Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*

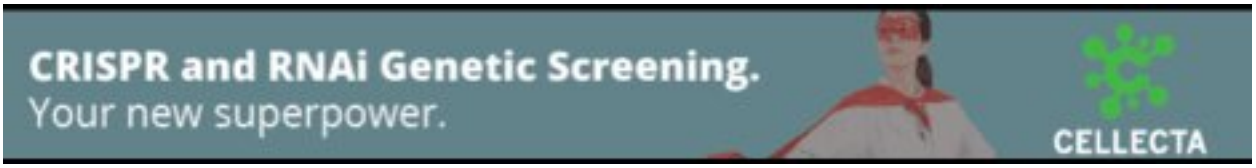
Quinn M. Mitrovich, Brian B. Tuch, Christine Guthrie, et al.

Genome Res. published online March 9, 2007
Access the most recent version at doi:10.1101/gr.6111907

P<P Published online March 9, 2007 in advance of the print journal.

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Copyright © 2007, Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

Letter

Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*

Quinn M. Mitrovich,[1] Brian B. Tuch, Christine Guthrie, and Alexander D. Johnson

Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94143-2200, USA

Candida albicans is the most common fungal pathogen of humans. Frequently found as a commensal within the digestive tracts of healthy individuals, *C. albicans* is an opportunistic pathogen that causes a wide variety of clinical syndromes in immuno-compromised individuals. A comprehensive annotation of the *C. albicans* genome sequence was recently published. Because many *C. albicans* coding sequences are interrupted by introns, proper intron annotation is essential for the accurate definition of genes in this pathogen. Intron annotation is also important for identifying potential targets of splicing regulation, a common mechanism of gene control in eukaryotes. In this study, we report an improved annotation of *C. albicans* introns. In addition to correcting the existing intron annotations, 25% of which were incorrect, we have used novel computational and experimental approaches to identify new introns, bringing the total to 415, almost double the number previously known. Our identification methods focus primarily on intron features rather than protein-coding features, overcoming biases of traditional intron annotation methods. Introns are not randomly distributed in *C. albicans*, and are over-represented in genes involved in specific cellular processes, such as splicing, translation, and mitochondrial respiration. This nonrandom distribution suggests functional roles for these introns, and we demonstrate that splicing of two transcripts whose introns have unusual sequence features is responsive to environmental factors.

[Supplemental material is available online at www.genome.org. The microarray data from this study are available at ArrayExpress (http://www.ebi.ac.uk), accession no. E-MEXP-1003.]

The yeast *Candida albicans* commonly inhabits the mucous membranes and digestive tracts of healthy individuals. Perturbation of a host's immune defenses, however, can cause a dramatic shift to invasive, pathogenic growth. Particularly susceptible are those receiving immunosuppressive therapies, such as cancer and transplant patients (Kullberg and Filler 2002), and individuals infected with HIV, more than 90% of whom will at some point suffer candidiasis (Fidel 2002). *C. albicans* exhibits remarkable adaptability, capable of successfully invading virtually every human organ and tissue (for review, see Odds 1988). During infection, *C. albicans* is most commonly associated with mucosal surfaces, but in its most devastating form—disseminated candidiasis—it spreads through the host bloodstream and invades multiple tissues, with an associated mortality in excess of 40% (for example, see Alonso-Valle et al. 2003).

Eukaryotic genes are often interrupted by introns, which must be spliced out of gene transcripts for coding sequences to be fully expressed. The regulation of intron splicing can also play an important role in controlling gene expression (for reviews, see Black 2003; Faustino and Cooper 2003; Blencowe 2006). Because introns are often highly prevalent in genomes—mammals, for example, have an average of more than seven introns per gene (Roy and Gilbert 2006)—identifying individual instances of splicing regulation is not always straightforward. The hemiascomycetous yeasts, which include *C. albicans* and the well-studied model organism *Saccharomyces cerevisiae*, in contrast, have experienced exceptionally high rates of intron loss (Dujon 2006). As a result, only 5% of genes in *S. cerevisiae* contain introns, and most of these genes contain only one (Davis et al. 2000). The loss of introns within the *S. cerevisiae* lineage has not been random, and introns are now highly over-represented within certain gene categories (for example, see Ares et al. 1999). Thus, hemiascomycetous genomes may be highly enriched for introns that are retained because they confer a selective advantage, such as those introns involved in gene regulation. Consistent with this idea, there are several examples of regulated splicing in *S. cerevisiae* (for examples, see Engebrecht et al. 1991; Li et al. 1995; Nakagawa and Ogawa 1999; Davis et al. 2000; Vilardell et al. 2000; Preker et al. 2002), and recent studies have shown that a large but specific subset of *S. cerevisiae* introns is regulated in response to amino acid starvation (J. Pleiss, G. Whitworth, M. Berkessel, and C. Guthrie, in prep.). As in *S. cerevisiae*, introns are relatively uncommon in *C. albicans*, and we therefore believe an accurate annotation of these introns will be a useful tool for studying the role of splicing regulation in pathogenesis and other facets of *C. albicans* biology.

The genome of *C. albicans* strain SC5314 was recently sequenced using a whole-genome shotgun approach (Jones et al. 2004). A community-based effort involving multiple laboratories subsequently produced a hand-curated annotation of *C. albicans* coding sequences (Braun et al. 2005). To generate an accurate set of protein predictions for a eukaryotic genome, the introns must be defined with absolute precision. While efforts were made by the community to annotate the introns of *C. albicans* (including contributions from one of our laboratories), we have re-examined these annotations and have found they were incomplete and often inaccurate.

Accurate intron identification has long been a challenge for genome annotators (for review, see Brent 2005). The most accu-

[1]**Corresponding author.**
E-mail quinn.mitrovich@ucsf.edu; fax (415) 502-4315.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6111907.

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

rate systematic method relies on spliced mRNA sequence data derived from large-scale expressed sequence tag (EST) projects. For comprehensive coverage, however, particularly for genes that are poorly expressed, this approach requires a scale that has been impractical for most organisms. In the absence of such data, large-scale intron identification has relied almost exclusively on computational methods. These methods rely primarily on alignments of putative coding sequences with homologous genes from other species, with alignment gaps identifying possible introns. Splice-site consensus sequences are used secondarily to support the existence of an intron and to define its precise boundaries. Other methods use de novo gene prediction algorithms, again focusing primarily on coding sequences. Despite substantial progress, however, such algorithms are still fairly inaccurate (for example, see Wei et al. 2005).

The primary reliance on coding-sequence alignments introduces serious biases to the identification of introns. In particular, introns in poorly conserved genes or near the ends of even well-conserved genes are easily overlooked using traditional alignment-based methods. This bias may be particularly problematic for the Hemiascomycetes; in *S. cerevisiae*, introns very close to the 5′-ends of coding sequences are quite common (Spingola et al. 1999).

Given the inherent difficulties with traditional methods of intron identification, it is not surprising that the existing *C. albicans* intron annotations are far from complete. In the course of correcting the existing intron annotations, it became apparent to us that many had been overlooked. This prompted us to undertake a genome-wide search for undiscovered introns, including in our analyses two methods that do not rely on strong coding-sequence conservation. Here, we present a high-confidence set of 415 introns, almost doubling the number of known introns in *C. albicans*. We believe that this represents a nearly complete catalog of *C. albicans* introns. We show that these introns are not randomly distributed, but are dramatically over-represented in genes within specific functional categories. This nonrandom distribution suggests that some introns are under selective pressure, perhaps for roles in regulating gene expression. To test this idea, we have examined two genes, *RPL30* and *SPR28*, whose introns have unusual sequence features, and have shown that their splicing is, indeed, responsive to environmental signals.

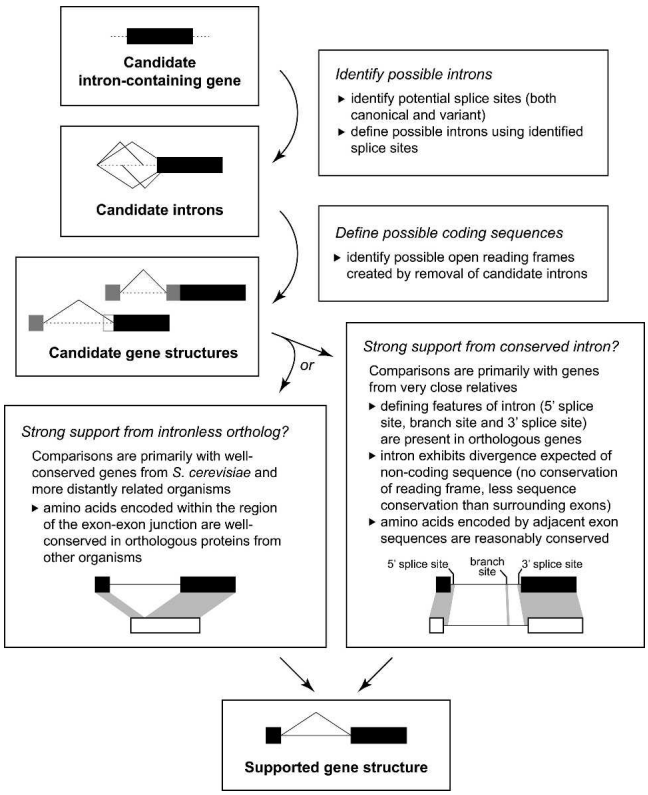


Figure 1. Refinement and confirmation of candidate introns. Once candidate intron-containing genes are identified by any of our methods, we use a manual bioinformatics approach to identify possible introns and determine whether any are strongly supported by available phylogenetic data. Because our approach does not require annotated sequences for comparison, we can rely on support from the sequenced genomes of other *Candida* species when protein conservation is insufficient for comparisons with more distant relatives. One of our two sets of criteria—strong support from either orthologous coding sequences or conserved introns—must be met for us to consider an intron supported. (See Methods for further details.)

Results

Refinement of intron predictions in published annotations

Based on the recently published annotation of the *C. albicans* genome (Braun et al. 2005), most genes appear to lack introns, as expected for a hemiascomycetous yeast (Bon et al. 2003). Once redundant features are removed, published annotations identify 232 introns in 215 genes (Hull and Johnson 1999; Braun et al. 2005; Chibana et al. 2005), roughly comparable to the number of introns in *S. cerevisiae*. Our initial inspection of the published intron annotations revealed several unanticipated features. In particular, many of the splice site sequences deviated substantially from the *S. cerevisiae* consensus. These sequences, which include the 5′-splice site, the 3′-splice site, and the internal branch site, are the primary elements by which the spliceosomal machinery precisely identifies introns for removal. We investigated introns with unusual splice sites to determine whether they might have been misannotated, using a general approach we used throughout our studies (Fig. 1). In this approach, we use splicing consensus sequences to identify candidate introns, build gene models around each of these introns, and use phylogenetic data to confirm or reject the individual models. For some of the questionable introns in the previously published annotations, the boundaries were miscalled, and required minor adjustment. In other cases, the adjustments were more substantial, such as reassignment of an entire exon or division of a single intron into two.

Many of the originally annotated introns did not pass our filters and were removed from our assignments. Most often these questionable introns were introduced by previous annotators to correct for stop codons or frameshifts within the coding sequences. DNA alignments with the close relative *Candida dubliniensis* suggest that many of these stop codons and frameshifts are simply the result of sequencing errors, without which the open reading frames would be continuous in the absence of splicing. Throughout our studies, we found that such errors are quite common in the *C. albicans* genome sequence, in contrast to the sequenced genomes of other *Candida* species, such as *C. dubliniensis* (http://www.sanger.ac.uk) and *Candida tropicalis* (http://www.broad.mit.edu).

Overall, we found that 59 of the published intron annotations (~25%) required adjustment or removal. Once these corrections were introduced, we were left with a high-confidence set of

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

224 introns in 201 genes (Supplemental Table S1). Our confidence in these assignments is based on conformity of splice-site sequences and strong phylogenetic support from introns and coding sequences conserved in other species. To test our predictions further, we selected four candidate introns and confirmed them experimentally by sequencing their corresponding cDNAs (Supplemental Table S1). As described below, several of our annotations have also been confirmed by other experimental techniques.

Candidate gene approach for intron identification

The annotation of introns in *S. cerevisiae* is fairly comprehensive, having received attention from numerous research groups (for examples, see Spingola et al. 1999; Davis et al. 2000; Cliften et al. 2003; Kellis et al. 2003). The orthologs of many intron-containing genes from *S. cerevisiae* also contain annotated introns in *C. albicans*. The *C. albicans* orthologs of the remaining *S. cerevisiae* intron-containing genes are therefore good candidates in which to search for undiscovered introns.

We were able to identify putative *C. albicans* orthologs for 70% of *S. cerevisiae* intron-containing genes. Because many genes have duplicated since the divergence of these two yeasts, particularly in the *S. cerevisiae* lineage (Byrne and Wolfe 2005), there is not always a one-to-one relationship between orthologs. We therefore used both reciprocal best BLAST hits and more complex ortholog maps (Tsong et al. 2006) to generate a list of *C. albicans* orthologs. Among the genes thus identified, 35% also contained annotated introns. We inspected the remaining 65% of genes manually, and were able to identify introns in one-quarter of them. We also identified introns in two additional genes through direct inspection of interesting candidates (Table 1; Supplemental Table S1).

Nine of the introns we identified using our candidate approach reside outside of the coding sequences, within the 5′-untranslated regions (UTRs) of the RNAs. 5′-UTR introns can be difficult to validate without direct experimental data, as they are not surrounded by the protein-coding sequences that normally lend support to an intron assignment. In *S. cerevisiae*, however, the few 5′-UTR introns that have been identified have strong phylogenetic and experimental support (Spingola et al. 1999; Davis et al. 2000; Cliften et al. 2003; Kellis et al. 2003). The 5′-UTR introns we found using our candidate approach are all conserved within the 5′-UTRs of their *S. cerevisiae* orthologs and are therefore well supported.

Experimental approach for intron identification

Thus far, our approaches to intron identification contain significant biases. The published genome annotations that provided our starting point relied heavily on well-conserved coding sequences and canonical splice sites for identification of intervening introns (Braun et al. 2005). Our candidate approach was restricted to identifiable orthologs of *S. cerevisiae* intron-containing genes. To complement our previous analysis, we therefore developed an experimental approach to identify introns without these inherent biases. This approach relies on the construction and analysis of a *C. albicans* mutant that accumulates intron sequences.

In the process of pre-mRNA splicing, introns are liberated as lariat structures, in which the 5′-end is joined by a 5′–2′-phosphodiester linkage to an internal branch point nucleotide (Fig. 2A). Lariat structures are subsequently linearized by a debranching enzyme, Dbr1, and rapidly degraded from their free 5′- and 3′-ends by exonucleases. Deletion of *DBR1* in *S. cerevisiae* results in a dramatic accumulation of branched intron sequences, resistant to exonucleolytic degradation because they lack free ends (Chapman and Boeke 1991). Branched introns have abnormal electrophoretic mobility and can be distinguished from other RNAs by two-dimensional gel electrophoresis (Ruskin et al. 1984). We constructed a *C. albicans* debranchase mutant by deleting both *DBR1* alleles and found that the distribution of accumulated intron sequences is similar to the distribution in a *S. cerevisiae dbr1*Δ mutant (data not shown).

To identify introns that accumulate in our *dbr1*Δ mutant, we designed a microarray with probes flanking all predicted *C. albicans* open reading frames (for details, see Methods). While such an array is not comprehensive, we reasoned that the most likely locations for undiscovered introns would be adjacent to predicted coding sequences. If a probe does hybridize to an expressed intron, it should show increased signal from *dbr1*Δ RNA compared to wild-type RNA, provided the probe is of sufficient sequence complexity to give a signal with specificity for the intron (Fig. 2B).

We synthesized cDNA from RNA from both wild-type and *dbr1*Δ mutant strains grown under standard conditions, labeled the two populations differentially, and hybridized them together on our microarray. We used the normalized ratio of *dbr1*Δ and wild-type intensity values for each microarray probe as a measure of differential expression. The vast majority of probes (99.6%) exhibited a less than twofold bias in intensity between the two strains. Among the remaining probes, nine exhibited up to a sixfold decrease in *dbr1*Δ cells, including two probes specific to *DBR1*, while 124 exhibited up to an 80-fold increase in *dbr1*Δ cells. Although the nature of our probe design usually precluded detection of previously annotated introns, this was not always the case. Indeed, 44% of the probes that exhibited a significant increase correspond to previously predicted introns, providing independent confirmation for these predictions.

We manually inspected the remaining sequences that exhibited a greater than twofold increase, using our phylogenetic criteria to look for independent evidence of introns in the nearby genes (Fig. 2B). We were able to identify 26 new introns with this approach (Supplemental Table S1). Many of the newly discovered introns were confirmed by hybridization to multiple probes. Only 15% of the probes with a greater than twofold increase identified regions in which we could find no independent evidence of an intron. The relative expression values for these false positives all fell within the bottom quartile of up-regulated probes.

In addition to identifying candidate introns, our microarray data also provide corroborating evidence that a sequence is

Table 1. Summary of *C. albicans* intron annotations

	Number of introns	Number of genes with introns	Proportion of genome
Published annotations	232	215	3.4%
After corrections	224	201	3.2%
New annotations, by approach			
S. cerevisiae candidates	+34	+32	
Direct inspection	+2	+2	
Experimental	+28	+27	
Computational	+116	+108	
D. hansenii candidates	+11	+11	
Current annotation	415	381	6.0%

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

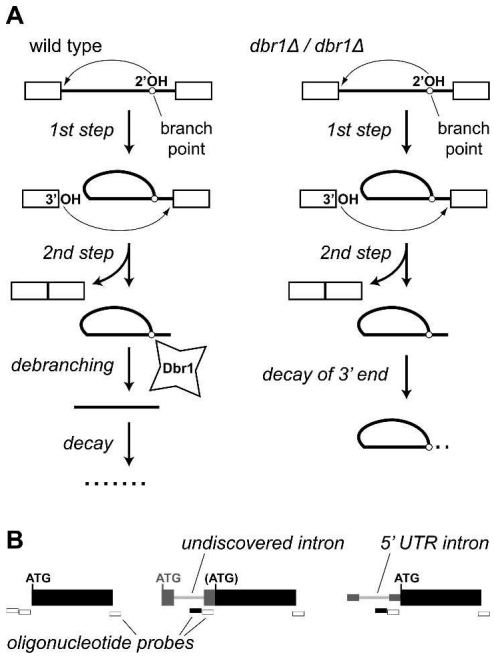


Figure 2. Sequences that accumulate in a debranchase mutant are used to identify new introns. (*A*) Intron removal occurs via two *trans*-esterification reactions (thin arrows), carried out by the 2′- and 3′-hydroxyls indicated. The resulting intron lariats are linearized by a debranching enzyme (Dbr1) and degraded by exonucleases (*left*). In a debranchase mutant (*right*), intron sequences upstream of the branch point are inaccessible to exonucleases, and accumulate to high levels. (*B*) When *dbr1* mutant cDNA is compared to wild-type cDNA on our microarrays, probes specific to the region between the 5′-splice site and the branch point of an expressed intron show an intensity bias (black probes), while most other probes do not (white probes). (Boxes) Exons; (lines) introns; (ATG) translation start codons; (gray items) previously undiscovered introns and exons.

spliced. Eleven of the 26 new introns we identified in our *dbr1*Δ strain were within 5′-UTRs. As discussed, 5′-UTR introns are only weakly supported by bioinformatics analyses alone. Direct experimental evidence from our microarrays is therefore important in establishing the validity of these intron assignments.

Because our original microarray compared only yeast growing in standard media at 30°C, it would fail to identify introns that are not spliced or whose genes are not expressed in this condition. We therefore grew wild-type and *dbr1*Δ yeast under a variety of conditions, pooled RNAs from the various conditions to synthesize cDNA, and compared the wild-type and *dbr1*Δ cDNAs on our microarray. The conditions we used included different carbon sources, different growth phases, different growth forms (filamentous, mating, and opaque phase cells), starvation, and nitric oxide stress (for details, see Supplemental Methods). This approach identified two new introns (Supplemental Table S1). In both cases, the increased *dbr1*Δ intron signal was clearly absent under standard conditions.

Global computational approach for intron identification

With our candidate and experimental approaches, we generated a high-confidence set of 288 introns. We found that the splice-site sequences tend to show relatively little deviation from the consensus. The introns tend to be short, with 90% of introns <500 nt in length, and they tend to be near the 5′-ends of transcripts. Because of the highly stereotyped nature of introns in *C. albicans*, we reasoned that their characteristics could be used to build a computational model for predicting new introns in other genes.

Our first approach was to model the *C. albicans* intron structure with a Hidden Markov Model (HMM), as has been done previously to highlight common features of introns in *S. cerevisiae* (Spingola et al. 1999). We trained several alternative HMM architectures on our high-confidence set of intron-containing genes using the Baum-Welch training algorithm (Durbin et al. 1998). None of the resulting HMMs, however, could reasonably distinguish our positive training set from a set of randomly chosen genes (data not shown). We therefore turned to a more direct approach that uses position-specific scoring matrices (PSSMs) to score splice-site sequences, and distance distributions to score the spacing between splice sites. This more explicitly defined model produces a combined score indicating the likelihood that a given intron is a true intron. The score is based on the characteristics of its putative 5′-splice site, 3′-splice site, branch site, and the relative distances between them. We trained the model on these characteristics from our high-confidence set of introns. Using the trained model, we developed an algorithm that scores all possible introns within a given sequence and returns the maximal-scoring candidate intron.

We generated a set of scored intron predictions, applying our algorithm to all predicted *C. albicans* open reading frames (ORFs). For every ORF, we included both the predicted coding sequence and 500 nt of upstream sequence. Because of the short length and 5′ bias of most introns, we reasoned this would likely capture the majority of undiscovered intron sequences. We ranked all genes according to their intron scores, and inspected the top-scoring genes manually for new introns (Fig. 1). For each gene, we used the intron predicted by our algorithm as a starting point, but did not limit our inspection to this sequence. Importantly, our algorithm effectively distinguished the majority of known intron-containing genes from most of the rest of the genome (Fig. 3). More than 90% of our previously identified intron-containing genes scored within the top 6% of genes.

We inspected the remaining genes that received high scores from our algorithm, beginning with the top score and working down until false positives predominated (see below). Among the genes that scored within the top ~6%, we were able to identify 94 new introns that passed our criteria for strong phylogenetic support (outlined in Fig. 1). The intron predictions that accompanied these scores accurately identified the precise splice junctions more than 85% of the time. In the remaining cases, the failure was almost always in predicting the correct 3′-splice site. This is not surprising, given that the 3′-splice site has the lowest information content, and accurate annotation often requires information from phylogenetic comparisons.

Because we included upstream sequence to generate our intron predictions, several of the top scores were due to introns in upstream genes. In two cases, this identified new genes that had been missed in the genome annotation. Both genes are <100 codons in length and, had we not identified the introns first, would have been easily overlooked. One, orf19.2965.1, is a homolog of *Homo sapiens SERF2*, and the other, orf19.3223.1, encodes the 12-kDa subunit of mitochondrial NADH-ubiquinone oxidoreductase.

Six of the intron predictions within the top 6% would reside within 5′-UTRs. Because these introns lack adjacent coding sequence, they failed to meet our criteria for strong phylogenetic support. To determine whether any of these were actually in-

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

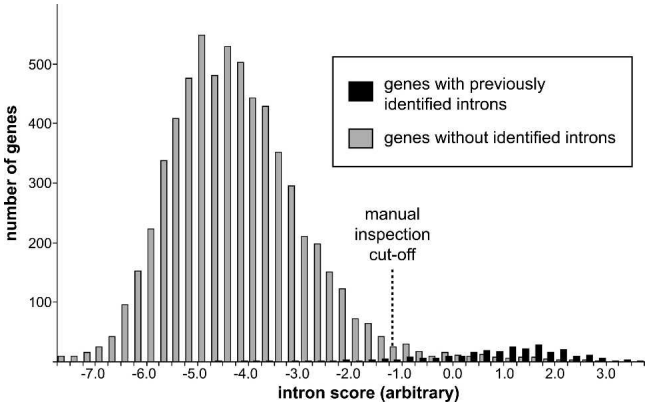


Figure 3. Distribution of computational intron scores. The histogram depicts the distribution of intron scores among all genes in our genomic data set. Scores are in arbitrary units, with higher numbers indicating a higher likelihood that a given gene contains an intron. Whether or not a gene is represented as containing a known intron reflects its status prior to our computational screen. (Dashed line) The cutoff used to define our set of candidate intron-containing genes.

trons, we tested them directly, using RT-PCR to span the predicted splice junctions. We found that all of them were, indeed, spliced (Supplemental Fig. S1A).

Finally, some of the predictions for the genes we examined are clearly not introns. These false predictions are generally inconsistent with phylogenetic data—the putative splice sites are not conserved in closely related species, and splicing of the putative introns would disrupt conserved coding sequences. Our algorithm does not predict a discrete boundary between genes that do and genes that do not contain introns (Fig. 3), and we therefore used these false positives to determine how far down our list we would search. In the end, we examined the top-scoring ~6% of genes from our whole genome list. Within this ranked subset of ~400 genes, false positives were entirely absent from the top half, but represented almost 90% of the predictions as we reached the bottom 4% of these top-scoring genes.

We used the new introns identified by our computational approach to refine our intron model, and applied the refined model to the *C. albicans* ORFs for a second round of predictions. To capture additional introns we may have missed in our first run, we also included 500 nt of downstream sequence for each ORF, and extended the upstream sequence from 500 to 1000 nt. Since the median lengths of upstream and downstream intergenic sequences are 600 nt and 300 nt, respectively (as calculated from annotations at http://www.candidagenome.org), this should be sufficient for most genes. Manual inspection of the top-scoring genes revealed 16 new introns. Two of these introns had unusually degenerate splice sites, and received higher scores in our second analysis because of the refinements of our model. The other 14 were detected because of the additional sequence we included with each ORF: two were within downstream sequence, and 12 were within upstream sequence, including one 5′-UTR intron we confirmed by RT-PCR (Supplemental Fig. S1A). As before, some of the upstream introns were within previously unannotated genes. These included homologs of the *S. cerevisiae* genes *ATP18*, *ATP19*, *DAD4*, *INO4*, and *SLX9*, and a gene with similarity to *H. sapiens BLOC1S2* (encoded by orf19.2018.1).

As an independent test of our predictions, we examined several candidate introns directly using RT-PCR. We concentrated primarily on the introns with particularly unusual splice sites (and therefore low scores from our algorithm), figuring these should represent the weakest of our predictions. Of the eight genes we examined, we were able to confirm the existence of an intron of the predicted size in every case (Supplemental Fig. S1B). Interestingly, most of these introns were inefficiently spliced, perhaps because of, in part, the unusual splice sites.

Secondary candidate approach for intron identification

Debaryomyces hansenii is a nonpathogenic yeast more closely related to *C. albicans* than is *S. cerevisiae*. The sequence divergence between *C. albicans* and *D. hansenii* is roughly half of that between *C. albicans* and *S. cerevisiae*, based on a phylogenetic tree inference (Tsong et al. 2006). The genome of *D. hansenii* has been sequenced and annotated (Sherman et al. 2004), and the annotations include several intron predictions. Because of the more recent shared ancestry of *C. albicans* and *D. hansenii*, we would expect a greater overlap in the genes that contain introns, compared to the overlap we observe between *C. albicans* and *S. cerevisiae*. The *D. hansenii* intron set thus provides a test (albeit not a rigorous one) of how thorough our *C. albicans* intron annotation has been. If a candidate approach, investigating the orthologs of *D. hansenii* intron-containing genes, were to reveal many new intron-containing genes in *C. albicans*, we would have to conclude that our annotation is far from complete.

Of the 338 *D. hansenii* genes whose annotations include introns, we found that 35 were likely misannotated, and probably did not contain introns. For three genes, we were unable to identify any *C. albicans* homolog. Of the remaining 300 genes, 176 had *C. albicans* homologs in which introns had already been identified, 56 of which were first identified as containing introns in our experimental and computational screens described above. For another 114 genes, we found that the *C. albicans* homologs did not contain introns.

Interestingly, seven of the remaining 10 *D. hansenii* genes identified a homolog (and in one case two homologs) that had been overlooked in the *C. albicans* gene annotations. These genes have short ORFs and are interrupted by introns, which is presumably why they had not been discovered previously. Unlike the eight new genes we discovered using our computational approach, these genes are not close enough to a previously annotated ORF to have been identified in our computational search.

The final three *D. hansenii* genes identify new introns in known *C. albicans* genes. These genes received scores below the cutoff in our computational approach (although two of the three were within the top 8%). The low scores were due to highly unusual features of the introns in *C. albicans*: one has a unique branch site sequence, while the other two have both a noncanonical splice site and an unusually long distance between the branch site and the 3′-splice site. Thus, there are almost certainly introns that remain to be discovered in *C. albicans*. We believe, however, that the small number of introns identified with this last approach suggests that our overall analysis has identified most.

Categories of intron-containing genes

As with *S. cerevisiae*, the introns of *C. albicans* are not randomly distributed. The genes that contain introns often fall into distinct functional classes. We performed a gene ontology (GO) term analysis on the intron-containing genes of *C. albicans* and *S. cerevisiae* to identify over-represented gene categories, some of which are shown in Figure 4. Some categories are common to

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

Categories shared with *S. cerevisiae*	*C. albicans* genes: Proportion of genes with introns	*C. albicans* genes: Corrected p-value	*S. cerevisiae* genes: Proportion of genes with introns	*S. cerevisiae* genes: Corrected p-value
Ribosomes				
large ribosomal subunit (GO:0015934)	32/54 (59%)	2.2e-19	57/85 (67%)	2.3e-56
small ribosomal subunit (GO:0015935)	22/38 (58%)	3.8e-13	45/62 (73%)	6.4e-47
Meiosis				
meiosis I (GO:0007127)	12/34 (35%)	3.8e-4	12/48 (25%)	2.5e-5
Categories specific to *C. albicans*				
Microtubules				
microtubule organizing center (GO:0044450)	5/10 (50%)	5.1e-3	0/15 (0%)	0.65
kinetochore microtubule (GO:0005828)	3/4 (75%)	4.8e-3	2/15 (13%)	0.31
Splicing				
tri-snRNP complex (U4/U6,U5) (GO:0046540)	8/20 (40%)	2.0e-3	3/28 (11%)	0.36
U6 snRNP (GO:0005688)	4/5 (80%)	6.1e-4	2/8 (25%)	0.10
Mitochondrial respiration				
ATP biosynthesis (GO:0006754)	7/12 (58%)	1.8e-4	0/17 (0%)	0.68
cytochrome c oxidase assembly (GO:0008535)	3/5 (60%)	1.4e-2	0/8 (0%)	0.54
Protein degradation				
regulation of protein catabolism (GO:0042176)	4/6 (67%)	2.5e-3	1/4 (25%)	0.22
proteasome complex (Eukaryota) (GO:0000502)	5/15 (33%)	3.7e-2	2/15 (13%)	0.31

p < 0.1 | p < 0.01 | p < 1e-3 | p < 1e-4

Figure 4. Intron-containing genes are biased for specific categories in *C. albicans* and *S. cerevisiae*. We performed an over-representation analysis on GO terms associated with genes from *C. albicans* or *S. cerevisiae*. A noncomprehensive list of GO terms significantly enriched in the *C. albicans* intron-containing set is shown. *P*-values, which have been subjected to multiple test correction, are shaded according to the cutoffs illustrated *below* the values.

both yeasts, such as ribosomal protein genes and genes involved in meiosis in *S. cerevisiae*, but of currently unknown function in *C. albicans*. Others are over-represented in *C. albicans* but not significantly over-represented in *S. cerevisiae*, such as genes involved with microtubules, splicing, mitochondrial respiration, and protein degradation.

The over-representation of introns within certain gene categories demonstrates that the widespread loss of introns within the hemiascomycetous lineage has not been random. This, in turn, suggests that the retention of certain introns is under selective pressure. One possible explanation is that these introns are involved in regulating the expression of the genes in which they reside, as has already been demonstrated for several *S. cerevisiae* introns (discussed above).

What predictions can we make about regulated splicing in *C. albicans* on the basis of sequence features alone? While not always the case, regulated splicing in *S. cerevisiae* often involves introns with noncanonical splice sites (Li et al. 1995; Spingola and Ares 2000; Vilardell et al. 2000; Preker et al. 2002). We therefore examined two genes with unusual splice sites—one ribosomal protein gene and one septin gene—to determine whether we could find evidence for regulated splicing in *C. albicans*.

Alternative splicing of *RPL30* pre-mRNA

S. cerevisiae ribosomal protein L30 (Rpl30) regulates its own expression by inhibiting the splicing of its own pre-mRNA (Vilardell et al. 2000). Excess Rpl30 can bind to a structure within its pre-mRNA that includes the 5′-splice site. This 5′-splice site is already suboptimal (GU**CA**GU), and binding by Rpl30 inhibits splicing, causing the accumulation of unspliced and therefore unproductive RNA.

The published annotation of the *RPL30* (orf19.3788.1) intron in *C. albicans* was incorrect, but the corrected sequence revealed that the Rpl30 RNA-binding site is conserved with *S. cerevisiae*. It appears that the splicing regulation may also have been conserved in *C. albicans*, as we were able to detect both *RPL30* mRNA and pre-mRNA by RT-PCR (Fig. 5A). While inspecting the intron sequence, however, we found a canonical 5′-splice site sequence in the center of the *RPL30* intron. Consistent with this observation, we detected an additional RT-PCR product of intermediate size, and sequencing identified it as an alternatively spliced mRNA arising through use of this internal 5′-splice site. As far as we are aware, this is the only known case of alternative splicing in *C. albicans*. Both the alternative mRNA and the unspliced pre-mRNA are unproductive transcripts, interrupted by non-coding sequence and therefore unable to produce Rpl30 protein. Interestingly, which of the two unproductive forms predominates appears to be influenced by growth temperature. When cells were shifted to 37°C, we detected more of the alternatively spliced form, and when cells were shifted to 16°C, we detected more of the unspliced form (Fig. 5A). This is in contrast to other transcripts we examined, for which splicing appears to be unaffected by growth temperature.

Splicing regulation of a septin pre-mRNA

SPR28 (orf19.4266) is one of seven genes that encode septin proteins in *C. albicans*. Septins are structural proteins that form a network of filaments at the inner surface of the plasma membrane, and are required for a broad range of dynamic membrane events (for review, see Douglas et al. 2005). Of the nonessential *C. albicans* septins, two are required for proper hyphal morphogenesis (Warenda and Konopka 2002) and for invasive growth and dissemination following successful colonization of mice (Warenda et al. 2003). The role of *SPR28*, however, is unclear. In *S. cerevisiae*, *SPR28* is expressed specifically during meiotic division and ascospore formation (De Virgilio et al. 1996), neither of which is known to occur in *C. albicans* (Johnson 2003).

The *SPR28* gene has several unusual features in *C. albicans*. While 93% of intron-containing genes have only a single intron, *SPR28* has two. The first has a noncanonical branch site (TA**T** TAAC) located an unusually long distance (54 nt) from the 3′-splice site. The second intron has a noncanonical 5′-splice site (GT**GA**GT) found in only two other introns. These features suggested to us that *SPR28* pre-mRNA might be inefficiently spliced. They also suggested that *SPR28* gene expression might be regulated via splicing.

To test these ideas, we examined *SPR28* splicing by RT-PCR, using primers whose PCR products span both introns. We detected four distinct products, corresponding to unspliced pre-mRNA, fully spliced mRNA, and partially spliced RNA in which either the first or the second intron has been retained (Fig. 5B). Even with the inherent bias of PCR against amplification of longer products, we detected predominantly unspliced RNA and

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

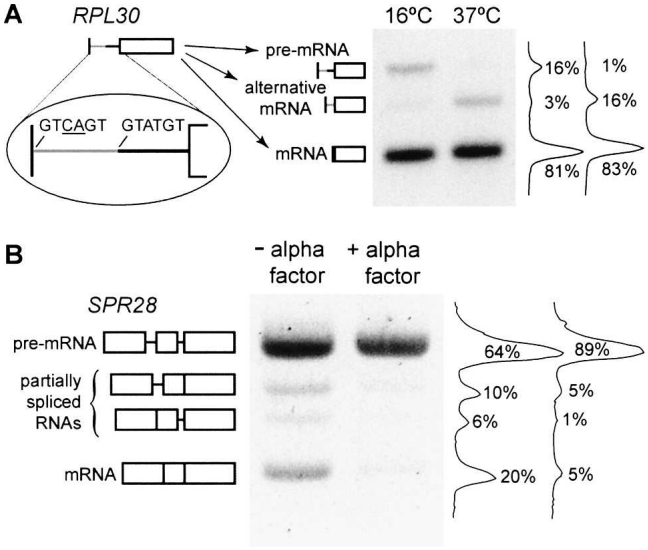


Figure 5. Splicing of *RPL30* and *SPR28* pre-mRNA is responsive to environmental signals. We resolved RT-PCR products spanning the introns of either *RPL30* or *SPR28* by gel electrophoresis. We observe the same effects in multiple experiments, and in multiple strain backgrounds (data not shown). Owing to inherent biases of PCR, volume quantitations of lane plots (*right*) do not necessarily reflect absolute RNA levels, but do reflect changes in proportion. (*A*) RNA for *RPL30* analysis was extracted from cells grown at either 16°C or 37°C. RT-PCR products were identified by sequencing and correspond to the RNAs diagrammed to the *left*. Labeled sequences correspond to 5′-splice sites, with noncanonical nucleotides underscored. (Boxes) Exons; (lines) introns; (gray lines) the portion of intron retained when the alternative 5′-splice site is used. (*B*) RNA for *SPR28* analysis was extracted from cells grown in the presence or absence of α-factor. The four products correspond in size to the RNAs diagrammed on the *left*. Identities of all forms, except for the shorter partially spliced form, were confirmed by sequencing.

relatively little spliced mRNA. This is in contrast to several genes with canonical splice sites, for which we detected very little or no unspliced RNA by RT-PCR (data not shown).

Expression of the splicing factors *LEA1* and *SLU7* is regulated by exposure of *C. albicans* cells to α-factor mating pheromone (Bennett and Johnson 2006). This suggests that splicing regulation may be a component of the response to mating pheromone. We measured the splicing of several introns in cells exposed to α-factor, including several with noncanonical splice sites, and found that while all others showed no effect, *SPR28* splicing was inhibited. In treated cells, unspliced *SPR28* RNA predominated even more, at the expense of both fully spliced and partially spliced RNA (Fig. 5B).

Discussion

Using both experimental and computational approaches, we have corrected and extended the existing *C. albicans* intron annotations, creating a high-confidence, and we believe nearly comprehensive set of 415 introns (Supplemental Table S1). Several introns we found join together exons previously thought to represent separate genes. In other cases, we found exons that were not previously annotated. Some of these identify new protein domains and therefore new functions for the genes that contain them. For example, three genes (orf19.1604, orf19.6781, and orf19.6888) have Gal4-like DNA-binding domains in their upstream exons, identifying them as probable transcription factors. We also found 16 genes that had been previously overlooked, presumably because their relatively short coding sequences are interrupted by introns. The curators of the *Candida* Genome Database have kindly agreed to host our intron annotations on their Web site (http://www.candidagenome.org).

Contrasts between the introns of *S. cerevisiae* and *C. albicans*

In many ways, the introns of *S. cerevisiae* and *C. albicans* are similar. *C. albicans* introns are present within a minority of genes, and genes that contain them usually have only one. The introns tend to be short, and their splice-site sequences conform to a strict consensus. A closer inspection, however, does reveal differences. A larger proportion of the *C. albicans* genome contains introns compared with *S. cerevisiae* (Fig. 6A). The distribution of intron sizes is also somewhat different (Fig. 6B), with a substantially greater number of very short introns (<80 nt) in *C. albicans*.

There are several differences in the characteristics of the splice-site sequences. The most dramatic is the distance between the branch point nucleotide (the final adenosine of the branch site) and the 3′-splice site, with the median distance in *S. cerevisiae* nearly twice as long (Fig. 6A). There are more subtle differences in the amount of variation within the splice-site sequences (Fig. 6C). *C. albicans* 5′-splice sites adhere slightly more to consensus, while the 3′-splice sites show more variation at the −3 position. It will be interesting to determine whether these differences correlate with changes in spliceosomal components involved in splice site recognition. We have already found numerous unexpected differences within the small nuclear RNAs (snRNAs) that function at the catalytic core of the spliceosome (Q. Mitrovich, A. Johnson, and C. Guthrie, in prep.).

Our GO term analysis identified several categories over-represented among the intron-containing genes of *C. albicans* (Fig. 4). Such over-representation suggests functional roles for these introns, perhaps in the regulation of gene expression. Categories in common with *S. cerevisiae* intron-containing genes may represent shared splicing regulatory pathways. Transcripts in both of the common categories listed—ribosomal protein genes and meiosis factors—are known to be regulated by splicing in *S. cerevisiae* (Engebrecht et al. 1991; Li et al. 1995; Nakagawa and Ogawa 1999; Davis et al. 2000; Vilardell et al. 2000). It is therefore likely that similar regulation exists in *C. albicans*, although, curiously, no one has been able to detect a meiotic cycle in this yeast (Johnson 2003). It is important to note that the over-representation of introns within these categories in *C. albicans* may be somewhat exaggerated, since the *S. cerevisiae* intron set was used to inform the annotation of introns in *C. albicans*. Even more interesting to us, however, are the categories specific to *C. albicans*, as these may reflect forms of regulation specific to *C. albicans* biology, and perhaps pathogenesis.

Introns with functional roles in gene expression

Introns were once thought of simply as nonfunctional sequences, requiring disposal before protein-coding sequences could be properly expressed. Many functional roles have since been ascribed to introns, including the evolution of novel combinations of protein domains (Roy and Gilbert 2006), generation of multiple protein variants through alternative splicing (Blencowe 2006), regulation of gene expression (Juneau et al. 2006 and references above), and hosting of snoRNAs (Maxwell and Fournier 1995). While alternative splicing and the hosting of snoRNAs within introns are notably uncommon in *S. cerevisiae*,

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

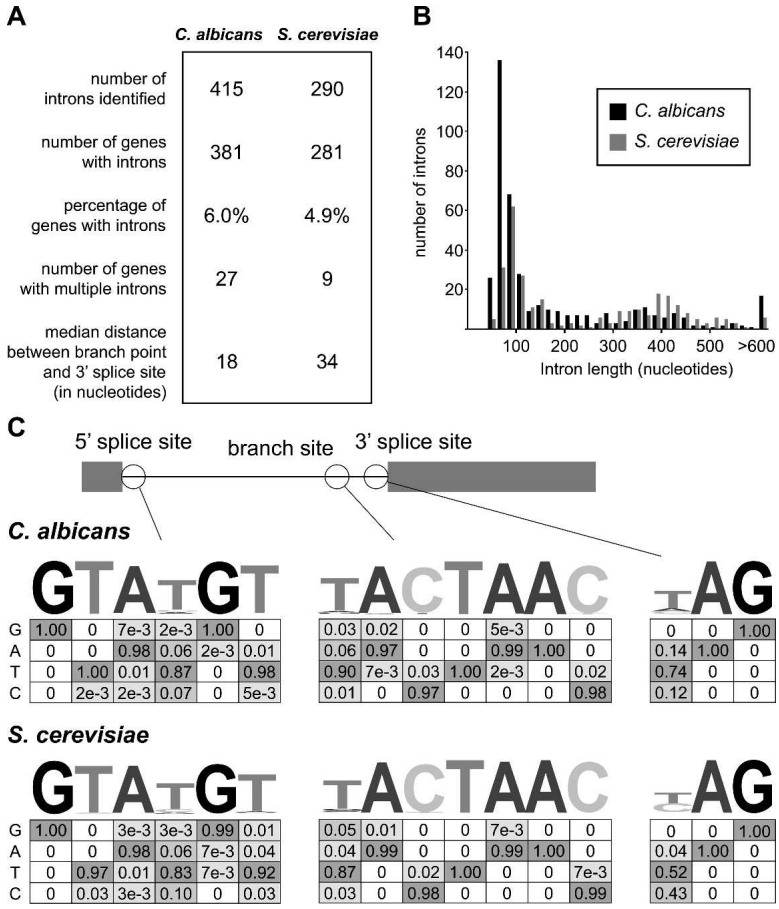


Figure 6. Intron comparisons between *C. albicans* and *S. cerevisiae*. For *C. albicans*, we used our high-confidence set of 415 introns. For *S. cerevisiae*, we compiled a set of 290 introns from various sources (http://www.yeastgenome.org, non-spliceosomal and five highly dubious introns removed; Ares lab yeast intron database at http://ribonode.ucsc.edu). (*A*) We derived intron statistics from the sequences in our sets directly. To compute the percentages of genes with introns, we used gene number estimates from Braun et al. (2005) and http://www.yeastgenome.org. (*B*) The distribution of intron lengths is shown as a histogram with a bin size of 20 nt. All introns longer than 600 nt are included in the final bin. (*C*) The nucleotide distributions within the 5′-splice sites, branch sites, and 3′-splice sites are shown as sequence logos (http://weblogo.berkeley.edu) and as numerical frequencies. Frequencies are shaded according to value, with higher frequencies shaded darker.

the control of gene expression via splicing regulation is not. We have found two examples of regulated splicing in *C. albicans*, both of which differ from *S. cerevisiae* in significant ways.

The first example of regulated splicing in *C. albicans* involves the intron in the ribosomal protein gene *RPL30*. In higher eukaryotes, excess ribosomal proteins can inhibit their own synthesis by promoting the use of unproductive alternative splice sites within their pre-mRNAs. This results in mRNAs whose coding sequences are disrupted by intron fragments and are therefore unable to express functional protein (Mitrovich and Anderson 2000; Cuccurese et al. 2005). In *S. cerevisiae*, however, ribosomal proteins can block splicing entirely, leading to accumulation of unspliced RNAs (Li et al. 1995; Vilardell et al. 2000). Regulation of *RPL30* in *C. albicans* appears to occur by both mechanisms, as we detect both unspliced and unproductive alternatively spliced forms of *RPL30* RNA. This difference with *S. cerevisiae RPL30* may reflect a more complex environmental regulation; interestingly, the form that predominates depends on growth temperature (Fig. 5A).

SPR28 is one of seven septin genes in *C. albicans*, and one of three that contain introns. In contrast, none of the *S. cerevisiae* septin genes contains introns. This over-representation in *C. albicans* suggests that splicing regulation may play an important role in septin biology. *SPR28* pre-mRNA contains two introns with unusual sequence features, and its splicing is highly inefficient. This inefficient splicing is inhibited further by exposure of cells to α-factor mating pheromone (Fig. 5B). In the future, it will be interesting to determine the roles of both *SPR28* and its negative regulation by α-factor, as well as the protein factors that mediate this response. It will also be interesting to determine whether there are any conditions in which the efficiency of *SPR28* splicing increases, which may provide insight into its biological function.

New approaches to intron discovery

The incompleteness of the previously published *C. albicans* intron annotations was likely due, in part, to the limitations inherent in traditional methods of intron identification. These methods rely primarily on well-conserved coding sequences to identify candidate introns and are particularly poor at identifying introns near the ends of coding sequences. Two of the methods we used—one experimental and one computational—overcome these limitations by focusing primarily on intron features rather than coding sequences. Once candidate introns are identified, we still rely on phylogenetic comparisons with other genomes to confirm and refine them, but these comparisons are secondary, and require only raw, unannotated genome sequence (Fig. 1). For this, it was helpful to have available a rich source of phylogenetic data in the sequenced genomes of the closely related yeasts *C. dubliniensis*, *C. tropicalis*, and *D. hansenii*.

We have shown that a debranchase mutant is a powerful tool for identifying introns experimentally. Previous studies have used other techniques to visualize the distribution of introns in *S. cerevisiae dbr1* mutants by gel electrophoresis (Ruskin et al. 1984; Spingola et al. 1999). By coupling a *dbr1* mutant to microarray analysis, we have been able to identify introns directly, on a global scale. This was especially useful for identifying 5′-UTR introns, which are difficult to validate using a purely computational approach. While they do not disrupt coding sequences, 5′-UTR introns can be just as effective as other introns in regulating gene expression, as failure to remove them usually introduces short upstream ORFs that prevent the ribosome from reaching the protein-coding ORF.

Our global computational approach was very successful, yielding 116 previously undiscovered introns. This approach is

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

particularly powerful for the Hemiascomycetes, in which alternative splicing is exceedingly rare (Davis et al. 2000), 5′-splice site and branch site sequences are highly stereotyped, and the distances between splice sites are short (Fig. 6; Bon et al. 2003). Nonetheless, because our algorithm was not strictly confined to the canonical sequences, we were able to identify many introns with deviant splice sites (Fig. 6C; Supplemental Table S1). In fact, 47% of the introns identified by our computational approach had a noncanonical 5′-splice site, branch site, or both, compared to only 21% among introns identified prior to this. These noncanonical splice sites may have contributed to the introns being overlooked in previous annotation efforts. As mentioned, splicing regulation often involves noncanonical splice sites. Identification of these introns is therefore interesting not only because it leads to more accurate protein predictions, but also because of the roles these introns may play in gene regulation.

The overall success of our approaches suggests that they will be useful for generating high-confidence intron annotations for newly sequenced genomes, as well as for refining the intron sets of previously annotated genomes. For organisms with highly stereotyped introns, our computational approach focusing primarily on intron features rather than coding-sequence alignments should be particularly useful. For organisms in which the debranchase enzyme can be deleted or inactivated, our experimental approach holds great potential. The availability of increasingly high-density microarrays should allow for more comprehensive use of this method in the future. Finally, as comprehensive intron annotations are made available, they will inform the annotations of related genomes by providing candidate intron sets. We believe that our annotation of *C. albicans* introns will thus provide an additional benefit by informing the annotations of emerging *Candida* genomes.

Methods

See Supplemental Methods for details regarding yeast strains and cell culture, RNA extraction and cDNA synthesis, RT-PCR analysis, and GO term analysis.

Phylogenetic refinement and validation of intron predictions

Candidate intron-containing genes identified by any of our methods were subjected to the same general analysis to identify and validate any introns, as outlined in Figure 1. First, we identify candidate introns by searching for canonical or degenerate splice sites (Fig. 6), and by using any other available information as a starting point (e.g., locations of known introns in other species, the sequence coordinates of probes from our microarray, candidate introns predicted by our algorithm). Next, we examine the surrounding coding sequences, and rule out introns that would destroy the ORF of the gene if spliced. Introns that lie upstream of a previously annotated coding sequence must allow extension of that coding sequence upstream of the intron, once it is spliced. (Exceptions to this rule are 5′-UTR introns, which require experimental support to be considered valid.)

Once candidate introns are identified, we examine homologous genes from other species to determine whether there is strong phylogenetic support for any of the introns. We identify homologs using BLAST; the specific program we use is determined by phylogenetic distance from *C. albicans* (BLASTN for *C. dubliniensis*, TBLASTN for *C. tropicalis*, and BLASTP for *D. hansenii* and more distantly related species). For us to consider an intron validated, it must be supported by at least one of two sets of criteria, as follows:

1. The intron is clearly conserved in the orthologous gene of at least one other species. Such support generally comes from very closely related yeasts, in which conservation is clear at the nucleotide level. The intron must disrupt the ORF in at least one species. The sequence features that define the intron—the 5′-splice site, the branch site, and the 3′-splice site—must be present in the ortholog. Aside from these features, the intron must exhibit divergence typical of non-coding sequence—nucleotide sequence is more divergent than in the surrounding protein-coding region, and the sizes of insertions/deletions are not biased toward multiples of three (i.e., the size of a codon). The continuous ORF created by removal of the intron in *C. albicans* must be conserved in the ortholog, and the amino acids encoded by the exon sequences surrounding the intron must be well enough conserved to demonstrate that the position of the intron is the same.
2. The intron is precisely absent in the most similar homolog of at least one other species. Such support generally comes from more distantly related species, and requires that the encoded proteins be very well-conserved, so that alignment within the region surrounding the intron is unambiguous despite the phylogenetic distance. The intron in *C. albicans* must disrupt the open reading frame, and splicing of the intron must generate an exon–exon junction whose coding sequence aligns precisely (at the level of the encoded protein) with the protein from the other species.

Microarray design, hybridization, and analysis

We designed custom Agilent printed microarrays using their 44,000 probe microarray format. We compiled a list of all predicted *C. albicans* ORFs derived from the orf19 contig sequences (http://www.candidagenome.org), including ORFs that had been discarded as spurious in the final annotation (Braun et al. 2005). For much of the genome, sequence from both allelic chromosomes is available (Jones et al. 2004), and we included all predicted ORFs from both alleles, where available. For every ORF, we designed three distinct 60-mer probes. The 3′-probe spans the predicted stop codon, with 32 nt downstream. To design the first 5′-probe, we began at the annotated start codon, scanned upstream for the first occurrence of a stop codon in the same reading frame, and tiled our probe just downstream of this stop codon. The second 5′-probe is the sequence immediately upstream of the first. For allelic ORFs, we used only the 3′-probe and the upstream 5′-probe.

cDNAs for microarrays were labeled with either Cy3 (*DBR1*) or Cy5 (*dbr1*Δ) mono-reactive dye (Amersham Biosciences). cDNAs were first desiccated in aliquots derived from 10 µg of total RNA and suspended in 5 µL of 100 mM sodium bicarbonate (pH 9.0). To this was added ~20 µg of dye suspended in 5 µL of DMSO. Coupling was carried out by incubation in the dark for 20 min at 60°C, followed by purification on DNA Clean & Concentrator spin columns (Zymo Research) using the manufacturer's protocol. Hybridization to microarrays and subsequent washing were carried out using the manufacturer's protocols (Agilent). Microarrays were scanned using a GenePix 4000B scanner with a 5-µm pixel resolution.

Microarray data were extracted from scans using GenePix Pro v5.1. Data were subjected to intensity-dependent loess normalization using Goulphar v1.1 (http://www.transcriptome.ens.fr/goulphar) and ranked by normalized intensity ratios. Probes exhibiting a greater than twofold *dbr1*Δ bias identified candidates for further analysis. The intron in orf19.3354 fell slightly below our twofold cutoff, but was identified in a preliminary experiment with an earlier version of our microarray. Microarray data are available at ArrayExpress (http://www.ebi.ac.uk), accession no. E-MEXP-1003.

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

Intron search algorithm

We developed our search algorithm using intron training sets from our previous analyses. From each training set, 5′-splice site, branch site, and 3′-splice site features were extracted and converted to three frequency matrices (of lengths 11, 7, and 15, respectively) by counting the occurrences of each nucleotide in each column of these features. A single nucleotide pseudocount was added to each position, except for those positions known to be invariant (G in the first position of the 5′-splice site matrix, A in the sixth position of the branch site matrix, and A and G in the last two positions of the 3′-splice site matrix). Each matrix was transformed to a log-odds position-specific scoring matrix (PSSM) by the single nucleotide frequencies of all *C. albicans* ORFs. Two distance features, the 5′-splice site to branch site distance, and the branch site to 3′-splice site distance, were also extracted from the training set and converted to frequency matrices by counting occurrences in each of 20 and 15 evenly spaced bins, respectively. A single pseudocount was added to each distance bin. Distance constraint scores were derived by log-transforming the frequency in each bin.

The score for a putative intron sequence is calculated as the sum of its three PSSM scores and two distance constraint scores. Each input sequence was searched exhaustively for a maximally scoring putative intron. Finally, the log of the squared input sequence length was subtracted from each score to correct for the increased likelihood of false positives in longer ORFs.

Acknowledgments

We thank Jing Zhu for assistance with *C. albicans* ortholog identification, Burk Braun for technical assistance and sharing of unpublished data, Oliver Homann and Hao Li for advice on computational analyses, and members of the Johnson and Guthrie labs for many helpful discussions and critical reading of our manuscript. Funding for this work was provided by NIH research grants AI49187 (A.J.) and GM21119 (C.G.), an NSF Predoctoral Graduate Fellowship (B.T.), and a Sandler Postdoctoral Research Fellowship (Q.M.).

References

Alonso-Valle, H., Acha, O., Garcia-Palomo, J.D., Farinas-Alvarez, C., Fernandez-Mazarrasa, C., and Farinas, M.C. 2003. Candidemia in a tertiary care hospital: Epidemiology and factors influencing mortality. *Eur. J. Clin. Microbiol. Infect. Dis.* **22:** 254–257.

Ares Jr., M., Grate, L., and Pauling, M.H. 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5:** 1138–1139.

Bennett, R.J. and Johnson, A.D. 2006. The role of nutrient regulation and the Gpa2 protein in the mating pheromone response of *C. albicans*. *Mol. Microbiol.* **62:** 100–119.

Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72:** 291–336.

Blencowe, B.J. 2006. Alternative splicing: New insights from global analyses. *Cell* **126:** 37–47.

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuveglise, C., Munsterkotter, M., Guldener, U., Mewes, H.W., Van Helden, J., Dujon, B., et al. 2003. Molecular evolution of eukaryotic genomes: Hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.* **31:** 1121–1135.

Braun, B.R., van Het Hoog, M., d'Enfert, C., Martchenko, M., Dungan, J., Kuo, A., Inglis, D.O., Uhl, M.A., Hogues, H., Berriman, M., et al. 2005. A human-curated annotation of the *Candida albicans* genome. *PLoS Genet.* **1:** 36–57.

Brent, M.R. 2005. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.* **15:** 1777–1786.

Byrne, K.P. and Wolfe, K.H. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15:** 1456–1461.

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

Chibana, H., Oka, N., Nakayama, H., Aoyama, T., Magee, B.B., Magee, P.T., and Mikami, Y. 2005. Sequence finishing and gene mapping for *Candida albicans* chromosome 7 and syntenic analysis against the *Saccharomyces cerevisiae* genome. *Genetics* **170:** 1525–1537.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Cuccurese, M., Russo, G., Russo, A., and Pietropaolo, C. 2005. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res.* **33:** 5965–5977.

Davis, C.A., Grate, L., Spingola, M., and Ares Jr., M. 2000. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. *Nucleic Acids Res.* **28:** 1700–1706.

De Virgilio, C., DeMarini, D.J., and Pringle, J.R. 1996. *SPR28*, a sixth member of the septin gene family in *Saccharomyces cerevisiae* that is expressed specifically in sporulating cells. *Microbiology* **142:** 2897–2905.

Douglas, L.M., Alvarez, F.J., McCreary, C., and Konopka, J.B. 2005. Septin function in yeast model systems and pathogenic fungi. *Eukaryot. Cell* **4:** 1503–1512.

Dujon, B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.* **22:** 375–387.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Engebrecht, J.A., Voelkel-Meiman, K., and Roeder, G.S. 1991. Meiosis-specific RNA splicing in yeast. *Cell* **66:** 1257–1268.

Faustino, N.A. and Cooper, T.A. 2003. Pre-mRNA splicing and human disease. *Genes & Dev.* **17:** 419–437.

Fidel Jr., P.L. 2002. Immunity to *Candida*. *Oral Dis.* **8:** 69–75.

Hull, C.M. and Johnson, A.D. 1999. Identification of a mating type-like locus in the asexual pathogenic yeast *Candida albicans*. *Science* **285:** 1271–1275.

Johnson, A. 2003. The biology of mating in *Candida albicans*. *Nat. Rev. Microbiol.* **1:** 106–116.

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., et al. 2004. The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci.* **101:** 7329–7334.

Juneau, K., Miranda, M., Hillenmeyer, M.E., Nislow, C., and Davis, R.W. 2006. Introns regulate RNA and protein abundance in yeast. *Genetics* **174:** 511–518.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kullberg, B.J. and Filler, S.G. 2002. Candidemia. In *Candida and Candidiasis* (ed. R.A. Calderone), pp. 327–340. ASM Press, Washington, DC.

Li, Z., Paulovich, A.G., and Woolford Jr., J.L. 1995. Feedback inhibition of the yeast ribosomal protein gene *CRY2* is mediated by the nucleotide sequence and secondary structure of *CRY2* pre-mRNA. *Mol. Cell. Biol.* **15:** 6454–6464.

Maxwell, E.S. and Fournier, M.J. 1995. The small nucleolar RNAs. *Annu. Rev. Biochem.* **64:** 897–934.

Mitrovich, Q.M. and Anderson, P. 2000. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in *C. elegans*. *Genes & Dev.* **14:** 2173–2184.

Nakagawa, T. and Ogawa, H. 1999. The *Saccharomyces cerevisiae MER3* gene, encoding a novel helicase-like protein, is required for crossover control in meiosis. *EMBO J.* **18:** 5714–5723.

Odds, F.C. 1988. *Candida and Candidosis*. Baillière Tindall, London.

Preker, P.J., Kim, K.S., and Guthrie, C. 2002. Expression of the essential mRNA export factor Yra1p is autoregulated by a splicing-dependent mechanism. *RNA* **8:** 969–980.

Roy, S.W. and Gilbert, W. 2006. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat. Rev. Genet.* **7:** 211–221.

Ruskin, B., Krainer, A.R., Maniatis, T., and Green, M.R. 1984. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell* **38:** 317–331.

Sherman, D., Durrens, P., Beyne, E., Nikolski, M., and Souciet, J.L. 2004. Genolevures: Comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Res.* **32:** D315–D318.

Spingola, M. and Ares Jr., M. 2000. A yeast intronic splicing enhancer and Nam8p are required for Mer1p-activated splicing. *Mol. Cell* **6:** 329–338.

Spingola, M., Grate, L., Haussler, D., and Ares Jr., M. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5:** 221–234.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. 2006. Evolution of

Downloaded from genome.cshlp.org on June 10, 2026 . Published by Cold Spring Harbor Laboratory Press

alternative transcriptional circuits with identical logic. *Nature* **443:** 415–420.
Vilardell, J., Chartrand, P., Singer, R.H., and Warner, J.R. 2000. The odyssey of a regulated transcript. *RNA* **6:** 1773–1780.
Warenda, A.J. and Konopka, J.B. 2002. Septin function in *Candida albicans* morphogenesis. *Mol. Biol. Cell* **13:** 2732–2746.
Warenda, A.J., Kauffman, S., Sherrill, T.P., Becker, J.M., and Konopka, J.B. 2003. *Candida albicans* septin mutants are defective for invasive growth and virulence. *Infect. Immun.* **71:** 4045–4051.
Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M., and Brent, M.R. 2005. Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* **15:** 577–582.

Received November 9, 2006; accepted in revised form February 7, 2007.