Analysis of fungal genomes reveals commonalities of intron loss/gain and functions in intron-poor species

Chun Shen Lim¹, Brooke N. Weinstein²,³, Scott W. Roy²,³,* and Chris M. Brown¹,*

¹ Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
² Quantitative & Systems Biology, School of Natural Sciences, University of California-Merced, Merced, CA, USA
³ Department of Biology, San Francisco State University, San Francisco, CA, USA
* Corresponding authors. Emails: chris.brown@otago.ac.nz, scottwroy@gmail.com

ABSTRACT (249 words)
Current evolutionary reconstructions predict that early eukaryotic ancestors including both the last common ancestor of eukaryotes and of all fungi had intron-rich genomes. However, some extant eukaryotes have few introns, raising the question as to why these few introns are retained. Here we have used recently available fungal genomes to address this question. Evolutionary reconstruction of intron presence and absence using 263 diverse fungal species support the idea that massive intron loss has occurred in multiple clades. The intron densities estimated in the fungal ancestral states differ from zero to 8.28 introns per one kbp of protein-coding gene. Massive intron loss has occurred not only in microsporidian parasites and saccharomycetous yeasts (0.01 and 0.05 introns/kbp on average, respectively), but also in diverse smuts and allies (e.g. Ustilago maydis, Meira miltonrushii and Malassezia globosa have 0.06, 0.10 and 0.20 introns/kbp, respectively). To investigate the roles of introns, we searched for their special characteristics using 1302 orthologous genes from eight intron-poor fungi. Notably, most of these introns are found close to the translation initiation codons. Our transcriptome and translatome data analyses showed that these introns are from genes with both higher mRNA expression and translation efficiency. Furthermore, these introns are common in specific classes of genes (e.g. genes involved in translation and Golgi vesicle transport), and rare in others (e.g. base-excision repair genes). Our study shows that fungal introns have a complex evolutionary history and underappreciated roles in gene expression.

INTRODUCTION
Spliceosomal introns are ubiquitous in eukaryotes. They are removed from all regions of the transcripts including the untranslated regions (UTRs) as well as coding sequences (CDS) (De Conti et al. 2013; Shi 2017; Lim et al. 2018). Early studies proposed that introns may be involved in
generating multi-domain genes by exon shuffling (Logsdon et al. 1995; Patthy 2003; Stoltzfus 2004; Sverdlov et al. 2005), and promoting intragenic recombination for higher fitness (Gilbert 1978; Tonegawa et al. 1978; Comeron and Kreitman 2000; Duret 2001). Notable experimentally supported roles of introns in eukaryotes include: (i) generating protein diversity by alternative splicing (Kempken 2013; Irimia and Roy 2014), (ii) harboring noncoding RNA (ncRNA) genes, such as snoRNAs and microRNAs (Chorev and Carmel 2012; Jo and Choi 2015), (iii) maintaining genome stability by decreasing the formation of DNA-RNA hybrids called R-loops (Niu 2007; Bonnet et al. 2017), (iv) intron-mediated enhancement of gene expression (Niu and Yang 2011; Gallegos and Rose 2015; Laxa 2016; Shaul 2017), (v) harboring binding sites for transcriptional or posttranscriptional regulators of gene expression (Rose 2018), (vi) allowing for an additional level of post-transcriptional regulation through regulation of RNA splicing (Witten and Ule 2011), and (vii) triggering nonsense-mediated decay in unspliced or partially spliced mRNAs through exon junction complexes (EJCs) (Mekouar et al. 2010; Grützmann et al. 2014; Zhang and Sachs 2015; Hellens et al. 2016).

Recently, we have uncovered an unexpected relationship between introns and translation, suggesting a role of 5′UTR introns in promoting translation of upstream open reading frames (Lim et al. 2018).

The most well-studied introns are those that interrupt the protein-coding regions of genes. Extensive computational studies suggest that the last eukaryotic common ancestor (LECA) had a density of introns of about 4 introns/kbp (the number of introns per one kbp of CDS on average) (Stajich et al. 2007; Csuros et al. 2011; Koonin et al. 2013; Irimia and Roy 2014). Notably, a study of 99 eukaryotic genomes has revealed a surprising variability of intron densities, ranging from 0.1 introns/kbp in the baker's yeast Saccharomyces cerevisiae to 7.8 introns/kbp in Trichoplax adhaerens (Csuros et al. 2011). Counterintuitively, T. adhaerens is one of the simplest free-living multicellular animals (Srivastava et al. 2008). The large variability of intron densities owes to remarkable differences in rates of intron loss through eukaryotic evolution (Roy and Gilbert 2005; Csuros et al. 2011) and may, in part, be due to the transposable properties of some spliceosomal introns (Roy 2004; Worden et al. 2009; Huff et al. 2016; Wu et al. 2017). Several models have also been proposed for intron loss, in particular, through genomic deletion (Loh et al. 2008; Yenerall et al. 2011; Zhu and Niu 2013a) and recombination of cDNA with genomic DNA (Fink 1987; Roy and Gilbert 2005; Zhang et al. 2010; Zhu and Niu 2013b).

As of April 2020, a total of 6,337 fungal genome assemblies were available in NCBI Genome. Fungi and their genomes are of interests for many reasons, notably as food, and plant/animal pathogens/symbionts, and for biotechnology applications (Sapountzis et al. 2015; Wheeler et al. 2017; Chan et al. 2018; Kijpornyongpan et al. 2018; Uhse et al. 2018). As fungi belong to a diverse group of organisms evolving over the past 900 million years (Dornburg et al. 2017; Kumar et al. 2017), some fungal clades have undergone massive loss of introns, in particular, the intracellular parasites.
microsporidia as well as saccharomycetous yeasts (Byrne and Wolfe 2005; Neuvéglise et al. 2011; Hooks et al. 2014; Corradi 2015; Han and Weiss 2017; Whelan et al. 2019; Priest et al. 2020; Wang et al. 2020). For instance, only 4% of *S. cerevisiae* genes have introns. In contrast, some other fungi, for example, the facultative pathogen *Cryptococcus neoformans*, have a relatively high intron density of 4 introns/kbp (Stajich et al. 2007; Csuros et al. 2011).

Previous results have suggested that frequent intron loss events, relatively few instances of intron gain, and the retention of ancestral introns characterize the evolution of introns throughout most fungal lineages (Csuros et al. 2007; Stajich et al. 2007; Csuros et al. 2011). With thousands of fungal genomes available to date (Priest et al. 2020), it is timely to revisit the ancestral states and scale of intron gain or loss in the fungal kingdom. Our analysis includes representatives from nearly all phylum-level clades, including the early-diverging Blastocladiomycota, Chytridiomycota, Mucoromycota, Zoopagomycota, Cryptomycota, and Microsporidia phyla. The diversity of intron-exon structures and the wealth of kingdom-wide genomic resources of fungi make them excellent models for studying the intron gain and loss dynamics and the functional roles of introns (Priest et al. 2020). Here we analyzed the introns of 644 fungal genomes, inferring ancestral states, conservation, and functional processes.

**RESULTS**

**Evolutionary reconstruction reveals high ancestral intron densities and a general bias towards intron loss over intron gain**

We aligned protein sequences and mapped corresponding intron positions for 1445 sets of orthologous genes from 263 fungal species. We reconstructed the evolutionary history of intron gain or loss among these species (Figure 1; see Materials and Methods). These 263 species represented a wide variety of intron densities, from various intronless Microsporidia to 4.8 introns/kbp in the chytrid *Gonapodya prolifera*. This reconstruction revealed a remarkably dynamic and diverse history of intron loss and, with many episodes of massive intron loss and/or gain coupled to general stasis within large clades of organisms (e.g., very low intron densities within all Microsporidia and similar intron densities among nearly all Pezizomycotina). Most strikingly, we reconstructed very high intron densities (Figure 2; Table 1), with some 8.3 introns/kbp reconstructed in the fungal ancestor. While it may be counterintuitive that the ancestral fungus harbored nearly twice as many introns as any modern fungus in this dataset, this finding is in keeping with previous results showing a general bias towards intron loss over intron gain in many lineages, and echoes the finding of considerably higher intron densities in alveolate ancestors than in modern alveolates (Csuros et al. 2011). While these results are in general agreement with previous studies that inferred intron-rich ancestral fungi (Stajich et al. 2007; Csuros et al. 2011; Grau-Bové et al. 2017), our inferred densities are considerably higher,
likely due to improved model specification made possible by greater species density. Interestingly, our reconstructed value is close to the inferred intron content of the animal ancestor (8.8 introns/kbp) in a study using the same reconstruction method on a smaller, eukaryote-wide dataset (Csuros et al. 2011).

In contrast to intron-rich ancestral states, almost three-quarters of fungi have maintained less than 10% of the intron density of the last fungal ancestor (191 of 263 species; Figure 1; see also Supplementary Table S1 and S2).

These results also illuminate the history of massive intron loss in these same two lineages. Many studies have found that the obligate intracellular microsporidian parasites have zero or few introns (Keeling et al. 2010; Cuomo et al. 2012; Peyretaillade et al. 2012; Corradi 2015; Desjardins et al. 2015; Han and Weiss 2017; Mikhailov et al. 2017; Ndikumana et al. 2017) and that saccharomycetous yeasts have lost most of their introns (Stajich et al. 2007; Csuros et al. 2011; Hooks et al. 2014). For both remarkable groups, our analysis includes newly available genomes including relatively intron-rich sister species (Rozella allomycis (2.7 introns/kbp) for Microsporidia and Lipomyces starkeyi (1.2 introns/kbp) for Saccharomycotina), allowing for improved resolution of the history of these organisms. In both lineages our reconstructions reveal a massive intron loss event leading to the ancestor of a large clade of intron-poor organisms. However, whereas in Microsporidia this loss event occurred in the ancestor of the group after divergence from Cryptomycota, for saccharomycetous yeast this massive loss event occurs within the group, after divergence of L. starkeyi.

A general bias towards intron loss punctuated by several independent episodes of intron gain

A bias towards intron loss over intron gain is seen across the fungal tree. This is evident not only in Microsporidia and Saccharomycotina but also in groups with more moderate intron densities, including the filamentous ascomycetes Pezizomycotina (120 of 122 species) and smuts/allies Ustilaginomycotina (15 of 20 species). Indeed, we find a striking bias towards intron loss over gain. Among branches estimated to have undergone at least 5% change in intron density, ten times as many have more loss than gain. Remarkably, a bias is seen even for lineages with very little change, in which intron loss outweighs gain three-fold (Supplementary Figure S1).

While ongoing intron loss is characteristic of most lineages, our results indicate several substantial episodes of intron gain. Within Basidiomycotina, we estimated a 26% increase in intron density leading to the ancestor of Ustilaginomycotina and an 18% increase in the ancestor of Pucciniomycotina. The most substantial intron gains occurred, unexpectedly, within the famously intron-poor lineages Microsporidia and Saccharomycotina. We inferred substantial, secondary independent intron gain in two extant microsporidians, (Nosema bombycis and Nosema apis) and four saccharomycetous yeasts (Scheffersomyces stipitis, Candida maltosa, Pichia kudriavzevii, and...
Spathaspora passalidarum). While preliminary analysis suggests the reality of some of these gains, it is worthy of note that, given the small absolute number of gains involved (leading to <1 intron/kbp), further detailed analysis will be necessary to confirm these episodes.

No relationship of genome size to intron number

Given that genome size has been argued to relate to intron number, organismal complexity, population size and generation time (Lynch and Conery 2003; Koonin et al. 2013), we examined the relationship between genome size and intron density using phylogenetic independent contrasts. Remarkably, we found no evidence for a positive relationship between genome size and intron number — indeed, the correlation is slightly and non-significantly negative (Figure 3; Spearman’s rho = −0.070, p-value = 0.41).

Functional biases of intron-containing genes in intron-poor species

To better understand the evolutionary forces responsible for maintenance of introns through evolution, we chose eight intron-poor species (with intron densities <10% of the fungal ancestor), identified orthologous genes, and analyzed the selective pressures on intron-containing and intronless genes. These species comprised of S. cerevisiae and Candida dubliniensis in Saccharomycotina, Cyphellophora europaea and Sporothrix schenckii in Pezizomycotina, and Ustilago maydis, Pseudozyma hubeiensis, Meira miltonrushii and Malassezia sympodialis in Ustilaginomycotina (Figure 1, green filled circles), representing six separate massive reductions in intron number. Comparison of intron-containing genes with intronless genes revealed a number of differences. We found that intron-containing genes are less likely to have undergone recent positive selection on their protein-coding meaning (Figure 4A, one-sided Fisher’s exact tests, p-value < 0.05).

We also found an association with gene duplication. Significantly higher proportions of the intron-containing genes in S. cerevisiae and U. maydis are duplicated (Figure 4B, two-sided Fisher’s exact tests, p-value < 0.05). While this finding in S. cerevisiae could largely be explained by the previously-noted concentration of introns in ribosomal protein genes, in which introns facilitate cross-regulation among paralogous genes (Pleiss et al. 2007; Parenteau et al. 2011; Petibon et al. 2016; Parenteau and Abou Elela 2019), it cannot explain the bias in U. maydis.

Concordance of the presence of introns in orthologs in species with independent massive intron loss

If introns carry useful functions, we hypothesize that they should be maintained in the orthologs of the intron-poor species. We determined ratios of the intron-containing orthologs among the intron-poor species and compared these ratios with null expectations (i.e. the proportions of orthologs with introns...
for the intron-poor species; Figure 5A). Indeed, these orthologs tend to harbor introns concordantly
(Figure 5B). Strikingly, two orthologous genes, RPL7B and NOG2 have conserved intron positions in
all eight studied intron-poor species (Figure 6). RPL7B encodes a ribosomal 60S protein whereas
NOG2 encodes a putative GTPase-associated pre-60S ribosomal subunit, in which their introns harbor
a box C/D and a box H/ACA snoRNAs, respectively [snR59 and snR191 in *S. cerevisiae*,
respectively; *Saccharomyces* Genome Database (Cherry et al. 2012)]. This suggests that introns with
conserved positions have functions (e.g. as snoRNAs). Interestingly, this conservation may not be
trackable to LECA as snoRNA genes are mobile (Weber 2006; Luo and Li 2007; Schmitz et al. 2008;
Hoeppner and Poole 2012), however this suggests an ancient association within fungi.

We next asked whether the distances of introns with respect to translation initiation codons
are conserved. We compared the distribution of the first introns in the coding genes with two null
distributions (Figure 7). Both null distributions were generated using *S. cerevisiae* genes. The first
null distribution was generated based on our observation that *S. cerevisiae* has only 1 or 2 introns in
its coding genes (GCA_000146045, total 282 spliceosomal introns, 273 of 6619 coding genes have
introns). Therefore, the expected distances of the first introns from the start codons are either half or
one-third of the CDS lengths (Figure 7, dotted lines in light gray). The second null distribution was
generated in a similar way but including UTR introns and centering at the transcription start or
termination sites (Figure 8, dashed lines in dark gray). We then divided the genes into two groups,
genomes related to translation and other classes of genes using gene ontology (GO) terms (Cherry et al.
2012).

We observed that introns are closer to the initiation codons than these null distributions
(Figure 7), which is in agreement with previous studies (Bon et al. 2003; Mourier and Jeffares 2003;
Russell et al. 2005; Franzén et al. 2013). This observation is consistent irrespective of the roles of
intron-containing genes in translation, supporting the idea that introns may have regulatory roles in
both transcription and translation (Lim et al. 2018).

**Roles of introns in gene expression**

Previous studies have shown that introns are common in the ribosomal protein genes (e.g. *RPL7B*) of
intron-poor protozoa and saccharomycetous yeasts (Bon et al. 2003; Russell et al. 2005; Franzén et al.
2013). However, the abundance of introns in other classes of genes is less well-known. We examined
the GO terms of the orthologs of the intron-poor species. We found that introns are highly abundant
not only in genes involved in cytoplasmic translation (e.g. ribosomal proteins) but also in genes
involved in Golgi vesicle transport (Figure 7). In contrast, introns are depleted in genes involved in
base-excision repair and peptidyl-proline modification. The reasons for these biases are still unclear.
These findings prompted us to compare the transcription level and translation efficiency between intron-containing and intronless genes. We analyzed the matched RNA-seq and ribosome profiling datasets for the fungal species that are publicly available — *S. cerevisiae* (Heyer and Moore 2016), *Candida albicans* (Muzzey et al. 2014), *S. pombe* (Subtelny et al. 2014), and *Neurospora crassa* (Yu et al. 2015) (Supplementary Table S3 and Materials and Methods).

Interestingly, intron-containing genes tend to have higher mRNA expression and translation efficiency than that of intronless genes (Figure 9). Overall, our results provide independent evidence of diverse roles of fungal introns in transcription and translation.

**DISCUSSION**

**Widespread intron loss in the fungal kingdom**

This study has shown that intron loss was widespread in the fungal kingdom during evolution (Figure 1). The most extreme cases are microsporidian parasites, which have lost all, or nearly all introns. Microsporidian parasites have the smallest eukaryotic genomes and coding capacities known to date (Corradi 2015; Han and Weiss 2017). Intriguingly, Chytridiomycota and Mucaromycota, two other early-diverging phyla, are instead characterized by the retention of ancestral introns and maintain relatively high intron densities. *Gonapodya prolifera*, a chytrid fungus, has the highest intron density of all the fungi in our analysis (4.8 introns/kbp), 73% of the intron density of humans *Homo sapiens*.

As previously observed, ascomycetes have lost more introns than basidiomycetes, which is in agreement with a study of 99 eukaryotic species (Csuros et al. 2011). Many ascomycetes are unicellular fungi or yeasts, known to have low numbers of introns (in particular, the subphylum Saccharomycotina) (Byrne and Wolfe 2005; Neuvéglise et al. 2011; Hooks et al. 2014). Our analysis has further shown substantial intron loss in Pezizomycotina, notably the *Cyphellophora* and Sporothrix* sp.* (below 10% of the intron densities of the last fungal ancestor). These species are conidia producing fungi that have a yeast or yeast-like stage as part of their life cycle (Barros et al. 2011; Feng et al. 2012).

In contrast, many basidiomycetes are fruit-body producing fungi that are known to have relatively higher numbers of introns. However, prominent intron loss has also occurred in Ustilaginomycotina (Kämper et al. 2006; Stajich et al. 2007). Many of these smuts and allies have evolved into plant pathogens and evolved a yeast state as part of their life cycle (Rush and Aime 2013; Wang et al. 2015; Benevenuto et al. 2018; Kijpornyongpan et al. 2018).

However, not all yeasts or yeast-like fungi have low intron densities. For example, *Pneumocystis murina* and *Cryptococcus neoformans*, which are ascomycetous and basidiomycetous yeasts, have high intron densities (3.7 and 3.6 introns/kbp).
Concerted evolution’ of introns and their host genes

What is the function of introns? The role of most introns is unclear as they are mostly dispensable (Niu 2008). To address this we chose eight ascomycetes and basidiomycetes with extensive intron loss for in-depth analysis. These intron-poor species all have a yeast or yeast-like stage in their life cycle. Our evolutionary and statistical approaches have shown that remaining introns are unlikely to be conserved by chance (Figure 4, 5 and 6).

Several studies have shown that the 5’ splice sites of intron-poor species are more conserved than that of intron-rich species (Irimia et al. 2007; Skelly et al. 2009; Neuvéglise et al. 2011). In addition, previous studies have shown that deleting most introns in S. cerevisiae does not significantly compromise growth but starvation resistance (Parenteau et al. 2008; Parenteau et al. 2011; Parenteau et al. 2019). These support our idea that introns are retained because of their useful functions.

Interestingly, S. cerevisiae and U. maydis have significantly higher proportions of duplicated genes with introns (Figure 4). Functional divergence might have occurred in one of the paralogs (and their introns) as shown in S. cerevisiae (Kellis et al. 2004; Pleiss et al. 2007; Parenteau et al. 2011; Petibon et al. 2016; Parenteau and Abou Elela 2019). It would be interesting to see whether the introns of duplicated genes in U. maydis have similar roles as that of S. cerevisiae.

Regulatory roles of introns in transcription and translation

Notably, most of the first introns are located near the translation initiation codons (Figure 7). Indeed, intron loss near the 3’ end of a gene was previously found to be prevalent in some protozoa and fungi, probably due to reverse transcriptase-mediated intron loss (Fink 1987; Roy and Gilbert 2005; Russell et al. 2005; Lee et al. 2010; Zhang et al. 2010; Franzén et al. 2013; Koonin et al. 2013; Zhu and Niu 2013a; Zhu and Niu 2013b; Irimia and Roy 2014). This location bias of introns has also been found in intron-rich metazoa and plants (Lim et al. 2018).

Introns are also more abundant in ancient genes, in particular, ribosomal protein genes (Figure 8). This is in agreement with a previous study on seven saccharomycetous yeasts (Bon et al. 2003). In addition, introns are more abundant in genes that have higher mRNA expression and translation efficiency, irrespective of their cellular functions (Figure 9). This extends previous analyses of global gene expression of S. cerevisiae (Juneau et al. 2006; Hoshida et al. 2017). In metazoa and plants, introns may enhance transcription or translation, in part, through EJCs (Wiegand et al. 2003; Diem et al. 2007; Chazal et al. 2013; Le Hir et al. 2016). EJCs deposit at about 20-24 bases upstream of the exon-exon junctions upon splicing, carrying over the ‘memory’ of splicing events to cytoplasmic translation. However, S. cerevisiae has no EJCs, unlike complex eukaryotes or even the fission yeast S. pombe. It remains unclear how intron enhances transcription and translation in Saccharomycetes (Moabbi et al. 2012; Hoshida et al. 2017).
We propose that highly conserved intron positions are indicative of functional importance, e.g. the ncRNA gene snR191 embedded in the intron of NOG2 gene (Figure 6). This intron was previously found to be highly conserved in the family Saccharomycetaceae (Hooks et al. 2014; Hooks et al. 2016). Some other introns may harbor functional structured RNA elements, such as the introns of RPL18A and RPS22B pre-mRNAs that promote RNase III-mediated degradation, and the GLC7 intron that modulates gene expression during salt stress (Danin-Kreiselman et al. 2003; Juneau et al. 2006; Parenteau et al. 2008; Hooks et al. 2016).

Concluding remarks
By encompassing an unprecedented number of species, from a single group of eukaryotes with a range of very different evolutionary histories, these results allow us to better understand commonalities of intron evolution. We find a remarkable trend towards intron number reduction across lineages, as well as highly predictable patterns of intron retention in intron-poor species at the level of gene function, specific gene, specific intron, and genic position. These results provide explanations for long-observed patterns, while revealing previously unknown patterns to be explained by future studies.

MATERIALS AND METHODS
Genome sequences and annotations
We retrieved 633 fungal genomes (FASTA and GTF files) from the Ensembl Fungi release 34 (Zerbino et al. 2018). In addition, the L. starkeyi and Neolecta irregularis genomes were retrieved from Ensembl Fungi 42 and NCBI Genome, respectively, whereas seven Ustilaginomycotina and two Taphrinomycotina genomes from JGI MycoCosm (Cissé et al. 2013; Grigoriev et al. 2014; Riley et al. 2016; Mondo et al. 2017; Nguyen et al. 2017; Kijpornyongpan et al. 2018). Detailed information can be found in Supplementary Table S1.

Redundant species were filtered by assembly level (ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/assembly_summary_genbank.txt) (Kitts et al. 2016). Complete genomes were retained, otherwise the assemblies at the chromosome, scaffold, or contig levels. For redundant assemblies, only the assemblies with the highest numbers of CDS were retained. For outgroups, the genomes of Homo sapiens and the cellular slime mold Fonticula alba were downloaded from Ensembl 95 and Ensembl Protists 42, respectively.

The annotation of the UTR and UTR introns of S. cerevisiae was retrieved from YeastMine (Balakrishnan et al. 2012). The GO terms of S. cerevisiae were retrieved from the Saccharomyces Genome Database (Cherry et al. 2012).
Taxonomic and phylogenetic trees

We chose a subset of 263 fungi for the inference of ancestral introns then pruned an 1100 taxa tree from concatenated analyses (J. Stajich, personal communication, December 24, 2018). *Homo sapiens* and *Fonticula alba* were included as outgroups. For visualization, the tips and nodes were color-coded by inferred intron densities using the R package ggtree v1.16.6 (Yu et al. 2017).

For phylogenetic independent contrasts analysis, we retrieved a phylogenetic tree from the SILVA database release-111 (Yarza et al. 2017). The tip labels were replaced using AfterPhylo.pl and the tree was pruned using filter_tree.py (Caporaso et al. 2010; Zhu 2014).

Orthology analysis

For the inference of ancestral introns, orthologous genes were identified using HMMER v3.1b2 (Johnson et al. 2010). Profile hidden markov models (HMMs) were retrieved from the 1000 Fungal Genomes Project (1KFG) and fuNOG (eggNOG v4.5) (Huerta-Cepas, Szklarczyk, et al. 2016; Bewick et al. 2019). A HMM database was built using hmmpress. Homology sequences were detected using hmmsearch. For species that have multiple hits per HMM, only the top hit was retained. To remove false positives, hits with bit scores over 276.48 were retained. This threshold was estimated from the distribution of bit scores (bimodal lognormal) using the R package cutoff v0.1.0 (Choisy 2015). Only the orthologs that captured over 80% (212/265) of the species were used in the subsequent analyses (1445 orthologs).

Eight intron-poor species were selected for analysis of intron functions, including *S. cerevisiae* and *C. dubliniensis* in Saccharomycotina, *C. europaea* and *S. schenckii* in Pezizomycotina, and *U. maydis*, *P. hubeiensis*, *M. miltornrushii* and *M. sympodialis* in Ustilaginomycotina. The orthologs of these intron-poor species were identified using proteinortho5 (using parameter -synteny) (Lechner et al. 2011). A total of 1302 orthologs were identified. In contrast to the above approach, this approach is less scalable but unrestricted by a predefined set of orthologs (HMMs).

Duplicated genes were identified using SkewGD v1 (Tian 2018). This pipeline includes sequence clustering and ‘age’ estimation using $K_s$ (the number of synonymous substitutions per synonymous site) (Blanc and Wolfe 2004; Vanneste et al. 2013).

Intron alignment

For the inference of ancestral introns, protein sequences were aligned using Clustal Omega v1.2.4 (using parameter --hmm-in) (Sievers and Higgins 2018). Annotations of intron positions were extracted from GTF/GFF files using ReSplicer (by calling the splice.extractAnnotations class) (Séton Bocco and Csűrös 2016). The alignments were realigned using IntronAlignment (Csurös et al. 2007).
The orthologs of the intron-poor species were aligned using MUSCLE v3.8.31 (Edgar 2004).

The protein sequences were realigned using ReSplicer and IntronAlignment as above. Introns were then re-annotated using ReSplicer, by calling a series of java classes splice.extractAnnotations, splice.collectStatistics, and splice.checkSites. Realignment was repeated using re-annotated intron positions.

**Inference of ancestral introns**

We inferred ancestral introns from 1445 orthologs of 263 fungal genomes using Malin (Csúrös 2008). Firstly, we generated a table of intron presence/absence in the orthologs using Malin. It included 46,381 intron sites allowing a maximum of 53 ambiguous characters per site.

Failure to account for variation in intron loss rate across sites can lead to an underestimation in intron density of eukaryotic ancestors (Stajich et al. 2007), and previous experiments with rate variation models across sites in Malin showed that model fit was significantly impacted solely by variation in loss rate across intron sites (Csúrös et al. 2011). Here, intron gain and loss rates were optimized in Malin using maximum likelihood using the constant rate and rate-variation models starting from the standard null model and running 1000 optimization rounds (likelihood convergence threshold = 0.001). For the constant rate model, each intron site has only a branch-specific gain and loss rate. In contrast, for the rate-variation model, intron sites additionally belong to one of two discrete rate loss categories.

Malin calculates gain/loss rates and intron density at the root by numerical optimization of the likelihood. For both the constant rate and rate-variation models, we used 100 bootstrap replicates of the intron table to assess uncertainty about inferred rate parameters and intron site history for every node. For model comparison, the likelihood-ratio test statistic calculated as \( \Delta = -2 \times (L_1 - L_2) \), where \( L_1 \) is the log-likelihood of the constant rate model (\( L_1 = -402882 \)) and \( L_2 \) is the log-likelihood of the rate-variation model (\( L_2 = -397548 \)). The likelihood-ratio test statistic is 10,668, which was then compared to a \( \chi^2 \) distribution with one degree of freedom. In this comparison, we obtained a \( p \)-value of 0.0. Therefore, we rejected the constant rate results and chose the more complex rate-variation model. Finally, we inferred ancestral densities by using Dollo parsimony (Farris 1977).

For all analyses, we scaled the number of inferred introns to intron density by multiplying by 0.37 and dividing by 322, where 0.37 and 322 are intron density and the number of introns in *Schizosaccharomyces pombe* in the orthologous dataset, respectively. *S. pombe* was used as a reference because it has a high-quality annotation and over an order of magnitude higher intron density than *S. cerevisiae*(Csuros et al. 2011; Lock et al. 2018).
Phylogenetic independent contrasts analysis

Three features (intron density, genome size, and the number of CDS) were first examined for normality using different transformation functions in the R package bestNormalize v1.4.3 (Peterson 2018). Ordered quantile transformation was chosen. Phylogenetic independent contrasts analysis was carried out using the R package caper v1.0.1 (Orme et al. 2018).

Branch-site test

The orthologous protein sequences were aligned using PRANK v.150803 (Löytynoja and Goldman 2008; Jeffares et al. 2015). The aligned protein sequences were converted to aligned DNA sequences using PAL2NAL (Suyama et al. 2006). These aligned DNA sequences were used to build phylogenetic trees using RaxML v8.2.9 (using parameters -f a -x 1181 -N 1000 -m GTRGAMMA) (Stamatakis 2014). To identify positively selected genes, branch-site tests were performed using both the aligned DNA sequences and phylogenetic trees using ETE toolkit v3.1.1 (ete-evol, a CodeML wrapper) (Yang 2007; Huerta-Cepas, Serra, et al. 2016). The positive selection (bsA, alternative hypothesis) and relaxation (bsA1, null hypothesis) evolutionary models were fit to the orthologous dataset. This involved modeling each branch by recursively marking the remaining branches as the foreground branches, and comparing them using likelihood-ratio tests (using parameters --models M0 bsA bsA1 --leaves --tests bsA,bsA1).

snoRNA prediction

The Stockholm alignment files of fungal snoRNA families were downloaded from http://www.bioinf.uni-leipzig.de/publications/supplements/17-001 (Canzler et al. 2018). These files were used to build HMMs or covariance models using Infernal v1.1.2 (Nawrocki and Eddy 2013). These models were used to detect the snoRNA genes encoded by introns. The functional elements in the snoRNAs were predicted using snoscan v0.2b and the snoGPS web server (Lowe and Eddy 1999; Schattner et al. 2005).

Gene ontology analysis

Functional annotation of S. cerevisiae genes was performed using the Bioconductor packages clusterProfiler v3.0.5 and org.Sc.sgd.db v3.4.0 (Yu et al. 2012; Huber et al. 2015; Carlson 2017). Redundant GO terms were removed using the simplify function with default settings. Genes were grouped by GO terms and normalized/plotted using the R package massageR v0.7.2 (Stanstrup 2017).

RNA-seq and ribosome profiling data analyses
List of RNA-seq and ribosome profiling datasets used are available in Supplementary Table S3. The genome and annotation files of *Candida albicans* and *Schizosaccharomyces pombe* were downloaded from the *Candida* Genome Database assembly 22 and PomBase release 30, respectively (Skrzypek et al. 2017; Lock et al. 2018).

Reads were aligned to ncRNAs using STAR v2.5.2.b (Dobin et al. 2013). Unmapped reads were then aligned to the genome with transcript annotation. Uniquely mapped reads were counted using featureCounts v1.5.0-p3 (Liao et al. 2014).

For RNA-seq, count data were normalized to Reads Per Kilobase per Million (RPKM) mapped reads. RPKM = \( \frac{\text{read_counts}}{(\text{gene_length}/1000)/(\text{total_read_counts}/10^6)} \). For ribosome profiling, count data were used to calculate translation efficiency. Translation efficiency = \( \frac{\text{ribosome_footprints/total_footprint_counts}}{\text{RNA-seq_read_counts/total RNA-seq_read_counts}} \).

We detected the *S. cerevisiae* orthologs in other species using proteinortho5 (using parameter -synteny) (Lechner et al. 2011). We found 3063, 2506, and 2541 *S. cerevisiae* orthologs in *C. albicans*, *S. pombe*, and *N. crassa*, respectively. The orthologs were grouped by introns presence/absence and GO terms. These data were then mapped to the RNA-seq and ribosome profiling results by gene.

**Statistical analysis**

Statistical analysis and plotting were performed using R ≥3.4 (R Core Team 2018). Fisher’s exact test, the chi-square test, Welch two sample t-test and Spearman’s rank correlation were calculated using the base R system. Computation of binomial confidence intervals using Bayesian inference was performed using the binom package (Dorai-Raj 2014). All \( p \)-values obtained from multiple testing were adjusted using the Bonferroni correction to avoid false positives (Armstrong 2014), unless otherwise mentioned. Plots were constructed using ggplot2 v2.2.1 (Wickham 2016), unless otherwise stated.

**Data availability**

Code and data for this study are available at https://github.com/Brookesloci/fungi_intron_paper_2020/.

**ACKNOWLEDGMENTS**

CMB and CSL were funded by the University of Otago. CSL was a recipient of a Dr. Sulaiman Daud 125th Jubilee Postgraduate Scholarship. SWR and BNW were supported by the National Science Foundation award #1616878.
REFERENCES

Andrews S. 2016. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Armstrong RA. 2014. When to use the Bonferroni correction. Ophthal. Physiol. Opt. 34:502–508.

Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, Sullivan J, Micklem G, Cherry JM. 2012. YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database 2012:bar062.

Barros MB d. L, M B d, de Almeida Paes R, Schubach AO. 2011. Sporotrichosis. Clin. Microbiol. Rev. 24:633–654.

Benevenuto J, Teixeira-Silva NS, Kuramae EE, Croll D, Monteiro-Vitorello CB. 2018. Comparative genomics of smut pathogens: Insights from orphans and positively selected genes into host specialization. Front. Microbiol. 9:660.

Bewick AJ, Hofmeister BT, Powers RA, Mondo SJ, Grigoriev IV, James TY, Stajich JE, Schmitz RJ. 2019. Diversity of cytosine methylation across the fungal tree of life. Nat Ecol Evol 3:479–490.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678.

Bon E, Casaregola S, Blandin G, Llorente B, Neuvèglise 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.

Bonnet A, Grosso AR, Elkaoutari A, Coleno E, Presle A, Sridhar SC, Janbon G, Géli V, de Almeida SF, Palancade B. 2017. Introns protect eukaryotic genomes from transcription-associated genetic instability. Mol. Cell 67:608–621.e6.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15:1456–1461.

Canzler S, Stadler PF, Schor J. 2018. The fungal snoRNAome. RNA 24:342–360.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JJ, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7:335–336.

Carlson M. 2017. org.Sc.sgd.db: Genome wide annotation for Yeast. Available from: http://bioconductor.org/packages/org.Sc.sgd.db/

Chan LG, Cohen JL, de Moura Bell JMLN. 2018. Conversion of agricultural streams and food-processing by-products to value-added compounds using filamentous fungi. Annu. Rev. Food Sci. Technol. 9:503–523.

Chazal P-E, Daguenet E, Wendling C, Ulryck N, Tomasetto C, Sargueil B, Le Hir H. 2013. EJC core component MLN51 interacts with eIF3 and activates translation. Proc. Natl. Acad. Sci. U. S. A.
110:5903–5908.
Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40:D700–D705.
Choisy M. 2015. cutoff: Identify a cutoff value from bimodal data. Available from: https://github.com/choisy/cutoff
Chorev M, Carmel L. 2012. The function of introns. Front. Genet. 3:55.
Cissé OH, Almeida JMGCF, Fonseca A, Kumar AA, Salojärvi J, Overmyer K, Hauser PM, Pagni M. 2013. Genome sequencing of the plant pathogen Taphrina deformans, the causal agent of peach leaf curl. MBio 4:e00055–13.
Comeron JM, Kreitman M. 2000. The correlation between intron length and recombination in Drosophila. Dynamic equilibrium between mutational and selective forces. Genetics 156:1175–1190.
Corradi N. 2015. Microsporidia: Eukaryotic intracellular parasites shaped by gene loss and horizontal gene transfers. Annu. Rev. Microbiol. 69:167–183.
Csuros M. 2008. Malin: maximum likelihood analysis of intron evolution in eukaryotes. Bioinformatics 24:1538–1539.
Csuros M, Holey JA, Rogozin IB. 2007. In search of lost introns. Bioinformatics 23:i87–i96.
Csuros M, Rogozin IB, Koonin EV. 2011. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput. Biol. 7:e1002150.
Cuomo CA, Desjardins CA, Bakowski MA, Goldberg J, Ma AT, Becnel JJ, Didier ES, Fan L, Heiman DI, Levin JZ, et al. 2012. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. 22:2478–2488.
Danin-Kreiselman M, Lee CY, Chanfreau G. 2003. RNAse III-mediated degradation of unspliced pre-mRNAs and lariat introns. Mol. Cell 11:1279–1289.
De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA 4:49–60.
Desjardins CA, Sanscrainte ND, Goldberg JM, Heiman D, Young S, Zeng Q, Madhani HD, Becnel JJ, Cuomo CA. 2015. Contrasting host-pathogen interactions and genome evolution in two generalist and specialist microsporidian pathogens of mosquitoes. Nat. Commun. 6:7121.
Diem MD, Chan CC, Younis I, Dreyfuss G. 2007. PYM binds the cytoplasmic exon-junction complex and ribosomes to enhance translation of spliced mRNAs. Nat. Struct. Mol. Biol. 14:1173–1179.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.
Dorai-Raj S. 2014. binom: Binomial confidence intervals for several parameterizations. Available
from: https://cran.r-project.org/web/packages/binom/index.html

Dornburg A, Townsend JP, Wang Z. 2017. Maximizing power in phylogenetics and phylogenomics: A perspective illuminated by fungal big data. *Adv. Genet.* 100:1–47.

Duret L. 2001. Why do genes have introns? Recombination might add a new piece to the puzzle. *Trends Genet.* 17:172–175.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Farris JS. 1977. Phylogenetic Analysis Under Dollo’s Law. *Syst. Biol.* 26:77–88.

Feng P, Lu Q, Najafzadeh MJ, van den Ende AHGG, Sun J, Li R, Xi L, Vicente VA, Lai W, Lu C, et al. 2012. *Cyphellophora* and its relatives in *Phialophora*: biodiversity and possible role in human infection. *Fungal Divers.* 65:17–45.

Fink GR. 1987. Pseudogenes in yeast? *Cell* 49:5–6.

Franzén O, Jerlström-Hultqvist J, Einarsson E, Ankarklev J, Ferella M, Andersson B, Svärd SG. 2013. Transcriptome profiling of *Giardia intestinalis* using strand-specific RNA-seq. *PLoS Comput. Biol.* 9:e1003000.

Gallegos JE, Rose AB. 2015. The enduring mystery of intron-mediated enhancement. *Plant Sci.* 237:8–15.

Gilbert W. 1978. Why genes in pieces? *Nature* 271:501–501.

Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I. 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. *Elife* [Internet] 6. Available from: http://dx.doi.org/10.7554/eLife.26036

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42:D699–D704.

Grützmann K, Szafranski K, Pohl M, Voigt K, Petzold A, Schuster S. 2014. Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. *DNA Res.* 21:27–39.

Han B, Weiss LM. 2017. Microsporidia: Obligate intracellular pathogens within the fungal kingdom. *Microbiol Spectr* 5:FUNK – 0018–2016.

Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC. 2016. The emerging world of small ORFs. *Trends Plant Sci.* 21:317–328.

Heyer EE, Moore MJ. 2016. Redefining the translational status of 80S monosomes. *Cell* 164:757–769.

Hoeppner MP, Poole AM. 2012. Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC Evol. Biol.* 12:183.
Hooks KB, Delneri D, Griffiths-Jones S. 2014. Intron evolution in Saccharomycetaceae. Genome Biol. Evol. 6:2543–2556.

Hooks KB, Naseeb S, Parker S, Griffiths-Jones S, Delneri D. 2016. Novel intronic RNA structures contribute to maintenance of phenotype in Saccharomyces cerevisiae. Genetics 203:1469–1481.

Hoshida H, Kondo M, Kobayashi T, Yarimizu T, Akada R. 2017. 5′-UTR introns enhance protein expression in the yeast Saccharomyces cerevisiae. Appl. Microbiol. Biotechnol. 101:241–251.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, et al. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12:115–121.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635–1638.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44:D286–D293.

Huff JT, Zilberman D, Roy SW. 2016. Mechanism for DNA transposons to generate introns on genomic scales. Nature 538:533–536.

Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. Trends Genet. 23:321–325.

Irimia M, Roy SW. 2014. Origin of spliceosomal introns and alternative splicing. Cold Spring Harb. Perspect. Biol. 6:a016071.

Jeffares DC, Tomiczek B, Sojo V, dos Reis M. 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. Methods Mol. Biol. 1201:65–90.

Jo B-S, Choi SS. 2015. Introns: The functional benefits of introns in genomes. Genomics Inform. 13:112–118.

Johnson LS, Eddy SR, Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431.

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

Kämper J, Kahmann R, Bölker M, Ma L-J, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O, et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. Nature 444:97–101.

Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, Akiyoshi DE, Tzipori S. 2010. The reduced genome of the parasitic microsporidian Enterocytozoon bieneusi lacks genes for
core carbon metabolism. *Genome Biol. Evol.* 2:304–309.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617–624.

Kempken F. 2013. Alternative splicing in ascomycetes. *Appl. Microbiol. Biotechnol.* 97:4235–4241.

Kijpornyongpan T, Mondo SJ, Barry K, Sandor L, Lee J, Lipzen A, Pangilinan J, LaButti K, Hainaut M, Henrisat B, et al. 2018. Broad genomic sampling reveals a smut pathogenic ancestry of the fungal clade *Ustilaginomycotina*. *Mol. Biol. Evol.* 35:1840–1854.

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapoijnkov V, Smith RG, Tatusova T, Xiang C, Zherikov A, et al. 2016. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44:D73–D80.

Koonin EV, Csuros M, Rogozin IB. 2013. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *Wiley Interdiscip. Rev. RNA* 4:93–105.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34:1812–1819.

Laxa M. 2016. Intron-mediated enhancement: A tool for heterologous gene expression in plants? *Front. Plant Sci.* 7:1977.

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12:124.

Lee RCH, Gill EE, Roy SW, Fast NM. 2010. Constrained intron structures in a microsporidian. *Mol. Biol. Evol.* 27:1979–1982.

Le Hir H, Saulière J, Wang Z. 2016. The exon junction complex as a node of post-transcriptional networks. *Nat. Rev. Mol. Cell Biol.* 17:41–54.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930.

Lim CS, T. Wardell SJ, Keffmann T, Brown CM. 2018. The exon–intron gene structure upstream of the initiation codon predicts translation efficiency. *Nucleic Acids Res.* 46:4575–4591.

Lock A, Rutherford K, Harris MA, Wood V. 2018. PomBase: The scientific resource for fission yeast. *Methods Mol. Biol.* 1757:49–68.

Logsdon JM Jr, Tyschenko MG, Dixon C, D-Jafari J, Walker VK, Palmer JD. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci. U. S. A.* 92:8507–8511.

Loh Y-H, Brenner S, Venkatesh B. 2008. Investigation of loss and gain of introns in the compact genomes of pufferfishes (*Fugu* and *Tetraodon*). *Mol. Biol. Evol.* 25:526–535.

Lowe TM, Eddy SR. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science*
Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320:1632–1635.

Luo Y, Li S. 2007. Genome-wide analyses of retrogenes derived from the human box H/ACA snoRNAs. *Nucleic Acids Res.* 35:559–571.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–1404.

Mekouar M, Blanc-Lenfle I, Ozanne C, Da Silva C, Cruaud C, Wincker P, Gaillardin C, Neuvéglise C. 2010. Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. *Genome Biol.* 11:R65.

Mikhailov KV, Simdyanov TG, Aleoshin VV. 2017. Genomic survey of a hyperparasitic microsporidian *Amphialmblys* sp. (Metchnikovellidae). *Genome Biol. Evol.* 9:454–467.

Moabbi AM, Agarwal N, El Kaderi B, Ansari A. 2012. Role for gene looping in intron-mediated enhancement of transcription. *Proc. Natl. Acad. Sci. U. S. A.* 109:8505–8510.

Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, Haridas S, Kuo A, Salamov A, Ahrendt SR, et al. 2017. Widespread adenine N6-methylation of active genes in fungi. *Nat. Genet.* 49:964–968.

Mourier T, Jeffares DC. 2003. Eukaryotic intron loss. *Science* 300:1393.

Muzzey D, Sherlock G, Weissman JS. 2014. Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. *Genome Res.* 24:963–973.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933–2935.

Ndikumana S, Pelin A, Williot A, Sanders JL, Kent M, Corradi N. 2017. Genome analysis of *Pseudoloma Neurophilia*: A microsporidian parasite of zebrafish (*Danio Rerio*). *J. Eukaryot. Microbiol.* 64:18–30.

Neuvéglise C, Marck C, Gaillardin C. 2011. The intronome of budding yeasts. *C. R. Biol.* 334:662–670.

Nguyen TA, Cissé OH, Yun Wong J, Zheng P, Hewitt D, Nowrousian M, Stajich JE, Jedd G. 2017. Innovation and constraint leading to complex multicellularity in the Ascomycota. *Nat. Commun.* 8:14444.

Niu D-K. 2007. Protecting exons from deleterious R-loops: a potential advantage of having introns. *Biol. Direct* 2:11.

Niu D-K. 2008. Exon definition as a potential negative force against intron losses in evolution. *Biol. Direct* 3:46.

Niu D-K, Yang Y-F. 2011. Why eukaryotic cells use introns to enhance gene expression: Splicing
reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. *Biol. Direct* 6:24.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2018. caper: Comparative Analyses of Phylogenetics and Evolution in R. Available from: https://cran.r-project.org/web/packages/caper/index.html

Parenteau J, Abou Elela S. 2019. Introns: Good Day Junk Is Bad Day Treasure. *Trends Genet.* 35:923–934.

Parenteau J, Durand M, Morin G, Gagnon J, Lucier J-F, Wellingier RJ, Chabot B, Elela SA. 2011. Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* 147:320–331.

Parenteau J, Durand M, Véronneau S, Lacombe A-A, Morin G, Guérin V, Cecez B, Gervais-Bird J, Koh C-S, Brunelle D, et al. 2008. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol. Biol. Cell* 19:1932–1941.

Parenteau J, Maignon L, Berthoumieux M, Catala M, Gagnon V, Abou Elela S. 2019. Introns are mediators of cell response to starvation. *Nature* 565:612–617.

Patthy L. 2003. Modular assembly of genes and the evolution of new functions. *Genetica* 118:217–231.

Peterson RA. 2018. Estimating normalization transformations with bestNormalize. Available from: https://github.com/petersonR/bestNormalize

Petibon C, Parenteau J, Catala M, Elela SA. 2016. Introns regulate the production of ribosomal proteins by modulating splicing of duplicated ribosomal protein genes. *Nucleic Acids Res.* 44:3878–3891.

Peyretaillade E, Parisot N, Polonais V, Terrat S, Denonfoux J, Dugat-Bony E, Wawrzyniak I, Biderre-Petit C, Mahul A, Rimour S, et al. 2012. Annotation of microsporidian genomes using transcriptional signals. *Nat. Commun.* 3:1137.

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol. Cell* 27:928–937.

Priest SJ, Yadav V, Heitman J. 2020. Advances in understanding the evolution of fungal genome architecture. *F1000Res.* 9:776.

R Core Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org/

Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, Salamov AA, Wisecaver JH, Long TM, Calvey CH, et al. 2016. Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci. U. S. A.* 113:9882–9887.

Rose AB. 2018. Introns as Gene Regulators: A Brick on the Accelerator. *Front. Genet.* 9:672.
Roy SW. 2004. The origin of recent introns: transposons? *Genome Biol.* 5:251.

Roy SW, Gilbert W. 2005. The pattern of intron loss. *Proc. Natl. Acad. Sci. U. S. A.* 102:713–718.

Rush TA, Aime MC. 2013. The genus *Meira*: phylogenetic placement and description of a new species. *Antonie Van Leeuwenhoek* 103:1097–1106.

Russell AG, Shutt TE, Watkins RF, Gray MW. 2005. An ancient spliceosomal intron in the ribosomal protein L7a gene (*Giardia lamblia*). *BMC Evol. Biol.* 5:45.

Sapountzis P, Gruntjes T, Otani S, Estevez J, da Costa RR, Plunkett G 3rd, Perna NT, Poulsen M. 2015. The enterobacterium *Trabulsiella odontotermitis* presents novel adaptations related to its association with fungus-growing termites. *Appl. Environ. Microbiol.* 81:6577–6588.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33:W686–W689.

Schmitz J, Zemann A, Churakov G, Kuhl H, Grützner F, Reinhardt R, Brosius J. 2008. Retroposed SNOFall--a mammalian-wide comparison of platypus snoRNAs. *Genome Res.* 18:1005–1010.

Sêton Bocco S, Csûrös M. 2016. Splice sites seldom slide: Intron evolution in oomycetes. *Genome Biol. Evol.* 8:2340–2350.

Shaul O. 2017. How introns enhance gene expression. *Int. J. Biochem. Cell Biol.* 91:145–155.

Shi Y. 2017. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat. Rev. Mol. Cell Biol.* 18:655–670.

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 27:135–145.

Skelly DA, Ronald J, Connelly CF, Akey JM. 2009. Population genomics of intron splicing in 38 *Saccharomyces cerevisiae* genome sequences. *Genome Biol. Evol.* 1:466–478.

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

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, et al. 2008. The *Trichoplax* genome and the nature of placozoans. *Nature* 454:955–960.

Stajich JE, Dietrich FS, Roy SW. 2007. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol.* 8:R223.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stamstrup J. 2017. massageR: Functions for data massaging in R. Available from: https://gitlab.com/R_packages/massageR

Stoltzfus A. 2004. Molecular evolution: Introns fall into place. *Curr. Biol.* 14:R351–R352.
Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* 508:66–71.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2005. Conservation versus parallel gains in intron evolution. *Nucleic Acids Res.* 33:1741–1748.

Tian L. 2018. SkewGD: Summation of Ks pairs for Exploration of Whole Genome Duplications. Available from: https://github.com/LongTianPy/SkewGD_v1

Tonegawa S, Maxam AM, Tizard R, Bernard O, Gilbert W. 1978. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. *Proc. Natl. Acad. Sci. U. S. A.* 75:1485–1489.

Uhse S, Pflug FG, Stirnberg A, Ehrlinger K, von Haeseler A, Djamei A. 2018. In vivo insertion pool sequencing identifies virulence factors in a complex fungal-host interaction. *PLoS Biol.* 16:e2005129.

Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* 30:177–190.

Wang B, Liang X, Gleason ML, Hsiang T, Zhang R, Sun G. 2020. A chromosome-scale assembly of the smallest Dothideomycete genome reveals a unique genome compaction mechanism in filamentous fungi. *BMC Genomics* 21:321.

Wang Q-M, Begerow D, Groenewald M, Liu X-Z, Theelen B, Bai F-Y, Boekhout T. 2015. Multigene phylogeny and taxonomic revision of yeasts and related fungi in the *Ustilaginoidea*. *Stud. Mycol.* 81:55–83.

Weber MJ. 2006. Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.* 2:e205.

Wiegand HL, Lu S, Cullen BR. 2003. Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. *Proc. Natl. Acad. Sci. U. S. A.* 100:11327–11332.

Witten JT, Ule J. 2011. Understanding splicing regulation through RNA splicing maps. *Trends Genet.* 27:89–97.

Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E,
Everett MV, et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science 324:268–272.

Wu B, Macielog AI, Hao W. 2017. Origin and spread of spliceosomal introns: Insights from the fungal clade Zymoseptoria. Genome Biol. Evol. 9:2658–2667.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yarza P, Yilmaz P, Panzer K, Glöckner FO, Reich M. 2017. A phylogenetic framework for the kingdom Fungi based on 18S rRNA gene sequences. Mar. Genomics 36:33–39.

Yenerall P, Krupa B, Zhou L. 2011. Mechanisms of intron gain and loss in Drosophila. BMC Evol. Biol. 11:364.

Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y. 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol. Cell 59:744–754.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. McInerny G, editor. Methods Ecol. Evol. 8:28–36.

Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16:284–287.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. Nucleic Acids Res. 46:D754–D761.

Zhang L-Y, Yang Y-F, Niu D-K. 2010. Evaluation of models of the mechanisms underlying intron loss and gain in Aspergillus fungi. J. Mol. Evol. 71:364–373.

Zhang Y, Sachs MS. 2015. Control of mRNA stability in fungi by NMD, EJC and CBC factors through 3'UTR introns. Genetics 200:1133–1148.

Zhu Q. 2014. AfterPhylo: a Perl script for manipulating trees after phylogenetic reconstruction. Available from: https://github.com/qiyunzhu/AfterPhylo

Zhu T, Niu D-K. 2013a. Frequency of intron loss correlates with processed pseudogene abundance: a novel strategy to test the reverse transcriptase model of intron loss. BMC Biol. 11:23.

Zhu T, Niu D-K. 2013b. Mechanisms of intron loss and gain in the fission yeast Schizosaccharomyces. PLoS One 8:e61683.
Table 1. Intron densities of the ancestral and current states of fungal clades. See also Figure 1 and 2.

| Clade                  | Number of Species | Ancestral state<sup>a</sup> | Current state<sup>b</sup> |
|------------------------|-------------------|-----------------------------|---------------------------|
|                        |                   | Ancestral state<sup>a</sup> | Current state<sup>b</sup> | Mean  | Median |
| Cryptomycota           | 1                 | NA                          |                           | 2.70  | 2.70   |
| (Rozella allomyces)    |                   |                             |                           |       |        |
| Microsporidia          | 15                | 0.06                        | 0.01                      | 0.00  |        |
| Chytridiomycota        | 2                 | 5.92                        | 4.25                      | 4.25  |        |
| Blastocladiomycota     | 1                 | NA                          | 1.43                      | 1.43  |        |
| (Allomyces macrogynus) |                   |                             |                           |       |        |
| Entomophthoromycotina  | 1                 | NA                          | 1.70                      | 1.70  |        |
| (Conidiobolus coronatus)|                 |                             |                           |       |        |
| Mucaromycota           | 5                 | 6.05                        | 2.67                      | 2.20  |        |
| Pucciniomycotina       | 8                 | 4.20                        | 3.09                      | 3.03  |        |
| Ustilaginomycotina     | 20                | 2.93                        | 0.56                      | 0.16  |        |
| Agaricomycotina        | 39                | 4.67                        | 3.57                      | 3.79  |        |
| Taphrinomycotina       | 13                | 4.51                        | 1.45                      | 1.24  |        |
| Saccharomycotina       | 36                | 1.68                        | 0.05                      | 0.01  |        |
| Pezizomycotina         | 122               | 1.06                        | 0.52                      | 0.52  |        |

<sup>a</sup> Obtained from the inference of intron gain/loss.

<sup>b</sup> Arithmetic mean or median inferred introns/kbp of the species within a clade.

Introns/kbp, the number of introns per one kbp of protein-coding sequence; NA, not applicable.
Figure 1. Widespread loss of introns during the evolution of *Fungi*. Ancestral introns were inferred from 1445 orthologs in 263 fungal species using Malin (Csuros, 2008), a Markov model with rates across sites and branch-specific gain and loss rates. Branches are color-coded with intron densities from the median posterior distribution for each node. A list of full names and intron densities are available in Supplementary Table S2. See also related Table 1 and Figure 2. Introns/kbp, the number of introns per one kbp of protein-coding sequence. Green filled circles denote eight intron poor species selected for additional analysis.
Figure 2. Intron densities of the fungal ancestral states derived from a Monte-Carlo approximation of 100 bootstrap distributions. Dotted lines denote the ancestral intron densities inferred from Dollo parsimony (blue) and maximum likelihood (ML, green) models. See also related Figure 1 and Table 1. Introns/kbp, the number of introns per one kbp of protein-coding sequence.
**Figure 3.** Intron density weakly correlates with genome size. Phylogenetic independent contrasts analysis of genome size versus intron density and the number of protein-coding genes. CDS, coding sequence; introns/kbp, the number of introns per one kbp of protein-coding sequence; $r_s$, Spearman’s rho.

**Figure 4.** Features of intron-containing genes in intron-poor fungal species. Proportion of intronless and intron-containing genes that have undergone (A) positive selection and (B) gene duplication. $^*$, $p < 0.05$ (Fisher’s exact test, adjusted using the Bonferroni correction); Cdb, *Candida dubliniensis*; Ceur, *Cyphellophora europaea*; Mmil, *Meira miltonrushii*; Msym, *Malassezia sympodiali*; Phub, *Pseudozyma hubeiensis*; Scer, *Saccharomyces cerevisiae*; Ssch, *Sporothrix schenckii*; Umay, *Ustilago maydis*. 
**Figure 5. Orthologous genes harbor introns concordantly.** (A) Schematic example of a pairwise comparison of intron-containing orthologs among three species. (B) The ratios of intron-containing orthologs in a pairwise comparison in contrast to null expectations (solid horizontal colored lines). Non-significant chi-square tests (dotted circles, Bonferroni adjusted $p$-value > 0.01) suggest that introns can be retained in any genes. As a result, 34 of 56 comparisons between species are statistically significant. The binomial confidence intervals (95%) were estimated from these ratios using Bayesian inference with 1000 iterations (vertical colored lines). Cdub, *Candida dubliniensis*; Ceur, *Cyphellophora europaea*; Mmil, *Meira miltonrushii*; Msym, *Malassezia sympodialis*; Phub, *Pseudozyma hubeiensis*; Scer, *Saccharomyces cerevisiae*; Ssch, *Sporothrix schenckii*; Umay, *Ustilago maydis*. 
Figure 6. Introns of RPL7B and NOG2 have conserved positions. The introns of RPL7B and NOG2 encode box C/D and box H/ACA snoRNAs (snR59 and snR191 in S. cerevisiae, respectively). The predictions of stem-loop 2 and antisense element (ASE) of the M. miltonrushi box H/ACA snoRNA are of low confidence. 5′ SS and 3′ SS denote 5′ and 3′ splice-sites, respectively. Cdb, Candida dubliniensis; Ceur, Cyphellophora europaea; Mmil, Meira miltonrushi; Msym, Malassezia sympodiali; Phub, Pseudozyma hubeiensis; Scer, Saccharomyces cerevisiae; Ssch, Sporothrix schenckii; Umay, Ustilago maydis.

Figure 7. First introns are located near the translation initiation codons. Plus and minus signs denote translation-associated genes and other classes of genes, respectively. Gene counts are shown in the right panel. Dotted lines in light gray denote a null distribution of the first CDS introns of S. cerevisiae. Dashed lines in dark gray denote a null distribution of the actual first introns (including UTR introns) of S. cerevisiae centered at the transcription start or termination sites. CDS, coding...
sequence; Cdb, Candida dubliniensis; Ceur, Cyphellophora europaea; GO, gene ontology; Mmil, Meira miltonrushii; Msym, Malassezia symmodiali; Phub, Pseudozyma hubeiensis; Scer, Saccharomyces cerevisiae; Ssch, Sporothrix schenckii; Umay, Ustilago maydis; UTR, untranslated regions.

**Figure 8.** Introns are more abundant in specific classes of genes. Cdb, Candida dubliniensis; Ceur, Cyphellophora europaea; Mmil, Meira miltonrushii; Msym, Malassezia symmodiali; Phub, Pseudozyma hubeiensis; Scer, Saccharomyces cerevisiae; Ssch, Sporothrix schenckii; Umay, Ustilago maydis.

**Figure 9.** Intron-containing genes have higher mRNA expression and translation efficiency. (A) RNA-seq and (B) ribosome profiling results of *S. cerevisiae* orthologs. Translation efficiency was determined by the ratio of ribosome-protected fragments and mRNA read counts normalized to respective library sizes. *C. albicans*, Candida albicans; *N. crassa*, Neurospora crassa; *p*, the p-values of Welch two sample t-tests were adjusted with Bonferroni correction; RPKM, Reads Per Kilobase per Million mapped reads; *S. cerevisiae*, Saccharomyces cerevisiae; *S. pombe*, Schizosaccharomyces pombe.