UNIVERSITY OF CALIFORNIA, MERCED

Minor Spliceosomal Introns: What, How and Where

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Quantitative and Systems Biology

by

Graham E. Larue

Committee in charge:

Professor David Ardell, Chair
Professor Gorden Bennett
Professor Steven E. Brenner
Professor Scott W. Roy

2022

The dissertation of Graham E. Larue is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Gordon Bennett, Ph.D.

Steven E. Brenner, Ph.D.

Scott W. Roy, Ph.D., Advisor

David Ardell, Ph.D., Chair

University of California, Merced

2022

DEDICATION

This dissertation is dedicated to my patient and supportive partner, Morgan, and our two young daughters Acacia and Hazel—the former born during my PhD, the latter born during both a global pandemic *and* my PhD (arguably separate phenomena).

EPIGRAPH

When I was young, I admired clever people.
Now that I am old, I admire kind people
—Abraham Joshua Heschel

TABLE OF CONTENTS

LIST OF FIGURES

Figure 3.1: (Continued from previous page) (B) Presence of BPS motif in various groups of *P. polycephalum* introns. (Main) Occurrence of TTTGA motif as a function of number of nucleotides upstream of the 3′SS), for U12-like ([AG]TATCCTT-A[CG] or [AG]TATCTTT-A[CG] splice sites for "U12-like" and "U12-like +6T", respectively), U2-like (GTNNG-AG), and conserved U12-like ([AG]TATC[CT]-NN and conserved as a U12-type intron in another species). (Inset) The same data as a cumulative bar plot for positions -14 through -8. See also Figure 3.4B. (C) Intron type classification and associated motifs. The main plot shows BPS-vs-5′SS log-ratio z-scores for all *P. polycephalum* introns, with conserved U12-type introns highlighted in blue. The dashed green line indicates the approximate U2-U12 score boundary (section 3.9, see also Figure 3.4C). Below the scatter plot are sequence logos showing motif differences between the two groups (top, U12-type, n = 20,899; bottom, U2-type, n = 154,299). (D) Conservation status of *P. polycephalum* introns in other species, showing substantially lower U12- than U2-type conservation. For each species, the pair of bars shows the fractions of *P. polycephalum* introns of each intron type (U12-type, un-hashed; U2-type, hashed) that are conserved as either U12-type (red) or U2-type (yellow) introns, or not conserved (gray). Total numbers of *P. polycephalum* introns assessed are given at right. (E) Comparison of U12-type intron density (fraction of introns that are U12-type) in genes of different age categories for *P. polycephalum* (PhyPol), *Homo sapiens* (HomSap) and *Arabidopsis thaliana* (AraTha), relative to expectation (blue/red = below/above expectation). U12-type intron densities in *P. polycephalum* are significantly overrepresented in newer genes, in contrast to the pattern seen in both human and *Arabidopsis*. Significance assessed by Fisher's exact tests corrected for multiple testing using the Holm step-down method. Full species names listed in Table S1, `https://doi.org/10.6084/m9.figshare.20483790`. 48

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Scott Roy for his (many) years of scientific guidance, support and camaraderie. In addition, I am grateful to all members of my thesis committee for their support, feedback and advice during my time in the graduate program (and in particular to Steven Brenner, who has sat on committees for both of my advanced degrees). I should also take this opportunity to thank both my wonderful parents, who have been instrumental in so many aspect of my life, along with all the members of the Roy lab, past and present, and the other friends who have stuck with me along the way. I would be remiss if I didn't also give special mention to my good friend Jacob Stanley, an accomplished scientist whom I greatly admire and who probably understands my work better than anyone else save my advisor. He has been an invaluable sounding board for issues of all sorts throughout my graduate career, and I'm very glad to have him in my corner.

Much of the work presented herein was produced in collaboration with other talented and thoughtful scientists with whom I was lucky enough to get to work. By chapter, they are:

- *Chapter 2*

  Devlin C. Moyer (co-first author in first position) Devlin was responsible for writing a large portion of the draft manuscript, and was entirely responsible for establishing the database backend for the IAOD. My involvement in the project was initially intended to be mostly in the production of intron classification data, but over time various technical challenges developed and I became more involved with generation of additional data and analyses, to the point that we decided I should be made co-first author along with Devlin.

  Courtney E. Hershberger

  Richard A. Padgett

- *Chapter 3*

Marek Eliáš

The work in this dissertation was supported in part by the National Science Foundation, award numbers 1616878 and 1751372 to Scott W. Roy.

VITA

2008	B. S. in Biology, University of California Santa Cruz
2015	M. S. in Molecular and Cell Biology, San Francisco State University
2022	Ph. D. in Quantitative and Systems Biology, University of California Merced

PUBLICATIONS

Larue, G. E., Eliáš, M., & Roy, S. W. (2021). Expansion and transformation of the minor spliceosomal system in the slime mold *Physarum polycephalum*. *Current Biology: CB, 31*(14), 3125–3131.e4. https://doi.org/10.1016/j.cub.2021.04.050

Moyer, D. C.[†], **Larue, G. E.**[†], Hershberger, C. E., Roy, S. W., & Padgett, R. A. (2020). Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Research, 48(13), 7066–7078. https://doi.org/10.1093/nar/gkaa464

Slabodnick, M. M., Ruby, J. G., Reiff, S. B., Swart, E. C., Gosai, S., Prabakaran, S., Witkowska, E., **Larue, G. E.**, Fisher, S., Freeman, R. M., Jr, Gunawardena, J., Chu, W., Stover, N. A., Gregory, B. D., Nowacki, M., Derisi, J., Roy, S. W., Marshall, W. F., & Sood, P. (2017). The Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell. *Current Biology: CB, 27*(4), 569–575. https://doi.org/10.1016/j.cub.2016.12.057

[†]co-first authors

ABSTRACT OF THE DISSERTATION

Minor Spliceosomal Introns: What, How and Where

by

Graham E. Larue
Doctor of Philosophy in Quantitative and Systems Biology
University of California Merced, 2022
Professor David Ardell, Chair

Spliceosomal introns, sequences interspersed throughout genes that are removed during mRNA production by machinery called the spliceosome, are a hallmark feature of eukaryotic genomes and gene structure. They are ancient genomic elements which can vary widely in number and size between species, and have remained a puzzling aspect of genome evolution in the decades since they were first discovered. To make the situation more complicated, there are in fact two separate spliceosomal systems, termed major (or U2) and minor (U12), which are responsible for the removal of $\geq 99.5\%$ and less than half a percent of introns in most genomes, respectively. Minor introns in particular display a puzzling evolutionary pattern: in many cases, they are maintained with high degrees of conservation between deeply-diverged species, yet in others are either mostly or entirely missing. A large amount of the foundational work on minor introns was completed before the advent of next-generation sequencing, and almost all of the existing comparative genomics work in the field involves fewer than a dozen or so model organisms. In addition, the primary resources used by the field to identify minor introns are static databases containing information on a limited number of species. This dissertation contributes to our understanding of minor introns over the course of four distinct but related projects: First, it describes an effective and accessible method for identifying minor introns in intron sequence data, as well as the creation of a database containing the largest collection of minor intron orthology data available. Second,

this dissertation highlights an extraordinary case of minor intron enrichment in a slime mold genome, reshaping our understanding of the limits of minor intron evolution. Third, it contains a sweeping analysis of minor intron diversity across more than 3000 eukaryotic genomes, uncovering a number of novel aspects of minor intron evolution and providing a rich substrate for future work. Finally, this dissertation describes a novel finding of a fungal genome with significant numbers of minor introns, and uses it to attempt to clarify a set of longstanding theories about the role of minor introns in eukaryotic evolution.

Chapter 1

Introduction to the Dissertation

1.1 Background on introns

One of the ways that eukaryotic gene structure differs from that of prokaryotes (like bacteria) is the presence of spliceosomal introns (hereafter just "introns")—stretches of genomic sequence that interrupt the portions of genes that ultimately constitute the mature messenger RNA during gene expression (Berget, C. Moore, and Sharp, 1977; Chow et al., 1977; Knapp et al., 1978). These sequences complicate the production of downstream gene products because they must be removed, or "spliced" before the mRNA can exit the nucleus and go on to make, say, a protein. The discovery of introns in eukaryotic organisms fundamentally altered our understanding of eukaryotic cell biology, and since that time an enormous amount of work has gone into understanding the manifold cellular processes involved in spliceosome assembly and gene splicing (Konarska et al., 1985; Gilbert, 1978; Will, Lührmann, and Luhrmann, 2011; X. Zhang et al., 2017). Other groups have focused more on the evolutionary aspect of introns, and there are plenty of interesting questions pertaining to introns in that domain. Introns are peculiar in their evolutionary dynamics, and display an overwhelming variety in many different aspects of their biology, from variation in size (the introns of the ciliate protist *Stentor coeruleus* are almost all 15-16 bp (Slabodnick et al., 2017), whereas there are introns in the human genome that are millions of bp long) to variation in number (there are hundreds of thousands of introns in vertebrate

genomes, but only ∼300 in brewer's yeast) and average density within individual genes (Scott William Roy and Walter Gilbert, 2006).

Well before I began working in the field, there was a relatively infamous debate regarding how long introns had been in the genome. The two primary hypotheses were that a) introns had been widespread in the genome before the ancestor of eukaryotes and were in fact instrumental in the formation of proteins ("introns-early") (Gilbert, 1987; Gilbert, 1978; Gilbert, Souza, and Long, 1997), or b) that they were a more recent (though still ancient by any reasonable standard) eukaryotic innovation, having forced existing genes into pieces rather than having helped create them to begin with ("introns-late") (Doolittle and Stoltzfus, 1993; Stoltzfus et al., 1994; Stoltzfus, 1994; Arlin Stoltzfus et al., 1997; Logsdon, 1998). Eventually the dust more or less settled and something of a consensus formed, incorporating bits of both theories and supporting the idea that introns were present in the eukaryotic ancestor at densities comparable to those of extant species (Koonin, 2006; Koonin, Csuros, and Rogozin, 2013; Scott W Roy and Walter Gilbert, 2005a; Irimia and Scott William Roy, 2014), and predated that ancestor by some time but at perhaps significantly lower densities (Koonin, 2006; Koonin, Csuros, and Rogozin, 2013; Scott William Roy and Walter Gilbert, 2006; Rogozin et al., 2012). It is now largely accepted that aside from a small number of possible concerted intron gain events at the roots of a number of deep eukaryotic nodes (e.g., the animal ancestor), the dominant modality driving intron evolution is loss (Scott William Roy and Penny, 2007b; Scott William Roy and Penny, 2006; Scott W Roy, Fedorov, and Walter Gilbert, 2003; Mourier and Jeffares, 2003; Rogozin et al., 2012). In general, intron loss appears to occur primarily via removal of the sequence itself, either by direct genomic deletion (via a potentially large number of specific mechanisms or recombination with a reverse-transcriptase product of spliced mRNA (Ma et al., 2022; Cohen, Shen, and Carmel, 2012; K. Lin and D.-Y. Zhang, 2005; Scott W Roy and Walter Gilbert, 2005b; Scott William Roy and Walter Gilbert, 2006; Jeffares, Mourier, and Penny, 2006). The characterization of intron gain remains ongoing, with proposed mechanisms including transposable element insertion, reverse-transcriptase mediated processes and the "intronization"

of exonic sequence among many others (Sharpton et al., 2008; Yenerall and Zhou, 2012; Fedorov, S. Roy, et al., 2003; Larue, Eliáš, and Scott W Roy, 2021; Scott William Roy, Gozashti, et al., 2020; Scott William Roy and Penny, 2007a; Scott William Roy, 2016; Scott W Roy, 2004; Jeffares, Mourier, and Penny, 2006; Huff, Zilberman, and Scott W Roy, 2016).

Roughly a decade after introns were first characterized in eukaryotes, a new type of intron was discovered that looked very distinct—most introns described up until that point had started with the dinucleotide pair "GT" and had ended with "AG", whereas this new type of intron had AT-AC termini and a longer, less variable motif at the $5'$ end, which matched sequences in previously-described but poorly understood small nuclear ribonucleoproteins (snRNPs) (Montzka and J A Steitz, 1988; Jackson, 1991; S. L. Hall and R A Padgett, 1994). Eventually, it was shown that an entirely separate (but functionally analogous) splicing machinery was responsible for the removal of these introns, which was termed the "minor" spliceosome and the introns, minor introns (with the "regular" introns being called major introns) (Patel and Joan A Steitz, 2003). Both spliceosomes are evolutionarily ancient, appearing to date back to at least the last eukaryotic common ancestor (Russell et al., 2006), and while major intron dynamics are themselves quite varied, minor intron evolution is even more so. While most eukaryotic lineages still maintain at least some major introns, there are many instances throughout the eukaryotic tree where minor introns have been entirely lost, along with the minor splicing components (Turunen, Niemelä, et al., 2013). One intriguing aspect of minor intron evolution specifically is the fact that minor intron losses have happened again and again in many different groups, while other large clades (vertebrates, land plants) have retained relatively high numbers of minor introns for hundreds of millions of years (Rogozin et al., 2012; Moyer et al., 2020). Minor introns have been found in animals, plants and fungi, although published species-specific information is lacking for all but $\sim$30 model organisms (Moyer et al., 2020; Alioto, 2007; Baumgartner, Drake, and Kanadia, 2019).

A persistent question throughout the intron evolution literature is, basically, "Why introns?". This question does not appear to admit of a straightforward

or overarching answer; rather, introns appear to be involved in many different cellular phenomena depending on their local genetic context, including both direct and indirect modulation of gene expression (Sands, Yun, and Mendenhall, 2021; Chung et al., 2006; Shaul, 2017; Sands, Yun, and Mendenhall, 2021; Lu et al., 2008; Abou Alezz et al., 2020; Younis et al., 2013; Castillo-Davis et al., 2002; Rose, 2018; Patel, McCarthy, and Joan A Steitz, 2002; Abou Alezz et al., 2020; Parenteau, Durand, et al., 2008; Parenteau, Maignon, et al., 2019), nonsense-mediated decay (Behringer and D. W. Hall, 2016; Weischenfeldt et al., 2012; Hirose, Shu, and J A Steitz, 2004; Lewis, Green, and Brenner, 2003) and intron retention (Middleton et al., 2017; Wong and Schmitz, 2022; Inoue et al., 2021; Monteuuis et al., 2019; Buckley et al., 2011; Niemelä, Oghabian, et al., 2014; Braunschweig et al., 2014; Li, Xiao, and Y. X. Zhu, 2014; Mao et al., 2014; Gault et al., 2017). More recently, minor introns in particular have been implicated in the regulation of cell-cycle progression (Gault et al., 2017; Doggett et al., 2018; König et al., 2007; Meinke et al., 2020; Bai et al., 2019) and the development of certain disease states in both animals and plants (Richard A Padgett, 2012; Inoue et al., 2021; Levesque, Salazar, and Scott William Roy, 2022; Lotti et al., 2012; Edery et al., 2011; Olthof, Hyatt, and Kanadia, 2019; Doggett et al., 2018; Gault et al., 2017). At a high level, whatever the original factors may have been that drove the spread of introns throughout the early genome, their prevalence in eukaryotic genes has allowed them to act as substrate for an increase in genomic plasticity (via, for example, alternative splicing) and the evolution of various networks of regulation and gene expression within the genomes of extant species (Jo and Choi, 2015; Chorev and Carmel, 2012; I V Poverennaya and Roytberg, 2020; Rogozin et al., 2012).

1.2 Research overview

This dissertation comprises four distinct but related projects concerning various aspects of minor intron biology. When I began working on minor introns in 2013, next-generation sequencing had only been around for a few years, and had not yet been integrated into most genome annotation pipelines. At the same time,

there were far fewer complete genome assemblies in need of annotation—according to NCBI statistics, fewer than 200 genomes had been processed by their system when I began reviewing the minor intron literature in search of a research topic. The most cited minor intron resources at that point were the U12DB (Alioto, 2007) and SpliceRack (Sheth et al., 2006), which acted (and often still act) as the gold standard for information on minor introns. However, neither resource provides a mechanism for annotating *new* minor introns, and over the years the various published analyses of minor introns have usually been accomplished with one-off, group-specific bioinformatic algorithms which were never broadly released (Bartschat and Samuelsson, 2010; C. B. Burge, R A Padgett, and Sharp, 1998; Sheth et al., 2006). Because minor intron evolution does not seem to be driven by a single functional paradigm, a great deal of insight can be gained simply by obtaining a more complete understanding of the overall diversity of minor introns in different lineages.

It was this context that prompted me to begin work on the first project included in my dissertation, the design of a general-purpose pipeline to identify minor introns from primary sequence data. I wrote an initial version of the pipeline during my master's degree, but the first implementation had a number of deficits and was never published. After starting my PhD, I re-wrote the entire pipeline from the ground up, using a fundamentally different machine-learning-based approach to sequence classification, and in collaboration with Rick Padgett's lab at Cleveland Clinic published the classification algorithm `intronIC` (`https://github.com/glarue/intronIC`), and an associated database in *Nucleic Acids Research*[1]. My co-first author on that paper, Devlin Moyer—then an undergraduate student in the Padgett lab—was the architect (in conjunction with Rick) of the general research approach as well as the design and implementation of the IAOD database. Devlin also carried out many of the downstream analyses in the paper (e.g., phase bias, intron length distributions) and wrote a first draft of the manuscript, while I was responsible for generating all of the substrate intron

[1]Moyer, D. C., Larue, G. E., Hershberger, C. E., Roy, S. W. & Padgett, R. A. Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res. 48, 7066–7078 (2020)

classification and orthology data, as well as writing much of the Methods section, revising the draft manuscript (along with the other authors) and responding to reviewer comments. A great deal of work was put into designing `intronIC` to be both extensible and easy to use; it is installable via the Python package manager `pip`, and can be run using only a genome and annotation file, or directly on a plain-text file of intronic sequences. It has been built to allow any future datasets of verified minor introns to be incorporated into its classification models, and there is a good chance that spliceosome profiling data (Burke et al., 2018) for the minor spliceosome will at some point provide an empirically-verified ground-truth set of minor introns that should improve future classification performance. More generally, my experience writing and maintaining an open-source piece of academic software has highlighted the degree to which the field of bioinformatics especially is in need of a better long-term funding model for these sorts of projects; without the ability for the maintainers of the algorithms the field relies upon to be supported in that work, many worthwhile bioinformatics projects will wither on the vine, and (to mix metaphors) the algorithmic wheel will need to be reinvented time and time again.

The next project came about serenditpitously—a protistologist in Czechia, Marek Eliáš, contacted our lab after discovering some unusual non-canonical intron sequences in a number of GTPase genes in the slime mold *Physarum polycephalum*. We began following up on them as possible minor introns based upon early results from our intron classification pipeline, and I began the tedious process of updating the *Physarum* gene annotations using RNA-seq. At the start of the project, I had overlooked the fact that another group had found putative minor introns in the same species a number of years earlier (Bartschat and Samuelsson, 2010), although their dataset was very limited and they were commensurately cautious in their conclusions. Noticing this encouraged us to pursue the project in additional depth; upon first inspection, I found many intron sequences with minor-intron-like sequence motifs, but no annotated AT-AC introns. Based on my experience looking for minor introns across different species, this pattern (many putative minor introns, but no introns with AT-AC termini) is often the result of an annotation

pipeline biased against detection of minor introns, usually due to conservative constraints on intron terminal dinucleotide pairs (often only GT/C-AG, regardless of support for a given splice junction). There was a relatively small amount of publicly-available RNA-seq for *Physarum* at the time, so I used all of it to update the original annotations by a variety of methods (detailed in chapter 3). With regard to minor introns specifically, there is probably still a great deal of low-hanging fruit in terms of species with older annotations that could be updated to dramatically increase the number of annotated minor introns. The number of such species, however, appears to be diminishing with improvements to annotation pipelines used by e.g., NCBI, and as allowances for AT-AC termini are integrated into commonly-used annotation tools such as `AUGUSTUS` Stanke et al., 2008. After updating the annotations, I spent a significant amount of time modifying the `intronIC` algorithm to account for some of the apparently unique sequence features of *Physarum*, like the highly position-specific branch point sequence of its minor introns. Because `intronIC` was not initially written with the facility to provide additional weighting to BPS motifs based on their location, the high degree of positional enrichment in *Physarum* required a fair amount of re-engineering of the plumbing involved in how `intronIC` identifies the highest-scoring BPS subsequence in an intron based on a given motif. While this functionality could have been integrated into the main `intronIC` codebase, given the niche requirements of *Physarum* and my philosophical caution around catering to such edge-cases, I decided to provide the *Physarum*-specific code on Zenodo, and exclude the position-based BPS weighting from the main `intronIC` code branch. Publishing and maintaining open-source academic software is a balancing act between one's interest and one's time, and even in my own narrow field I have frequently suffered the consequences of trying to do too much in too many ways and trying to anticipate the needs of specialized end users. The Unix philosophy of writing software to do one thing, well, is a useful guiding principle in this regard, if not one I have ever managed to fully realize in my own work. Ultimately, my coauthors and I were able to publish[2] a fairly compelling story of massive minor intron gain as well as minor

[2]Larue, G. E., Eliáš, M. & Roy, S. W. Expansion and transformation of the minor spliceosomal system in the slime mold *Physarum polycephalum*. *Curr. Biol.* 31, 3125–3131.e4 (2021)

splicing system changes in *Physarum*, highlighting it as an example of extreme evolution in the minor intron landscape. The initial draft of the manuscript was written by me, and collaboratively refined by myself and Scott, with input from our collaborator Prof. Eliáš.

Over the course of completing these first two projects, I had generated a lot of ancillary minor intron data as a result of various comparative analyses related to the previous projects as well as looking for additional species which might harbor interesting minor intron stories. Given the dearth of published minor intron data in the many more recently-annotated genomes, it seemed like a good opportunity to use the various tools and pipelines I had built up to undertake a large-scale survey of minor intron diversity in as many genomes as possible. Due to the scale of the project and as laid out in Chapter 4, this involved developing a number of new skillsets related to data management, organization and interpretation and resulted in fairly comprehensive overview of minor intron presence/absence/features in thousands of eukaryotic genomes. Part of my motivation in completing this project, in addition to being uniquely well-positioned to do so, was in service of leaving the field with more information of general use than it had when I started my graduate career. I made extensive use of a number of important papers describing early analyses of minor intron diversity, and it felt like a nice capstone on my time in the field to contribute as much data as I could to inform future investigations by those coming after me. The entirety of Chapter 4 was written by me.

The final chapter in this dissertation, Chapter 5, is of much narrower scope than the last, although a great deal of it touches on an important aspect of the current thinking on minor introns. There is growing consensus around the idea that minor introns may have some regulatory role in cell-cycle progression and differentiation (Gault et al., 2017; Doggett et al., 2018; König et al., 2007; Meinke et al., 2020), a finding which has been shown in both animals and plants (Gault et al., 2017; Bai et al., 2019). The fact that minor splicing appears to have this association in both lineages, in addition to the functional enrichment of "information processing" genes in minor intron-containing gene (MIG) sets, raises the intriguing possibility that

this role may be ancestral. In order to be more confident in this result, however, additional lineages (aside from animals and plants) would ideally be examined, but this hasn't been possible until now because there were no additional lineages known to contain sufficient numbers of minor introns. In this chapter, we make use of our discovery of significant enrichment of minor introns in the Glomeromycete fungus *Rhizophagus irregularis* to gain further insight into the association between minor introns/splicing and cell cycle regulation/cellular proliferation. Using previously-published data on proliferation-associated genes (Sandberg et al., 2008), cell-type specific sequencing in *Rhizophagus* (Kameoka et al., 2019) as well as our data on MIGs, we were able to characterize the relationship between minor splicing and proliferating cells in an additional major lineage of eukaryotes. We do not find a strong association between minor splicing and proliferation in *Rhizophagus*, in contrast to previous results in animals and plants, weakening (or at least, suggesting more investigation is needed into) the hypothesis that the association found in other lineages is indicative of an ancestral role for minor splicing in cell-cycle regulation. Furthermore, we show that the enrichment of "information processing" genes in minor introns, found in GO analyses by various groups for the last two decades, may largely be explained by the old age of MIGs generally, rather than something particular to the biology of MIGs and minor splicing. Scott wrote the initial outline of this chapter (as a manuscript), I wrote Section 5.6, Section 5.5 was written by Scott, and the remainder was written/modified collaboratively as I completed analyses and enumerated the methods.

The work I have completed over the course of my PhD has contributed to the field of minor intron evolution in a number of concrete and, hopefully, important ways. First, it has produced a novel, publicly-available and open-source tool for identifying minor introns (`https://github.com/glarue/intronIC`) that requires minimal domain expertise from the end user (Chapter 2). My hope is that this tool will facilitate a broader exploration of minor intron diversity in current and yet-to-be sequenced genomes, and provide an accessible and reproducible method for producing and documenting these findings. Second, my work has described the only known case of significant minor intron gain in any species, a result which

suggests that minor intron evolution can be more dynamic in more ways than were previously appreciated and establishes a reference point for our current understanding of the extreme bounds of minor intron splicing. Finally, it has produced the largest compendium of minor intron data so far available, providing fertile ground for future related investigations. Additionally, this substantial increase in available data has highlighted many underexamined aspects of minor intron dynamics in various lineages, and has called into question a number of longstanding results related to minor intron conservation, their intra-gene distribution (Chapter 4) and the functional biases of minor intron-containing genes (Chapter 5).

1.3 Dissertation template and formatting

This dissertation was prepared using LaTeX, modified from a template from the University of California San Diego. At the time of writing, the template was available on Overleaf at `https://www.overleaf.com/latex/templates/university-of-california-san-diego-ucsd-thesis-template/jkzzvcrbssfg`. Modifications for the University of California Merced included the addition of committee member names to the signature page, and packages for equation formatting, cross-referencing and other formatting helpers including:

```
cleveref
hyperref
seqsplit
caption
subcaption
amsmath
siunitx
array
booktabs
tabularx
```

Chapter 2

Comprehensive Database and Evolutionary Dynamics of U12-Type Introns

2.1 Prior publication note

A version of this chapter of the dissertation has been published in *Nucleic Acids Research*:

> Moyer, D. C., Larue, G. E., Hershberger, C. E., Roy, S. W. & Padgett, R. A. Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res. 48, 7066–7078 (2020)

2.2 Abstract

During nuclear maturation of most eukaryotic pre-messenger RNAs and long non-coding RNAs, introns are removed through the process of RNA splicing. Different classes of introns are excised by the U2-type or the U12-type spliceosomes, large complexes of small nuclear ribonucleoprotein particles and associated proteins. We created `intronIC`, a program for assigning intron class to all introns in a given genome, and used it on 24 eukaryotic genomes to create the Intron Annotation and Orthology Database (IAOD). We then used the data in the IAOD

to revisit several hypotheses concerning the evolution of the two classes of spliceosomal introns, finding support for the class conversion model explaining the low abundance of U12-type introns in modern genomes.

2.3 Introduction

The process of RNA splicing is a necessary step in the maturation of nearly all eukaryotic pre-messenger RNAs and many long non-coding RNAs. During this process, introns are excised from primary RNA transcripts, and the flanking exonic sequences are joined together to form functional, mature messenger RNAs (Turunen, Niemelä, et al., 2013; Chen and M. J. Moore, 2014). In most organisms, introns can be excised through two distinct pathways: by the major (greater than 99% of introns in most organisms) or minor (less than 1% in most organisms, with some organisms lacking minor class introns altogether) spliceosomes. Despite the existence of eukaryotic species lacking the minor spliceosome, many reconstructions have shown that all eukaryotes descended from ancestors that contained minor class introns in their genomes, all the way back to the last eukaryotic common ancestor (Russell et al., 2006; Bartschat and Samuelsson, 2010). The minor class introns have consensus splice site and branch point sequences distinct from the major class introns (S. L. Hall and R A Padgett, 1996; Rogozin et al., 2012). It was originally thought that the two classes of introns were distinguished by their terminal dinucleotides, with introns recognized by the major spliceosome beginning with GT and ending with AG, and introns recognized by the minor spliceosome beginning with AT and ending with AC. However, it was later shown that introns in both classes can have either sets of terminal dinucleotides and that longer sequence motifs recognized by the snRNA components unique to each spliceosome distinguish the two classes of introns, hence the designations of "U2-type" for the major and "U12-type" for the minor spliceosomes (Dietrich, Incorvaia, and Richard A Padgett, 1997).

The large-scale and well-organized online databases of genomic data, like Ensembl (Zerbino et al., 2018), UCSC (Casper et al., 2018), and RefSeq (O'Leary

et al., 2016), do not provide extensive annotation information of intronic sequence in particular. Many databases focusing primarily on intron annotation information were created in the early 2000s, but most are no longer accessible (P. J. Lopez and Séraphin, 2000; M. Sakharkar et al., 2000; M. K. Sakharkar, Kangueane, et al., 2000; Saxonov et al., 2000; Fedorov, Stombaugh, et al., 2005; Bhasi et al., 2009), and the ones that remain accessible have not been updated in many years (Burset, Seledtsov, and Solovyev, 2001; Alioto, 2007). The Exon-Intron Database (EID) (Saxonov et al., 2000) was one of the most comprehensive and robust databases in this group, and served as a basis for many further investigations into the peculiarities of introns (Fedorov, Merican, and Walter Gilbert, 2002; Fedorov, S. Roy, et al., 2003; Chamary and Hurst, 2005), including other, more niche intron annotation databases (Fedorov, Stombaugh, et al., 2005; M. K. Sakharkar, Tan, and Souza, 2001). EID was maintained for at least six years, as it was updated in 2006 (Shepelev and Fedorov, 2006), but it is no longer accessible. Some more recent databases have been created, like ERISdb (Szcześniak et al., 2013), JuncDB (Chorev, Guy, and Carmel, 2016) and MIDB (Olthof, Hyatt, and Kanadia, 2019), but they are relatively narrow in scope: ERISdb only annotates splice sites in a selection of plant genomes; JuncDB annotates splice sites in a wide variety of genomes, but does not have any other easily-accessible intron annotation information; MIDB only annotates U12-type introns in the human and mouse genomes. Of all of the databases mentioned above, U12DB (Alioto, 2007), ERISdb and MIDB are the only databases that annotate intron class. Since U12DB has very old annotation data and ERISdb and MIDB only annotate introns in a small number of genomes, there is presently no publicly available source of current U12-type intron annotation for an evolutionarily diverse array of organisms.

Many features of eukaryotic introns have been examined for clues about their evolutionary history. Introns can be assigned to one of three phases based on their position relative to the codons of the flanking exonic sequence: phase 0 introns fall directly between two codons, phase 1 introns fall between the first and second nucleotides of a single codon, and phase 2 introns fall between the second and third nucleotides of a single codon. It has long been noted that introns are not

evenly distributed between the three phases (Long, Rosenberg, and Gilbert, 1995; Long, Souza, and Gilbert, 1995). In conjunction with sequence biases on the exonic sides of splice sites, the phase biases were frequently cited by both sides of the debate between the proponents of the "exon theory of genes" (the idea that primordial genes arose through exon shuffling and introns originally came into existence to facilitate this) (Gilbert, 1987) and those who argued that spliceosomal introns are descended from group II introns that invaded the ancestral eukaryotic genome, preferentially inserting themselves into so-called "proto-splice sites" (Dibb and Newman, 1989; Dibb, 1991; Sverdlov et al., 2004). Shortly after the discovery of U12-type introns (Jackson, 1991; S. L. Hall and R A Padgett, 1996), it was noted that the distribution of U12-type introns in the human genome was nonrandom, further complicating the debate around models explaining the origins of introns by requiring them to explain the presence of two classes of introns, the large discrepancy in the numbers of introns in each class, and the nonrandom distribution of U12-type introns (C. B. Burge, R A Padgett, and Sharp, 1998). Furthermore, the phase biases in U12-type introns were noted to be different from the previously-documented phase biases in U2-type introns (Levine and Durbin, 2001; Sheth et al., 2006).

In an effort to address some of the many open questions about intron evolution, we created the Intron Annotation and Orthology Database (IAOD), a database of intron information for all annotated introns in 24 genomes, including plant, fungal, mammalian, and insect genomes. It also uniquely annotates orthologous introns, and assigns intron class using the `intronIC` (`https://github.com/glarue/intronIC`) algorithm described herein (2.4). The website is publicly accessible at `https://introndb.lerner.ccf.org`.

2.4 Materials and Methods

2.4.1 Classifying intron type with `intronIC`

To begin, `intronIC` identifies all intron sequences in an annotation file by interpolating between coding features (CDS or exon) within the longest isoform

of each annotated gene. For each intron, sequences corresponding to the 5′ splice site (5′SS, from -3 to +9 relative to the first base of the intron) and branch point sequence (BPS) region (from -55 to -5 relative to the last base of the intron) are scored using a set of position weight matrices (PWMs) representing canonical sequence motifs for both U2-type and U12-type human introns. A small "pseudo-count" frequency value of 0.001 is added to all matrix positions to avoid zero division errors while still providing a significant penalty for low-frequency bases. For all scored motifs, the binary logarithm of the U12/U2 score ratio (the log ratio) is calculated, resulting in negative scores for introns with U2-like motifs, and positive scores for introns with U12-like motifs. Because U2-type introns are not known to contain an extended BPS motif, the PWM for the U2-type BPS is derived empirically using the best-scoring U12-type BPS motifs from all introns in the final dataset whose 5′SS U12-type scores are below the 95th percentile. To identify the most likely BPS for each intron, all 12-mer sequences within the BPS region are scored and the one with the highest U12-type log ratio score is chosen. This initial scoring procedure follows the same general approach used by a variety of different groups for bioinformatic identification of U12-type introns. (Bartschat and Samuelsson, 2010; C. B. Burge, R A Padgett, and Sharp, 1998; Levine and Durbin, 2001; Sheth et al., 2006; C.-F. Lin et al., 2010).

As originally shown by Burge et al. (C. B. Burge, R A Padgett, and Sharp, 1998), the 5′SS and BPS scores together are sufficient to produce good binary clustering of introns into putative types, due to strong correspondence between the 5′SS and BPS scores in U12-type introns (Bartschat and Samuelsson, 2010; C. B. Burge, R A Padgett, and Sharp, 1998). While this general feature of the data has often been employed in the identification of U12-type introns, a variety of different techniques have been used to define the specific scoring criteria by which an intron is categorized as U2- or U12-type. Here, we have implemented a machine learning method which uses support vector machine (SVM) classifiers (Cortes and Vapnik, 1995) to assign intron types, an approach which produces good results across a diverse set of species and provides an easy-to-interpret scoring metric.

Our classification method relies upon two pieces of data: PWMs describing

sequence motifs for the different subtypes of U2-type (GT-AG/GC-AG) and U12-type (GT-AG/AT-AC) introns, and sets of high-confidence U2- and U12-type intron sequences with which to train the SVM classifier (Figure 2.1A). Due to the scarcity of bona fide, experimentally-verified U2- and U12-type introns, a certain amount of curation was required to compile type-specific classifier training and scoring data. For the U12-type set, introns from six previously-published studies (Alioto, 2007; Madan et al., 2015; Niemelä and Mikko J Frilander, 2014; Nojima et al., 2018; Pineda and Bradley, 2018; Cologne et al., 2019) as well as highly-conserved introns from a number of multi-species ortholog alignments were scored using SpliceRack (Sheth et al., 2006) PWMs, and those with 5′SS scores > 0 (i.e., 5′SS motifs more similar to U12- than U2-type) present in at least three different sources were kept for use as U12-type training data. Combining these introns with branch point data from (Pineda and Bradley, 2018), we identified likely U12-type BPS motifs which were then used to generate BPS PWMs, requiring an A at either position +9 or +10 (following (Sheth et al., 2006)). For the U2-type set, we first collected intron sequences from the yeast *Schizosaccharomyces pombe*, a species which is believed to lack U12-type introns. These introns were then filtered using data from (Burke et al., 2018) to include only those with direct evidence of splicing, and scored against human SpliceRack PWMs to establish an upper bound for SpliceRack U12-type PWM scores on high-confidence non-U12-type introns. Finally, using a set of introns conserved between human, zebrafish and horseshoe crab we identified human introns found in orthologous groups where every constituent intron had a 5′SS SpliceRack PWM score less than the *S. pombe* U2-type threshold. These human U2-type introns were combined with the U12-type set to build an updated collection of PWMs, and to define positive (U12-type) and negative (U2-type) training sequences for the SVM (Figure 2.1C). In order to establish U2-type BPS PWMs specific to each unique set of input introns (e.g., each different species), U12-type PWMs are first used to find the highest-scoring BPS motifs for all introns whose 5′SS U12-type scores are lower than the 95th percentile (i.e., introns unlikely to be U12-type). These sequences are then used to create U2-type BPS PWMs, making the overall BPS scoring more conservative by defining the

U2-type BPS PWMs using the most U12-like BPS motifs found in the empirical data (similar to the approach described in ref. (Bartschat and Samuelsson, 2010)).

Because clear discrimination between U12-type and U2-type introns can be achieved by considering only two scoring dimensions (Bartschat and Samuelsson, 2010; C. B. Burge, R A Padgett, and Sharp, 1998; Sheth et al., 2006), we use a relatively simple SVM classifier with a linear kernel as implemented in the `scikit-learn` Python library (Pedregosa, 2011). The SVM is trained on a set of two-dimensional vectors, corresponding to the 5′SS and BPS scores of the introns in the training data, which are labeled by intron type. For linear classifiers there is only a single free hyperparameter to be adjusted, C, which is (roughly) the degree to which misclassification of data in the training set is penalized during the creation of an optimized (i.e., wide) margin separating the positive and negative classes. To our knowledge there is no single, standard approach for establishing the best value of C; we therefore chose to optimize C using an iterative cross-validation method which starts with a wide range of logarithmically-distributed values and narrows that range based upon the best-performing (highest balanced accuracy score) value of C in each validation round. After several such iterations, the mean of the resulting range is taken as the final value of C to be used to train the classifier. Balanced accuracy is used as a performance metric due to the highly imbalanced nature of the training data, where the negative class (U2-type) greatly outnumbers the positive class (U12-type). Because the human training data is very well-separated, when applied to intron sequences in the human genome values of $C \geq 10$ perform equally well during cross-validation (Figure 2.1D). Given the broad range of good parameter values, taking the average of all best-performing values results in a more conservative margin (larger C) than taking the default "best" parameter value via the `scikit-learn` API, which simply returns the first rank-1 parameter value found. For the human genome, this approach results in a classifier which performs perfectly on the training sets, with both F1 and precision-recall AUC scores of 1.0 on held-out training data (examples of final scores for human introns in Figs 2.1B,E).

For the purpose of populating the IAOD, `intronIC` was slightly modified to pro-

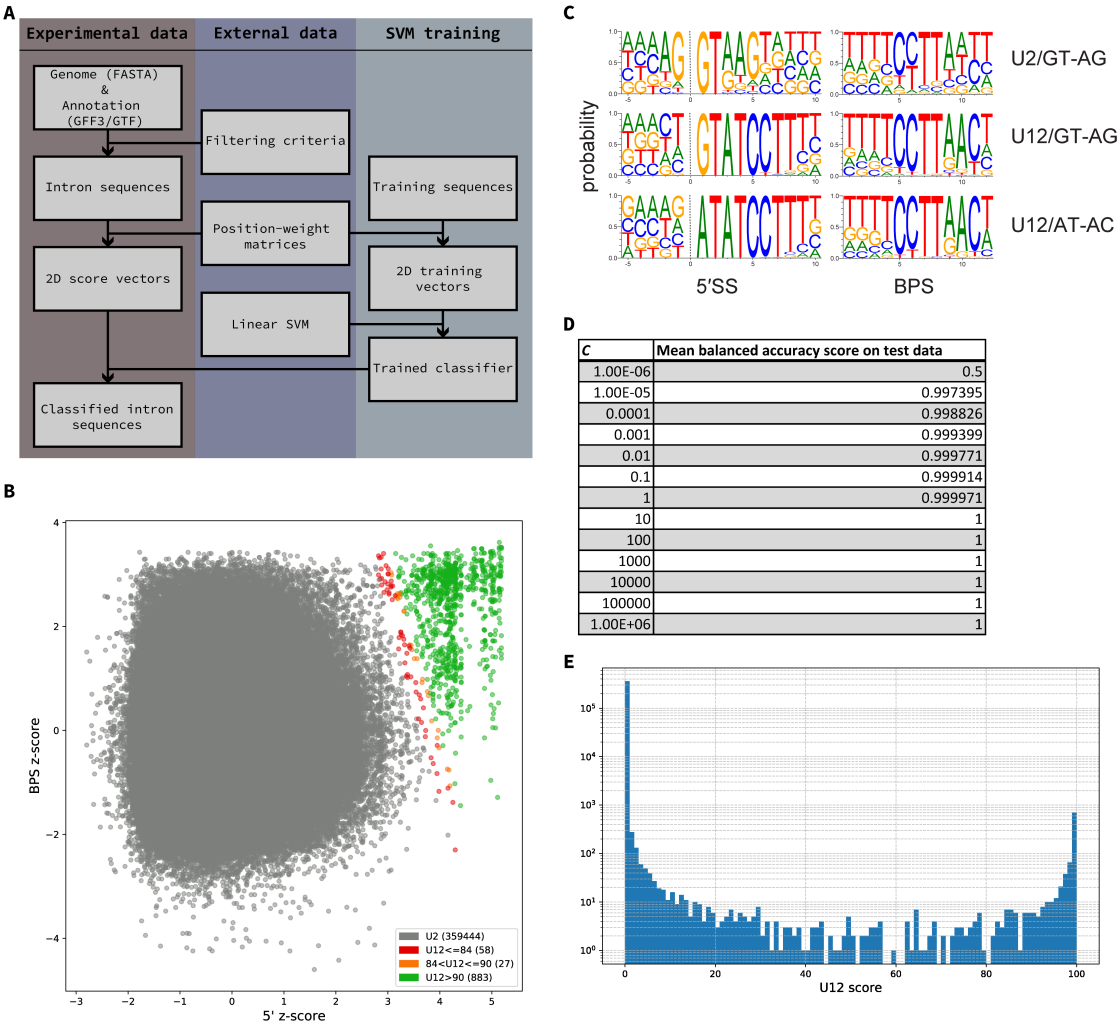


Figure 2.1: **(A)** Overview of the major steps of the `intronIC` algorithm. **(B)** Scatter plot of all classified introns in the human genome; gray: U2-type introns, red: U12-type introns with probability scores $\leq 84\%$; yellow: U12-type introns with probability scores from 84–90%, green: U12-type introns with probability scores $> 90\%$, our chosen scoring threshold. **(C)** Sequence logos of the 5′SS and BPS PWMs for GT-AG/AT-AC U2- and U12-type introns. **(D)** Balanced accuracy performance of the classifier with different values of hyperparameter C on test sets during the first round of the cross-validation process. **(E)** Histogram (with logarithmic scale y-axis) of probability scores for the human data shown in part B.

duce a single output file containing all of the annotation information recorded in the IAOD for each intron—the default and actively-developed version is available for download from (`https://github.com/glarue/intronIC`) and the modified version used for this application is available at (`https://github.com/Devlin-Moyer/IAOD`).

The annotation and sequence files provided as input to `intronIC` were downloaded from release 92 of Ensembl (with the exception of the FUGU5 assembly of the *Takifugu rubripes* genome, which was downloaded from release 94), release 39 of Ensembl Metazoa, or release 40 of Ensembl Plants (Zerbino et al., 2018). Data was obtained for every genome annotated by U12DB with the addition of *Zea mays*, *Oryza sativa*, *Glycine max*, and *Schizosaccharomyces pombe* to increase the evolutionary diversity of the represented genomes.

2.4.2 Annotating introns in non-coding transcripts / regions

To annotate introns, `intronIC` can use either exon or CDS entries in a GFF3 or GTF file. When using exon entries to define introns, intron phase is undefined. In order to get complete annotation of both introns within open reading frames and within untranslated regions or non-coding transcripts, `intronIC` was run twice on each genome analyzed, once producing exon-defined introns and once producing CDS-defined introns. A custom Python script then compared both lists of introns to produce a single list where the CDS-defined intron annotation information was used if the intron was in a coding region and the exon-defined information was used otherwise.

2.4.3 Finding gene symbols

The output of `intronIC` includes the Ensembl gene ID but not the gene symbol for all introns in a genome using an annotation file from Ensembl. Ensembl maintains vast databases of genomic data which are accessible with BioMart (Durinck, Moreau, et al., 2005). `biomaRt` (Durinck, Spellman, et al., 2009) is an R package

for interacting with these databases. A custom R script submitted a list of all of the Ensembl gene IDs in each genome in the database to `biomaRt` and obtained gene symbols for all of those genes.

2.4.4 Assigning orthologous introns

Coding sequences for every annotated transcript in each of the 24 genomes were extracted and translated into their corresponding protein sequences. These sequences were aligned with `DIAMOND` (Buchfink, Xie, and Huson, 2015) to identify sets of best reciprocal hits—considered orthologs going forward—between every pairwise combination of species, using an E-value cutoff of 10^{-10} and `--min-orfs` set to 1. Every pair of orthologous transcripts was then globally aligned at the protein level using `ClustalW` (v2.1; ref. (Larkin et al., 2007)), and all introns in regions of good local alignment between pairs ($\geq$ 40% matching amino acid sequence $\pm$10 residues around each intron) were extracted using custom Python scripts (following the approach of ref. (Scott W Roy, Fedorov, and Walter Gilbert, 2003)). Lastly, conservative clustering of the pairwise orthologous intron sets was performed through identification of all complete subgraphs where every member is an ortholog of every other member (i.e., maximal clique listing) to produce the final intron groups (e.g., `A-B`, `A-C`, `B-C`, `B-D` $\rightarrow$ `A-B-C`, `B-D`).

2.4.5 Database creation

A custom Python script created a PostgreSQL database using the output of `intronIC`, the lists of gene symbols from BioMart, and the list of orthologous groups of introns. All of the orthologous groups were inserted in a table with two columns: a unique numeric ID for each group and a list of all intron labels belonging to that group. One table for each genome contains, for each intron: the abbreviated sequence (see above), taxonomic and common names of the organism, name of the genome assembly, `intronIC` score, intron class (determined from the `intronIC` score), `intronIC` label, chromosome, start coordinate, stop coordinate, length, strand, rank in transcript, phase, terminal dinucleotides, upstream exonic

sequence (50 nt), 3′ terminus with the branch point region enclosed with brackets (40 nt), downstream exonic sequence (50 nt), full intron sequence, Ensembl gene ID, Ensembl transcript ID, and gene symbol. Another table with identical fields contains all U12-type introns from all genomes.

2.4.6 Website design

The website was constructed using Django 2.0 , an open-source Python web development framework, and Bootstrap 4.0.0, an open-source framework for front-end web development. The search engines use the Django ORM to interact with the PostgreSQL database.

There are four search engines on the website: the main, advanced, U12, and orthologous searches. The main and advanced search interfaces have input fields corresponding to individual columns in the database, so the text input in each field can easily be matched with the appropriate column using the Django ORM. The U12 search engine uses PostgreSQL search vectors to allow users to make full text queries against the database, i.e., users can input a string containing one term corresponding to as many fields as they like and get a result. However, if the search query contains, e.g., the names of two different species or genes, no results will be returned, since no single record (intron) in the database corresponds to multiple species or genes. This limits the number of possible queries, but allows for a simple user interface for simple queries concerning U12-type introns. Since the main and advanced search engines require users to specify which field of the database each term of their query corresponds to, it does not need to use full text search vectors, and can consequently accept multiple search terms for each field. The homolog search engine also makes use of PostgreSQL search vectors to find the row of the homolog table containing the intron ID input by the user.

2.4.7 Assessing Randomness of Distribution of U12-type Introns

If U12-type introns were randomly inserted a genome, we would expect the distribution of U12-type introns per gene to be binomial with parameters $n =$ number of genes with at least one U12-type intron and $p = 1-(1-x)^{m-1}$, where x is the proportion of U12-type introns in the genome and m is the average number of introns in the genome. Data from the IAOD was used to obtain n, x, and m for each genome in the database that contained at least one U12-type intron. The dbinom function in R was used to compute the probability of observing the observed number of genes with multiple U12-type introns in each genome. Table 2.3 lists the parameters passed to the `dbinom` function.

To ensure that the observed clustering of U12-type introns in the same genes was not an artifact of U12-type introns with alternative splice sites being recorded as distinct U12-type introns, all intron coordinates listed by `intronIC` were used to create a graph where each node corresponded to a position within each genome (e.g., `GRCh38+chr1+492045` corresponds to base pair 492045 on chromosome 1 in assembly GRCh38 of the human genome) and two nodes are joined with an edge if they appear in the same row of the list of intron coordinates. In this graph, alternatively-spliced introns are evident as clusters of more than 2 nodes, so each cluster represents a single intron, regardless of how many alternative splice sites it possesses. A single edge from each cluster was selected and the corresponding coordinates were matched to the original `intronIC` output to get accurate counts of the total number of unique introns in each class in all genomes annotated in the IAOD.

2.5 Results and Discussion

2.5.1 Identification of U12-type introns across 24 eukaryotic species

We developed the method textttintronIC (see 2.4), implemented in Python, and used it to perform genome-wide identification of U2- and U12-type spliceosomal introns in 24 eukaryotic species including 14 vertebrate animals, 5 invertebrate animals, 4 plants and two yeasts. For each species, type-specific position-weight matrices (PWMs) for the $5'$ splice site and branch point sequences were used to create score vectors for every intron in each genome. These score vectors were then compared against corresponding vectors in high-confidence training sets from Homo sapiens using a machine-learning (SVM) classifier to assign each intron a probability of being U12-type (see Figure 2.1 for an overview of the major steps in the algorithm and examples of classifier performance on human data). Once trained on the conserved intron data, the classifier assigned every intron in the experimental set a probability of being U12-type. Introns with at least a 90% probability of being U12-type were classified as U12-type, which produces classifications in good agreement with previously-reported findings for well-studied species. For example, running our method on U12-type intron sequences from the U12DB (Alioto, 2007) results in equivalent classifications for 96% (381/398) of the U12DB introns in chicken, 97% (535/554) in mouse, 94% (15/16) in *Drosophila melanogaster* and 95% (656/691) in human. In *Arabidopsis thaliana*, our method matches the calls in the U12DB 94% of the time (223/238), with similar results (269/292, 92%) for U12-type introns from the plant-specific database ERISdb (Szcześniak et al., 2013). Furthermore, in each test species listed above `intronIC` identifies additional putative U12-type introns not present in existing databases, likely due to a combination of newer annotation data and our method's sensitivity. In *Caenorhabditis elegans*, a well-annotated species believed to have lost all of its U12-type introns, when run on all introns (not just those from the longest isoform per gene) our method categorized only 1/116241 introns as U12-type, suggesting a false-positive rate of less than 0.001%. A total of 8,967 U12-type introns were identified using this

technique. Groups of analyzed introns in conserved regions of homologous genes were also annotated; collectively, these data constitute the Intron Annotation and Orthology Database (IAOD).

Figure 2.2 compares the number of U12-type introns annotated in each species in the IAOD with the numbers of U12-type introns annotated by previous databases annotating U12-type introns: U12DB (Alioto, 2007), SpliceRack (Sheth et al., 2006), and ERISdb (Szcześniak et al., 2013). The IAOD often annotates many more introns than U12DB, likely due to the different approaches to annotating intron class and the quality of the genome assemblies used. In U12DB, U12-type introns were annotated by mapping a set of reference introns from Homo sapiens, *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Ciona intestinalis* to the whole genomes of every other organism in the database, while introns in the IAOD were annotated directly from every genome in the database using the intron-classifying program `intronIC` (see 2.4 for details). U12DB primarily annotates U12-type introns in all represented species that are orthologous to the reference U12-type introns (Alioto, 2007), while the IAOD annotates U12-type introns in all genomes independently, using the species-specific annotations for each genome. Furthermore, the genome assemblies and annotations used to identify introns in the present study are all several versions newer than those used in U12DB, so part of the discrepancy in the number of U12-type introns annotated is likely due to an increase in the number of annotated genes and splice sites since the creation of U12DB. While `intronIC` itself does not provide homology information about the annotated introns, the IAOD also annotates intron orthologs: of the 3,645,636 total introns in the IAOD, 54% (1,989,840) have at least one other intron annotated as being in a conserved region of a homologous gene in another genome in the IAOD.

As shown in Figure 2.2, there are substantially fewer U12-type introns in the analyzed invertebrate animals than in the vertebrates, and none in either species of yeast analyzed, consistent with earlier findings (C. B. Burge, R A Padgett, and Sharp, 1998; C.-F. Lin et al., 2010). The numbers of introns determined by `intronIC` to be U12-type in a few species deserve special attention. The numbers of U12-type introns in *A. thaliana*, *O. sativa* and *Z. mays* are noteworthy because

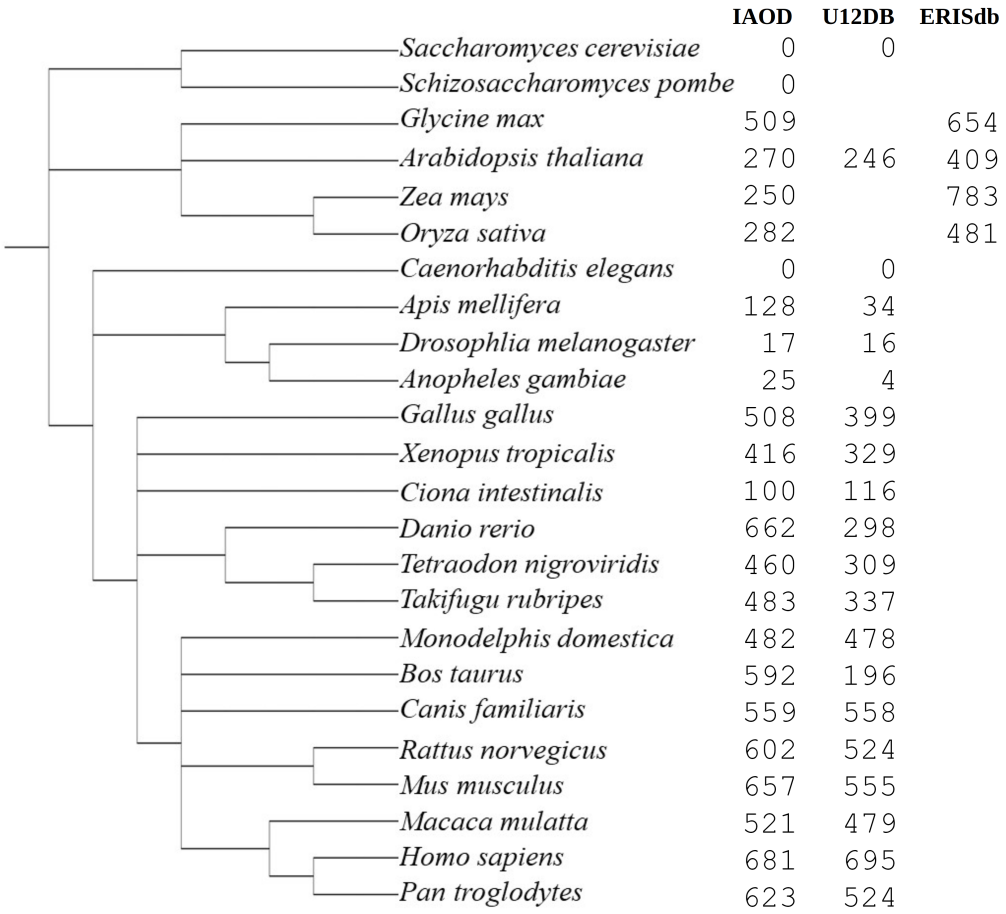


Figure 2.2: Phylogenetic distribution of U12-type introns in all species annotated by the Intron Annotation and Orthology Database (IAOD), U12DB (Alioto, 2007), SpliceRack (Sheth et al., 2006) and ERISdb (Szcześniak et al., 2013). Blank entries in the table represent organisms not represented in the respective database. Counts of U12-type introns in the IAOD only represent introns flanked by coding exons. The NCBI Taxonomy Browser (Federhen, 2012) and Integrative Tree of Life (Letunic and Bork, 2016) were used to create the phylogenetic tree.

there are substantially fewer U12-type splice sites annotated in the IAOD than in ERISdb, but inspection of the U12-type splice sites annotated in ERISdb reveals many duplicate sequences. These duplicates arise from the fact that ERISdb counts each set of U12-type splice sites from every transcript of every gene as a distinct set of U12-type splice sites. In the case of *A. thaliana*, of the 414 U12-type splice sites annotated in ERISdb, there are only 292 unique sequences, which is much closer to the 269 annotated in the IAOD.

2.5.2 Phase bias of U12-type introns is consistent with the conversion hypothesis

The phase biases observed in the IAOD (Figure 2.3) agree with the results of previous studies and extend them to many more organisms: an excess of phase 0 introns among U2-type introns (Long, Souza, and Gilbert, 1995; Long, Rosenberg, and Gilbert, 1995; C. B. Burge, R A Padgett, and Sharp, 1998; Levine and Durbin, 2001; Nguyen, Yoshihama, and Kenmochi, 2006), and a bias against phase 0 introns among U12-type introns (C. B. Burge, R A Padgett, and Sharp, 1998; Sheth et al., 2006) are seen in all studied lineages, and the presence of these biases in both plant and animal genomes suggest a deep evolutionary source. Multiple explanations for the overrepresentation of phase 0 U2-type introns have been proposed, including exon shuffling (Long and Rosenberg, 2000), insertion of introns into proto-splice sites (Dibb and Newman, 1989; Nguyen, Yoshihama, and Kenmochi, 2006), and preferential loss of phase 1 and 2 introns (Rogozin et al., 2012). These models do not consider or explain the underrepresentation of phase 0 U12-type introns.

To account for the phase biases present in U12-type introns, we propose that the observed phase biases in both classes of introns can be explained by an extension of the class-conversion hypothesis proposed by Burge et al. (C. B. Burge, R A Padgett, and Sharp, 1998). This hypothesis arose from the observation that U12-type introns in human genes were often found to have U2-type introns at orthologous positions in *C. elegans* genes. Dietrich et al. (Dietrich, Incorvaia, and Richard A Padgett, 1997) showed that U12-type introns could be converted to U2-type introns with as few as two point mutations. These results also suggest

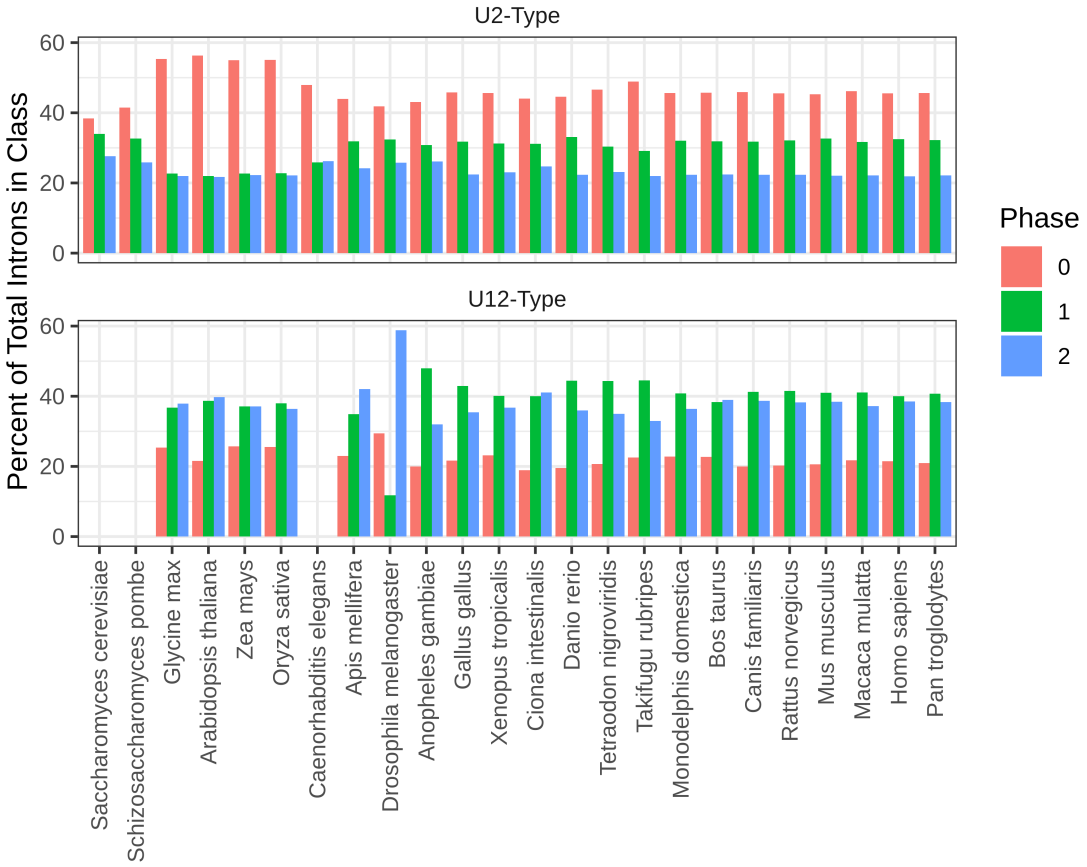


Figure 2.3: Phase distribution of introns within each class in all genomes annotated in the IAOD. Organisms are grouped by phylogeny. The bias against phase 0 U12-type introns is statistically significant in all organisms but *G. max*, *O. sativa*, *X. tropicalis* and *Z. mays* (chi-squared; $P < 0.10$). The bias toward phase 0 U2-type introns is statistically significant in all organisms but *A. mellifera*, *D. melanogaster*, *S. cerevisiae* and *S. pombe* (chi-squared; $P < 0.10$).

that class conversion is likely to only proceed from U12-type to U2-type, a hypothesis for which we find support in the distinctly U12-like phase distribution among U2-type introns with > 1 U12-type ortholog (i.e., putative U12-type-to-U2-type conversions) (Figure 2.6). In light of this, one possible explanation for the current data is that, at an early stage in eukaryotic evolution, there were many more U12-type introns than are currently observed in any characterized genome, and the phase bias arose as phase 0 U12-type introns were preferentially converted into U2-type introns, producing both an overrepresentation of phase 0 U2-type introns and an underrepresentation of phase 0 U12-type introns. This selectivity for phase 0 introns in the class conversion process rests on the function of the -1 nucleotide relative to the $5'$ splice site in both spliceosomes.

As shown in Figure 2.4, there is a large excess of G at the -1 position of U2-type $5'$ splice sites, across all three phases, in agreement with earlier investigations (Long and Rosenberg, 2000; Mount, 1982). Figure 2.4 also shows that there is an excess of -1U in U12-type introns in all three phases. There also appears to be a bias against -1 A and G in phase 1 and phase 2 U12-type introns, but not in phase 0 U12-type introns. Interestingly, when introns are grouped by terminal dinucleotides, these biases are only found in U12-type introns with GT-AG terminal dinucleotides and not in U12-type AT-AC introns (Figure 2.5). The preference for -1G at U2-type $5'$ splice sites appears to be due to the fact that the -1 nucleotide pairs with a C on the U1 snRNA (Pomeranz Krummel et al., 2009; Kondo et al., 2015). The preference for -1U at U12-type $5'$ splice sites is more mysterious, as no snRNAs are known to bind to this position. It was previously shown that the U11/U12-48K protein interacts with the +1, +2 and +3 nucleotides at the U12-type $5'$ splice site in a sequence-specific fashion, but the specificity of the interaction with the -1 position was not studied (Turunen, Will, et al., 2008). As noted above, the bias against -1G in U12-type introns with GT-AG terminal dinucleotides could be a consequence of the gradual conversion of many U12-type introns into U2-type introns. The lack of consistent -1 nucleotide biases in AT-AC introns of either class may be due to the fact that AT-AC introns are poorly recognized by the U2-type spliceosome (Kondo et al., 2015) and were thus largely unaffected by the class conversion process.

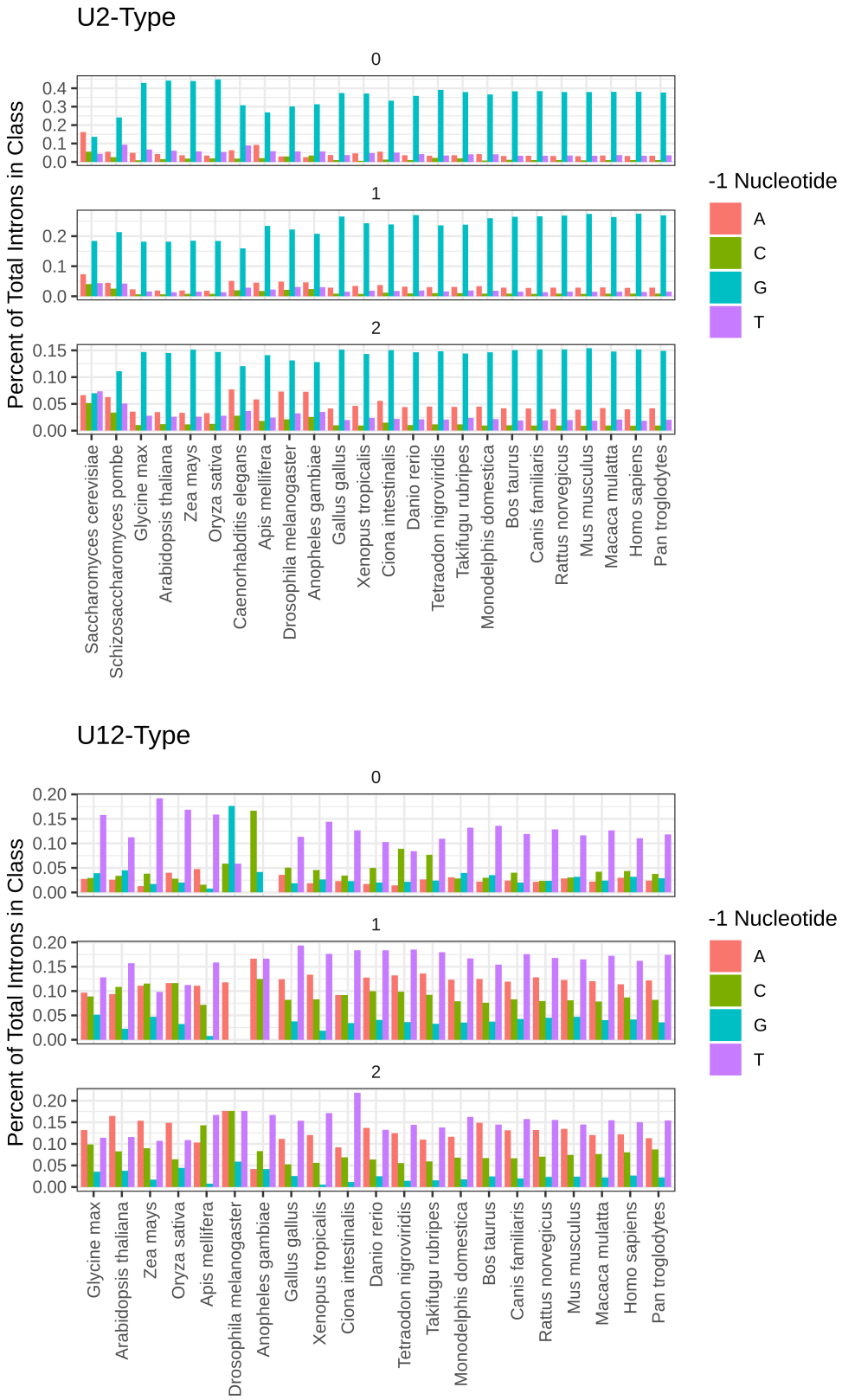


Figure 2.4: Percentages of introns with the specified nucleotide immediately upstream of the $5'$ splice site in each phase of both classes of introns. Organisms are grouped by phylogeny.

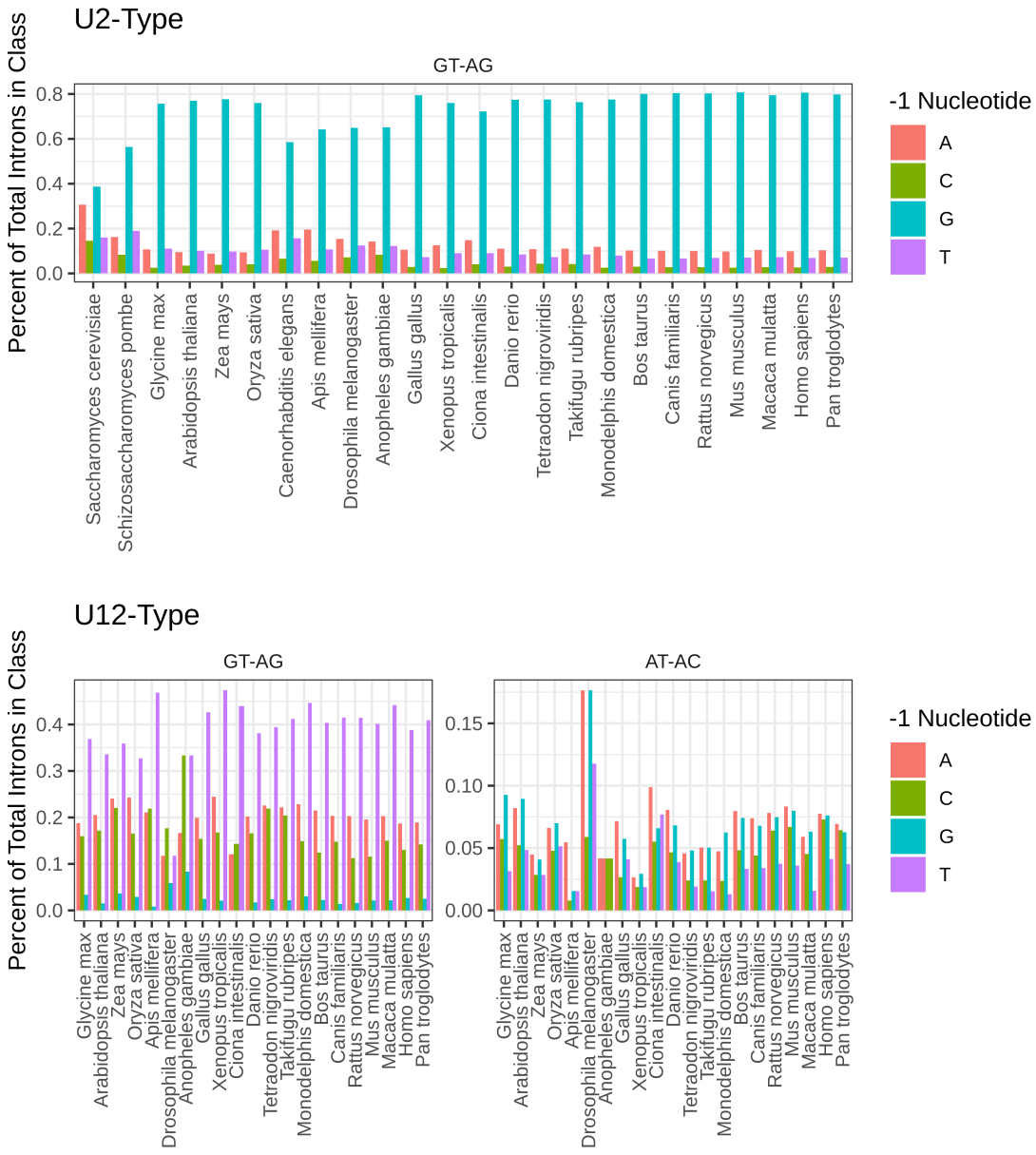


Figure 2.5: Nucleotide biases at the -1 position relative to the 5′ splice site for all organisms (excluding those lacking U12-type introns) annotated in the IAOD, grouped by terminal dinucleotides and intron class.

We propose that this preference for conversion of phase 0 U12-type introns is due to the fact that introns with a G at the -1 position relative to the 5′ splice site bind more strongly to the U1 snRNA (Kondo et al., 2015; Turunen, Will, et al., 2008), and the -1 nucleotide of phase 0 introns is the final wobble position of the corresponding codon and can be a G in 13 of 20 codon families. Thus, the -1 nucleotides of phase 0 U12-type introns were more free to mutate to G and increase the affinity of the U2-type spliceosome for their 5′ splice sites, gradually accumulating mutations in the other sequences required for recognition by the U2-type spliceosome (Dietrich, Incorvaia, and Richard A Padgett, 1997). Table 2.1 contains some examples of orthologous introns of different classes that demonstrate the class conversion process. This unidirectional conversion process also provides an explanation for the low abundance of U12-type introns in modern eukaryotic genomes.

2.5.3 U12-type introns are non-randomly distributed across genes

Multiple previous surveys of U12-type introns have revealed that the distribution of U12-type introns in the human genome is non-random, i.e., there is a statistically significant tendency for U12-type introns to cluster together in the same genes (C. B. Burge, R A Padgett, and Sharp, 1998; Levine and Durbin, 2001; Sheth et al., 2006). Repeating this analysis for all genomes in the IAOD (except *S. cerevisiae*, *S. pombe*, and *C. elegans*, as they lack U12-type introns) replicated their findings in all 21 genomes ($P < 0.05$ for all genomes; see 2.4 and Table 2.3). Many explanations for this nonrandom distribution have been proposed, including the fission-fusion model of intron evolution (C. B. Burge, R A Padgett, and Sharp, 1998); a difference in the speed of splicing of U12-type and U2-type introns (Sheth et al., 2006; Niemelä and Mikko J Frilander, 2014); and the idea that the U12-type introns arose during an invasion of group II introns after U2-type introns had already seeded the ancestral eukaryotic genome, meaning the new U12-type introns could only be inserted in certain locations (Lynch and Richardson, 2002). The fission-fusion model posits that two separate lineages of the proto-eukaryote

Table 2.1: Groups of orthologous introns of different classes. Introns with nothing in the gene column came from transcripts with no gene name annotated by Ensembl. Vertical bars in the sequence column denote splice sites, and the middle sequence flanked by ellipses is the putative branch point region as annotated by `intronIC`

Organism	Gene	Intron Class	Phase	Sequence
Zea mays		U12-type	0	GCAAAG|GTATCCTTTT...TCCTCCTAAACT...TGCAG|TCCTCC
Oryza sativa		U2-type	0	GCCAAG|GTAATTTATA...TTAATGTTTAAT...TGCAG|TTAATG
Zea mays		U2-type	0	AAGCGG|GTATGTCTAG...TTGATCTCACCT...ATCAG|TTGATC
Glycine max		U12-type	0	AAGCGT|GTATCCTTCA...TTGTCCTTGACC...GAAAG|TTGTCC
Zea mays		U2-type	1	TCAACA|GTACGCAACA...TCCTTCTTAATT...TGTAG|TCCTTC
Oryza sativa		U12-type	1	TCAACA|GTATCCATCA...TTTTTCTTAACT...TGTAG|TTTTTC
Arabidopsis thaliana		U2-type	1	TCAACA|GTAAATTTTC...TTTCTCTTGACC...TGCAG|TTTCTC
Canis familiaris	SMYD2	U12-type	1	ACAAAT|ATATCCTTTA...CTTTCCTTGACA...AGCAC|CTTTCC
Homo sapiens	SMYD2	U12-type	1	ATAAAT|ATATCCTTTA...CTTTCCTTGACT...AGCAC|CTTTCC
Mus musculus	Smyd2	U12-type	1	ACAAAT|ATAACCTTTC...GTTTCCTTGACG...AGCAC|GTTTCC
Macaca mulatta	SMYD2	U2-type	1	ACAACT|GCCCTGATGG...GTTTCCTTGACT...CACAG|GTTTCC
Pan troglodytes	SMYD2	U12-type	1	ATAAAT|ATATCCTTTA...GTTTCCTTGACT...AGCAC|GTTTCC
Rattus norvegicus	Smyd2	U12-type	1	ACAAAT|ATAACCTTTC...GTTTCCTTGACG...AGCAC|GTTTCC
Anopheles gambiae		U2-type	2	TAATCC|GTATGTAACC...TGTTTCTCCTTT...TGTAG|TGTTTC
Monodelphis domestica	RNF121	U12-type	2	TAACCC|GTATCCTTTT...TTTTCTTTAACC...TGAAG|TTTTCT
Rattus norvegicus	Rnf121	U12-type	2	CAATCC|GTATCCTTTG...TGATCCTTAACA...GACAG|TGATCC
Homo sapiens	UFD1	U12-type	2	AGCCGT|GTATCTTTTT...GTTGCCTTGACA...TGCAG|GTTGCC
Pan troglodytes	UFD1	U12-type	2	AGCCGT|GTATCTTTTT...GTTGCCTTGACA...TGCAG|GTTGCC
Tetraodon nigroviridis	ufd1l	U2-type	2	AGCAGT|GTAAGAACGA...GAATTGTTTTCT...TGCAG|GAATTG

evolved distinct spliceosomes and then fused their genomes such that all genes originally contained either only U2-type introns or only U12-type introns. Thus, modern U2-type introns in genes also containing U12-type introns were originally U12-type introns that were subjected to the class conversion process discussed above (C. B. Burge, R A Padgett, and Sharp, 1998). An alternative argument for the low abundance of U12-type introns is that they are excised more slowly than U2-type introns, so genes that contain U12-type introns contain them because those genes need to be expressed slowly for some reason (Sheth et al., 2006; Niemelä and Mikko J Frilander, 2014). However, it has since been shown that the rate of excision of U12-type introns is not sufficiently different from the rate of excision of U2-type introns to produce a meaningful impact on the expression of transcripts containing U12-type introns (Singh and Richard A Padgett, 2009). Furthermore, recent evidence suggests that the rates of both types of splicing are sufficiently fast that most introns will be excised cotranscriptionally (Nojima et al., 2018).

2.5.4 Low conservation of U12-type intron positions between animals and plants

Basu et al. (Basu, Rogozin, and Koonin, 2008) argued that the number of U12-type introns present in the ancestral eukaryotic genome was unlikely to be substantially larger than the largest number of U12-type introns observed in any modern genome, thus suggesting that the process of class conversion is a minor evolutionary force. However, the basis of their argument is the finding that the positions of U12-type introns are more highly conserved than the positions of U2-type introns between humans and *Arabidopsis thaliana*, a result that the present data do not support: we find that out of the 93 U12-type introns in the human genome in regions of good alignment to *A. thaliana*, only 8 (9%) are in conserved positions, while out of the 9,527 U2-type introns in such regions of the human genome, 2,098 (22%) are in conserved positions in *A. thaliana*. Thus, our comparative analysis is consistent with U12-type intron enrichment in the ancestral eukaryotic genome relative to the most U12-intron-rich extant lineages.

2.5.5 Splicing boundaries of U12-type introns

The majority of introns annotated in the IAOD in both classes begin with GT and end with AG (Table 2.2), in agreement with previous studies (Burset, Seledtsov, and Solovyev, 2001; C. B. Burge, R A Padgett, and Sharp, 1998; Sheth et al., 2006). A substantial minority of U2-type introns, but almost no U12-type introns, were found to have GC-AG as their terminal dinucleotides in many of the analyzed genomes, reflecting their previously documented role in alternative 5′ splice site selection in U2-type splicing in many organisms (Rogozin et al., 2012; Sheth et al., 2006; Thanaraj and Clark, 2001; Farrer et al., 2002; Churbanov et al., 2008). Several previous studies have found numerous introns with other non-canonical terminal dinucleotides in multiple genomes, sometimes with functional roles in regulation of alternative splicing (Burset, Seledtsov, and Solovyev, 2001; Sheth et al., 2006; Thanaraj and Clark, 2001; Szafranski et al., 2007), but `intronIC` has annotated many thousands of U2-type introns with non-canonical terminal dinucleotides in certain organisms, such as *Gallus gallus* and *Tetraodon nigroviridis* (Table 2.2). Inspection of these introns reveals that the vast majority of these splice sites are only a few nucleotides away from a conventional U2-type splice site with canonical terminal dinucleotides; these splice sites with non-canonical dinucleotides were likely annotated on the basis of conserved exon boundaries, without regard for the precise placement of the splice sites. The proportion of U12-type introns with non-canonical terminal dinucleotides (Table 2.2) largely agrees with previous investigations (Sheth et al., 2006; C.-F. Lin et al., 2010; Dietrich, Fuller, and Richard A Padgett, 2005b).

2.5.6 Distribution of intron lengths of U12- and U2-type introns

Figure 2.7 shows the distributions of intron lengths in six of the genomes annotated in the IAOD, representing each general type of length distribution observed in the IAOD. In accordance with previous studies, when plotted on a log scale, there are two distinct peaks in the distribution of intron lengths in U2-type in-

Table 2.2: Percentages of introns with various terminal dinucleotides in each class in all annotated genomes. Organisms are sorted in the phylogenetic order shown in Figure 2.2

Intron class	U2-type				U12-type			
Terminal dinucleotides	GT-AG	GC-AG	AT-AC	Other	GT-AG	GC-AG	AT-AC	Other
Saccharomyces cerevisiae	912	2.7	0	5.7	0	0	0	0
Schizosaccharomyces pombe	100	0.12	0	0.01	0	0	0	0
Glycine max	98	1.6	0	0	76	0.01	23	0
Arabidopsis thaliana	99	1.0	0.01	0.08	73	0	26	1.1
Zea mays	99	0.50	0.05	0.41	86	0	13	1.2
Oryza sativa	97	0.30	0.01	2.5	74	0	22	3.8
Caenorhabditis elegans	99	0.64	0	0.18	0	0	0	0
Apis mellifera	99	0.65	0.01	0.03	91	0.71	8.6	0
Drosophila melanogaster	99	0.75	0.01	0.07	47	5.3	47	0
Anopheles gambiae	100	0.26	0	0.09	88	4.2	8.3	0
Ciona intestinalis	91	0.70	0.04	8.7	65	0	23	12
Gallus gallus	97	2.2	0.01	0.56	77	0.97	18	3.4
Xenopus tropicalis	85	0.90	0.07	14	78	0.22	8	14
Danio rerio	97	1.2	0.02	1.4	75	0	22	2.7
Tetraodon nigroviridis	86	1.4	0.03	13.0	77	0.80	12	11
Takifugu rubripes	91	5.7	0	3.4	79	4.3	12	5.0
Monodelphis domestica	98	0.50	0.01	1.1	82	0.20	14	4.1
Bos taurus	94	0.89	0.03	5.1	69	0.81	21	10
Canis familiaris	95	0.86	0.02	3.7	69	0.35	20	11
Rattus norvegicus	97	0.78	0.02	2.1	69	0.49	23	6.5
Mus musculus	99	0.78	0.01	0.16	69	0.47	25	5.4
Macaca mulatta	97	3.4	0.01	0.02	78	2.8	17	2
Pan troglodytes	97	2.8	0.01	0.14	72	1.7	21	4.7
Homo sapiens	99	0.77	0.01	0.15	68	0.61	26	5.2

trons in humans and chicken while the distribution of U12-type intron lengths has only one peak (Levine and Durbin, 2001; Vinogradov, 1999) (these peaks are not apparent when length is plotted on a linear scale). A previous study considered the distribution of intron lengths amongst several eukaryotic genomes collectively (Dietrich, Fuller, and Richard A Padgett, 2005b), producing a distribution similar to those observed in the human and chicken genomes in Figure 2.7. However, Figure 2.7 demonstrates great diversity in the distributions of intron lengths amongst eukaryotes; zebrafish have two distinct peaks of comparable size of intron lengths in both classes, while corn, honeybee and fugu have large peaks of shorter introns and very small peaks of longer introns in both classes. The significance of these variations is unclear; differing distributions of intron lengths in the two classes of introns have previously been used to argue that U12-type introns are recognized through intron definition, while U2-type introns are recognized by exon definition (Patel and Joan A Steitz, 2003). However, Figures 2.8 and 2.9 show that the mean intron lengths in both classes of intron in all 24 genomes annotated in the IAOD correlate strongly with genome size (Pearson's r: 0.87 for U12-type introns and 0.93 for U2-type introns), consistent with previous findings (Vinogradov, 1999; Deutsch and Long, 1999; Lynch and Conery, 2003). This correlation suggests that mean intron lengths in both classes are generally a function of genome size and not a reflection of intron definition imposing a restriction on the size of U12-type introns.

Interestingly, Figures 2.8 and 2.9 show that the relationship between mean intron length and genome size differs between vertebrates, insects, and plants. In insects, the total genome size remains very small and the mean intron length does not appear to correlate with total genome size. This may be related to the greater prevalence of intron definition in splicing in insects than in vertebrates (Patel and Joan A Steitz, 2003; De Conti, Baralle, and Buratti, 2013). In plants, mean intron length does appear to correlate with total genome size, but mean intron length increases much more slowly with total genome size than in vertebrates. Similar correlations are observed between mean intron length and gene number, with a much more prominent difference between the slope of the correlation in plants and

vertebrates (data not shown). The significance of this remains unclear.

2.6 Concluding Remarks

We have created a database of intron annotation and homology information and used it to investigate several evolutionary hypotheses regarding the two classes of spliceosomal introns in eukaryotes. We have also created a web-based interface for querying this database to facilitate further investigations, and have made publicly available the intron classification algorithm used to populate the database. The relationships between intron class, phase, terminal dinucleotides and -1 nucleotides at the 5′ splice site and the nonrandom distribution of U12-type introns annotated in the IAOD do not support many previous models that explain these patterns (Rogozin et al., 2012; Dibb and Newman, 1989; Nguyen, Yoshihama, and Kenmochi, 2006; Long and Rosenberg, 2000), but do support an extension of the class conversion model previously proposed (C. B. Burge, R A Padgett, and Sharp, 1998).

2.7 Data Availability

The IAOD is publically accessible at `https://introndb.lerner.ccf.org/` and all code used to create the database and run the website is available at the following GitHub repository `https://github.com/Devlin-Moyer/IAOD`. The standalone `intronIC` algorithm is available at `https://github.com/glarue/intronIC`.

2.8 Funding

This work was supported by the National Institutes of Health [R01GM104059 to RAP]; the National Science Foundation [1616878 and 1751372 to SWR].

2.9 Author Contributions

DM planned, designed, and created the website and all scripts used to generate and process data except `intronIC`. GL designed and wrote `intronIC`, and generated underlying intron and orthology data. DM wrote the manuscript with input from all authors.

2.10 Acknowledgments

Michael Weiner provided invaluable technical support by hosting and managing the website. Rosemary Dietrich provided extensive background on splicing and general advice on various aspects of data interpretation. Daniel Blankenburg provided valuable feedback on various aspects of the web design and method of annotating orthologous introns.

2.11 Supplementary materials

2.11.1 Supplementary tables

Table 2.3: Probabilities that U12-type introns were randomly inserted into each genome along with all parameters used to calculate those probabilities. If U12-type introns were randomly inserted a genome, one would expect the distribution of U12-type introns per gene to be binomial with parameters n = number of genes with a U12-type intron and $p = 1 - (1 - x)^{m-1}$, where x is the proportion of U12-type introns in the genome and m is the average number of introns in the genome. Organisms are grouped by phylogeny. *S. cerevisiae*, *S. pombe*, and *C. elegans* were omitted from this analysis as they lack U12-type introns. *D. melanogaster* was omitted from this analysis as there are no genes with multiple U12-type introns in that genome.

Organism	Number of U12-Type Introns	Genes with Multiple U12-Type Introns	Proportion of Introns That Are U12-Type	Probability That U12-Type Introns Are Randomly Distributed
Glycine max	521	25	0.0025	2.3E-10
Arabidopsis thaliana	274	16	0.0025	3.6E-08
Zea mays	282	8	0.0017	0.0011
Oryza sativa	281	12	0.0024	7.5E-06
Apis mellifera	140	4	0.0021	0.026
Anopheles gambiae	24	1	6.0E-4	0.036
Ciona intestinalis	104	4	0.0011	0.0027
Gallus gallus	503	33	0.0034	1.8E-08
Xenopus tropicalis	422	85	0.0026	8.6E-47
Danio rerio	643	43	0.0030	2.9E-10
Tetraodon nigroviridis	474	25	0.0028	1.4E-4
Takifugu rubripes	521	23	0.0032	0.0025
Monodelphis domestica	470	28	0.0028	2.6E-08
Bos taurus	574	40	0.0035	1.3E-06
Canis familiaris	542	44	0.0035	7.7E-09
Rattus norvegicus	583	42	0.0034	7.7E-10
Mus musculus	630	49	0.0032	2.5E-17
Macaca mulatta	550	37	0.0033	9.9E-09
Pan troglodytes	639	45	0.0036	5.7E-09
Homo sapiens	674	48	0.0033	4.8E-16

2.11.2 Supplementary figures

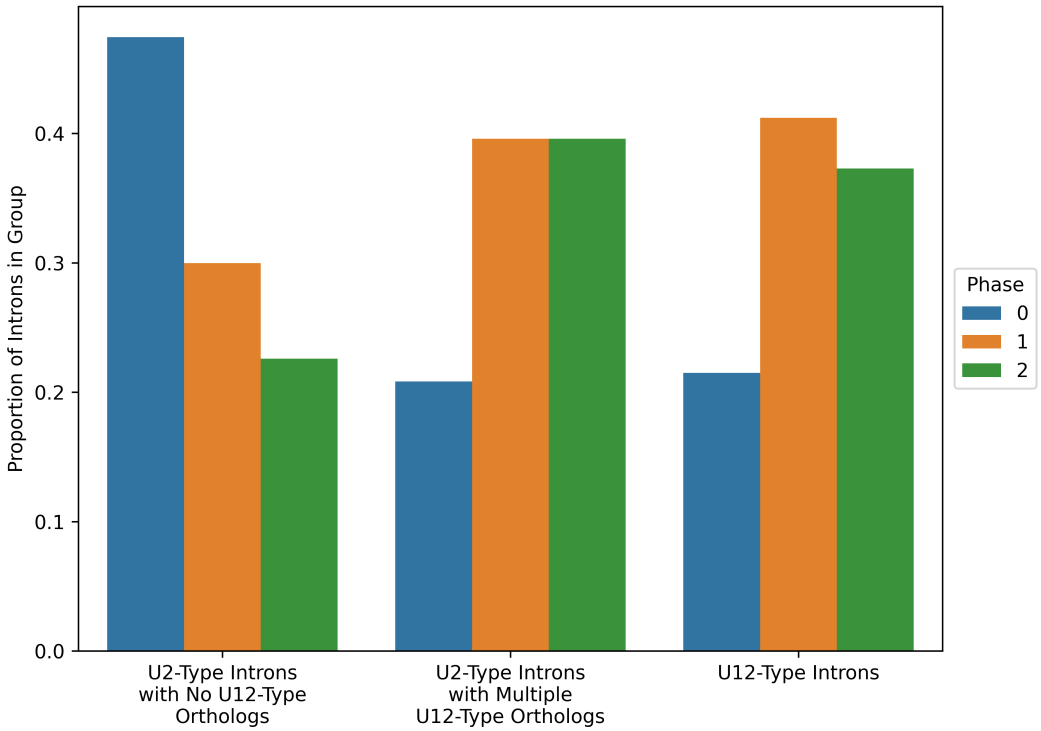


Figure 2.6: Distributions of intron phases (0, 1, or 2) across different sets of introns, showing similarities between putative U12-type → U2-type conversions (i.e., U2-type introns from orthologous groups containing multiple U12-type introns) and U12-type introns. (left) U2-type introns in orthologous groups where no orthologous intron is called as U12-type (n=3,348,724); (middle) U2-type introns in orthologous groups where at least two other members are called as U12-type (n=437); (right) called U12-type introns without U2-type orthologs (n=7,820).

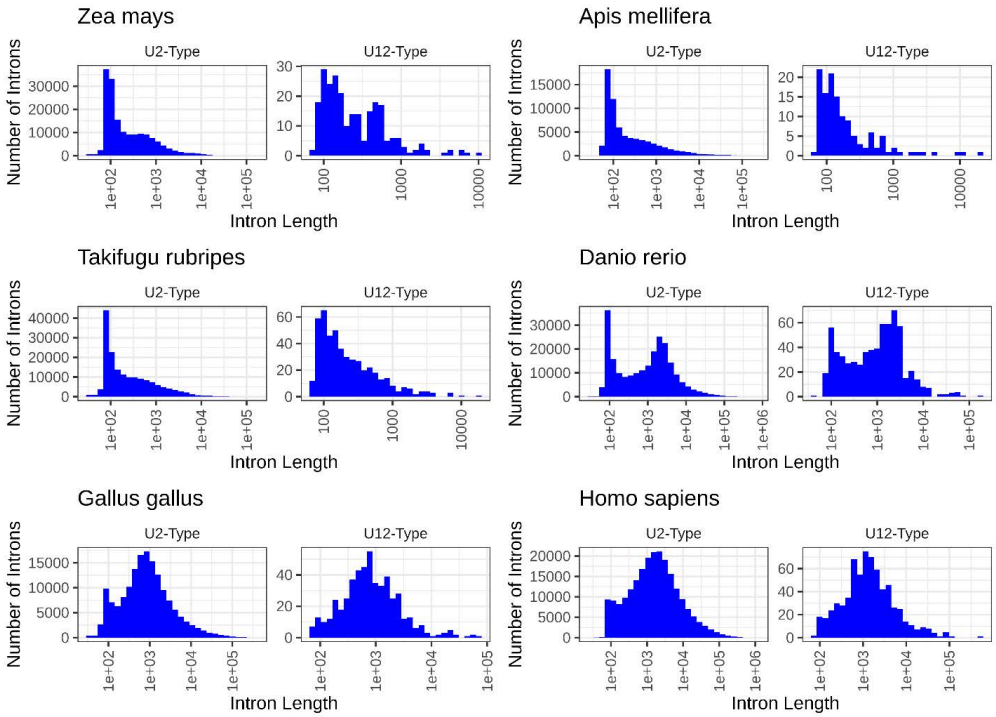


Figure 2.7: Distributions of intron lengths in both classes of intron in six of the genomes annotated in the IAOD. The x-axis of each plot is a log scale.

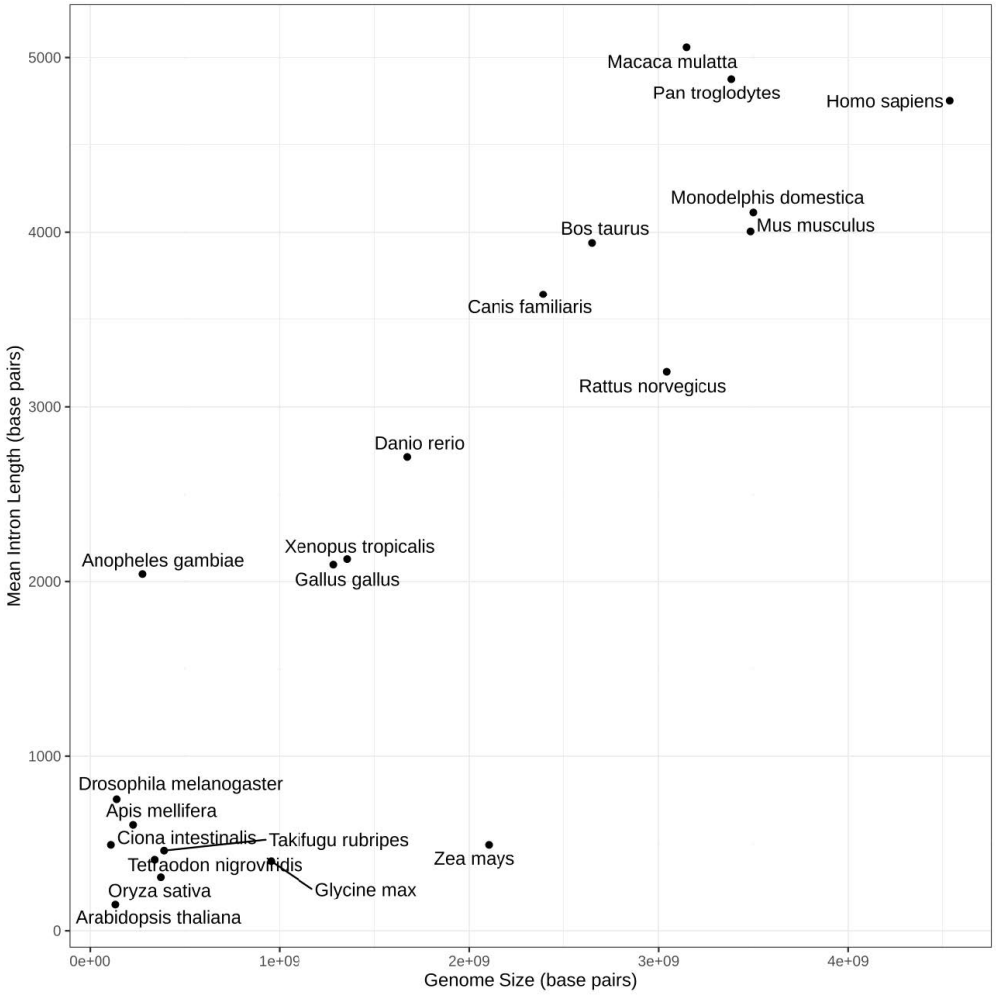


Figure 2.8: Relationship of genome size and mean U12-type intron length in genomes annotated in the IAOD. *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans* are not shown in this figure as they lack U12-type introns.

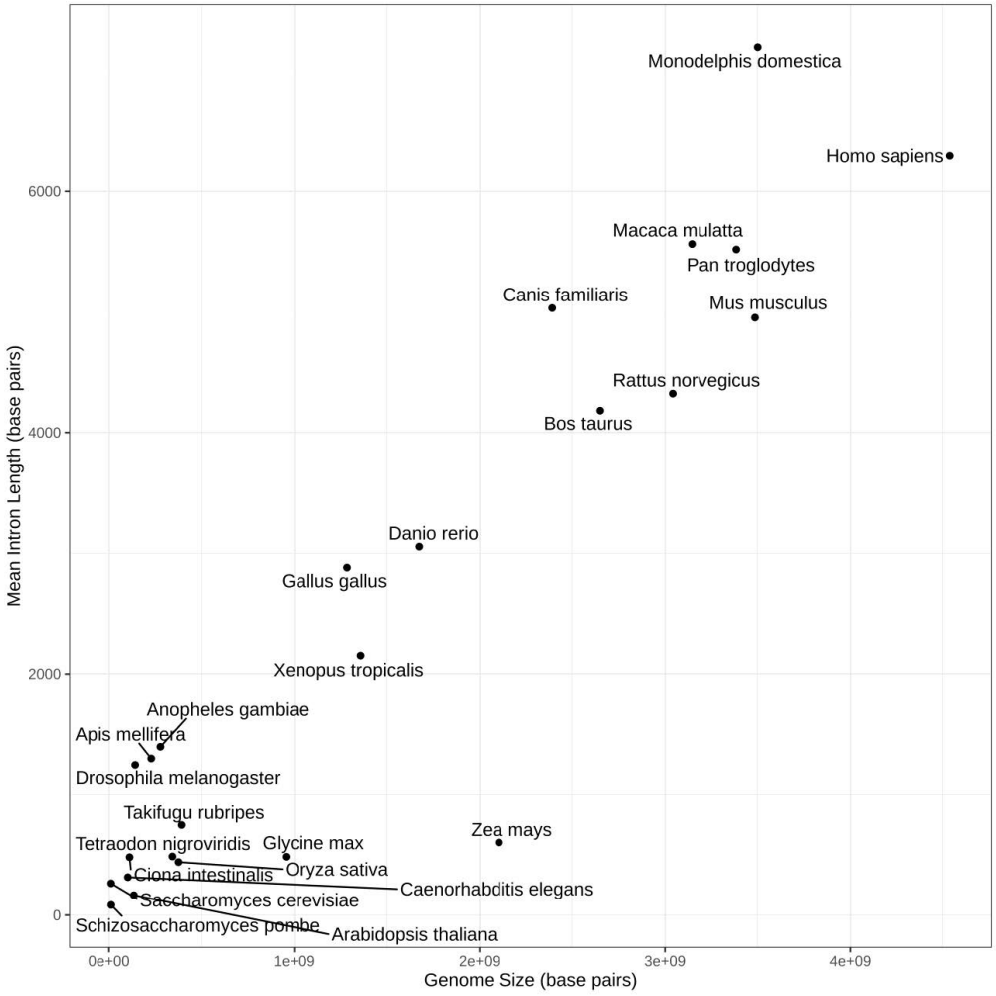


Figure 2.9: Relationship of genome size and mean U2-type intron length in genomes annotated in the IAOD.

Chapter 3

Expansion and transformation of the minor spliceosomal system in the slime mold *Physarum polycephalum*

3.1 Prior publication note

A version of this chapter of the dissertation has been published in *Current Biology*:

> Larue, G. E., Eliáš, M. & Roy, S. W. Expansion and transformation of the minor spliceosomal system in the slime mold *Physarum polycephalum*. *Curr. Biol.* 31, 3125–3131.e4 (2021)

3.2 Abstract

Spliceosomal introns interrupt nuclear genes and are removed from RNA transcripts ("spliced") by machinery called spliceosomes. While the vast majority of spliceosomal introns are removed by the so-called major (or "U2") spliceosome, diverse eukaryotes also contain a rare second form, the minor ("U12") spliceosome, and associated ("U12-type") introns (Jackson, 1991; S. L. Hall and R A

Padgett, 1994; S. L. Hall and R A Padgett, 1996). In all characterized species, U12-type introns are distinguished by several features, including being rare in the genome (0.5% of all introns) (Turunen, Niemelä, et al., 2013; Alioto, 2007; Sheth et al., 2006), containing extended evolutionary-conserved splicing sites (Turunen, Niemelä, et al., 2013; Alioto, 2007; C.-F. Lin et al., 2010; Moyer et al., 2020), being generally ancient (Russell et al., 2006; C. B. Burge, R A Padgett, and Sharp, 1998) and being inefficiently spliced (Niemelä and Mikko J Frilander, 2014; Patel, McCarthy, and Joan A Steitz, 2002; Younis et al., 2013). Here, we report a remarkable exception in the slime mold *Physarum polycephalum*. The *P. polycephalum* genome contains > 20,000 U12-type introns—25 times more than any other species—enriched in a diversity of non-canonical splice boundaries as well as transformed splicing signals that appear to have co-evolved with the spliceosome due to massive gain of efficiently spliced U12-type introns. These results reveal an unappreciated dynamism of minor spliceosomal introns and spliceosomal introns in general.

3.3 Results

3.3.1 U12-type intron enrichment in *Physarum*

During manual annotation of GTPase genes in the genome of the slime mold *Physarum polycephalum*, we observed several introns lacking typical GT/C-AG boundaries, including both AT-AC and non-canonical introns (i.e., neither G[T/C]-AG nor AT-AC; Figure 3.1A). Most of these atypical introns also contained extended U12-like 5′SS motifs ([G/A]TATC[C/T]TTT), consistent with previous evidence of U12 splicing in this species (Bartschat and Samuelsson, 2010; M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008). However, genome-wide analysis of the current *P. polycephalum* genome annotation (Schaap et al., 2015; Glöckner and Marwan, 2017) revealed that all annotated introns have GY-AG boundaries, a pattern suggesting non-GY-AG introns may have been discarded by the annotation pipeline (Schaap et al., 2015; Stanke et al., 2008). Indeed, an RNA-seq based genome reannotation combining de novo transcriptome assembly, spliced

transcript alignment and ab initio annotation steps while explicitly allowing for non-GY-AG introns (section 3.9) improved overall annotation quality (73.3% versus 60.1% `BUSCO` (Simão et al., 2015) broadly-conserved gene sets present), and revealed a large number of previously unannotated introns, including a substantial number of introns with AT-AC splice boundaries (1,830 AT-AC, 54,816 GY-AG).

Our updated *P. polycephalum* annotation contains 3,648 introns with perfect matches to the canonical U12-type 5′SS motif (3,021 with GTATCCTT, 627 with ATATCCTT). In contrast, far fewer introns exhibit the classic U12-type BPS motif (561 with CCTT[G/A]AC present in the last 45 bases out of all introns, and only 20 of the 3,648 introns with perfect U12-type 5′SS motifs), and standard position weight matrix (PWM) methods (following the general methods of (Sheth et al., 2006; Moyer et al., 2020; Bartschat and Samuelsson, 2010; C. Burge and Sharp, 1997)) failed to clearly identify U12-type introns (section 3.9 and Figure 3.4A). Lack of classic U12-type branchpoints were confirmed for a subset of conserved U12-type introns (those with U12-like 5′SS motifs found at positions that match those of U12-type introns in other species (section 3.9). Instead, we noted the motif TTTGA falling within a short region near the 3′SS (terminal A 9-12 bp upstream of splice site) in many of these introns, a feature also common in the manually identified non-GY-AG introns (Figure 3.1A). Genome-wide analysis of the 5′SS and TTTGA motifs showed a clear correspondence: TTTGA motifs are present 9-12 bp upstream of the 3′SS in 59% (41/70) of conserved U12-type introns, as well as 42% (3,107/7,462) of GTATCYTT-AG introns and 67% (417/625) of ATATCYTT-AC introns, but only 6% (10,313/167,111) of other introns (Figure 3.1B). Consistent with a functional role in splicing, among introns with U12-like 5′ splice sites, introns containing the TTTGA motif had lower average retention than those without it (Figure 3.4B). Interestingly, the TTTGA motif was particularly enriched in non-canonical introns with otherwise U12-like motifs, perhaps reflecting a compensatory response to ensure recognition by the minor splicing machinery despite altered terminal dinucleotides in these introns (Supplementary Figure 3.7).

Combining this position-specific atypical branchpoint motif with species-specific

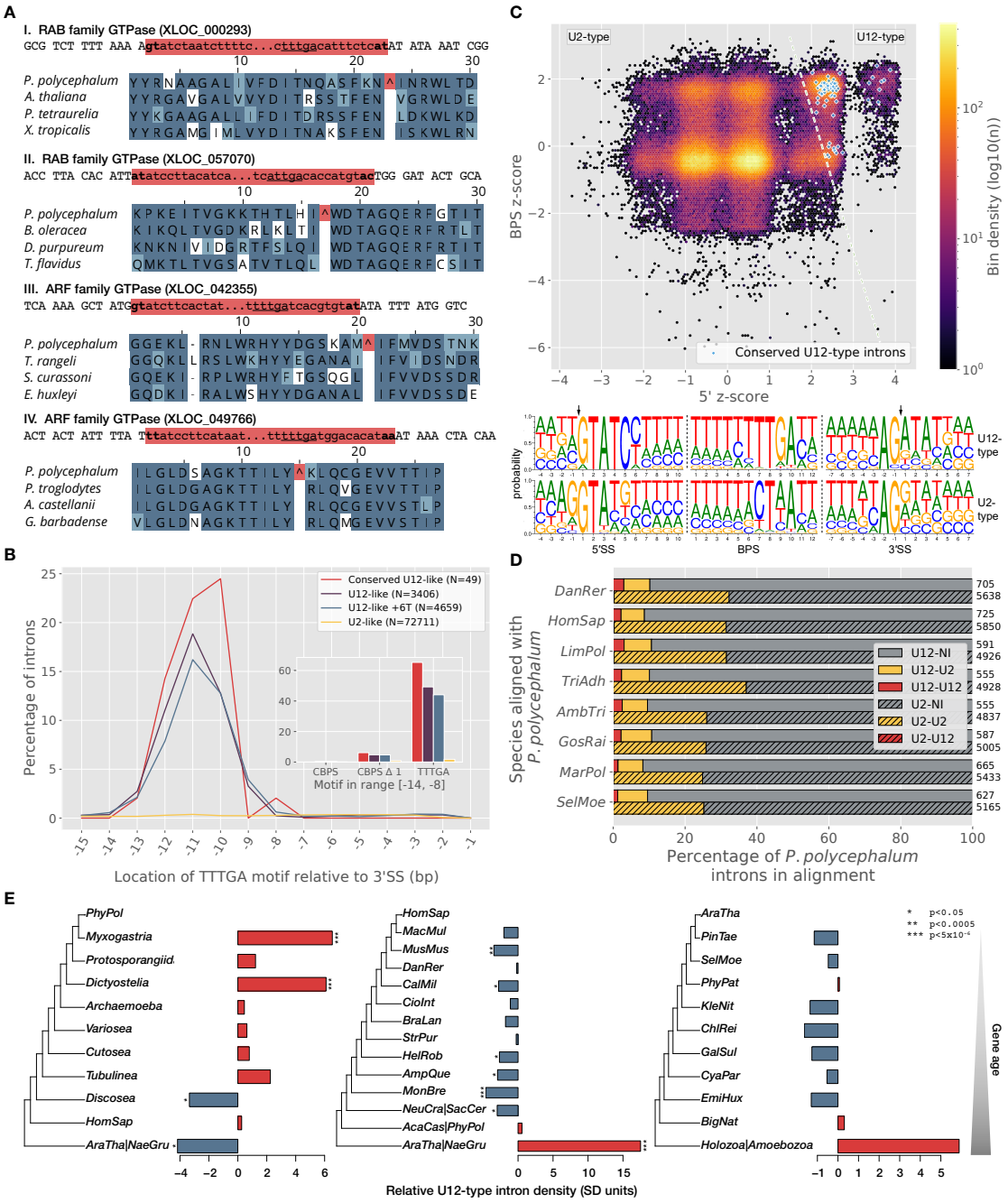


Figure 3.1: Evidence of massive U12-type intron gain in *Physarum polycephalum*. (A) Canonical and non-canonical U12-like introns in conserved *P. polycephalum* GTPase genes. Intron positions in alignments represented by carets (^). Lowercase red characters indicate intron sequence, with terminal dinucleotides in bold and putative BPS motifs underlined. (Caption continued on next page)

Figure 3.1: (Continued from previous page) (B) Presence of BPS motif in various groups of *P. polycephalum* introns. (Main) Occurrence of TTTGA motif as a function of number of nucleotides upstream of the 3′SS), for U12-like ([AG]TATCCTT-A[CG] or [AG]TATCTTT-A[CG] splice sites for "U12-like" and "U12-like +6T", respectively), U2-like (GTNNG-AG), and conserved U12-like ([AG]TATC[CT]-NN and conserved as a U12-type intron in another species). (Inset) The same data as a cumulative bar plot for positions -14 through -8. See also Figure 3.4B. (C) Intron type classification and associated motifs. The main plot shows BPS-vs-5′SS log-ratio z-scores for all *P. polycephalum* introns, with conserved U12-type introns highlighted in blue. The dashed green line indicates the approximate U2-U12 score boundary (section 3.9, see also Figure 3.4C). Below the scatter plot are sequence logos showing motif differences between the two groups (top, U12-type, n = 20,899; bottom, U2-type, n = 154,299). (D) Conservation status of *P. polycephalum* introns in other species, showing substantially lower U12- than U2-type conservation. For each species, the pair of bars shows the fractions of *P. polycephalum* introns of each intron type (U12-type, un-hashed; U2-type, hashed) that are conserved as either U12-type (red) or U2-type (yellow) introns, or not conserved (gray). Total numbers of *P. polycephalum* introns assessed are given at right. (E) Comparison of U12-type intron density (fraction of introns that are U12-type) in genes of different age categories for *P. polycephalum* (PhyPol), *Homo sapiens* (HomSap) and *Arabidopsis thaliana* (AraTha), relative to expectation (blue/red = below/above expectation). U12-type intron densities in *P. polycephalum* are significantly overrepresented in newer genes, in contrast to the pattern seen in both human and *Arabidopsis*. Significance assessed by Fisher's exact tests corrected for multiple testing using the Holm step-down method. Full species names listed in Table S1, `https://doi.org/10.6084/m9.figshare.20483790`.

splice site motifs in intronIC (Moyer et al., 2020) (section 3.9) led to a clearer separation of putative U12- and U2-type introns (Figures 1C, S1C). Using a conservative criterion of 95% U12-type probability (section 3.9), we identified 20,899 putative U12-type introns in *P. polycephalum* (leaving 154,299 putative U2-type introns with U12-type scores $\leq$95%), representing 11.9% of all 175,198 annotated introns and 25 times more than has been observed in any other species. The true U12-type nature of these introns was further supported by two additional findings. First, comparisons of 8,267 pairs of *P. polycephalum* paralogs showed strong conservation of U12-type character: among intron positions shared between paralogs, an intron was 34-45 times more likely to be predicted to be U12-type if its paralogous intron was predicted to be U12-type (section 3.9, Figure 3.4D). Second, putative *P. polycephalum* U12-type introns as a group are strongly biased away from phase

0 (26% compared with 39% for U2-type introns; phase is not part of the scoring process), consistent with the phase bias observed in other species (C. B. Burge, R A Padgett, and Sharp, 1998; Levine and Durbin, 2001) (Figure 3.5).

3.3.2 Evolution of *Physarum* U12-type introns

To investigate the evolutionary dynamics of U12-type introns in *P. polycephalum*, we performed multiple-sequence alignments of *P. polycephalum* genes with their orthologs in a variety of species, which allowed us to characterize the conservation status of the associated introns (Moyer et al., 2020; Scott W Roy, Fedorov, and Walter Gilbert, 2003). Interestingly, very few *P. polycephalum* U12-type intron positions in conserved coding regions are shared with distantly-related species (e.g., only 9% of *P. polycephalum* U12-type introns are found as either U2- or U12-type introns in humans, far fewer than the 31% of U2-type intron positions so-conserved; Figure 3.1D), indicating either massive U12-type intron gain in *P. polycephalum* or commensurate loss in other species. There is, however, no evidence for widespread loss of U12-type introns in other species, and previous results have attested to significant U12-type intron conservation across long evolutionary distances (C.-F. Lin et al., 2010; Moyer et al., 2020; Basu, Makalowski, et al., 2008). Indeed, among U12-type introns conserved between *P. polycephalum* and plants and/or animals (i.e., ancestral U12-type introns), 63% are retained as either U2- or U12-type in the variosean amoeba *Protostelium aurantium*, and 70% are similarly retained in the discosean *Acanthamoeba castellanii* (Figure 3.5B).

That *P. polycephalum* has recently gained many U12-type introns is also supported by the fact that putatively recently evolved *P. polycephalum* genes (i.e., those lacking homology to genes outside of those in closely related species) show substantial U12-type intron densities (Figure 3.1E). This finding is not expected from retention of ancestral U12-type introns and is in clear contrast to the low U12-type intron densities in young human and plant genes (Figure 3.1E). In those species, the oldest category of genes (those whose conservation across deeply-diverged eukaryotes suggests their presence in the last common ancestor of extant eukaryotes) has dramatically elevated U12-type intron densities, consistent with a

substantial fraction of plant and animal U12-type introns dating to early eukaryotic evolution; by contrast, *P. polycephalum* shows a very different pattern, with the oldest class of genes instead containing lower densities of U12-type introns. For instance, the oldest classes of genes in human and *Arabidopsis* show overrepresentations in U12-type intron densities of 144% and 56%, respectively (p-values 3.8×10^{-18} and 0.143, Fisher's exact test), whereas in *P. polycephalum*, the oldest class shows a 4.15% underrepresentation ($p = 0.038$).

3.3.3 Features of the U12 system in *Physarum*

Analysis of the highly expanded U12 spliceosomal system in *P. polycephalum* revealed a variety of other surprising characteristics. In contrast to the remarkable consistency of splice sites in most eukaryotic genomes (e.g., 99.85% GY-AG or AT-AC in human), we found many noncanonical introns in *P. polycephalum* (section 3.9). After filtering for likely reverse transcriptase artifacts (section 3.9), 1,425 introns (0.8% of all introns) had non-canonical terminal dinucleotides. Remarkably, 71% (1,014/1,425) of non-canonical introns were classified as either confident (60%) or likely (11%) U12-type introns (Figures 3.2A, 3.5C). These non-canonical U12-type introns were dominated by boundary pairs with a single difference from canonical pairs, in particular AT-AG (29%), AT-AA (27%), GT-AT (17%), and AT-AT (8%). As with the canonical U12-type introns described earlier in this paper, the intronic and U12-type character of these introns was supported by conservation across *P. polycephalum* paralogs (section 3.9, Figure 3.5D-E).

We also scrutinized components of the U12 spliceosome in *P. polycephalum*. A genomic search using Infernal (Nawrocki and Eddy, 2013) revealed a single candidate for the U12 snRNA (as previously reported in (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008)), the component which basepairs with the branchpoint. Strikingly, this sequence exhibits two transition mutations relative to the core branchpoint binding motif (underlined): GCAAAGAA, which produce basepairing potential with the putative TTTGA branchpoint with a bulged A, comparable to the canonical structure (Figure 3.2B). This apparent complementary evolution of core U12 spliceosomal machinery and branchpoint sequence represents a rare

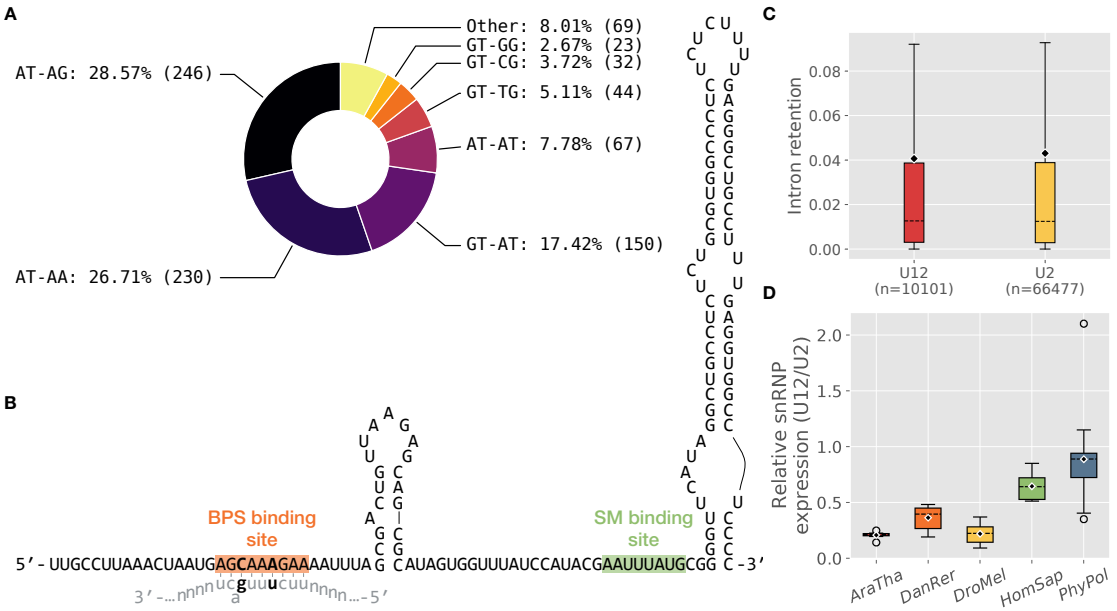


Figure 3.2: Transformed features of the *P. polycephalum* minor splicing system. (A) Non-canonical U12-type intron splice boundaries. (B) Putative U12 snRNA sequence and secondary structure (structure based on (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008)). Highlighted are the BPS binding site (orange), SM binding site (green) and the consensus intronic branchpoint motif (lowercase). The BPS binding site contains two changes relative to the canonical U12 snRNA (bold) which exactly complement changes in the putative TTTGA BPS motif relative to the canonical motif (also bold). (C) Comparison of average (mean) intron retention in RNA-seq data for U12-type and U2-type introns. In contrast to mammals (Figure 3.6B), average intron retention of U12-type introns is not higher than that of U2-type introns in *P. polycephalum* ($p = 0.89$, Mann-Whitney U test). (D) Increased expression of the U12 spliceosome in *P. polycephalum*. The average (mean) expression of U12 spliceosomal components, relative to U2 spliceosomal components, is significantly higher in *P. polycephalum* than other species (section 3.9). For both C and D, dashed line = median, diamond = mean, whiskers = 1.5 IQR.

instance of coevolution of complementary changes in core intronic splicing motifs and core spliceosomal snRNAs. Length distributions for the two intron types are very similar (the U2-type distribution has a longer, narrow tail; data not shown), and the median lengths are almost identical (251 bp for U12-type, 252 bp for U2-type). Interestingly, we find that the median U12-type intron position within transcripts is skewed slightly toward the 3′ end when compared to U2-type introns (median of 50.8% as a fraction of coding sequence for U12-type vs 46.8% for U2-type, $p = 8.6 \times 10^{-46}$, Kruskal-Wallis H test), which runs opposite the pattern reported in ancestral U12-type introns shared between human and plants (Basu, Makalowski, et al., 2008) as well as in the vast majority of vertebrates, invertebrates and plants (where generally no significant difference is found and otherwise, U12-type introns are usually 5′-biased; GEL and SWR, unpublished data). This inverted positional bias is, however, consistent with relatively young U12-type introns in *P. polycephalum* having inserted preferentially into 3′ regions, perhaps due to those regions' lower density of older introns (K. Lin and D.-Y. Zhang, 2005; Scott W Roy and Walter Gilbert, 2005b).

U12-type introns in other species have been reported to have lower splicing efficiency than U2-type introns (Patel, McCarthy, and Joan A Steitz, 2002; Younis et al., 2013; Niemelä, Oghabian, et al., 2014; Pessa, Ruokolainen, and Mikko J Frilander, 2006; Woan Y Tarn and J. a. Steitz, 1996); in *P. polycephalum*, inefficient splicing of such a large number of introns would appear to pose a substantial cost, raising the question of how its genome copes with ubiquitous U12-type introns. To investigate, we used RNA-seq and IRFinder (Middleton et al., 2017) to calculate intron retention levels, as well as estimating splicing efficiency by comparing fractions of spliced and unspliced junction support between U2- and U12-type introns (section 3.9). Surprisingly, U12-type introns in *P. polycephalum* show slightly lower average intron retention (and higher average splicing efficiency) when compared to U2-type introns either en masse (Figures 3.2C, 3.6A-B) or in matched pairwise comparisons with neighboring introns from the same genes (Figure 3.6C). Consistent with increased efficiency of U12-type splicing in this lineage, we also found that the difference in average expression between the U12 and U2 spliceosomal

components was smaller in *P. polycephalum* than is the case in species with lower U12-type intron densities (Figure 3.2D). These data raise the possibility that minor spliceosomal kinetics are not inherently inefficient, and suggest that minor intron splicing may have been optimized in *P. polycephalum* in concert with and/or in response to the spread of U12-type introns. While additional work is needed to support this hypothesis, if true it raises interesting questions about the processes governing the less efficient splicing of U12-type introns in other species.

3.3.4 U12-type intron creation in *Physarum*

The near absence of U12-type intron creation in most lineages has been argued to reflect the low a priori likelihood of random appearance of the strict U12-type splicing motifs at a given locus (C. B. Burge, R A Padgett, and Sharp, 1998; Dietrich, Incorvaia, and Richard A Padgett, 1997). How, then, did *P. polycephalum* acquire so many U12-type introns? Inspired by cases of U2-type intron creation by insertion of DNA transposable elements (Huff, Zilberman, and Scott W Roy, 2016; Henriet et al., 2019) as well as a number of other recent reports of intron gain (Gumińska et al., 2018; Milanowski et al., 2016), we scrutinized U12-type splice sites in *P. polycephalum*. We observed that many *P. polycephalum* U12-type introns carry sequences that resemble the signature of DNA transposable elements, namely inverted repeats (rtatcttt...aaagATAT) flanked by a direct repeat of a TA motif. This suggests the possibility that *P. polycephalum* U12-type introns could have been created by a novel DNA transposable element with TCTTT-AAAGA termini and a TA insertion site (Figure 3.3). It is of note that *P. polycephalum* U12-type introns differ at two sites from the corresponding classic motif (TCCTT-NYAGA), where both changes increase the repeat character. An ancestral decrease in the length and stringency of the branchpoint motif could have increased the probability of de novo evolution of a DNA transposable element carrying sufficiently U12-like splice sites for new insertions to be recognized by the U12 spliceosome.

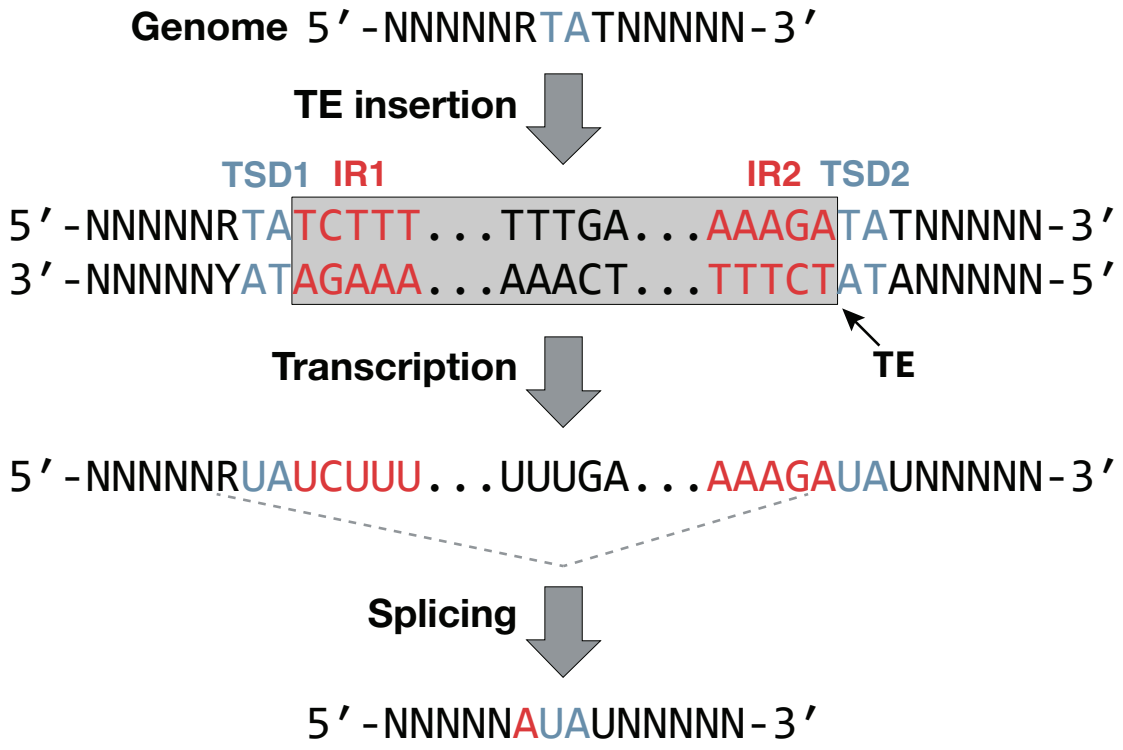


Figure 3.3: Proposed mechanism for transposon-driven creation of U12-type introns in *P. polycephalum*. Insertion of a transposable element (TE, gray box) carrying inverted repeats (IR1/IR2, red) leads to duplication of a TA target side (TSD1/TSD2, blue). Splicing at RT-AG boundaries leads to a spliced transcript with a sequence identical or nearly identical to the initial gene sequence with loss of an R (G/A) nucleotide and gain of the 3′ A from the TE, maintaining the original reading frame.

3.4 Discussion

Over the nearly two decades since the surprising discovery of the existence of the U12 spliceosomal system (Jackson, 1991; S. L. Hall and R A Padgett, 1994), U12-type introns have consistently been defined by a number of hallmark characteristics distinct from their U2-type counterparts. First, in all lineages examined U12-type introns are either rare or absent, ranging from 700 (0.36% of all introns) in humans (Alioto, 2007; Sheth et al., 2006; Moyer et al., 2020) to 19 (0.05%) in fruitflies (Janice et al., 2012) to complete absence in diverse lineages (Turunen, Niemelä, et al., 2013; M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008). Second, U12-type introns show distinct extended splicing motifs at the 5′ splice site (5′SS) ([G/A]TATCCTT) and branchpoint sequence (BPS) (TTCCTT[G/A]AC, $\lesssim$ 45 bases from the 3′ splice site (3′SS)) which exactly basepair with complementary stretches of core non-coding RNAs in the splicing machinery (S. L. Hall and R A Padgett, 1996; W Y Tarn and J A Steitz, 1996; Pineda and Bradley, 2018). Third, U12-type introns are typically ancient (e.g., 94% of human U12-type introns are conserved as U12-type in chicken [7]), implying low rates of U12-type intron creation through evolution (Turunen, Niemelä, et al., 2013; C.-F. Lin et al., 2010; Russell et al., 2006; W. Zhu and Brendel, 2003). Finally, U12-type introns show slow rates of splicing, suggesting inherently low efficiency of the U12 spliceosomal reaction (Niemelä and Mikko J Frilander, 2014; Patel, McCarthy, and Joan A Steitz, 2002; Younis et al., 2013; Singh and Richard A Padgett, 2009). In contrast to this portrait of the U12 spliceosomal system as rare, ancient, static and suboptimal, the results presented here expand our understanding of U12 diversity, by (i) increasing the upper bound of U12-type intron density per species by two orders of magnitude; (ii) showing that U12-type introns have been gained en masse through eukaryotic evolution; and (iii) showing that U12-type splicing is not necessarily less efficient than U2-type splicing. *P. polycephalum* provides promise for an understanding of the flexibility of U12 splicing, a potentially important role given the increasing appreciation of U12 splicing errors in development and human disease. In addition, the availability of *P. polycephalum* and related species (once additional genomic data becomes available) as models for studying the evolution of U12-type

introns represents an exciting opportunity to examine the mechanisms by which new U12-type introns are created, potentially shedding light both on the origins of U12-type introns in very early eukaryotes as well as their broader functional roles and implications across the tree of life.

Our results inform a number of remaining questions about U12-type introns and spliceosomal introns in general. First, our finding that peculiarities of U12-type intron phase extend to newly-created introns suggests that this pattern could reflect biases in the process of initial U12-type intron creation (rather than secondary differential intron loss or conversion to U2-type), as has been reported elsewhere for U2-type introns (Huff, Zilberman, and Scott W Roy, 2016; Sverdlov et al., 2004; Scott William Roy, Gozashti, et al., 2020). Second, several of our results raise questions about the functional roles of U12-type introns. In particular, biases in the genic distribution of U12-type introns (Moyer et al., 2020; C. B. Burge, R A Padgett, and Sharp, 1998), evidence for regulation of cell cycle genes by the U12 spliceosome (Baumgartner, Olthof, et al., 2018) and evidence for regulation of U12 splicing in differentiation (Younis et al., 2013; Meinke et al., 2020) all suggest distinct functional roles; the presence of U12-type introns in such a large fraction of *P. polycephalum* genes complicates this pattern. Whether this reflects more specialized regulation of subsets of U12-type introns in *P. polycephalum* or restriction of U12-specific functions to a subset of U12-type intron containing lineages remains to be determined. Finally, another unanswered question involves how the cell has accommodated the invasion of tens of thousands of introns of a type thought to be ancestrally inefficiently spliced. Timing the various transformations described herein (intron invasion, changes in splicing motifs, changes in splicing efficiencies) through the study of related species should help to shed further light on this dynamic evolutionary history.

3.5 Acknowledgements

G.E.L. and S.W.R. were supported by the National Science Foundation (award no. 1616878 to S.W.R.). M.E. was supported by the Czech Science Foundation

project no. 18-18699S and the project "CePaViP" (CZ.02.1.01/0.0/0.0/16_019/0000759) provided by ERD Funds.

3.6 Author contributions

M.E. supplied initial motivating data, G.E.L. and S.W.R. conceived of the study, G.E.L. performed computational analysis, data processing and visualization, G.E.L. and S.W.R. analyzed and interpreted the results, G.E.L and S.W.R. wrote the manuscript with input from M.E..

3.7 Declaration of interests

The authors declare no competing interests.

3.8 Resource availability

3.8.1 Lead Contact

Any requests for additional data/resources related to this paper should be addressed to the Lead Contact, Scott W. Roy (scottwroy@gmail.com).

3.8.2 Data and Code Availability

The *Physarum polycephalum* genome and annotation file used in our analyses are available in the following Zenondo archive: `https://doi.org/10.5281/zenodo.4086119`. Intron coordinates and U12-type probability scores for all *P. polycephalum* introns in our annotation have been archived here: `https://doi.org/10.5281/zenodo.4099156`. The modified version of `intronIC` used herein for classifying introns in *P. polycephalum* has been archived at `https://doi.org/10.5281/zenodo.4265109`, and the standard version of `intronIC` used for all additional species is open-source and available on GitHub: `https://www.github.com/glarue/intronIC`.

3.8.3 Experimental model and subject details

Genome and annotation files used in this study were downloaded from a variety of publicly-available resources including Ensembl, RefSeq, GenBank and JGI as well as a number of other taxa-specific sources (Table S1, `https://doi.org/10.6084/m9.figshare.20483790`). Annotated coding sequences were extracted from each genome using custom Python software (Materials and Methods). RNA-seq samples for all species were downloaded from the NCBI SRA database.

3.9 Materials and Methods

3.9.1 Reannotation of the *P. polycephalum* genome

We downloaded the *Physarum polycephalum* genome assembly and annotation from `http://www.physarum-blast.ovgu.de/`, and RNA-seq from NCBI's SRA database (accession numbers DRR047256, ERR089824-ERR089827, ERR557103-ERR557120) (Schaap et al., 2015; Glöckner and Marwan, 2017; Glöckner, Golderer, et al., 2008; Barrantes, Leipzig, and Marwan, 2012). To reannotate the genome, we combined several *de novo* and reference-based approaches. First, we generated a *de novo* transcriptome from the aggregate RNA-seq data using `Trinity` (Grabherr et al., 2011) (v2.5.1). We also separately mapped the reads to the genome using `HISAT2` (Kim et al., 2019) (v2.1.0), allowing for non-canonical splice sites (`-pen-noncansplice 0`), followed by `StringTie` (M. Pertea et al., 2016) (v1.3.3) to incorporate the mapped reads with the existing annotations and generate additional putative transcript structures.

Coding-sequence annotations for the assembled transcripts, informed by additional homology information from the SwissProt (UniProt Consortium, 2008) protein database, were generated using `TransDecoder` (Brian J Haas et al., 2013) (v5.0.2), and further refined with the *de novo* transcriptome via `PASA` (B J Haas, 2003) (v2.2.0). In addition, an `AUGUSTUS` (Stanke et al., 2008) (v3.3) annotation was generated ab initio from the mapped reads using `BRAKER1` (Hoff et al., 2015) (v2.1.0) explicitly allowing for AT-AC splice boundaries (`--allow_hinted_splic`

esites=atac). Lastly, the AUGUSTUS- and StringTie-based gene predictions were merged using gffcompare (G. Pertea, n.d.) (v0.10.5), and processed again using TransDecoder. To gauge the quality of our annotations compared to those previously available, we performed a BUSCO (Simão et al., 2015) (v3.0.1) analysis against conserved eukaryotic genes. Where the previous annotations contained matches to 60.1% of eukaryotic BUSCO groups (54.5% single-copy; 27.1% fragmented; 12.8% missing), our annotation increased this percentage to 73.3% (64.4% single-copy; 18.5% fragmented; 8.2% missing).

3.9.2 Classification of intron types

All annotated intron sequences from our updated *P. polycephalum* genome annotation were collected and analyzed using a modified version of intronIC (Moyer et al., 2020). Briefly, we first obtained high-confidence sets of U12- and U2-type *P. polycephalum* introns as follows: High-confidence U2-type introns were defined as introns classified as U2-type under default settings and conserved as U2-type in at least three other species. Due to the low evolutionary conservation of putative *P. polycephalum* U12-type introns, the confident U12-type intron set was assembled from introns with U12-type probabilities $> 95\%$ conserved as U12-type in one or more species, introns with perfect 5′SS motifs ([GA]TATCCTT) interrupting coding sequences in regions of good alignment to orthologs in one or more species, introns with near-perfect 5′SS motifs in addition to the TTTGA BPS motif 10-12 bp upstream of the 3′SS, and AT-AC introns (less likely to be false positives) with strong 5′SS consensus motifs in conserved eukaryotic genes (defined as genes with BUSCO matches).

Sub-sequences of each intron corresponding to the 5′SS (from -3 to +8, where +1 is the first intronic base) and all 12mers within the branchpoint region (-45 to -5 where -1 is the last intronic base) were scored against position-weight matrices (PWMs) derived from the sets of high-confidence *P. polycephalum* U2- and U12-type introns to obtain U12/U2 log ratio scores for each motif. These log ratios were normalized to z-scores for each motif (5′SS and BPS) and used to construct two-dimensional vector representations of each intron's score. In addition, to account

for the narrow window of occurrence of the non-canonical TTTGA BPS, intronIC was modified to weight the branchpoint scores of introns whose BPS adenosines were found within the range [-12, -10] of the 3′SS, with the additional weight equal to the frequency of occurrence of the BPS adenosine at the same position within confident U12-type introns. Finally, except where explicitly stated otherwise, we used a more conservative U12-type probability score of 95% for classifying introns in *P. polycephalum* (versus intronIC's default U12-type classification threshold of 90%, used for all other species). The prominently separated "cloud" in the upper-right corner of Figure 3.1D is composed mainly of AT-AC U12-type introns, whose 5′SS scores are more distinct than U12-type introns with other splice boundaries.

3.9.3 Identification of homologs and conserved introns

Genomes and annotations for all additional species were downloaded from various online resources (Table S1, `https://doi.org/10.6084/m9.figshare.20483790`), and in cases where sufficient RNA-seq was available and we suspected that U12-type introns had been systematically suppressed (e.g., zero or very few AT-AC introns annotated), we performed RNA-seq based annotation updates using `Trinity` and `PASA` (Grabherr et al., 2011; Brian J Haas et al., 2013). For each genome, annotated coding sequences were extracted and translated via a custom Python script (`https://github.com/glarue/cdseq`). Annotated intron sequences were collected and scored using `intronIC` (Moyer et al., 2020) with default settings. Under these settings, only introns defined by CDS features from the longest isoform of each gene were included, and introns with U12-type probability scores $> 90\%$ ($> 95\%$ for *P. polycephalum*) were classified as U12-type. Furthermore, introns shorter than 30 nt and/or introns with ambiguous (`N`) characters within scored motif regions were excluded.

Between *P. polycephalum* and each other species (or, in the case of paralogs, itself), we performed pairwise reciprocal `BLASTP` (Altschul et al., 1990; Camacho et al., 2009) (v2.6.0+) searches (E-value cutoff of 1×10^{-10}), and parsed the results to retrieve reciprocal best-hit pairs (defined by bitscore) using a custom Python script (`https://github.com/glarue/reciprologs`). Pairs of homologous

sequences were globally aligned at the protein level using `ClustalW` (Larkin et al., 2007) (v2.1), and introns occurring at the same position in regions of good local alignment ($\geq$4/10 shared amino acid residues on both sides of the intron) were considered to be conserved (based on the approach in (Scott W Roy, Fedorov, and Walter Gilbert, 2003)).

3.9.4 Calculation of dS values between paralogs

We identified 8267 pairs of paralogs in *P. polycephalum* using the same approach as for other homologs. Each pair sharing at least one intron position was globally aligned at the protein level using `Clustal Omega` (Sievers and Higgins, 2014) (v1.2.4), and then back-translated to the original nucleotide sequence using a custom Python script. Maximum likelihood dS values for each aligned sequence pair were computed using `PAML` (Yang, 2007) (v4.9e) (`runmode=-2,seqtype=1,model=0`), with dS values greater than 3 treated as equal to 3 in subsequent analyses (as dS values > 3 are not meaningfully differentiable in this context) (Figure 3.4D).

3.9.5 Relative gene ages

For the three focal species (FS) *P. polycephalum*, human and *Arabidopsis thaliana*, sets of node-defining species (NDS) were selected to represent a range of evolutionary distances from the FS based on established phylogenetic relationships. In the case of *P. polycephalum*, we used data from Kang et al. (2017) and their amoebozoan phylogeny; for the other two FS, we downloaded corresponding NDS genomes and annotation files from a combination of the publicly-available resources Ensembl, JGI and NCBI (Table S1, `https://doi.org/10.6084/m9.figshare.20483790`). We then performed one-way `BLASTP` (Altschul et al., 1990; Camacho et al., 2009) (v2.8.0+) searches (E-value cutoff 1×10^{-10}) of each FS transcriptome against the transcriptomes of its NDS set to establish an oldest node for each gene, defined as the ancestral node of the FS and the most-distantly-related NDS where one or more `BLASTP` hits to the gene were found. For example, a human gene would be assigned to the human-Danio rerio ancestral node if a `BLASTP` hit to the gene

were found in *Danio rerio* (and optionally, any more closely related NDS) but not in any other more distantly related NDS.

Once gene ages were assigned, for each FS we examined the difference of the observed and expected number of U12-type introns at each node using an expected value based on the aggregate density of U12-type introns in all other nodes, and scaled the observed-minus-expected value by dividing by the node's expected standard deviation, which we calculated as follows: For a given node with n introns and expected U12-type intron frequency p (based on the aggregate frequency from all other nodes), per the binomial theorem the expected standard deviation $SD = \sqrt{np(1-p)}$. The significance of the observed numbers of U12-type introns at each node was calculated with a Fisher's exact test (`SciPy` (Virtanen et al., 2020) v1.5.2), and p-values were corrected for multiple-testing using the Holm step-down method in the Python library `statsmodels` (Seabold and Perktold, 2010) (v0.11.1).

3.9.6 Intron splicing efficiency and retention

For each annotated intron defined by CDS features from the longest isoform of each gene, splice junctions for the spliced (5′ exon + 3′ exon) and retained (5′ exon + intron, intron + 3′ exon) structures were created in silico using a custom Python script. RNA-seq reads (accession numbers listed in the reannotation section) were then mapped in single-end mode to the junction constructs using `Bowtie` v1.2.2 (Langmead, 2010) with parameter -m 1 to exclude multiply-mapped reads. Reads overlapping a junction by ≥5 nt were counted and corrected by the number of mappable positions on the associated junction construct. For each RNA-seq dataset, introns with no read support for the spliced form were excluded from the analysis, as were introns with no junctions supported by at least 10 reads. Efficiency was calculated as the ratio of splice-supporting read coverage (C_s) over the total read coverage, which is just C_s plus the average of the retention-supporting read coverage (C_r), expressed as a percentage, i.e., $\frac{C_s}{\frac{C_r}{2}+C_s} \cdot 100\%$. For each intron with sufficient junction support in at least two RNA-seq samples, splicing efficiency was then computed as the mean efficiency—weighted by the sum of read support for

the spliced/unspliced junctions—across all samples.

To help validate our splicing efficiency results, we also employed an established method to evaluate intron retention using the same RNA-seq data. We obtained intron retention values for all annotated *P. polycephalum* introns with `IRFinder` (Middleton et al., 2017) (v1.3.0), which produced an equivalent (inverted) pattern to our splicing efficiency metric (Figure 3.6A,B). Introns with `IRFinder` warnings of "LowSplicing" and "LowCover" were excluded.

3.9.7 Paralogous and non-canonical U12-type introns

Introns conserved across *P. polycephalum* paralogs were identified as described for homologous introns. We then examined all intron positions conserved between paralogs and tabulated the intron types at each position. To determine the relative likelihood of a given U12-type intron being conserved as U12-type across paralogs, we calculated the relative probability of an intron A being U12-type conditioned on its paralogous intron B being U12-type or U2-type as $\frac{P(A_{U12}|B_{U12})}{P(A_{U12}|B_{U2})}$, which results in a likelihood fold-increase of $(\frac{866}{1074})/(\frac{208}{8695}) \approx \frac{0.806}{0.024} \approx 34$. This value is most likely conservative, as decreasing the stringency of U12-type classification results in a further increase in the relative likelihood.

To avoid inclusion of spurious intron annotations representing artifacts of the RT-PCR process ("RTfacts", (Scott William Roy and Irimia, 2008)) in our non-canonical intron analysis (where, given that we are concerned with unusual introns, we wanted to be conservative to errors likely to generate non-canonical splice boundaries), we used a fairly simple heuristic to detect unexpectedly high similarities between extended sequences around the 5′ and 3′ splice sites (5′SS, 3′SS). For each intron, we considered regions of 24 bp centered around the 5′SS and 3′SS (12 bp from the exon and 12 bp from the intron in each case) and used a 12 bp sliding window to compare every 5′SS 12mer against every 3′SS 12mer. For each 12mer pair, we defined their pairwise similarity s as $s = \frac{1-d}{l}$, where d is the Hamming distance between the two strings and l is their length in bp (i.e., 12), and treated the highest value found as the overall similarity score. Introns with similarity scores $\geq$0.916 (corresponding to one mismatch between the pair of

splice-site 12mers) were considered possible RTfacts and were excluded (n=1,624, 0.93% of 175,198 total introns).

In our survey of non-canonical introns in *P. polycephalum*, we took advantage of the greater number of conserved U12-type intron positions within paralogs (versus with other species) to gauge support for non-canonical U12-type intron boundaries present in our annotations. Of the non-canonical U12-type introns found in regions of good alignment between paralogs, 66% (42/63) contained the U12-type BPS motif 9-12 bp upstream of the 3′SS; in the smaller set conserved as introns between paralogs, the same motif was present in 73% (22/30). The BPS motif enrichment within these introns supports their identity as genuine non-canonical U12-type, and the distribution of the most common boundaries found within paralogs is consistent with the broader set of non-canonical U12-type introns (Figure 3.5D).

3.9.8 Relative expression of snRNPs

Orthologs for components of the major and minor spliceosome (major: SF3a120/SAP114, SF3a60/SAP61, U1-70K, U1 A, U2 A'; minor: U11/U12 20K, 25K, 25K, 31K, 35K, and 65K) were identified via reciprocal `BLASTP` searches (as described in the section on ortholog identification) using the components' annotated human transcripts as queries (Table S2, `https://doi.org/10.6084/m9.figshare.20483790`). For each species, a series of RNA-seq samples (curated by size and wild-type status; Table S3, `https://doi.org/10.6084/m9.figshare.20483790`) were aligned to the coding sequences of all available components using `HISAT2`, and the output processed with `StringTie` using the `-A` option to obtain per-transcript TPM values. Then, for each RNA-seq run mean per-species TPM values for the U12- and U2-type components were compared to calculate the U12/U2 expression ratios shown in Figure 3.2D. An ANOVA test was performed on the aggregate group of ratios ($p = 8.8 \times 10^{-13}$) to justify further comparisons, followed by pairwise Mann-Whitney U tests between all combinations of ratios. The difference between *P. polycephalum* and every other species was significant at $p < 0.05$ following multiple-testing correction.

3.9.9 Quantification and statistical analysis

Details of the statistical methods used in this study including sample sizes, sub-setting criteria and statistical tests are given in either the figures/legends or in the corresponding Results or section 3.9 sections. General information about our statistical workflow follows.

Statistical analysis software and general practices

All statistical analyses were performed in Python 3, primarily using the `SciPy` (Virtanen et al., 2020) package. Figures were generated using `Matplotlib` (Hunter, 2007) (v3.1.1) apart from Figure 3.3, which was created using graphic design software, and Figure 3.1E which was manually assembled using output from the R package `phytools` (Revell, 2012) (v0.7.47) and `Matplotlib`. All formal statistical tests used (e.g., t-test, ANOVA, etc.) were done in `SciPy`, and multiple-testing correction was performed where appropriate using the Holm step-down method as implemented in the Python library `statsmodels` v0.11.1 (Seabold and Perktold, 2010). Unless otherwise noted, all pairwise tests were two-tailed (where applicable). All t-tests were performed with Welch's correction to allow for unequal variances, and ANOVAs were run on grouped data first to justify additional pairwise comparisons where appropriate.

3.10 Supplemental materials

3.10.1 Supplementary figures

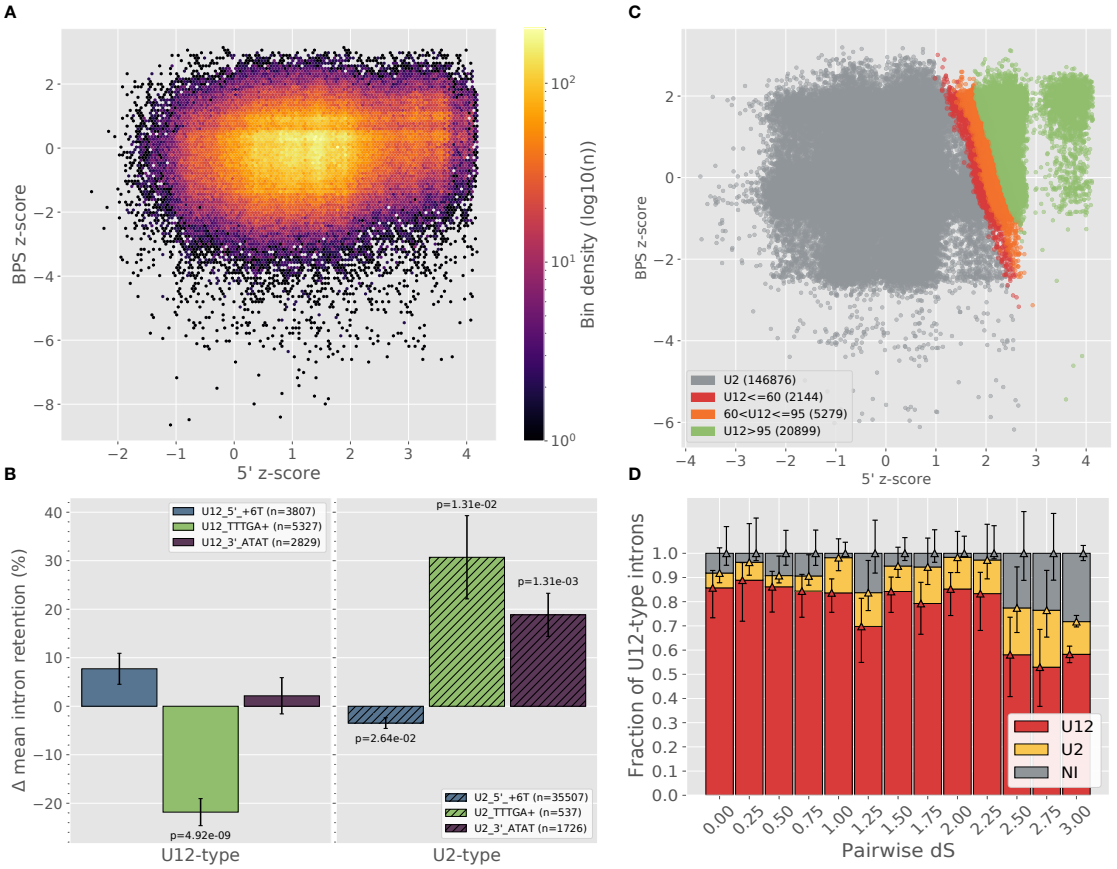


Figure 3.4: (A) The default PWMs used by intronIC are derived from human introns, and for divergent motifs like those present in *P. polycephalum* (especially the BPS motif) they fail to produce clear differentiation (i.e., separation of U12-type introns into a distinct cloud in the first quadrant). Curation of species-specific PWMs for *P. polycephalum* resulted in clearer differentiation along both axes (as in Figure 3.1C). (B) Relative intron retention for U12- (left) and U2-type (right) introns based on sequence features. Differences from the mean for each category are relative to all other introns of the same type. A negative/positive value indicates that introns with the given feature exhibit more/less efficient splicing relative to other introns of the same type. Features shown are "5′_+6T", introns with a T at position +6 in the intron; "TTTGA+", introns with the TTTGA motif within the last 55 bases of the intron; "3′_ATAT", introns with the motif ATAT immediately downstream of the 3′SS. (Caption continued on next page)

Figure 3.4: (Continued from previous page) (C) BPS-vs-5′SS score plot with assigned classifications for all *P. polycephalum* introns. The same underlying data as 3.1C, where each point represents an intron, and the color indicates the U12-type probability classification (U2-type: gray; U12-type with probability $\leq$ 60%: red; U12-type with probability 60-95%: orange; U12-type with probability $>$ 95%: green). (D) Between-paralog comparison provides little evidence for ongoing U12-type intron gain in *P. polycephalum*. For U12-type intron-containing paralog pairs sharing at least one intron of either type (to exclude recent retrogenes), pairwise dS values were used to bin all pairs into the range [0, 3]; dS values $\geq$ 3 were binned together. Within a given bin, each U12-type intron has one of three possible conservation states in its corresponding paralog: U12-type (red), U2-type (yellow) or no intron present ("no intron", gray). These data suggest that there have not been major U12-type intron gains in *P. polycephalum* since a time corresponding to at least dS $\approx$ 2.5. Whiskers represent the binomial proportion confidence intervals (Wilson score intervals) for the three categories (indicated by color of associated diamond).

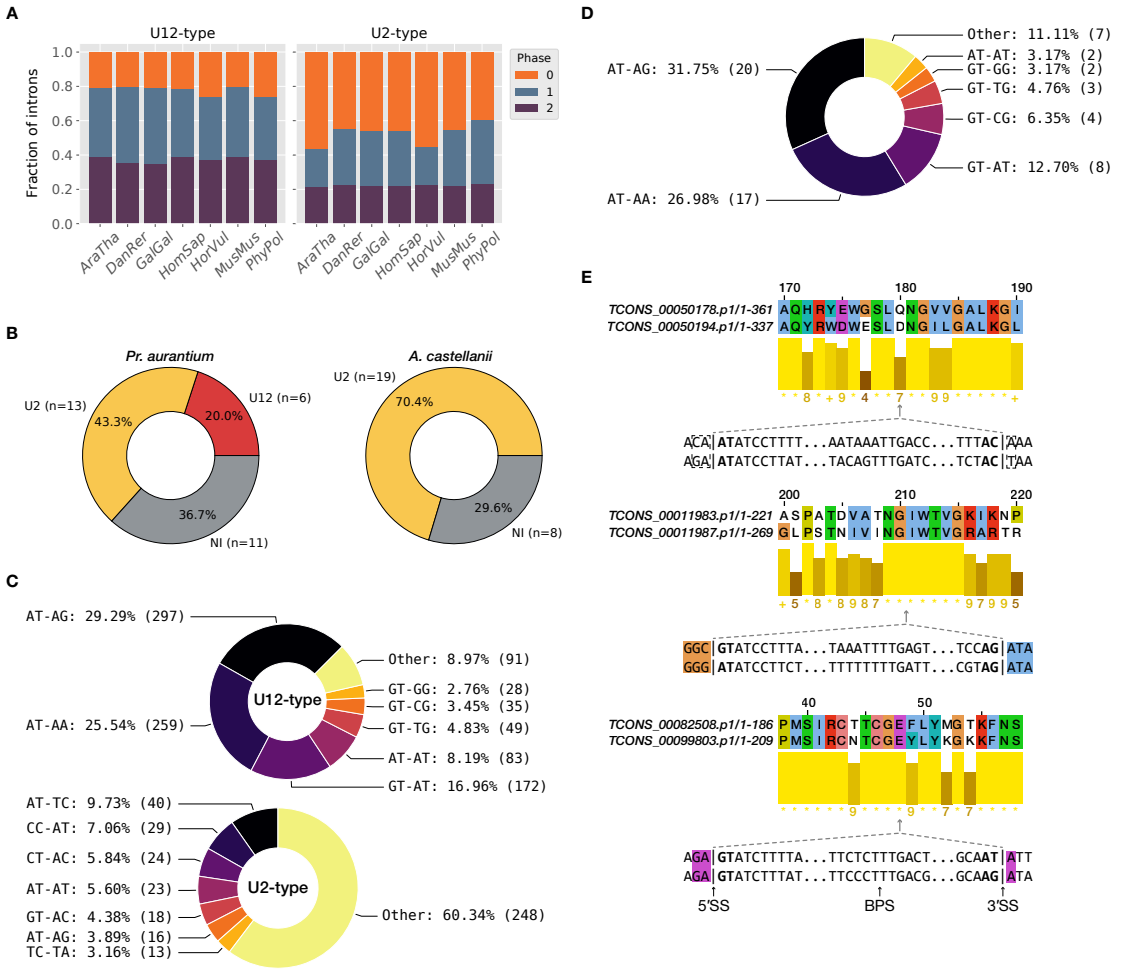


Figure 3.5: (A) Phase distribution of U12- (left) and U2-type (right) introns across different species. U12-type introns in *P. polycephalum* (PhyPol), as in other species, display a bias away from phase 0 whereas U2-type introns show a bias against phase 2. For each species, only introns interrupting coding sequence from the longest isoform of each gene were included. See Table S1, `https://doi.org/10.6084/m9.figshare.20483790` for additional species abbreviations. (B) Ancestral U12-type introns in *P. polycephalum* are conserved as introns in other amoebozoans. Each pie chart shows the conservation status (red, U12-type; yellow, U2-type; gray, no intron) of the same ancestral set of *P. polycephalum* U12-type introns (introns conserved as U12-type with one or more non-amoebozoans) in the variosean amoeba *Protostelium aurantium* (left) and the discosean amoeba *Acanthamoeba castellanii* (right). In each case, a significant majority of the U12-type introns are conserved as introns. These data suggest that these species have not undergone massive loss of U12-type introns; thus, the unprecedented number of U12-type introns in *P. polycephalum* likely represents significant U12-type intron creation in *P. polycephalum* rather than commensurate loss in related species. (Caption continued on next page)

Figure 3.5: (Continued from previous page) (C) U12- (top) and U2-type (bottom) non-canonical intron subtypes in *P. polycephalum* (using a 60% probability threshold for the U12/U2-type classification instead of the 95% threshold used elsewhere e.g., Figure 3.2A, thereby including "likely" U12-type introns), highlighting the degree to which non-canonical U12-type introns are greatly enriched for a subset of boundary pairs. By contrast, the U2-type non-canonical subtype distribution is much more diffuse. (D) Distribution of the subset of non-canonical U12-type introns which are found in regions of good alignment between pairs of *P. polycephalum* paralogs (but not necessarily conserved as introns between pairs)—increasing confidence that they are real introns—showing general consistency with the data in part C. (E) Example alignments of *P. polycephalum* paralogs, showing conserved U12-type introns (canonical and non-canonical). Coloring is based on chemical properties of the amino acids, and bars underneath each alignment represent chemical similarities of the aligned amino acids. Colored nucleotides before and after the intron splice sites correspond to the colors of the amino acid(s) in the alignment that are interrupted by the shared intron position. Transcript names appear in italics.

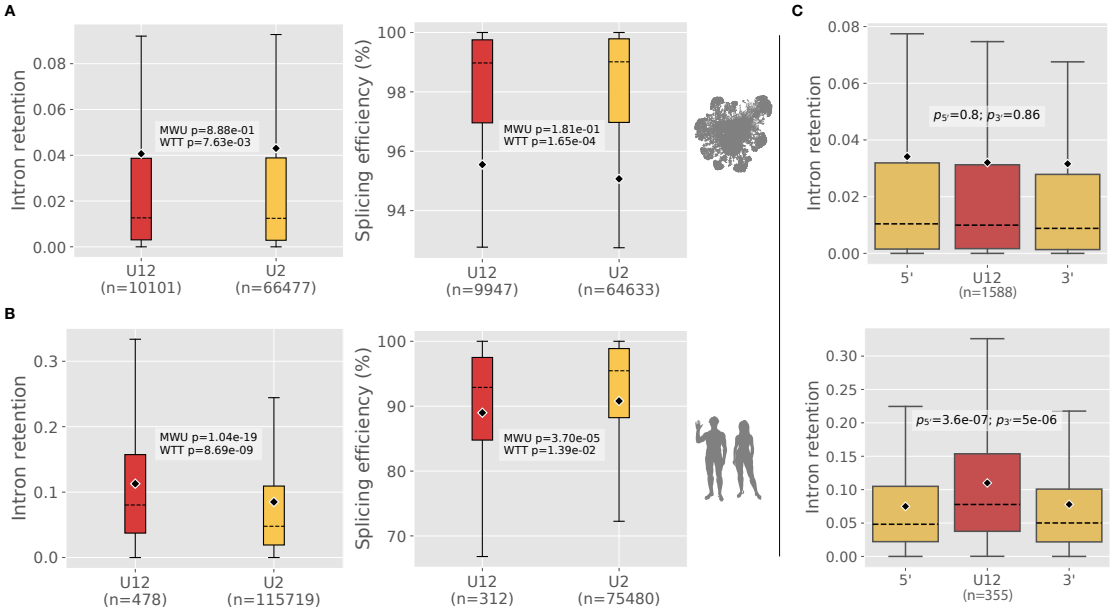


Figure 3.6: (A) Comparison of intron retention (left) and splicing efficiency (right) in *P. polycephalum* and human. Box plot of average intron retention and splicing efficiency data for *P. polycephalum* introns, showing that U12-type introns are neither more retained nor less efficiently-spliced than U2-type introns. Note that although the differences in means between U12- and U2-type introns are significant, this difference is inverted relative to data in other species. The left panel is the same as Figure 3.2C. (B) As in (A), but for Homo sapiens. Here, by both statistical measures shown there are significant differences between the two types of introns, with U12-type introns being more retained/less-efficiently spliced as has been reported elsewhere. MWU: Mann-Whitney U test; WTT: Welch's t-test. (D) U12-type intron retention is not significantly different from that of neighboring U2-type introns in *P. polycephalum* (top), unlike in human (bottom). Each plot represents aggregate data from multiple RNA-seq samples (total unique intron count listed below each plot), showing the distribution of intron retention values for U12-type (red; $>$ 95% U12-type probability in Physarum, $>$ 90% in human) and neighboring U2-type (yellow; $\leq$ 5% U12-type probability in Physarum, $\leq$10% in human) introns on either side (left: 5′, right: 3′). For each plot, pairwise U12- vs U2-type p-values were obtained via Mann-Whitney U tests, and corrected for multiple testing using the Holm step-down method (reported as $p_{5'}$ and $p_{3'}$ for the 5′ and 3′ U2-type data, respectively). For all parts, dashed line = median, diamond = mean, whiskers = 1.5 IQR. Note that y-axis scales differ between plots.

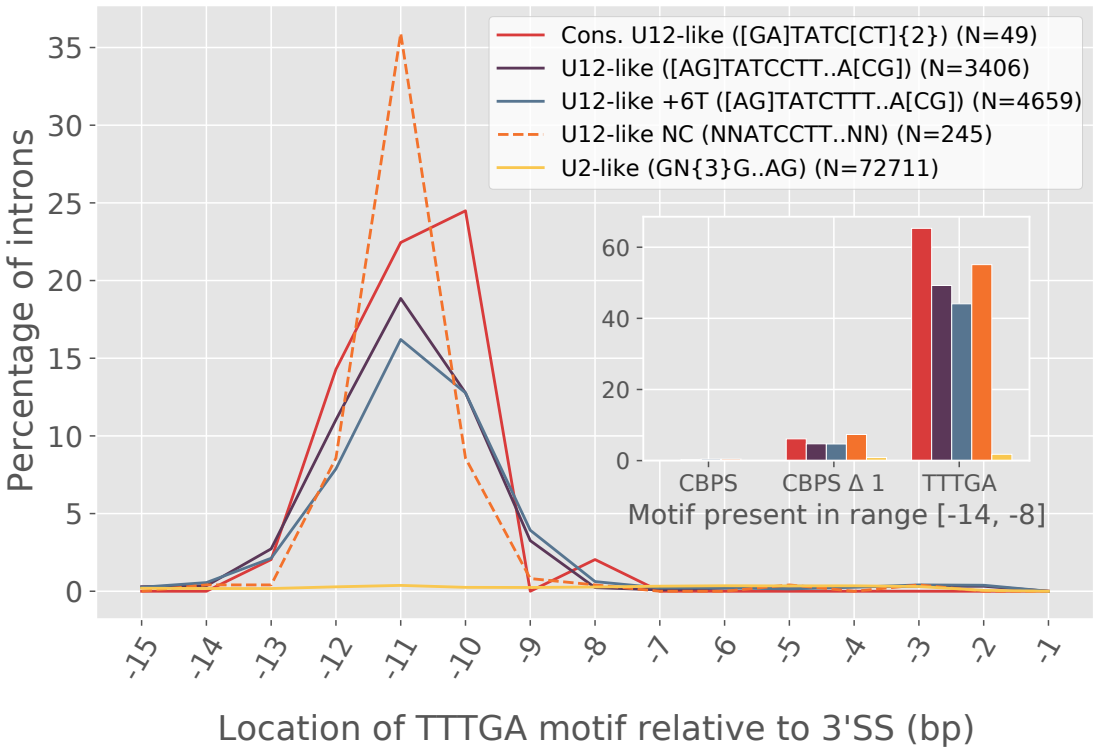


Figure 3.7: Enrichment of *Physarum*-specific BPS motif in non-canonical introns with U12-like sequence motifs.

Chapter 4

Where the Minor Things Are: A Census of Minor Spliceosomal Introns Across Thousands of Eukaryotic Genomes

4.1 Abstract

Spliceosomal introns are segments of eukaryotic pre-mRNA that are removed ("spliced") during creation of mature mRNA, and are one of the defining and domain-specific features of gene structure in eukaryotes. Introns are spliced by a large, multi-subunit ribonucleoprotein machinery called the spliceosome, and most eukaryotic genomes contain two distinct sets of this machinery—the major (or U2-type) spliceosome is responsible for removal of the vast majority (usually $> 99\%$) of introns in a given genome, with the minor (or U12-type) spliceosome processing the remaining minuscule fraction. Despite in some cases only being responsible for the removal of single-digit numbers of introns in entire genomes, the minor splicing system has been maintained over the roughly 1.5 billion years of eukaryotic evolution since the last eukaryotic common ancestor, and a number of recent studies have suggested that minor introns may be involved in certain aspects

of cell cycle regulation and cancer development. It is in this context that we present a broad bioinformatic survey of minor introns across more than 3000 eukaryotic genomes, providing a dramatic increase in information about their distribution in extant species, descriptions of their evolutionary dynamics and features across the eukaryotic tree and estimates of the minor intron complements of various ancestral nodes.

4.2 Introduction

Spliceosomal introns are sequences in eukaryotic genes that are removed (spliced) from the pre-mRNA transcripts of genes prior to maturation and nuclear export of the final mRNA, so named by association with the machinery that performs the splicing, the spliceosome (Will, Lührmann, and Luhrmann, 2011; Matera and Wang, 2014; Jurica, 2008). For the better part of a decade after spliceosomal introns (hereafter simply introns) were first characterized in eukaryotic genomes (Jeffreys and Flavell, 1977; Gilbert, 1978; Brack and Tonegawa, 1977; Breathnach and Chambon, 1981; R A Padgett et al., 1986), it was assumed that all introns shared a fixed set of consensus dinucleotide termini—GT at the beginning (5′ side) and AG at the end (3′ side)—and were processed in the same way (Mount, 1982; Jackson, 1991). This view was revised after the discovery of a small number of introns with AT-AC termini (Jackson, 1991; S. L. Hall and R A Padgett, 1994), and shortly thereafter an entirely separate spliceosome was described that could process these aberrant introns (Woan Y Tarn and J. a. Steitz, 1996; W Y Tarn and J A Steitz, 1996), termed the minor spliceosome. The minor spliceosome, like its counterpart now known as the major spliceosome, has origins early in eukaryotic evolution (Russell et al., 2006). Since minor introns were first documented as having AT-AC termini, it has been shown that the majority of minor introns in most species are in fact of the GT-AG subtype, although an increasing diversity of termini (so-called "non-canonical" introns, with boundaries that aren't GT-AG, GC-AG or AT-AC) seem to be able to be processed in certain contexts and to varying degrees by both spliceosomes (Frey and Pucker, 2020; Burset, Seledtsov, and

Solovyev, 2000; Pucker and Brockington, 2018; Sibley, Blazquez, and Ule, 2016; Moyer et al., 2020). Until very recently (Larue, Eliáš, and Scott W Roy, 2021), in every genome investigated minor introns have been found to comprise only a tiny fraction (<∼0.05%) of the total set of introns; despite this, they have also been found to be well-conserved over hundreds of millions of years of evolution (e.g., 96% of minor introns in human are conserved in chicken, 83% in octopus 4.2).

Because minor introns possess sequence motifs distinct from major introns (C. Burge and Sharp, 1997; C. B. Burge, R A Padgett, and Sharp, 1998), it is possible to try to identify them using sequence-based bioinformatic methods (Moyer et al., 2020; Alioto, 2007; C. B. Burge, R A Padgett, and Sharp, 1998; Bartschat and Samuelsson, 2010; Sheth et al., 2006; Levine and Durbin, 2001). Using various unpublished tools, previous studies have cataloged the presence/absence of minor introns/spliceosome components across multiple eukaryotic genomes (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008; Bartschat and Samuelsson, 2010; Sheth et al., 2006; Alioto, 2007), but many of these studies were necessarily constrained in their analyses by the limits of the data available at the time, and by the lack of a published or otherwise convenient computational method to identify minor introns. In this work, with the substantially larger and more diverse genomic datasets now publicly accessible coupled with the intron classification program `intronIC` (Moyer et al., 2020), we have been able to aggregate minor intron presence/absence data with higher fidelity than earlier works across a much larger sample of eukaryotic species, and characterize various aspects of their primary sequences, containing genes and evolutionary dynamics. For the purposes of this investigation we have limited our analyses to include only introns in protein-coding regions of genes, yet many genes contain introns in their untranslated regions (UTRs), at least some of which appear to be involved in the regulation of gene expression (Bicknell et al., 2012; Chung et al., 2006; Stark et al., 2005; Sharangdhar et al., 2017; Lu et al., 2008). UTR introns generally are an understudied class of introns, and almost nothing is known about minor introns in UTRs; exploring this in more detail would certainly be an exciting avenue for future research.

4.3 Materials and methods

4.3.1 Data acquisition

Genomes and annotations for all species were downloaded from the online databases hosted by NCBI (RefSeq and GenBank), JGI and Ensembl using custom Python scripts, and taxonomic information for each species was retrieved from the NCBI Taxonomy Database (Federhen, 2012) using a Python script. We used NCBI as our primary resource, since it contains the largest number of species and in many cases serves as the upstream source for a number of other genome resources. Additional species available only from JGI and Ensembl were included for completeness, as were GenBank genomes with standard annotation files (GTF or GFF; species with only GBFF annotations were excluded). GenBank annotations in particular are of highly variable quality and may be preliminary, draft or otherwise incomplete, which is one of the reasons we chose to also include annotation `BUSCO` scores for all species. Accession numbers (where available) and retrieval dates of the data for each species are available under the following DOI: `https://doi.org/10.6084/m9.figshare.20483655`.

4.3.2 Identification of spliceosomal snRNAs

Each annotated transcriptome (genome-based searches at this scale would have been time-prohibitive with our computational resources) was searched for the presence of the minor snRNAs U11, U12, U4atac and U6atac using `INFERNAL` v1.3.3 (Nawrocki and Eddy, 2013) with covariance models from Rfam (RF00548, RF00007, RF00618, RF00619). The default E-value cutoff of 0.01 was used to identify positive snRNA matches, and any snRNA with at least one match below the E-value cutoff was considered present.

4.3.3 Classification of minor introns

For every genome with annotated introns,`intronIC` v1.2.3 (Moyer et al., 2020) was used to identify putative minor introns using default settings which includes

introns defined by exon features (e.g., introns in UTRs) and excludes any introns shorter than 30 nt.

4.3.4 Identification of orthologous introns

A set of custom software was used to identify orthologs between various species as described previously (Larue, Eliáš, and Scott W Roy, 2021). Briefly, transcriptomes for each species in a group were generated using the longest isoforms of each gene (`https://github.com/glarue/cdseq`). Then, the program reciprologs (`https://github.com/glarue/reciprologs`) was used in conjunction with `DIAMOND` v2.0.13 (flags: `--very-sensitive -evalue 1e-10`) to identify reciprocal best hits (RBHs) between all pairs of species, and construct an undirected graph using the RBHs as edges. The maximal cliques of such a graph represent orthologous clusters where all members are RBHs of one another. In certain cases, when only orthologous MIGs were required (as opposed to all orthologs), reciprologs was run as part of a pipeline using the –subset argument in combination with separately generated lists of MIGs for each species, which dramatically decreases runtime by constraining the query lists to only include MIGs (producing identical results to the subset of a full alignment containing MIGs). Full ortholog searches were required for all ancestral density reconstructions as well as all comparisons of minor and major intron conservation/loss (Figure 4.5).

Orthologous groups were aligned at the protein level using a combination of `MAFFT` v7.453 and `Clustal Omega` v1.2.4, and intron positions within the alignments were computed using a custom Python pipeline (following the approach in (Scott W Roy, Fedorov, and Walter Gilbert, 2003)). Local alignment quality of $\geq 40\%$ conserved amino acid identity (without gaps) over a window of 10 residues both upstream and downstream of each intron position was required. Introns of the same type in the same position within aligned orthologs were considered conserved. For the analyses of putative intron type conversions (e.g., minor-to-major), major introns were required to have scores ≤ -30 instead of the default threshold of 0 to minimize the inclusion of minor introns with borderline scores as major-type, and intron alignments containing introns with such borderline scores (a tiny

fraction of the total alignments) were excluded. Intron sliding (the shifting of an individual intron's boundaries within a gene versus its ancestral location) (Arlin Stoltzfus et al., 1997) is not explicitly accounted for by our pipeline (an intron sliding event would be categorized as intron loss in the containing gene); however, this phenomenon is at best very rare and unlikely to meaningfully affect our results (Irina V Poverennaya, Potapova, and Spirin, 2020; S. W. S. W. Roy, 2009; Arlin Stoltzfus et al., 1997; Sêton Bocco and Csűrös, 2016).

4.3.5 Intron position within transcripts and intron phase

Included in the output of `intronIC` (Moyer et al., 2020) is information about the relative position of each intron within its parent transcript, represented as a percentage of the total length of coding sequence, as well as intron phase (for introns defined by CDS features). These were extracted for every species and used in the associated analyses.

4.3.6 Non-canonical minor introns

Species were first analyzed to assess the number of putative non-canonical minor introns they contained, and those with the highest numbers of non-canonical minor introns were used to perform multiple alignments of different pairs of species. From these alignments, all orthologous intron pairs with a minor intron (minor-minor or minor-major) were collected, and used to build clusters (subgraphs) of orthologous introns. For animals, human was used in a majority of the alignments to facilitate the generation of larger subgraphs (where the same intron shared between different pairwise alignments will group the alignments together); for plants, *Elaeis guineensis* served the same purpose.

4.3.7 BUSCO analysis

Translated versions of the transcriptomes of all species were searched for broadly-conserved eukaryotic genes using `BUSCO` v5.3.2 (Simão et al., 2015) and the `BUSCO`

lineage `eukaryota_odb10`. Complete `BUSCO` scores were then integrated into the overall dataset (e.g., Figure 4.1 and 4.4).

4.3.8 Curation of minor intron data/edge cases

Due to the number of genomes involved in our analyses, there will be some number of introns that appear superficially similar to minor introns simply by chance, and `intronIC` has no way of filtering these introns out because it cannot account for additional factors like local context, presence/lack of snRNAs, etc. In general this is not an issue, as the number of false-positive minor introns is usually very small. However, when summarizing an enormous amount of data and attempting to provide a resource to be used as a reference by others, we felt that some amount of curation was warranted to avoid obvious red herrings.

In service of that, we used the following heuristics in deciding whether to designate a given species as either confidently containing or not containing minor introns—species not meeting either set of criteria were assigned a minor intron density color of gray in Figure 4.1. First, it is important to note that `intronIC` will automatically try to adjust intron boundaries by a short distance if the starting boundaries are non-canonical and there is a strong minor 5′SS motif within ∼10 bp of the annotated 5′SS. In some poorly-annotated species, or species with otherwise aberrant intron motifs this can lead to increased false positives in the form of introns with "corrected" splice boundaries. Such introns are tagged by `intronIC` in the output, so it is possible to determine their proportion in the final number of minor introns reported. The criteria for presence of minor introns in our dataset is, a corrected / total minor intron fraction of ≤ 0.25 and any of the following: a) ≥ 3 called minor introns and at least two minor snRNAs, b) between 5 and 25 called minor introns and at least three minor snRNAs, c) ≥ 3 called minor introns and all four minor snRNAs.

The criteria for absence of minor introns (assigned the color black in the minor intron density color strip in Figure 4.1) is any of the following: a) ≤ 3 called minor introns and fewer than two minor snRNAs, b) ≤ 5 called minor introns and fewer than two minor snRNAs and fewer than five uncorrected AT-AC minor introns

and either annotated in RefSeq or with a `BUSCO` score greater than or equal to $B_{Q1} - (1.5 \times B_{IQR})$, where B is all the `BUSCO` scores of the broad RefSeq category to which the species belongs (i.e., "vertebrates", "invertebrates", "plants", "protozoa", "fungi"), B_{Q1} is the first quartile of such scores and B_{IQR} is the inner quartile range of such scores. The idea behind this metric is to only assign confident minor intron loss to species whose `BUSCO` scores aren't extremely low; very low `BUSCO` scores could indicate real gene loss or incomplete annotations, and neither of those scenarios forecloses on the possibility of the species having minor introns (whereas a species with a high `BUSCO` score and a very low number of minor introns/minor snRNAs is more likely to be genuinely lacking either/both). Finally, species with very low numbers of minor introns and minor snRNAs but very high minor intron densities ($\geq 1\%$) were categorized as uncertain to account for a small number of edge cases with massive intron loss and spurious false positives that, due to the low number of total introns, appear to be cases of outstandingly high minor intron density (e.g., *Leishmania martiniquensis*). Importantly, Figure 4.1 still includes the raw values for each species matching the above criteria; it is only the minor intron density color which is adjusted to indicate lack of confidence.

4.3.9 Calculation of summary statistics (introns/kbp CDS, transcript length, etc.)

Transcriptomes for all species were generated using a custom Python script (`https://github.com/glarue/cdseq`). Briefly, each annotated transcript's length was calculated as the sum of its constituent CDS features, and the longest isoform for each gene was selected using this metric. The number of introns per transcript was computed based on the same data, and combined with the transcript length to calculate introns/kbp coding sequence for each gene. Intron lengths were extracted directly from `intronIC` output, as was intron phase and intron position as a fraction of transcript length (where the position of each intron, taken as the point position in the coding sequence where the intron occurs, is calculated as the cumulative sum of the preceding coding sequence divided by the total length of coding sequence in the transcript). For comparisons of intron densities and gene

lengths of MIGs and non-MIGs, species with fewer than ten putative minor introns were excluded to avoid inclusion of spurious minor intron calls.

4.3.10 Ancestral intron density reconstruction

Reconstructions of ancestral intron complements in different nodes was performed as described in (Scott W Roy and Walter Gilbert, 2005a). Briefly, for a set of three species α, β and γ where γ is an outgroup to α and β (i.e., α and β are sister with respect to γ), introns shared between any pair of species are (under the assumption of negligible parallel intron gain) *a priori* part of the set of introns in the ancestor of α and β. For all introns shared between a given species pair, for example α and γ (but not necessarily β) $N_{\alpha\gamma}$, the probability of an intron from that set being found in β (in other words, the fraction of ancestral introns retained in β) is

$$\hat{P}_\beta = \frac{N_{\alpha\beta\gamma}}{N_{\alpha\gamma}},$$

where $N_{\alpha\beta\gamma}$ is the number of introns shared between all three species. Deriving these fractions of ancestral introns for each of the aligned species, we then define N_Ω as the total number of ancestral introns, and its relationship to the conservation states of introns in the alignments of the three species as

$$N_{\alpha\beta\gamma} = N_\Omega(\hat{P}_\alpha \cdot \hat{P}_\beta \cdot \hat{P}_\gamma),$$

the product of the ancestral intron number and the fraction of ancestral introns present in each species. Finally, solving for the number of ancestral introns we get the estimate

$$\hat{N}_\Omega = \frac{N_{\alpha\beta} \cdot N_{\alpha\gamma} \cdot N_{\beta\gamma}}{(N_{\alpha\beta\gamma})^2}.$$

Performing the above procedure for both major and minor introns in a given alignment allowed us to estimate the ancestral minor intron density for the corresponding node as

$$\hat{\rho}_{minor} = \frac{\hat{N}_{\Omega_{minor}}}{\hat{N}_{\Omega_{minor}} + \hat{N}_{\Omega_{major}}} \cdot 100\%$$

. However, without a point of reference this number is difficult to interpret, as the genes included in the alignments are not an especially well-defined set—because

these genes are simply all of the orthologs found between a given trio of species, their composition is likely to change at least somewhat for each unique group of aligned species. We deal with this by normalizing to a chosen reference species included in each group. For example, in our reconstructions of intron densities in the ancestor of Diptera, human was used as the outgroup and was therefore present in all alignments. After calculating the estimated minor intron density in the Dipteran ancestor, we then divided that value by the minor intron density in the human genes present in the same alignments to produce the estimated ancestral minor intron density relative to the corresponding minor intron density in human. Because using human as the outgroup for reconstructions of fungal and plant ancestors results in very small absolute numbers of minor introns, kingdom-specific outgroups were chosen instead: the estimates of ancestral fungal densities are relative to *Rhizophagus irregularis*, and those for plants are relative to *Lupinus angustifolius*. Because multiple species combinations were used to estimate the minor intron density at each ancestral node, we report the mean value over all n estimates for each node

$$\overline{\rho}_{minor} = \frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{minor_i},$$

$\pm$ the standard error (Figure 4.18).

As has been pointed out in other contexts, ancestral state reconstructions may be confounded by several different factors (Holland et al., 2020; Duchêne and Lanfear, 2015; Cunningham, 1999). Many such concerns are minimized in our specific application given that a) the traits under consideration are not complex, but rather the simple binary presence/absence of discrete genetic elements, b) calculations are restricted to introns present in well-aligning regions of orthologs (thereby avoiding issues with missing gene annotations in a given species, since alignments must include sequences from all species to be considered) and c) the contribution of parallel intron gain, especially of minor introns, is likely to be very small (Scott William Roy, 2016; Sverdlov et al., 2005; Carmel et al., 2007). There are a number of other potential sources of bias in our analyses, however, which are worth addressing. First, our ancestral intron density estimates are (to a large, though not complete, extent) dependent upon the accuracy of the phylogenetic relationships

in Figure 4.1. Ideally, we would have perfect confidence in all of the relationships underlying each node's reconstruction, but such an undertaking is beyond both the scope of this paper and the expertise of its authors. While we have done our best to be assiduous in choosing nodes with well-resolved local phylogenies—which is one reason we have not provided similar reconstructions for a much larger number of nodes with less-confident phylogenetic relationships—it remains the case that our reconstructions are only fully informative with respect to the tree upon which they are based. That being said, unless the phylogeny for a given node is so incorrect as to have mistaken one of the ingroups for the outgroup (i.e., the chosen outgroup was not in fact an outgroup), the reconstruction should still represent the ancestor of the two ingroup species. Second, we are relying on the correct identification of minor introns within each species to allow us to identify conserved/non-conserved minor introns in multi-species alignments. Although the field in general lacks a gold-standard set of verified minor introns upon which to evaluate classifier performance, the low empirical false-positive rate of `intronIC` (as determined by the number of minor introns found in species with compelling evidence of a lack of minor splicing) and the high degree of correspondence of its classifications with previously-published data suggests that our analyses are capturing the majority of the minor introns in each alignment. There is also the possibility that many minor introns are unannotated in many genomes (and in fact, for certain annotation pipelines we know that this has historically been the case). This concern is mediated somewhat by the fact that, because we are only considering gene models that produce well-aligning protein sequences across multiple species, our alignments are unlikely to contain unannotated introns of either type. Other unannotated minor introns, necessarily residing in completely unannotated genes, would of course not be considered in our analyses. The end result of this type of bias would be a shrinking of the total number of orthologous genes compared, resulting in the specter of the law of small numbers, whereby the samples can no longer be relied upon to represent the complete data with sufficient confidence. We have done what we can to combat this by choosing species with annotations of high quality (as assessed by `BUSCO` completeness, for example), and by using multiple combinations of species

to reconstruct each node—for reconstructions based upon a large number of different alignments, the low standard errors of the estimates give us some confidence that this kind of missing data is unlikely to qualitatively change our results.

4.4 Results

4.4.1 Minor intron diversity in thousands of eukaryotic genomes

In order to better assess the landscape of minor intron diversity in eukaryotes, we used the intron classification program `intronIC` (Moyer et al., 2020) (described in Materials and Methods) to process ∼270 million intron sequences and uncover minor intron numbers for over 3000 publicly-available eukaryotic genomes, representing to our knowledge the largest and most diverse collection of minor intron data assembled to date (Figure 4.1, underlying plain text data available at `https://doi.org/10.6084/m9.figshare.20483655`), one we hope will prove useful in informing future investigations into minor intron evolution.

Of the 1844 genera represented in our data, 1172 (64%) have well-supported evidence of minor introns in at least one species (see section 4.3 for details), with the remaining 672 appearing to lack minor introns in all available constituent species. Consistent with previous studies (C. B. Burge, R A Padgett, and Sharp, 1998; Bartschat and Samuelsson, 2010; Janice et al., 2012; Sheth et al., 2006; Alioto, 2007; Szcześniak et al., 2013; C.-F. Lin et al., 2010; Moyer et al., 2020; Turunen, Niemelä, et al., 2013), minor intron numbers and densities (fractions of introns in a given genome classified as minor type) vary dramatically across the eukaryotic tree; average values are highest in vertebrates and other animals, while variation between species appears to be lowest within land plants. Conservation of minor introns between different pairs of species is largely consistent with previously-published results (Alioto, 2007; Bartschat and Samuelsson, 2010; C.-F. Lin et al., 2010; Moyer et al., 2020) Pairwise minor intron conservation between various species. Bottom number is the number of minor introns conserved

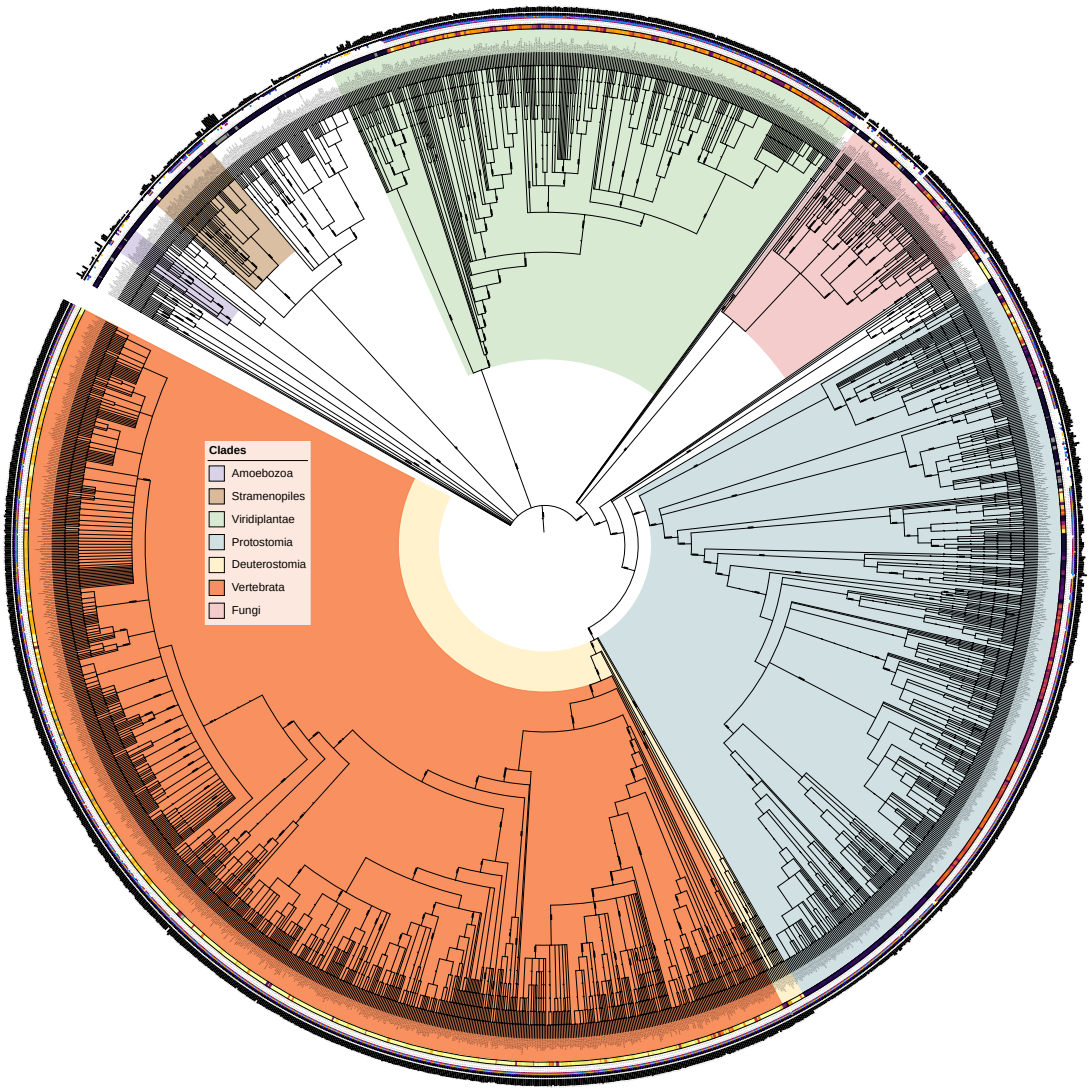


Figure 4.1: Minor intron densities for thousands of eukaryotic species. The colored strip following the species name represents the relative minor intron density (darker = lower, brighter = higher, gray values indicate species for which the estimated values are less confident, and may be enriched for false positives; see Materials and methods). Additional data from inside to outside is as follows: minor intron density (%), number of minor introns, presence/absence of minor snRNAs in the annotated transcriptome (red: U11, light blue: U12, yellow: U4atac, purple: U6atac), `BUSCO` score versus the eukaryotic `BUSCO` gene set, average overall intron density in introns/kbp coding sequence. Taxonomic relationships based upon data from the NCBI Taxonomy Database (Federhen, 2012); tree generated with iTOL (Letunic and Bork, 2021)

.

between the pair; top number is the number of conserved minor introns as a percentage of the minor introns present in the alignments for the associated species (the row species). For example, there are eight minor introns conserved between *D. melanogaster* and *L. polyphemus*, which is 88.9% of the *Drosophila* minor introns present in the alignment, but only 4.3% of *Limulus* minor introns. Full names of species are as follows: *Homo sapiens*, *Gallus gallus*, *Xenopus tropicalis*, *Latimeria chalumnae*, *Asterias rubens*, *Limulus polyphemus*, *Ixodes scapularis*, *Apis mellifera*, *Drosophila melanogaster*, *Priapulus caudatus*, *Lingula anatina*, *Octopus sinensis*, *Acropora millepora*, *Basidiobolus meristosporus*, *Rhizophagus irregularis*, *Arabidopsis thaliana*, *Lupinus angustifolius*, *Nicotiana tabacum*, *Zea mays*, *Amborella trichopoda*, *Sphagnum fallax*. The intriguing pattern of punctuated wholesale loss of minor introns is apparent within many larger clades in our data, along with a number of striking cases of minor intron enrichment in otherwise depauperate groups.

Minor intron enrichment

A number of cases of minor-intron-rich lineages are worth highlighting. As shown in Figure 4.3, the highest known minor intron density is found within the Amoebozoa; our recently-reported data in the slime mold *Physarum polycephalum* (Larue, Eliáš, and Scott W Roy, 2021) dwarfs all other known instances of local minor intron enrichment and appears to be an extremely rare example of significant minor intron gain. In the present study, we also find relatively high numbers of minor introns (compared to other amoebozoan species) in both the flagellar amoeba *Pelomyxa schiedti* (n=90) and the variosean amoeba *Protostelium aurantium* (labeled *Planoprotostelium fungivorum*[1] in Figure 4.1) (n=265). Although the numbers of minor introns in these species conserved with other lineages (e.g., humans) are very low, in all cases we find at least some degree of conservation. For example, in alignments between human and *P. aurantium* orthologs, 11% of

[1]Incorrectly labeled in the NCBI database as *Planoprotostelium fungivorum*; originally described as *Planoprotostelium fungivorum* in (Hillmann et al., 2018) but subsequently corrected in the main text (the incorrect usage remains in the supplemental materials); see (Shadwick, Silberman, and Spiegel, 2018) for supporting evidence of its classification as *Pr. aurantium*. Credit to Marek Eliáš for this addendum.

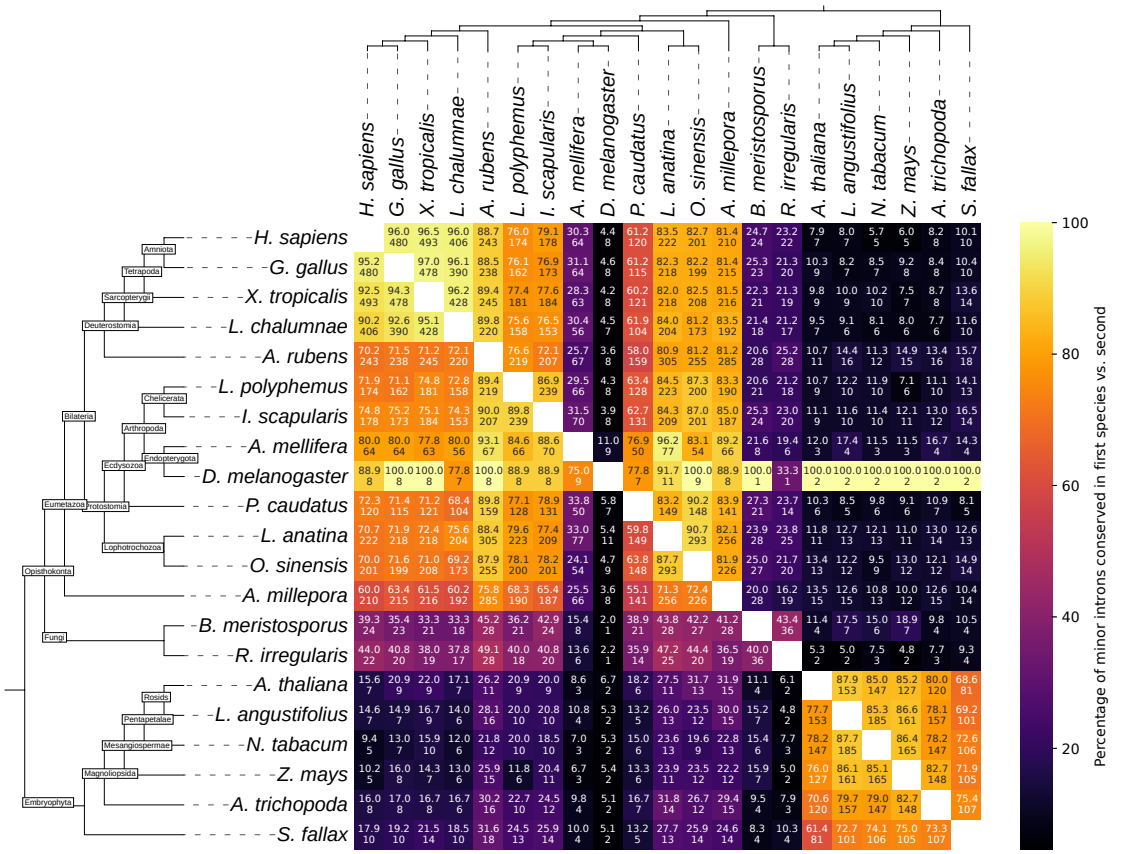


Figure 4.2: Pairwise minor intron conservation between various species. Bottom number is the number of minor introns conserved between the pair; top number is the number of conserved minor introns as a percentage of the minor introns present in the alignments for the associated species (the row species). For example, there are eight minor introns conserved between *D. melanogaster* and *L. polyphemus*, which is 88.9% of the *Drosophila* minor introns present in the alignment, but only 4.3% of *Limulus* minor introns. Full names of species are as follows: *Homo sapiens*, *Gallus gallus*, *Xenopus tropicalis*, *Latimeria chalumnae*, *Asterias rubens*, *Limulus polyphemus*, *Ixodes scapularis*, *Apis mellifera*, *Drosophila melanogaster*, *Priapulus caudatus*, *Lingula anatina*, *Octopus sinensis*, *Acropora millepora*, *Basidiobolus meristosporus*, *Rhizophagus irregularis*, *Arabidopsis thaliana*, *Lupinus angustifolius*, *Nicotiana tabacum*, *Zea mays*, *Amborella trichopoda*, *Sphagnum fallax*

human minor introns are conserved as minor introns in *P. aurantium*, comparable to proportions shared between human and many plant species (Moyer et al., 2020); in alignments with *P. schiedti* the proportion of conserved human minor introns is closer to 2.5%, although this appears to largely be due to massive minor-to-major conversion of ancestral minor introns in *P. schiedti*, as 69% of the human minor introns in those alignments are paired with major introns in *P. schiedti*. As reported by Gentekaki et al., the parasitic microbe *Blastocystis sp. subtype 1* within the stramenopiles contains hundreds of minor introns (Gentekaki et al., 2017), although our pipline identifies ∼45% fewer (n=253) than previously described. Interestingly, the *Blastocystis sp. subtype 1* minor introns we identify are highly enriched for the AT-AC subtype (77% or 196/253, where AT-AC introns are only ∼26% of all minor introns in human), and the classic minor intron bias away from phase 0 is inverted, with 49% (124/253) of the putative minor introns in *Blastocystis* being phase 0. *Blastocystis* also has the shortest average minor intron length in the data we analyzed at just under 42 bp (median 39 bp) (although introns shorter than 30 bp were systematically excluded from all species).

Surprisingly, we find unusually high minor intron densities in a number of fungal species, a kindgom which until now was not known to contain significant numbers of minor introns. In particular, the Glomeromycete species *Rhizophagus irregularis* has a minor intron density comparable to that of humans (0.272%, n=205), and *Basidiobolus meristosporus*, in Zoopagomycota, has one of the highest minor intron densities outside of the Amoebozoa (0.554%, n=249) (Figure 4.4). We do not find any convincing support for minor introns in either of the two largest fungal groups, Ascomycota and Basidiomycota, which seem to have lost most if not all of the required minor snRNAs in the vast majority of species as has been previously reported (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008; Bartschat and Samuelsson, 2010). Our analysis confirms the presence of a small number of minor introns in the oomycete genus *Phytopthora* as reported by other groups (Russell et al., 2006; Bartschat and Samuelsson, 2010) (Figure 4.4), and in addition we find that members of the stramenopile water mould genus *Saprolegnia* contain dozens of minor introns each. While any species with a very low number reported

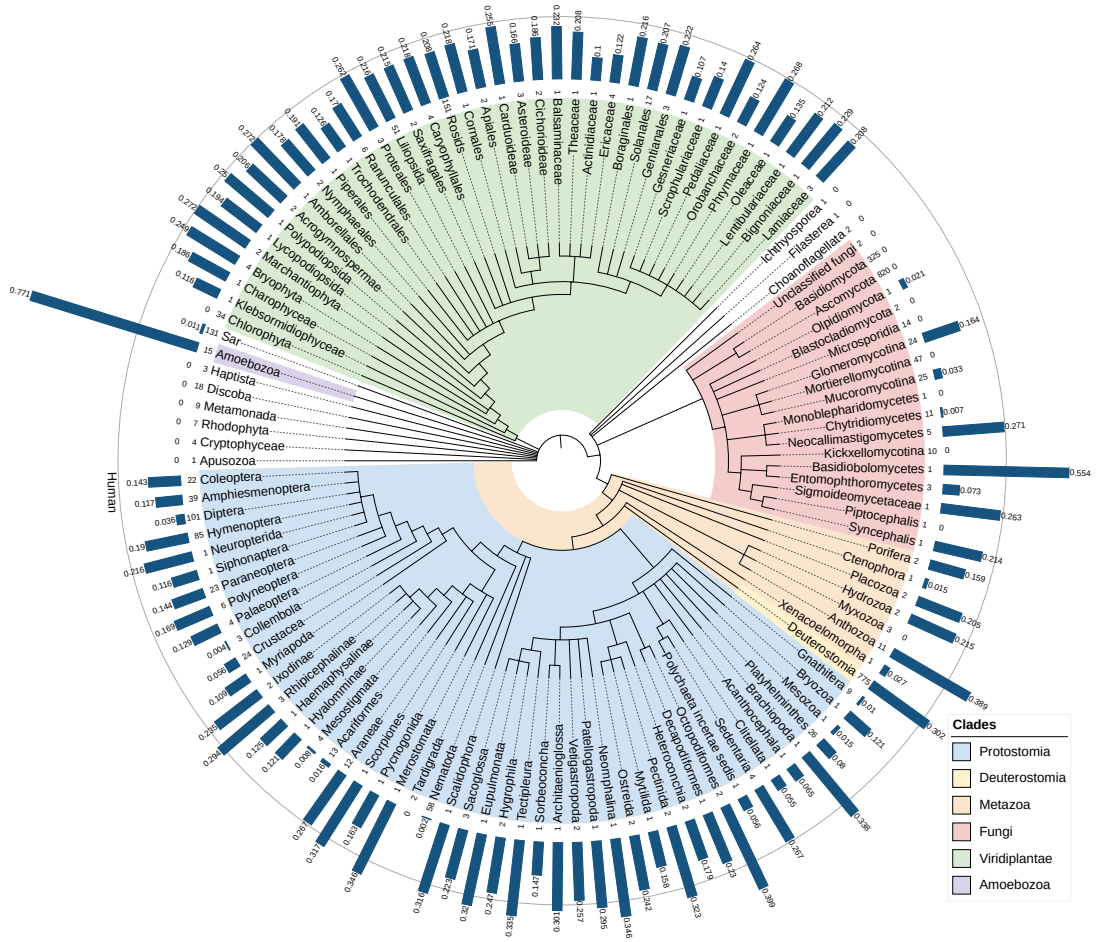


Figure 4.3: Average minor intron density (percentage of introns which are minor type; blue bars) in various eukaryotic clades. Integer numbers following clade names denote number of species represented. Outer circle indicates the minor intron density of the human genome. Taxonomic relationships based upon data from the NCBI Taxonomy Database (Federhen, 2012).

number of minor introns raises concerns about false positives, subsets of minor introns from each of these lineages have been found in conserved positions with minor introns in distantly-related species in our analyses, and minor snRNAs in each of the aforementioned genomes provides further evidence for the existence of bona fide minor introns in these species (Figure 4.4). Interestingly, given its sister placement to the broadly minor-intron-poor nematode clade, the cactus worm *Priapulus caudatus* appears to be quite minor-intron rich (n=330, 0.316%), with substantial minor intron conservation to other metazoan lineages.

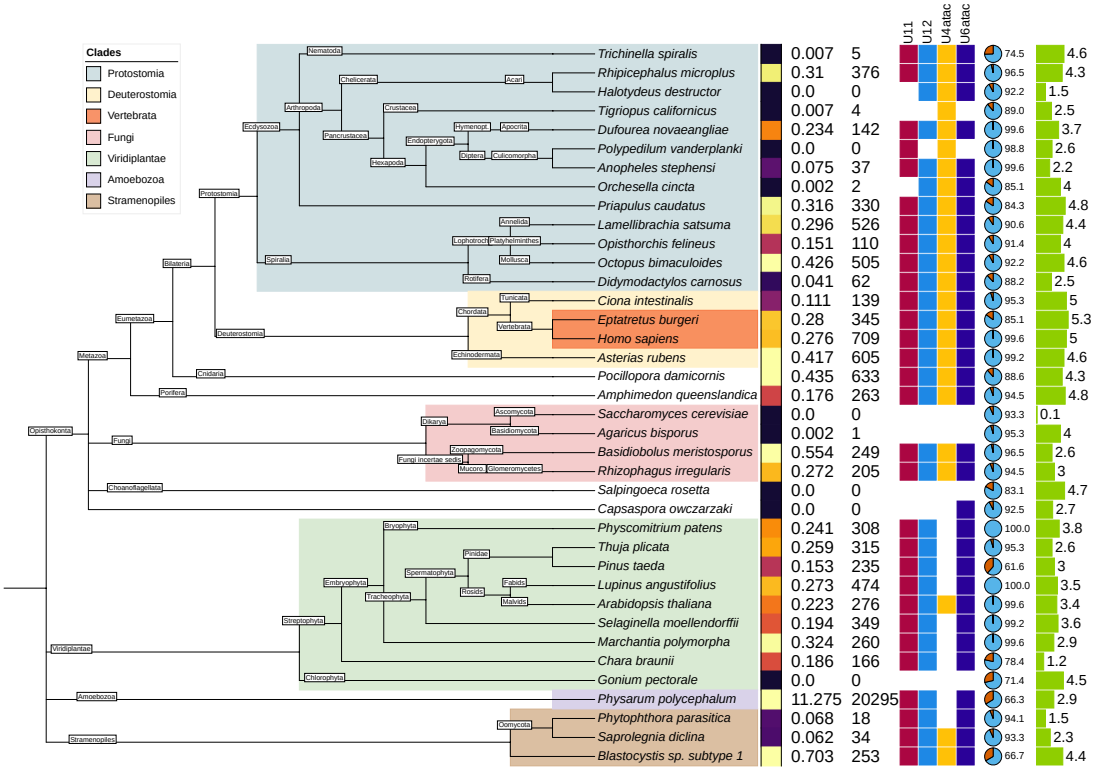


Figure 4.4: Minor intron densities and other metadata for selected species of interest. Graphical elements are as described in Figure 4.1.

Within the protostomes, one of the two sister clades of bilateria, there are cases of relative minor intron enrichment in both arachnids (Arachnida) and molluscs (Mollusca) (Figure 4.1), as well as in the brachiopod species *Lingula anatina* and the horseshoe crab *Limulus polyphemus*. The order of ticks Ixodida, including *Ixodes scapularis*, *Dermacentor silvarum* and *Rhipicephalus*, has a much higher

average minor intron density than other groups within Acari, which includes both mites and ticks and has seen substantial loss of minor introns in many of its lineages. On the other side of the bilaterian tree in deuterostomes, minor intron densities are far more homogeneous. Vertebrates have consistently high (∼0.3%) minor intron densities, with only a handful of exceptions in our data that are very likely due to incomplete or otherwise problematic annotations (for an example, see *Liparis tanakae* in Figure 4.1, an individual species with dramatically lower minor intron densities than surrounding taxa, with a low `BUSCO` score and no indication of minor spliceosome loss). The remaining deeply-diverging clades within deuterostomes have minor intron densities comparable to vertebrates (the starfish *Asterias rubens* being on the high side of vertebrate densities, for example) with the exception of tunicates, which appear to have lost a significant number of their ancestral minor intron complements (and minor splicing apparatus, in the case of the highly-transformed species *Oikopleura dioica*).

In their seminal paper examining spliceosomal snRNAs in various eukaryotic lineages, Dávila López et al. (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008) report a number of clades without some/any minor snRNAs which, based upon our larger dataset, it now seems clear have both minor introns and most if not all of the canonical minor snRNAs. These include the *Acropora* genus of coral, which has an average minor intron density higher than that of most vertebrates; within the fungal phylum Chytridiomycota the Chytridiomycete species *Spizellomyces punctatus* as well as a number of Neocallimastigomycetes including *Piromyces finnis* and *Neocallimastix californiae*; the genus of blood flukes *Schistosoma*; and all of the species of Streptophyta included in their analysis (see Fig. 1 in (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008)). Notably, we also find minor introns (confirmed by comparative genomics) in the green algal species *Chara braunii* (n=166) and *Klebsormidium nitens* (n=110), representatives of a group which until now was thought to lack minor splicing entirely (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008; Bartschat and Samuelsson, 2010; Russell et al., 2006), as well as in the Glaucophyte alga *Cyanophora paradoxa* (n=77) (which may have transformed minor splicing machinery, as we only find significant

hits to the U11 snRNA in that species).

Minor intron depletion

Punctuated and dramatic loss of minor introns is a hallmark feature of the minor splicing landscape, and it remains an outstanding question why certain lineages undergo either partial or complete loss of their ancestral minor intron complements (Turunen, Niemelä, et al., 2013). Previous work has delineated many groups that appear to lack either minor introns, minor splicing components or both (Bartschat and Samuelsson, 2010; Russell et al., 2006; M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008), but the diversity and scope of more recently-available data motivated us to revisit this topic. In addition to the underlying data presented in Figure 4.1, there are a number of cases of severe and/or complete minor intron loss that we highlight here. First, the amoebozoan *Acanthamoeba castellanii* has been found to contain both minor splicing apparatus as well as a limited number of minor-like intron sequences (Russell et al., 2006). While it remains likely that this species contains a small number of minor introns based upon previous evidence, we do not find conservation of any of the twelve *Acanthamoeba* introns our pipeline classified as minor in either human or the more closely-related amoeobozoan *Protostelium aurantium*. This may not be particularly surprising, given the low absolute number of minor introns under consideration—between *Protostelium aurantium* and human, for example, ∼23% of *Protostelium* minor introns are conserved, and there are only two minor introns from *Acanthamoeba* in regions of good alignment with human orthologs. Furthermore, we do find a single shared minor intron position between *Acanthamoeba* and human when we disregard the local alignment quality and simply consider all introns in identical positions within aligned regions, which amounts to 20% of *Acanthamoeba* minor introns in such alignments.

Among clades with extreme but incomplete loss (a classic case in animals being Diptera), notable examples include the Acari (ticks and mites), bdelloid rotifers, the springtail (Collembola) subclass of hexapods, We find no evidence for minor introns in the following taxa, some of which corroborate earlier stud-

ies: tardigrades (e.g., *Hypsibius exemplaris*, Discoba (e.g., *Trypanosoma*, *Leishmania*) (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008), Orchrophyta (stramenopiles), Alveolata (protists) (Bartschat and Samuelsson, 2010; M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008). In addition to an overall extreme reduction in minor introns throughout the clade generally, the Acari also contains a number of cases of apparent (by comparative genomic analysis) complete loss in the parasitic mite *Tropilaelaps mercedesae* (though minor introns are present in sister taxa) and the earth mite *Halotydeus destructor*. We also report two other novel cases of apparent complete minor intron loss outside of Acari. First, in the Dipteran clade Chironomidae, we find scant evidence of minor introns in *Clunio marinus*, *Polypedilum vanderplanki* or *Belgica antarctica*, all of which also appear to be missing between half and three-quarters of their minor snRNAs. Second, the copepod species of crustaceans *Tigriopus californicus* and *Eurytemora affinis* each lack both conserved minor introns and 75% of the minor snRNA set.

4.4.2 Minor introns have lower average conservation than major introns

A persistent result in the minor intron literature is that minor introns are more highly conserved than major introns (specifically, between animals and plants and even more specifically, between human and *Arabidopsis thaliana*) (Basu, Makalowski, et al., 2008), although this assertion has been contradicted by at least one more recent analysis (Moyer et al., 2020). The claim that minor intron conservation exceeds major intron conservation rests upon the numbers of introns of both types found in identical positions within 133 alignments of orthologous human-*Arabidopsis* sequences, as reported in Table 1 of Basu et al. (Basu, Makalowski, et al., 2008). For major (U2-type) introns, they find 115 conserved as major in aligned ortholog pairs, and 1391 either not present in one of the two orthologs or present as a minor intron; for minor introns (U12-type), they report 20 conserved and 135 missing/converted. For each intron type, taking the number conserved and dividing by the total number of introns of that type present in the alignments results in conservation percentages of 7.6% ($\frac{115}{115+1391}$) for major introns and 12.9%

($\frac{20}{20+135}$) for minor introns, leading to the conclusion (although the aforementioned values are not explicitly stated in the text) that minor introns are more highly conserved between human and *Arabidopsis* than are major introns. To the extent that we correctly understand their approach, however, we believe there may be a complication with this analysis.

Examining the ortholog pairs the authors provide in the supplementary data, it is evident that many *Arabidopsis* sequences are present in multiple ortholog pairs, which suggests that a standard reciprocal-best-hit criteria for ortholog identification was not employed and that certain introns will be counted multiple times within the orthologous alignments. As many minor introns occur in larger gene families, this methodology could lead to artificial inflation of the calculated minor intron conservation, especially given the small absolute number of minor introns at issue. To attempt to more thoroughly address the question of minor vs. major conservation, we identified orthologs in many different pairs of species across a range of evolutionary distances (see 4.3), and calculated intron conservation using the same metric as above. Within more than 100 such comparisons between animals and plants (and more than 60 between animals and fungi), we find no cases where minor intron conservation exceeds major intron conservation (Figure 4.5).

Furthermore, we observe only a handful of cases where minor intron conservation marginally exceeds major intron conservation in alignments of more closely-related species (∼3% greater between the starfish *Asterias rubens* and the stony coral *Orbicella faveolata*, for example). In the specific case of human-*Arabidopsis* addressed by Basu et al., our data show minor intron conservation to around half that of major intron conservation (Table 4.1). Thus, in the final analysis we find no compelling support for the idea that minor introns are in general more conserved than major introns and in fact, the opposite seems to be true in the vast majority of cases.

4.4.3 Minor intron loss vs. conversion

When an ancestral minor intron ceases to be a minor intron, it is thought to happen primarily in one of two ways: the entire intron sequence could be lost via,

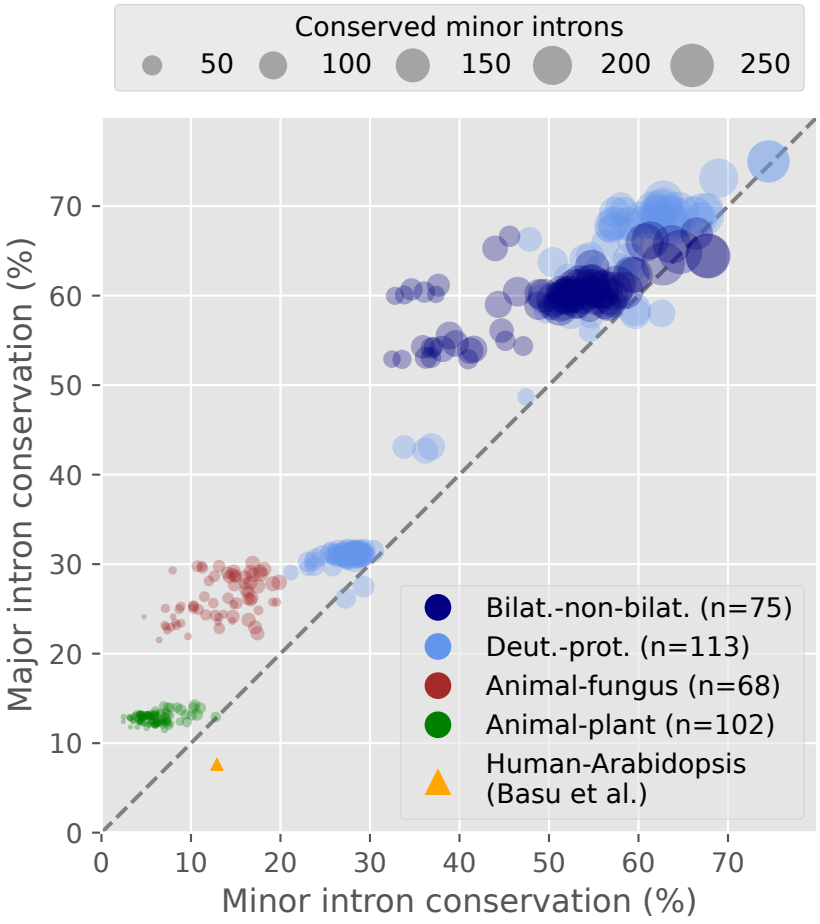


Figure 4.5: Comparison of major (y-axis) vs. minor (x-axis) intron conservation across hundreds of pairs of species. Bilat.-non-bilat.: bilaterian vs. non-bilaterian (animal); Deut.-prot.: deuterostome vs. protostome. The yellow triangle indicates levels of conservation of major and minor introns between *Homo sapiens* and *Arabidopsis thaliana* as reported by Basu et al. (Basu, Makalowski, et al., 2008). Size of markers indicates number of minor introns conserved between each pair.

Table 4.1: Comparison of major and minor intron conservation between human and *Arabidopsis thaliana*. $N_{conserved}$ indicates the number of introns of each type conserved as the same type in both human and *Arabidopsis*. $N_{variable}$ indicates the total number of introns (of both species) present in the alignments where the corresponding position in the opposing sequence either does not contain an intron, or contains an intron of the other type.

Major			Minor			p_{Fisher}
$N_{conserved}$	$N_{variable}$	conservation (%)	$N_{conserved}$	$N_{variable}$	conservation (%)	
2052	14162	12.7	7	120	5.5	0.015

for example, reverse transcriptase-mediated reinsertion of spliced mRNA (C.-F. Lin et al., 2010; Cohen, Shen, and Carmel, 2012; Turunen, Niemelä, et al., 2013; Irimia and Scott William Roy, 2014), or the intron could undergo sequence changes sufficient to allow it to be recognized instead by the major spliceosome (C. B. Burge, R A Padgett, and Sharp, 1998; Dietrich, Incorvaia, and Richard A Padgett, 1997; Dietrich, Fuller, and Richard A Padgett, 2005a; M J Frilander and J A Steitz, 1999). From first-principles arguments based on the greater information content of the minor intron motifs (C. B. Burge, R A Padgett, and Sharp, 1998; C. Burge and Sharp, 1997; Dietrich, Incorvaia, and Richard A Padgett, 1997) along with empirical analyses (C.-F. Lin et al., 2010), it is assumed that intron conversion proceeds almost universally unidirectionally from minor to major. Previous work has also shown that the paradigm of full intron loss (sequence deletion) appears to dominate over conversion in minor introns (C.-F. Lin et al., 2010); we wondered whether any exceptions to this might exist.

We first assembled a manually-curated sample of species with significant/complete minor intron loss, along with a number of species with much higher minor intron conservation for comparison. For each selected species, we chose an additional species as well as a species to serve as an outgroup, and then identified orthologs between all members of the group to allow us to identify ancestral major/minor introns (see 4.3 for details) and estimate fractions of each intron type retained. Considering loss to include both sequence deletion as well as type conversion (which we assume to be unidirectional from minor to major, as discussed above), Figure 4.6

shows that minor intron loss is more pronounced than major intron loss in the species we examined (also shown more generally in Figure 4.5).

We can, however, also decompose the phenomena contributing to the higher degree of loss of minor introns and ask whether sequence deletion specifically, for example, differs between the two types. Somewhat surprisingly, we find that this form of intron loss is very similar between the two types of introns in species which have lost significant fractions of their minor introns (Figure 4.7). Because the selected species were chosen based upon putative loss of minor introns and the sample size is low, it is difficult to interpret the apparent bias toward minor intron deletion in the vertebrates and plants in Figure 4.7. For the other species, however, this data suggests that there is not a particular selective pressure toward removing minor intron sequences themselves, at least not any more than there is pressure to remove intron sequences generally, in instances of pronounced minor intron upheaval.

We can also look at the other side of the intron loss coin, conversion from minor to major type rather than sequence deletion. Here, we find that in many instances loss does indeed outstrip conversion (as reported by (C.-F. Lin et al., 2010)), sometimes dramatically so, but there are interesting exceptions—in particular, the leech *Helobdella robusta* (HelRob), which seems to have retained a large fraction of its ancestral major introns, has lost ∼80% of its minor introns primarily through conversion to major-type (Figure 4.8). By contrast, the annelid worm *Dimorphilus gyrociliatus* (DimGyr), found in a clade (Polychaeta) sister to *Helobdella robusta*, has undergone a seemingly independent loss of minor introns of similar proportion to *Helobdella* under a very different modality, with loss (deletion) outweighing conversion (Figure 4.8). It is unclear what forces are responsible for the relative contributions of each mechanism; in *Helobdella*, the major intron sequences are slightly more degenerate at the 5′SS end than in e.g., human, which might lower the barrier to entry for would-be minor-to-major converts but this is purely speculative and more work is needed to better characterize these dynamics. It should be noted that under the current analysis we cannot differentiate between losses and conversions followed by subsequent loss; our conversion estimates, therefore, should

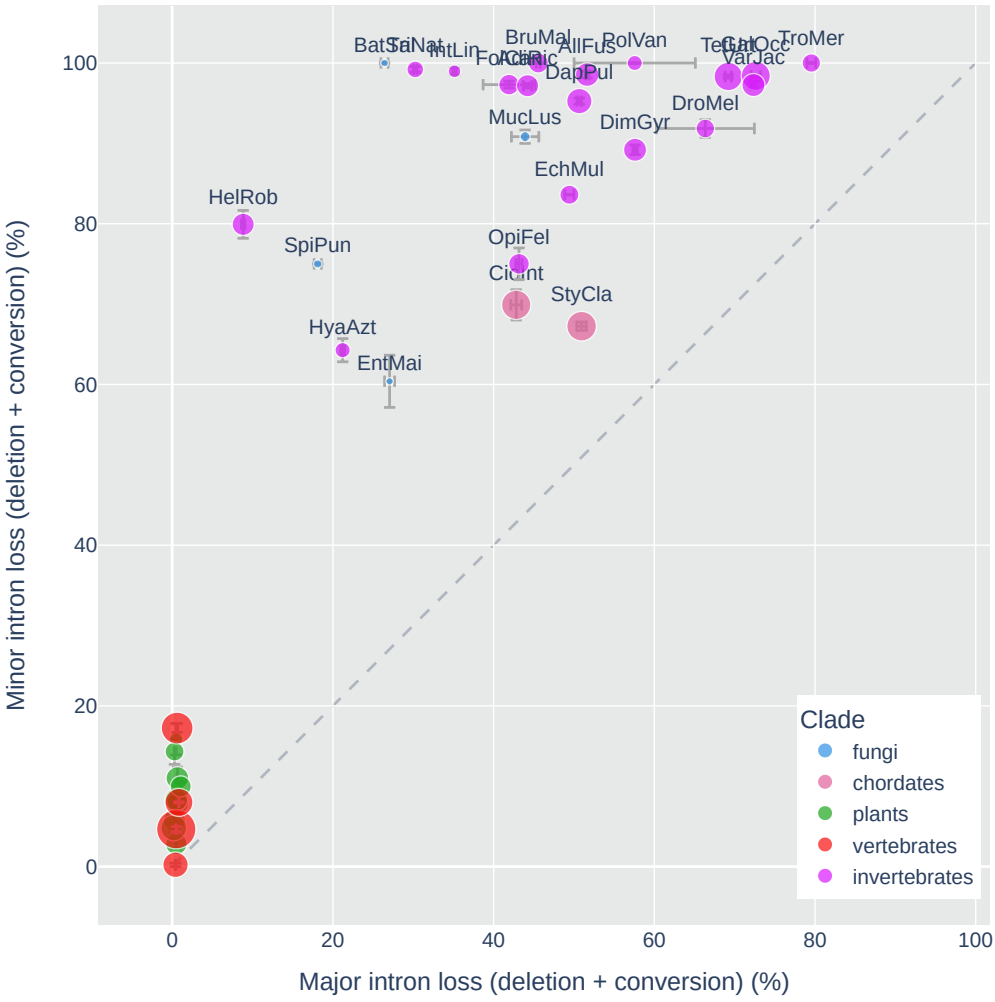


Figure 4.6: Major vs. minor intron loss, where "loss" includes both sequence deletion and conversion to an intron of the other type. Species abbreviations are as follow: AdiRic: *Adineta ricciae*, AllFus: *Allacma fusca*, BatSal: *Batrachochytrium salamandrivorans*, BruMal: *Brugia malayi*, CioInt: *Ciona intestinalis*, CluMar: *Clunio marinus*, DapPul: *Daphnia pulicaria*, DimGyr: *Dimorphilus gyrociliatus*, DroMel: *Drosophila melanogaster*, EchMul: *Echinococcus multilocularis*, EntMai: *Entomophaga maimaiga*, FolCan: *Folsomia candida*, GalOcc: *Galendromus occidentalis*, HelRob: *Helobdella robusta*, HyaAzt: *Hyalella azteca*, IntLin: *Intoshia linei*, MucLus: *Mucor lusitanicus*, OpiFel: *Opisthorchis felineus*, PolVan: *Polypedilum vanderplanki*, SpiPun: *Spizellomyces punctatus*, StyCla: *Styela clava*, TetUrt: *Tetranychus urticae*, TriNat: *Trichinella nativa*, TroMer: *Tropilaelaps mercedesae*, VarJac: *Varroa jacobsoni*

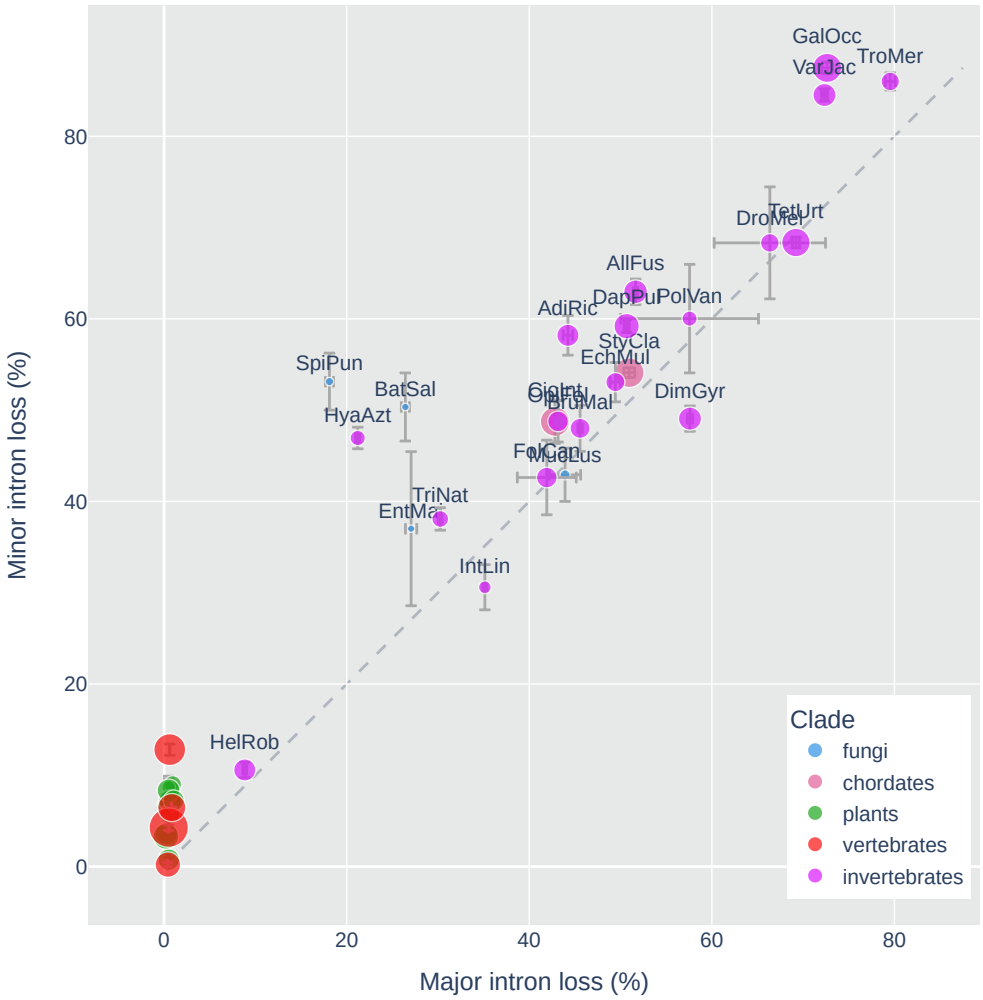


Figure 4.7: Major vs. minor intron loss, where "loss" represents actual deletion of the intron sequence. For species abbreviations see Figure 4.6

be seen as lower bounds.

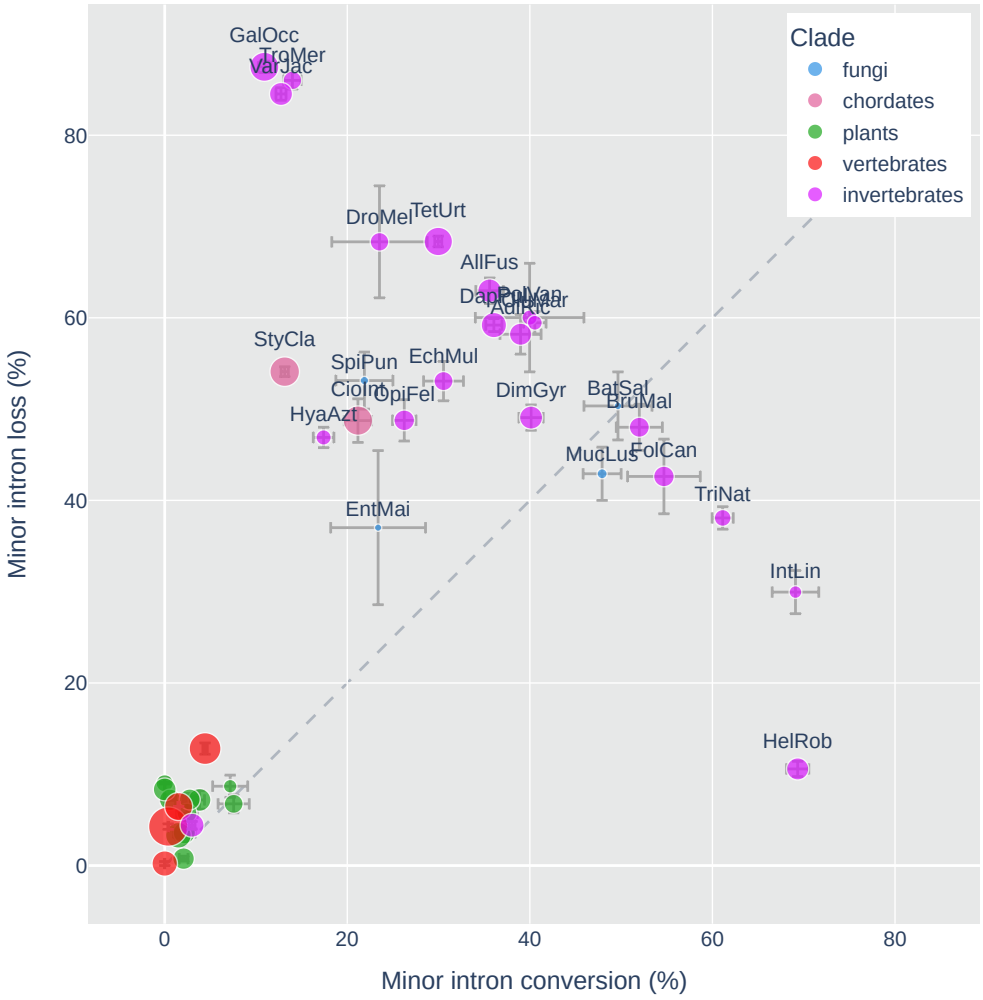


Figure 4.8: Minor intron loss vs. conversion, where "loss" represents actual deletion of the intron sequence. For species abbreviations see Figure 4.6

4.4.4 Positional biases of major and minor introns

It has been known for many years that introns in many species exhibit a 5′ bias in their positions within transcripts (Mourier and Jeffares, 2003; K. Lin and D.-Y. Zhang, 2005; Sakurai et al., 2002). This can be explained in large part due to biased intron loss: because a primary mechanism of intron loss (and, to a

more limited extent, gain) is thought to occur via reverse-transcriptase mediated insertion of spliced mRNA (Scott W Roy and Walter Gilbert, 2005b; Scott William Roy and Walter Gilbert, 2006; Derr and Strathern, 1993), and because this loss appears to be biased toward the 3′ end of transcripts, over time such a process would result in higher concentrations of introns closer to the 5′ end of transcripts.

Less attention has been paid to the positional biases of minor introns specifically, although at least one study (Basu, Makalowski, et al., 2008) found that minor introns appear to be especially over-represented in the 5′ half of transcripts in both human and *Arabidopsis thaliana*. We were curious to see whether the same patterns were present in our own data and whether they generalized beyond the two species so far examined.

We selected two sets of species to highlight—for the first, we chose lineages with substantial numbers of minor introns from a variety of groups; for the second, we picked species with significant inferred amounts of minor intron loss to investigate whether any 5′ bias might be more extreme in the remaining minor introns. In our analysis, we confirm the 5′ bias as previously described (Basu, Makalowski, et al., 2008) in *Arabidopsis thaliana* (Figure 4.9a), although we do not find the same difference in major and minor intron positions in human.

More broadly, our results point to a less-clear picture than earlier work might suggest—while we do find a number of cases of cases in animals where minor introns are more 5′-biased than major introns (Figure 4.9b, *Amphibalanus amphitrite* and *Trichinella spiralis*), the pattern is not broadly significant and is occasionally reversed (e.g *Ixodes scapularis*), albeit in animal species with less-dramatic minor intron loss (Figure 4.9A). Within plants, however, we appear to see a clearer pattern, with a much higher fraction of plants species in both groups displaying a clear 5′ bias in their minor introns. To determine how widespread this pattern of greater relative 5′ bias in minor introns is, we searched our entire dataset for species with: 1) significant differences in minor intron occurrence between the 5′ and 3′ halves of trancripts as assessed by a two-tailed exact binomial test, where presence in the 5′ half of a transcript was considered a success (as used in ref. (Basu, Makalowski, et al., 2008)); 2) significant differences between major and minor positions as de-

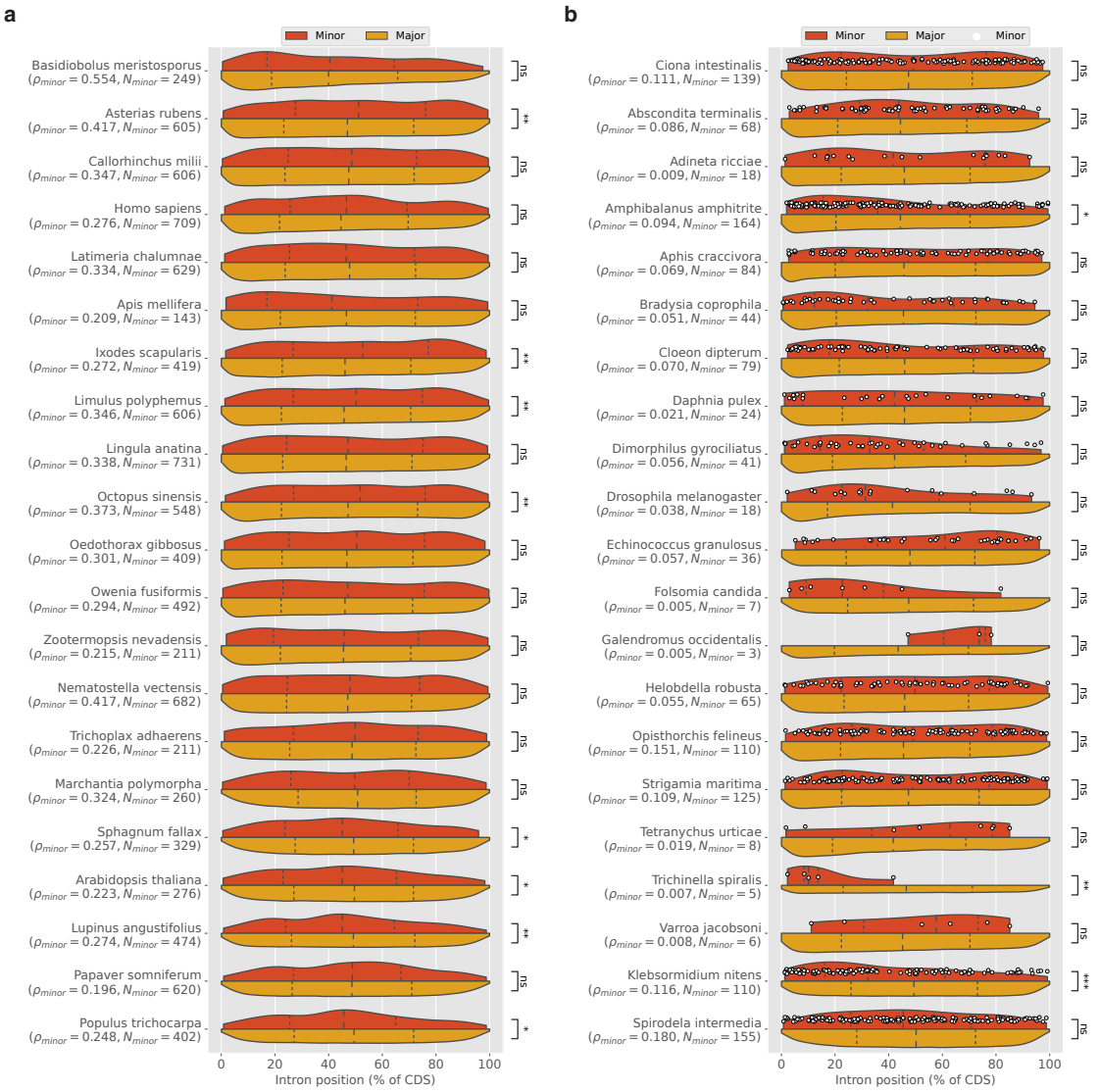


Figure 4.9: Intron position distributions for major (red) and minor (yellow) introns in selected species. (a) Species enriched in minor introns. (b) Species with significant inferred minor intron loss; white dots represent individual minor introns. For both plots: Dashed lines represent the first, second and third quartiles of each distribution. Statistically significant differences between minor and major introns are indicated with asterisks (two-tailed Mann-Whitney U test; * $p < 0.05$; ** $p < 0.001$; *** $p < 0.0001$; `ns`=not significant). Note that in some cases of significant differences between the two intron types, e.g., within animals, the set with greater 5′ bias is the *major* introns.

termined via a two-tailed Mann-Whitney U test; 3) median minor intron positions more 5′ biased than median major intron positions. Among such species, plants are highly over-represented (Table 4.2, $p = 8.9 \times 10^{-68}$ by a Fisher's exact test).

Table 4.2: Proportion of species in various groups with statistically-significant 5′ bias ($N_{5'MIB}$) of minor intron positions within transcripts.

Clade	N_{total}	$N_{5'MIB}$	%
Streptophyta	290	112	38.6
Fungi	63	2	3.2
Metazoa	1204	21	1.7
Stramenopiles	16	0	0.0
Evosea	4	0	0.0
Discosea	1	0	0.0

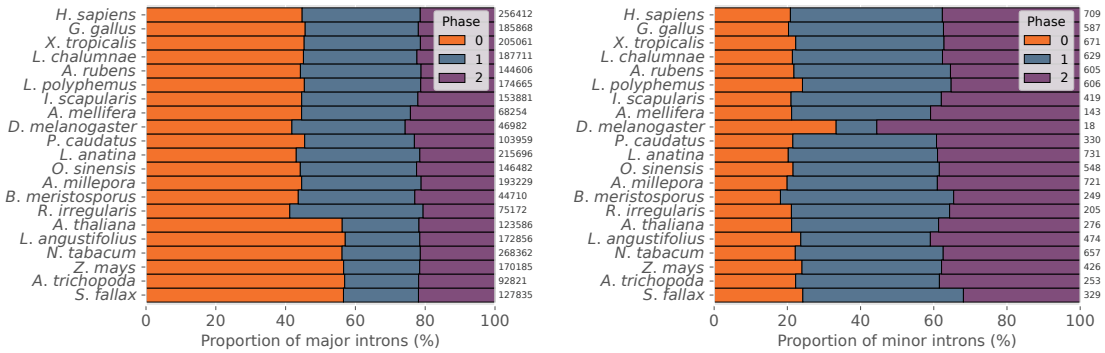
It is possible that this pattern, taken together with the higher degree of stability of minor intron densities in plants, reflects an ancient loss of minor introns in the plant ancestor, the signature of which is now shared broadly among extant species. It might also suggest a unique and/or more consistent paradigm for minor intron loss in plants, distinct from the more haphazard process seemingly at work within other parts of the eukaryotic tree where losses have occurred more recently and more frequently.

4.4.5 Phase biases of minor introns

Spliceosomal introns may occur at one of three positions relative to protein-coding sequence: between codons (phase 0), after the first nucleotide of a codon (phase 1) or after the second (phase 2). In most species, major introns display a bias toward phase 0 (Nguyen, Yoshihama, and Kenmochi, 2006; Long, Rosenberg, and Gilbert, 1995) 4.10a[2], while minor introns are biased away from phase 0 (C. B. Burge, R A Padgett, and Sharp, 1998; Moyer et al., 2020) 4.10b. It remains an unsettled issue why minor introns are biased in this way—one theory proposed

[2]It is interesting to note that all ∼1000 of the introns in the yeast species *Candida maltosa* (which lacks minor introns) appear to be phase 0 (Figure 4.11a, bottom-right corner)

by Moyer et al. (Moyer et al., 2020) suggests that it could arise from preferential conversion of phase 0 minor introns to major type, which would over time lead to the observed pattern. Here, we wanted to make use of the size of our dataset to better characterize the diversity of intron phase more broadly, and identify any exceptions to the general rule.

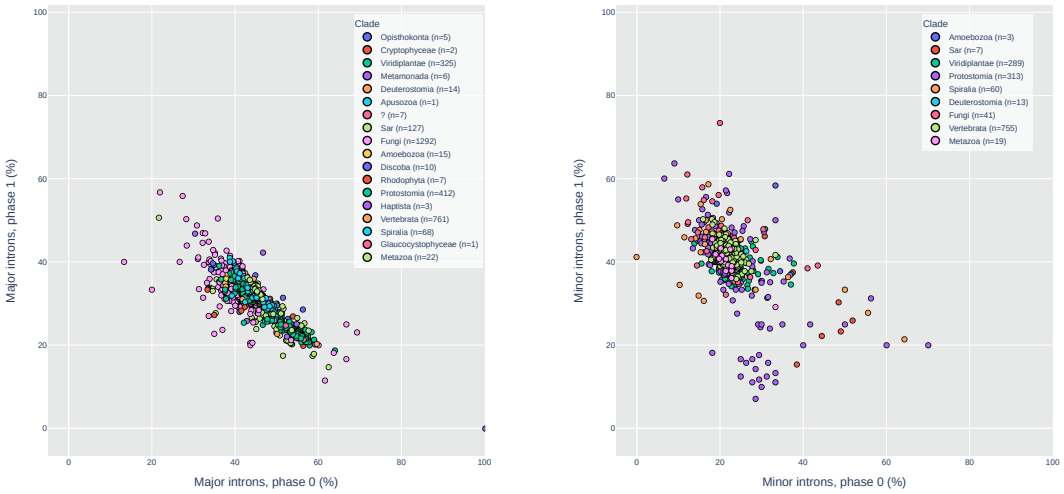


(a) **(b)**

Figure 4.10: Phase distributions for major (a) and minor (b) introns in various species. Numbers at the end of each bar represent the total number of constituent introns.

As shown in Figure 4.11a, the phase distributions of major introns are fairly tightly grouped. On average, for major introns in minor-intron-containing species, phase 0 makes up 47%, phase 1 30% and phase 2 23%. In addition, the proportions of phase 0 and phase 1 introns are quite highly correlated ($\rho_s = -0.81$, $p = 0.0$). Minor introns, on the other hand, are less consistent in their phase distribution and have a lower phase 0 to phase 1 correlation ($\rho_s = -0.48$, $p = 2.14 \times 10^{-87}$), although the majority cluster quite strongly around the average value of phase 0, 22% (Figure 4.11b).

It is intriguing that a small number of species appear to have much higher fraction of phase 0 minor introns (Figures 4.11b and 4.12; what's more, these species (with the notable exception of *Blastocystis sp. subtype 1*, addressed below) all have very low numbers of minor introns (Figure 4.12). While these data are not necessarily incompatible with the conversion paradigm mentioned above (which might predict minor introns in species with pronounced loss to show especially strong bias away from phase 0, although with such small numbers of remaining



(a) Proportions of phase 1 (y-axis) vs. phase 0 (x-axis) in major introns of various species (which contain minor introns). Correlation of phase 0 to phase 1: $\rho_s = -0.81$, $p = 0.0$.

(b) Proportions of phase 1 (y-axis) vs. phase 0 (x-axis) in minor introns of various species. Correlation of phase 0 to phase 1: $\rho_s = -0.48$, $p = 2.14 \times 10^{-87}$

Figure 4.11

minor introns it may simply be that stochasticity dominates, for example), it at least invites further investigation into the forces underlying the phase biases in minor introns generally.

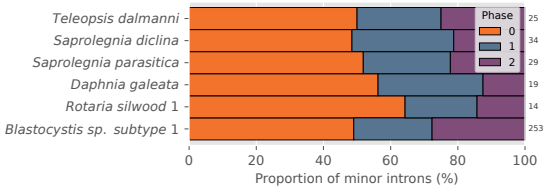


Figure 4.12: Unusually high proportions of phase 0 minor introns in certain species. Numbers at the end of each bar represent the total number of constituent introns.

The species *Blastocystis sp. subtype 1* is similar in its reduced minor intron bias away from phase 0 to the other unusual cases mentioned, but is remarkable for the number of minor introns involved (n=253). It is also interesting that its minor intron phase distribution is almost identical to the distribution of phases in its major introns (not shown). While this raises the possibility that the minor

introns in *Blastocystis sp. subtype 1* are false-positives, the fact that we find a) all four minor snRNAs in the genome, b) a (small but non-zero) number of minor introns conserved in *Lingula anatina* (not shown) and c) putative minor introns in a closely-related species (*Blastocystis hominis*) provides evidence that they are likely to be real. Assuming they are bona fide minor introns, another possible explanation for the phase 0 enrichment could be that they have been more recently gained, and (under the conversion hypothesis) have not yet had time to develop the phase bias present in older minor intron sets. More thorough comparative genomics work within the clade after additional species become available would help to clarify the evolutionary picture.

Finally, we also note that most *Drosophila* minor introns have a stronger bias away from phase 1 than average; this can be seen as the set of species roughly clustered around 30% phase 0, 10% phase 1 in Figure 4.11b, and in more detail in Figure 4.13. Given the consistency of the bias across the clade, it seems likely that its origin extends back to at least the last common ancestor of *Drosophila*. We do not have a specific working hypothesis for why this phase bias exists, but highlight it here as a possible avenue of future investigation.

4.4.6 Non-canonical minor intron splice boundaries

The vast majority (>98.5%) of major introns in most eukaryotic genomes begin with the dinucleotide pair GT, and end with the pair AG (Burset, Seledtsov, and Solovyev, 2000; Burset, Seledtsov, and Solovyev, 2001; Moyer et al., 2020; Sheth et al., 2006), with an additional much smaller contingent of GC-AG introns present in many genomes. When minor introns were first discovered, they were initially characterized largely by their distinct AT-AC termini (S. L. Hall and R A Padgett, 1994; Jackson, 1991). However, it was subsequently discovered that in fact the majority of minor introns in most species share the same terminal boundaries as major introns (Dietrich, Incorvaia, and Richard A Padgett, 1997; C. B. Burge, R A Padgett, and Sharp, 1998), although the AT-AC subtype may constitute a more significant fraction of minor introns in certain species (Turunen, Niemelä, et al., 2013; Rogozin et al., 2012; Moyer et al., 2020; Alioto, 2007; Bartschat and

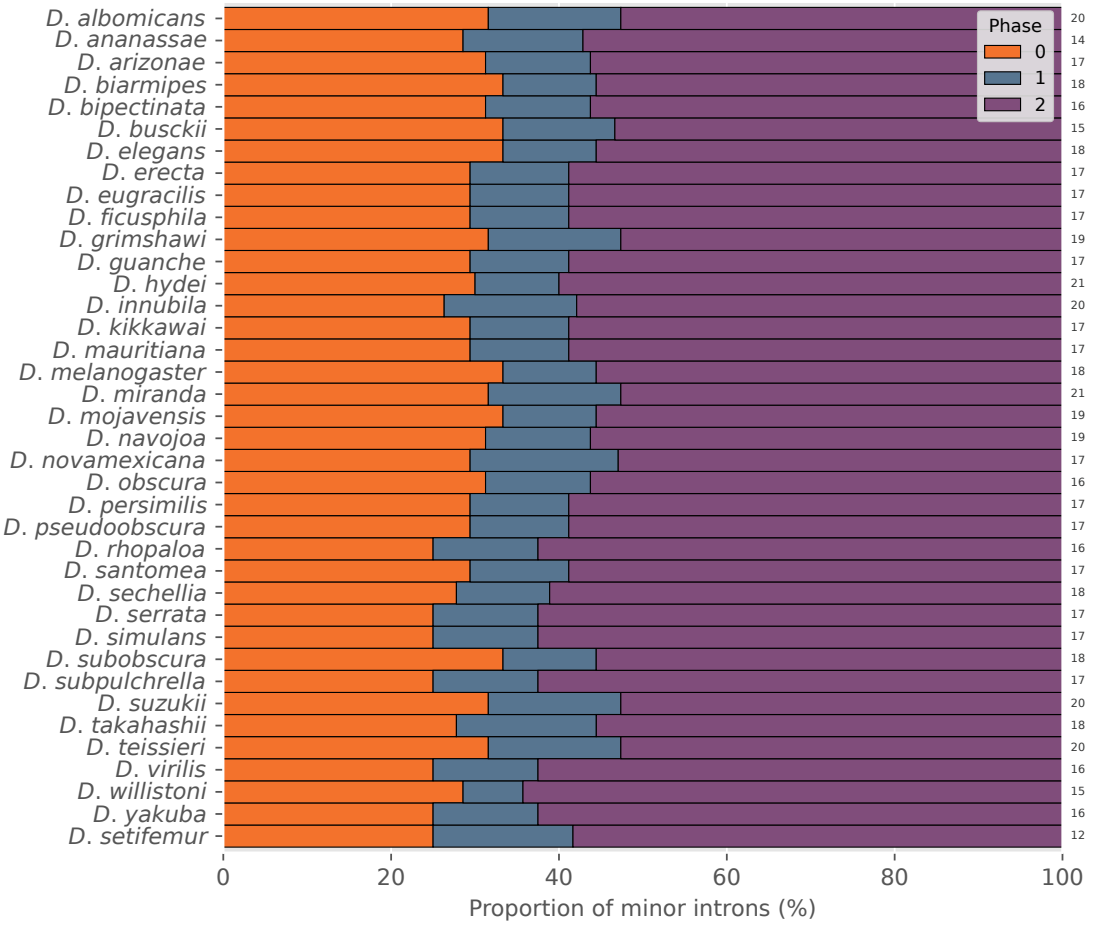
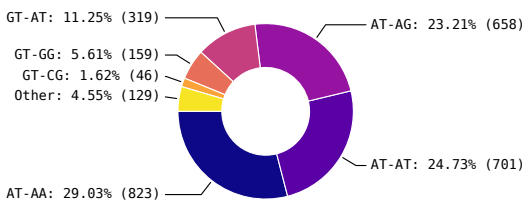


Figure 4.13: Phase distributions of minor introns in various *Drosophila* species, highlighting the reduced fraction of phase 1 introns relative to the normal pattern. Numbers at the end of each bar represent the total number of constituent introns.

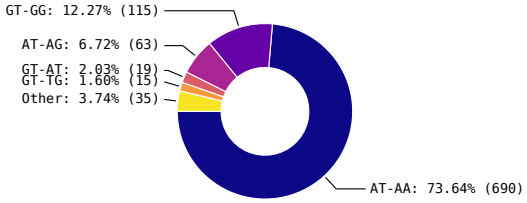
Samuelsson, 2010). Over time, additional non-canonical (i.e., not GT-AG, GC-AG or AT-AC) subtypes of minor introns have been identified in various organisms (Parada et al., 2014; Alioto, 2007; Moyer et al., 2020; Sheth et al., 2006; C.-F. Lin et al., 2010), but these analyses have been limited to species for which there were available minor intron annotations which until now were quite limited.

Because non-canonical introns do not, by definition, look like normal introns, it can be difficult to differentiate between biological insights and annotation errors when examining eukaryotic diversity at scale. For example, a recent report on non-canonical introns in diverse species described significant enrichment of CT-AC introns in fungi (Frey and Pucker, 2020). However, and as addressed briefly in the paper itself, CT-AC boundaries happen to be the exact reverse-complement of the canonical GT-AG boundaries—additional sequence features of these introns presented, such as a high occurrence of C 5 nt upstream of the 3′SS (which would perfectly match the hallmark +5G were the intron on the other strand) and an enrichment in +1C after the 3′SS (corresponding to the canonical -1G at the 5′SS on the other strand) make it very likely in our estimation that such introns are in fact incorrectly annotated due to some combination of technical errors and antisense transcripts. To combat issues of this sort, we first performed multiple within-kingdom alignments of various animal and plant species with high relative levels of annotated non-canonical minor intron boundaries. Conserved introns were then clustered across many different alignments to form conserved intron sets, which were then filtered to include only minor introns in sets where at least two minor introns were found (see 4.3 for details). These sets of introns are much less likely to contain spurious intron sequences, although they also may not fully represent more recent or lineage-specific boundary changes and they do not include introns from every species in our collected data.

Our results in animals (Figure 4.14a) and plants (Figure 4.14b) are largely consistent with previous data on non-canonical minor introns (Parada et al., 2014; Sheth et al., 2006; C.-F. Lin et al., 2010), although they differ to some degree in rank-order within each set. The set of plant non-canonical minor intron termini is both less-diverse than the animal set and more lopsided; while the most common



(a) Non-canonical intron termini found in conserved minor introns in animals.



(b) Non-canonical intron termini found in conserved minor introns in plants.

non-canonical termini is AT-AA in both kingdoms, almost 75% of all non-canonical minor introns we identify in plants are of the AT-AA subtype, versus less than half that proportion in animals. Interestingly, the second most common non-canonical termini in animals, AT-AT, is almost entirely absent in plants.

As can be seen in Table 4.3, the vast majority of non-canonical termini differ by a single nucleotide from a canonical terminus; only GT-TA, GT-CA, AT-GA, AT-CG, CT-AT, and AT-GT in animals and AT-TT, AT-CA, AT-CG, AT-GA, and AT-GT in plants differ by more than one nucleotide, and each are only tiny minorities of the total non-canonical set. Additionally, there are small differences between the consensus sequences outside of the terminal dinucleotides between the different subtypes of minor introns (Figure 4.15), and also within the same subtype between animals and plants. The most prominent examples of the latter are in the following subtypes: GT-GG (AT motif immediately preceeding the 3′SS, and ATG motif immediately following it in plants), AT-AG (`-1C` from the 3′SS in animals) and GT-AT (`-1A` in animals).

4.4.7 Minor intron-containing genes are longer and more intron-rich than genes with major introns only

Across the eukaryotic tree, genomes can vary widely in the number of introns contained in an average gene (Scott William Roy and Walter Gilbert, 2006). Some vertebrate genes have dozens or even hundreds of introns (e.g., the gene titin in human), whereas most genes of the yeast *Saccharomyces cerevisiae* lack introns entirely. Given the fact that minor introns appear to be arranged non-randomly

Table 4.3: Non-canonical minor intron termini in animals and plants. Termini with only a single occurrence are excluded.

Animals			Plants		
Termini	%	Count	Termini	%	Count
AT-AA	29	823	AT-AA	73.6	690
AT-AT	24.7	701	GT-GG	12.3	115
AT-AG	23.2	658	AT-AG	6.7	63
GT-AT	11.3	319	GT-AT	2	19
GT-GG	5.6	159	GT-TG	1.6	15
GT-CG	1.6	46	TT-AG	1.3	12
GT-TG	1	29	AT-CA	0.6	6
CT-AC	0.8	24	GT-CG	0.5	5
GG-AG	0.7	21	AT-AT	0.4	4
GA-AG	0.5	14	AT-TT	0.2	2
TT-AG	0.4	10	AT-CG	0.1	1
GT-TA	0.1	3	GT-AA	0.1	1
GT-CA	0.1	3	AT-GC	0.1	1
AT-GA	0.1	3	GT-AC	0.1	1
AT-GC	0.1	3	AT-GA	0.1	1
CT-AG	0.1	2	AT-GT	0.1	1
AT-CG	0.1	2			
CT-AT	0.1	2			
AT-CC	0.1	2			
AT-GT	0.1	2			

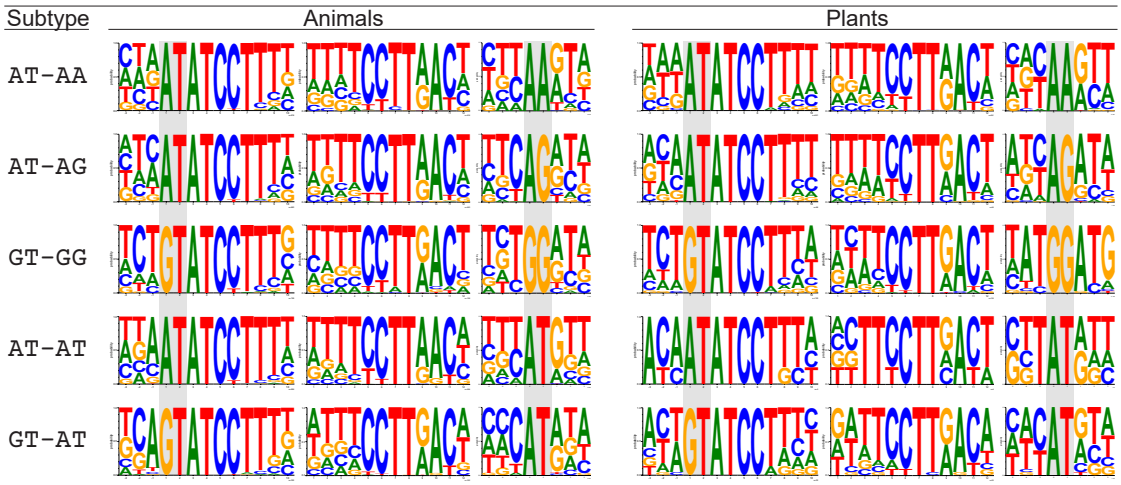
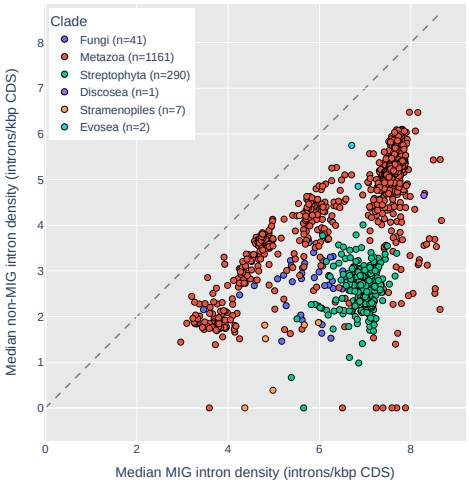


Figure 4.15: Sequence motifs of the 5′SS, BPS and 3′SS regions of non-canonical minor introns in animals and plants. The terminal dinucleotide pairs are highlighted in gray.

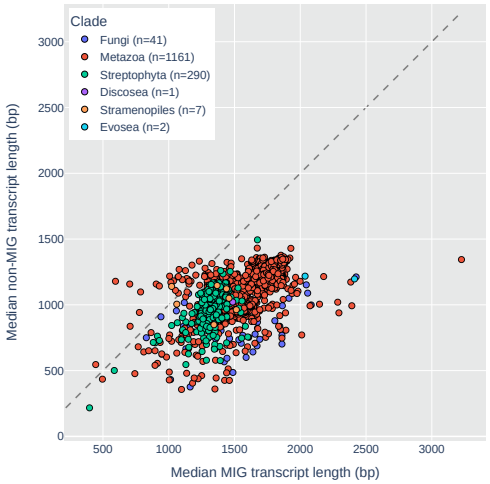
throughout genomes where they are found (Moyer et al., 2020; Levine and Durbin, 2001; Sheth et al., 2006; Alioto, 2007) a natural question to ask is, to what extent and in what ways are the genes where minor introns are found measurably different than those without minor introns? As far as we are aware, while this question has been addressed in a variety of other ways by different groups (Basu, Makalowski, et al., 2008; C. B. Burge, R A Padgett, and Sharp, 1998; Janice et al., 2012), relatively little attention has been paid to possible differences in a number of basic gene attributes, namely gene length (excluding introns) and number of introns per unit coding sequence or "genic intron density" (a coinage we will use here to distinguish from our more frequent usage in this paper of "intron density" to describe some number of introns in terms of their relative share of the total introns in the genome).

Strikingly, when we compare the genic intron density of minor intron-containing genes (MIGs) to all other genes in species with minor introns, we find that MIGs are universally more intron-dense than non-MIGs on average (Figure 4.16a). Furthermore, it appears that average MIG lengths (excluding intron sequences) are longer than other genes (Figure 4.16b). There are a number of cases where the median non-MIG gene length exceeds the median MIG gene length, although none of those differences are statistically significant (Mann-Whitney U test, $p \geq 0.05$).

An in-depth analysis of this qualitative finding is beyond the scope of the current paper, but it seems an underappreciated point of differentiation between the two intron types.



(a) Median genic intron density (introns/kbp coding sequence) in exclusively major intron genes (y-axis) vs. minor intron-containing genes (x-axis).



(b) Median gene length (sum of CDS) for exclusively major intron genes (y-axis) vs. minor intron-containing genes (x-axis).

4.4.8 Comparison of minor and major intron lengths

While a number of studies have compared the length distributions of the different intron types in a limited assortment of genomes (Levine and Durbin, 2001; Vinogradov, 1999; Moyer et al., 2020), without a large set of species containing minor introns to compare within it has been difficult to gauge the extent to which minor intron lengths might differ from major intron lengths. With the comprehensive minor intron data we have collected, we were able to ask a very basic question: what is the general relationship between average major and minor intron lengths? At a high level, the answer appears to be that the relationship is roughly linear (Figure 4.17)—species with longer average major intron length tend to also have longer average minor intron length (Spearman's $\rho = 0.625$ for median values, $p = 5.08 \times 10^{-16}$). One interesting aspect of the data in Figure 4.17 is

shown more clearly in the inset plot (which is simply the subset of the data in the main plot with length $\leq$ 1000 bp), and pertains to the species with significant minor intron loss (small markers) and large differences between average minor and major intron lengths. The set of species in that region is enriched for *Drosophila* (as it is a taxonomically over-represented genus in the sequence databases), but includes many additional insect species as well.

It is not clear what immediate conclusions there are to draw from this data, though some additional questions are raised: Were shorter minor introns especially selected against in these lineages for some reason, such that the only minor introns remaining are disproportionately long? What is driving variation within, for example, *Drosophila* such that in some species the difference between minor and major is relatively modest (*Drosophila busckii*, major=65 bp and minor=189 bp) and in others, it's much more stark (*Drosophila biarmipes*, major=77 bp and minor=677 bp)? It should be noted as well that for *Drosophila* specifically, almost all of the minor introns are conserved within the genus, so the previous example is made more interesting because 100% of the *D. busckii* minor introns are shared with *D. biarmipes*, yet are far longer in the latter than the former. It did occur to us to check whether minor introns in these outlier species happen to be (for whatever reason) in genes with longer-than-average intron size, and although we have not done so systematically we have checked a number of more extreme cases and have found the same pattern recapitulated between minor and major introns of the same genes. For example, in the black soldier fly *Hermetia illucens*, the median minor intron length is 4019 bp and the median major intron length is only 105 bp. Comparing minor to major within only the minor intron-containing genes changes things, but not qualitatively—the median major intron length becomes 399.5 bp, but the difference between minor and major is still significant ($p = 0.0025$ by one-tailed Mann-Whitney U test under the alternative hypothesis that minor intron lengths are longer).

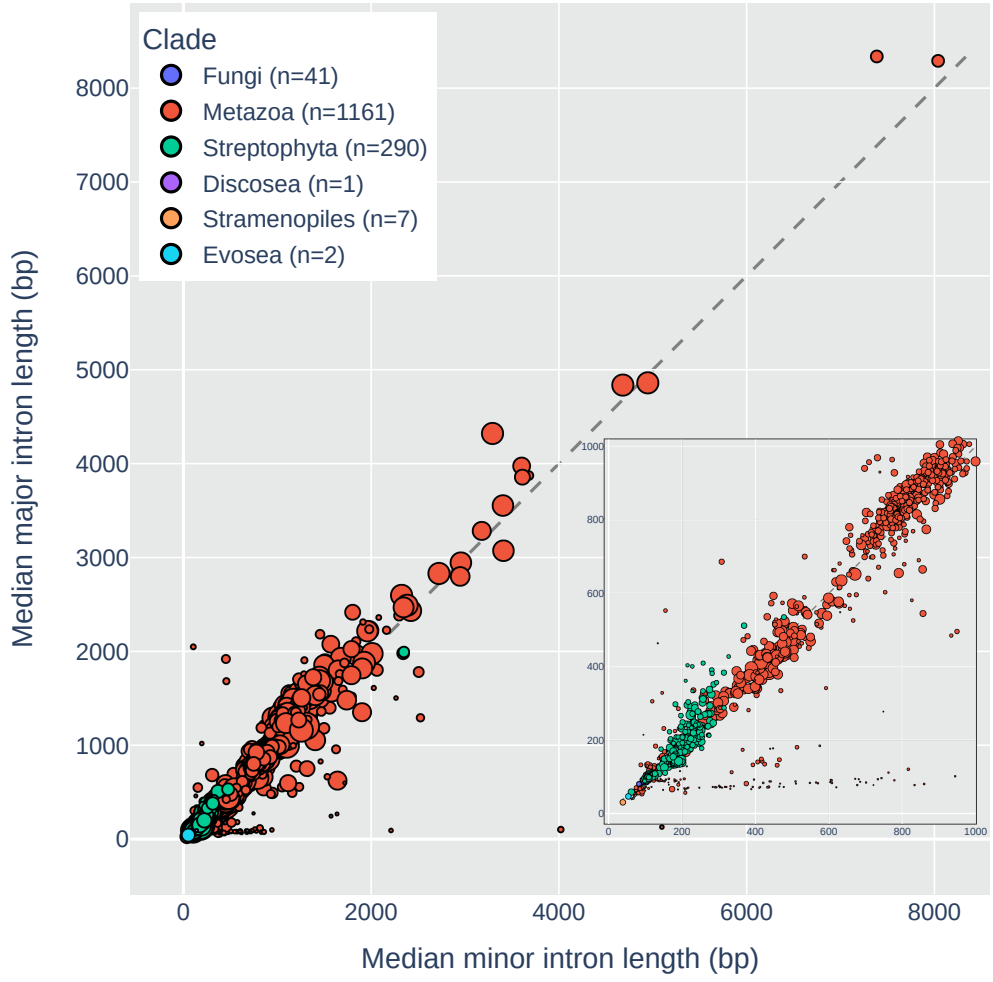


Figure 4.17: Median major intron length (y-axis) vs. median minor intron length (x-axis) for all species with high-confidence minor introns. Size of markers indicates number of minor introns in the genome. Inset: The subset of the data where the maximum median intron length is 1000 bp.

4.4.9 Reconstruction of ancestral minor intron densities

In an attempt to quantify some of the evolutionary dynamics leading to the variegated pattern of minor intron densities we see in extant lineages, we sought to estimate minor intron densities for certain ancestral nodes throughout the eukaryotic tree (see 4.3). For each selected node, we identified pairs of species for which the node is the most recent common ancestor and, in combination with an outgroup species, performed three-way protein-level alignments to allow us to identify intron states for each species within the alignments. Then, using the procedure described in (Scott W Roy and Walter Gilbert, 2005a), we calculated the number of minor and major introns estimated to have been present in the aligned regions in the ancestral genome, and repeated this process using many different combinations of species for each node to derive average values across all such comparisons. Because the absolute number of introns present in the aligned regions in the ancestor is not a particularly easy value to interpret, for reconstructions within a given kingdom we normalized the ancestral density of each intron type by a chosen reference species from that kingdom present in every alignment (see Materials and methods for details). The reference species for animals, fungi and plants were *Homo sapiens* (minor intron density 0.276%), *Rhizophagus irregularis* (minor intron density 0.272%) and *Lupinus angustifolius* (minor intron density 0.273%), respectively. Figure 4.18 shows both distributions of minor intron densities in constituent species from each terminal clade (violin plots), as well as estimates of ancestral minor intron densities at various nodes (colored boxes) as fractions of the density of minor introns in the aligned regions of the reference species (i.e., ancestral densities > 1 indicate minor intron enrichment relative to the reference species, and ancestral densities <1 indicate reduction).

As shown in Figure 4.18, ancestral minor intron densities were, in large part, modestly higher than minor intron densities in the relatively minor-intron-rich reference species, with the exception of a number of episodes of pronounced loss in the ancestors of Diptera, Pancrustacea and Zoopagomycota. The apparent enrichment of minor introns in the ancestor of Chelicerata is interesting, as it suggests there many have been some amount of minor intron gain along the path from the

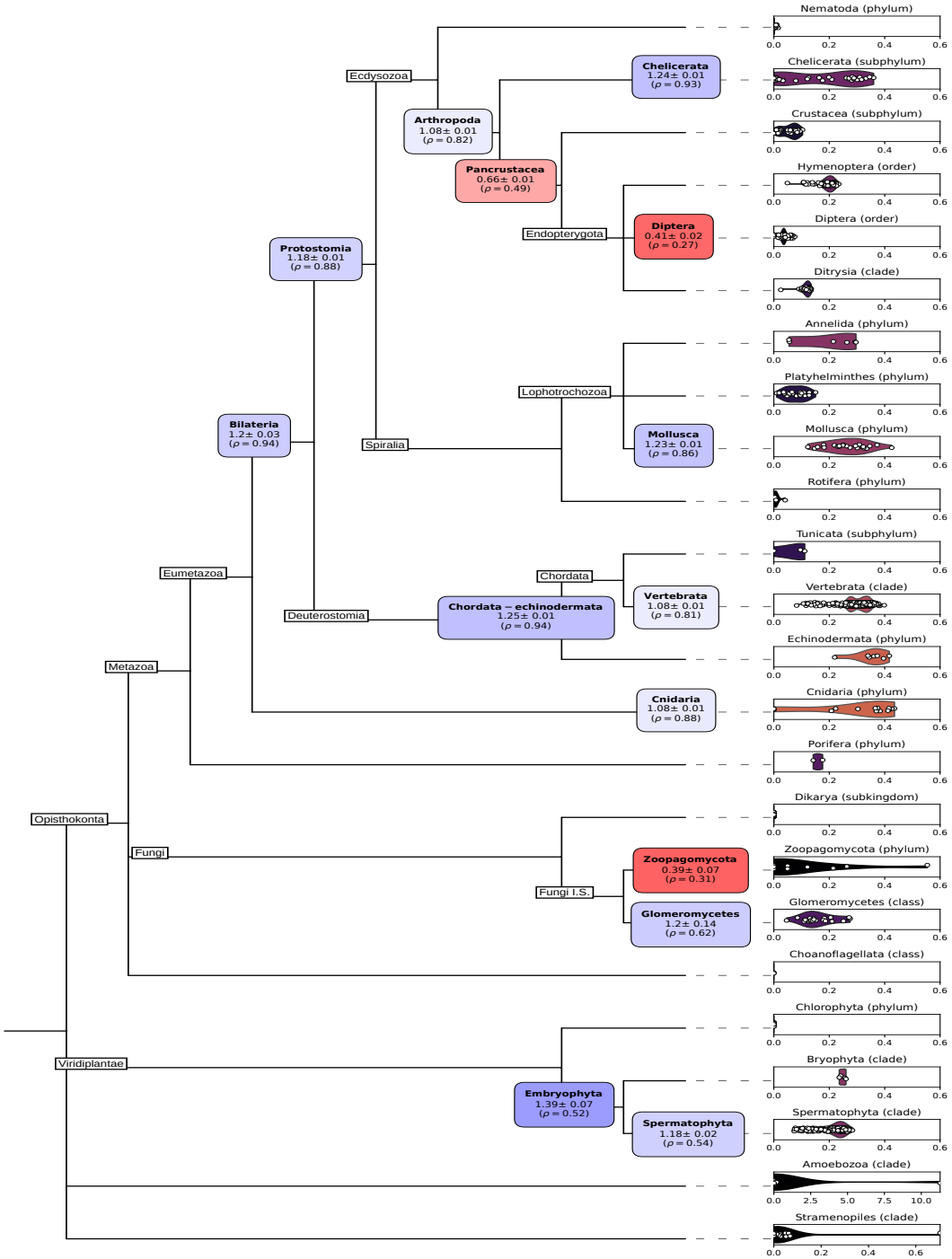


Figure 4.18: Minor intron density distributions in selected clades, and ancestral reconstructions of minor intron densities at selected nodes. Ancestral density node label color indicates relative enrichment (blue) or reduction (red) relative to the reference species in the alignments. For animals, the reference is *Homo sapiens*; for plants the reference is *Lupinus angustifolius*; for fungi the reference is *Rhizophagus irregularis*.

arthropod ancestor. This result needs to be qualified, however, by noting that in that region of the tree we were constrained by lack of available data to using only *Limulus polyphemus* for one of the two ingroup species, as well as the fact that in any given reconstruction, the calculated intron density is limited to the genes involved in the reconstruction. With similar caveats, the low inferred ancestral minor intron density of Zoopagomycota is notable as that group contains *Basidiobolus meristosporus*, which has the highest minor intron density so far discovered in fungi (0.554%). Overall, these results paint a picture of ancestral minor intron complements as generally analogous to those of minor intron rich extant species, and highlight the quixotic nature of minor intron loss dynamics throughout eukaryotic diversity. It would be interesting to have these results expanded upon once phylogenetic uncertainty has been reduced throughout the tree and even more diverse genomes are available for analysis.

4.5 Discussion

Over the last decade, and after many if not all of the most prominent papers examining minor introns in more than one species were published, there has been a marked increase in the number of annotated genomes publicly available for bioinformatic analysis. Ten years ago, for example, NCBI had annotated fewer than 60 genomes—it now lists over 800, counting only annotations performed by NCBI itself. The breadth of data now available enabled us to undertake a much more sweeping, if necessarily less focused, assessment of minor intron diversity than has ever been possible before, uncovering a wide variety of both novel and confirmatory information about minor intron dynamics across the eukaryotic tree.

Despite our best efforts to be as careful as possible in curating the data we have reported, as with any computational study of this scale there is bound to be some noise in the data, especially given our reliance on existing gene annotations derived from heterogeneous pipelines. One persistent issue in bioinformatic analyses of minor introns is the lack of a gold standard, empirically-verified set of minor intron sequences. While comparative genomics can do a great deal of heavy

lifting in this regard, it is often a time-consuming process at scale and the field in general would benefit greatly from a ground-truth set of minor introns based upon minor spliceosome profiling data or similar—we look forward to this type of data becoming available and being used to improve the accuracy of minor intron identification and as a result, our understanding of minor introns and their evolutionary dynamics.

We have shown for the first time the presence of substantial numbers of minor introns as well as minor spliceosomal snRNAs in a number of lineages thought to be lacking them including green algae, fungi and stramenopiles. In addition, we have produced results contradicting a number of longstanding results in the minor intron literature, and have highlighted various underappreciated differences between MIGs and other genes as well as broadening the scope of analyses done previously in more limited capacities. Aside from the merits of the specific analyses we have performed, we were drawn to this work in part because we hoped to be able to create a resource for the broader scientific community, and in that regard we hope that the detailed information we have collected here will help inform future exciting work on minor introns.

Chapter 5

A Minor Intron-Rich Fungus and Evidence That Neutral Evolution May Explain Biases in Minor Intron-Containing Genes

5.1 Abstract

Minor spliceosomal introns comprise ≤0.5% of all introns in the vast majority of eukaryotic genomes, and display a curious evolutionary pattern of high conservation in certain lineages (e.g., vertebrates) and dramatic reduction/wholesale loss in others (e.g., Diptera, nematodes). Some recent studies have shown associations between minor intron splicing and cell-cycle regulation and cellular differentiation in both animals and plants (Gault et al., 2017; Doggett et al., 2018; König et al., 2007; Meinke et al., 2020; Bai et al., 2019). To better assess whether these associations might represent an ancient functional role for minor splicing, we make use of the fungal species *Rhizophagus irregularis*, in which we report hundreds of previously-undescribed minor introns, to examine characteristics of minor splicing across different cell types in this lineage. Furthermore, we provide cursory evidence that the functional bias of minor introns described throughout the literature (C. B.

Burge, R A Padgett, and Sharp, 1998; Patel and Joan A Steitz, 2003; Turunen, Niemelä, et al., 2013) may be explained by neutral evolutionary processes.

5.2 Introduction

The existence of minor spliceosomal introns represents a long-standing puzzle of eukaryotic biology (C. B. Burge, R A Padgett, and Sharp, 1998; Rogozin et al., 2012). In many ways, minor spliceosomal introns resemble the better known major spliceosomal introns: they interrupt nuclear genes and require removal from RNA transcripts by a large molecular machinery comprising five RNA-protein complexes and diverse accessory proteins, many of which are even shared between minor and major machineries (Turunen, Niemelä, et al., 2013). However, the minor and major systems show quite different patterns and evolutionary histories: whereas major introns interrupt the majority of genes in most eukaryotes and are ubiquitous in many eukaryotes (Irimia and Scott William Roy, 2014), minor introns are orders of magnitude rarer (accounting for less than 1% of introns in humans), and have been repeatedly lost in diverse eukaryotic lineages (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008; Bartschat and Samuelsson, 2010; C.-F. Lin et al., 2010; Turunen, Niemelä, et al., 2013). While minor introns were almost certainly present in early eukaryotes (Russell et al., 2006) and are retained in a wide variety of eukaryotic lineages (M. D. Lopez, Alm Rosenblad, and Samuelsson, 2008; Moyer et al., 2020), to date only two lineages are known to retain more than a few dozen minor introns, namely animals and plants (Moyer et al., 2020). Interestingly, in contrast to the massive intron loss observed in many lineages, in some lineages minor intron complements are remarkably evolutionarily stable; for instance there is almost no loss within the evolutionary history of characterized vertebrates (C.-F. Lin et al., 2010). This contrasting pattern of retention versus massive loss raises a puzzle of minor spliceosomal intron function: if minor introns are not functionally important why are they almost entirely conserved over hundreds of millions of years in some lineages; yet if they are important, how can they be repeatedly decimated or lost entirely in other lineages? Two observations are particularly relevant to the

question of minor spliceosomal intron function.

First, over the past ten or so years, some studies have shown roles for minor introns in cellular differentiation, with decreased minor activity driving downregulation of minor intron-containing genes (MIGs) associated with cessation of cell cycling (Gault et al., 2017; Doggett et al., 2018; König et al., 2007). Most compellingly, a recent study showed that the splicing regulator SR10 is regulated at the level of minor splicing, with inefficient splicing leading to downregulation of other SR proteins whose pro-splicing activities promote cell cycle progression (Meinke et al., 2020). Interestingly, a negative association of minor splicing with cell differentiation has also been argued for in plants (Gault et al., 2017; Bai et al., 2019). This pattern is curious, given that the common ancestor of animals and plants is thought to have been unicellular and thus not to have undergone terminal differentiation (although recent findings of multicellular stages as well as differentiation-like processes across diverse eukaryotes may ultimately call this common assumption into question (Najle and Ruiz-Trillo, 2021; Brunet et al., 2019)).

Second, minor introns show functional biases, being disproportionately found in genes encoding various core cellular functions including DNA repair, RNA processing, transcription and cell cycle functions that largely appear to hold between plants and animals (Gault et al., 2017; C. B. Burge, R A Padgett, and Sharp, 1998) (though the strength of these associations is questioned somewhat in (Baumgartner, Olthof, et al., 2018)). Particularly given the above evidence that regulation of minor splicing regulates core cellular processes, these patterns would seem to be consistent with an ancient role for minor splicing in cell cycle regulation, that could have been secondarily recruited for multicellular differentiation separately in animals and plants. However, other explanations remain possible. What is needed to understand the evolutionary history and importance of minor spliceosomal introns is genomic and regulatory characterization of additional lineages with relatively large complements of minor spliceosomal introns. Here, we report our discovery of hundreds of minor spliceosomal introns in a mycorrhizal fungus, and characterize several features of the minor spliceosomal system across cell types.

5.3 Results

5.3.1 Unprecedented minor intron density in the fungus *Rhizophagus irregularis*

A broad bioinformatic survey of minor spliceosomal intron number and diversity across eukaryotes (to be presented elsewhere) revealed a large number of candidate minor spliceosomal introns in the mycorrhizal fungus *Rhizophagus* irregularis, a member of the Glomeromycota group of fungi. There was a clear correspondence between minor-versus-major spliceosomal sequence characteristics in the two primary differentiating parts of the introns, namely the 5′splice site and the 3′branchpoint structure (Figure 5.1a), and consensus sequence features closely followed those previously found in animals and plants (Figure 5.1b). A subset of minor introns were found at conserved gene positions with minor introns in other fungi, animals and plants (Figure 5.1c,d), further increasing our confidence that these introns represent bona fide minor spliceosomal introns. Searches of the genome further provided evidence for presence of many minor spliceosome-specific proteins as well as all four minor spliceosome-specific non-coding RNAs (U11, U12, U4atac, and U6atac) (Figure 5.1e,f). As found previously in animals and plants, we found a distinctive distribution of intron phase (position at which introns interrupt the coding codon series, whether between codons (phase 0) or after the first or second nucleotide of a codon (phase 1 and 2, respectively): whereas major introns typically show the pattern (ph0 > ph1 > ph2), minor introns in *R. irregularis* followed the minor pattern in animals and plants (ph1 > ph2 > ph0; (Levine and Durbin, 2001; Moyer et al., 2020)) (Figure 5.1g). In total, we predict that 199 introns in *R. irregularis* are minor-type (0.275% of 72,285 annotated introns), orders of magnitude higher than for other fungal species previously reported to contain minor introns (∼4 in *Rhizopus oryzae* and ∼20 in *Phycomyces blakesleeanus*) (Bartschat and Samuelsson, 2010).

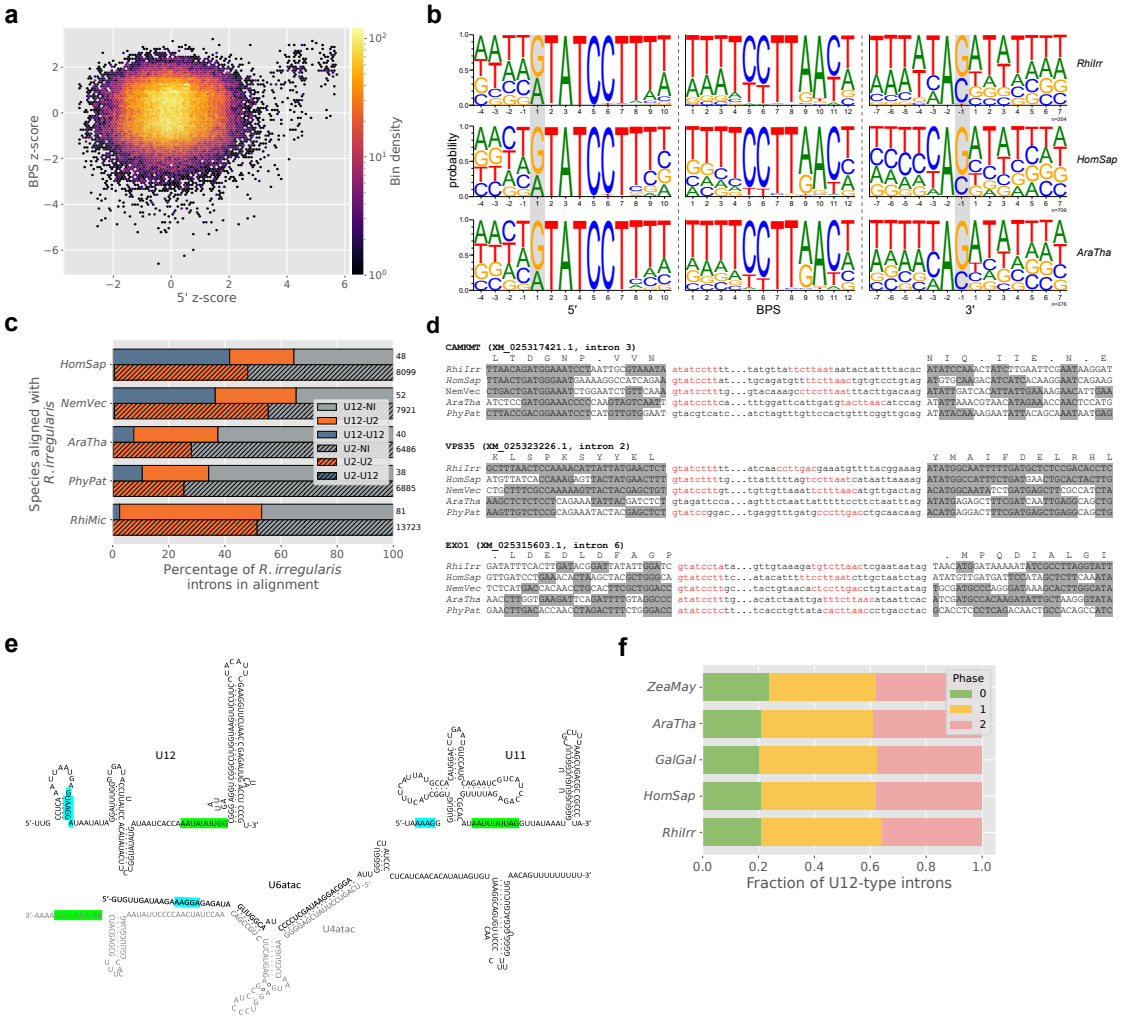


Figure 5.1: Evidence of minor introns and splicing machinery in *Physarum polycephalum*. (a) BPS vs. 5′SS scores for *Rhizophagus irregularis*, showing the expected cloud of introns with minor-intron-like 5′SS and BPS scores in the first quadrant. (b) Comparison of minor intron sequence motifs in *Rhizophagus*, human and *Arabidopsis*. (c) Conservation of *Rhizophagus* minor and major introns in different species. (d) Examples of minor introns in *Rhizophagus* in conserved alignments with minor introns in other species. (e) The four minor snRNAs U11, U12, U4atac and U6atac found in *Rhizophagus*. (f) Comparison of minor intron phase distributions in different species, showing the expected pattern in *Rhizophagus*. Species abbreviations are as follow: HomSap: *Homo sapiens*, NemVec: *Nematostella vectensis*, AraTha: *Arabidopsis thaliana*, PhyPat: *Physcomitrium patens*, RhiMic: *Rhizopus microsporus*, ZeaMay: *Zea mays*, GalGal: *Gallus gallus*.

5.3.2 No evidence for increased minor splicing in proliferating cells

We next sought to test whether *R. irregularis*, like animals and plants, upregulates splicing of minor introns in proliferating cells. We used published transcriptomic data from five cell types (four replicates each), andassessed likely proliferation profiles of the six cell types using the previously published proliferation index (PI) approach. Briefly, we first identified putative orthologs of genes known to be associated with cell proliferation in humans. For each such putative PI ortholog, z-scores were calculated for all 20 samples, and those z-scores were then used for comparison across cell types as well as for comparisons within cell types between putative PI orthologs and other genes. This allowed us to calculate relative proliferation scores for all five cell types. While $\frac{4}{5}$ cell types showed similar PI values, one cell type, immature spores, showed substantially and significantly higher values (Figure 5.2a), a pattern that also held when we look at the more straightforward metric of adjusted FPKM values (Figure 5.3). This overall significance notwithstanding, it should be noted that only a small fraction of genes included in the PI individually showed significant differences in expression between cell types. In addition, we noted that many non-PI genes are also overexpressed in immature spores relative to other cell types; while one interpretation of this result is that it reflects generally more active gene expression in proliferating cells, it does provide a caveat for the overall strength of the observed difference.

We then tested the association between markers of minor spliceosomal activity and these proliferation scores. We first looked for systematic differences in overall gene expression of MIGs between cell types with different proliferation scores, using various approaches. First, using the same z-score based approach as for the proliferation score (though with MIGs instead of putative PI orthologs), we found that MIGs were in fact more highly expressed in cell types with higher proliferation scores (Figure 5.2b). On the other hand, we found that very few MIGs reached significant levels of differential expression, and were in fact underrepresented among genes that showed significant differential expression in multiple comparisons between cell types of different proliferation index scores (e.g., 4.2%

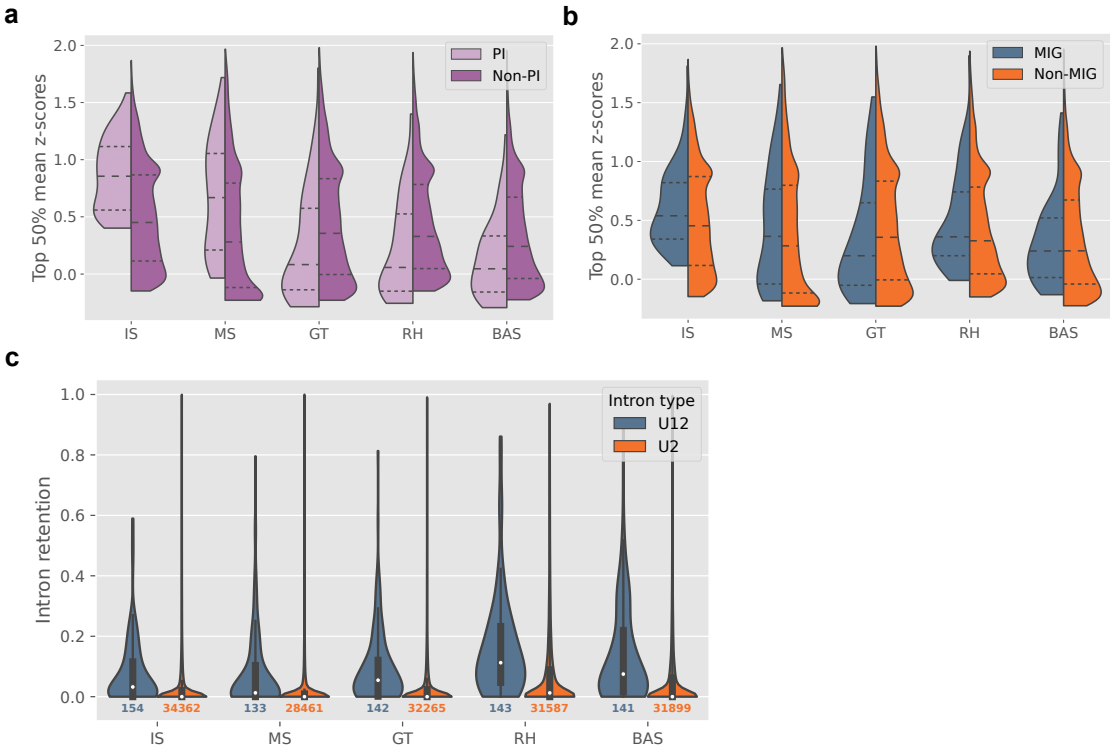


Figure 5.2: (a) Comparison of expression of proliferation-index genes (PI, light purple) and all other genes (Non-PI, dark purple) across cell types, n=70 PI and n=9276 non-PI in each cell type. (b) As in (a), but for minor intron-containing genes (MIGs) compared to non-MIGs; n=96 MIG and n=9249 non-MIG for each cell type. (c) Intron retention values across cell types for U12- (blue, left) and U2-type (orange, right) introns. Cell types are labeled as described in the text.

of minor intron-containing genes compared to 21.9% of other genes in the IS-MS comparison). In total, these results suggest that expression of MIGs shows a detectable but only moderate association with proliferation index in *R. irregularis*, in contrast to the robust results previously observed in humans.

We next compared the efficiency of minor intron splicing between cell types. Contrary to our hypothesis that minor splicing would be more active in proliferating cells, we found that minor intron retention was in fact significantly (though only modestly) higher in proliferating cells (Figure 5.2c). This result held whether we used z-score-based metrics or the intron retention values themselves, and whether we used splicing efficiency or intron retention as our metric. We also assessed expression of the minor splicing machinery itself (i.e., the known components of the minor spliceosome). In comparisons between immature spores and other cell types, no component individually showed higher expression, however collectively the machinery was 3.5x more highly expressed in immature spores than other cell types, reaching significance when considered collectively. However, the major spliceosomal machinery also showed a similar pattern (with 5x higher expression), and as such it seems that lower expression of the minor splicing machinery could be part of a larger pattern of up/regulation of core molecular functions in proliferating/quiescent cells.

The observed association between minor intron splicing and cell proliferation in animals resonates with the long-standing finding that minor introns are overrepresented in genes involved in core cellular processes. Given that minor intron splicing in *Rhizophagus* does not appear to be associated with cell proliferation, we probed these patterns more deeply.

Gene ontology analysis of *Rhizophagus* MIGs revealed a curious pattern in which GO results were highly dependent on the control dataset used. Because of the dearth of *Rhizophagus* functional annotations, GO analyses were necessarily run by identifying human orthologs of *Rhizophagus* MIGs. When GO analysis was run on these orthologs as a subset of all human genes, a number of overrepresented functional categories were found, in large part mirroring results for humans. However, we realized that there is a potential bias in this analysis: all human genes

Table 5.1: GO term enrichment for MIGs in *Rhizophagus*, compared to all human-*Rhizophagus* orthologs. E: expected, O/U: over/under, FE: fold enrichment, FDR: false-discovery rate.

GO term	Hs-Ri	RiMIGs	E	O/U	FE	FDR
vesicle-mediated transport (GO:0016192)	288	33	12.82	+	2.57	1.16E-02
intracellular transport (GO:0046907)	434	39	19.32	+	2.02	3.00E-02
establishment of localization in cell (GO:0051649)	476	42	21.19	+	1.98	2.84E-02
small molecule metabolic process (GO:0044281)	549	5	24.44	-	.20	6.81E-03
carboxylic acid metabolic process (GO:0019752)	309	1	13.75	-	.07	3.41E-02
oxoacid metabolic process (GO:0043436)	314	1	13.98	-	.07	4.43E-02
organic acid metabolic process (GO:0006082)	319	1	14.20	-	.07	3.34E-02

present in the *Rhizophagus* MIG ortholog set have *Rhizophagus* orthologs, thus excluding most human genes (only 14%, 3190/23257, had identified *Rhizophagus* orthologs), and in particular animal-specific genes. Remarkably, when we limited our GO analysis control group to human genes with *Rhizophagus* orthologs, we found much less functional overrepresentation (Table 5.1).

Notably, a similar concern applies to human MIGs in general: because nearly all human minor introns are quite old, human MIGs are commensurately old, which could drive functional correlations given known differences in functional categories between genes of different ages. Indeed, when we performed a GO analysis of human MIGs with *Rhizophagus* orthologs, limiting the reference set to human genes with *Rhizophagus* orthologs (a rough surrogate for gene age given that, unlike baker's yeast, *Rhizophagus* may not have lost many ancestral genes (Sales-Lee et al., 2021)), we found a much lower degree of functional enrichment

Table 5.2: GO term enrichment for human MIGs with *Rhizophagus* orthologs, compared to all human-*Rhizophagus* orthologs. E: expected, O/U: over/under, FE: fold enrichment, FDR: false-discovery rate.

GO term	Hs-Ri	HsMIGs w/RiO	E	O/U	FE	FDR
intracellular transport (GO:0046907)	434	54	29.65	+	1.82	4.27E-02
small molecule metabolic process (GO:0044281)	549	12	37.51	-	.32	1.67E-03
carboxylic acid metabolic process (GO:0019752)	309	2	21.11	-	.09	9.13E-04
oxoacid metabolic process (GO:0043436)	314	2	21.45	-	.09	8.72E-04
organic acid metabolic process (GO:0006082)	319	2	21.79	-	.09	1.12E-03

(Table 5.2). These results support the conclusion that the long-standing result that minor introns are functionally overrepresented in core cellular processes may be largely explained by the fact that minor introns fall primarily in evolutionarily older genes, which are overrepresented in core cellular functions. Interestingly, when we compared all human MIGs (given that minor intron presence strongly suggests that a gene is ancient) to human genes with *Rhizophagus* orthologs, we did see a significant number of overrepresented functional categories (`https://doi.org/10.6084/m9.figshare.20483841`). It is not entirely clear why all MIGs, but not MIGs with *Rhizophagus* orthologs, show substantial functional differences relative to all genes with *Rhizophagus* orthologs. Insofar as MIGs are ancient genes, MIGs without *Rhizophagus* orthologs likely represent losses in fungi; gene losses are likely to be functionally biased, perhaps explaining the observed pattern.

5.4 Discussion

5.4.1 A relatively simple organism with a large number of minor introns

Minor introns are primarily found in animals and plants and have been implicated in specifically multicellular phenomena, in particular cell differentiation, leading to the idea that minor introns are closely associated with organismal complexity (though that these minor introns' by-and-large appear to date to early, likely unicellular, eukaryotic ancestors complicates this narrative). Here, we report a mycorrhizal fungus with only a few cell types and a simple body plan, whose genome contains over 199 minor spliceosomal introns. Whereas two previously reported instances of relatively high minor intron densities outside of animals and plants—in the slime mold *Physarum polycephalum* (Larue, Eliáš, and Scott W Roy, 2021) and the protist *Blastocystis* (Gentekaki et al., 2017)—appear to be largely due to secondary minor intron creation ((Larue, Eliáš, and Scott W Roy, 2021), section 4.4 and G.E.L and S.W.R., unpublished data), the *R. irregularis* minor intron content likely largely reflects retention of ancestral introns, as evidenced by evolutionary conservation of the minor intron positions as well as functional biases of minor intron-containing genes that largely echo those in animals and plants.

Insofar as the major determinant of modern minor intron number across species appears to be the degree of minor intron loss from an ancestral complement, a major contributor to modern densities may be overall evolution rate, with faster-evolving lineages having lost more introns. If so, then it may be that *Rhizophagus*, like multicellular animals and plants, owes its large minor intron complement mostly to generally slower genomic evolution, as also could be the case for major intron complement, with modern intron densities larger in lineages with lower intron loss rates (e.g., (Carmel et al., 2007)).

5.4.2 Neutral evolution can explain functional biases in minor intron-containing genes

Two patterns of minor intron distribution have long lay in tension. On the one hand, minor introns' overrepresentation in genes with certain functions suggests a regulatory function for minor splicing, and conservation of many of these functional biases from animals to plants suggests these functions are ancient. On the other hand, minor introns have been lost en masse or entirely many times in eukaryotic evolution, suggesting their expendability.

Here, we show that functional biases in MIGs are almost entirely explained by biases of gene age: because most minor introns are old, most MIGs are old, and core cellular processes are overrepresented among old genes. This suggests that minor intron distributions across genes could simply reflect largely unbiased minor intron gain into ancestral genomes, and then a lack of minor gain in more recent times. It would be interesting to explore whether such a schema for the evolution of a functional bias without selection is relevant to other age-stratified phenomena.

5.4.3 How did splicing of animal minor introns become associated with cell cycle progression?

Prior to the current work, we perceived a chicken and egg problem of functional biases among MIGs (C. B. Burge, R A Padgett, and Sharp, 1998; C.-F. Lin et al., 2010) and control of cell proliferation by regulation of minor splicing ((Gault et al., 2017; Meinke et al., 2020; Bai et al., 2019): that is, how could the regulatory control evolve without the functional bias, by why would the functional bias evolve without the regulatory function? We thus sought to illuminate this question by studying a third minor intron-rich lineage. The current findings that the observed functional biases appear to be largely explained by minor introns' bias towards older genes, and older genes bias towards core cellular functions, suggest an answer. Thus, functional biases could have initially evolved due to these gene age biases, and this functional bias could then have secondarily been recruited to regulate cell proliferation in animals in plants.

While this scenario makes sense schematically, is remains a remarkable contention that decreased minor splicing could evolve a function in cell regulation; insofar as MIGs represent a quasi-random subset of ancient genes, it seems likely that a global reduction in minor splicing would have a wide variety of impacts, many of them likely costly. Thus how failure to process a quasi-randomly chosen set of ancestral genes could evolve as a regulatory mechanism remains puzzling, and will require additional work across diverse minor intron-containing lineages.

5.4.4 Limitations of the study

Possible caveats of this study arise from two surrogates that we have employed. First, to assess cell cycle activity/proliferation of cell types, we have used orthologs of human genes associated with proliferation. The possibility of turnover of gene expression patterns raises the concern that these genes are not an appropriate gene set to assess proliferation. Indeed, while clear statistical differences in proliferation are seen when PI genes are viewed collectively, only a small fraction (∼5-10%) individually show significantly different expression between cell types. However, similar comparisons with model fungi attest to generally good conservation of genes' association with proliferation, consistent with an ancient core of cell cycle regulation. Second, we have used available transcriptomic data not specifically generated for the purposes of comparing proliferation, potentially leading to noise in the data. However, the general pattern observed, in which developing spores show the highest proliferation index, mirrors intuitive expectations, suggesting that our proliferation scores are capturing at least some of the relevant biological phenomena. Testing of transcriptomic effects of direct manipulations of cell cycle would be very useful to confirm (or refute) our results.

5.5 Concluding remarks

Our results do not support the emerging dominant hypothesis for the existence of minor introns, namely that minor introns provide a means for regulation of cell cycle progression. The reported lack of cell cycle-regulated minor intron splicing in

fungi suggests that this association is not a general phenomenon, correspondingly weakening the hypothesis that such a function could explain the persistence of minor introns across eukaryotes generally. However, given the possibility that it may be fungi that are atypical, having secondarily lost this function, discovery and study of additional minor intron-rich lineages is a priority, as is development and testing of alternative hypotheses for the origins and functional biases of minor intron-containing genes.

5.6 Methods

5.6.1 Identification of minor introns

The RefSeq genome and annotation for *Rhizophagus irregularis* were downloaded from NCBI, and annotated introns were extracted and classified as major or minor using intronIC v1.3.2 (Moyer et al., 2020) (`https://github.com/glarue/intronIC`). Only introns defined by annotated CDS features were included for analysis.

5.6.2 Differential gene expression

Single-end RNA-seq reads from previously-published cell-type-specific sequencing of *Rhizophagus* irregularis (Kameoka et al., 2019) (four biological replicates per cell type) were pseudoaligned to a decoy-aware version of the transcriptome using Salmon v1.6.0 (Patro et al., 2017) (with non-default arguments `--seqBias --softclip`). The Salmon output was then formatted with tximport v1.14.2 (Soneson, Love, and Robinson, 2015), and differential gene expression (DGE) analysis was performed using DESeq2 v1.26.0 (Love, Huber, and Anders, 2014) with the following arguments: `test="LRT"`, `useT=TRUE`, `minReplicatesForReplace=Inf`, `minmu=1e-6`, `reduced=~1`. For each pairwise combination of cell types, genes with significant DGE values (Wald p-value < 0.05) were retained for further analysis.

5.6.3 Z-score metric

Following the methodology used by Sandberg et al. to assign a proliferation index to cell types (Sandberg et al., 2008), z-scores were calculated per feature (whether for gene expression or intron retention/splicing efficiency) across all cell-type replicates (n=20), and then summarized for each cell type by the mean value of the corresponding replicate z-scores (departing from the reference method in this aspect). Prior to conversion to z-scores, the raw gene expression data was normalized by running the output from `tximport` through the `fpkm()` function in `DESeq2`. For group z-score comparisons (e.g., proliferation-index genes, minor introns vs. major introns), the median of the top 50% of z-scores from each group was used. As the z-score calculation requires there to be variation across samples, certain genes/introns were necessarily omitted under this metric.

5.6.4 Intron retention and splicing efficiency

For each RNA-seq sample, `IRFinder-S v2.0` (Middleton et al., 2017) was used to compute intron retention levels for all annotated introns. Introns with warnings of "LowSplicing" and "LowCover" were excluded from downstream analyses. Across replicates within each cell type, a weighted mean retention value was calculated for each intron, with weights derived by combining the average number of reads supporting the two intron-exon junctions and the total number of reads supporting the exon-exon junction.

Intron splicing efficiency was calculated as previously described (Larue, Eliáš, and Scott W Roy, 2021). Briefly, RNA-seq reads were mapped to splice-junction sequence constructs using `Bowtie` v1.2.3 (Langmead, 2010) (excluding multiply-mapping reads using the non-default argument `-m 1`). Introns with fewer than five reads supporting either the corresponding exon-exon junction or one of the intron-exon junctions (or both) were excluded. For each intron, the proportion of reads mapped to the intron-exon junction(s) versus the exon-exon junction was used to assign a splicing efficiency value for each sample (see reference for details). Within each cell type, the weighted mean of replicate splicing efficiency values for each intron was calculated in the same manner as for intron retention.

5.6.5 Spliceosome-associated gene expression

Orthologs of human spliceosome components were found in *Rhizophagus irregularis* via a reciprocal-best-hit approach (`https://github.com/glarue/reciprologs`) using `BLAST` v2.9.0+ (Camacho et al., 2009) with an E-value cutoff of 1×10^{-10}. Four genes from each splicing system (major and minor) were identified in *Rhizophagus* by this approach, consisting of orthologs to human minor spliceosome genes ZMAT5 (U11/U12-20K), RNPC3 (U11/U12-65K), SNRNP35 (U11/U12-35K), and SNRNP25 (U11/U12-25K) and major spliceosome genes SF3A1 (SF3a120), SF3A3 (SF3a60), SNRNP70 (U1-70K) and SNRPA1 (U2 A′). Gene expression values generated by Salmon for each set of genes in each cell type were averaged across replicates, and pairwise comparisons between cell types were made for the same set of genes (e.g., minor spliceosome genes in IS vs. MS). The significance of differences in expression between paired gene sets from different cell types was assessed using a Wilcoxon signed-rank test, with p-values corrected for multiple testing by the Benjamini-Hochberg method.

5.7 Supplementary materials

5.7.1 Supplementary figures

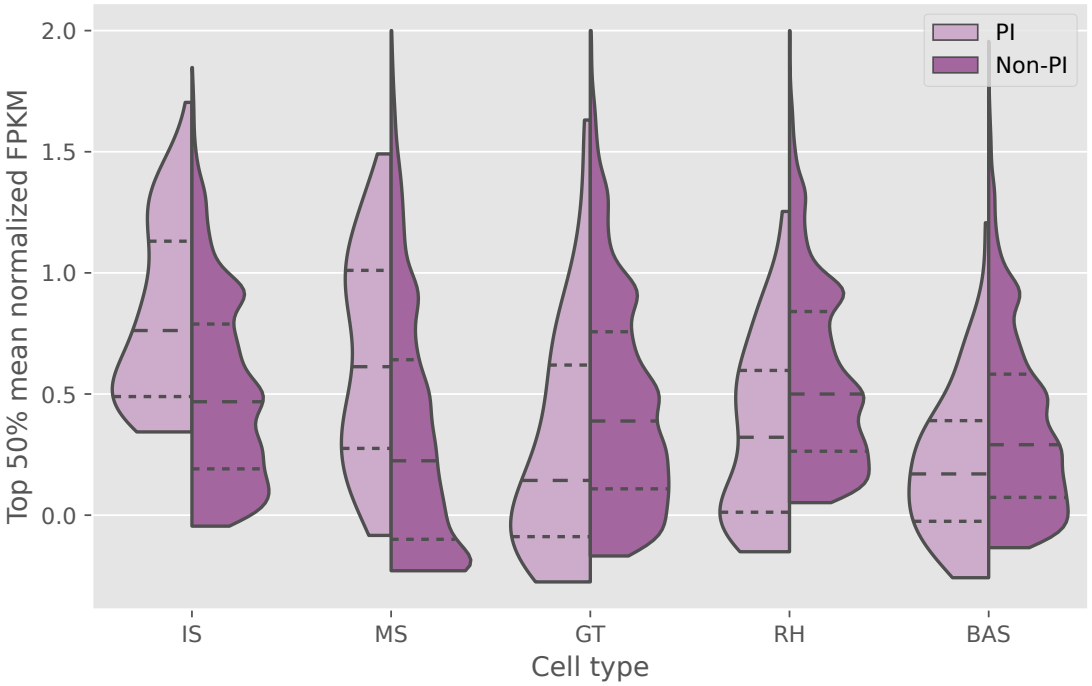


Figure 5.3: Comparison of expression (normalized FPKM from `DESeq2`, power-transformed) of proliferation-index genes (PI, light purple) and all other genes (Non-PI, dark purple) across cell types.

Bibliography

Abou Alezz, Monah et al. (2020). "GC-AG Introns Features in Long Non-coding and Protein-Coding Genes Suggest Their Role in Gene Expression Regulation". In: *Front. Genet.* 11, p. 488. ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00488. URL: https://www.frontiersin.org/article/10.3389/fgene.2020.00488.

Alioto, Tyler S (Jan. 2007). "U12DB: a database of orthologous U12-type spliceosomal introns". en. In: *Nucleic Acids Res.* 35.Database issue, pp. D110–D115. ISSN: 0305-1048. DOI: 10.1093/nar/gkl796. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl796.

Altschul, S F et al. (Oct. 1990). "Basic local alignment search tool". In: *J. Mol. Biol.* 215.3, pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2. URL: http://www.sciencedirect.com/science/article/pii/S0022283605803602.

Bai, Fang et al. (Mar. 2019). "RNA Binding Motif Protein 48 Is Required for U12 Splicing and Maize Endosperm Differentiation". en. In: *Plant Cell* 31.3, pp. 715–733. ISSN: 1040-4651, 1532-298X. DOI: 10.1105/tpc.18.00754. URL: http://dx.doi.org/10.1105/tpc.18.00754.

Barrantes, Israel, Jeremy Leipzig, and Wolfgang Marwan (2012). "A next-generation sequencing approach to study the transcriptomic changes during the differentiation of Physarum at the single-cell level". In: *Gene Regul. Syst. Bio.* 2012.6, pp. 127–137. ISSN: 1177-6250. DOI: 10.4137/GRSB.S10224. URL: http://dx.doi.org/10.4137/GRSB.S10224.

Bartschat, Sebastian and Tore Samuelsson (Feb. 2010). "U12 type introns were lost at multiple occasions during evolution". en. In: *BMC Genomics* 11.1,

p. 106. ISSN: 1471-2164. DOI: 10.1186/1471-2164-11-106. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-106.

Basu, Malay Kumar, Wojciech Makalowski, et al. (Jan. 2008). "U12 intron positions are more strongly conserved between animals and plants than U2 intron positions". In: *Biol. Direct* 3, p. 19. ISSN: 1745-6150. DOI: 10.1186/1745-6150-3-19. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2426677%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract.

Basu, Malay Kumar, Igor B Rogozin, and Eugene V Koonin (Nov. 2008). "Primordial spliceosomal introns were probably U2-type". en. In: *Trends Genet.* 24.11, pp. 525–528. ISSN: 0168-9525. DOI: 10.1016/j.tig.2008.09.002. URL: http://www.sciencedirect.com/science/article/B6TCY-4TJC7BS-1/2/574a92bfa57467305d46f05ee3b7e99d%5Cnhttp://ac.els-cdn.com/S0168952508002308/1-s2.0-S0168952508002308-main.pdf?_tid=3c7e996c-1343-11e2-874f-00000aab0f26&acdnat=1349919207_8cd254660d3f63ab8a4f6a.

Baumgartner, Marybeth, Kyle Drake, and Rahul N Kanadia (Nov. 2019). "An Integrated Model of Minor Intron Emergence and Conservation". en. In: *Front. Genet.* 10, p. 1113. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.01113. URL: http://dx.doi.org/10.3389/fgene.2019.01113.

Baumgartner, Marybeth, Anouk M Olthof, et al. (Aug. 2018). "Minor spliceosome inactivation causes microcephaly, owing to cell cycle defects and death of self-amplifying radial glial cells". en. In: *Development* 145.17. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.166322. URL: http://dx.doi.org/10.1242/dev.166322.

Behringer, Megan G and David W Hall (2016). "Selection on position of nonsense codons in Introns". In: *Genetics* 204.3, pp. 1239–1248. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.116.189894. URL: http://dx.doi.org/10.1534/genetics.116.189894.

Berget, S M, C Moore, and P A Sharp (Aug. 1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA". en. In: *Proc. Natl. Acad. Sci. U. S. A.*

74.8, pp. 3171–3175. ISSN: 0027-8424. DOI: 10.1073/pnas.74.8.3171. URL: http://dx.doi.org/10.1073/pnas.74.8.3171.

Bhasi, Ashwini et al. (Jan. 2009). "ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes". en. In: *Nucleic Acids Res.* 37.Database issue, pp. D703–11. ISSN: 0305-1048.

Bicknell, Alicia A et al. (2012). "Introns in UTRs: Why we should stop ignoring them". In: *Bioessays* 34.12, pp. 1025–1034. ISSN: 0265-9247. DOI: 10.1002/bies.201200073. URL: http://dx.doi.org/10.1002/bies.201200073.

Brack, C and S Tonegawa (Dec. 1977). "Variable and constant parts of the immunoglobulin light chain gene of a mouse myeloma cell are 1250 nontranslated bases apart". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5652–5656. ISSN: 0027-8424. DOI: 10.1073/pnas.74.12.5652. URL: http://dx.doi.org/10.1073/pnas.74.12.5652.

Braunschweig, Ulrich et al. (2014). "Widespread intron retention in mammals functionally tunes transcriptomes". In: *Genome Res.* 24.11, pp. 1774–1786. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.177790.114. URL: http://dx.doi.org/10.1101/gr.177790.114.

Breathnach, R and P Chambon (1981). "Organization and expression of eucaryotic split genes coding for proteins". en. In: *Annu. Rev. Biochem.* 50, pp. 349–383. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.50.070181.002025. URL: http://dx.doi.org/10.1146/annurev.bi.50.070181.002025.

Brunet, Thibaut et al. (Oct. 2019). "Light-regulated collective contractility in a multicellular choanoflagellate". en. In: *Science* 366.6463, pp. 326–334. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aay2346. URL: http://dx.doi.org/10.1126/science.aay2346.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson (Jan. 2015). "Fast and sensitive protein alignment using DIAMOND". en. In: *Nat. Methods* 12.1, pp. 59–60. ISSN: 1548-7091.

Buckley, Peter T et al. (2011). "Cytoplasmic Intron Sequence-Retaining Transcripts Can Be Dendritically Targeted via ID Element Retrotransposons". In:

Neuron 69.5, pp. 877–884. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.02.028. URL: http://dx.doi.org/10.1016/j.neuron.2011.02.028.

Burge, C and P A Sharp (1997). "Classification of introns: U2-type or U12-type". In: *Cell* 91.7, pp. 875–879. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)80479-1. URL: http://dx.doi.org/10.1016/S0092-8674(00)80479-1.

Burge, C B, R A Padgett, and P A Sharp (Dec. 1998). "Evolutionary fates and origins of U12-type introns". en. In: *Mol. Cell* 2.6, pp. 773–785. ISSN: 1097-2765. DOI: 10.1016/S1097-2765(00)80292-0. URL: http://dx.doi.org/10.1016/S1097-2765(00)80292-0.

Burke, Jordan E et al. (May 2018). "Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution". en. In: *Cell* 173.4, 1014–1030.e17. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.03.020. URL: http://dx.doi.org/10.1016/j.cell.2018.03.020.

Burset, M, I A Seledtsov, and V V Solovyev (Nov. 2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes". en. In: *Nucleic Acids Res.* 28.21, pp. 4364–4375. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/28.21.4364. URL: http://dx.doi.org/10.1093/nar/28.21.4364.

— (Jan. 2001). "SpliceDB: database of canonical and non-canonical mammalian splice sites". en. In: *Nucleic Acids Res.* 29.1, pp. 255–259. ISSN: 0305-1048.

Camacho, Christiam et al. (Dec. 2009). "BLAST+: architecture and applications". en. In: *BMC Bioinformatics* 10, p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421. URL: http://dx.doi.org/10.1186/1471-2105-10-421.

Carmel, Liran et al. (2007). "Patterns of intron gain and conservation in eukaryotic genes". In: *BMC Evol. Biol.* 7, p. 192. ISSN: 1471-2148. DOI: 10.1186/1471-2148-7-192. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2151770&tool=pmcentrez&rendertype=abstract.

Casper, Jonathan et al. (Jan. 2018). "The UCSC Genome Browser database: 2018 update". en. In: *Nucleic Acids Res.* 46.D1, pp. D762–D769. ISSN: 0305-1048.

Castillo-Davis, Cristian I et al. (2002). "Selection for short introns in highly expressed genes". In: *Nat. Genet.* 31.4, pp. 415–418. ISSN: 1061-4036. DOI: 10.1038/ng940. URL: http://dx.doi.org/10.1038/ng940.

Chamary, Jean-Vincent and Laurence D Hurst (May 2005). “Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else?” en. In: *Trends Genet.* 21.5, pp. 256–259. ISSN: 0168-9525.

Chen, Weijun and Melissa J Moore (Feb. 2014). “The spliceosome: disorder and dynamics defined”. en. In: *Curr. Opin. Struct. Biol.* 24, pp. 141–149. ISSN: 0959-440X.

Chorev, Michal and Liran Carmel (2012). “The function of introns”. In: *Front. Genet.* 3.APR, pp. 1–15. ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00055. URL: http://dx.doi.org/10.3389/fgene.2012.00055.

Chorev, Michal, Lotem Guy, and Liran Carmel (Jan. 2016). “JuncDB: an exon-exon junction database”. en. In: *Nucleic Acids Res.* 44.D1, pp. D101–9. ISSN: 0305-1048.

Chow, L T et al. (Sept. 1977). “An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA”. en. In: *Cell* 12.1, pp. 1–8. ISSN: 0092-8674. DOI: 10.1016/0092-8674(77)90180-5. URL: http://dx.doi.org/10.1016/0092-8674(77)90180-5.

Chung, Betty Y W et al. (May 2006). “Effect of 5’UTR introns on gene expression in Arabidopsis thaliana”. en. In: *BMC Genomics* 7, p. 120. ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-120. URL: http://dx.doi.org/10.1186/1471-2164-7-120.

Churbanov, Alexander et al. (July 2008). “Accumulation of GC donor splice signals in mammals”. en. In: *Biol. Direct* 3, p. 30.

Cohen, Noa E, Roy Shen, and Liran Carmel (Jan. 2012). “The role of reverse transcriptase in intron gain and loss mechanisms”. en. In: *Mol. Biol. Evol.* 29.1, pp. 179–186. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msr192. URL: http://dx.doi.org/10.1093/molbev/msr192.

Cologne, Audric et al. (Sept. 2019). “New insights into minor splicing-a transcriptomic analysis of cells derived from TALS patients”. en. In: *RNA* 25.9, pp. 1130–1149. ISSN: 1355-8382.

Cortes, Corinna and Vladimir Vapnik (Sept. 1995). “Support-vector networks”. In: *Mach. Learn.* 20.3, pp. 273–297. ISSN: 0885-6125. DOI: 10.1007/BF00994018. URL: https://doi.org/10.1007/BF00994018.

Cunningham, Clifford W (1999). “Some Limitations of Ancestral Character-State Reconstruction When Testing Evolutionary Hypotheses”. In: *Syst. Biol.* 48.3, pp. 665–674. ISSN: 1063-5157, 1076-836X. URL: http://www.jstor.org/stable/2585333.

De Conti, Laura, Marco Baralle, and Emanuele Buratti (Jan. 2013). “Exon and intron definition in pre-mRNA splicing”. en. In: *Wiley Interdiscip. Rev. RNA* 4.1, pp. 49–60. ISSN: 1757-7004. DOI: 10.1002/wrna.1140. URL: http://dx.doi.org/10.1002/wrna.1140.

Derr, L K and J N Strathern (Jan. 1993). “A role for reverse transcripts in gene conversion”. en. In: *Nature* 361.6408, pp. 170–173. ISSN: 0028-0836. DOI: 10.1038/361170a0. URL: http://dx.doi.org/10.1038/361170a0.

Deutsch, M and M Long (Aug. 1999). “Intron-exon structures of eukaryotic model organisms”. en. In: *Nucleic Acids Res.* 27.15, pp. 3219–3228. ISSN: 0305-1048.

Dibb, N J (Aug. 1991). “Proto-splice site model of intron origin”. en. In: *J. Theor. Biol.* 151.3, pp. 405–416. ISSN: 0022-5193.

Dibb, N J and A J Newman (July 1989). “Evidence that introns arose at proto-splice sites”. en. In: *EMBO J.* 8.7, pp. 2015–2021. ISSN: 0261-4189.

Dietrich, Rosemary C, John D Fuller, and Richard A Padgett (Sept. 2005a). “A mutational analysis of U12-dependent splice site dinucleotides”. en. In: *RNA* 11.9, pp. 1430–1440. ISSN: 1355-8382.

— (2005b). “A mutational analysis of U12-dependent splice site dinucleotides A mutational analysis of U12-dependent splice site dinucleotides”. In: *Spring*, pp. 1430–1440. DOI: 10.1261/rna.7206305. URL: http://dx.doi.org/10.1261/rna.7206305.

Dietrich, Rosemary C, Robert Incorvaia, and Richard A Padgett (Dec. 1997). “Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns”. English. In: *Mol. Cell* 1.1, pp. 151–160. ISSN: 1097-2765. DOI: 10.1016/S1097-2765(00)80016-7. URL: http://linkinghub.

elsevier.com/retrieve/pii/S1097276500800167%5Cnhttp://ac.els-cdn.com/S1097276500800167/1-s2.0-S1097276500800167-main.pdf?_tid=06fb8d62-1209-11e2-bf6e-00000aacb35d&acdnat=1349784255_a887e0227a9aed594e325340d7d7b7f5.

Doggett, Karen et al. (Dec. 2018). "Early developmental arrest and impaired gastrointestinal homeostasis in U12-dependent splicing-defective Rnpc3-deficient mice". en. In: *RNA* 24.12, pp. 1856–1870. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.068221.118. URL: http://dx.doi.org/10.1261/rna.068221.118.

Doolittle, W F and A Stoltzfus (Feb. 1993). "Molecular evolution. Genes-in-pieces revisited". en. In: *Nature* 361.6411, p. 403. ISSN: 0028-0836. DOI: 10.1038/361403a0. URL: http://dx.doi.org/10.1038/361403a0.

Duchêne, Sebastian and Robert Lanfear (Sept. 2015). "Phylogenetic uncertainty can bias the number of evolutionary transitions estimated from ancestral state reconstruction methods". en. In: *J. Exp. Zool. B Mol. Dev. Evol.* 324.6, pp. 517–524. ISSN: 1552-5007, 1552-5015. DOI: 10.1002/jez.b.22638. URL: http://dx.doi.org/10.1002/jez.b.22638.

Durinck, Steffen, Yves Moreau, et al. (Aug. 2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". en. In: *Bioinformatics* 21.16, pp. 3439–3440. ISSN: 1367-4803.

Durinck, Steffen, Paul T Spellman, et al. (July 2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". en. In: *Nat. Protoc.* 4.8, pp. 1184–1191. ISSN: 1754-2189.

Edery, Patrick et al. (2011). "Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA". In: *Science* 332.6026, pp. 240–243. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1202205. URL: http://dx.doi.org/10.1126/science.1202205.

Farrer, Tracy et al. (Aug. 2002). "Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing". en. In: *Nucleic Acids Res.* 30.15, pp. 3360–3367. ISSN: 0305-1048.

Federhen, Scott (Jan. 2012). “The NCBI Taxonomy database”. en. In: *Nucleic Acids Res.* 40.Database issue, pp. D136–43. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkr1178. URL: http://dx.doi.org/10.1093/nar/gkr1178.

Fedorov, Alexei, Amir Feisal Merican, and Walter Gilbert (Dec. 2002). “Large-scale comparison of intron positions among animal, plant, and fungal genes”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 99.25, pp. 16128–16133. ISSN: 0027-8424.

Fedorov, Alexei, Scott Roy, et al. (Oct. 2003). “Mystery of intron gain”. en. In: *Genome Res.* 13.10, pp. 2236–2241. ISSN: 1088-9051.

Fedorov, Alexei, Jesse Stombaugh, et al. (Aug. 2005). “Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database”. en. In: *Nucleic Acids Res.* 33.14, pp. 4578–4583. ISSN: 0305-1048.

Frey, Katharina and Boas Pucker (Feb. 2020). “Animal, Fungi, and Plant Genome Sequences Harbor Different Non-Canonical Splice Sites”. en. In: *Cells* 9.2. ISSN: 2073-4409. DOI: 10.3390/cells9020458. URL: http://dx.doi.org/10.3390/cells9020458.

Frilander, M J and J A Steitz (Apr. 1999). “Initial recognition of U12-dependent introns requires both U11/5’ splice-site and U12/branchpoint interactions”. en. In: *Genes Dev.* 13.7, pp. 851–863. ISSN: 0890-9369. DOI: 10.1101/gad.13.7.851. URL: http://dx.doi.org/10.1101/gad.13.7.851.

Gault, Christine M et al. (2017). “Aberrant splicing in maize rough endosperm3 reveals a conserved role for U12 splicing in eukaryotic multicellular development”. In: *Proceedings of the National Academy of Sciences* 114.11, E2195–E2204. ISSN: 0027-8424. DOI: 10.1073/pnas.1616173114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1616173114.

Gentekaki, Eleni et al. (Sept. 2017). “Extreme genome diversity in the hyper-prevalent parasitic eukaryote Blastocystis”. en. In: *PLoS Biol.* 15.9, e2003769. ISSN: 1544-9173, 1545-7885. DOI: 10.1371/journal.pbio.2003769. URL: http://dx.doi.org/10.1371/journal.pbio.2003769.

Gilbert, W (1978). *Why genes in pieces?* DOI: 10.1038/271501a0. URL: http://dx.doi.org/10.1038/271501a0.

Gilbert, W (1987). “The exon theory of genes”. en. In: *Cold Spring Harb. Symp. Quant. Biol.* 52, pp. 901–905. ISSN: 0091-7451.

Gilbert, W, S J de Souza, and M Long (July 1997). “Origin of genes”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 94.15, pp. 7698–7703. ISSN: 0027-8424. DOI: 10.1073/pnas.94.15.7698. URL: http://dx.doi.org/10.1073/pnas.94.15.7698.

Glöckner, Gernot, Georg Golderer, et al. (2008). “A first glimpse at the transcriptome of Physarum polycephalum”. In: *BMC Genomics* 9, p. 6. ISSN: 1471-2164. DOI: 10.1186/1471-2164-9-6. URL: http://dx.doi.org/10.1186/1471-2164-9-6.

Glöckner, Gernot and Wolfgang Marwan (Sept. 2017). “Transcriptome reprogramming during developmental switching in Physarum polycephalum involves extensive remodeling of intracellular signaling networks”. en. In: *Sci. Rep.* 7.1, p. 12304. ISSN: 2045-2322. DOI: 10.1038/s41598-017-12250-5. URL: http://dx.doi.org/10.1038/s41598-017-12250-5.

Grabherr, Manfred G et al. (July 2011). “Full-length transcriptome assembly from RNA-Seq data without a reference genome”. In: *Nat. Biotechnol.* 29.7, pp. 644–652. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.1883. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3571712%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract.

Gumińska, Natalia et al. (Oct. 2018). “Order of removal of conventional and non-conventional introns from nuclear transcripts of Euglena gracilis”. en. In: *PLoS Genet.* 14.10, e1007761. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.1007761. URL: http://dx.doi.org/10.1371/journal.pgen.1007761.

Haas, B J (Oct. 2003). “Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies”. In: *Nucleic Acids Res.* 31.19, pp. 5654–5666. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkg770. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg770.

Haas, Brian J et al. (2013). “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis”. In: *Nat. Protoc.* 8.8, pp. 1494–1512. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2013.084. URL: http://dx.doi.org/10.1038/nprot.2013.084.

Hall, S L and R A Padgett (1994). “Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites”. In: *J. Mol. Biol.* 239.3, pp. 357–365. ISSN: 0022-2836. DOI: 10.1006/jmbi.1994.1377. URL: http://www.sciencedirect.com/science/article/pii/S0022283684713775.

— (Mar. 1996). “Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns”. en. In: *Science* 271.5256, pp. 1716–1718. ISSN: 0036-8075. DOI: 10.1126/science.271.5256.1716. URL: http://dx.doi.org/10.1126/science.271.5256.1716.

Henriet, Simon et al. (Oct. 2019). “Evolution of the U2 Spliceosome for Processing Numerous and Highly Diverse Non-canonical Introns in the Chordate Fritillaria borealis”. en. In: *Curr. Biol.* 29.19, 3193–3199.e4. ISSN: 0960-9822, 1879-0445. DOI: 10.1016/j.cub.2019.07.092. URL: http://dx.doi.org/10.1016/j.cub.2019.07.092.

Hillmann, Falk et al. (Feb. 2018). “Multiple Roots of Fruiting Body Formation in Amoebozoa”. en. In: *Genome Biol. Evol.* 10.2, pp. 591–606. ISSN: 1759-6653. DOI: 10.1093/gbe/evy011. URL: http://dx.doi.org/10.1093/gbe/evy011.

Hirose, T, M-D Shu, and J A Steitz (2004). “Splicing of U12-type introns deposits an exon junction complex competent to induce nonsense-mediated mRNA decay”. In: *Proceedings of the National Academy of Sciences* 101.52, pp. 17976–17981. ISSN: 0027-8424. DOI: 10.1073/pnas.0408435102. URL: http://dx.doi.org/10.1073/pnas.0408435102.

Hoff, Katharina J et al. (2015). “BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS”. In: *Bioinformatics*. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv661. URL: http://dx.doi.org/10.1093/bioinformatics/btv661.

Holland, Barbara R et al. (May 2020). “Accuracy of ancestral state reconstruction for non-neutral traits”. en. In: *Sci. Rep.* 10.1, p. 7644. ISSN: 2045-2322. DOI: 10.1038/s41598-020-64647-4. URL: http://dx.doi.org/10.1038/s41598-020-64647-4.

Huff, Jason T, Daniel Zilberman, and Scott W Roy (2016). "Mechanism for DNA transposons to generate introns on genomic scale s". In: ISSN: 0028-0836. DOI: 10.1038/nature20110. URL: http://dx.doi.org/10.1038/nature20110.

Hunter, John D (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3, pp. 90–95. ISSN: 1521-9615, 1558-366X. DOI: 10.1109/mcse.2007.55. URL: http://dx.doi.org/10.1109/mcse.2007.55.

Inoue, Daichi et al. (Apr. 2021). "Minor intron retention drives clonal hematopoietic disorders and diverse cancer predisposition". en. In: *Nat. Genet.* ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-021-00828-9. URL: http://dx.doi.org/10.1038/s41588-021-00828-9.

Irimia, Manuel and Scott William Roy (2014). "Origin of spliceosomal introns and alternative splicing". In: *Cold Spring Harb. Perspect. Biol.* 6.6. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a016071. URL: http://dx.doi.org/10.1101/cshperspect.a016071.

Jackson, I J (July 1991). "A reappraisal of non-consensus mRNA splice sites". In: *Nucleic Acids Res.* 19.14, pp. 3795–3798. ISSN: 0305-1048. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=328465&tool=pmcentrez&rendertype=abstract.

Janice, J et al. (Jan. 2012). "U12-type Spliceosomal Introns of Insecta". In: *Int. J. Biol. Sci.* 8.3, pp. 344–352. ISSN: 1449-2288. DOI: 10.7150/ijbs.3933. URL: http://dx.doi.org/10.7150/ijbs.3933.

Jeffares, Daniel C, Tobias Mourier, and David Penny (2006). "The biology of intron gain and loss". In: *Trends Genet.* 22.1, pp. 16–22. ISSN: 0168-9525. DOI: 10.1016/j.tig.2005.10.006. URL: http://dx.doi.org/10.1016/j.tig.2005.10.006.

Jeffreys, A J and R A Flavell (Dec. 1977). "The rabbit beta-globin gene contains a large large insert in the coding sequence". en. In: *Cell* 12.4, pp. 1097–1108. ISSN: 0092-8674. DOI: 10.1016/0092-8674(77)90172-6. URL: http://dx.doi.org/10.1016/0092-8674(77)90172-6.

Jo, Bong-Seok and Sun Shim Choi (Dec. 2015). "Introns: The Functional Benefits of Introns in Genomes". en. In: *Genomics Inform.* 13.4, pp. 112–118. ISSN:

1598-866X. DOI: 10.5808/GI.2015.13.4.112. URL: http://dx.doi.org/10.5808/GI.2015.13.4.112.

Jurica, Melissa S (June 2008). "Detailed close-ups and the big picture of spliceosomes". en. In: *Curr. Opin. Struct. Biol.* 18.3, pp. 315–320. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2008.05.005. URL: http://dx.doi.org/10.1016/j.sbi.2008.05.005.

Kameoka, Hiromu et al. (Oct. 2019). "Structure-Specific Regulation of Nutrient Transport and Metabolism in Arbuscular Mycorrhizal Fungi". en. In: *Plant Cell Physiol.* 60.10, pp. 2272–2281. ISSN: 0032-0781, 1471-9053. DOI: 10.1093/pcp/pcz122. URL: http://dx.doi.org/10.1093/pcp/pcz122.

Kang, Seungho et al. (2017). "Between a Pod and a Hard Test: The Deep Evolution of Amoebae". In: *Mol. Biol. Evol.* 34.9, pp. 2258–2270. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msx162. URL: http://dx.doi.org/10.1093/molbev/msx162.

Kim, Daehwan et al. (Aug. 2019). "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". en. In: *Nat. Biotechnol.* 37.8, pp. 907–915. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0201-4. URL: http://dx.doi.org/10.1038/s41587-019-0201-4.

Knapp, G et al. (June 1978). "Transcription and processing of intervening sequences in yeast tRNA genes". en. In: *Cell* 14.2, pp. 221–236. ISSN: 0092-8674. DOI: 10.1016/0092-8674(78)90109-5. URL: http://dx.doi.org/10.1016/0092-8674(78)90109-5.

Konarska, M M et al. (1985). "Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors". en. In: *Nature* 313.6003, pp. 552–557. ISSN: 0028-0836. DOI: 10.1038/313552a0. URL: http://dx.doi.org/10.1038/313552a0.

Kondo, Yasushi et al. (Jan. 2015). "Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition". en. In: *Elife* 4. ISSN: 2050-084X. DOI: 10.7554/eLife.04986. URL: http://dx.doi.org/10.7554/eLife.04986.

König, Harald et al. (Nov. 2007). "Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation". en. In: *Cell* 131.4, pp. 718–729. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.09.043. URL: http://dx.doi.org/10.1016/j.cell.2007.09.043.

Koonin, Eugene V (Aug. 2006). "The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?" en. In: *Biol. Direct* 1, p. 22. ISSN: 1745-6150. DOI: 10.1186/1745-6150-1-22. URL: http://dx.doi.org/10.1186/1745-6150-1-22.

Koonin, Eugene V, Miklos Csuros, and Igor B Rogozin (Jan. 2013). "Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes". en. In: *Wiley Interdiscip. Rev. RNA* 4.1, pp. 93–105. ISSN: 1757-7004, 1757-7012. DOI: 10.1002/wrna.1143. URL: http://dx.doi.org/10.1002/wrna.1143.

Langmead, Ben (2010). "Aligning short sequencing reads with Bowtie". In: *Curr. Protoc. Bioinformatics* 32.1, pp. 11–17. ISSN: 1934-3396. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1107s32.

Larkin, M A et al. (Nov. 2007). "Clustal W and Clustal X version 2.0". en. In: *Bioinformatics* 23.21, pp. 2947–2948. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btm404. URL: http://dx.doi.org/10.1093/bioinformatics/btm404.

Larue, Graham E, Marek Eliáš, and Scott W Roy (July 2021). "Expansion and transformation of the minor spliceosomal system in the slime mold Physarum polycephalum". en. In: *Curr. Biol.* 31.14, 3125–3131.e4. ISSN: 0960-9822, 1879-0445. DOI: 10.1016/j.cub.2021.04.050. URL: http://dx.doi.org/10.1016/j.cub.2021.04.050.

Letunic, Ivica and Peer Bork (July 2016). "Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees". en. In: *Nucleic Acids Res.* 44.W1, W242–5. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw290. URL: http://dx.doi.org/10.1093/nar/gkw290.

— (July 2021). "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation". en. In: *Nucleic Acids Res.* 49.W1, W293–W296.

ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkab301. URL: http://dx.doi.org/10.1093/nar/gkab301.

Levesque, Lauren, Nicole Salazar, and Scott William Roy (2022). *Distinct Minor Splicing Patterns across Cancers*. DOI: 10.3390/genes13020387. URL: http://dx.doi.org/10.3390/genes13020387.

Levine, Aaron and Richard Durbin (Oct. 2001). "A computational scan for U12-dependent introns in the human genome sequence". en. In: *Nucleic Acids Res.* 29.19, pp. 4006–4013. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/29.19.4006. URL: http://nar.oupjournals.org/cgi/content/abstract/29/19/4006%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC60238/pdf/gke532.pdf.

Lewis, Benjamin P, Richard E Green, and Steven E Brenner (2003). "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans". In: *Proc. Natl. Acad. Sci. U. S. A.* 100.1, pp. 189–192. ISSN: 0027-8424. DOI: 10.1073/pnas.0136770100. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=140922%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract.

Li, Qin, Guanghui Xiao, and Yu Xian Zhu (2014). "Single-nucleotide resolution mapping of the gossypium raimondii transcriptome reveals a new mechanism for alternative splicing of introns". In: *Mol. Plant* 7.5. ISSN: 1674-2052, 1752-9867. DOI: 10.1093/mp/sst175. URL: http://dx.doi.org/10.1093/mp/sst175.

Lin, Chiao-Feng et al. (Feb. 2010). "Evolutionary dynamics of U12-type spliceosomal introns". en. In: *BMC Evol. Biol.* 10, p. 47. ISSN: 1471-2148. DOI: 10.1186/1471-2148-10-47. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2831892&tool=pmcentrez&rendertype=abstract.

Lin, Kui and Da-Yong Zhang (Nov. 2005). "The excess of 5' introns in eukaryotic genomes". en. In: *Nucleic Acids Res.* 33.20, pp. 6522–6527. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gki970. URL: http://dx.doi.org/10.1093/nar/gki970.

Logsdon Jr, J M (Dec. 1998). "The recent origins of spliceosomal introns revisited". en. In: *Curr. Opin. Genet. Dev.* 8.6, pp. 637–648. ISSN: 0959-437X. DOI: 10.

1016/s0959-437x(98)80031-2. URL: http://dx.doi.org/10.1016/s0959-437x(98)80031-2.

Long, M and C Rosenberg (Dec. 2000). "Testing the "proto-splice sites" model of intron origin: evidence from analysis of intron phase correlations". en. In: *Mol. Biol. Evol.* 17.12, pp. 1789–1796. ISSN: 0737-4038.

Long, M, C Rosenberg, and W Gilbert (Dec. 1995). "Intron phase correlations and the evolution of the intron/exon structure of genes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 92.26, pp. 12495–12499. ISSN: 0027-8424.

Long, M, S J de Souza, and W Gilbert (Dec. 1995). "Evolution of the intron-exon structure of eukaryotic genes". en. In: *Curr. Opin. Genet. Dev.* 5.6, pp. 774–778. ISSN: 0959-437X.

Lopez, Marcela Davila, Magnus Alm Rosenblad, and Tore Samuelsson (May 2008). "Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components". In: *Nucleic Acids Res.* 36.9, pp. 3001–3010. ISSN: 0305-1048. DOI: 10.1093/nar/gkn142. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkn142.

Lopez, P J and B Séraphin (Jan. 2000). "YIDB: the Yeast Intron DataBase". en. In: *Nucleic Acids Res.* 28.1, pp. 85–86. ISSN: 0305-1048.

Lotti, Francesco et al. (2012). "An SMN-dependent U12 splicing event essential for motor circuit function". In: *Cell* 151.2, pp. 440–454. ISSN: 0092-8674. DOI: 10.1016/j.cell.2012.09.012. URL: http://dx.doi.org/10.1016/j.cell.2012.09.012.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12, p. 550. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

Lu, Jianli et al. (June 2008). "Gene expression enhancement mediated by the 5 UTR intron of the rice rubi3 gene varied remarkably among tissues in transgenic rice plants". In: *Mol. Genet. Genomics* 279.6, pp. 563–572. ISSN: 1617-4615, 1617-4623. DOI: 10.1007/s00438-008-0333-6. URL: https://doi.org/10.1007/s00438-008-0333-6.

Lynch, Michael and John S Conery (Nov. 2003). “The origins of genome complexity”. en. In: *Science* 302.5649, pp. 1401–1404. ISSN: 0036-8075. DOI: 10.1126/science.1089370. URL: http://dx.doi.org/10.1126/science.1089370.

Lynch, Michael and Aaron O Richardson (Dec. 2002). “The evolution of spliceosomal introns”. en. In: *Curr. Opin. Genet. Dev.* 12.6, pp. 701–710. ISSN: 0959-437X.

Ma, Ming-Yue et al. (June 2022). “Intron losses and gains in the nematodes”. en. In: *Biol. Direct* 17.1, p. 13. ISSN: 1745-6150. DOI: 10.1186/s13062-022-00328-8. URL: http://dx.doi.org/10.1186/s13062-022-00328-8.

Madan, Vikas et al. (Jan. 2015). “Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome”. en. In: *Nat. Commun.* 6.May 2014, p. 6042. ISSN: 2041-1723. DOI: 10.1038/ncomms7042. URL: http://dx.doi.org/10.1038/ncomms7042.

Mao, Rui et al. (2014). “Comparative analyses between retained introns and constitutively spliced introns in Arabidopsis thaliana using random forest and support vector machine”. In: *PLoS One* 9.8, pp. 1–12. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0104049. URL: http://dx.doi.org/10.1371/journal.pone.0104049.

Matera, A Gregory and Zefeng Wang (Feb. 2014). “A day in the life of the spliceosome”. en. In: *Nat. Rev. Mol. Cell Biol.* 15.2, pp. 108–121. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm3742. URL: http://dx.doi.org/10.1038/nrm3742.

Meinke, Stefan et al. (Apr. 2020). “Srsf10 and the minor spliceosome control tissue-specific and dynamic SR protein expression”. en. In: *Elife* 9. ISSN: 2050-084X. DOI: 10.7554/eLife.56075. URL: http://dx.doi.org/10.7554/eLife.56075.

Middleton, Robert et al. (Mar. 2017). “IRFinder: assessing the impact of intron retention on mammalian gene expression”. en. In: *Genome Biol.* 18.1, p. 51. ISSN: 1465-6906. DOI: 10.1186/s13059-017-1184-4. URL: http://dx.doi.org/10.1186/s13059-017-1184-4.

Milanowski, Rafał et al. (Feb. 2016). “Intermediate introns in nuclear genes of euglenids - are they a distinct type?” en. In: *BMC Evol. Biol.* 16, p. 49. ISSN: 1471-2148. DOI: 10.1186/s12862-016-0620-5. URL: http://dx.doi.org/10.1186/s12862-016-0620-5.

Monteuuis, Geoffray et al. (Dec. 2019). “The changing paradigm of intron retention: regulation, ramifications and recipes”. en. In: *Nucleic Acids Res.* 47.22, pp. 11497–11513. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz1068. URL: http://dx.doi.org/10.1093/nar/gkz1068.

Montzka, K A and J A Steitz (Dec. 1988). “Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 85.23, pp. 8885–8889. ISSN: 0027-8424. DOI: 10.1073/pnas.85.23.8885. URL: http://dx.doi.org/10.1073/pnas.85.23.8885.

Mount, S M (Jan. 1982). “A catalogue of splice junction sequences”. en. In: *Nucleic Acids Res.* 10.2, pp. 459–472. ISSN: 0305-1048.

Mourier, Tobias and Daniel C Jeffares (May 2003). “Eukaryotic intron loss”. en. In: *Science* 300.5624, pp. 1393–1393. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1080559. URL: http://dx.doi.org/10.1126/science.1080559.

Moyer, Devlin C et al. (July 2020). “Comprehensive database and evolutionary dynamics of U12-type introns”. en. In: *Nucleic Acids Res.* 48.13, pp. 7066–7078. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa464. URL: http://dx.doi.org/10.1093/nar/gkaa464.

Najle, Sebastián R and Iñaki Ruiz-Trillo (Apr. 2021). “The protistan origins of animal cell differentiation”. In: *Origin and Evolution of Metazoan Cell Types*. CRC Press, pp. 13–26. ISBN: 9781315388229. DOI: 10.1201/b21831-2. URL: https://www.taylorfrancis.com/books/9781315388212/chapters/10.1201/b21831-2.

Nawrocki, Eric P and Sean R Eddy (Nov. 2013). “Infernal 1.1: 100-fold faster RNA homology searches”. en. In: *Bioinformatics* 29.22, pp. 2933–2935. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btt509. URL: http://dx.doi.org/10.1093/bioinformatics/btt509.

Nguyen, Hung D, Maki Yoshihama, and Naoya Kenmochi (Sept. 2006). "Phase distribution of spliceosomal introns: implications for intron origin". en. In: *BMC Evol. Biol.* 6, p. 69.

Niemelä, Elina H and Mikko J Frilander (2014). "Regulation of gene expression through inefficient splicing of U12-type introns". en. In: *RNA Biol.* 11.11, pp. 1325–1329. ISSN: 1547-6286, 1555-8584. DOI: 10.1080/15476286.2014.996454. URL: http://www.tandfonline.com/doi/abs/10.1080/15476286.2014.996454.

Niemelä, Elina H, Ali Oghabian, et al. (2014). "Global analysis of the nuclear processing of transcripts with unspliced U12-type introns by the exosome". In: *Nucleic Acids Res.* 42.11, pp. 7358–7369. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gku391. URL: http://dx.doi.org/10.1093/nar/gku391.

Nojima, Takayuki et al. (Oct. 2018). "RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing". en. In: *Mol. Cell* 72.2, 369–379.e4. ISSN: 1097-2765, 1097-4164. DOI: 10.1016/j.molcel.2018.09.004. URL: http://dx.doi.org/10.1016/j.molcel.2018.09.004.

O'Leary, Nuala A et al. (Jan. 2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". en. In: *Nucleic Acids Res.* 44.D1, pp. D733–45. ISSN: 0305-1048.

Olthof, Anouk M, Katery C Hyatt, and Rahul N Kanadia (Aug. 2019). "Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns". en. In: *BMC Genomics* 20.1, pp. 1–19. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6046-x. URL: http://dx.doi.org/10.1186/s12864-019-6046-x.

Padgett, R A et al. (1986). "Splicing of messenger RNA precursors". en. In: *Annu. Rev. Biochem.* 55, pp. 1119–1150. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.55.070186.005351. URL: http://dx.doi.org/10.1146/annurev.bi.55.070186.005351.

Padgett, Richard A (2012). “New connections between splicing and human disease”. In: *Trends Genet.* 28.4, pp. 147–154. ISSN: 0168-9525. DOI: 10.1016/j.tig.2012.01.001. URL: http://dx.doi.org/10.1016/j.tig.2012.01.001.

Parada, G E et al. (2014). “A comprehensive survey of non-canonical splice sites in the human transcriptome”. In: *Nucleic Acids Res.* 42.16, pp. 10564–10578. ISSN: 0305-1048. DOI: 10.1093/nar/gku744. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku744.

Parenteau, Julie, Mathieu Durand, et al. (May 2008). “Deletion of many yeast introns reveals a minority of genes that require splicing for function”. en. In: *Mol. Biol. Cell* 19.5, pp. 1932–1941. ISSN: 1059-1524, 1939-4586. DOI: 10.1091/mbc.e07-12-1254. URL: http://dx.doi.org/10.1091/mbc.e07-12-1254.

Parenteau, Julie, Laurine Maignon, et al. (2019). “Introns are mediators of cell response to starvation”. In: *Nature* 565.7741. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0859-7. URL: http://dx.doi.org/10.1038/s41586-018-0859-7.

Patel, Abhijit A, Matthew McCarthy, and Joan A Steitz (2002). “The splicing of U12-type introns can be a rate-limiting step in gene expression”. In: *EMBO J.* 21.14, pp. 3804–3815. ISSN: 0261-4189. DOI: 10.1093/emboj/cdf297. URL: http://dx.doi.org/10.1093/emboj/cdf297.

Patel, Abhijit A and Joan A Steitz (Dec. 2003). “Splicing double: insights from the second spliceosome”. en. In: *Nat. Rev. Mol. Cell Biol.* 4.December, pp. 960–970. ISSN: 1471-0072. DOI: 10.1038/nrm1259. URL: http://dx.doi.org/10.1038/nrm1259.

Patro, Rob et al. (Apr. 2017). “Salmon provides fast and bias-aware quantification of transcript expression”. en. In: *Nat. Methods* 14.4, pp. 417–419. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4197. URL: http://dx.doi.org/10.1038/nmeth.4197.

Pedregosa, Fabian (2011). “Scikit-learn: Machine Learning in Python”. In: *J. Mach. Learn. Res.* 12, pp. 2825–2830. ISSN: 1532-4435.

Pertea, Geo (n.d.). *gffcompare*. URL: https://github.com/gpertea/gffcompare.

Pertea, Mihaela et al. (2016). “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown”. In: *Nat. Protoc.* 11.9, pp. 1650–1667. ISSN: 1754-2189. DOI: 10.1038/nprot.2016.095. URL: http://dx.doi.org/10.1038/nprot.2016.095%5Cnhttp://10.1038/nprot.2016.095%5Cnhttp://www.nature.com/nprot/journal/v11/n9/abs/nprot.2016.095.html#supplementary-information.

Pessa, Heli, Annukka Ruokolainen, and Mikko J Frilander (2006). “The abundance of the spliceosomal snRNPs is not limiting the splicing of U12-type introns”. In: *RNA* 12, pp. 1883–1892. ISSN: 1355-8382. DOI: PMC1581978. URL: http://www.rnajournal.org/cgi/content/abstract/12/10/1883.

Pineda, Jose Mario Bello and Robert K Bradley (Apr. 2018). “Most human introns are recognized via multiple and tissue-specific branchpoints”. en. In: *Genes and Development* 32.7-8, pp. 577–591. ISSN: 0890-9369. DOI: 10.1101/gad.312058.118. URL: http://dx.doi.org/10.1101/gad.312058.118.

Pomeranz Krummel, Daniel A et al. (Mar. 2009). “Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution”. en. In: *Nature* 458.7237, pp. 475–480. ISSN: 0028-0836.

Poverennaya, I V and M A Roytberg (July 2020). “Spliceosomal Introns: Features, Functions, and Evolution”. In: *Biochemistry* 85.7, pp. 725–734. ISSN: 1608-3040. DOI: 10.1134/S0006297920070019. URL: https://doi.org/10.1134/S0006297920070019.

Poverennaya, Irina V, Nadezhda A Potapova, and Sergey A Spirin (Dec. 2020). “Is there any intron sliding in mammals?” en. In: *BMC Evol. Biol.* 20.1, p. 164. ISSN: 1471-2148. DOI: 10.1186/s12862-020-01726-0. URL: http://dx.doi.org/10.1186/s12862-020-01726-0.

Pucker, Boas and Samuel F Brockington (Dec. 2018). “Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes”. en. In: *BMC Genomics* 19.1, p. 980. ISSN: 1471-2164. DOI: 10.1186/s12864-018-5360-z. URL: http://dx.doi.org/10.1186/s12864-018-5360-z.

Revell, Liam J (Apr. 2012). “phytools: an R package for phylogenetic comparative biology (and other things)”. en. In: *Methods Ecol. Evol.* 3.2, pp. 217–223. ISSN:

2041-210X. DOI: 10.1111/j.2041-210x.2011.00169.x. URL: http://dx.doi.org/10.1111/j.2041-210x.2011.00169.x.

Rogozin, Igor B et al. (Apr. 2012). "Origin and evolution of spliceosomal introns". en. In: *Biol. Direct* 7.1, p. 11. ISSN: 1745-6150. DOI: 10.1186/1745-6150-7-11. URL: http://www.biology-direct.com/content/7/1/11.

Rose, Alan B (2018). "Introns as Gene Regulators: A Brick on the Accelerator". en. In: *Front. Genet.* 9, p. 672. ISSN: 1664-8021. DOI: 10.3389/fgene.2018.00672. URL: http://dx.doi.org/10.3389/fgene.2018.00672.

Roy, S W Scott W (Jan. 2009). "Intronization, de-intronization and intron sliding are rare in Cryptococcus". In: *BMC Evol. Biol.* 9, p. 192. ISSN: 1471-2148. DOI: 10.1186/1471-2148-9-192. URL: http://www.biomedcentral.com/1471-2148/9/192/.

Roy, Scott W (Jan. 2004). "The origin of recent introns: transposons?" In: *Genome Biol.* 5.12, p. 251. ISSN: 1465-6906, 1465-6914. DOI: 10.1186/gb-2004-5-12-251. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545789&tool=pmcentrez&rendertype=abstract.

Roy, Scott W, Alexei Fedorov, and Walter Gilbert (June 2003). "Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.12, pp. 7158–7162. ISSN: 0027-8424. DOI: 10.1073/pnas.1232297100. URL: http://dx.doi.org/10.1073/pnas.1232297100.

Roy, Scott W and Walter Gilbert (Feb. 2005a). "Complex early genes". In: *Proc. Natl. Acad. Sci. U. S. A.* 102.6, pp. 1986–1991. ISSN: 0027-8424. DOI: 10.1073/pnas.0408355101. URL: http://www.ncbi.nlm.nih.gov/pubmed/24057835.

— (2005b). "The pattern of intron loss". In: *Proc. Natl. Acad. Sci. U. S. A.* 102.3, pp. 713–718. ISSN: 0027-8424. DOI: 10.1073/pnas.0408274102. URL: http://dx.doi.org/10.1073/pnas.0408274102.

Roy, Scott William (Aug. 2016). "How Common Is Parallel Intron Gain? Rapid Evolution Versus Independent Creation in Recently Created Introns in Daphnia". en. In: *Mol. Biol. Evol.* 33.8, pp. 1902–1906. ISSN: 0737-4038, 1537-1719.

DOI: 10.1093/molbev/msw091. URL: http://dx.doi.org/10.1093/molbev/msw091.

Roy, Scott William and Walter Gilbert (Mar. 2006). "The evolution of spliceosomal introns: patterns, puzzles and progress". In: *Nat. Rev. Genet.* 7.3, pp. 211–221. ISSN: 1471-0056. DOI: 10.1038/nrg1807. URL: http://dx.doi.org/10.1038/nrg1807.

Roy, Scott William, Landen Gozashti, et al. (Oct. 2020). "Massive Intron Gain in the Most Intron-Rich Eukaryotes is Driven by Introner-Like Transposable Elements of Unprecedented Diversity and Flexibility". DOI: 10.2139/ssrn.3721721. URL: https://papers.ssrn.com/abstract=3721721.

Roy, Scott William and Manuel Irimia (2008). "When good transcripts go bad: Artifactual RT-PCR 'splicing' and genome analysis". In: *Bioessays* 30.6, pp. 601–605. ISSN: 0265-9247. DOI: 10.1002/bies.20749. URL: http://dx.doi.org/10.1002/bies.20749.

Roy, Scott William and David Penny (Dec. 2006). "Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses". In: *Mol. Biol. Evol.* 23.12, pp. 2259–2262. ISSN: 0737-4038. DOI: 10.1093/molbev/msl098. URL: http://dx.doi.org/10.1093/molbev/msl098.

— (July 2007a). "A very high fraction of unique intron positions in the intron-rich diatom Thalassiosira pseudonana indicates widespread intron gain". en. In: *Mol. Biol. Evol.* 24.7, pp. 1447–1457. ISSN: 0737-4038. DOI: 10.1093/molbev/msm048. URL: http://dx.doi.org/10.1093/molbev/msm048.

— (2007b). "Patterns of intron loss and gain in plants: Intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana". In: *Mol. Biol. Evol.* 24.1, pp. 171–181. ISSN: 0737-4038. DOI: 10.1093/molbev/msl159. URL: http://dx.doi.org/10.1093/molbev/msl159.

Russell, Anthony G et al. (Oct. 2006). "An early evolutionary origin for the minor spliceosome". en. In: *Nature* 443.7113, pp. 863–866. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature05228. URL: http://dx.doi.org/10.1038/nature05228.

Sakharkar, M et al. (Jan. 2000). “ExInt: an Exon/Intron database”. en. In: *Nucleic Acids Res.* 28.1, pp. 191–192. ISSN: 0305-1048.

Sakharkar, M K, P Kangueane, et al. (Dec. 2000). “IE-Kb: intron exon knowledge base”. en. In: *Bioinformatics* 16.12, pp. 1151–1152. ISSN: 1367-4803.

Sakharkar, M K, T W Tan, and S J de Souza (Aug. 2001). “Generation of a database containing discordant intron positions in eukaryotic genes (MIDB)”. en. In: *Bioinformatics* 17.8, pp. 671–675. ISSN: 1367-4803.

Sakurai, A et al. (Oct. 2002). “On biased distribution of introns in various eukaryotes”. en. In: *Gene* 300.1-2, pp. 89–95. ISSN: 0378-1119. DOI: 10.1016/s0378-1119(02)01035-1. URL: http://dx.doi.org/10.1016/s0378-1119(02)01035-1.

Sales-Lee, Jade et al. (Nov. 2021). “Coupling of spliceosome complexity to intron diversity”. en. In: *Curr. Biol.* 31.22, 4898–4910.e4. ISSN: 0960-9822, 1879-0445. DOI: 10.1016/j.cub.2021.09.004. URL: http://dx.doi.org/10.1016/j.cub.2021.09.004.

Sandberg, Rickard et al. (June 2008). “Proliferating Cells Express mRNAs with Shortened 3’ Untranslated Regions and Fewer MicroRNA Target Sites”. en. In: *Science* 320.5883, pp. 1643–1647. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1155390. URL: https://science.sciencemag.org/content/320/5883/1643.

Sands, Bryan, Soo Yun, and Alexander R Mendenhall (Nov. 2021). “Introns control stochastic allele expression bias”. en. In: *Nat. Commun.* 12.1, p. 6527. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26798-4. URL: http://dx.doi.org/10.1038/s41467-021-26798-4.

Saxonov, S et al. (Jan. 2000). “EID: the Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes”. en. In: *Nucleic Acids Res.* 28.1, pp. 185–190. ISSN: 0305-1048.

Schaap, Pauline et al. (2015). “The Physarum polycephalum Genome Reveals Extensive Use of Prokaryotic Two-Component and Metazoan-Type Tyrosine Kinase Signaling”. In: *Genome Biol. Evol.* 8.1, pp. 109–125. ISSN: 1759-6653. DOI: 10.1093/gbe/evv237. URL: http://www.ncbi.nlm.nih.gov/pubmed/

26615215%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4758236.

Seabold, Skipper and Josef Perktold (2010). "Statsmodels: Econometric and statistical modeling with python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 57, p. 61. URL: http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf.

Sêton Bocco, Steven and Miklós Csűrös (Aug. 2016). "Splice Sites Seldom Slide: Intron Evolution in Oomycetes". en. In: *Genome Biol. Evol.* 8.8, pp. 2340–2350. ISSN: 1759-6653. DOI: 10.1093/gbe/evw157. URL: http://dx.doi.org/10.1093/gbe/evw157.

Shadwick, John D L, Jeffery D Silberman, and Frederick W Spiegel (May 2018). "Variation in the SSUrDNA of the Genus Protostelium Leads to a New Phylogenetic Understanding of the Genus and of the Species Concept for Protostelium mycophaga (Protosteliida, Amoebozoa)". en. In: *J. Eukaryot. Microbiol.* 65.3, pp. 331–344. ISSN: 1066-5234, 1550-7408. DOI: 10.1111/jeu.12476. URL: http://dx.doi.org/10.1111/jeu.12476.

Sharangdhar, Tejaswini et al. (Oct. 2017). "A retained intron in the 3'-UTR of Calm3 mRNA mediates its Staufen2- and activity-dependent localization to neuronal dendrites". en. In: *EMBO Rep.* 18.10, pp. 1762–1774. ISSN: 1469-221X, 1469-3178. DOI: 10.15252/embr.201744334. URL: http://dx.doi.org/10.15252/embr.201744334.

Sharpton, Thomas J et al. (Jan. 2008). "Mechanisms of intron gain and loss in Cryptococcus". en. In: *Genome Biol.* 9.1, R24. ISSN: 1465-6906. DOI: 10.1186/gb-2008-9-1-r24. URL: http://dx.doi.org/10.1186/gb-2008-9-1-r24.

Shaul, Orit (Oct. 2017). "How introns enhance gene expression". en. In: *Int. J. Biochem. Cell Biol.* 91.Pt B, pp. 145–155. ISSN: 1357-2725, 1878-5875. DOI: 10.1016/j.biocel.2017.06.016. URL: http://dx.doi.org/10.1016/j.biocel.2017.06.016.

Shepelev, Valery and Alexei Fedorov (June 2006). "Advances in the Exon-Intron Database (EID)". en. In: *Brief. Bioinform.* 7.2, pp. 178–185. ISSN: 1467-5463.

Sheth, Nihar et al. (Aug. 2006). "Comprehensive splice-site analysis using comparative genomics". en. In: *Nucleic Acids Res.* 34.14, pp. 3955–3967. ISSN: 0305-1048. DOI: 10.1093/nar/gkl556. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1557818%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract.

Sibley, Christopher R, Lorea Blazquez, and Jernej Ule (2016). *Lessons from non-canonical splicing*. DOI: 10.1038/nrg.2016.46. URL: http://dx.doi.org/10.1038/nrg.2016.46.

Sievers, Fabian and Desmond G Higgins (2014). "Clustal Omega". In: *Curr. Protoc. Bioinformatics* 2014, pp. 3.13.1–3.13.16. ISSN: 1934-3396, 1934-340X. DOI: 10.1002/0471250953.bi0313s48. URL: http://dx.doi.org/10.1002/0471250953.bi0313s48.

Simão, Felipe A et al. (2015). "BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics*. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv351. URL: http://dx.doi.org/10.1093/bioinformatics/btv351.

Singh, Jarnail and Richard A Padgett (Nov. 2009). "Rates of in situ transcription and splicing in large human genes". en. In: *Nat. Struct. Mol. Biol.* 16.11, pp. 1128–1133. ISSN: 1545-9993, 1545-9985. DOI: 10.1038/nsmb.1666. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2783620&tool=pmcentrez&rendertype=abstract.

Slabodnick, Mark M et al. (Feb. 2017). "The Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell". en. In: *Curr. Biol.* 27.4, pp. 569–575. ISSN: 0960-9822, 1879-0445. DOI: 10.1016/j.cub.2016.12.057. URL: http://dx.doi.org/10.1016/j.cub.2016.12.057.

Soneson, Charlotte, Michael I Love, and Mark D Robinson (Dec. 2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences". en. In: *F1000Res.* 4, p. 1521. ISSN: 2046-1402. DOI: 10.12688/f1000research.7563.2. URL: http://dx.doi.org/10.12688/f1000research.7563.2.

Stanke, Mario et al. (Mar. 2008). “Using native and syntenically mapped cDNA alignments to improve de novo gene finding”. en. In: *Bioinformatics* 24.5, pp. 637–644. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btn013. URL: http://dx.doi.org/10.1093/bioinformatics/btn013.

Stark, Alexander et al. (2005). “Animal microRNAs confer robustness to gene expression and have a significant impact on 3???UTR evolution”. In: *Cell* 123.6, pp. 1133–1146. ISSN: 0092-8674. DOI: 10.1016/j.cell.2005.11.023. URL: http://dx.doi.org/10.1016/j.cell.2005.11.023.

Stoltzfus, A (June 1994). “Origin of introns–early or late”. en. In: *Nature* 369.6481, 526–7, author reply 527–8. ISSN: 0028-0836. DOI: 10.1038/369526b0. URL: http://dx.doi.org/10.1038/369526b0.

Stoltzfus, A et al. (July 1994). “Testing the exon theory of genes: the evidence from protein structure”. en. In: *Science* 265.5169, pp. 202–207. ISSN: 0036-8075. DOI: 10.1126/science.8023140. URL: http://dx.doi.org/10.1126/science.8023140.

Stoltzfus, Arlin et al. (1997). “Intron “sliding” and the diversity of intron positions”. In: *Proceedings of the National Academy of Sciences* 94.20, pp. 10739–10744. DOI: 10.1073/pnas.94.20.10739. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.94.20.10739. URL: https://www.pnas.org/doi/abs/10.1073/pnas.94.20.10739.

Sverdlov, Alexander V et al. (Aug. 2004). “Reconstruction of ancestral protosplice sites”. en. In: *Curr. Biol.* 14.16, pp. 1505–1508. ISSN: 0960-9822. DOI: 10.1016/j.cub.2004.08.027. URL: http://dx.doi.org/10.1016/j.cub.2004.08.027.

— (2005). “Conservation versus parallel gains in intron evolution”. In: *Nucleic Acids Res.* 33.6, pp. 1741–1748. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gki316. URL: http://dx.doi.org/10.1093/nar/gki316.

Szafranski, Karol et al. (2007). “Violating the splicing rules: TG dinucleotides function as alternative 3’ splice sites in U2-dependent introns”. en. In: *Genome Biol.* 8.8, R154. ISSN: 1465-6906.

Szcześniak, Michał Wojciech et al. (Feb. 2013). "ERISdb: a database of plant splice sites and splicing signals". en. In: *Plant Cell Physiol.* 54.2, e10. ISSN: 0032-0781. DOI: 10.1093/pcp/pct001. URL: http://dx.doi.org/10.1093/pcp/pct001.

Tarn, W Y and J A Steitz (Sept. 1996). "Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns". en. In: *Science* 273.5283, pp. 1824–1832. ISSN: 0036-8075. DOI: 10.1126/science.273.5283.1824. URL: http://dx.doi.org/10.1126/science.273.5283.1824.

Tarn, Woan Y and Joan a Steitz (1996). "A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro". In: *Cell* 84, pp. 801–811. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(00)81057-0. URL: http://dx.doi.org/10.1016/S0092-8674(00)81057-0.

Thanaraj, T A and F Clark (June 2001). "Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions". en. In: *Nucleic Acids Res.* 29.12, pp. 2581–2593. ISSN: 0305-1048.

Turunen, Janne J, Elina H Niemelä, et al. (Jan. 2013). "The significant other: splicing by the minor spliceosome". en. In: *Wiley Interdiscip. Rev. RNA* 4.1, pp. 61–76. ISSN: 1757-7004. DOI: 10.1002/wrna.1141. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3584512%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract.

Turunen, Janne J, Cindy L Will, et al. (May 2008). "The U11-48K protein contacts the 5' splice site of U12-type introns and the U11-59K protein". en. In: *Mol. Cell. Biol.* 28.10, pp. 3548–3560. ISSN: 0270-7306.

UniProt Consortium (Jan. 2008). "The universal protein resource (UniProt)". en. In: *Nucleic Acids Res.* 36.Database issue, pp. D190–5. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkm895. URL: http://dx.doi.org/10.1093/nar/gkm895.

Vinogradov, Alexander E (Sept. 1999). "Intron–Genome Size Relationship on a Large Evolutionary Scale". In: *J. Mol. Evol.* 49.3, pp. 376–384. ISSN: 0022-2844.

Virtanen, Pauli et al. (Mar. 2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. In: *Nat. Methods* 17.3, pp. 261–272. ISSN: 1548-7091,

1548-7105. DOI: 10.1038/s41592-019-0686-2. URL: http://dx.doi.org/10.1038/s41592-019-0686-2.

Weischenfeldt, Joachim et al. (Jan. 2012). "Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns". In: *Genome Biol.* 13.5, R35. ISSN: 1465-6906, 1465-6914. DOI: 10.1186/gb-2012-13-5-r35. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3446288&tool=pmcentrez&rendertype=abstract.

Will, Cindy L, Reinhard Lührmann, and R Luhrmann (July 2011). "Spliceosome structure and function". In: *Cold Spring Harb. Perspect. Biol.* 3.7, pp. 1–2. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a003707. URL: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a003707.

Wong, Justin J-L and Ulf Schmitz (Aug. 2022). "Intron retention: importance, challenges, and opportunities". en. In: *Trends Genet.* 38.8, pp. 789–792. ISSN: 0168-9525. DOI: 10.1016/j.tig.2022.03.017. URL: http://dx.doi.org/10.1016/j.tig.2022.03.017.

Yang, Ziheng (2007). "PAML 4: Phylogenetic analysis by maximum likelihood". In: *Mol. Biol. Evol.* 24.8, pp. 1586–1591. ISSN: 0737-4038. DOI: 10.1093/molbev/msm088. URL: http://dx.doi.org/10.1093/molbev/msm088.

Yenerall, Paul and Leming Zhou (Sept. 2012). "Identifying the mechanisms of intron gain: progress and trends". en. In: *Biol. Direct* 7, p. 29. ISSN: 1745-6150. DOI: 10.1186/1745-6150-7-29. URL: http://dx.doi.org/10.1186/1745-6150-7-29.

Younis, Ihab et al. (Jan. 2013). "Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA". In: *Elife* 2, e00780. ISSN: 2050-084X. DOI: 10.7554/eLife.00780. URL: http://elifesciences.org/lookup/doi/10.7554/eLife.00780.

Zerbino, Daniel R et al. (Jan. 2018). "Ensembl 2018". en. In: *Nucleic Acids Res.* 46.D1, pp. D754–D761. ISSN: 0305-1048.

Zhang, Xiaofeng et al. (May 2017). "An Atomic Structure of the Human Spliceosome". en. In: *Cell* 169.5, 918–929.e14. ISSN: 0092-8674, 1097-4172. DOI: 10.

1016/j.cell.2017.04.033. URL: http://dx.doi.org/10.1016/j.cell.2017.04.033.

Zhu, Wei and Volker Brendel (2003). "Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome". In: *Nucleic Acids Res.* 31.15, pp. 4561–4572. ISSN: 0305-1048. DOI: 10.1093/nar/gkg492. URL: http://dx.doi.org/10.1093/nar/gkg492.