Updated Database and Evolutionary Dynamics of U12-Type Introns

Devlin C. Moyer[1,†,*], Graham E. Larue[2,†,*], Courtney E. Hershberger[1], Scott W. Roy[3*], Richard A. Padgett[1]

[1] Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute, Cleveland Clinic Lerner College of Medicine, Cleveland Clinic and Department of Molecular Medicine, Case Western Reserve University, Cleveland, Ohio 44106, USA; [2] Department of Molecular and Cell Biology, University of California, Merced, California 95343, USA; [3] Department of Biology, San Francisco State University, San Francisco, California 94132, USA

* To whom correspondence should be addressed. Email: devmoy@gmail.com. Correspondence may also be addressed to GEL egrahamlarue@gmail.com and SWR (scottwroy@gmail.com).

† Joint Authors

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.

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 (1,2). 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) spliceosome. 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 (3,4). The minor class introns have consensus splice site and branch point sequences distinct from the major class introns (5,6). 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 (7).

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 (8,9). 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)[9] 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" (11—13). Shortly after the discovery of U12-type introns (5), 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 (14). Furthermore, the phase biases in U12-type introns were noted to be different from the previously-documented phase biases in U2-type introns (14,15).

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 intron class. The website is publically accessible at introndb.lerner.ccf.org.

MATERIALS AND METHODS

Annotating intron class with intronIC

To begin, intronIC identifies all intron sequences in an annotation file by interpolating between coding features (CDS or exon) within 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 -45 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 most U2 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 BPS motifs from all introns in the final dataset whose 5'SS U12 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 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 (4,14—17).

As originally shown by Burge et al., the 5'SS and BPS scores together are sufficient to produce fairly good binary clustering of introns into putative types, due to strong correspondence between the 5'SS and BPS scores in U12-type introns (4,14). While this general feature of the data has often been employed in the identification of U12-type introns, a variety of different approaches 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 (18) to assign intron

types, an approach which produces good results across a diverse set of species and provides an easy-to-interpret scoring metric.

This 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. Initially, PWMs from SpliceRack (16) were used to establish strong U12-type intron candidates in human, whose parent genes were then compared to orthologs across a variety of species to identify human introns with robust conservation as U12- or U2-type. In addition, U12-type introns from previously-published reports (19—21) with good matches to the U12-type 5′SS consensus were collected. Together, these high-confidence U12-type and U2-type intron sets were used to build an updated collection of PWMs, and to define positive and negative training data for the SVM.

Because clear discrimination between U12-type and U2-type introns can be achieved by considering only two scoring dimensions (4,14,15), we use a relatively simple SVM classifier with a linear kernel as implemented in the scikit-learn Python library (22). 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 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 values 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.

Once trained on the conserved intron data, the SVM classifier is used to assign to every intron in the experimental set a probability from 0-100% of being U12-type (a transformation achieved within the scikit-learn library via Platt scaling). In an attempt to minimize false-positive calls in our classification process, we define as U12-type only those introns with assigned U12 probabilities greater than 90%. Although this threshold is likely to be at least somewhat conservative to the full diversity of U12-type introns, it produces classifications in good agreement with previously-reported findings while also increasing the absolute numbers of U12-type introns overall. For example, running our method on U12-type intron sequences from the U12DB (21) resulted in equivalent classifications for 98% (393/399) of the U12DB introns in chicken, 98% (545/554) in mouse, 100% (16/16) in *Drosophila melanogaster* and 98% (676/692) in human. In *Arabidopsis thaliana*, our method matched the calls in the U12DB 96% of the time (228/238), with very similar results (277/289) for U12-type introns from the plant-specific database ERISdb (23). Furthermore, in each test species listed above intronIC identified additional putative U12-type introns not present in existing databases, likely due to some combination of newer annotation data

and our method's sensitivity. To empirically estimate a false-positive rate for our method, we considered *Caenorhabditis elegans*, a well-annotated species believed to have lost all of its U12-type introns, where we found 4/108073 introns categorized as U12-type. These introns have been previously reported (16) as having spuriously U12-like 5′SS motifs, and therefore suggest a false-positive rate on the order of 0.0037%.

For the purpose of populating the IAOD, intronIC was slightly modified to produce a single output file containing all of the annotation information recorded in the IAOD for each intron—the default 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 (24). 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.

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 it uses exon entries to define introns, it cannot annotate intron phases. 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.

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 (25). BiomaRt (26) is an R (27) 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.

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 (28) to identify sets of best reciprocal hits—considered as 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 Clustal W (v2.1; ref. 29), and all introns in regions of good local alignment between pairs (≥40% matching amino acid sequence ±10 residues around each intron) were extracted using custom Python scripts (following the approach of ref. 30). 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 → A-B-C, B-D).

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 (80 nt), 3' terminus with the branch point region enclosed with brackets (40 nt), downstream exonic sequence (80 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.

Website design

The website was constructed using Django 2.0 (31), an open-source Python web development framework, and Bootstrap 4.0.0 (https://getbootstrap.com/), 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.

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 (27) was used to compute the probability of observing the observed number of genes with multiple U12-type introns in each genome. Table S1 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 artefact 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.

RESULTS AND DISCUSSION

We used intronIC to perform genome-wide identification of U2- and U12-type spliceosomal introns in 25 eukaryotic species including 14 vertebrate animals, 5 invertebrate animals, 4 plants and two yeasts. High-confidence U2- and U12-type introns in human were curated from multiple sources and used to generate type-specific position-weight matrices (PWMs) for the 5' splice site and branch point sequences. These PWMs were used to create score vectors for every intron in each genome, which were then compared against the high-confidence sets using a machine-learning classifier to assign each intron a probability of being U12-type (see Methods). Introns scoring above 90% were considered putative U12-type introns in subsequent analyses. A total of 8,967 U12-type introns were identified using this technique. These data are compiled in the IAOD. Figure 1 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 (21), SpliceRack (16), and ERISdb (23). The IAOD generally annotates many more introns than U12DB in all species, 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 (21), while introns in the IAOD were annotated directly from every genome in the database using the intron-classifying program intronIC (see Methods for details). U12DB primarily annotates U12-type introns in all represented species that are orthologous to the reference U12-type introns (21), while the IAOD annotates U12-type introns in all genomes irrespective of orthology. 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 does not provide orthology 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 annotated ortholog.

As shown in Figure 1, 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 (14,17). The numbers of introns determined by intronIC to be U12-type in a few species deserve special attention. 5 introns in *C. elegans* were annotated by intronIC as being U12-type, despite the well-documented absence of the U12-type spliceosome in *C. elegans* (14,16,17). Further inspection of these introns reveals that their intronIC scores represent borderline cases, with all four scores falling just above the threshold for discriminating the two classes (see Methods). All of these introns were previously identified by Sheth et al. 2006 as having consensus U12-type 5' splice sites but very poor matches to the consensus U12-dependent branch point regions. Thus, these are almost certainly false positives arising from the inherent limitations of a sequence-based scoring system. The numbers of U12-type introns in *A. thaliana*, *O. sativa* and *Z. mays* are noteworthy because 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 409 U12-type splice sites annotated in ERISdb, there are only 286 unique sequences, which is much closer to the 293 annotated by the IAOD.

The phase biases observed in the IAOD (Figure 2) agree with the results of previous studies and extend them to many more organisms: an excess of phase 0 introns among U2-type introns (8,9,14,15,32,33), and a bias against phase 0 introns among U12-type introns (14,15) 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 (34), insertion of introns into proto-splice sites (11,33), and preferential loss of phase 1 and 2 introns (6). 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. (14). 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. (7) showed that U12-type introns could be converted to U2-type introns with as few as two point mutations. These results also suggest that class conversion is likely to only proceed from U12-type to U2-type. 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 3, 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 (32,35). Figure 3 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 4). 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 (36,37). 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 (38). 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 (37) and were thus largely unaffected by the class conversion process.

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 (36-37), 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 U12-type spliceosome (7). Table 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.

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 (14—16). 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 Methods and Table S1). Many explanations for this nonrandom distribution have been proposed, including the fission-fusion model of intron evolution (14); a difference in the speed of splicing of U12-type and U2-type introns (16,20); 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 (39). The fission-fusion model posits that two separate lineages of the proto-eukaryote 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 (14). 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 (15,20). 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. (40). Furthermore, recent evidence suggests that the rates of both types of splicing are sufficiently fast that most introns will be excised cotranscriptionally (41).

Basu et al. (42) 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.

The majority of introns annotated in the IAOD in both classes begin with GT and end with AG (Table 2), in agreement with previous studies (14,16,43). 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 (6,15,44—46). 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 (16,43,44,47), 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). 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) largely agrees with previous investigations (16,17,48).

Figure 5 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 introns in humans and chicken while the distribution of U12-type intron lengths has only one peak (15,49) (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 (50), producing a distribution similar to those observed in the human and chicken genome in Figure 5. However, Figure 5 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 (51). However, Figures 6 and 7 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 (49,50,52). 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 6 and 7 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 (53,54). 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.

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. 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 (6,11,33,34), but do support an extension of the class conversion model proposed in ref. 14.

DATA AVAILABILITY

The IAOD is publically accessible at www.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.

FUNDING

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

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. DM wrote the manuscript with input from all authors.

ACKNOWLEDGEMENTS

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.

REFERENCES

1. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. 2012. The significant other: splicing by the minor spliceosome. *WIREs RNA* **4**:61-76 doi: 10.1002/wrna.1141

2. Chen W and Moore MJ. 2014. The spliceosome: disorder and dynamics defined. **24**:141-149 doi: 10.1016/j.sbi.2014.01.009

3. Russell AG, Charette JM, Spencer DF, Gray MW. 2006. An early evolutionary origin for the minor spliceosome. *Nature* **443**:863-866 doi: 10.1038/nature05228

4. Bartschat S and Samuelsson T. 2010. U12 type introns were lost at multiple occasions during evolution. *BMC Genomics* **11**:106 doi: 10.1186/1471-2164-11-106

5. Hall SL and Padgett RA. 1996. Requirement of U12 snRNA for in Vivo Splicing of a Minor Class of Eukaryotic Nuclear Pre-mRNA Introns. *Science* **271**:1716-1718 doi: 10.1126/science.271.5256.1716

6. Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biol Direct* **7**:11 doi: 10.1186/1745-6150-7-11

7. Dietrich RC, Incorvaia R, Padgett RA. 1997. Terminal Intron Dinucleotide Sequences Do Not Distinguish between U2- and U12-Dependent Introns. *Mol Cell* **1**:151-160 doi: 10.1016/S1097-2765(00)80016-7

8. Long M, Rosenberg C, Gilbert W. 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc Natl Acad Sci USA* **92**:12495-124999 doi: 10.1073/pnas.92.26.12495

9. Long M, de Souza SJ, Gilbert W. 1995. Evolution of the intron-exon structure of genes. *Curr Opin Genet Dev* **5**:774-778 doi: 10.1016/0959-437X(95)80010-3

10. Gilbert W. 1987. The Exon Theory of Genes. *Cold Spring Harb Symp Quant Biol* **52**:901-905

11. Dibb NJ, Newman AJ.1989. Evidence that introns arose at proto-splice sites. *EMBO J* **8**:2015-2021 doi: 10.1002/j.1460-2075.1989.tb03609.x

12. Dibb NJ. 1991. Proto-splice Site Model of Intron Origin. *J Theor Biol* **151**:405-416

13. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2004. Reconstruction of Ancestral Protosplice Sites. *Curr Biol* **14**:1505-1508 doi: 10.1016/j.cub.2004.08.027

14. Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary Fates and Origins of U12-Type Introns. *Mol Cell* **2**:773-785 doi: 10.1016/S1097-2765(00)80292-0

15. Levine A and Durbin R. 2001. A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res* **29**:4006-4013 doi: 10.1093/nar/29.19.4006

16. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34**:3955-3967 doi: 10.1093/nar/gkl556

17. Lin CF, Mount SM, Jarmołowski A, Makałowski W. 2010. Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evol Biol* **10**:47 doi: 10.1186/1471-2148-10-47

18. Cortes C and Vapnik V. 1995. Support-vector networks. *Mach Learn* **20**:273–297

19. Madan V, Kanojia D, Li J, Okamoto R, Sato-Otsubo A, Kohlmann A, Sanada M, Grossman V, Sundaresan J, Shiraishi Y et al. 2015. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat Commun* **6**:1–14 doi: 10.1038/ncomms7042

20. Niemelä EH and Frilander MJ. 2014. Regulation of gene expression through inefficient splicing of U12-type introns. *RNA Biol* **11**:1325–1329 doi: 10.1080/15476286.2014.996454

21. Alioto TS. 2006. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Research* **35**:D110-D115 doi: 10.1093/nar/gkl796

22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P Weiss R, Dubourg V et al. 2012 Scikit-learn: Machine Learning in Python. *J Mach Learn Res* doi: 10.1007/s13398-014-0173-7.2

23. Szcześniak MW, Kabza M, Pokrzywa R, Gudyś A, Makałowska I. 2013. ERISdb:A Database of Plant Splice Sites and Splicing Signals. *Plant Cell Physiol* **54**:e10 doi: 10.1093/pcp/pct001

24. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giro´n CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**:D754-D761 doi: 10.1093/nar/gkx1098

25. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005). BioMart and Bioconductor:a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**:3439-3440 doi: 10.1093/bioinformatics/bti525

26. Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**:1184-1191 doi: 10.1038/nprot.2009.97

27. R Core Team. 2017. R:A language and environment for statistical computing. *R Foundation for Statistical Computing*

28. Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**:59-60. Doi: 10.1038/nmeth.3176

29. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*. **23**:2947-8. doi: 10.1093/bioinformatics/btm404

30. Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci* **100**:7158-62 doi: 10.1073/pnas.1232297100

31. Django Core Team. 2017. Django Version 2.0. *Django Software Foundation*

32. Long M, de Souza SJ, Rosenberg C, Gilbert W. 1998. Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis. *Proc Natl Acad Sci USA* **95**:219-233 doi: 10.1073/pnas.95.1.219

33. Nguyen HD, Yoshihama M, Kenmochi N. 2006. Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evolutionary Biology* **6**:69 doi: 10.1186/1471-2148-6-69

34. Long M, Rosenberg C. 2000. Testing the “Proto-splice Sites” Model of Intron Origins: Evidence from Analysis of Intron Phase Correlations. *Mol Biol Evol* **17**:1789-1796 doi: 10.1093/oxfordjournals.molbev.a026279

35. Mount SM. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res* **10**:459-472 doi: 10.1093/nar/10.2.459

36. Krummel DAP, Oubridge C, Leung AKW, Li J, Nagai K. 2009. Crystal structure of human spliceosomal U1 snRNP at 5.5 □ resolution. *Nature* **458**:475-480 doi: 10.1038/nature07851

37. Kondo Y, Oubridge C, van Roon AMM, Nagai K. 2015. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5’ splice site recognition. *eLife* **4**:e04986 doi: 10.7554/eLife.04986

38. Turunen JJ, Will CL, Grote M, Lührmann R, Frilander MJ. 2008. The U11-48K Protein Contacts the 5’ Splice Site of U12-Type Introns and the U11-59K Protein. *Mol Cell Biol* **28**:3548-3560 doi: 10.1128/FMCB.01928-07

39. Lynch M and Richardson AO. 2002. The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12**:701-710 doi: 10.1016/S0959-437X(02)00360-X

40. Singh J and Padgett R. 2009. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**:1128-1133 doi: 10.1038/nsmb.1666

41. Nojima T, Rebelo K, Gomes T, Grosso AR, Proudfoot NJ, and Carmo-Fonseca M. (2018). RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. *Mol Cell* **72**:369–379 doi: 10.1016/j.molcel.2018.09.004

42. Basu MK, Rogozin IB, Koonin EV. 2008. Primordial spliceosomal introns were probably U12-type. *Trends Genet* **24**:525-528 doi: 10.1016/j.tig.2008.09.002

43. Burset M, Seledtsov IA, Solovyev VV. 2001. SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res* **29**:255-259 doi: 10.1093/nar/29.1.255

44. Thanaraj TA, Clark F. 2001. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res* **29**:2581-2593 doi: 10.1093/nar/29.12.2581

45. Farrer T. 2002. Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing. *Nucleic Acids Res* **30**:3360-3367 doi: 10.1093/nar/gkf465

46. Churbanov A, Winters-Hilt S, Koonin EV, Rogozin IB. 2008. Accumulation of GC donor splice signals in mammals. *Biol Direct* **3**:30 doi: 10.1186/1745-6150-3-30

47. Szafranski K, Schindler S, Taudien S, Hiller M, Huse K, Jahn N, Schreiber S, Backofen R, Platzer M. (2007). Violating the splicing rules: TG dinucleotides function as alternative 3' splice sites in U2-dependent introns. *Genome Biol* **8**:R154 doi: 10.1186/gb-2007-8-8-r154

48. Dietrich RC, Fuller JD, Padgett RA. 2005. A mutational analysis of U12-dependent splice site dinucleotides. *RNA* **11**:1430-1440 doi: 10.1261/rna.7206305

49. Vinogradov AE. 1999. Intron-Genome Size Relationship on a Large Evolutionary Scale. *J Mol Evol* **49**:376-384 10.1007/PL00006561

50. Deutsch M and Long M. 1999. Intron—exon structures of eukaryotic model organisms. *Nucleic Acids Res* **27**:3219-3228 doi: 10.1093/nar/27.15.3219

51. Patel AA, Steitz JA. 2003. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol* **4**:960-970 doi: 10.1038/nrm1259

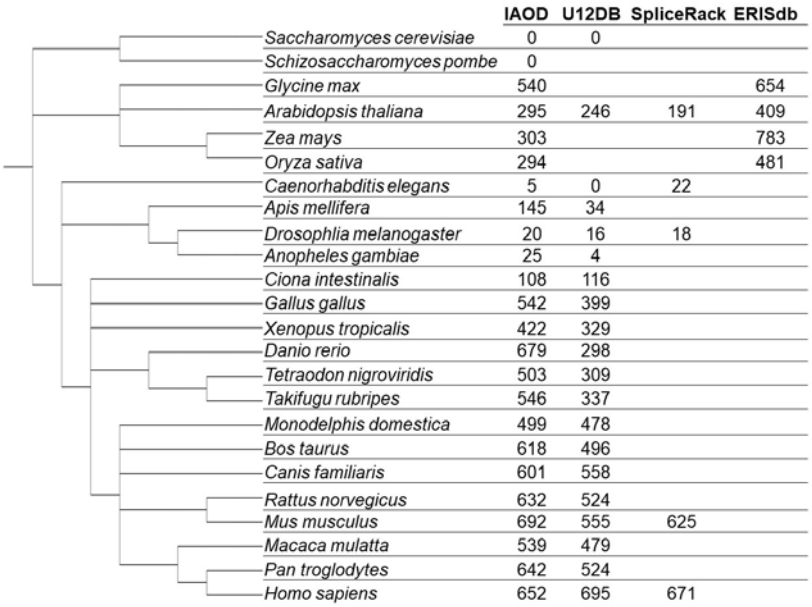
52. Lynch M and Conery JS. 2003. The Origins of Genome Complexity. *Science* **302**:1401-1404 doi: 10.1126/science.1089370

53. Fox-Walsh KL, Dou Y, Lam BJ, Hung S, Baldi PF, Hertel KJ. 2005. The architecture of pre-mRNAs affects mechanism of splice-site pairing. *Proc Natl Acad Sci USA* **102**:16176-16181 doi: 10.1073pnas.0508489102

54. DeConti L, Baralle M, Buratti E. 2012. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**:49-60. doi: 10.1002/wrna.1140

55. Federhen S. 2012. NCBI Taxonomy database. *Nucleic Acids Res* **40**:D136-D143 doi: 10.1093/nar/gkr1178

56. Letunic I and Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**:W242-W245 doi: 10.1093/nar/gkw290



	IAOD	U12DB	SpliceRack	ERISdb
Saccharomyces cerevisiae	0	0		
Schizosaccharomyces pombe	0			
Glycine max	540			654
Arabidopsis thaliana	295	246	191	409
Zea mays	303			783
Oryza sativa	294			481
Caenorhabditis elegans	5	0	22	
Apis mellifera	145	34		
Drosophlia melanogaster	20	16	18	
Anopheles gambiae	25	4		
Ciona intestinalis	108	116		
Gallus gallus	542	399		
Xenopus tropicalis	422	329		
Danio rerio	679	298		
Tetraodon nigroviridis	503	309		
Takifugu rubripes	546	337		
Monodelphis domestica	499	478		
Bos taurus	618	496		
Canis familiaris	601	558		
Rattus norvegicus	632	524		
Mus musculus	692	555	625	
Macaca mulatta	539	479		
Pan troglodytes	642	524		
Homo sapiens	652	695	671	

Figure 1. Phylogenetic distribution of U12-type introns in all species annotated by the Intron Annotation and Orthology Database (IAOD), U12DB (21), SpliceRack (16) and ERISdb (23). 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 (55) and Integrative Tree of Life (56) were used to create the phylogenetic tree.

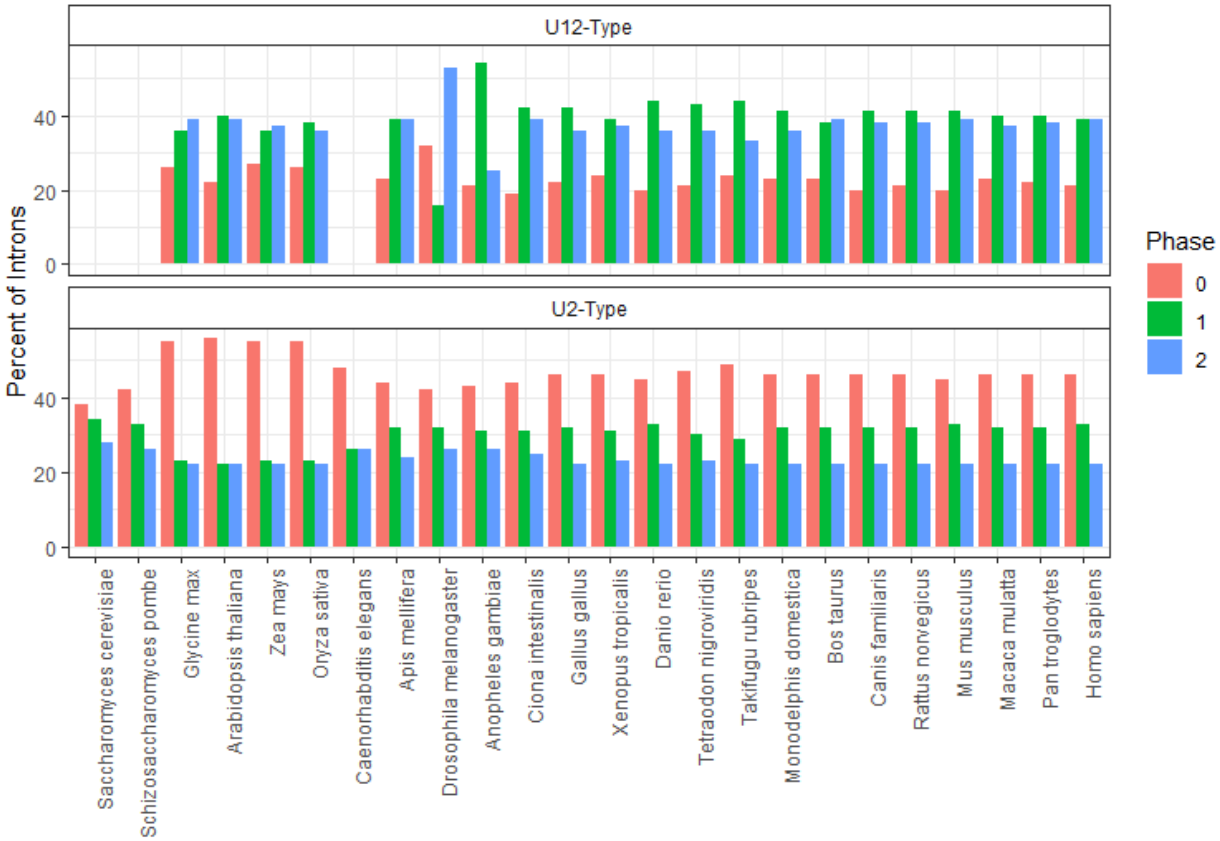


Figure 2. Phase distribution of introns within each class in all genomes annotated in the IAOD. Organisms are grouped by phylogeny. The bias against phase 0 U12-type introns is statistically significant in all organisms but *G. max*, *O. sativa*, *X. tropicalis*, and *Z. mays* (chi-squared; $p < 0.10$). The bias towards phase 0 U2-type introns is statistically significant in all organisms but *A. mellifera*, *D. melanogaster*, *S. cerevisiae*, and *S. pombe* (chi-squared; $p < 0.10$).

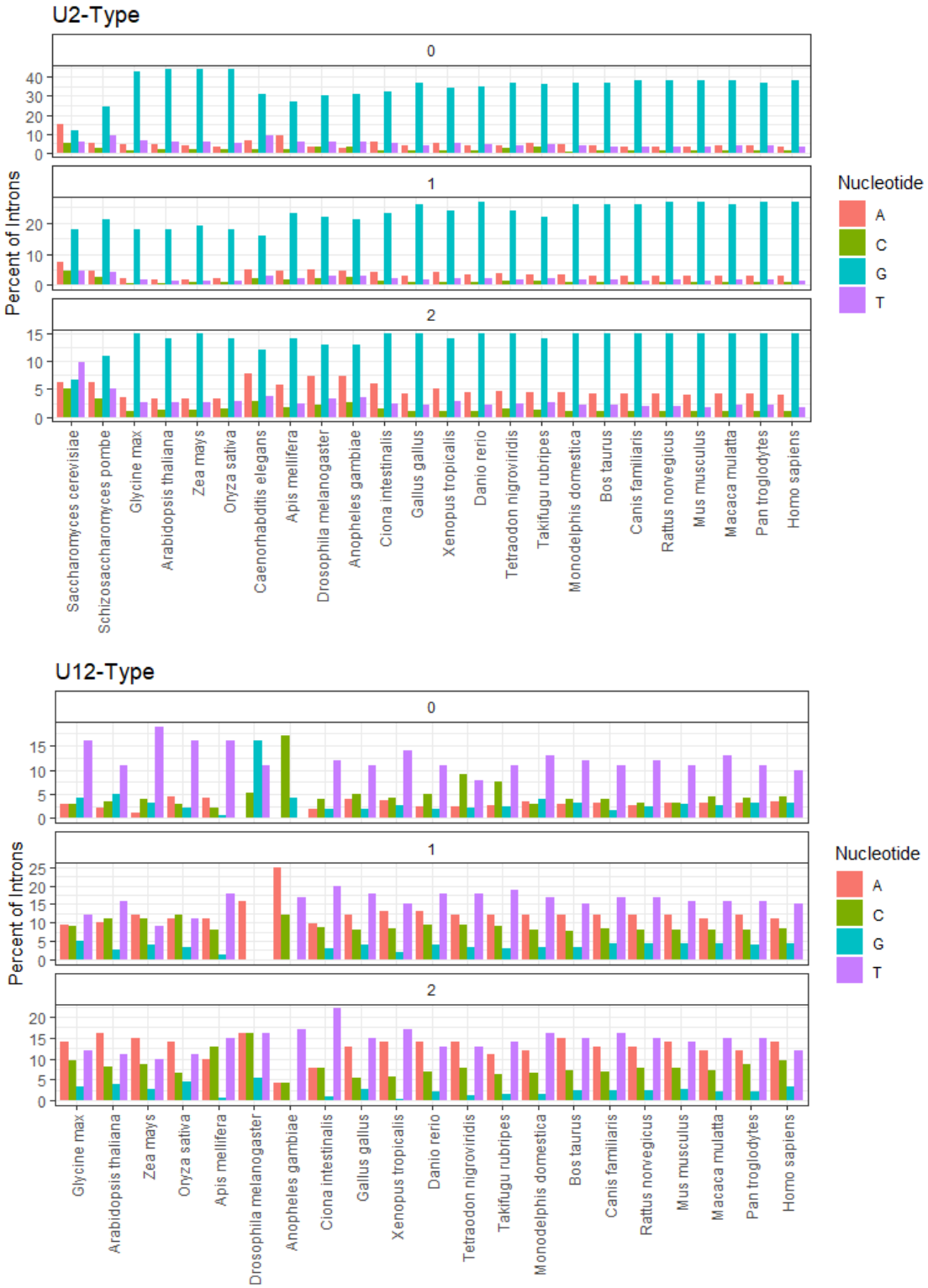


Figure 3. Percentages of introns with the specified nucleotide immediately upstream of the 5' splice site in each phase of both classes of introns. Organisms are grouped by phylogeny.

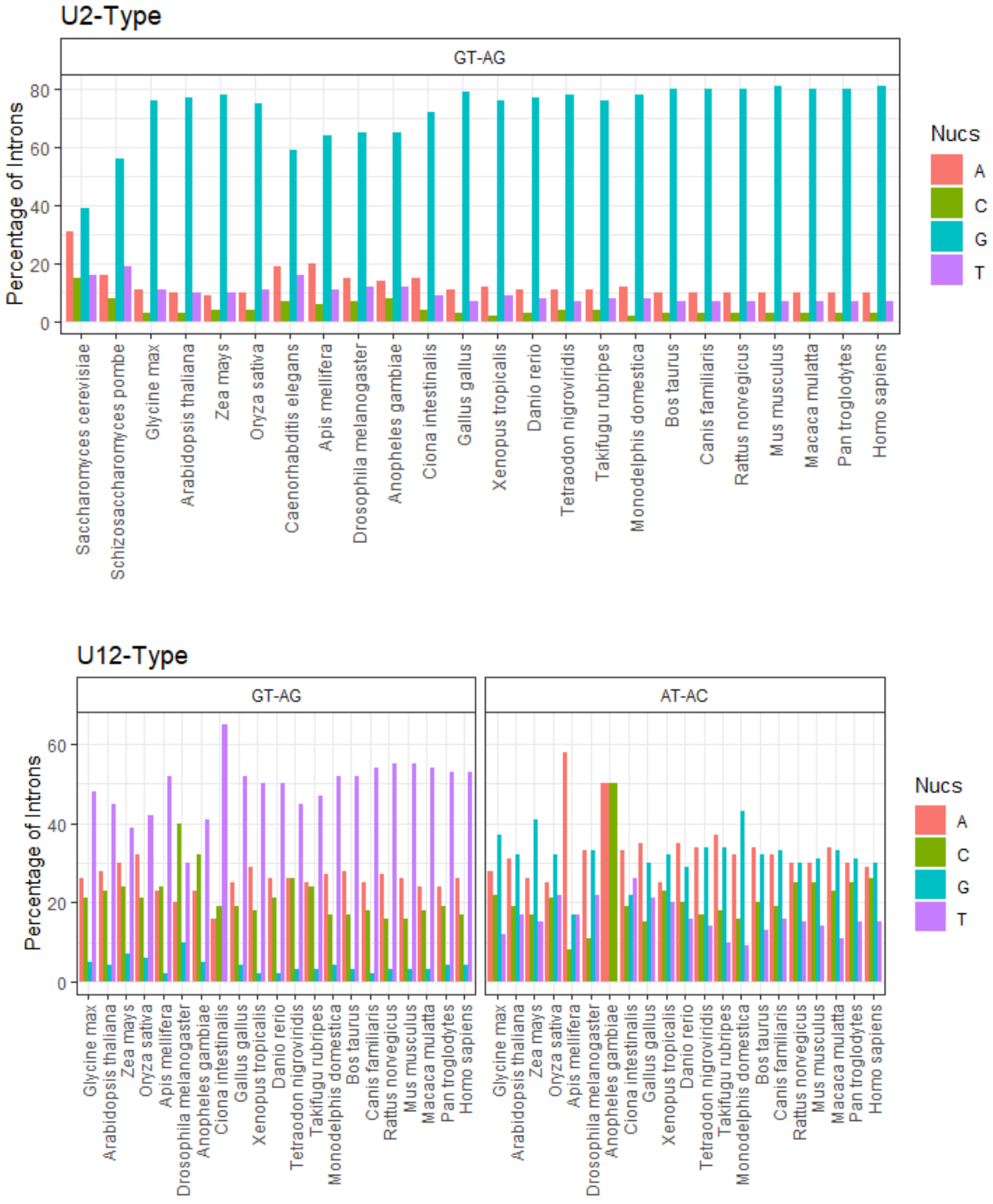


Figure 4. Nucleotide biases at the -1 position relative to the 5' splice site for all organisms (excluding those lacking U12-type introns) annotated in the IAOD, grouped by terminal dinucleotides and intron class.

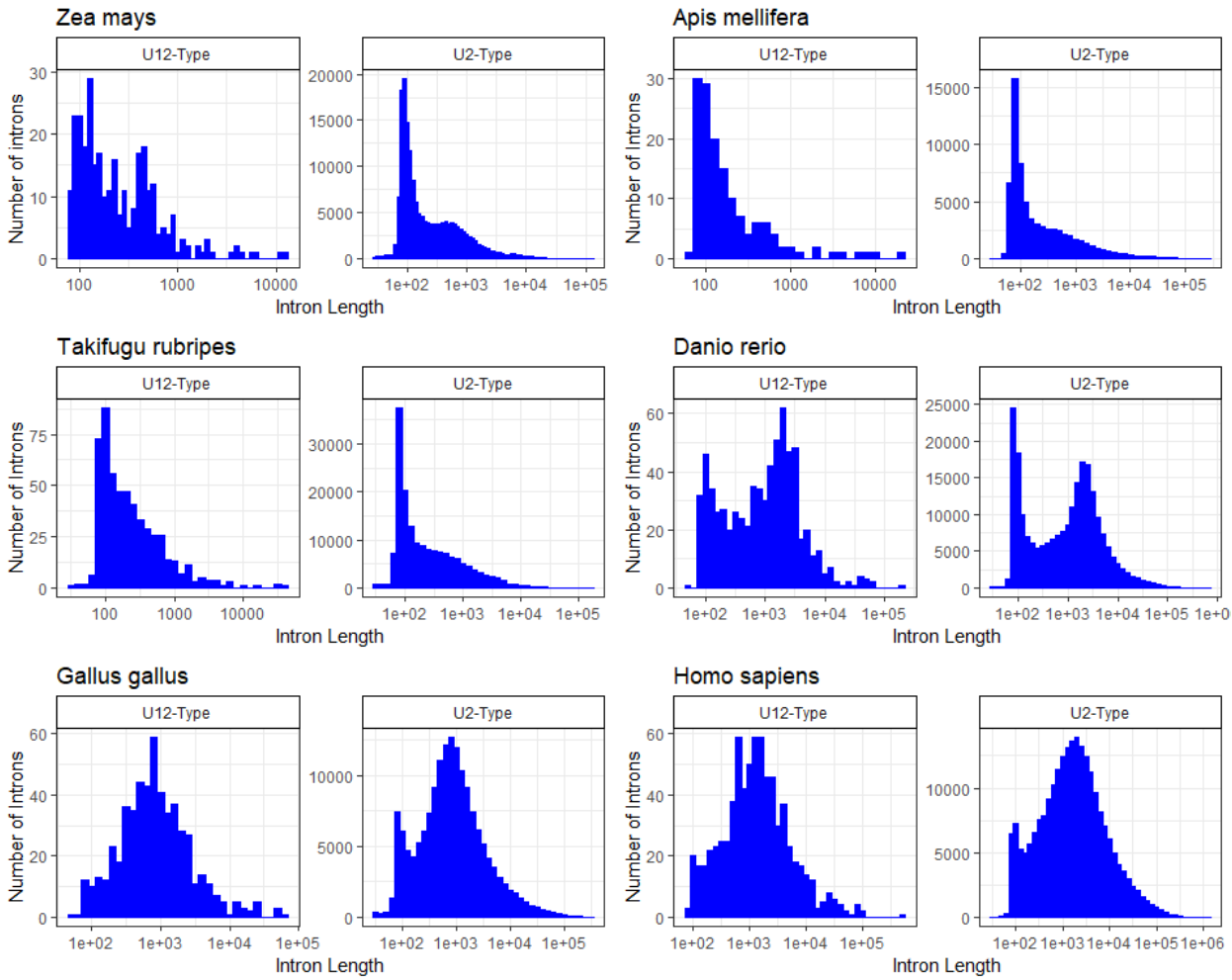


Figure 5. Distributions of intron lengths in both classes of intron in six of the genomes annotated in the IAOD. The x-axis of each plot is a log scale.

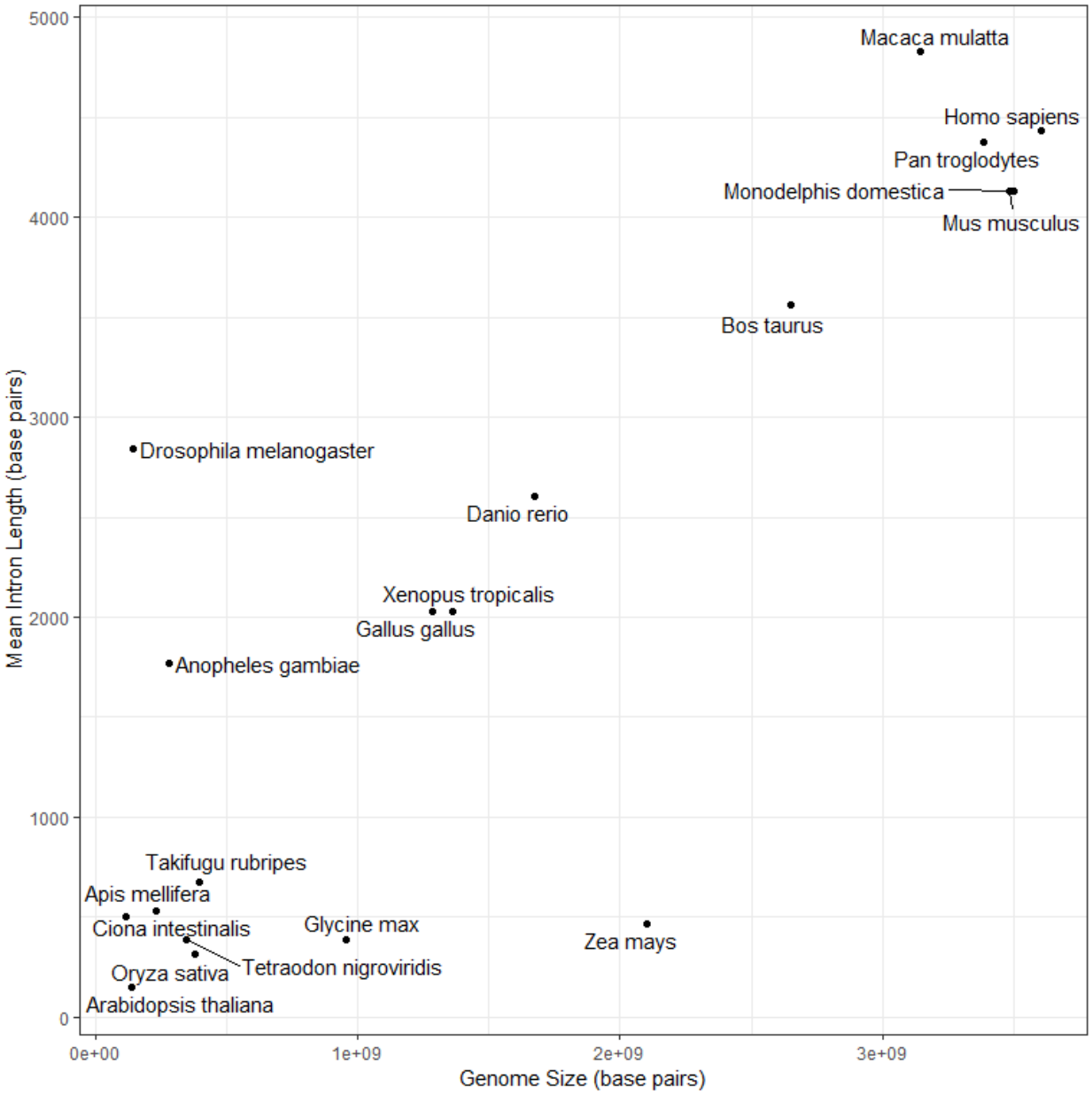


Figure 6. Relationship of genome size and mean U12-type intron length in genomes annotated in the IAOD. *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans* are not shown in this figure as they lack U12-type introns.

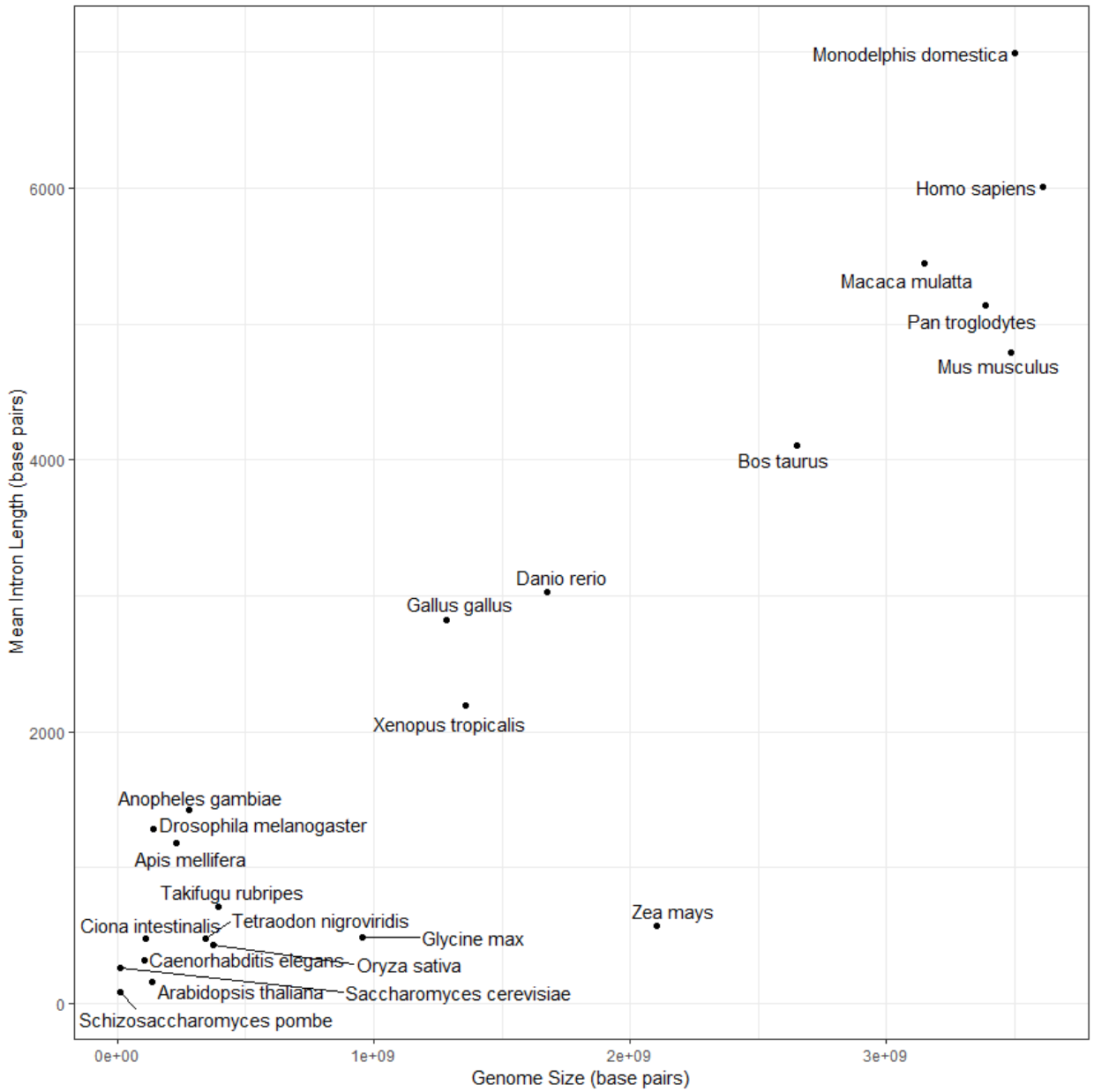


Figure 7. Relationship of genome size and mean U2-type intron length in genomes annotated in the IAOD.

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

Table 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.

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

Table 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 1.