# 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

## SUMMARY

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 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, orthologs of a number 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 a large fraction of 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 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 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 (Supplementary Table S1). 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 (0-9) 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

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 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 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 1C and Supplementary Table S1).

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 1D. For readability, we will use the human nomenclature for Cryptococcal spliceosome proteins throughout the manuscript (see Supplementary Table S1 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 2D is the impact of gene deletions on the levels of spliceosomal protein-encoding transcripts (see Supplementary Table S2 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 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 2B and 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. For the "change in Percent Splicing In" metric (Δψ), 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 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 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 3A (full dataset available in Supplementary Table S3). 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 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 lacking the ortholog of 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 3C). Again, cells lacking 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 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 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 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 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 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 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 5B). Numerous additional spliceosomal proteins were identified, including those characteristic of active C complex spliceosomes (Figure 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 Supplementary Table S4, 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 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 5B), indicating evolutionary conservation of this association. We also identified peptides corresponding to RBM5 (Figure 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 Supplementary Tables S5 and 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 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 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 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 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 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 mass 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.

### 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 Supplementary Table S7**)**. 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 30°C. 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 NaCl (low salt) or 300 mM NaCl (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 (Supplementary Table 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).

## FIGURE LEGENDS

**Figure 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 Supplementary Table S1.

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 Supplementary Table S1.

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 Supplementary Table S1.

**Figure 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 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 2C are plotted above the line while intron counts displaying negative changes in $\Delta\psi$ as defined in Figure 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. -$\log_{10}$(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. -$\log_{10}$(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. -$\log_{10}$(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:** 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 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 ($\log_{10}$) 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 ($\log_{10}$) 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 ($\log_{10}$) 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 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 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 $\Delta\psi$ (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 $\log_{10}$ 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 $\Delta\psi$ (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 $\log_{10}$ p-value. Strains were clustered by similarity.

## SUPPLEMENTARY TABLE LEGENDS

**Supplementary Table S1:** Core human spliceosomal proteins and fungal orthologs. Data are presented as an Excel file.

**Supplementary Table S2:** DESeq2 output of RNA-seq data. Data are presented as an Excel file.

**Supplementary Table S3:** Output of Junction Utilization Package (JUM) Analysis of 14 Mutant Strains. Data are presented as an Excel file.

**Supplementary Table S4:** Full TMT-MS Data for GPATCH1 immunopurifications. Data are presented as an Excel file. Raw, annotated, and spliceosomal protein data displayed in separate sheets.

**Supplementary Table S5:** Full TMT-MS Data for RBM17 immunopurifications. Data are presented as an Excel file. Raw, annotated, and spliceosomal protein data displayed in separate sheets.

**Supplementary Table S6:** 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.

**Supplementary Table S7:** Proteomes used for Evolutionary Analysis. Data are presented as an Excel file

## 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.

## REFERENCES

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

Grate, L., and Ares, M., Jr. (2002). Searching yeast intron data at Ares lab Web site. Methods Enzymol *350*, 380-392.

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.

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.

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.

Irimia, M., Penny, D., and Roy, S.W. (2007). Coevolution of genomic intron number and splice sites. Trends Genet *23*, 321-325.

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.

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.

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.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Pikielny, C.W., and Rosbash, M. (1985). mRNA splicing efficiency in yeast and the contribution of nonconserved sequences. Cell *41*, 119-126.

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.

Robert-Paganin, J., Rety, S., and Leulliot, N. (2015). Regulation of DEAH/RHA helicases by G-patch proteins. Biomed Res Int *2015*, 931857.

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.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

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.

Sharp, P.A. (1987). Splicing of messenger RNA precursors. Science *235*, 766-771.

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.

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.

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.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637.

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*.

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.

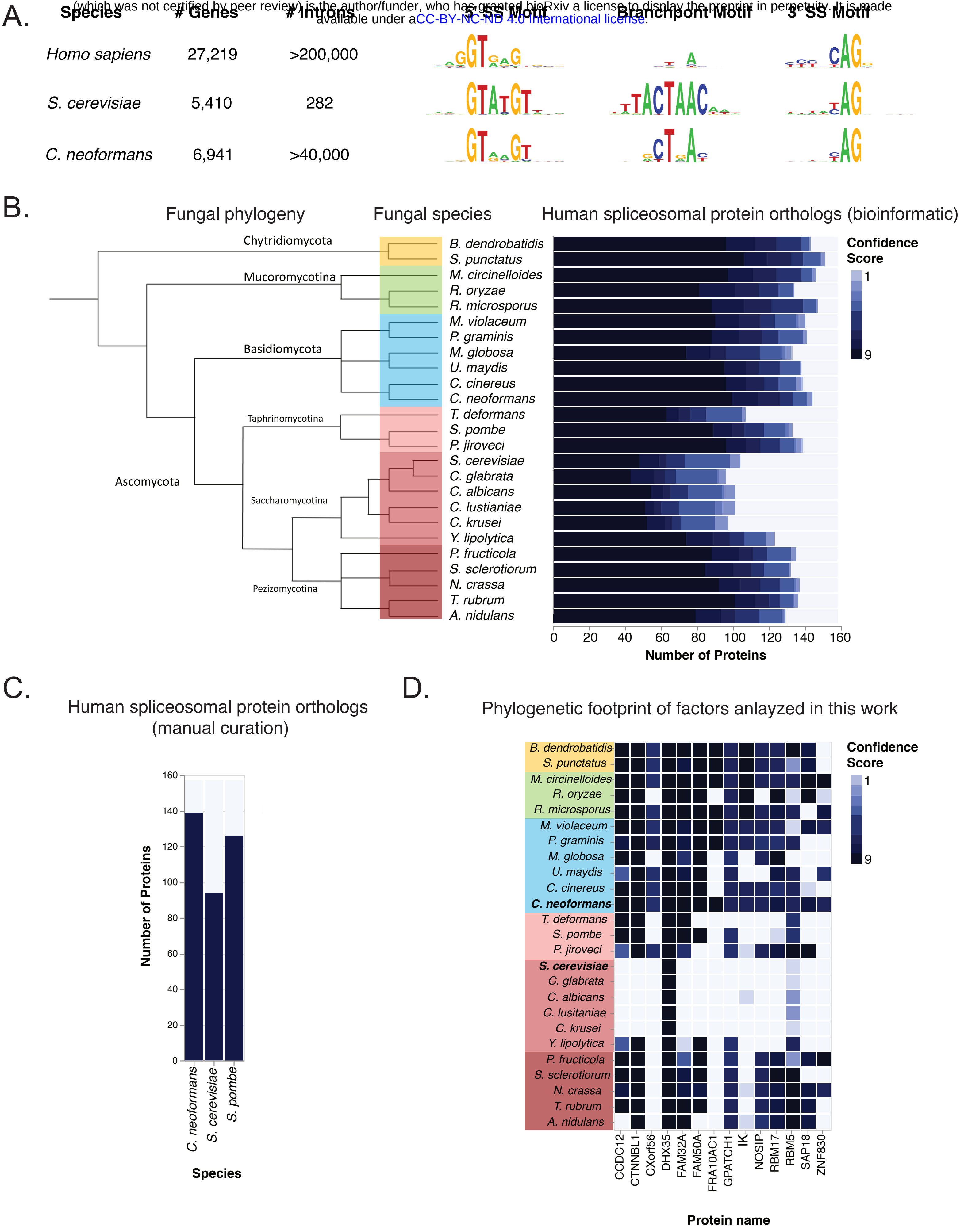
Wahl, M.C., and Luhrmann, R. (2015a). SnapShot: Spliceosome Dynamics I. Cell *161*, 1474-e1471.

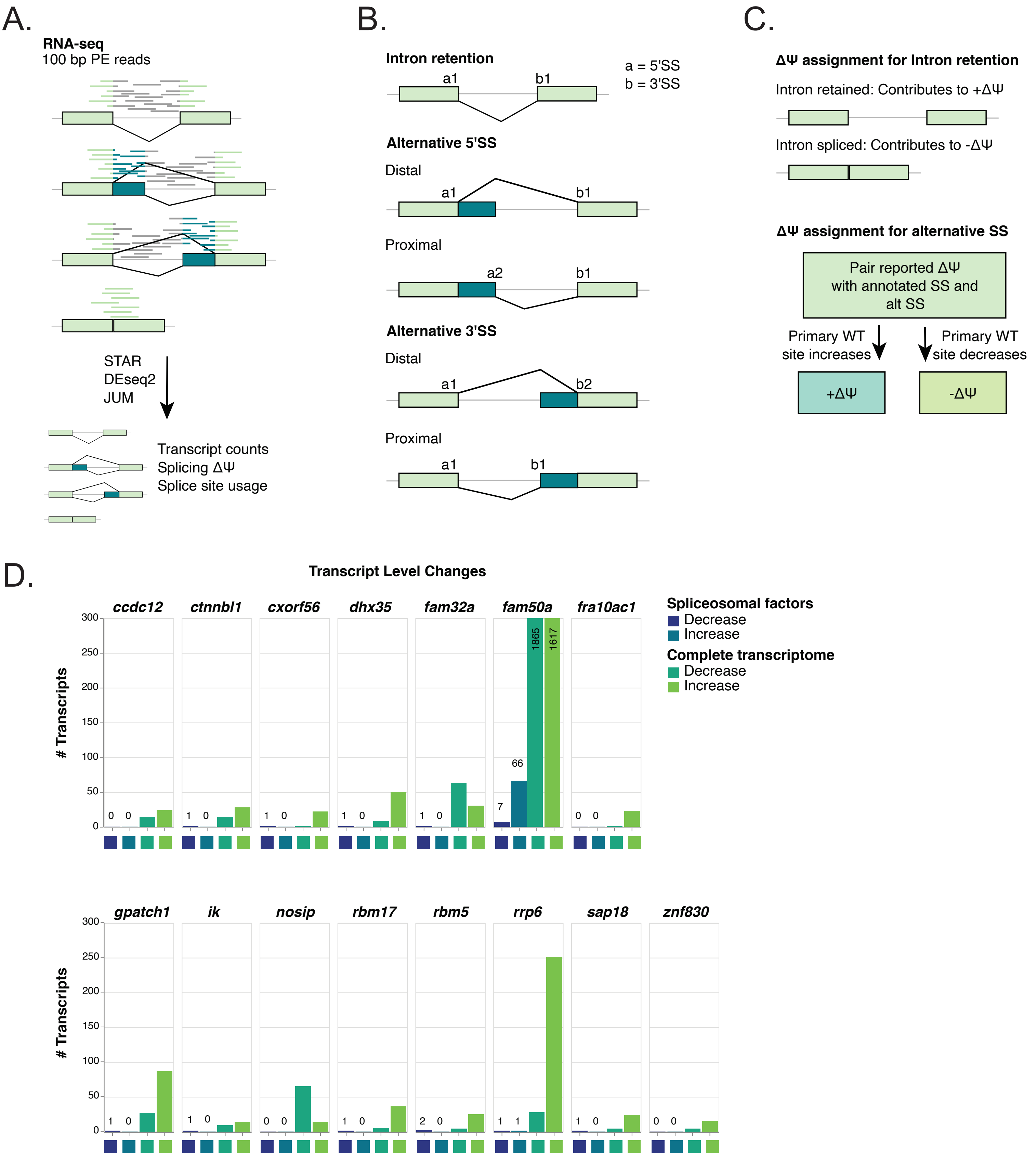
Wahl, M.C., and Luhrmann, R. (2015b). SnapShot: Spliceosome Dynamics II. Cell *162*, 456-456 e451.

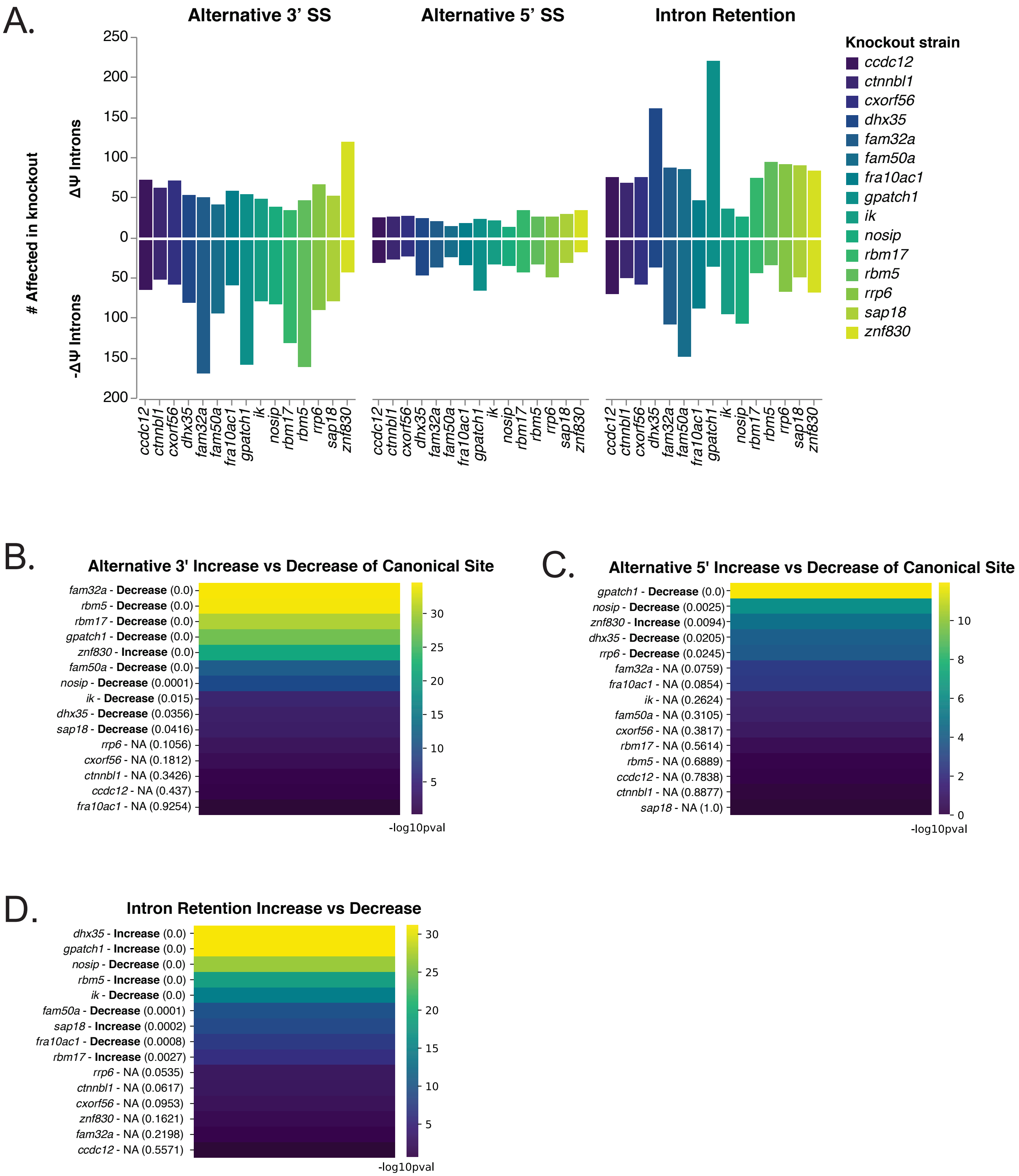
Wang, Q., and Rio, D.C. (2018). JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. Proc Natl Acad Sci U S A *115*, E8181-E8190.

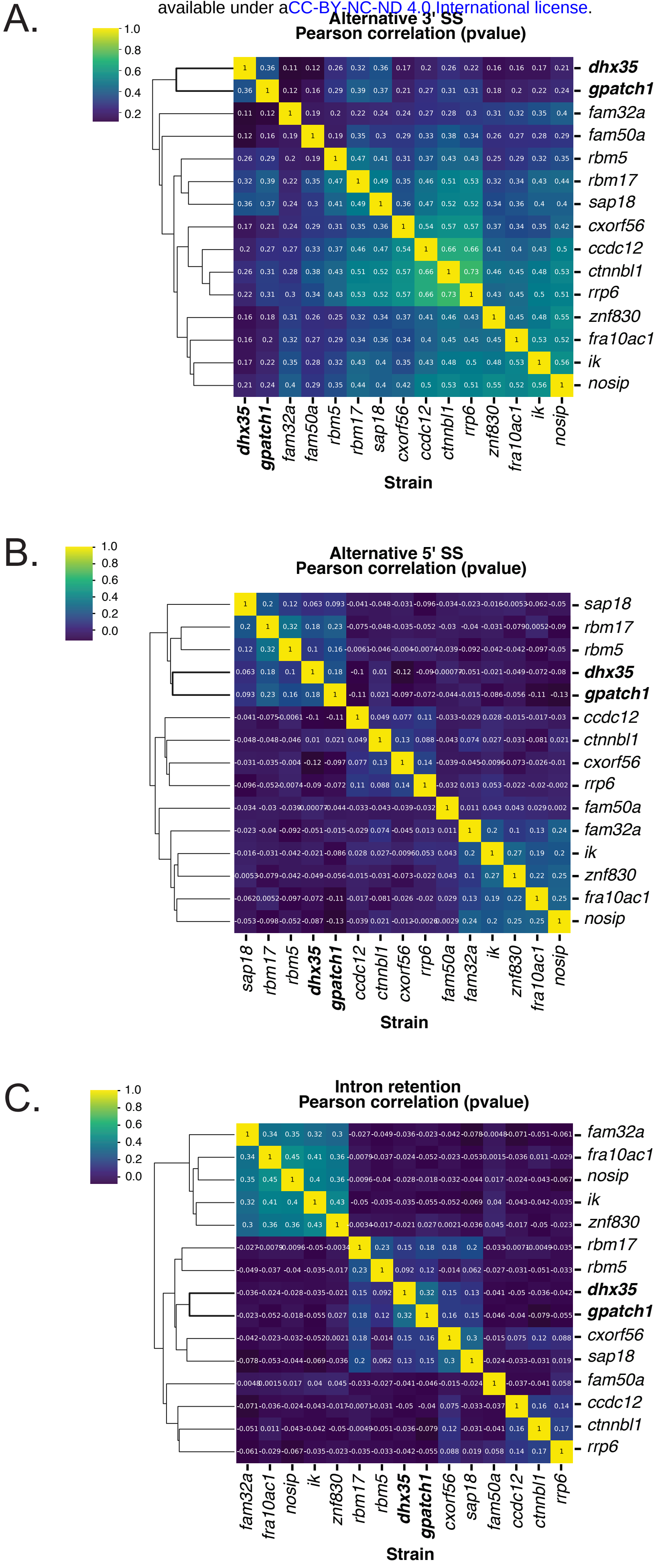
Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA Splicing by the Spliceosome. Annu Rev Biochem *89*, 359-388.

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.


A.
Species
# Genes
# Introns
5' SS Motif
Branchpoint Motif
3' SS Motif
Homo sapiens
27,219
>200,000
S. cerevisiae
5,410
282
C. neoformans
6,941
>40,000
B.
Fungal phylogeny
Fungal species
Human spliceosomal protein orthologs (bioinformatic)
Chytridiomycota
Mucoromycotina
Basidiomycota
Taphrinomycotina
Ascomycota
Saccharomycotina
Pezizomycotina
B. dendrobatidis
S. punctatus
M. circinelloides
R. oryzae
R. microsporus
M. violaceum
P. graminis
M. globosa
U. maydis
C. cinereus
C. neoformans
T. deformans
S. pombe
P. jiroveci
S. cerevisiae
C. glabrata
C. albicans
C. lustianiae
C. krusei
Y. lipolytica
P. fructicola
S. sclerotiorum
N. crassa
T. rubrum
A. nidulans
0
20
40
60
80
100
120
140
160
Number of Proteins
Confidence Score
1
9
C.
Human spliceosomal protein orthologs (manual curation)
Number of Proteins
0
20
40
60
80
100
120
140
160
C. neoformans
S. cerevisiae
S. pombe
Species
D.
Phylogenetic footprint of factors anlayzed in this work
B. dendrobatidis
S. punctatus
M. circinelloides
R. oryzae
R. microsporus
M. violaceum
P. graminis
M. globosa
U. maydis
C. cinereus
C. neoformans
T. deformans
S. pombe
P. jiroveci
S. cerevisiae
C. glabrata
C. albicans
C. lusitaniae
C. krusei
Y. lipolytica
P. fructicola
S. sclerotiorum
N. crassa
T. rubrum
A. nidulans
CCDC12
CTNNBL1
CXorf56
DHX35
FAM32A
FAM50A
FRA10AC1
GPATCH1
IK
NOSIP
RBM17
RBM5
SAP18
ZNF830
Protein name
Confidence Score
1
9


A.
RNA-seq
100 bp PE reads
STAR
DEseq2
JUM
Transcript counts
Splicing ΔΨ
Splice site usage
B.
Intron retention
a1
b1
a = 5'SS
b = 3'SS
Alternative 5'SS
Distal
a1
b1
Proximal
a2
b1
Alternative 3'SS
Distal
a1
b2
Proximal
a1
b1
C.
ΔΨ assignment for Intron retention
Intron retained: Contributes to +ΔΨ
Intron spliced: Contributes to -ΔΨ
ΔΨ assignment for alternative SS
Pair reported ΔΨ with annotated SS and alt SS
Primary WT site increases
Primary WT site decreases
+ΔΨ
-ΔΨ
D.
Transcript Level Changes
# Transcripts
0
50
100
150
200
250
300
ccdc12
0 0
ctnnbl1
1 0
cxorf56
1 0
dhx35
1 0
fam32a
1 0
fam50a
7
66
1865
1617
fra10ac1
0 0
Spliceosomal factors
Decrease
Increase
Complete transcriptome
Decrease
Increase
gpatch1
1 0
ik
1 0
nosip
0 0
rbm17
1 0
rbm5
2 0
rrp6
1 1
sap18
1 0
znf830
0 0

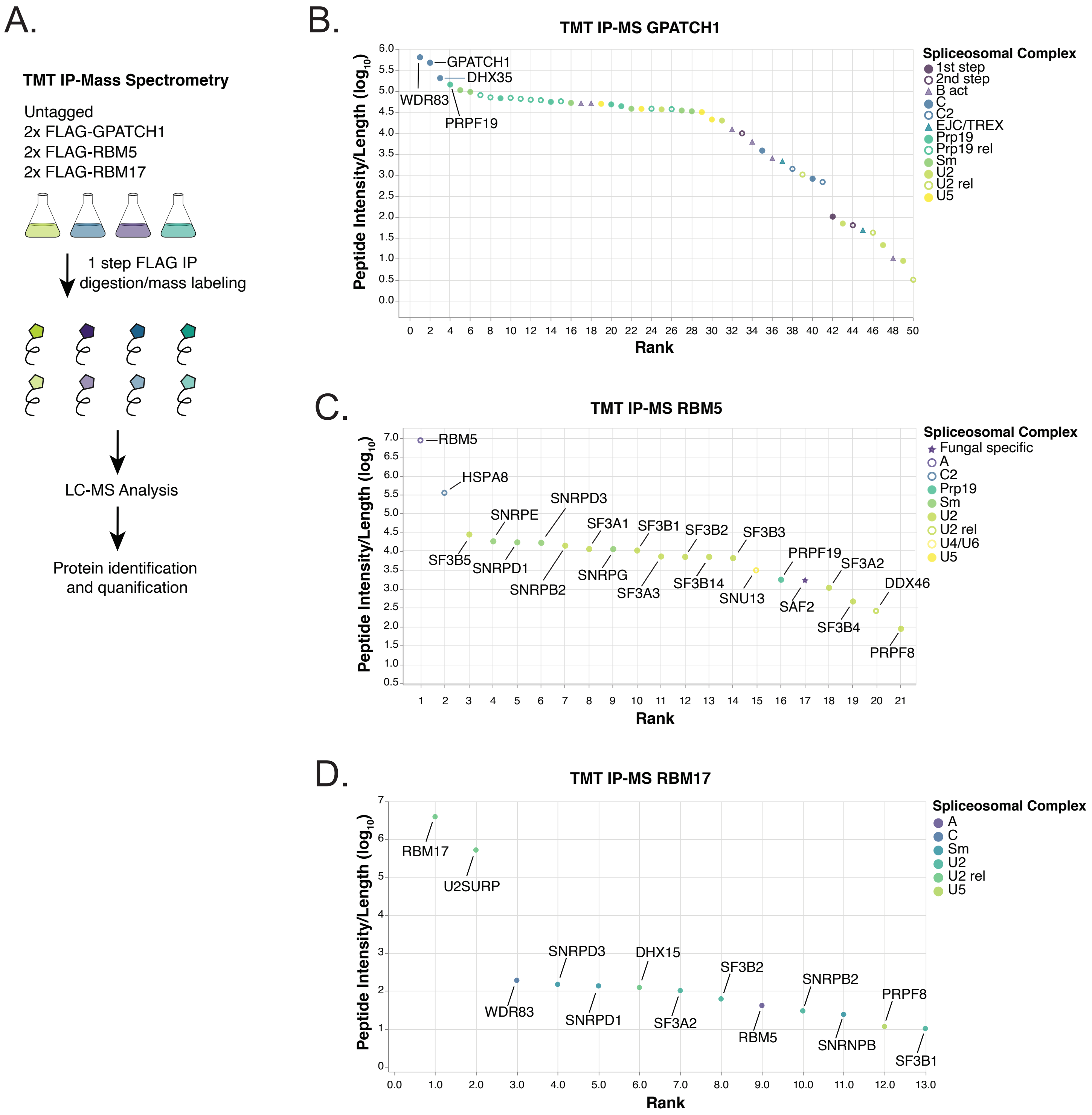

A.
Alternative 3' SS
Alternative 5' SS
Intron Retention
# Affected in knockout
ΔΨ Introns
-ΔΨ Introns
250
200
150
100
50
0
50
100
150
200
ccdc12
ctnnbl1
cxorf56
dhx35
fam32a
fam50a
fra10ac1
gpatch1
ik
nosip
rbm17
rbm5
rrp6
sap18
znf830
Knockout strain
B.
Alternative 3' Increase vs Decrease of Canonical Site
fam32a - Decrease (0.0)
rbm5 - Decrease (0.0)
rbm17 - Decrease (0.0)
gpatch1 - Decrease (0.0)
znf830 - Increase (0.0)
fam50a - Decrease (0.0)
nosip - Decrease (0.0001)
ik - Decrease (0.015)
dhx35 - Decrease (0.0356)
sap18 - Decrease (0.0416)
rrp6 - NA (0.1056)
cxorf56 - NA (0.1812)
ctnnbl1 - NA (0.3426)
ccdc12 - NA (0.437)
fra10ac1 - NA (0.9254)
30
25
20
15
10
5
-log10pval
C.
Alternative 5' Increase vs Decrease of Canonical Site
gpatch1 - Decrease (0.0)
nosip - Decrease (0.0025)
znf830 - Increase (0.0094)
dhx35 - Decrease (0.0205)
rrp6 - Decrease (0.0245)
fam32a - NA (0.0759)
fra10ac1 - NA (0.0854)
ik - NA (0.2624)
fam50a - NA (0.3105)
cxorf56 - NA (0.3817)
rbm17 - NA (0.5614)
rbm5 - NA (0.6889)
ccdc12 - NA (0.7838)
ctnnbl1 - NA (0.8877)
sap18 - NA (1.0)
10
8
6
4
2
0
-log10pval
D.
Intron Retention Increase vs Decrease
dhx35 - Increase (0.0)
gpatch1 - Increase (0.0)
nosip - Decrease (0.0)
rbm5 - Increase (0.0)
ik - Decrease (0.0)
fam50a - Decrease (0.0001)
sap18 - Increase (0.0002)
fra10ac1 - Decrease (0.0008)
rbm17 - Increase (0.0027)
rrp6 - NA (0.0535)
ctnnbl1 - NA (0.0617)
cxorf56 - NA (0.0953)
znf830 - NA (0.1621)
fam32a - NA (0.2198)
ccdc12 - NA (0.5571)
30
25
20
15
10
5
-log10pval

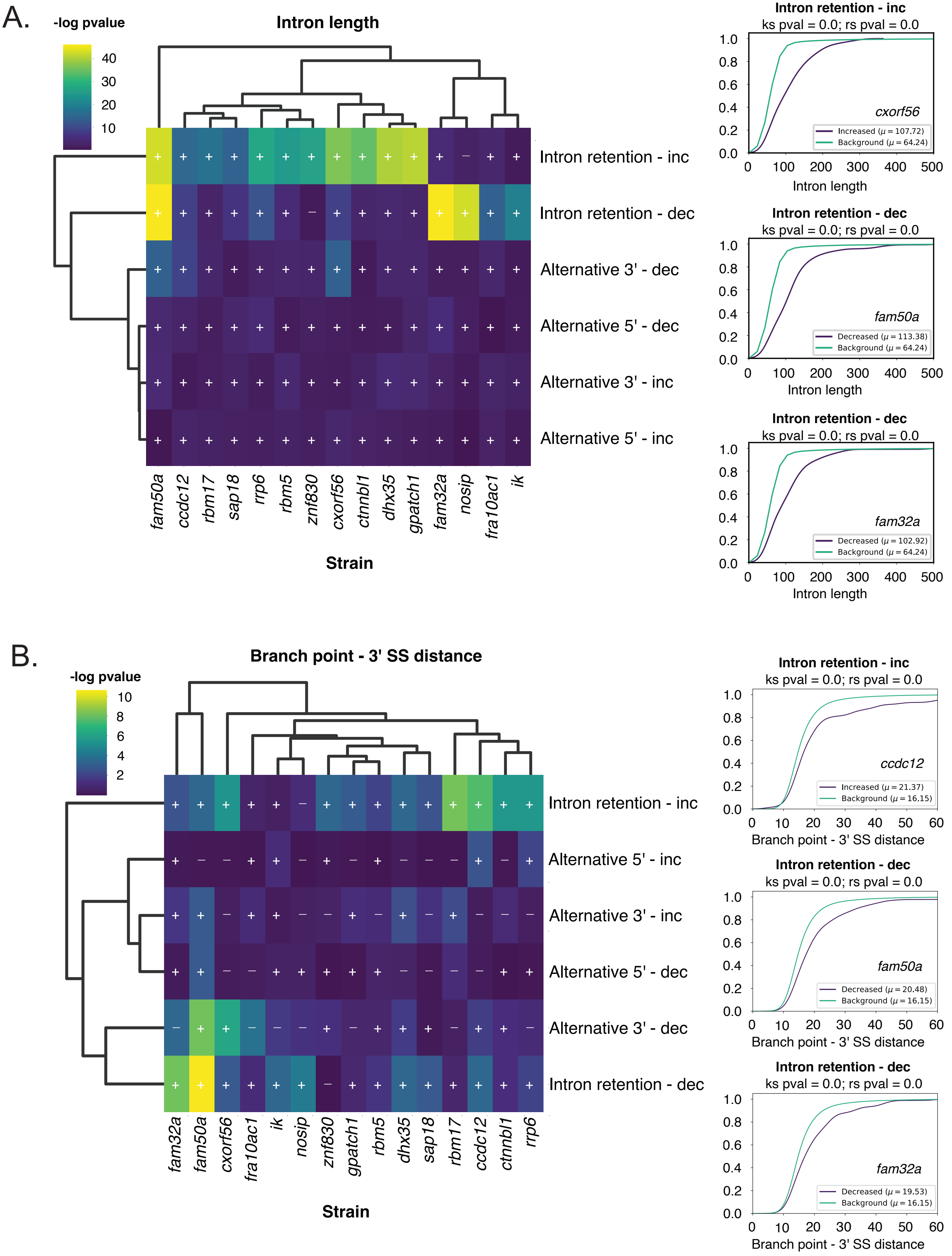


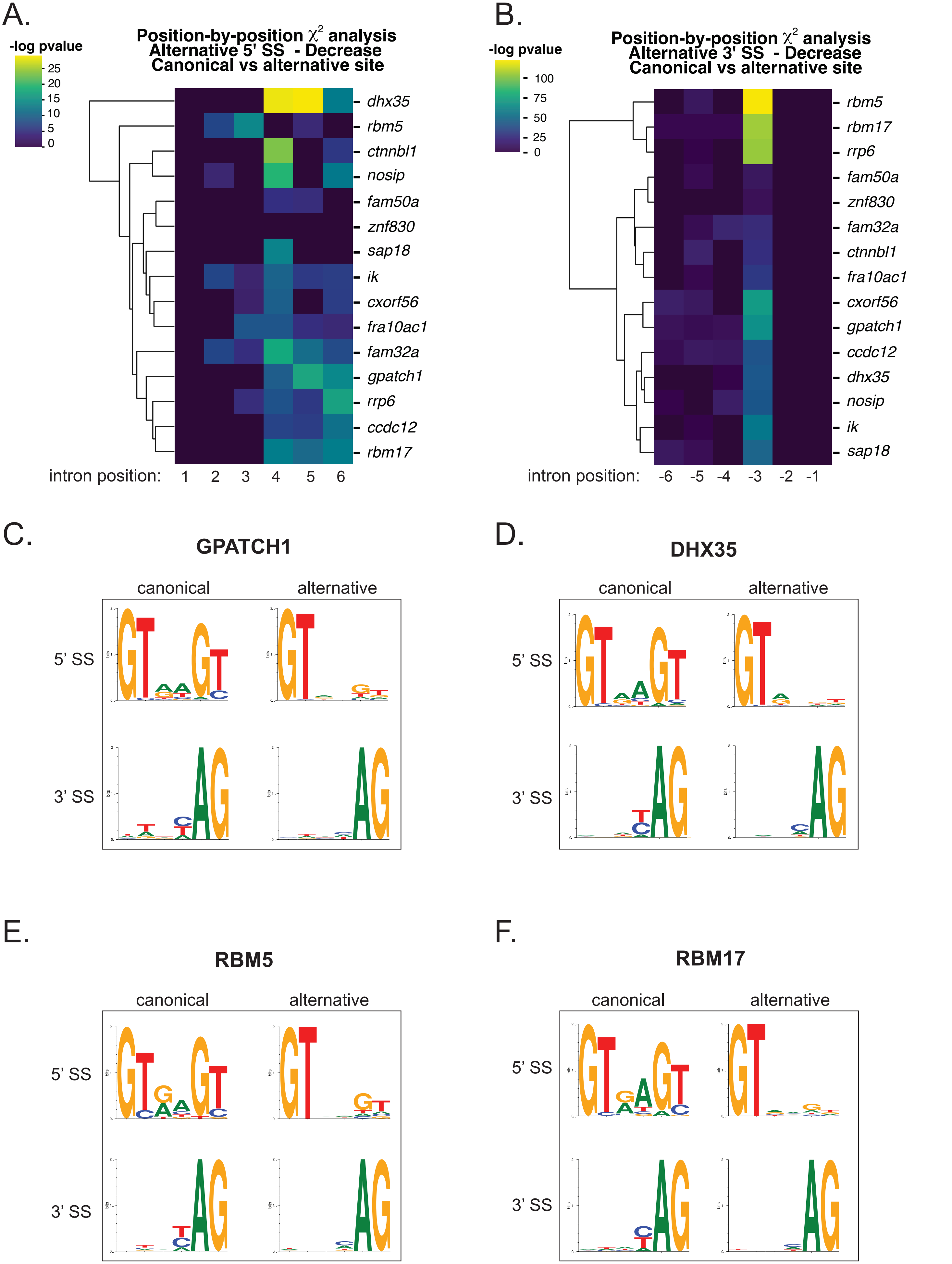
A. Alternative 3' SS
Pearson correlation (pvalue)

B. Alternative 5' SS
Pearson correlation (pvalue)

C. Intron retention
Pearson correlation (pvalue)


A.
TMT IP-Mass Spectrometry
Untagged
2x FLAG-GPATCH1
2x FLAG-RBM5
2x FLAG-RBM17
1 step FLAG IP
digestion/mass labeling
LC-MS Analysis
Protein identification
and quanification
B.
TMT IP-MS GPATCH1
Peptide Intensity/Length (log10)
Rank
GPATCH1
DHX35
WDR83
PRPF19
Spliceosomal Complex
1st step
2nd step
B act
C
C2
EJC/TREX
Prp19
Prp19 rel
Sm
U2
U2 rel
U5
C.
TMT IP-MS RBM5
Peptide Intensity/Length (log10)
Rank
RBM5
HSPA8
SNRPD3
SNRPE
SF3A1
SF3B1
SF3B2
SF3B3
SF3B5
SNRPD1
SNRPB2
SNRPG
SF3A3
SF3B14
SNU13
PRPF19
SF3A2
SAF2
DDX46
SF3B4
PRPF8
Spliceosomal Complex
Fungal specific
A
C2
Prp19
Sm
U2
U2 rel
U4/U6
U5
D.
TMT IP-MS RBM17
Peptide Intensity/Length (log10)
Rank
RBM17
U2SURP
WDR83
SNRPD3
SNRPD1
DHX15
SF3A2
SF3B2
RBM5
SNRPB2
SNRNPB
PRPF8
SF3B1
Spliceosomal Complex
A
C
Sm
U2
U2 rel
U5


A.
-log pvalue
40
30
20
10
Intron length
Intron retention - inc
Intron retention - dec
Alternative 3' - dec
Alternative 5' - dec
Alternative 3' - inc
Alternative 5' - inc
fam50a
ccdc12
rbm17
sap18
rrp6
rbm5
znf830
cxorf56
ctnnbl1
dhx35
gpatch1
fam32a
nosip
fra10ac1
ik
Strain
Intron retention - inc
ks pval = 0.0; rs pval = 0.0
cxorf56
Increased (μ = 107.72)
Background (μ = 64.24)
Intron length
Intron retention - dec
ks pval = 0.0; rs pval = 0.0
fam50a
Decreased (μ = 113.38)
Background (μ = 64.24)
Intron length
Intron retention - dec
ks pval = 0.0; rs pval = 0.0
fam32a
Decreased (μ = 102.92)
Background (μ = 64.24)
Intron length
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
B.
-log pvalue
10
8
6
4
2
Branch point - 3' SS distance
Intron retention - inc
Alternative 5' - inc
Alternative 3' - inc
Alternative 5' - dec
Alternative 3' - dec
Intron retention - dec
fam32a
fam50a
cxorf56
fra10ac1
ik
nosip
znf830
gpatch1
rbm5
dhx35
sap18
rbm17
ccdc12
ctnnbl1
rrp6
Strain
Intron retention - inc
ks pval = 0.0; rs pval = 0.0
ccdc12
Increased (μ = 21.37)
Background (μ = 16.15)
Branch point - 3' SS distance
Intron retention - dec
ks pval = 0.0; rs pval = 0.0
fam50a
Decreased (μ = 20.48)
Background (μ = 16.15)
Branch point - 3' SS distance
Intron retention - dec
ks pval = 0.0; rs pval = 0.0
fam32a
Decreased (μ = 19.53)
Background (μ = 16.15)
Branch point - 3' SS distance
0
10
20
30
40
50
60


A.
Position-by-position χ² analysis
Alternative 5' SS - Decrease
Canonical vs alternative site
-log pvalue
25
20
15
10
5
0
dhx35
rbm5
ctnnbl1
nosip
fam50a
znf830
sap18
ik
cxorf56
fra10ac1
fam32a
gpatch1
rrp6
ccdc12
rbm17
intron position:
1
2
3
4
5
6
B.
Position-by-position χ² analysis
Alternative 3' SS - Decrease
Canonical vs alternative site
-log pvalue
100
75
50
25
0
rbm5
rbm17
rrp6
fam50a
znf830
fam32a
ctnnbl1
fra10ac1
cxorf56
gpatch1
ccdc12
dhx35
nosip
ik
sap18
intron position:
-6
-5
-4
-3
-2
-1
C.
GPATCH1
canonical
alternative
5’ SS
3’ SS
D.
DHX35
canonical
alternative
5’ SS
3’ SS
E.
RBM5
canonical
alternative
5’ SS
3’ SS
F.
RBM17
canonical
alternative
5’ SS
3’ SS