Article

Current Biology

Coupling of spliceosome complexity to intron diversity

Highlights

- Phylogenetic analysis reveals the ancestral spliceosome was complex
- Human spliceosomal protein orthologs lost in *S. cerevisiae* are found in *C. neoformans*
- Functional analysis in *C. neoformans* demonstrates roles in splicing fidelity
- Proteomic and genetic analysis reveals functional modules

Authors

Jade Sales-Lee, Daniela S. Perry, Bradley A. Bowser, ..., John R. Yates III, Scott W. Roy, Hiten D. Madhani

Correspondence

scottwroy@gmail.com (S.W.R.), hitenmadhani@gmail.com (H.D.M.)

In brief

Sales-Lee et al. show that the ancestral spliceosome was complex and the intron-reduced yeast *S. cerevisiae* lost dozens of spliceosomal proteins maintained in intron-rich fungi and humans. Functional and proteomic analysis in the intron-rich yeast *C. neoformans* reveals roles for these proteins in splicing fidelity and uncovers functional modules.

Sales-Lee et al., 2021, Current Biology *31*, 4898–4910
November 22, 2021
https://doi.org/10.1016/j.cub.2021.09.004

CellPress

CellPress

Current Biology

Article

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,*] and Hiten D. Madhani[1,5,6,7,*]

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94158, USA
[2]Department of Molecular and Cellular Biology, University of California, Merced, Merced, CA 95343, USA
[3]Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA
[4]Department of Biology, San Francisco State University, San Francisco, CA 94132, USA
[5]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
[6]Twitter: @hitenmadhani
[7]Lead contact
*Correspondence: scottwroy@gmail.com (S.W.R.), hitenmadhani@gmail.com (H.D.M.)
https://doi.org/10.1016/j.cub.2021.09.004

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 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.[1] Two sequential transesterification steps mediate intron removal. Pre-mRNA splicing by the spliceosome seems complex for a process that removes a segment of RNA from a precursor. Splicing requires eight ATP-dependent steps and about 90 proteins in *Saccharomyces cerevisiae*. Much of our functional understanding derives from the analysis of conditional and null mutants in *S. cerevisiae*.[1] Human spliceosomes appear to contain about 60 additional proteins.[2,3] The reason for this added complexity is not understood.

Structures of the core portion of the spliceosome at various stages of its cycle have been elucidated using cryoelectron microscopy (cryo-EM).[1] Many structures have been obtained using *in-vitro*-assembled spliceosomes using extracts from the budding yeast *S. cerevisiae* or from HeLa cells. Although the structure of the 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. Although 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*.[4–6] 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*.[7] As in other lineages, such loss events correlate with intron signals moving toward optimal intron signals. Thus, as introns are lost, intron signals become homogeneous 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 describe below phylogenetic, functional, and proteomic investigations of this question.

RESULTS

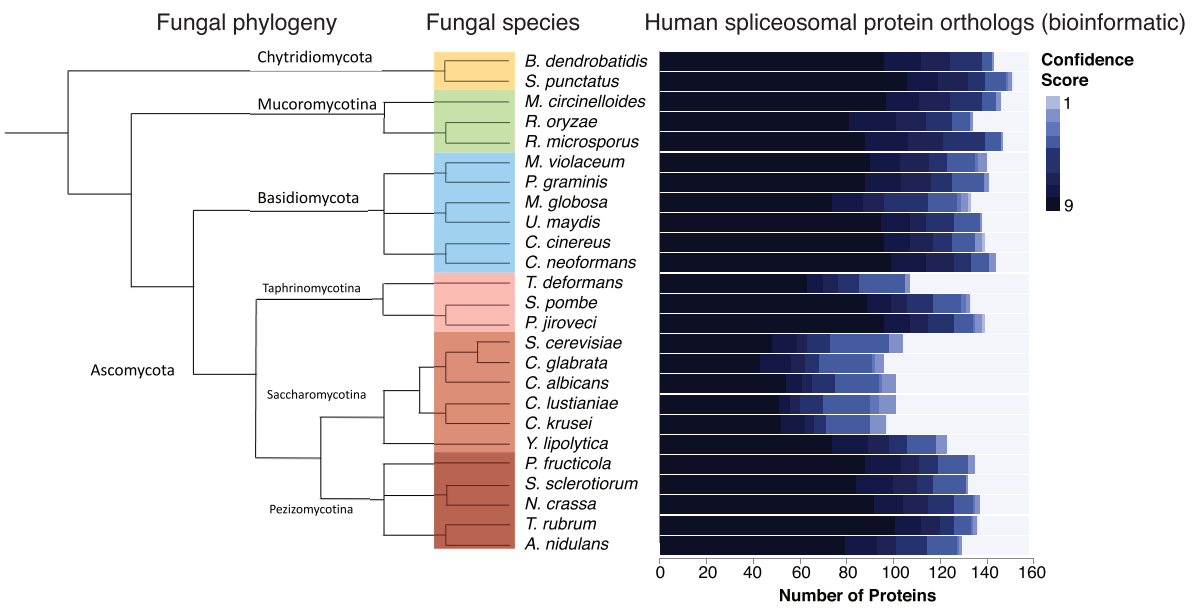
Maintenance of many dozens of human spliceosomal orthologs in fungal lineages

Cryptococcus neoformans offers a genetically tractable intron-rich haploid organism in which to investigate fundamental aspects of gene expression. We highlight the differences between

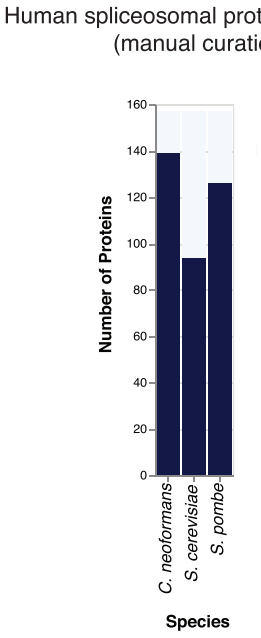
A

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



C



D

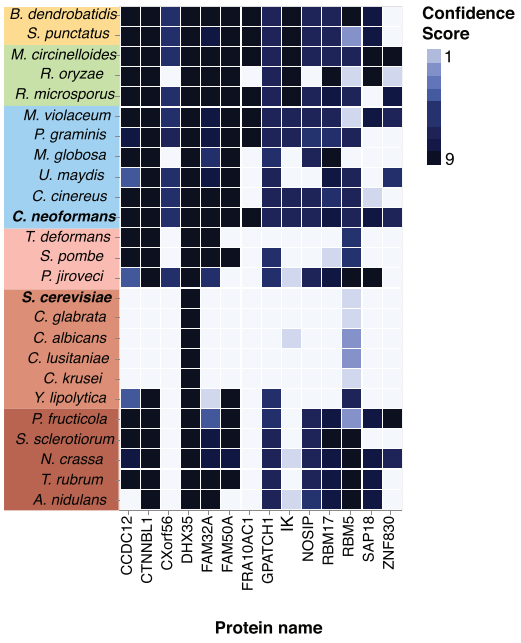


Figure 1. Massive loss of human spliceosomal protein orthologs in specific fungal lineages

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

(B) Evolutionary loss events. Phylogeny is based on James et al. (2020)[12]. See also Data S1 and Table S2.

(legend continued on next page)

the intron sequences and abundance between *S. cerevisiae*, *C. neoformans*, and *Homo sapiens* in Figure 1A. The *S. cerevisiae* genome is estimated to encode 282 introns[8] spread over 5,410 annotated genes (0.05 introns/gene), whereas *C. neoformans* (H99 strain) has 6,941 annotated protein-coding genes harboring over 40,000 introns,[9,10] comparable to humans (8 introns/gene), with 27,219 annotated genes and over 200,000 introns.[11] The sequences of *C. neoformans* 5′ splice sites and branchpoints are more variable than those of *S. cerevisiae*, suggesting its spliceosomes, like those of humans, may be more flexible in substrate utilization (Figure 1A).

We asked whether the loss of introns in the *Saccharomycotina*, the subphylum to which *S. cerevisiae* belongs, is accompanied by a loss of spliceosomal protein orthologs. We compiled a list of all spliceosome components reproducibly detected through mass spectrometry (MS), interaction studies, and/or purified and visualized in the spliceosome in structural biology studies.[2,3,13] This list includes 157 human proteins (Data 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. We generated a confidence score (0–9) for the presence of an ortholog in a given species (see STAR 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). This pipeline did not identify any duplicated paralogs. 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 of 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.[14] For three fungal species of interest (*S. cerevisiae*, *S. pombe*, and *C. neoformans*), we performed literature curation of the spliceosome (including our past studies of purified cryptococcal spliceosomes[15]) and an available experimentally curated database.[13] 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; Data S1).

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* 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 had no detectable effect on splicing but instead nucleolar ribosomal RNA processing.[16] We identified the cryptococcal ortholog of DHX35 previously in purified *C. neoformans* spliceosomes[15] and 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 (see Data 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 sequencing (RNA-seq). Samples were grown in duplicate and paired with wild-type samples grown on the same day to the same optical density. Paired-end 100-nt reads were obtained.

Limited impact of spliceosomal protein-null mutations on global transcript abundance

The overall scheme for the analysis of the RNA-seq data is shown in Figure 2, which includes analysis of transcript counts and splicing changes. 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.[17] Shown in Figure 2D is the impact of gene deletions on the levels of spliceosomal protein-encoding transcripts (see Table S1 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. Although FAM50A is a spliceosomal protein in humans, it has also been linked to transcription,[18] suggesting a pleiotropic role. Consistent with this, we observed that many genes display transcript-level changes in this mutant, whereas 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

To examine splicing changes (Figures 2A–2C), we used the junctional utilization method (JUM).[19] Using a stringent read count

(C) Numbers of human spliceosomal protein orthologs in *S. cerevisiae*, *S. pombe*, and *C. neoformans*.
(D) Spliceosomal factor orthologs for which null mutations in *C. neoformans* were obtained. Confidence scores for the presence of the indicated human spliceosomal protein orthologs in the indicated species are shown.

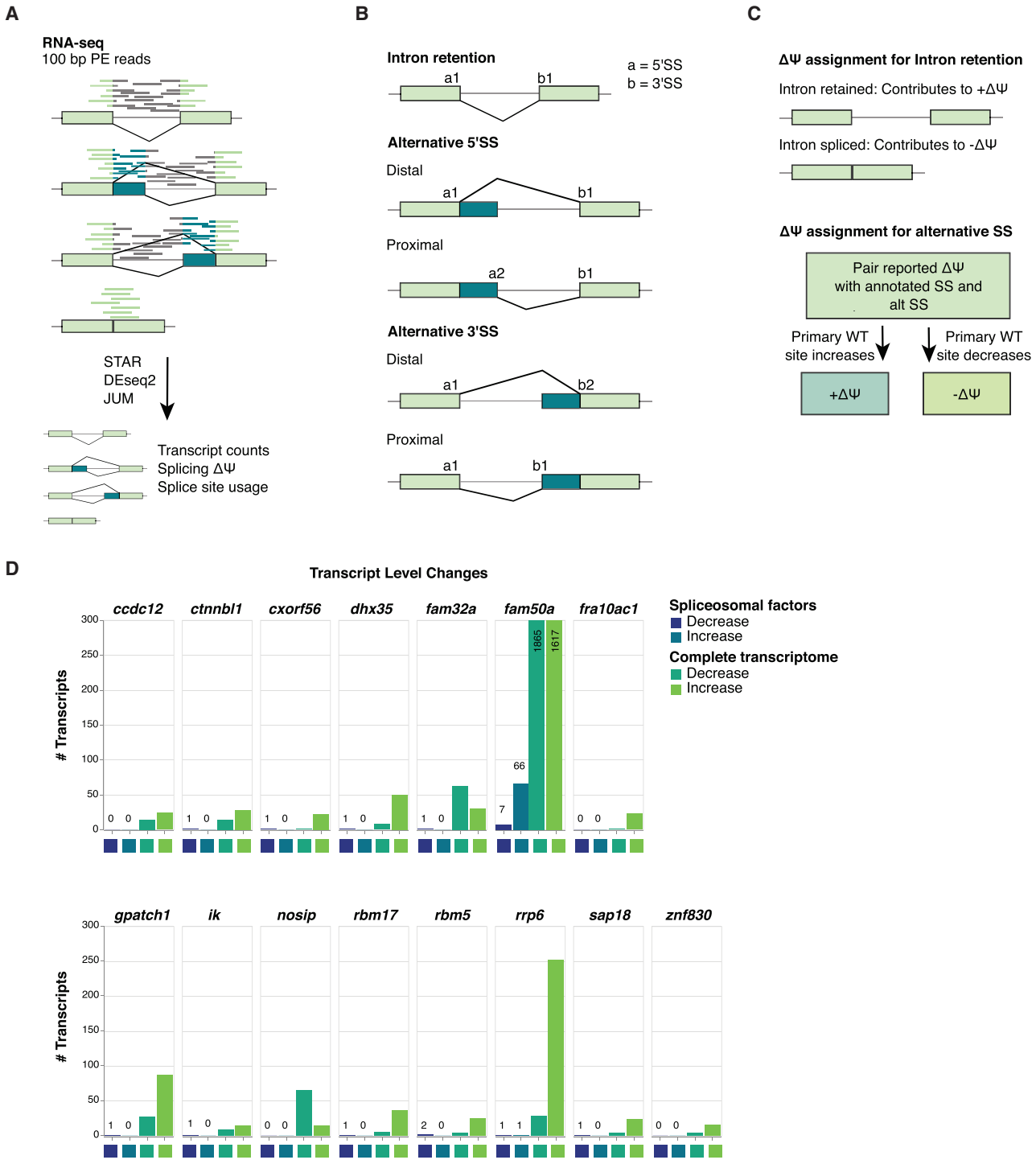


Figure 2. RNA-seq analysis of null mutations in 14 human spliceosomal protein orthologs

(A) Schematic of the RNA-seq pipeline.

(B) Definitions of the splicing events quantified.

(C) Definitions of changes to percent spliced in (delta PSI; Δψ) values for intron retention, alternative 3′ splice site, and alternative 5′ splice site events.

(D) Changes in transcript levels in mutants. Plotted is the total number of splicing factors changed in the RNA-seq data as well as total transcriptome changes. See also Table S1.

and p value cutoffs (see STAR Methods), we quantified splicing changes in each of the 14 gene deletion mutants described above. Because we did not identify instances of mutually exclusive exons and only a handful of cassette exons, we excluded these two categories, along with the "complex splicing" category, from our downstream analysis.

As diagrammed in Figures 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 number of retained intron transcripts (i.e., precursor) can be increased or decreased relative to mRNA. For the "change in percent splicing in" metric ($\Delta\psi$), a *positive* value corresponds to an *increase* in intron retention in a mutant, whereas a *decrease* in intron retention produces a *negative* $\Delta\psi$ value (Figures 2A–2C). For changes in the 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, whereas the alternative site was either unannotated or annotated as an alternative site in the current *C. neoformans* H99 strain genome annotation. For alternative 5′ or 3′ splice site changes, a *decrease* of usage of the preferred site in a mutant produces a *negative* $\Delta\psi$, whereas an *increase* in the usage of the preferred site in a mutant produces a *positive* $\Delta\psi$ value (Figures 2B and 2C).

For each mutant, we quantified the effects across alternative splicing events in the *C. neoformans* genome and tabulated these data across events. The results of this analysis are shown in Figure 3A (the dataset is available in Data S2; see Figure S1 for RT-PCR validation). Plotted is 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 such a way as to produce a positive $\Delta\psi$ value as defined above, whereas 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. 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 toward a negative $\Delta\psi$ for 3′ and 5′ splice site choice, indicating a decrease in the use of the canonical (preferred in wild-type) 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 toward increasing intron retention, consistent with increased splicing defects (increased pre-mRNA versus 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 a statistically significant bias toward 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 (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 toward 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 toward 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 toward 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, a 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).

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. We observed that GPATCH1 and DHX35 were among the mutants that clustered together in three matrices, suggesting they consistently impact overlapping intron sets. We also noticed that RBM5 and RBM17 also tended to cluster together in the alternative 5′ splice site usage data and the intron retention data (Figures 4B and 4C). Human GPATCH1 and DHX35 have both been identified in C complex spliceosomes assembled *in vitro*,[20] whereas RBM5 and RBM17 have been found in the early A complex that includes the U2 snRNP.[21] 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), a core component of active human spliceosomes recently visualized by cryo-EM,[22] 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 suggest 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) MS analysis (Figure 5A). We then ranked proteins based on relative peptide counts/protein lengths for all identified proteins. Remarkably, the next most abundant protein in the GPATCH1 IP was DHX35 (Figure 5B). The additional spliceosomal proteins were identified, including those characteristic of active C complex spliceosomes (Figure 5B), suggesting that, as in human cells,

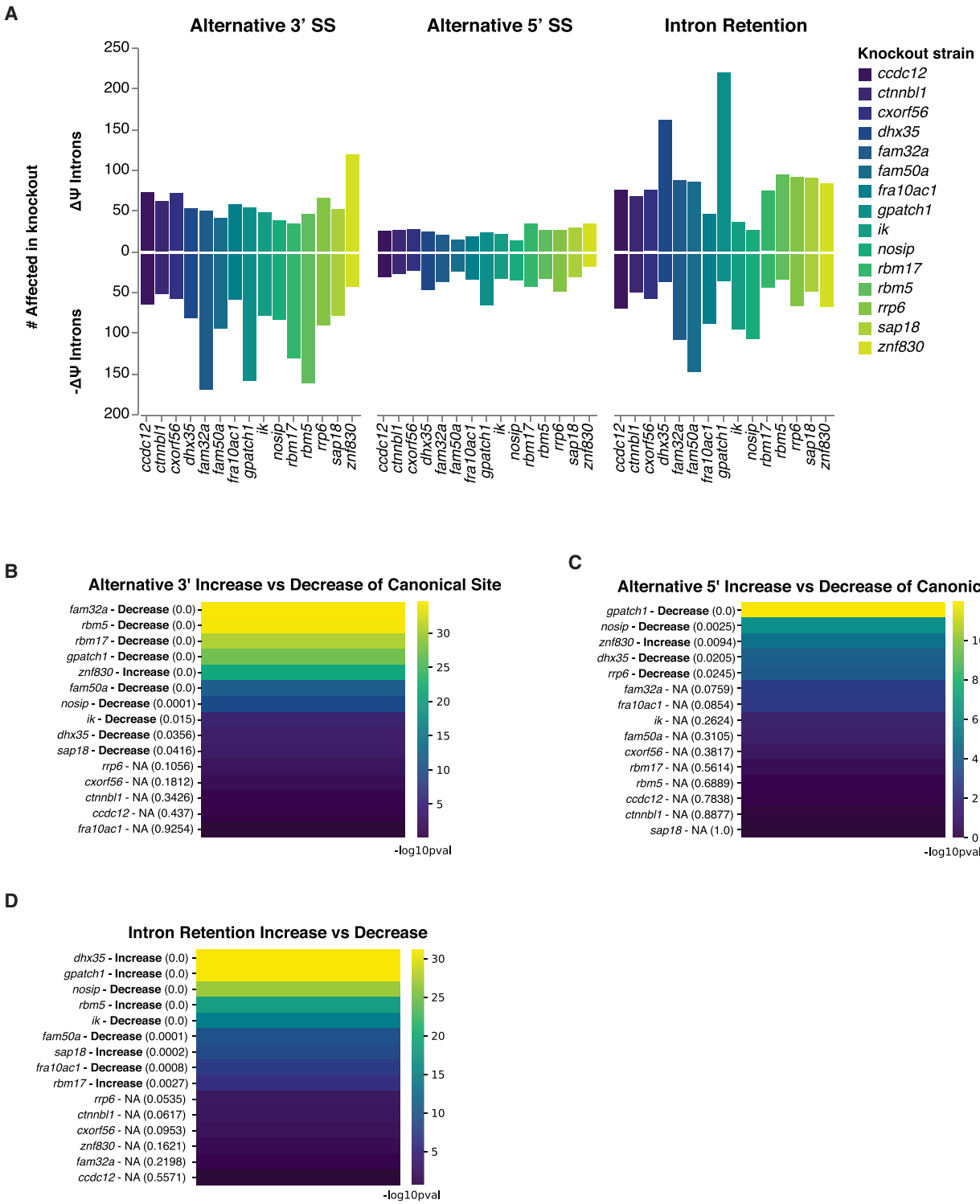


Figure 3. Quantification of altered pre-mRNA splicing in mutants lacking orthologs of human spliceosomal proteins

(A) Number of introns altered in pre-mRNA splicing in mutants. Changes in splicing are plotted as a count of the number of introns with significant $\Delta\psi$ ($p < 0.05$) values. See also Figure S1, Table S3, and Data S2.

(B) Binomial test for directionality of alternative 3′ splice site usage changes. Introns affected by each knockout (KO) strain were analyzed to test for a bias toward positive or negative $\Delta\psi$. The 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 and whether the bias reflects a decrease or increase in the canonical splice site in the mutant, with the p value shown in parentheses.

(C) Binomial test for directionality of alternative 5′ splice site usage changes.

(D) Binomial test for directionality of intron retention changes.

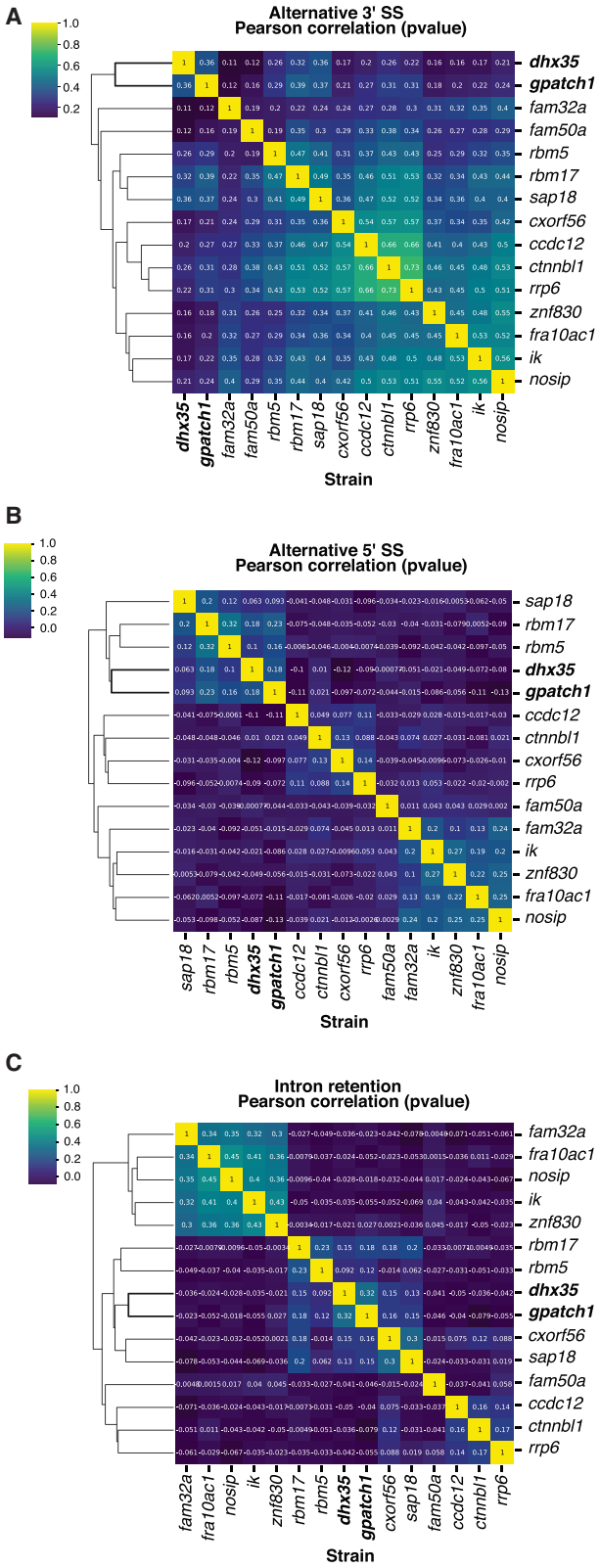


Figure 4. Correlation between phenotypic signatures of spliceosomal protein ortholog gene deletion mutants

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. 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.

GPATCH1 and DHX35 associate with active spliceosomes in *C. neoformans*. A poorly studied protein, WDR83, displayed higher normalized abundance than GPATCH1 (Figure 5B); the significance of this finding will require additional experimental work. The dataset can be found in Data S3.

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.[21] RBM17 has been found to associate with U2SURP and CHERP in IP-MS studies from human cells.[23] 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. However, because RBM17 was not identified in the RBM5 purification, the degree to which and mechanism by which they might act together remain to be determined. Datasets for the RBM5 and RBM17 purifications can be found in Data S4 and S5. These data indicate that the clustering of factors based on their impact on splicing choice and efficiency can be informative.

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 STAR 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. We asked whether the intron length distributions of affected versus unaffected introns differed 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 toward affecting introns with longer lengths (Figure 6A; a clustered heatmap of corrected p values is shown in the left panel and the 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

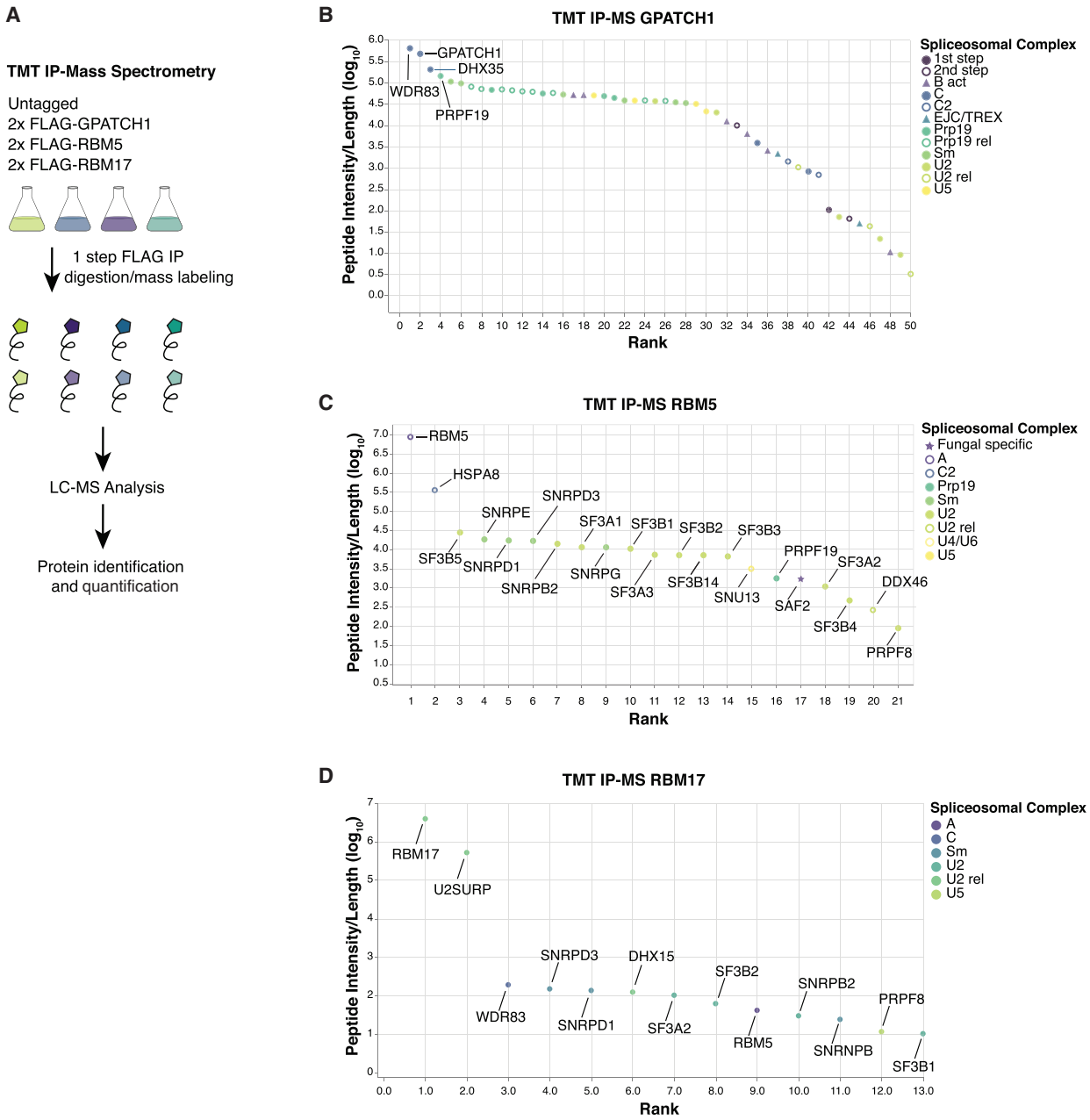


Figure 5. Purifications and TMT-MS analysis of endogenously tagged human spliceosomal protein orthologs

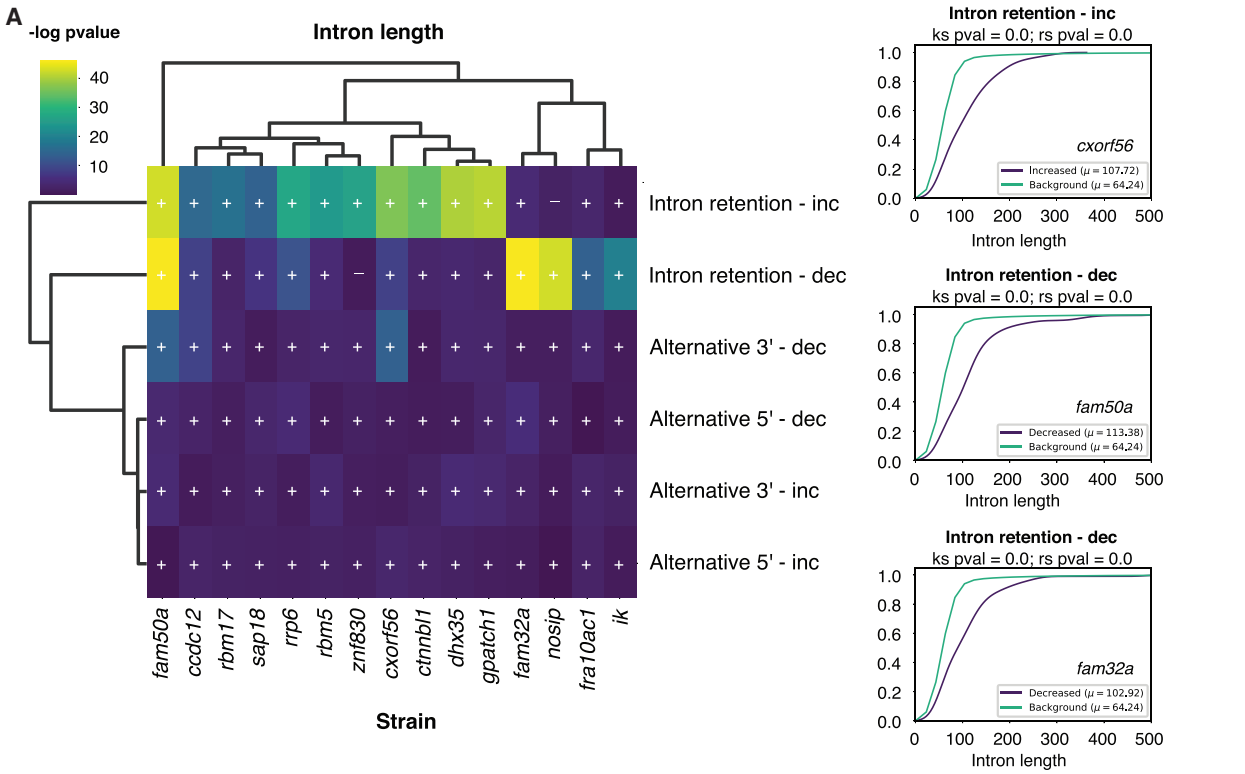
(A) Schematic of sample preparation for TMT-MS. Shown are relative normalized abundances of the sum of the low- and high-salt peptide intensities of the spliceosomal protein orthologs.

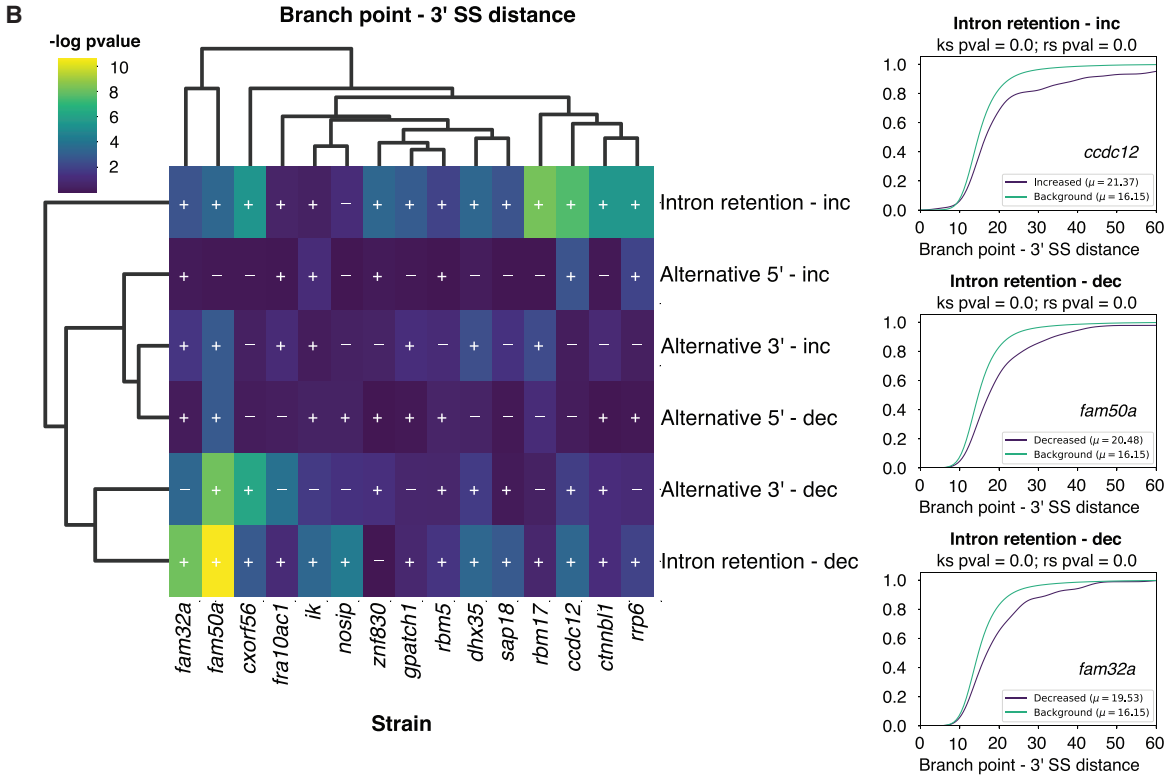
(B) IP-MS results for 2×FLAG GPATCH1. Plotted are TMT-MS data with length-normalized peptide intensity (log_{10}) on the y axis and rank on the x axis. See also Data S3.

(C) IP-MS results for 2×FLAG RBM5. Plotted are TMT-MS data with length-normalized peptide intensity (log_{10}) on the y axis and rank on the x axis. See also Data S4.

(D) IP-MS results for 2×FLAG RBM17. Plotted are TMT-MS data with length-normalized peptide intensity (log_{10}) on the y axis and rank on the x axis. See also Data S5.

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 as a "metazoan-specific" alternative step 2 factor in human spliceosome cryo-EM structures that promotes the splicing *in vitro* of an adenovirus substrate, harboring a relatively short branchpoint-to-3′ splice site distance.[24] 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*.





(legend on next page)

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

Because 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 positions 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-square test. Plotted in Figure 7A are the results for the first 6 nt of the intron for the cases of alternative 5′ slice site usage. We observed 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. We next generated sequence logo plots of the canonical and alternative sites. Shown in Figures 7C and 7D 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. The alternative splice site is consistently considerably weaker than the canonical and often 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 observed similar patterns in mutants of other factors. We conclude the spliceosomal proteins investigated here display a functional bias toward limiting the use of weak/alternative sites.

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 had not been investigated functionally *in vivo*. These factors are not essential for splicing per se, because they were lost in large numbers during the evolution of intron-reduced species, yet they have been conserved at least since the evolutionary divergence of fungi and humans several hundred million years ago. In the cases investigated by IP and MS, factors display biochemical interactions in *C. neoformans* that are similar to those of their human orthologs, suggesting conserved functional roles.

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,[25] with kinetic proofreading by ATPases limiting the splicing of mutant pre-mRNAs via discard and disassembly of substrates with kinetic defects during the catalytic stages of splicing.[26] 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 are maintained in other fungal lineages. We focused our attention on *C. neoformans*. 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. We also included the helicase DHX35 in this analysis, because it is found in *C. neoformans* spliceosomes but not in those of *S. cerevisiae*.[15] 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/post-catalytic complexes.[13] 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.[27,28]

GPATCH1 and DHX35 act together on active spliceosomes

Mutations in each of the 14 human spliceosome protein orthologs examined altered both splicing efficiency and choice. Clustering of the data demonstrated that mutants lacking orthologs of GPATCH1 and DHX35 consistently clustered together for multiple types of splicing changes. 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 likely that GPATCH1 functions to activate DHX35 in the spliceosome, although further biochemical work will be necessary. What the substrate of a GPATCH1-DHX35 complex might be is unclear. Based on the nature of the changes in splice site choice (see below), it may serve a role reminiscent of those of Prp16 and Prp22 in proofreading.[26] We note that, in human cells, GPATCH1 and DHX35 are found in catalytically active spliceosomal complexes,[20] and the MS data in *Cryptococcus* presented here and elsewhere[15] indicate 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 measurement of splicing efficiency.[29,30] A subset of mutants analyzed here display a bias in an increase in intron retention (versus decrease), indicating a tendency toward reducing the efficiency of splicing of specific substrates when mutated, reminiscent of classic pre-mRNA splicing mutants. These include mutants lacking orthologs of GPATCH1 and DHX35 as well as mutants in RBM5, SAP18, and RBM17.

Figure 6. Enrichment of divergent geometry parameters of introns whose splicing is altered in human spliceosomal protein ortholog mutants

(A) Enrichment of altered lengths in affected introns. Displayed in the heatmap is the negative log of the p value produced by a corrected Wilcoxon rank-sum test. Indicated by a "+" or "−" sign is the direction of effect. Right: cummulative density function plots and statistical test results for three gene deletion mutants/splicing change combinations.

(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.

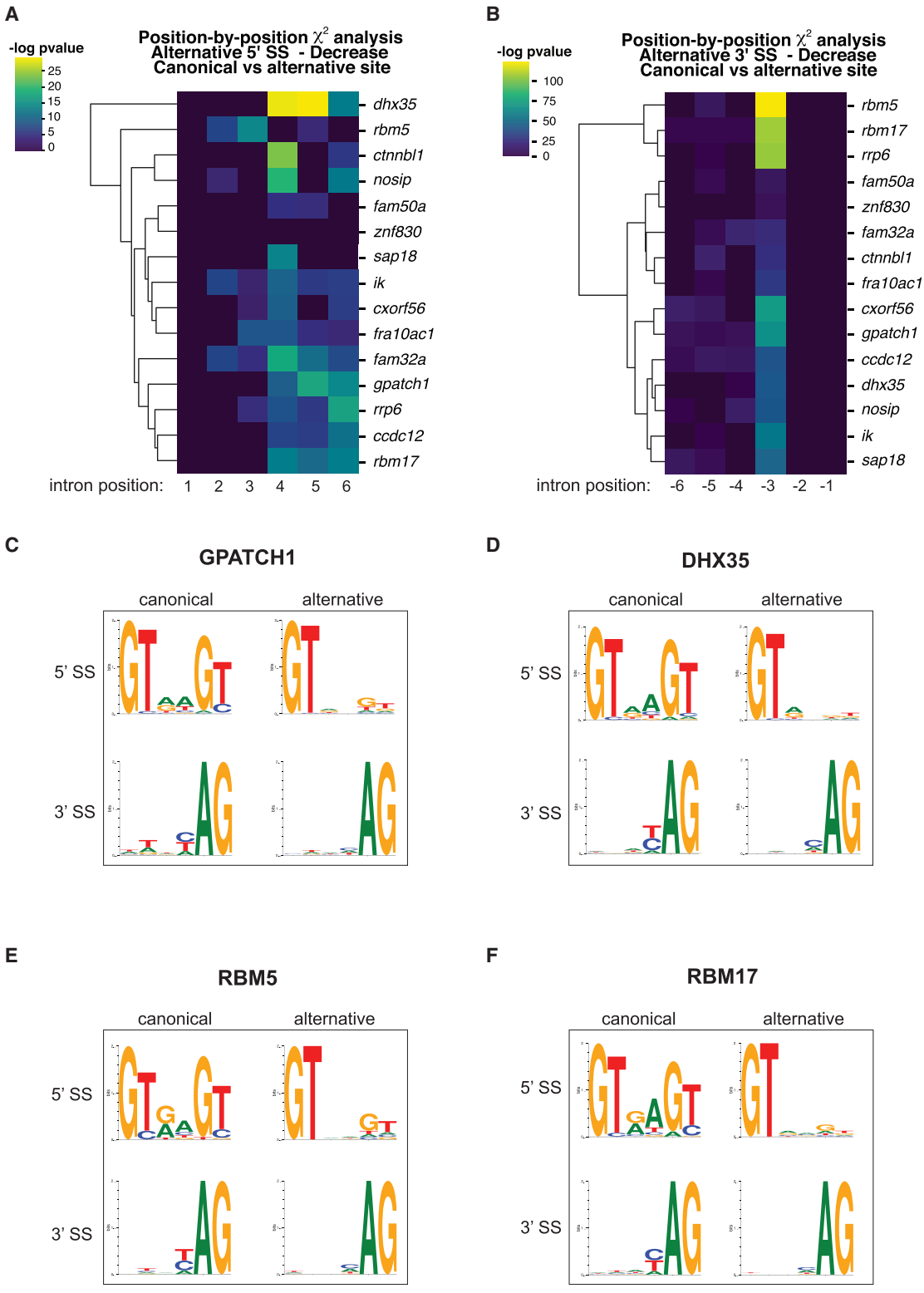


Figure 7. Activation of weak/cryptic alternative 5′ and 3′ splice sites in human spliceosomal protein ortholog mutants

(A) 5′ splice site bases showing significant differences in composition between canonical and alternative sites. Chi-square analysis of the first 6 nt of introns showing significantly decreased Δψ 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.

(C–F) Sequence logos for the canonical and alternative sites for the indicated mutants for introns that display a decrease in the canonical site in the mutant.

Unexpectedly, four mutants tested show the opposite bias (a bias toward improving splicing efficiency when absent): NOSIP, IK, FAM50A, and FRA10AC1. The effects of accessory factors on splicing efficiency correlate with distinctive features of substrates, notably longer intron size and nonoptimal predicted branchpoint-to-3′ splice site distance. Many factors studied here are biased rather than purely unidirectional. For example, although knockout of the ortholog of human GPATCH1 is strongly biased toward 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[31,32]) or context-dependent roles influenced by complex differences in intron structure and sequence.

Accessory factors promote spliceosome fidelity

Nine factors analyzed here display functional signatures that are biased toward the suppression of the use of nearby, weak/cryptic 5′ splice sites whereas four factors are biased toward 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. An additional layer of proofreading might be necessary in organisms whose spliceosomes need to accommodate more variable intron consensus sequences, because 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.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Spliceosomal protein searches
  - *C. neoformans* cultivation
  - Immunoprecipitation and TMT-MS
  - RNA preparation
  - RNA-seq
  - RNA-seq data analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Binomial tests
  - Comparisons of distributions
  - Chi-square analysis (canonical versus alternative sites)
  - p value correlations
  - Seqlogos

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.09.004.

ACKNOWLEDGMENTS

This work was supported by NIH grants 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.

AUTHOR CONTRIBUTIONS

J.S.-L. and H.D.M., conceptualization; B.A.B. and S.W.R., phylogenetic analysis; J.S.-L., RNA-seq, affinity purification, DESeq2 analysis, and JUM analysis; J.K.D. and J.R.Y., MS; B.R., tagged strain construction and RT-PCR validation; I.B., spliceosome annotation and RT-PCR validation; H.D.M., manuscript preparation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 26, 2021
Revised: August 17, 2021
Accepted: September 1, 2021
Published: September 22, 2021

REFERENCES

1. Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA splicing by the spliceosome. Annu. Rev. Biochem. *89*, 359–388.
2. Wahl, M.C., and Lührmann, R. (2015). SnapShot: spliceosome dynamics II. Cell *162*, 456.e1.
3. Wahl, M.C., and Lührmann, R. (2015). SnapShot: spliceosome dynamics I. Cell *161*, 1474.e1.
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. 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.
7. Irimia, M., Penny, D., and Roy, S.W. (2007). Coevolution of genomic intron number and splice sites. Trends Genet. *23*, 321–325.
8. Grate, L., and Ares, M., Jr. (2002). Searching yeast intron data at Ares lab web site. Methods Enzymol. *350*, 380–392.
9. Janbon, G., Ormerod, K.L., Paulet, D., Byrnes, E.J., III, 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.
10. 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.
11. 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.
12. James, T.Y., Stajich, J.E., Hittinger, C.T., and Rokas, A. (2020). Toward a fully resolved fungal tree of life. Annu. Rev. Microbiol. *74*, 291–313.
13. Cvitkovic, I., and Jurica, M.S. (2013). Spliceosome database: a tool for tracking components of the spliceosome. Nucleic Acids Res. *41*, D132–D141.

14. 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.
15. Burke, J.E., Longhurst, A.D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J.J., Yates, J.R., III, Li, J.J., and Madhani, H.D. (2018). Spliceosome profiling visualizes operations of a dynamic RNP at nucleotide resolution. Cell *173*, 1014–1030.e17.
16. 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.
17. 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.
18. 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.
19. 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. USA *115*, E8181–E8190.
20. 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.
21. Hartmuth, K., Urlaub, H., Vornlocher, H.P., Will, C.L., Gentzel, M., Wilm, M., and Lührmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. Proc. Natl. Acad. Sci. USA *99*, 16719–16724.
22. Townsend, C., Leelaram, M.N., Agafonov, D.E., Dybkov, O., Will, C.L., Bertram, K., Urlaub, H., Kastner, B., Stark, H., and Lührmann, R. (2020). Mechanism of protein-guided folding of the active site U2/U6 RNA during spliceosome activation. Science *370*, eabc3753.
23. 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.e7.
24. 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.
25. 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.
26. Koodathingal, P., and Staley, J.P. (2013). Splicing fidelity: DEAD/H-box ATPases as molecular clocks. RNA Biol. *10*, 1073–1079.
27. Robert-Paganin, J., Réty, S., and Leulliot, N. (2015). Regulation of DEAH/RHA helicases by G-patch proteins. BioMed Res. Int. *2015*, 931857.
28. Studer, M.K., Ivanović, L., Weber, M.E., Marti, S., and Jonas, S. (2020). Structural basis for DEAH-helicase activation by G-patch proteins. Proc. Natl. Acad. Sci. USA *117*, 7159–7170.
29. 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.
30. Pikielny, C.W., and Rosbash, M. (1985). mRNA splicing efficiency in yeast and the contribution of nonconserved sequences. Cell *41*, 119–126.
31. 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.
32. 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.
33. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
34. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.; SciPy 1.0 Contributors (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.
35. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.
36. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.
37. Clyde, M.A., and Parmigiani, G. (1998). Protein construct storage: Bayesian variable selection and prediction with mixtures. J. Biopharm. Stat. *8*, 431–443.
38. McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51–56.
39. He, L., Diedrich, J., Chu, Y.Y., and Yates, J.R., III. (2015). Extracting accurate precursor information for tandem mass spectra by RawConverter. Anal. Chem. *87*, 11361–11367.
40. 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.
41. Tabb, D.L., McDonald, W.H., and Yates, J.R., III. (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*, 21–26.
42. 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.
43. 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.
44. Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.
45. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.
46. Bembon, O. (2017) seqLogo: Sequence logos for DNA sequence. R package version 1.44.0.
47. Seabold, S., Perktold, J., et al. (2010). statsmodels: econometric and statistical modeling with python.. Proceedings of the 9th Python in Science Conference. Edited by Stefan van der Walt and Jarrod Millman. (SciPy2010), pp. 92–96.
48. Weber, M.O.. https://github.com/webermarcolivier/statannot.
49. Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423–3424.
50. Altschul, S.F., Madden, T.L., Schäffer, 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.
51. 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.
52. Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431.
53. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631–637.
54. 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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-FLAG M2 affinity gel	Sigma-Aldrich	A2220
Chemicals, peptides, and recombinant proteins		
3X FLAG peptide	Sigma-Aldrich	F4799
T4 RNA Ligase 2, truncated K227Q	New England Biolabs	M0351S
TURBO DNA-free Kit	Thermo Fisher Scientific	AM1907
T4 PNK	New England Biolabs	M0201S
Superscript III First-strand Synthesis System	Thermo Fisher Scientific	18080051
RNaseOUT Recombinant Ribonuclease Inhibitor	Thermo Fisher Scientific	10777109
Proteinase K	Sigma-Aldrich	P6556
Pierce Protease Inhibitor Tablets, EDTA-free	Thermo Fisher Scientific	88266
Critical commercial assays		
PolyATtract® mRNA Isolation System	Promega	Z5300
NEBNext Ultra Directional RNA Library Prep Kit for Illumina	New England Biolabs	E7420S
RNA 6000 Pico Kit	Agilent	5067-1513
High Sensitivity DNA Kit	Agilent	5067-4626
Deposited data		
RNA-seq data	GSE168814	
Experimental models: Organisms/strains		
C. neoformans: GPATCH1-CBP-2XFLAG	This Study	CM2047
C. neoformans: RBM17-CBP-2XFLAG	This Study	CM2046
C. neoformans: RBM5-CBP-2XFLAG	This Study	CM2048
C. neoformans: untagged	Madhani Laboratory	CM025
C. neoformans: cnag_05579Δ:: NEO	Madhani Laboratory	CK5778
C. neoformans: cnag_00761Δ:: NEO	Madhani Laboratory	CK3285
C. neoformans: cnag_00294Δ:: NEO	Madhani Laboratory	CK1844
C. neoformans: cnag_03665Δ:: NEO	Madhani Laboratory	CK2692
C. neoformans: cnag_02401Δ:: NEO	Madhani Laboratory	CK1446
C. neoformans: cnag_04679Δ:: NEO	Madhani Laboratory	CK5311
C. neoformans: cnag_02260Δ:: NEO	Madhani Laboratory	CK4072
C. neoformans: cnag_05845Δ:: NEO	Madhani Laboratory	CK5920
C. neoformans: cnag_06616Δ:: NEO	Madhani Laboratory	CK954
C. neoformans: cnag_05030Δ:: NEO	Madhani Laboratory	CK5488
C. neoformans: cnag_01058Δ:: NEO	Madhani Laboratory	CK3448
C. neoformans: cnag_02340Δ:: NEO	Madhani Laboratory	CK4114
C. neoformans: cnag_05307Δ:: NEO	Madhani Laboratory	CK5607
C. neoformans: cnag_02773Δ:: NEO	Madhani Laboratory	CM
C. neoformans: cnag_03031Δ:: NEO	Madhani Laboratory	CK4458
Software and algorithms		
Samtools	[33]	http://samtools.sourceforge.net/
SciPy	[34]	https://github.com/scipy/scipy

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Cutadapt	N/A	https://github.com/marcelm/cutadapt
FastX toolkit	N/A	http://hannonlab.cshl.edu/fastx_toolkit
HTSeq	35	https://pypi.python.org/pypi/HTSeq
BEDTools	N/A	https://github.com/arq5x/bedtools2
Integrative Genomics Viewer	36	http://software.broadinstitute.org/software/igv/
BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling (R)	37	https://cran.r-project.org/web/packages/BAS/index.html
Pandas (version 1.1.1)	38	https://zenodo.org/record/4067057#.X7RG3NNKi34
SP Pipeline	15	https://github.com/jeburke/SPTools/
IP2	N/A	http://www.integratedproteomics.com
RawConverter	39	http://fields.scripps.edu/rawconv/
ProLucid	40	http://fields.scripps.edu/downloads.php
DTASelect	41	http://fields.scripps.edu/yates/wp/
Census 2	42	http://fields.scripps.edu/yates/wp/?page_id=824
STAR	43	https://github.com/alexdobin/STAR
Matplotlib (version 3.3.3)	44	https://matplotlib.org/
Numpy	45	https://numpy.org
Seqlogo	46	https://github.com/betteridiot/seqlogo
Statsmodels (version 0.13.0)	47	https://www.statsmodels.org/dev/index.html
Statannot	48	https://github.com/webermarcolivier/statannot
Pybedtools	49	https://daler.github.io/pybedtools/
JUM (version 2.0.2)	19	https://github.com/qqwang-berkeley/JUM

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Hiten Madhani (hitenmadhani@gmail.com)

Materials availability

C. neoformans strains are available without restriction from the lead contact.

Data and code availability

RNA-seq data is publically available at the NCBI GEO database: GSE168814.

This paper does not report original code.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All experiments were performed in *Cryptococcus neoformans* in the KN99 strain background. Strains were cultivated in YPAD media (Difco) at 30°C.

METHOD DETAILS

Spliceosomal protein searches

Spliceosomal protein searches were performed on proteome assemblies available from NCBI and UniProt (See Table S2). 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}).[50] 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.[51–53] To avoid bias in protein domain content, domains used for HMM searches were defined as described.[54] 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. See Data S1.

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[15] 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.[15] See Data S3, S4, and S5.

RNA preparation

Polyadenylated RNA was prepared exactly as described.[15]

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 to the *C. neoformans* H99 genome sequence (NCBI ID: GCA_00149245.3) using STAR.[43] A minimum of 12M read/strain/replicate were obtained (Table S3). 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 $\Delta\psi$. See Data S2,

QUANTIFICATION AND STATISTICAL ANALYSIS

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-square analysis (canonical versus 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 versus unaffected introns as the columns. In some cases where less than 5 counts are observed in a category, a chi-square 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).