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

Graham E. Larue[1], Marek Eliáš[2], Scott W. Roy[1,3,4,*]

[1]Department of Molecular and Cell Biology, University of California, Merced, Merced, CA 95343, USA.

[2]Department of Biology and Ecology Faculty of Science, University of Ostrava, Ostrava, Czech Republic.

[3]Department of Biology, San Francisco State University, San Francisco, CA 94132, USA.

[4]Lead contact

*Correspondence to: scottwroy@gmail.com (S.W.R.)

Summary

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 [1–3]. In all characterized species, U12-type introns are distinguished by several features, including being rare in the genome (~0.5% of all introns) [4–6], containing extended evolutionary-conserved splicing sites [4,5,7,8], being generally ancient [9,10] and being inefficiently spliced [11–13]. 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.

Keywords

U12, minor spliceosome, minor introns, intron evolution, genomics, intron gain, U12-type introns, evolution, bioinformatics, comparative genomics

Results

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 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 [14,15]. However, genome-wide analysis of the current *P. polycephalum* genome annotation [16,17] revealed that all annotated introns have GY-AG boundaries, a pattern suggesting non-GY-AG introns may have been discarded by the annotation pipeline [16,18]. 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 (STAR Methods) improved overall annotation quality (73.3% versus 60.1% BUSCO [19] 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 [6,8,14,20]) failed to clearly identify U12-type introns (STAR Methods and Figure S1A). 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 (STAR Methods)). 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 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 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 S1B).

Combining this position-specific atypical branchpoint motif with species-specific splice site motifs in intronIC [8] (STAR Methods) led to a clearer separation of putative U12- and U2-type introns (Figures 1C, S1C). Using a conservative criterion of 95% U12-type probability (STAR Methods), we identified 20,899 putative U12-type introns in *P. polycephalum* (leaving

154,299 putative U2-type introns with U12-type scores ≤ 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 (STAR Methods, Figure S1D). 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 [10,21] (Figure S2A).

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 [8,22]. 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 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 [7,8,23]. 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 S2B).

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 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 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% *under*representation ($p = 0.038$).

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* (STAR Methods). After filtering for likely reverse

transcriptase artifacts (STAR Methods), 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 2A, S2C). 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 (STAR Methods, Figure S2D-E).

We also scrutinized components of the U12 spliceosome in *P. polycephalum*. A genomic search using Infernal [24] revealed a single candidate for the U12 snRNA (as previously reported in [15]), the component which basepairs with the branchpoint. Strikingly, this sequence exhibits two transition mutations relative to the core branchpoint binding motif (underlined): G<u>C</u>AA<u>A</u>GAA, which produce basepairing potential with the putative <u>T</u>TT<u>G</u>A branchpoint with a bulged A, comparable to the canonical structure (Figure 2B). This apparent complementary evolution of core U12 spliceosomal machinery and branchpoint sequence represents a rare 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 [23] 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 [25,26].

U12-type introns in other species have been reported to have lower splicing efficiency than U2-type introns [12,13,27–29]; 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 [30] 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 (STAR Methods). 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 2C, S3A-B) or in matched pairwise comparisons with neighboring introns from the same genes (Figure S3C). 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 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.

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 [10,31]. 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 [32,33] as well as a number of other recent reports of intron gain [34,35], 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). 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.

Discussion

Over the nearly two decades since the surprising discovery of the existence of the U12 spliceosomal system [1,2], 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 [5,6,8] to 19 (0.05%) in fruitflies [36] to complete absence in diverse lineages [4,15]. 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, ≲ 45 bases from the 3′ splice site (3′SS)) which exactly basepair with complementary stretches of core non-coding RNAs in the splicing machinery [3,37,38]. 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 [4,7,9,39]. Finally, U12-type introns show slow rates of splicing, suggesting inherently low efficiency of the U12 spliceosomal reaction [11–13,40].

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 [32,41,42]. 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 [8,10], evidence for regulation of cell cycle genes by the U12 spliceosome [43] and evidence for regulation of U12 splicing in differentiation [13,44] 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.

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.

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..

Declaration of interests

The authors declare no competing interests.

Figure Legends

Figure 1. Evidence of massive U12-type intron gain in *Physarum polycephalum*. Full species names listed in Table S1. **(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. **(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] SS's 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 S1B. **(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 (STAR Methods, see also Figure S1C). 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.

Figure 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 [15]). 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 S3B), 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 (STAR Methods). For both **C** and **D**, dashed line = median, diamond = mean, whiskers = 1.5 IQR.

Figure 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.

Resource availability

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).

Materials Availability

This study did not generate any new or unique reagents.

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.

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). Annotated coding sequences were extracted from each genome

using custom Python software (Method Details). RNA-seq samples for all species were downloaded from the NCBI SRA database.

Method details

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, and ERR557103-ERR557120) [16,17,45,46]. 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 [47] (v2.5.1). We also separately mapped the reads to the genome using HISAT2 [48] (v2.1.0), allowing for non-canonical splice sites (--pen-noncansplice 0), followed by StringTie [49] (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 [50] protein database, were generated using TransDecoder [51] (v5.0.2), and further refined with the *de novo* transcriptome via PASA [52] (v2.2.0). In addition, an AUGUSTUS [18] (v3.3) annotation was generated *ab initio* from the mapped reads using BRAKER1 [53] (v2.1.0) explicitly allowing for AT-AC splice boundaries (--allow_hinted_splicesites=atac). Lastly, the AUGUSTUS- and StringTie-based gene predictions were merged using gffcompare [54] (v0.10.5), and processed again using TransDecoder. To gauge the quality of our annotations compared to those previously available, we performed a BUSCO [19] (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).

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 [8]. 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 of Figure 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.

Identification of homologs and conserved introns

Genomes and annotations for all additional species were downloaded from various online resources (Table S1), 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 [47,51]. 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 [8] 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 [55,56] (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 [57] (v2.1), and introns occurring at the same position in regions of good local alignment (≥ 4/10 shared amino acid residues on both sides of the intron) were considered to be conserved (based on the approach in [22]).

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 [58] (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 [59] (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 S1D).

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. [60] 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). We then performed one-way BLASTP [55,56] (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 [61] v1.5.2), and p-values were corrected for multiple-testing using the Holm step-down method in the Python library statsmodels [62] (v0.11.1).

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 [63] 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., $\frac{C_s}{(C_r/2) + C_s} \cdot 100\%$. 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 [30] (v1.3.0), which produced an equivalent (inverted) pattern to our splicing efficiency metric (Figure S3A-B). Introns with IRFinder warnings of "LowSplicing" and "LowCover" were excluded.

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 $\frac{(866/1074)}{(208/8695)} \approx \frac{0.806}{0.024} \approx 34$. 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", [64]) 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 $s = 1 - \frac{d}{l}$, 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 ≥ 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 S2D).

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). For each species, a series of RNA-seq samples (curated by size and wild-type status; Table S3) 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 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.

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 STAR Methods 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 [61] package. Figures were generated using Matplotlib [65] (v3.1.1) apart from Figure 3, which was created using graphic design software, and Figure 1E which was manually assembled using output from the R package phytools [66] (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 [62]. 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.

Supplemental information

Table S1. Binomial name abbreviations and genome and annotation metadata for all species.

References

1. Jackson, I.J. (1991). A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. *19*, 3795–3798.

2. Hall, S.L., and Padgett, R.A. (1994). Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J. Mol. Biol. *239*, 357–365.

3. Hall, S.L., and Padgett, R.A. (1996). Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science *271*, 1716–1718.

4. Turunen, J.J., Niemelä, E.H., Verma, B., and Frilander, M.J. (2013). The significant other: Splicing by the minor spliceosome. Wiley Interdiscip. Rev. RNA *4*, 61–76.

5. Alioto, T.S. (2007). U12DB: a database of orthologous U12-type spliceosomal introns. Nucleic Acids Res. *35*, D110–D115.

6. Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. *34*, 3955–3967.

7. Lin, C.-F., Mount, S.M., Jarmołowski, A., and Makałowski, W. (2010). Evolutionary dynamics of U12-type spliceosomal introns. BMC Evol. Biol. *10*, 47.

8. Moyer, D.C., Larue, G.E., Hershberger, C.E., Roy, S.W., and Padgett, R.A. (2020). Comprehensive database and evolutionary dynamics of U12-type introns. Nucleic Acids Res. Available at: http://dx.doi.org/10.1093/nar/gkaa464.

9. Russell, A.G., Charette, J.M., Spencer, D.F., and Gray, M.W. (2006). An early evolutionary origin for the minor spliceosome. Nature *443*, 863–866.

10. Burge, C.B., Padgett, R. a., and Sharp, P. a. (1998). Evolutionary fates and origins of U12-type introns. Mol. Cell *2*, 773–785.

11. Niemelä, E.H., and Frilander, M.J. (2014). Regulation of gene expression through inefficient splicing of U12-type introns. RNA Biol. *11*, 1325–1329.

12. Patel, A.A., McCarthy, M., and Steitz, J.A. (2002). The splicing of U12-type introns can be a rate-limiting step in gene expression. EMBO J. *21*, 3804–3815.

13. Younis, I., Dittmar, K., Wang, W., Foley, S.W., Berg, M.G., Hu, K.Y., Wei, Z., Wan, L., and Dreyfuss, G. (2013). Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. Elife *2*, e00780.

14. Bartschat, S., and Samuelsson, T. (2010). U12 type introns were lost at multiple occasions during evolution. BMC Genomics *11*, 106.

15. Lopez, M.D., Alm Rosenblad, M., and Samuelsson, T. (2008). Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. Nucleic Acids Res. *36*, 3001–3010.

16. Schaap, P., Barrantes, I., Minx, P., Sasaki, N., Anderson, R.W., Bénard, M., Biggar, K.K., Buchler, N.E., Bundschuh, R., Chen, X., *et al.* (2015). The Physarum polycephalum Genome Reveals Extensive Use of Prokaryotic Two-Component and Metazoan-Type Tyrosine Kinase Signaling. Genome Biol. Evol. *8*, 109–125.

17. Glöckner, G., and Marwan, W. (2017). Transcriptome reprogramming during developmental switching in Physarum polycephalum involves extensive remodeling of intracellular signaling networks. Sci. Rep. *7*, 12304.

18. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.

19. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. Available at: http://dx.doi.org/10.1093/bioinformatics/btv351.

20. Burge, C., and Sharp, P.A. (1997). Classification of introns: U2-type or U12-type. Cell *91*, 875–879.

21. Levine, A., and Durbin, R. (2001). A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Res. *29*, 4006–4013.

22. Roy, S.W., Fedorov, A., and Gilbert, W. (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc. Natl. Acad. Sci. U. S. A. *100*, 7158–7162.

23. Basu, M.K., Makalowski, W., Rogozin, I.B., and Koonin, E.V. (2008). U12 intron positions are more strongly conserved between animals and plants than U2 intron positions. Biol. Direct *3*, 19.

24. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

25. Lin, K., and Zhang, D.-Y. (2005). The excess of 5’ introns in eukaryotic genomes. Nucleic Acids Res. *33*, 6522–6527.

26. Roy, S.W., and Gilbert, W. (2005). The pattern of intron loss. Proc. Natl. Acad. Sci. U. S. A. *102*, 713–718.

27. Niemelä, E.H., Oghabian, A., Staals, R.H.J., Greco, D., Pruijn, G.J.M., and Frilander, M.J. (2014). Global analysis of the nuclear processing of transcripts with unspliced U12-type introns by the exosome. Nucleic Acids Res. *42*, 7358–7369.

28. Pessa, H., Ruokolainen, A., and Frilander, M.J. (2006). The abundance of the spliceosomal snRNPs is not limiting the splicing of U12-type introns. RNA *12*, 1883–1892.

29. Tarn, W.Y., and Steitz, J. a. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell *84*, 801–811.

30. Middleton, R., Gao, D., Thomas, A., Singh, B., Au, A., Wong, J.J.-L., Bomane, A., Cosson, B., Eyras, E., Rasko, J.E.J., *et al.* (2017). IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol. *18*, 51.

31. Dietrich, R.C., Incorvaia, R., and Padgett, R.A. (1997). Terminal Intron Dinucleotide Sequences Do Not Distinguish between U2- and U12-Dependent Introns. Mol. Cell *1*, 151–160.

32. Huff, J.T., Zilberman, D., and Roy, S.W. (2016). Mechanism for DNA transposons to generate introns on genomic scales. Available at: http://dx.doi.org/10.1038/nature20110.

33. Henriet, S., Colom Sanmartí, B., Sumic, S., and Chourrout, D. (2019). Evolution of the U2 Spliceosome for Processing Numerous and Highly Diverse Non-canonical Introns in the Chordate Fritillaria borealis. Curr. Biol. *29*, 3193-3199.e4.

34. Gumińska, N., Płecha, M., Zakryś, B., and Milanowski, R. (2018). Order of removal of conventional and nonconventional introns from nuclear transcripts of Euglena gracilis. PLoS Genet. *14*, e1007761.

35. Milanowski, R., Gumińska, N., Karnkowska, A., Ishikawa, T., and Zakryś, B. (2016). Intermediate introns in nuclear genes of euglenids - are they a distinct type? BMC Evol. Biol. *16*, 49.

36. Janice, J., Pande, A., Weiner, J., Lin, C.F., and Makalowski, W. (2012). U12-type Spliceosomal Introns of Insecta. Int. J. Biol. Sci. *8*, 344–352.

37. Tarn, W.Y., and Steitz, J.A. (1996). Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science *273*, 1824–1832.

38. Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. Genes and Development *32*, 577–591.

39. Zhu, W., and Brendel, V. (2003). Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. Nucleic Acids Res. *31*, 4561–4572.

40. Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat. Struct. Mol. Biol. *16*, 1128–1133.

41. Sverdlov, A.V., Rogozin, I.B., Babenko, V.N., and Koonin, E.V. (2004). Reconstruction of ancestral protosplice sites. Curr. Biol. 14, 1505–1508.

42. Roy, S.W., Gozashti, L., Bowser, B.A., Weinstein, B.N., and Larue, G.E. (2020). Massive Intron Gain in the Most Intron-Rich Eukaryotes is Driven by Introner-Like Transposable Elements of Unprecedented Diversity and Flexibility. bioRxiv 2020.10.14.339549

43. Baumgartner, M., Olthof, A.M., Aquino, G.S., Hyatt, K.C., Lemoine, C., Drake, K., Sturrock, N., Nguyen, N., Al Seesi, S., and Kanadia, R.N. (2018). Minor spliceosome

inactivation causes microcephaly, owing to cell cycle defects and death of self-amplifying radial glial cells. Development 145.

44. Meinke, S., Goldammer, G., Weber, A.I., Tarabykin, V., Neumann, A., Preussner, M., and Heyd, F. (2020). Srsf10 and the minor spliceosome control tissue-specific and dynamic SR protein expression. Elife 9.

45. Glöckner, G., Golderer, G., Werner-Felmayer, G., Meyer, S., and Marwan, W. (2008). A first glimpse at the transcriptome of Physarum polycephalum. BMC Genomics *9*, 6.

46. Barrantes, I., Leipzig, J., and Marwan, W. (2012). A next-generation sequencing approach to study the transcriptomic changes during the differentiation of Physarum at the single-cell level. Gene Regul. Syst. Bio. *2012*, 127–137.

47. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D. a., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

48. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

49. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. *11*, 1650–1667.

50. UniProt Consortium (2008). The universal protein resource (UniProt). Nucleic Acids Res. *36*, D190-5.

51. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., *et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.

52. Haas, B.J. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666.

53. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2015). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. Available at: http://dx.doi.org/10.1093/bioinformatics/btv661.

54. Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Res. *9*, 304.

55. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

56. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

57. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

58. Sievers, F., and Higgins, D.G. (2014). Clustal Omega. Curr. Protoc. Bioinformatics *2014*, 3.13.1-3.13.16.

59. Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

60. Kang, S., Tice, A.K., Spiegel, F.W., Silberman, J.D., Pánek, T., Cepicka, I., Kostka, M., Kosakyan, A., Alcântara, D.M.C., Roger, A.J., *et al.* (2017). Between a Pod and a Hard Test: The Deep Evolution of Amoebae. Mol. Biol. Evol. *34*, 2258–2270.

61. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

62. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (Austin, TX), p. 61.

63. Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinformatics *32*, 11–17.

64. Roy, S.W., and Irimia, M. (2008). When good transcripts go bad: Artifactual RT-PCR "splicing" and genome analysis. Bioessays *30*, 601–605.

65. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95.

66. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223.

A

I. RAB family GTPase (XLOC_000293)

GCG TCT TTT AAA A**gt**atctaatcttttc...ctttgacatttctc**at**AT ATA AAT CGG

	10	20	30
P. polycephalum	Y Y R N A A G A L I V F D I T N Q A S F K N ^ I N R W L T D		
A. thaliana	Y Y R G A V G A L V V Y D I T R S S T F E N V G R W L D E		
P. tetraurelia	Y Y K G A A G A L L I F D I T D R S S F E N L D K W L K D		
X. tropicalis	Y Y R G A M G I M L V Y D I T N A K S F E N I S K W L R N		

II. RAB family GTPase (XLOC_057070)

ACC TTA CAC ATT**at**atccttacatca...tcattgacaccatgt**ac**TGG GAT ACT GCA

	10	20	30
P. polycephalum	K P K E I T V G K K T H T L H I ^ W D T A G Q E R F G T I T		
B. oleracea	K I K Q L T V G D K R L K L T I W D T A G Q E R F R T L T		
D. purpureum	K N K N I V I D G R T F S L Q I W D T A G Q E R F R T I T		
T. flavidus	Q M K T L T V G S A T V T L Q L W D T A G Q E R F C S I T		

III. ARF family GTPase (XLOC_042355)

TCA AAA GCT ATG**gt**atcttcactat...ttttgatcacgtgt**at**ATA TTT ATG GTC

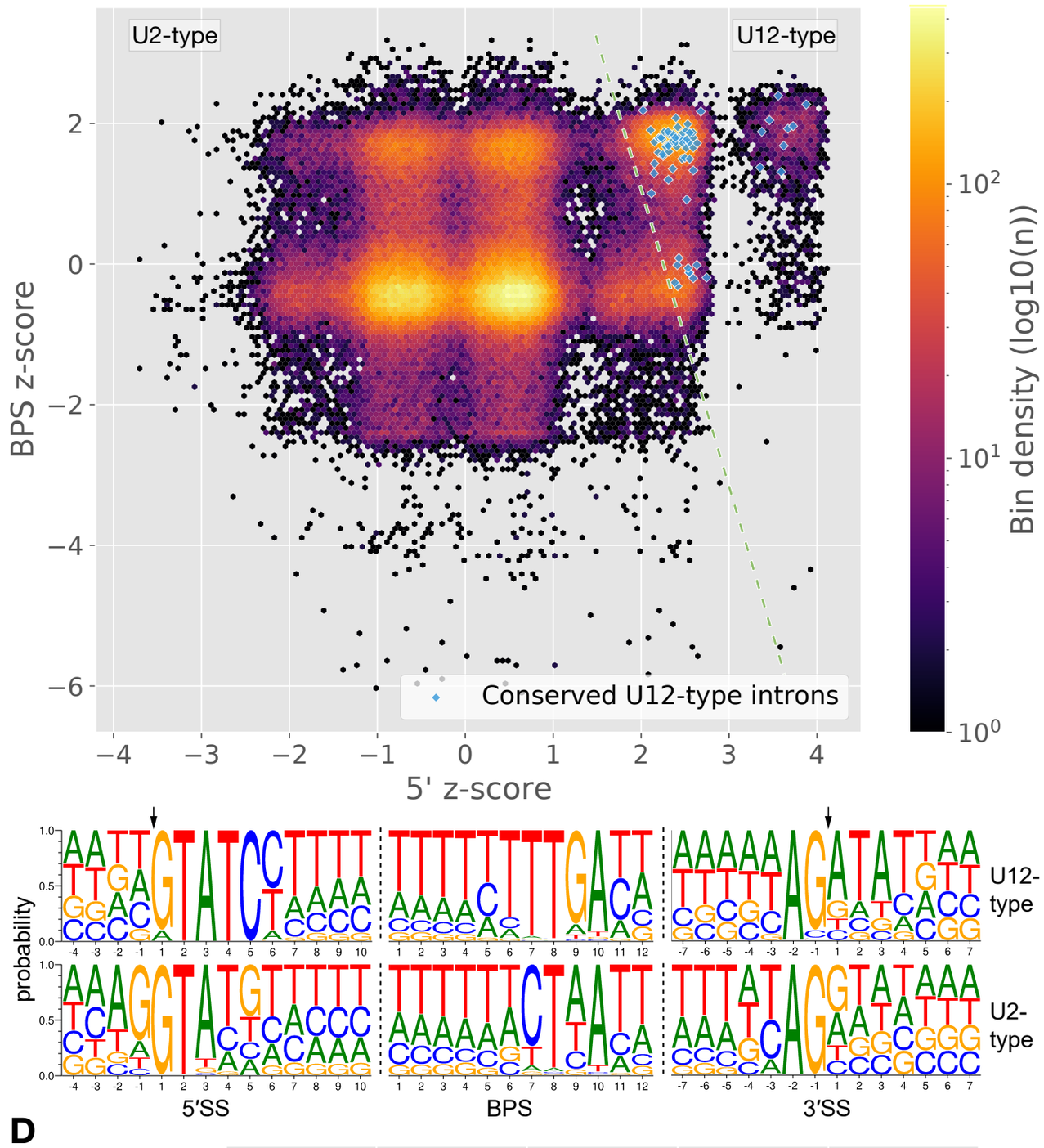
	10	20	30
P. polycephalum	G G E K L - R N L W R H Y Y D G S K A M ^ I F M V D S T N K		
T. rangeli	G G Q K L L R S L W K H Y Y E G A N A I I F V I D S N D R		
S. curassoni	G Q E K I - R P L W R H Y F T G S Q G L I F V V D S S D R		
E. huxleyi	G Q D K I - R A L W S H Y Y D G A N A L I F V V D S S D E		

IV. ARF family GTPase (XLOC_049766)

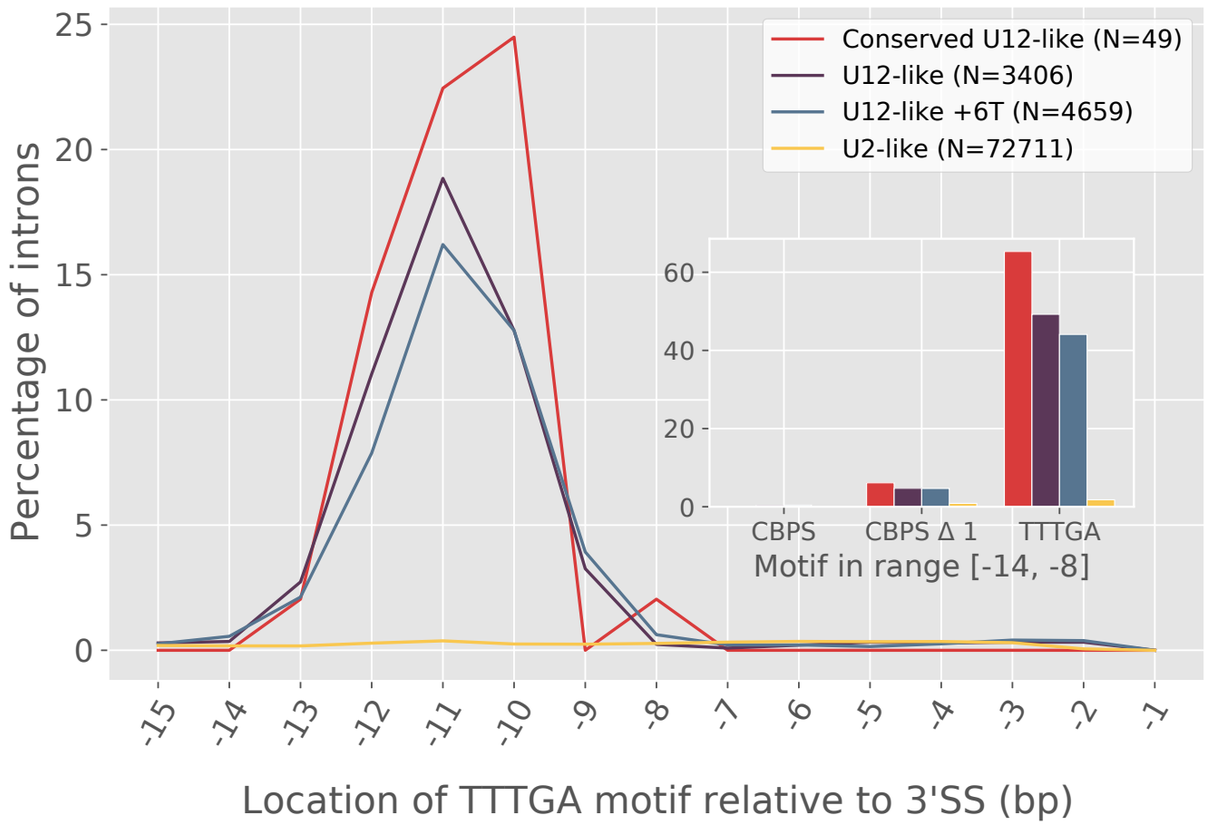
ACT ACT ATT TTA T**tt**atccttcataat...ttttgatggacacat**aa**AT AAA CTA CAA

	10	20
P. polycephalum	I L G L D S A G K T T I L Y ^ K L Q C G E V V T T I P	
P. troglodytes	I L G L D G A G K T T I L Y R L Q V G E V V T T I P	
A. castellanii	I L G L D G A G K T T I L Y R L Q C G E V V S T L P	
G. barbadense	V L G L D N A G K T T I L Y R L Q M G E V V S T I P	

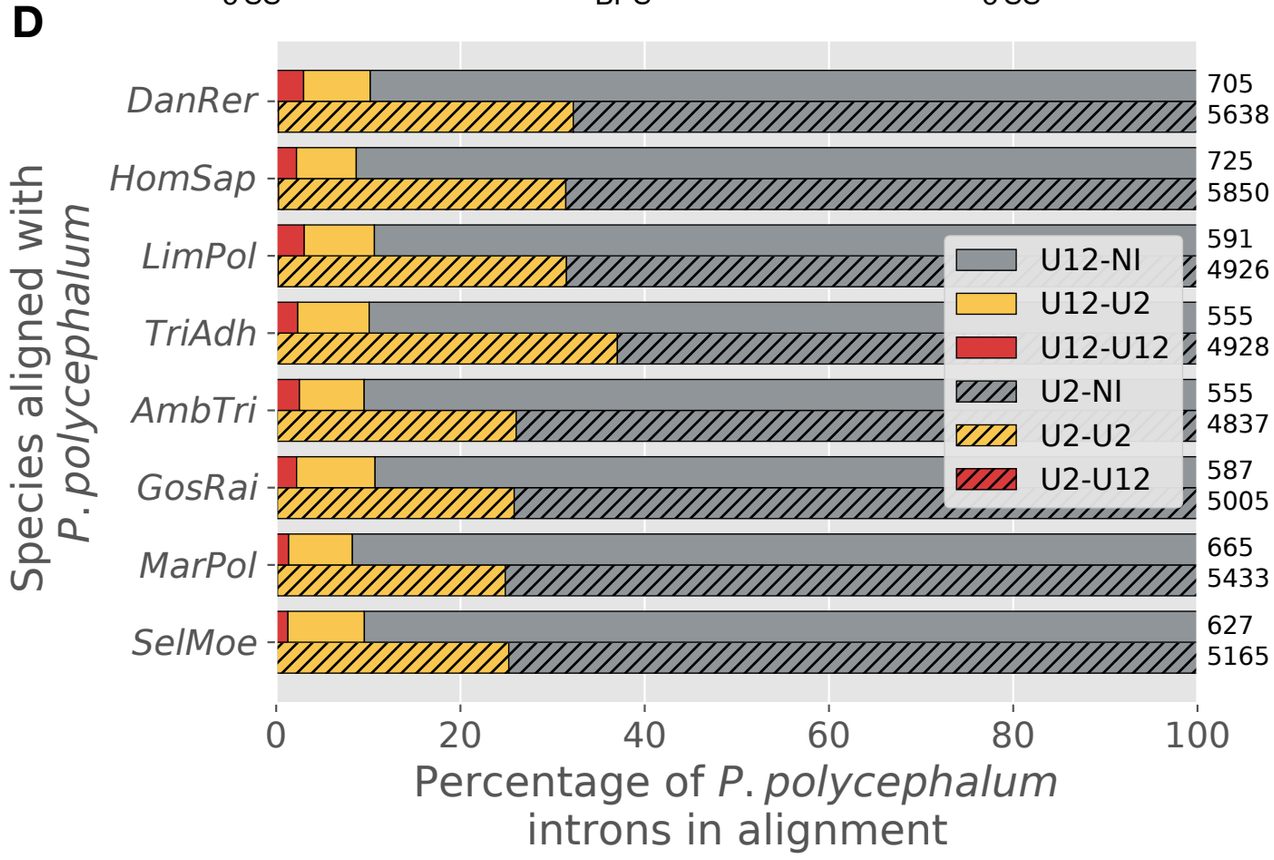
C



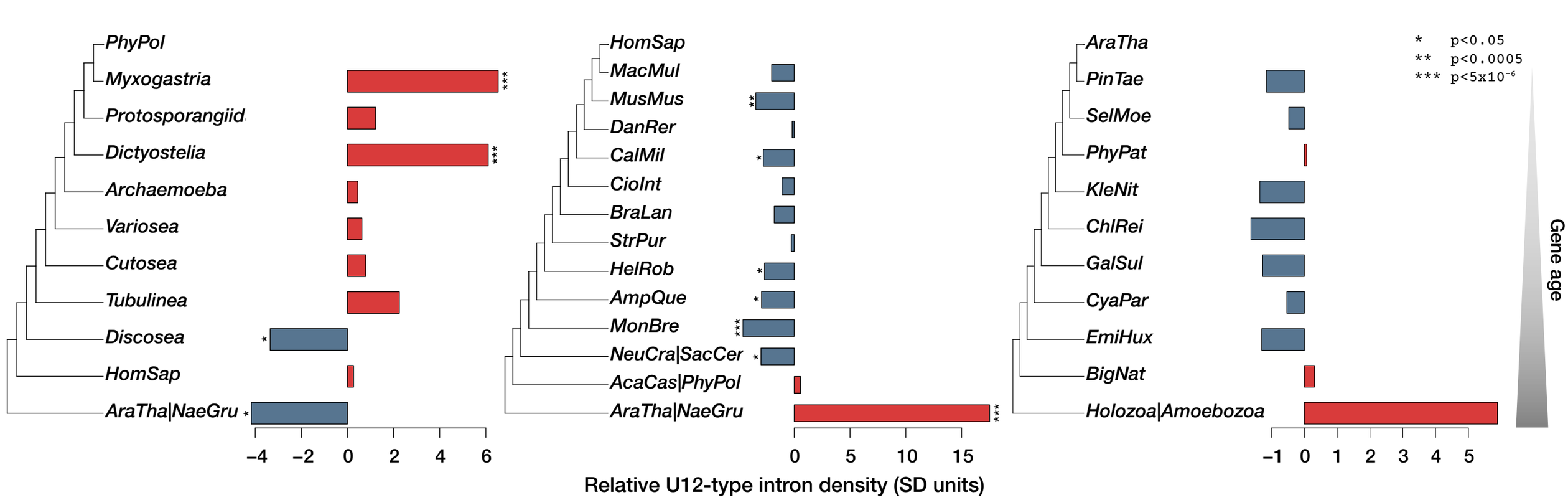
B

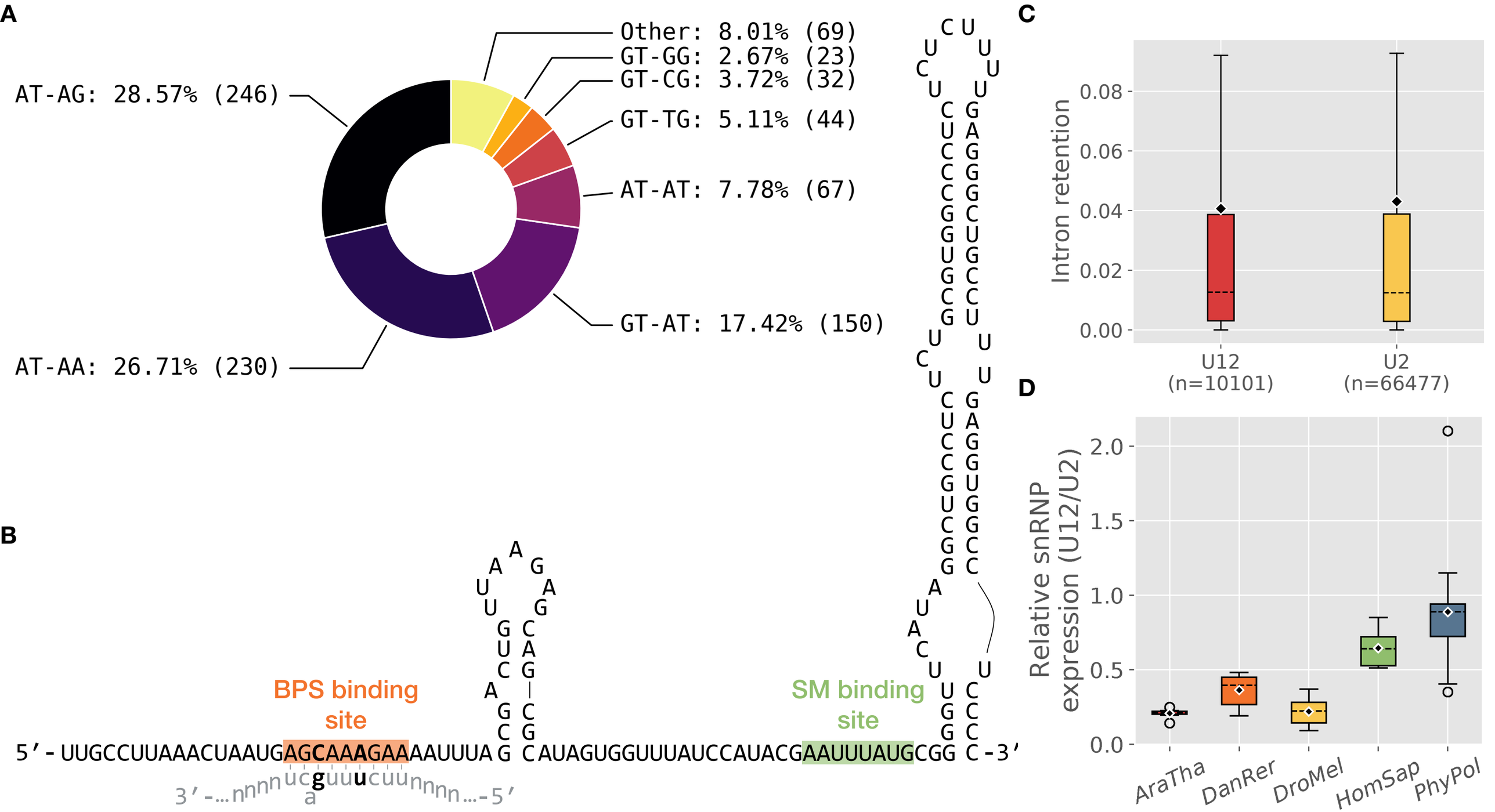


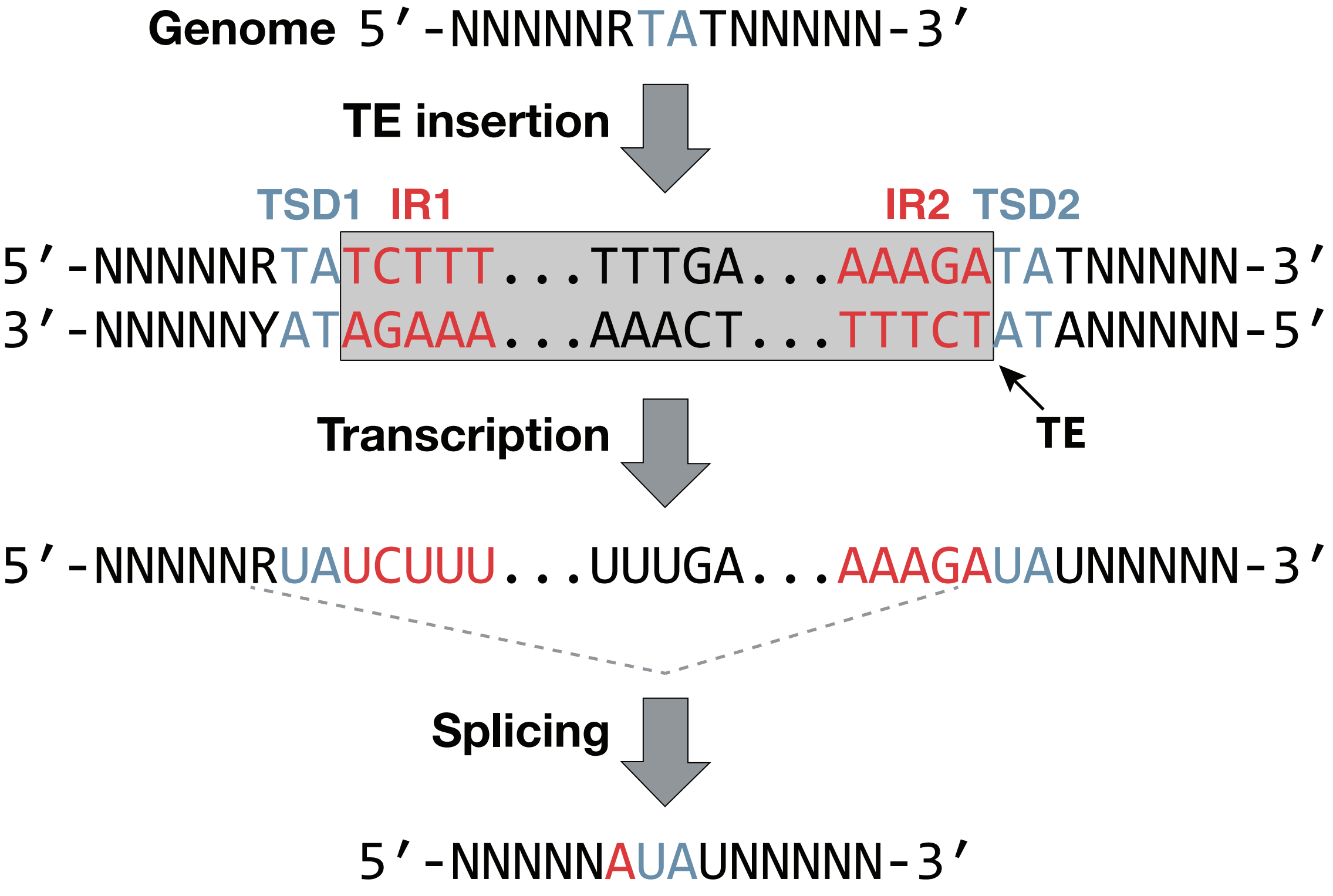
D



E




A
Other: 8.01% (69)
GT-GG: 2.67% (23)
GT-CG: 3.72% (32)
GT-TG: 5.11% (44)
AT-AT: 7.78% (67)
GT-AT: 17.42% (150)
AT-AA: 26.71% (230)
AT-AG: 28.57% (246)
B
BPS binding site
SM binding site
5′-UUGCCUUAAACUAAUGAGCAAAGAAAAUUUAGCAUAGUGGUUUAUCCAUACGAAUUUAUGCGGC-3′
3′-…nnnnucguuucuunnnnn…-5′
a
C
Intron retention
0.00
0.02
0.04
0.06
0.08
U12 (n=10101)
U2 (n=66477)
D
Relative snRNP expression (U12/U2)
0.0
0.5
1.0
1.5
2.0
AraTha
DanRer
DroMel
HomSap
PhyPol


Genome 5′-NNNNNRTATNNNNN-3′
TE insertion
TSD1 IR1
IR2 TSD2
5′-NNNNNRTATCTTT...TTTGA...AAAGATATNNNNN-3′
3′-NNNNNYATAGAAA...AAACT...TTTCTATANNNNN-5′
TE
Transcription
5′-NNNNNRUAUCUUU...UUUGA...AAAGAUAUNNNNN-3′
Splicing
5′-NNNNNAUAUNNNNN-3′

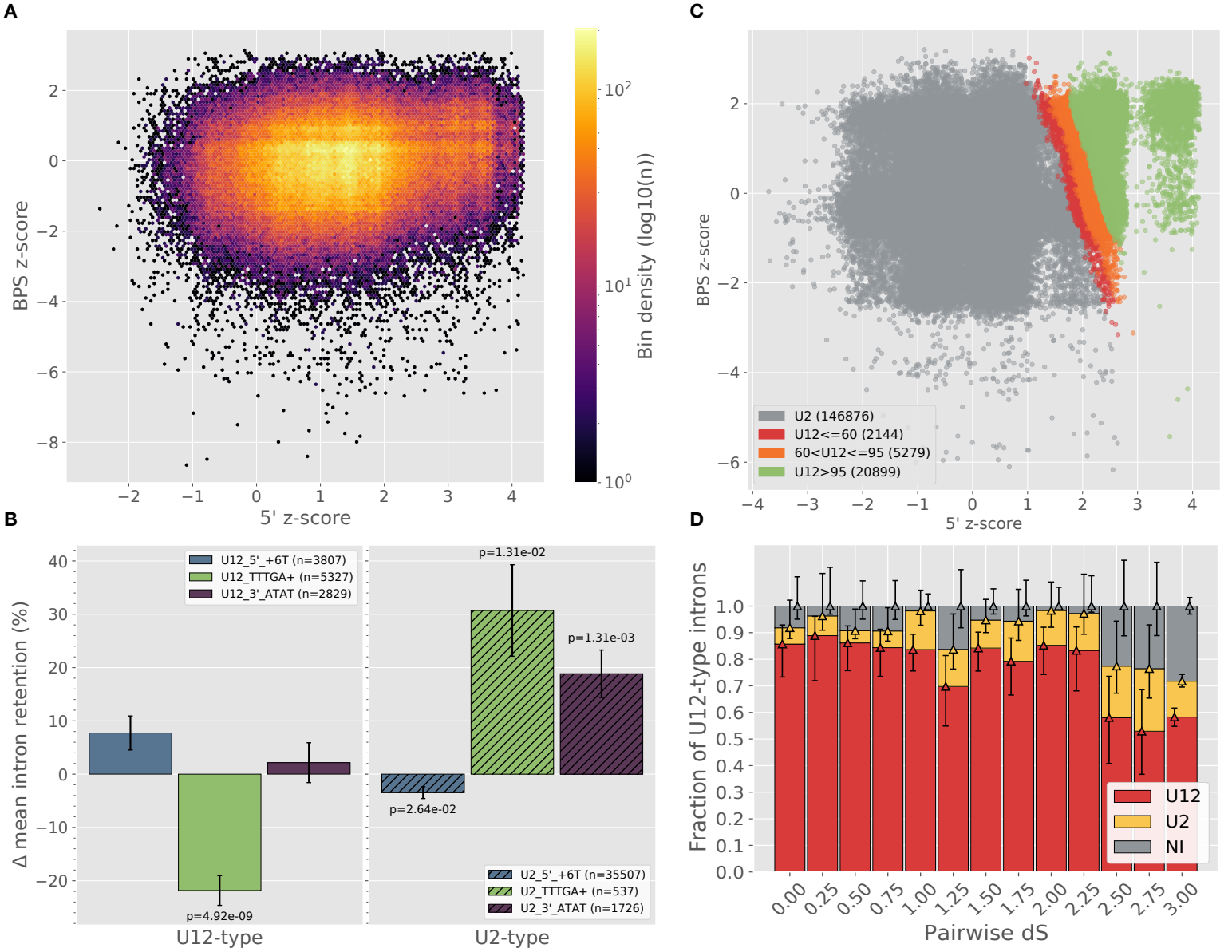


Figure S1. Classifier details, motif-specific intron retention and U12-type intron conservation within paralogs, related to Figures 1B,C and the paralog analysis section of the STAR methods (A) The default PWMs used by intronIC are derived from human introns, and for divergent motifs like those present 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 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. **(C)** BPS-vs-5'SS score plot with assigned classifications for all *P. polycephalum* introns. The same underlying data as Figure 1C, where each point represents an intron, and the color indicates the U12-type probability classification (U2-type: gray; U12-type with probability ≤ 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 ≥ 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 ("NI", 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 ≈ 2.5. Whiskers represent the binomial proportion confidence intervals (Wilson score intervals) for the three categories (category indicated by color of associated triangle).

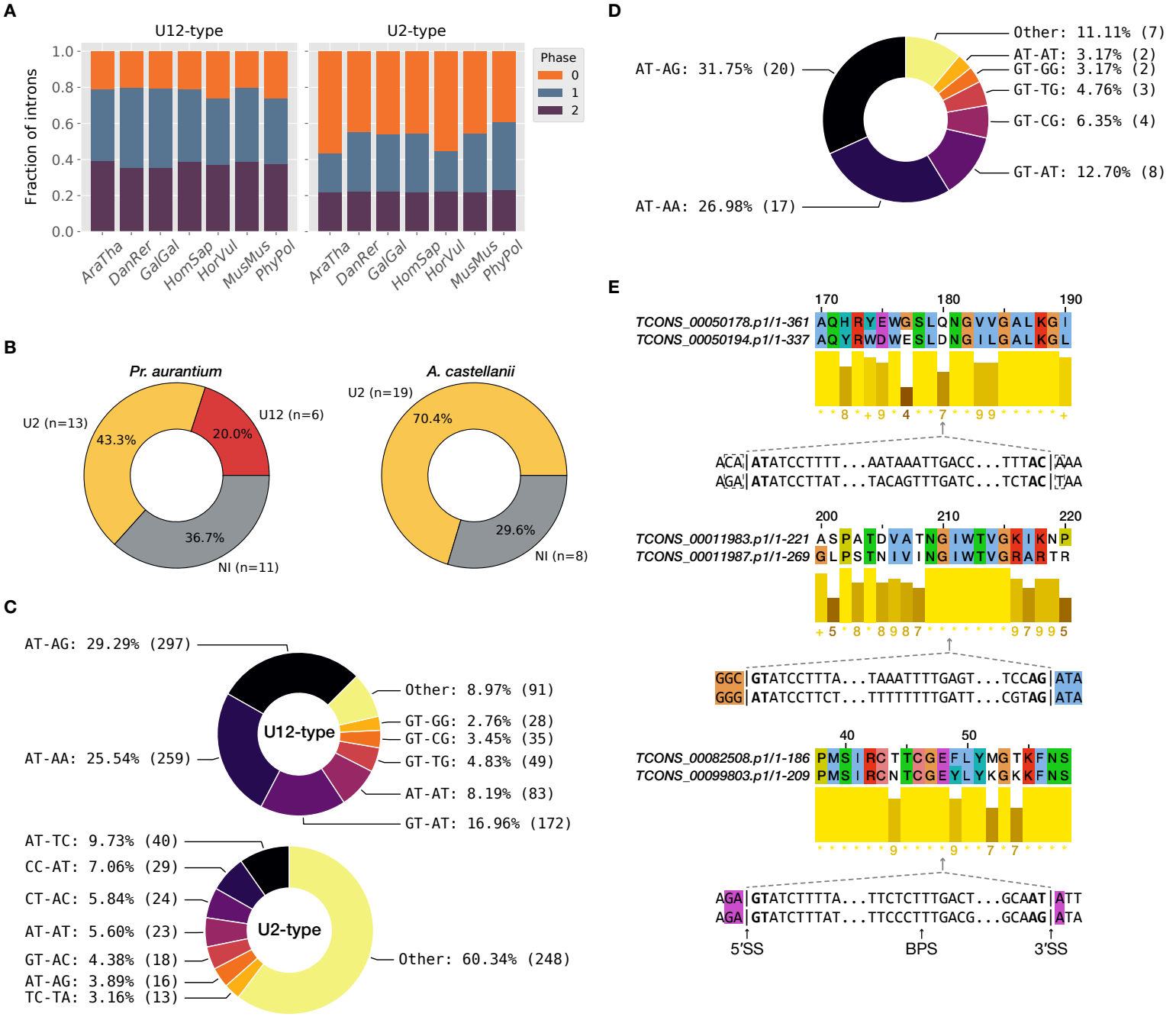


Figure S2. Features and comparisons of U12-type introns in *P. polycephalum*, related to Figures 1D, 2A and the Results and STAR methods sections (A) Phase distribution of U12- (left) and U2-type (right) introns across different species. U12-type introns in *P. polycephalum* (*PhyPol*), as in other species, 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 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. **(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 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)—thus 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.

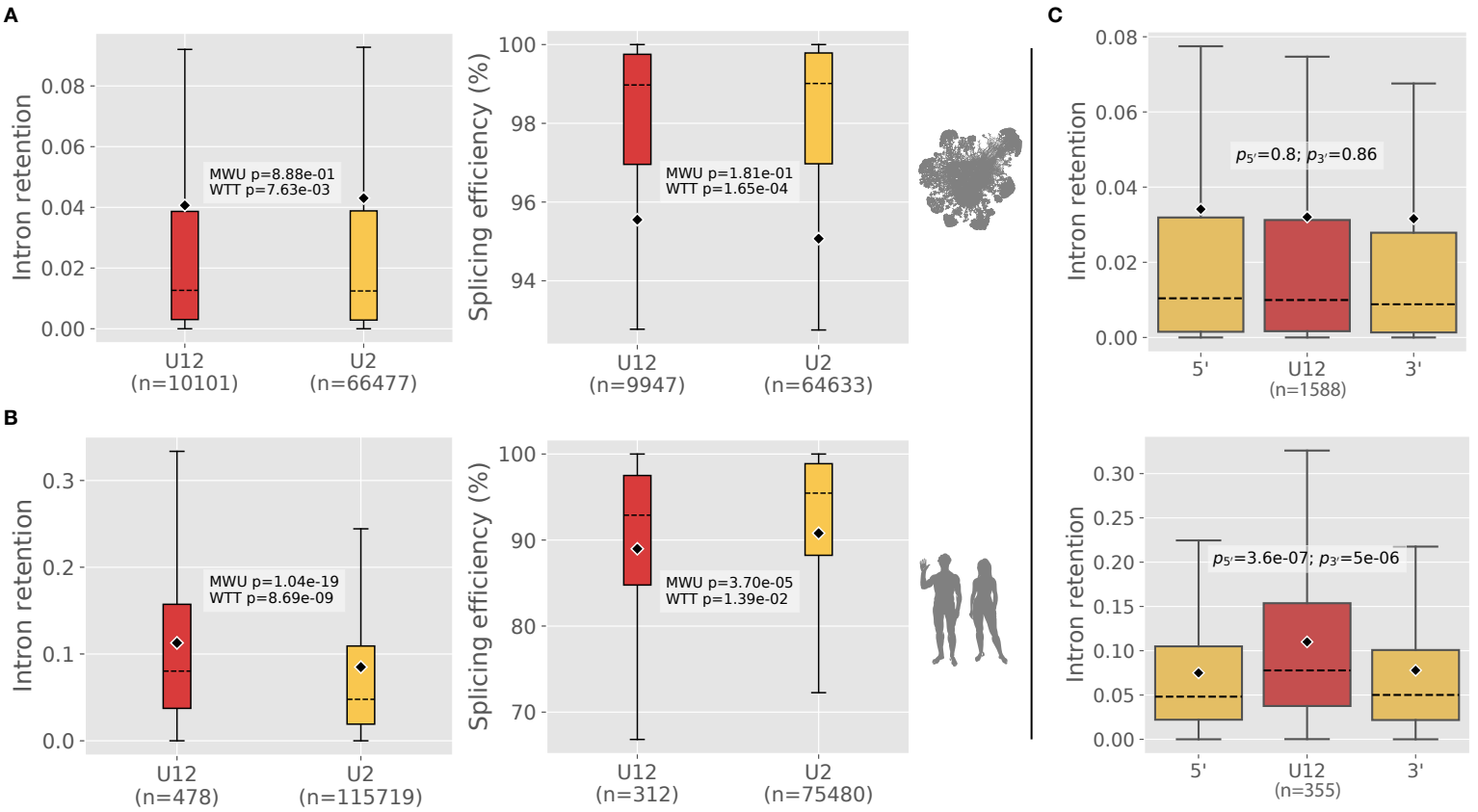


Figure S3. Intron retention and splicing efficiency, related to Figure 2C and the intron retention/splicing efficiency section of the STAR methods (A) Comparison of average intron retention (left) and splicing efficiency (right) data for *P. polycephalum* introns, showing that U12-type introns are neither 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 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. **(C)** 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 *P. polycephalum*, > 90% in human) and neighboring U2-type (yellow; ≤ 5% U12-type probability in *P. polycephalum*, ≤ 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.

snRNA gene information			species-specific transcript ID				
spliceosomal system	**RefSeq protein accession**	**gene name**	***Arabidopsis thaliana***	***Danio rerio***	***Drosophila melanogaster***	***Homo sapiens***	***Physarum polycephalum***
major	NP_005868	SF3a120, SAP114	AT1G14650.1	ENSDART00000061378	FBtr0083374	ENST00000215793	TCONS_00075169.p1
major	NP_006793	SF3a60, SAP61	AT5G06160.1	ENSDART00000160433	FBtr0078751	ENST00000373019	TCONS_00069287.p1
major	NP_004587	U1 A	-	ENSDART00000012018	-	ENST00000243563	TCONS_00026432.p1
major	NP_003080	U1-70K	AT3G50670.1	-	FBtr0079355	ENST00000598441	TCONS_00063511.p1
major	NP_003081	U2 A'	AT1G09760.1	ENSDART00000128530	FBtr0088941	ENST00000254193	TCONS_00018540.p1
minor	NP_061976	U11/U12-20K	AT5G26749.2	ENSDART00000144510	FBtr0077965	ENST00000344318	TCONS_00023854.p1
minor	NP_078847	U11/U12-25K	AT3G07860.1	ENSDART00000006783	-	ENST00000383018	TCONS_00025616.p1
minor	NP_149105	U11/U12-31K (MADP1)	AT3G10400.1	ENSDART00000067172	-	ENST00000266529	TCONS_00052334.p1
minor	NP_851030	U11/U12-35K	AT2G43370.1	ENSDART00000172191	-	ENST00000412157	TCONS_00095609.p1
minor	NP_060089	U11/U12-65K	AT1G09230.1	ENSDART00000017427	FBtr0339103	ENST00000423855	TCONS_00087173.p1

Table S2. Gene/annotated transcript metadata for minor/major spliceosomal components used in the snRNP relative expression analysis, related to Figure 2D.

species	RNA-seq accession numbers used in snRNP analysis
Arabidopsis thaliana	SRR5197911, SRR5197985, SRR5197986, SRR5820083, SRR6874226, SRR7663610, SRR7726615, SRR9019675, SRR9265357, SRR934391, SRR9995072
Danio rerio	ERR3365998, SRR4017373, SRR5274769, SRR6384886, SRR8441448, SRR8922974, SRR8944755, SRR9966394, SRR9966395, SRR9966396
Drosophila melanogaster	ERR1145740, ERR1145741, ERR1145742, ERR1145743, ERR1145746, ERR1145750, SRR10005788, SRR6652839, SRR6652840, SRR6665463, SRR7450865, SRR8949126
Homo sapiens	SRR1617450, SRR1617451, SRR1617452, SRR1617454, SRR1617461, SRR5442314, SRR5442315, SRR5442317

Table S3. RNA-seq accession numbers used in the snRNP relative expression analysis, related to Figure 2D. Accession numbers for *Physarum polycephalum* are listed in the STAR methods (all available RNA-seq samples were used).