towards limiting the use of weak/alternative sites, thereby enhancing the precision of pre-mRNA splicing.

Discussion

Our work defines a large group of spliceosomal proteins conserved between fungi and humans that enable the splicing of divergent introns while promoting fidelity. Most of these proteins have not been investigated functionally *in vivo* in any system. These factors are not essential for splicing *per se* as they were lost in large numbers during the evolution of intron-reduced species. Nonetheless, they have been conserved at least since the evolutionary divergence of fungi and humans several hundred million years ago. In the cases investigated by immunoprecipitation and mass spectrometry, factors display biochemical interactions in *C. neoformans* that are similar to those of their human orthologs, suggesting conserved functional roles in pre-mRNA splicing. This is important because orthologs of several of the proteins investigated here are involved in disease. Mutations in FAM50A cause Armfield X-linked Intellectual Disability (Lee et al., 2020). RBM17 has been strongly implicated in the pathogenesis of Spinocerebellar Ataxia type 1 (De Maio et al., 2018; Lai et al., 2011; Lim et al., 2008; Tan et al., 2016). In addition, CXorf56 has been shown to be mutated in another inherited X-linked intellectual disability syndrome (Rocha et al., 2020; Verkerk et al., 2018). Mutations in the RBM5 paralog RBM10 cause TARP syndrome, an X-linked congenital pleiotropic developmental syndrome (Johnston et al., 2010). Finally, an inherited mutation in CTNNBL1 causes common variable immunodeficiency associated with autoimmune cytopenia (Kuhny et

al., 2020). The findings discussed below provide a resource for understanding the underlying molecular pathologies of these diseases.

Massive evolutionary loss of spliceosomal proteins in the Saccharomycotina

Prior experimental work has shown that *S. cerevisiae* spliceosomes are not very tolerant of mutations of intronic sequences away from consensus (Lesser and Guthrie, 1993), with kinetic proofreading by ATPases Prp16 (human DHX38) and Prp22 (human DHX8) limiting the splicing of mutant pre-mRNAs via discard and disassembly of substrates with kinetic defects during the catalytic stages of splicing (Koodathingal and Staley, 2013). How the spliceosomes of organisms tolerate diversity in intron splicing signals and geometries is not understood. We reasoned that spliceosomal proteins whose genes were lost during evolution of organisms undergoing intron loss/homogenization might correspond to factors and processes that promote the use of divergent introns. Our analysis suggests orthologs of about a third of human spliceosomal proteins cannot be identified in *S. cerevisiae*. However, most of these are maintained in other fungal lineages. We focused our attention on *C. neoformans*, an experimentally tractable haploid yeast whose genome is intron-rich and well-annotated. Our analysis revealed 45 genes in *C. neoformans* that encode orthologs of human spliceosomal proteins that do not appear in the *S. cerevisiae* genome. Of these, we identified 13 for which deletion alleles had been generated as part of a gene deletion effort in our laboratory. We also included the helicase DHX35 in this analysis as it is found in *C. neoformans* spliceosomes but not in those of *S. cerevisiae* (Burke et al., 2018). The human orthologs of the encoded proteins studied here associate with spliceosomes at stages ranging from early complexes such as the A complex to late catalytic/postcatalytic complexes (Cvitkovic and Jurica, 2013). Three of the proteins investigated here harbor a G-patch motif, which is found in proteins that activate superfamily 2 helicases including two involved in splicing in yeast (Robert-Paganin et al., 2015; Studer et al., 2020).

GPATCH1 and DHX35 act together on active spliceosomes

RNA-seq analysis indicated that mutations in each of the 14 of the human spliceosome protein orthologs examined altered both splicing efficiency and choice. Clustering of the data based on the impacted introns in each mutant demonstrated that mutants lacking orthologs of GPATCH1 and DHX35 displayed consistently the most correlated signatures for multiple types of splicing changes (intron retention, alternative 5' splice site choice, and alternative 3' splice site choice). Affinity purification of a FLAG-tagged allele of GPATCH1 identified DHX35 as a top hit. Given that G-patch proteins are known activators of helicases, it seems very likely that GPATCH1 functions to activate DHX35 in the spliceosome, although further biochemical work will be necessary to confirm this hypothesis. What the substrate of a GPATCH1/DHX35 complex might be is unclear, but, based on the nature of the changes in splice site choice (see below), a role reminiscent to those of Prp16 and Prp22 in proofreading during the catalytic stages of splicing seems possible (Koodathingal and Staley, 2013). In this regard, we note that, in human cells, GPATCH1 and DHX35 are found in catalytically active spliceosomal complexes (Ilagan et al., 2013), and the spectrometry data in *Cryptococcus* presented here and elsewhere (Burke et al., 2018) indicates that this pattern of association is conserved in fungi.

Accessory factors impact the processing of genes with divergent geometries

The mRNA-to-precursor ratio is a classic measurement of splicing efficiency (Pikielny and Rosbash, 1985; Rymond et al., 1990). As such, what is referred to as an increase in intron retention is equivalent to a decrease in mRNA splicing efficiency. A subset of mutants analyzed here display a bias in an increase in intron retention (versus a decrease), indicating a tendency towards reducing the efficiency of splicing of specific substrates when mutated, reminiscent of classic pre-mRNA splicing mutants. These include mutants lacking ortholog of GPATCH1 and

DHX35 as well as mutants in NOSIP, RBM5, SAP18, and RBM17. Unexpectedly, four mutants tested either show the opposite bias (a bias towards improving splicing efficiency when absent): NOSIP, IK, FAM50A, and FRA10AC1. The effects of accessory factors on splicing efficiency correlates with distinctive features of substrates, notably longer intron size and nonoptimal predicted branchpoint-to-3' splice site distance. We note that the impact of many of the factors studied here is biased rather than purely unidirectional. For example, while knockout of the ortholog of human GPATCH1 is strongly biased towards causing reduced use of canonical 5' and 3' splice sites in favor of poor alternative sites, in a minority of cases, the opposite effect is observed. This may reflect a combination of direct and indirect effects [such as competition of 'hungry' spliceosomes for introns (Munding et al., 2013; Talkish et al., 2019)]. Alternatively, they may represent context-dependent roles that are influenced by complex differences in intron structure and sequence. Ultimately, *in vitro* reconstitution studies (either in human or *C. neoformans* extracts) and structural studies will be required for a full mechanistic understanding of the observations described here.

Accessory factors promote spliceosome fidelity

A notable finding of this work is that nine factors analyzed display functional signatures that are biased towards to the suppression of the use of nearby weak/cryptic 5' while four factors are biased towards suppression of nearby, weak 3' splice sites. Orthologs of GPATCH1 and DHX35 are notable in that they display this function for both 5' and 3' sites. This phenotype further suggests that these factors may act in a manner akin to the *S. cerevisiae* fidelity factors, Prp16 and Prp22. We propose that such an additional layer of proofreading might be necessary in organisms whose spliceosomes need to accommodate more variable intron consensus sequences as such flexible spliceosomes are likely to be more error prone. Other factors, such as the G-patch proteins RBM5 and RBM17 may have similar roles in earlier spliceosomal complexes. In this regard, we note that DHX15 (*Sc.* Prp43), the spliceosomal disassembly

helicase, has consistently been observed as a component of U2 snRNP (Cvitkovic and Jurica, 2013; Hartmuth et al., 2002). Our studies suggest that much remains to be learned about the spliceosome, and points towards the key roles of functional studies in tractable non-reduced organisms in complementing studies from *S. cerevisiae*.

Methods

Spliceosomal protein searches

Spliceosomal protein searches were performed on proteome assemblies available from NCBI and UniProt (See Table 4.7**)**. A curated list of relevant human spliceosomal proteins was used as queries in local BLASTp (version 2.9.0+) searches against independent Fungal proteome databases (initial e-value threshold of 10^{-6}) (Altschul et al., 1997). The results from the BLAST searches were further screened by analyzing domain content (HMMsearch, HMMer 3.1b2 – default parameters), size comparisons against human protein sequence length (within 25% variation), and reciprocal best-hit BLAST searches (RBH) to the query proteome (Bork et al., 1998; Johnson et al., 2010; Tatusov et al., 1997). To avoid bias in protein domain content, domains used for HMM searches were defined as described(Hudson et al., 2019). Briefly, a conserved set of domains for each spliceosomal protein was assembled by using only those domains present in all three of the human, yeast, and *Arabidopsis* orthologs. Fungal ortholog candidates in this study were scored and awarded a confidence value of 0-9 based on passing the above criteria. Scores were calculated by starting at 9 and penalizing candidates for falling outside of the expected size range (-1 point), missing HMM domain calls (-2 points), and failing to strictly pass RBH (-5 points). A score of 0.5 was given to candidates that failed all criteria but still had BLAST hits after the initial human query to separate from those that had no BLAST hits.

C. neoformans cultivation

Two-liter liquid cultures of all strains were grown in YPAD medium (Difco) by inoculation at low density (0.002-0.004 OD600 nm) followed by overnight growth with shaking 30C. For RNA-seq experiments, cells were harvested at OD_{600} of ~ 1. For TMT-MS experiments, an additional 1% glucose was added when the cultures reached OD_{600} of 1. Cells were harvested at OD_{600} of 2.

Immunoprecipitation and TMT-MS

Strains harboring a C-terminal (GPATCH1 and RBM17) or an N-terminal (RBM5) CBP-2XFLAG tag were generated by homologous replacement. Immunoprecipitations were performed exactly as described (Burke et al, 2018) with the following modifications: lysis and wash buffers were adjusted to either 150 mM KCl (low salt) or 300 mM KCl (high salt). Two untagged samples and two tagged samples (one at each salt concentration) was produced. The four samples were then subject to TMT-MS exactly as described (Burke et al., 2018).

RNA preparation

Polyadenylated RNA was prepared exactly as described (Burke et al., 2018).

RNA-seq

RNA-seq libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina. Samples were sequenced using an Illumina HiSeq 4000 instrument. Paired-end 100 nt reads were obtained. Data are available at the NCBI GEO database: GSE168814

RNA-seq data analysis

All reads were analyzed using FastQC and reads with more than 80% of quality scores below 25 were thrown out. Reads were aligned using STAR (Dobin et al., 2013). A minimum of 12M read/strain/replicate were obtained (Table 4.S8). Differentially spliced introns were called using JUM (version 2.0.2). Differential events with a p-value of greater than 0.05 were set to 1. An additional 5 read minimum was imposed. To further minimize false positive, differential splicing events called by JUM that do not have an isoform harboring a start and end corresponding with an annotated intron were also removed. Spot-checking of differential events was accomplished by manual browsing of the data. Alternative 3' and 5' splicing events containing more than two alternative endpoints were also removed. In few cases, JUM called introns as significantly alternatively spliced with a $\Delta\psi$ of 0; those introns were also removed. Next, each intron is classified as increased or decreased and proximal or distal based on the observed canonical endpoint and associated $\Delta\psi$. General data analysis, plotting and statistical testing were performed using Python and the SciPy stack as follows:

Binomial tests: Introns were grouped by splicing event and strain. Within each strain a binomial test (scipy.stats.binom_test) was conducted to see if there was significantly more or fewer introns with increased splicing.

Comparisons of distributions: Introns are grouped by strain and condition and each subset is compared to unaffected introns. The resulting two distributions are compared for each attribute. A Wilcoxon rank-sum test (scipy.stats.ranksums) is conducted to determine if the means of the two distributions are significantly different. (A Kolmogorov–Smirnov test is conducted to compare the distributions themselves (scipy.stats.kstest).) All results are multiple-test corrected using the FDR correction (statsmodels.stats.multitest.fdrcorrection).

Chi-squared analysis (canonical vs alternative sites): Introns were separated by strain and condition and the first and last six nucleotides of the canonical sequence of affected introns was compared to the non-canonical sequence of affected introns. Each position in the endpoints was treated as an independent Fisher exact test (FisherExact.fisherexact) or chi-square test (scipy.stats.chisquare) with (4-1)*(2-1)= 3 degrees of freedom performed on a contingency table with nucleotides in the rows and affected vs. unaffected introns as the columns. In some cases where less than 5 counts are observed in a category, a chi-squared test becomes inappropriate and the Fisher exact test is used.

P-value correlations: Treating all strains as a vector of p-values of affected introns, a Pearson correlation matrix is computed. (pandas.DataFrame.corr).

Seqlogos: Affected introns are grouped by splice type, condition (increased or decreased), and strain. Seqlogos are generated from the first and last six nucleotides (seqlogo.seqlogo).

Acknowledgements

This work was supported by R01 GM71801 and R01 AI00272 to H.D.M. B.A.B. and S.W.R. are supported by NSF Award 1751372 to S.W.R. I.B. is supported by a Swiss National Foundation Fellowship (191929). We thank Qingqing Wang for assistance with installation and usage of JUM scripts.

Figures

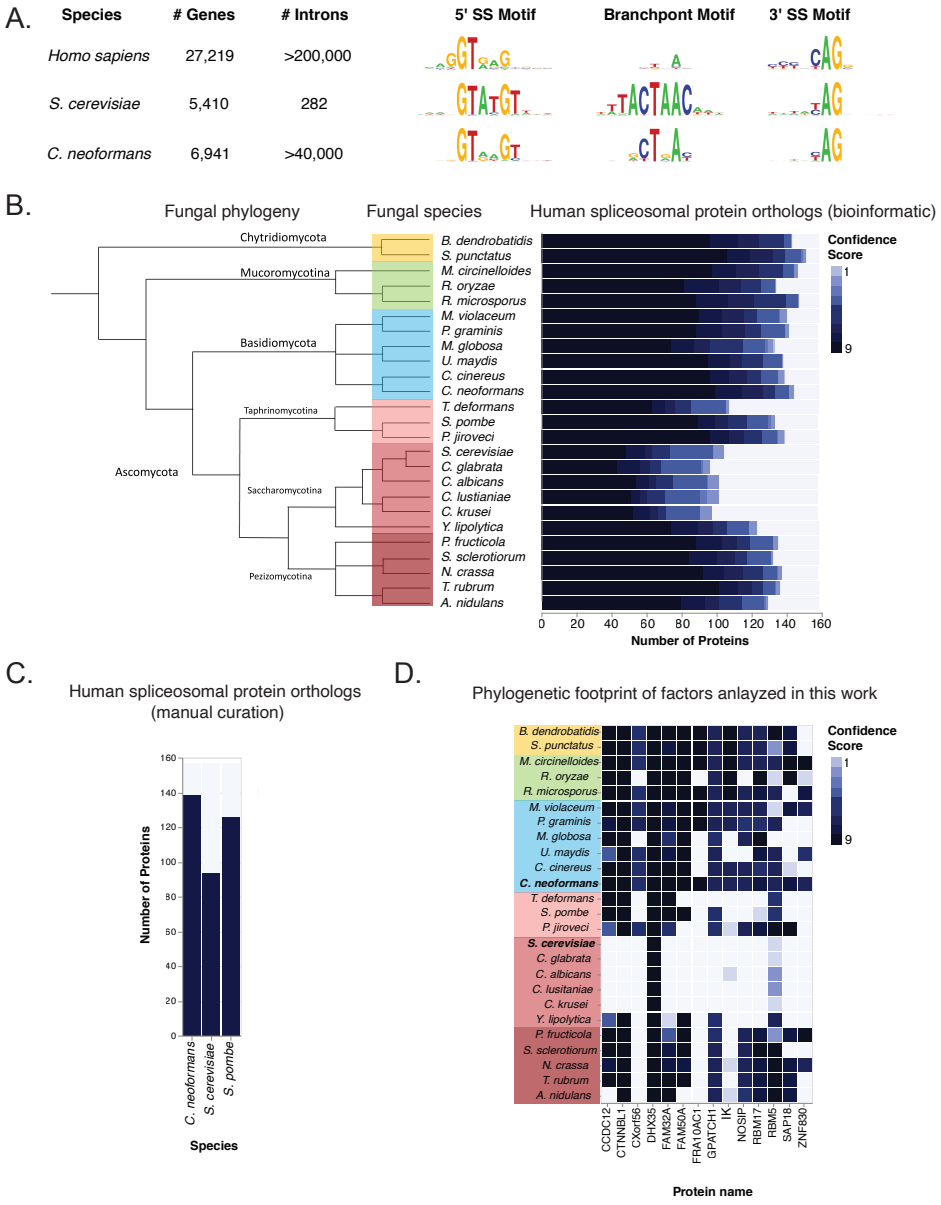


Figure 4.1: Massive loss of human spliceosomal protein orthologs in specific fungal lineages.

A. Comparison of intron number and properties in humans versus the yeasts S. cerevisiae and C. neoformans.

B. Evolutionary loss events. Shown is a phylogeny of the indicated fungal species, the groups to which they belong and the human spliceosomal protein orthologs that can be identified by a semi-automated bioinformatic process at the indicated levels of confidence (see Methods). Phylogeny is based on James et al. (2020). See also Table 4.1.

C. Numbers of human spliceosomal protein orthologs in S. cerevisiae, S. pombe and C. neoformans. Based on extensive additional literature curation, the numbers of human spliceosomal orthologs in the indicated species are plotted. See also Table 4.1.

D. Spliceosomal factor orthologs for which null mutations in C. neoformans were obtained. Plotted are the confidence scores for the presence of the indicated human spliceosomal protein orthologs in the indicated species. See also Table 4.1.

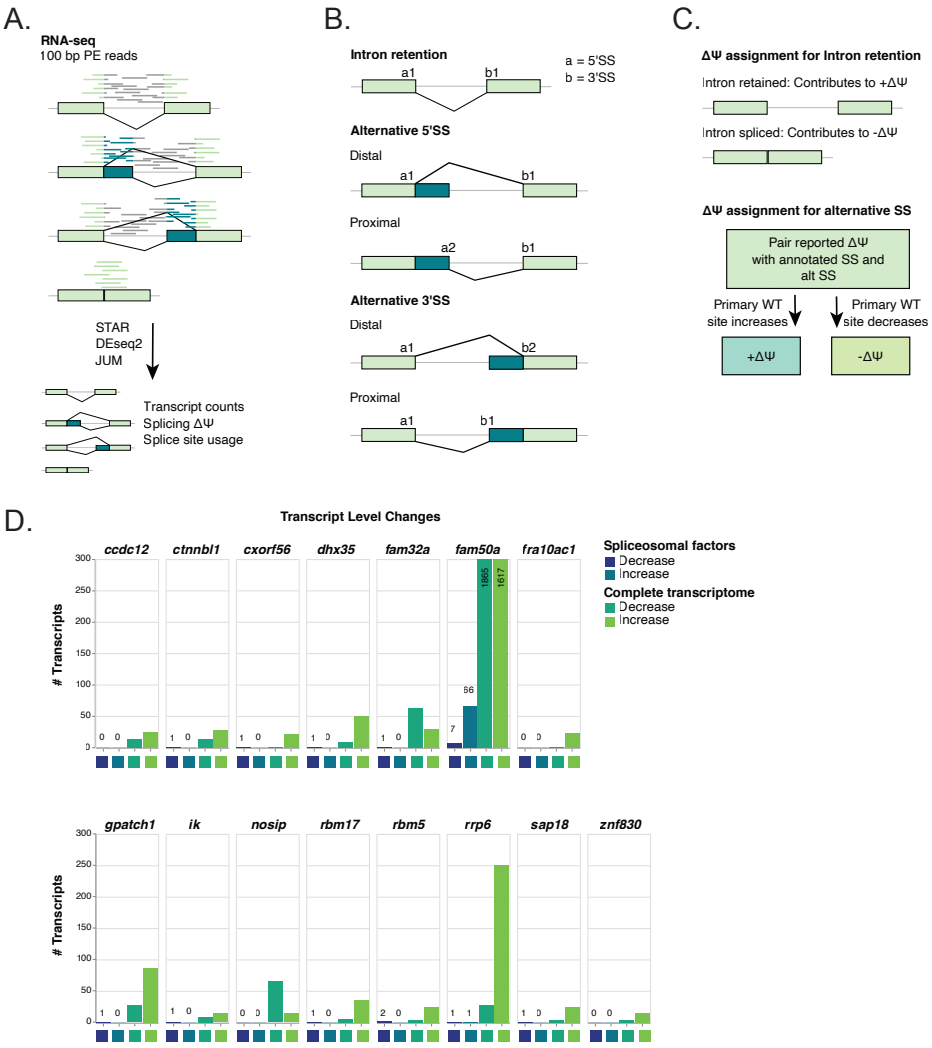


Figure 4.2: RNA-seq analysis of null mutations in 14 human spliceosomal protein orthologs.

A. Schematic of the RNA-seq pipeline. Replicate mutant and paired replicate wild-type samples grown on the same day were subjected to 100nt paired-end stranded RNA-seq analysis. Initially, an annotation-free method is used to identify splicing events which was then filtered to require that at least one of the alternative events corresponded to an annotated intron. This approach produced results that could be hand-validated through direct visualization of read data.

B. Definitions of splicing events quantified.

C. Definitions of changes to Percent Spliced In (delta PSI or $\Delta\psi$) values for intron retention, alternative 3' splice site, and alternative 5' spice site events. For the intron retention category, introns that display an increased in unspliced precursor (relative to spliced mRNA) in the mutant genotype are assigned a positive $\Delta\psi$, while those that show a relative decrease in unspliced precursor are assigned a negative $\Delta\psi$. For alternative 3' and 5' events, the primary splice site used in wild type is used to assign direction of the splicing change. An increase in primary site usage was assigned positive $\Delta\psi$, and a decrease was assigned negative $\Delta\psi$.

D. Changes in transcript levels in mutants. 100 bp paired-end RNA-seq results for 15 knockout strains and KN99 wildtype were analyzed using DEseq2 with a 2-fold change cutoff and an adjusted p-value cutoff of 0.01. All strains were grown in duplicate with duplicate wild-type strains grown on the same day. Significantly changed gene IDs were compared to a list of human splicing protein orthologs to determine the number of splicing factors affected by the KO. Plotted is the total number of splicing factors changed in the RNA-seq data as well as total transcriptome changes. (the genotype of strains with no called decreases the expression of annotated splicing factors were confirmed by direct examination of read data to show no reads corresponding to the region deleted).

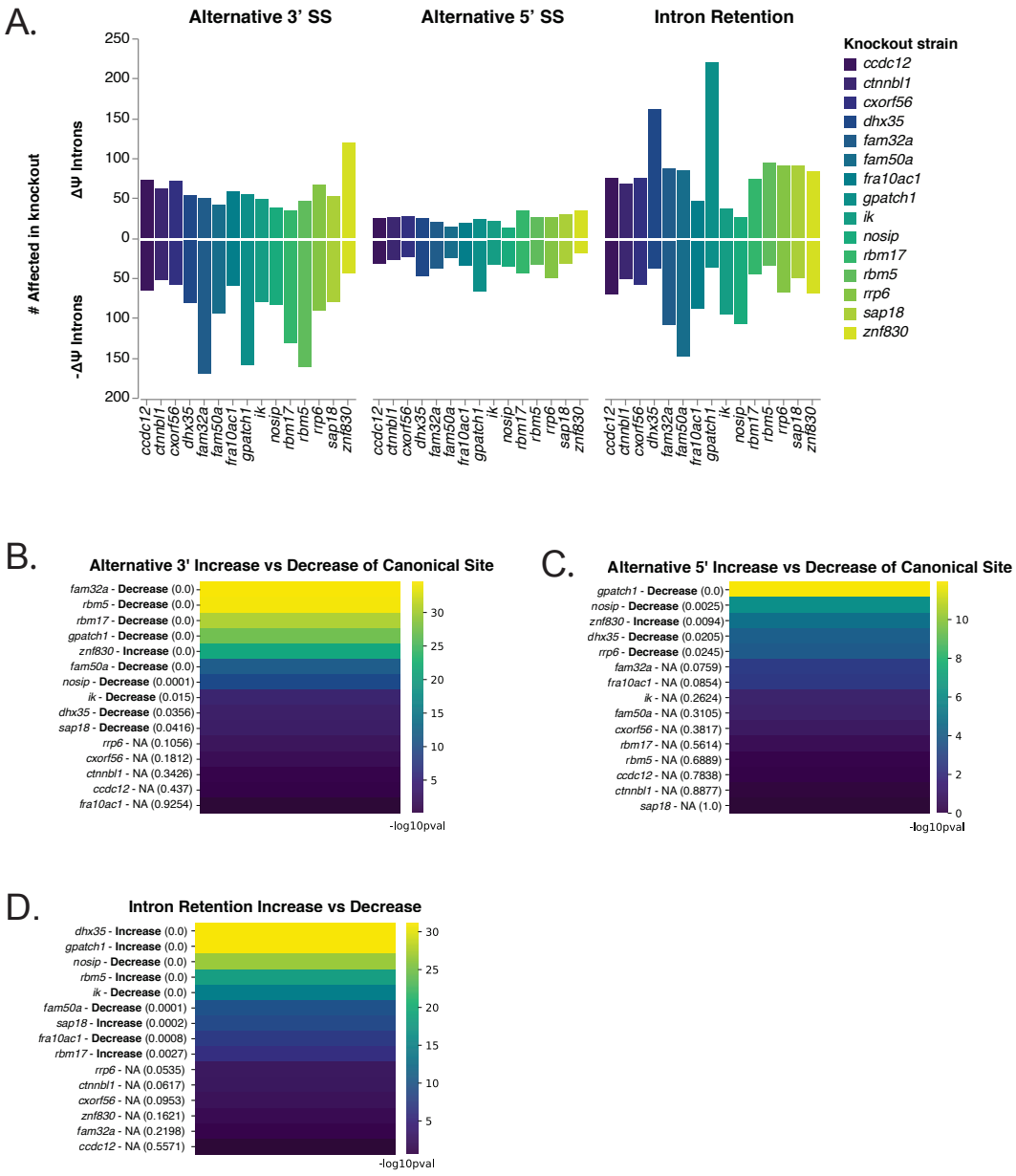


Figure 4.3: Quantification of altered pre-mRNA splicing in mutant lacking orthologs of human spliceosomal proteins.

A. Number of introns altered in pre-mRNA splicing in mutants. Changes in splicing is here plotted as a count of the number of introns with significant $\Delta\psi$ ($p<0.05$) values. Counts for each KO strain are reported for each of three alternative splicing types. Intron counts displaying positive changes in $\Delta\psi$ as defined in Figure 4.2C are plotted above the line while intron counts displaying negative changes in $\Delta\psi$ as defined in Figure 4.2C are plotted below the line.

B. Binomial test for directionality of alternative 3' splice site usage changes. Introns affected by each KO strain were analyzed to test for a bias towards positive or negative $\Delta\psi$. KO name is

reported followed by direction and p-value. -log10(p-value) is displayed and colored as indicated. The labels on the left indicate the mutant, whether the bias reflects a decrease or increase in the canonical splice site in the mutant with the p-value shown in parenthesis.

C. Binomial test for directionality of alternative 5' splice site usage changes. Introns affected by each KO strain were analyzed to test for a bias towards positive or negative $\Delta\psi$. KO name is reported followed by direction and p-value. -log10(p-value) is displayed and colored as indicated. The labels on the left indicate the mutant, whether the bias reflects a decrease or increase in the canonical splice site in the mutant with the p-value shown in parenthesis.

D. Binomial test for directionality of intron retention changes. Introns affected by each KO strain were analyzed to test for a bias towards positive or negative $\Delta\psi$. KO name is reported followed by direction and p-value. -log10(p-value) is displayed and colored as indicated. The labels on the left indicate the mutant, whether the bias reflects a decrease or increase precursor accumulation in the mutant with the p-value shown in parenthesis.

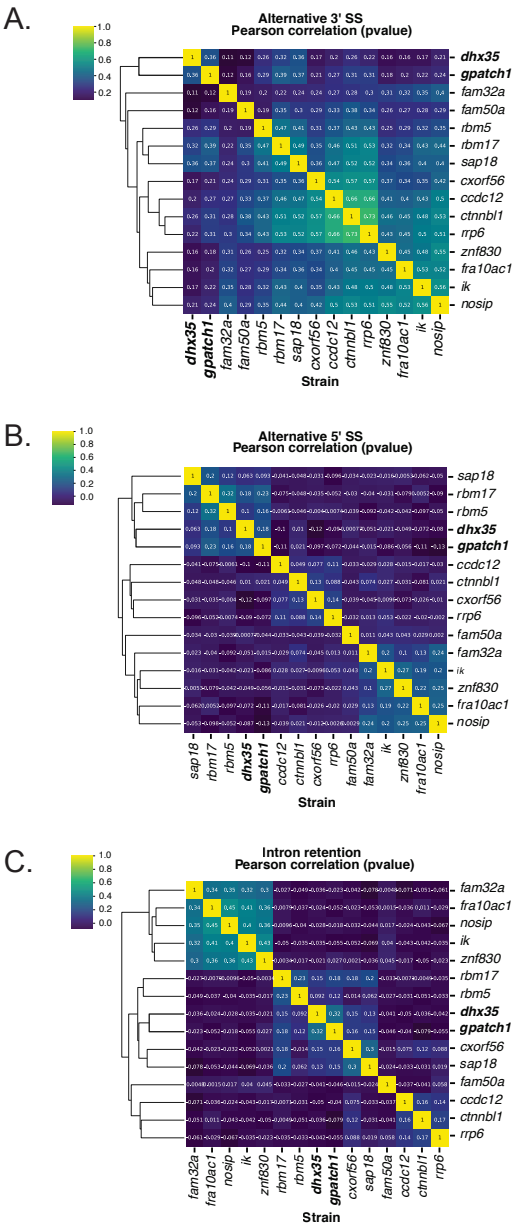


Figure 4.4: Correlation between phenotypic signatures of spliceosomal protein ortholog gene deletion mutants.

P-values (corrected for multiple hypothesis testing) for changes in splicing were treated as vectors and used to generate an autocorrelation matrix for each type of splicing event. P-values greater than 0.05 were set to 1. Pearson correlation was used as the distance metric (values shown in boxes – color bar shows absolute values). Data are organized by hierarchical clustering.

A. Mutant autocorrelation matrix based on significant alternative 3’ splice site changes

B. Mutant autocorrelation matrix based on significant alternative 5’ splice site changes

C. Mutant autocorrelation matrix based on significant intron retention changes

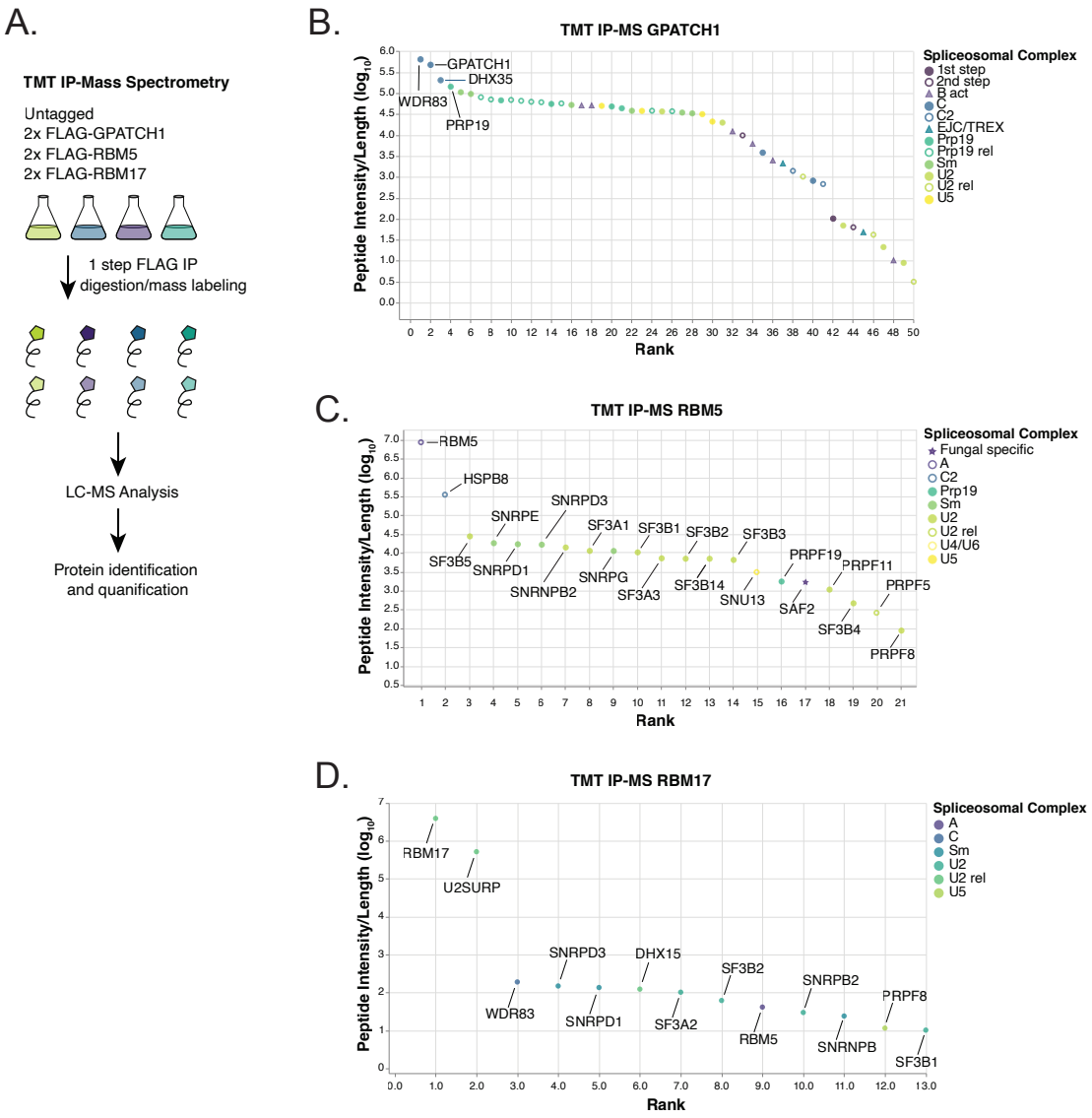


Figure 4.5: Purifications and TMT-MS analysis of endogenously tagged human spliceosomal protein orthologs.

A. Schematic of sample preparation for TMT-MS. Strains harboring 2x FLAG-tagged alleles were grown to OD600nm of 2 before being harvested and lysed (see Methods). Purifications were performed on the soluble fraction in untagged and tagged strains at low and high salt (four purifications for each tagged strains). Shown are relative normalized abundances of the sum of the low- and high-salt peptide intensities of the spliceosomal protein orthologs.

B. IP-MS results for 2x-FLAG GPATCH1. Plotted are TMT-MS data with length-normalized peptide intensity (log10) on the Y-axis and rank on the X-axis. Color and shape indicate the spliceosomal complex associated with the human orthologue. The top four hits are labeled with their orthologous human protein names.

C. IP-MS results for 2x-FLAG RBM17. Plotted are TMT-MS data with length-normalized peptide intensity (log10) on the Y-axis and rank on the X-axis. Color and shape indicate the spliceosomal complex associated with the human orthologue. All hits are labeled with their orthologous human protein names.

D. IP-MS results for 2x-FLAG RBM5. Plotted are TMT-MS data with length-normalized peptide on the Y-axis (log10) and rank on the X-axis. Color and shape indicate the spliceosomal complex associated with the human orthologue. All hits are labeled with their orthologous human protein names.

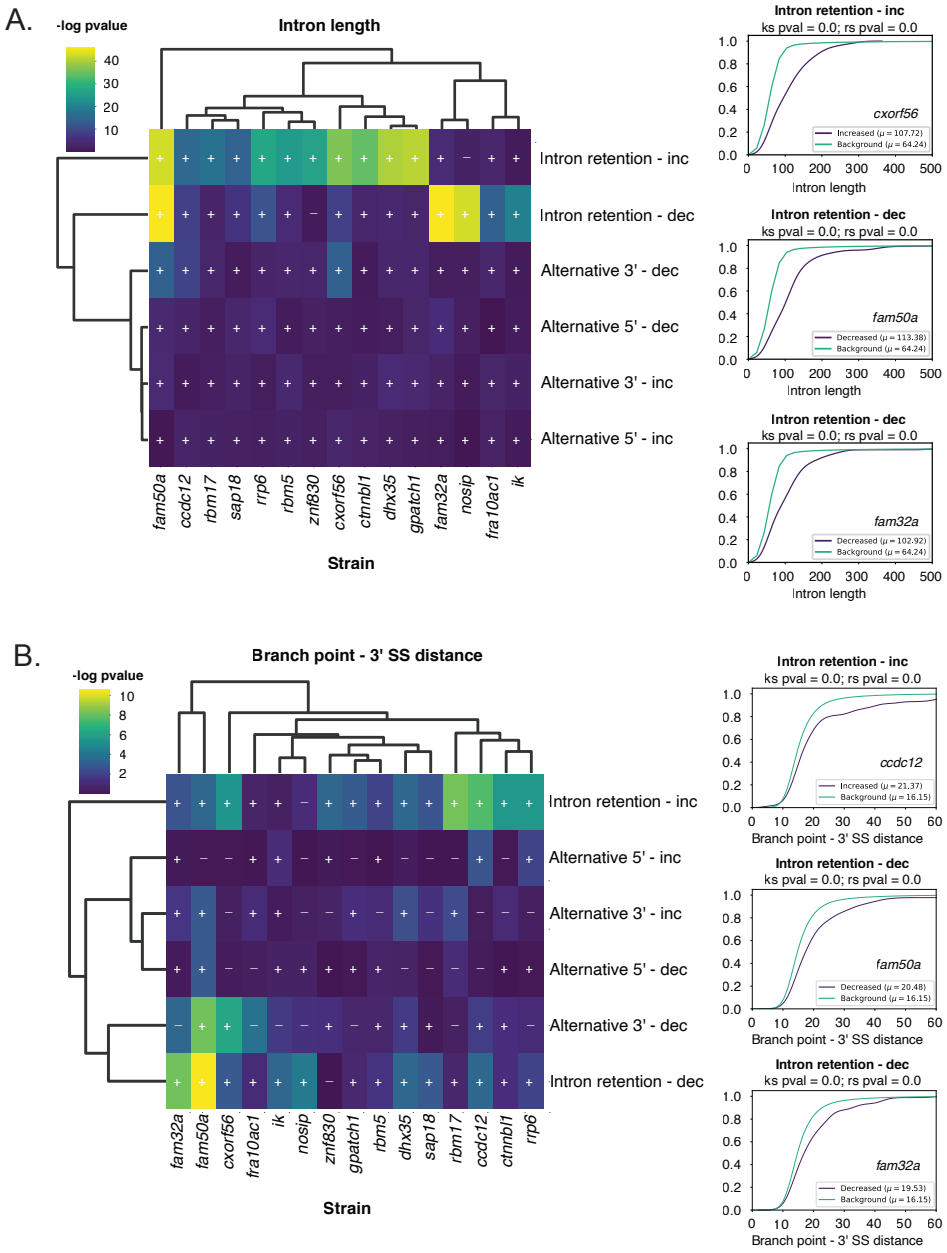


Figure 4.6: Enrichment of divergent geometry parameters of introns whose splicing is altered in human spliceosomal ortholog mutants.

A. Enrichment of altered lengths in affected introns. Shown in the heatmap is the negative log of the p-value generated produced by a Wilcoxon Rank Sum Test (corrected for multiple hypothesis testing) comparing affected and unaffected introns for a given gene deletion strain and type of splicing change. Also indicated by a (+) or (-) sign is the direction of effect. Shown on the right are CDF plots and statistical test results for three gene deletion mutants/splicing change combinations that display the most significant effects.

B. Enrichment of altered predicted branchpoint-to-3' splice site distances. Analysis was performed as in A. Branchpoint-to-3' splice site distances were predicted by using C. neoformans branchpoint consensus to predict branchpoints computationally.

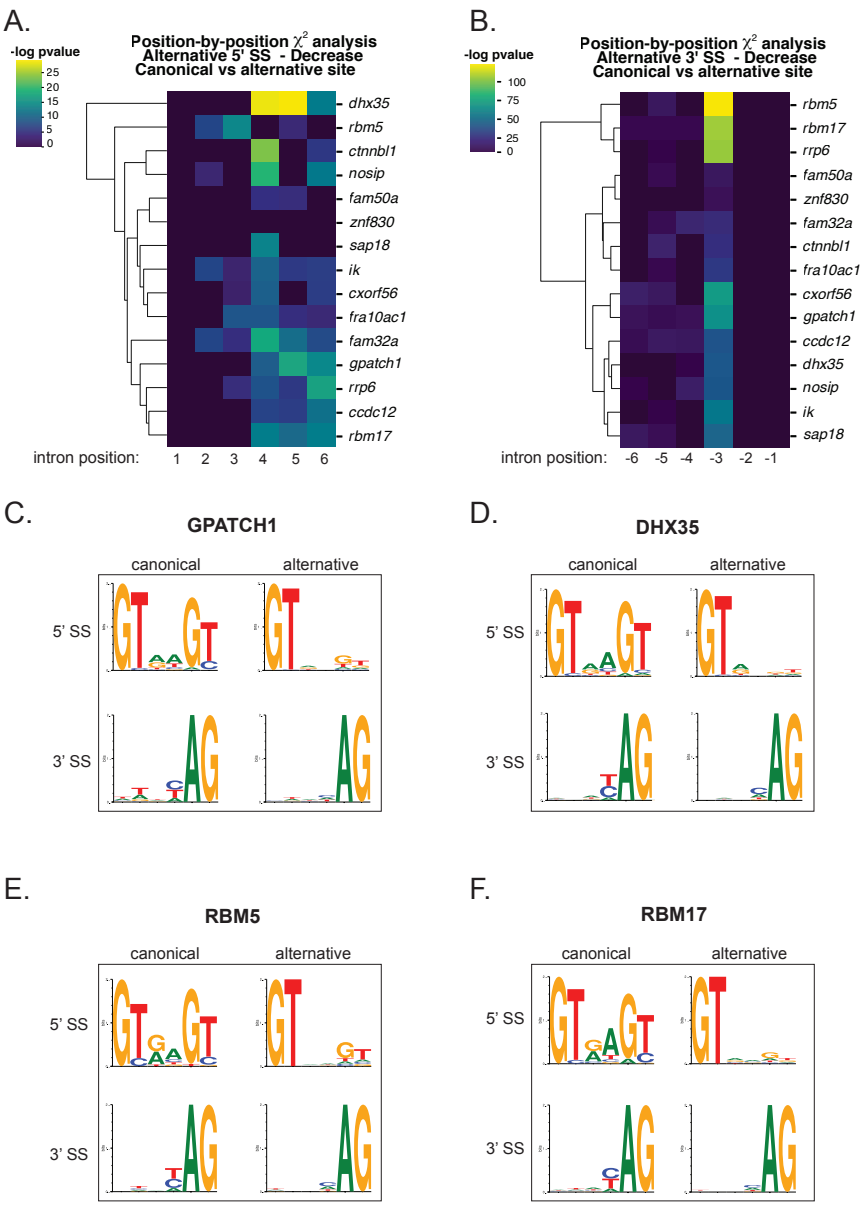


Figure 4.7: Activation of weak/cryptic alternative 5' and 3' splice sites in human spliceosomal ortholog mutants.

A. 5' splice site bases showing significant differences in composition between canonical and alternative sites. Chi-squared analysis (corrected for multiple hypothesis testing) of the first six nucleotides of introns showing significantly decreased Δψ (reduced use of the canonical site and increased use of the alternative site) for alternative 5' splice sites in the mutant. Plotted is the negative log10 p value. Strains were clustered by similarity.

B. 3' splice site bases showing significant differences in composition between canonical and alternative sites. Chi-squared analysis (corrected for multiple hypothesis testing) of the last six

nucleotides of introns showing significantly decreased Δψ (reduced use of the canonical site and increased use of the alternative site) for alternative 3' splice sites in the mutant. Plotted is the negative log10 p-value. Strains were clustered by similarity.

Tables

Table 4.1

Core human spliceosomal proteins and fungal orthologs. Data are presented as an Excel file.

DOI https://doi.org/10.7272/Q6T72FP1

Table 4.2

DESeq2 output of RNA-seq data. Data are presented as an Excel file.

DOI https://doi.org/10.7272/Q6T72FP1

Table 4.3

Output of Junction Utilization Package (JUM) Analysis of 15 Mutant Strains. Data are presented as an Excel file.

DOI https://doi.org/10.7272/Q6T72FP1

Table 4.4

Full TMT-MS Data for GPATCH1 immunopurifications. Data are presented as an Excel file. Raw, annotated, and spliceosomal protein data displayed in separate sheets.

DOI https://doi.org/10.7272/Q6T72FP1

Table 4.5

Full TMT-MS Data for RBM17 immunopurifications. Data are presented as an Excel file. Raw, annotated, and spliceosomal protein data displayed in separate sheets.

DOI https://doi.org/10.7272/Q6T72FP1

Table 4.6

Full TMT-MS Data for RBM5 immunopurifications. Data are presented as an Excel file. Raw, annotated, and spliceosomal protein data displayed in separate sheets. For readability, ribosomal proteins were filtered to generate the annotated sheet.

DOI https://doi.org/10.7272/Q6T72FP1

Table 4.7

Proteomes used for Evolutionary Analysis. Data are presented as an Excel file

DOI https://doi.org/10.7272/Q6T72FP1

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.
2. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. J Mol Biol *283*, 707-725.
3. Burke, J.E., Longhurst, A.D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J.J., Yates, J.R., 3rd, Li, J.J., and Madhani, H.D. (2018). Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. Cell *173*, 1014-1030 e1017.
4. Chen, W., Shulha, H.P., Ashar-Patel, A., Yan, J., Green, K.M., Query, C.C., Rhind, N., Weng, Z., and Moore, M.J. (2014). Endogenous U2.U5.U6 snRNA complexes in S. pombe are intron lariat spliceosomes. RNA *20*, 308-320.
5. Cipakova, I., Jurcik, M., Rubintova, V., Borbova, M., Mikolaskova, B., Jurcik, J., Bellova, J., Barath, P., Gregan, J., and Cipak, L. (2019). Identification of proteins associated with splicing factors Ntr1, Ntr2, Brr2 and Gpl1 in the fission yeast Schizosaccharomyces pombe. Cell Cycle *18*, 1532-1536.
6. Colley, A., Beggs, J.D., Tollervey, D., and Lafontaine, D.L. (2000). Dhr1p, a putative DEAH-box RNA helicase, is associated with the box C+D snoRNP U3. Mol Cell Biol *20*, 7238-7246.
7. Cvitkovic, I., and Jurica, M.S. (2013). Spliceosome database: a tool for tracking components of the spliceosome. Nucleic Acids Res *41*, D132-141.
8. De Maio, A., Yalamanchili, H.K., Adamski, C.J., Gennarino, V.A., Liu, Z., Qin, J., Jung, S.Y., Richman, R., Orr, H., and Zoghbi, H.Y. (2018). RBM17 Interacts with U2SURP and CHERP to Regulate Expression and Splicing of RNA-Processing Proteins. Cell Rep *25*, 726-736 e727.

9. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
10. Dumesic, P.A., Natarajan, P., Chen, C., Drinnenberg, I.A., Schiller, B.J., Thompson, J., Moresco, J.J., Yates, J.R., 3rd, Bartel, D.P., and Madhani, H.D. (2013). Stalled spliceosomes are a signal for RNAi-mediated genome defense. Cell *152*, 957-968.
11. Fica, S.M., Oubridge, C., Wilkinson, M.E., Newman, A.J., and Nagai, K. (2019). A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. Science *363*, 710-714.
12. Grate, L., and Ares, M., Jr. (2002). Searching yeast intron data at Ares lab Web site. Methods Enzymol *350*, 380-392.
13. Hartmuth, K., Urlaub, H., Vornlocher, H.P., Will, C.L., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. Proc Natl Acad Sci U S A *99*, 16719-16724.
14. Hudson, A.J., McWatters, D.C., Bowser, B.A., Moore, A.N., Larue, G.E., Roy, S.W., and Russell, A.G. (2019). Patterns of conservation of spliceosomal intron structures and spliceosome divergence in representatives of the diplomonad and parabasalid lineages. BMC Evol Biol *19*, 162.
15. Ilagan, J.O., Chalkley, R.J., Burlingame, A.L., and Jurica, M.S. (2013). Rearrangements within human spliceosomes captured after exon ligation. RNA *19*, 400-412.
16. Irimia, M., Penny, D., and Roy, S.W. (2007). Coevolution of genomic intron number and splice sites. Trends Genet *23*, 321-325.
17. Irimia, M., and Roy, S.W. (2008). Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet *4*, e1000148.

18. Ishigaki, Y., Li, X., Serin, G., and Maquat, L.E. (2001). Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. Cell *106*, 607-617.
19. Janbon, G., Ormerod, K.L., Paulet, D., Byrnes, E.J., 3rd, Yadav, V., Chatterjee, G., Mullapudi, N., Hon, C.C., Billmyre, R.B., Brunel, F.*, et al.* (2014). Analysis of the genome and transcriptome of Cryptococcus neoformans var. grubii reveals complex RNA expression and microevolution leading to virulence attenuation. PLoS Genet *10*, e1004261.
20. Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431.
21. Johnston, J.J., Teer, J.K., Cherukuri, P.F., Hansen, N.F., Loftus, S.K., Center, N.I.H.I.S., Chong, K., Mullikin, J.C., and Biesecker, L.G. (2010). Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Am J Hum Genet *86*, 743-748.
22. Kim, Y., Hur, S.W., Jeong, B.C., Oh, S.H., Hwang, Y.C., Kim, S.H., and Koh, J.T. (2018). The Fam50a positively regulates ameloblast differentiation via interacting with Runx2. J Cell Physiol *233*, 1512-1522.
23. Koodathingal, P., and Staley, J.P. (2013). Splicing fidelity: DEAD/H-box ATPases as molecular clocks. RNA Biol *10*, 1073-1079.
24. Kuhny, M., Forbes, L.R., Cakan, E., Vega-Loza, A., Kostiuk, V., Dinesh, R.K., Glauzy, S., Stray-Pedersen, A., Pezzi, A.E., Hanson, I.C.*, et al.* (2020). Disease-associated CTNNBL1 mutation impairs somatic hypermutation by decreasing nuclear AID. J Clin Invest *130*, 4411-4422.
25. Lai, S., O'Callaghan, B., Zoghbi, H.Y., and Orr, H.T. (2011). 14-3-3 Binding to ataxin-1(ATXN1) regulates its dephosphorylation at Ser-776 and transport to the nucleus. J Biol Chem *286*, 34606-34616.

26. Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M.J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J *20*, 4987-4997.
27. Lee, Y.R., Khan, K., Armfield-Uhas, K., Srikanth, S., Thompson, N.A., Pardo, M., Yu, L., Norris, J.W., Peng, Y., Gripp, K.W.*, et al.* (2020). Mutations in FAM50A suggest that Armfield XLID syndrome is a spliceosomopathy. Nat Commun *11*, 3698.
28. Lesser, C.F., and Guthrie, C. (1993). Mutational analysis of pre-mRNA splicing in Saccharomyces cerevisiae using a sensitive new reporter gene, CUP1. Genetics *133*, 851-863.
29. Lim, J., Crespo-Barreto, J., Jafar-Nejad, P., Bowman, A.B., Richman, R., Hill, D.E., Orr, H.T., and Zoghbi, H.Y. (2008). Opposing effects of polyglutamine expansion on native protein complexes contribute to SCA1. Nature *452*, 713-718.
30. Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A.*, et al.* (2005). The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science *307*, 1321-1324.
31. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.
32. McDonald, W.H., Ohi, R., Smelkova, N., Frendewey, D., and Gould, K.L. (1999). Myb-related fission yeast cdc5p is a component of a 40S snRNP-containing complex and is essential for pre-mRNA splicing. Mol Cell Biol *19*, 5352-5362.
33. Munding, E.M., Shiue, L., Katzman, S., Donohue, J.P., and Ares, M., Jr. (2013). Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. Mol Cell *51*, 338-348.
34. Pikielny, C.W., and Rosbash, M. (1985). mRNA splicing efficiency in yeast and the contribution of nonconserved sequences. Cell *41*, 119-126.

35. Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., and Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes *12*, 315.
36. Robert-Paganin, J., Rety, S., and Leulliot, N. (2015). Regulation of DEAH/RHA helicases by G-patch proteins. Biomed Res Int *2015*, 931857.
37. Rocha, M.E., Silveira, T.R.D., Sasaki, E., Sas, D.M., Lourenco, C.M., Kandaswamy, K.K., Beetz, C., Rolfs, A., Bauer, P., Reardon, W.*, et al.* (2020). Novel clinical and genetic insight into CXorf56-associated intellectual disability. Eur J Hum Genet *28*, 367-372.
38. Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.
39. Rymond, B.C., Pikielny, C., Seraphin, B., Legrain, P., and Rosbash, M. (1990). Measurement and analysis of yeast pre-mRNA sequence contribution to splicing efficiency. Methods Enzymol *181*, 122-147.
40. Sharp, P.A. (1987). Splicing of messenger RNA precursors. Science *235*, 766-771.
41. Studer, M.K., Ivanovic, L., Weber, M.E., Marti, S., and Jonas, S. (2020). Structural basis for DEAH-helicase activation by G-patch proteins. Proc Natl Acad Sci U S A *117*, 7159-7170.
42. Talkish, J., Igel, H., Perriman, R.J., Shiue, L., Katzman, S., Munding, E.M., Shelansky, R., Donohue, J.P., and Ares, M., Jr. (2019). Rapidly evolving protointrons in Saccharomyces genomes revealed by a hungry spliceosome. PLoS Genet *15*, e1008249.
43. Tan, Q., Yalamanchili, H.K., Park, J., De Maio, A., Lu, H.C., Wan, Y.W., White, J.J., Bondar, V.V., Sayegh, L.S., Liu, X.*, et al.* (2016). Extensive cryptic splicing upon loss of RBM17 and TDP43 in neurodegeneration models. Hum Mol Genet *25*, 5083-5093.
44. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637.
45. Townsend, C., Leelaram, M.N., Agafonov, D.E., Dybkov, O., Will, C.L., Bertram, K., Urlaub, H., Kastner, B., Stark, H., and Luhrmann, R. (2020). Mechanism of protein-guided folding of the active site U2/U6 RNA during spliceosome activation. Science *370*.

46. Verkerk, A., Zeidler, S., Breedveld, G., Overbeek, L., Huigh, D., Koster, L., van der Linde, H., de Esch, C., Severijnen, L.A., de Vries, B.B.A.*, et al.* (2018). CXorf56, a dendritic neuronal protein, identified as a new candidate gene for X-linked intellectual disability. Eur J Hum Genet *26*, 552-560.
47. Wahl, M.C., and Luhrmann, R. (2015a). SnapShot: Spliceosome Dynamics I. Cell *161*, 1474-e1471.
48. Wahl, M.C., and Luhrmann, R. (2015b). SnapShot: Spliceosome Dynamics II. Cell *162*, 456-456 e451.
49. Wang, Q., and Rio, D.C. (2018). JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. Proc Natl Acad Sci U S A *115*, E8181-E8190.
50. Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA Splicing by the Spliceosome. Annu Rev Biochem *89*, 359-388.
51. Zhou, Z., Luo, M.J., Straesser, K., Katahira, J., Hurt, E., and Reed, R. (2000). The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature *407*, 401-405.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Jade Sales-Lee
4D2BD3EA2AC9458...
Author Signature

3/14/2021
Date