Grau-Bové et al. Genome Biology (2018) 19:135
https://doi.org/10.1186/s13059-018-1499-9

Genome Biology

RESEARCH Open Access

Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture

Xavier Grau-Bové[1,2], Iñaki Ruiz-Trillo[1,2,3*] and Manuel Irimia[4,5*]

Abstract

Background: Alternative splicing, particularly through intron retention and exon skipping, is a major layer of pre-translational regulation in eukaryotes. While intron retention is believed to be the most prevalent mode across non-animal eukaryotes, animals have unusually high rates of exon skipping. However, when and how this high prevalence of exon skipping evolved is unknown. Since exon skipping can greatly expand proteomes, answering these questions sheds light on the evolution of higher organismal complexity in metazoans.

Results: We used RNA-seq data to quantify exon skipping and intron retention frequencies across 65 eukaryotic species, with particular focus on early branching animals and unicellular holozoans. We found that only bilaterians have significantly increased their exon skipping frequencies compared to all other eukaryotic groups. Unlike in other eukaryotes, however, exon skipping in nearly all animals, including non-bilaterians, is strongly enriched for frame-preserving sequences, suggesting that exon skipping involvement in proteome expansion predated the increase in frequency. We also identified architectural features consistently associated with higher exon skipping rates within all studied eukaryotic genomes. Remarkably, these architectures became more prevalent during animal evolution, indicating co-evolution between genome architectures and exon skipping frequencies.

Conclusion: We suggest that the increase of exon skipping rates in animals followed a two-step process. First, exon skipping in early animals became enriched for frame-preserving events. Second, bilaterian ancestors dramatically increased their exon skipping frequencies, likely driven by the interplay between a shift in their genome architectures towards more exon definition and recruitment of frame-preserving exon skipping events to functionally diversify their cell-specific proteomes.

Keywords: Alternative splicing, Exon skipping, Intron retention, Ancestral reconstruction, Gene architecture, Evolution of transcriptome regulation

Background

Alternative splicing (AS) is a pre-translational process that allows the creation of multiple messenger RNA (mRNA) transcripts from a single gene by differentially selecting splice sites in multi-exonic sequences [1]. This phenomenon can contribute to the regulation of gene expression [2–7], or the creation of multiple protein isoforms per gene, increasing the proteomic repertoire of eukaryotic genomes [8, 9] and potentially leading to key evolutionary innovations [10–12].

The main forms of AS among eukaryotes are the exclusion of specific exons and the retention of introns in the final transcripts [1, 13], referred to as exon skipping (ES) and intron retention (IR), respectively. These sources of transcript variation are widespread in eukaryotes, but initial studies revealed that the prevalence of each AS mode varied across lineages: animals show higher rates of ES than other eukaryotes, whereas IR is frequent across all eukaryotic groups, including animals, fungi, plants, and various protist lineages [14–17]. This contrast led to the proposition that ES-rich AS profiles were a major contributor to the increased phenotypic

* Correspondence: inaki.ruiz@ibe.upf-csic.es; mirimia@gmail.com
[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain
[4]Centre de Regulació Genòmica, Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain
Full list of author information is available at the end of the article

complexity of animals, since richer proteomes can provide an expanded tool-kit needed to sustain multicellularity [8, 18]. Consistently, and although the extent to which ES transcripts are translated and functional is still under debate [19, 20], many ES-derived isoforms have been found to be physiologically relevant in animals (reviewed in [8, 21]), for example, by tuning protein–protein interaction networks [22–24]. In contrast, IR events have been linked to down-regulation of gene expression via the nonsense-mediated decay (NMD) pathway [4–6, 25], nuclear retention [7] or intron detention [3] in a wide variety of eukaryotes.

The evolutionary origin of AS can be tracked to the last eukaryotic common ancestor (LECA), which already had an intron-dense genome [26, 27] with heterogeneous splice sites [28–30] and all the essential splicing machinery (the spliceosome, a complex of small nuclear RNAs and dozens of assisting protein factors) [27, 31, 32]. These observations have allowed inferring that the earliest eukaryotes already exhibited splicing-rich transcriptomes yielding multiple mRNA variants per gene, mostly by IR [18, 33, 34]. However, it remains unclear when and how animal transcriptomes shifted towards higher frequencies of ES and recruited this mode of AS as a mechanism to expand their proteomes. First, the sampling of early-branching species—poriferans, cnidarians, placozoans, and ctenophores—is scarce. Second, no comprehensive comparative study using high-throughput RNA sequencing (RNA-seq) data has been performed to date. Third, relative increases in ES frequency have also been identified in other phylogenetically scattered eukaryotes—e.g. in plants, *Volvox carteri*, or the chlorarachniophyte *Bigelowiella natans* [10, 35–38].

Here, we address these questions by analyzing RNA-seq-derived AS profiles for 65 eukaryotic species. Using this comprehensive dataset of joint transcriptomic and genomic data, we track the frequency of both main modes of AS (IR and ES) across all major eukaryotic lineages (Fig. 1a, Additional file 1: Figure S1), uncovering the phylogenetic patterns behind AS evolution. Specifically, we investigate the transition towards high ES frequencies in multicellular lineages (animals and plants) by comparing their AS profiles and genome architectures with their closest unicellular relatives. We find that the frequency of ES rose mainly in bilaterians, with only mild increases in non-bilaterians. However, we show that recruitment of ES for proteome expansion predated the bilaterian increase in ES and occurred early in metazoan evolution. Furthermore, we uncover a set of sequence and architectural features that influence the frequency of ES and IR in transcripts across eukaryotes, suggesting the existence of a soft pan-eukaryotic *cis*-regulatory code for AS determination. Using this code and reconstruction of ancestral intron–exon architectures we evaluated the step-wise increase of ES along animal evolution.

Results

ES frequency increased largely in bilaterian ancestors

We quantified the frequencies of ES and IR at the single exon and intron levels for 65 eukaryotic species (Fig. 1a), including a wide range of intron–exon architectures (Fig. 1b, c; Additional file 1: Figures S1 and S2). For this purpose, we compiled a large dataset of available RNA-seq data (Additional file 1: Figure S1) and performed de novo RNA-seq for three phylogenetically key species: the placozoan *Trichoplax adhaerens*, the holozoan *Sphaeroforma*

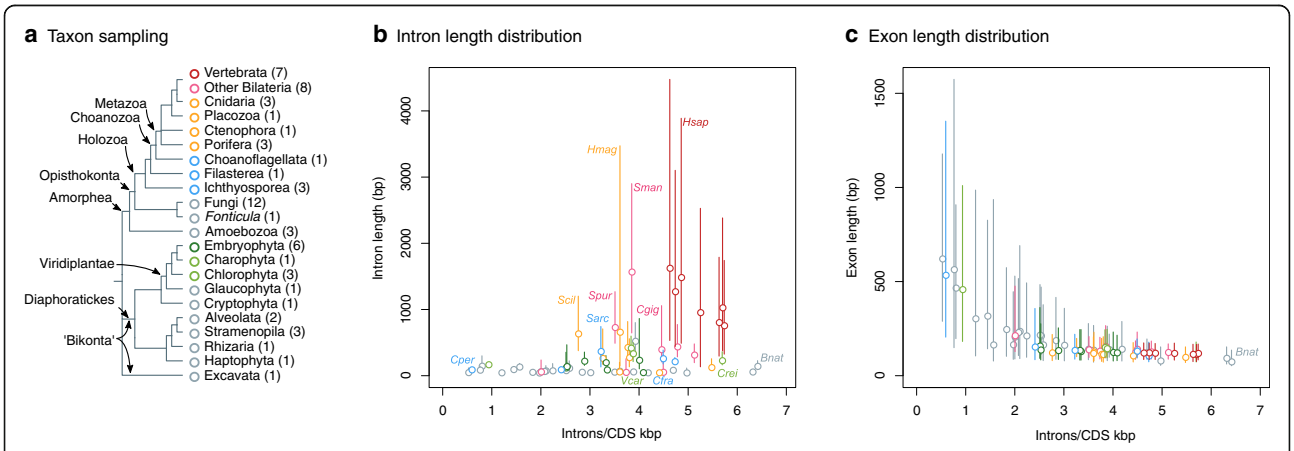

Fig. 1 Intron/exon structure across eukaryotes. **a** Summary of the taxon sampling used in this study and the phylogenetic relationships between species (Additional file 1: Figure S1). **b, c** Comparison of intron density (introns/CDS kbp) and intron (**b**) and exon (**c**) length distribution (in bp) in 65 eukaryotic genomes. *Dots* represent median intron/exon lengths and *vertical lines* delimit the first and third quartiles (Additional file 1: Figure S2). Color-coded according to taxonomy

arctica, and the intron-rich excavate *Naegleria gruberi*. For the analysis of ES events (Additional file 1: Figure S3), we compiled a dataset of exon triplets from 2.93×10^6 internal exons from 5.08×10^5 multi-exonic genes with transcriptomic support. Each internal exon was classified as ES-negative (0–10% exon skipping rate [r_{ES}] and sufficient read coverage; 89.29% of the dataset), ES-positive (r_{ES} = 10–90% and sufficient read coverage; 0.74% of the data), or undetermined (all other cases; Additional file 1: Figure S4). For the analysis of IR (Additional file 1: Figure S3), an analogous dataset was built with 1.98×10^6 introns from 3.88×10^5 multi-exonic genes, which were classified as IR-negative (0–10% inclusion rate [r_{IR}] and sufficient read coverage; 71.09% of the data), IR-positive (r_{IR} = 10–90%, and sufficient read coverage; 5.27% of the data), or undetermined (Additional file 1: Figure S4).

Next, we examined the frequency of each AS mode at the species level by averaging exon- and intron-specific ES and IR rates across 100 subsets of exons or introns with normalized RNA-seq coverage ($F_{ES,sp}$ and $F_{IR,sp}$, respectively; see "Methods"; Additional file 1: Figures S3 and S5). This analysis produced three major results. First, we found evidence of ES and IR in transcriptomes from all studied eukaryotic lineages—namely, animals, fungi, opisthokont protists, amoebozoans, Viridiplantae, the cryptophyte *Guillardia theta*, SAR, the haptophyte *Emiliania huxleyi*, and the excavate *Naegleria gruberi* (Fig. 2, Additional file 1: Figures S4 and S6). Second, IR frequencies exceeded ES in all but one species (Fig. 2b). This result is in line with previous reports highlighting the dominance of IR-based AS across eukaryotes, but challenges initial views of animal transcriptomes as being dominated by ES. A possible explanation for this disagreement is the association between high IR rates and low transcript expression levels (see below), which hinders the detection of retained introns (particularly in studies based on EST data [14, 15, 39]).

Third, we found a clear phylogenetic pattern behind ES frequencies (Fig. 2a): animals, particularly bilaterians, had the highest frequencies, followed by non-bilaterians, plants, and a handful of other scattered eukaryotes. The consistency of our quantifications at the species level was assessed with replicate transcriptomic datasets for six selected ES-rich taxa obtained from independent studies (Additional file 1: Figure S7). Species-level $F_{ES,sp}$ values were highly consistent between pairs of multi-organ transcriptomic datasets of adult human, frog, and tale cress and a comparison of developmental time series with various growth conditions for fruit fly ($p > 0.01$, Wilcoxon rank-sum test). Mild significant differences were observed between two multi-organ sets of mouse ($p = 0.0037$, Wilcoxon rank-sum test) and a comparison of developmental series of the sea anemone *Nematostella vectensis* ($p = 1.67e-08$, Wilcoxon rank-sum test), but in both cases the distributions were within their respective taxonomic ranges. Similarly, although pooled samples generally had higher ES levels than the individual tissues, the latter also fell within the taxonomic range (Additional file 1: Figure S8). This indicates that our approach for ES quantification yields robust results independently of the experimental approaches used in different RNA-seq experiments.

To further investigate when the shift towards ES-rich transcriptomes occurred in animal evolution, we next compared the aggregated ES frequencies of vertebrates, non-vertebrate bilaterians, non-bilaterians (cnidarians, poriferans, the ctenophore *Mnemiopsis leidyi*, and the placozoan *Trichoplax adhaerens*), and their closest unicellular relatives in the Holozoa (choanoflagellates, *Capsaspora owczarzaki*, and ichthyosporeans) (Fig. 2c–f; complete report of statistical comparisons in Additional file 1: Figure S9).

Taken together, bilaterian animals have significantly higher $F_{ES,sp}$ than non-bilaterians, unicellular holozoans, and other opisthokonts (Fig. 2c; $p < 1.0e-76$ in all comparisons, one-sided Kolmogorov-Smirnov test [oKS]). Furthermore, vertebrates exhibited an enrichment compared to other bilaterians ($p = 1.0e-03$). The grouping of non-bilaterians did not show significantly higher $F_{ES,sp}$ than unicellular holozoans (Fig. 2c), although different patterns were observed when comparing against individual groups (Fig. 2d): $F_{ES,sp}$ values were higher in cnidarians and the ctenophore *M. leidyi* and lower in poriferans and *T. adhaerens*. On the other hand, the ichthyosporean *Sphaeroforma arctica* had the highest $F_{ES,sp}$ among unicellular holozoans (Fig. 2e; $p < 1e-09$, oKS), which indicates a clear lineage-specific departure from the low incidence of ES in other unicellular holozoans [40, 41]. Importantly, these patterns were robust to different levels of read depth downsampling across species (Additional file 1: Figure S10). Therefore, these data show that (i) bilaterians, and vertebrates in particular, have a consistently higher ES frequency than their close relatives and other eukaryotes, and (ii) some non-bilaterian animals and unicellular holozoans have experienced relative increases in ES frequency as well.

In parallel, multicellular land plants also exhibited higher ES rates than other 'bikonts' (Dipahoratickes and *N. gruberi*), including their colonial and unicellular relatives in Chlorophyta and Charophyta (Fig. 2f; all $p < 1.0e-06$, oKS). However, the colonial chlorophyte *Volvox carteri* was a notable exception, with higher $F_{ES,sp}$ than other algae (including its close unicellular relative *Chlamydomonas reinhardtii* [38]) and most land plants ($p < 1e-13$ for all comparisons, oKS).

Finally, our analysis of $F_{ES,sp}$ levels in the chlorarachniophyte *Bigelowiella natans* showed contradictory results.

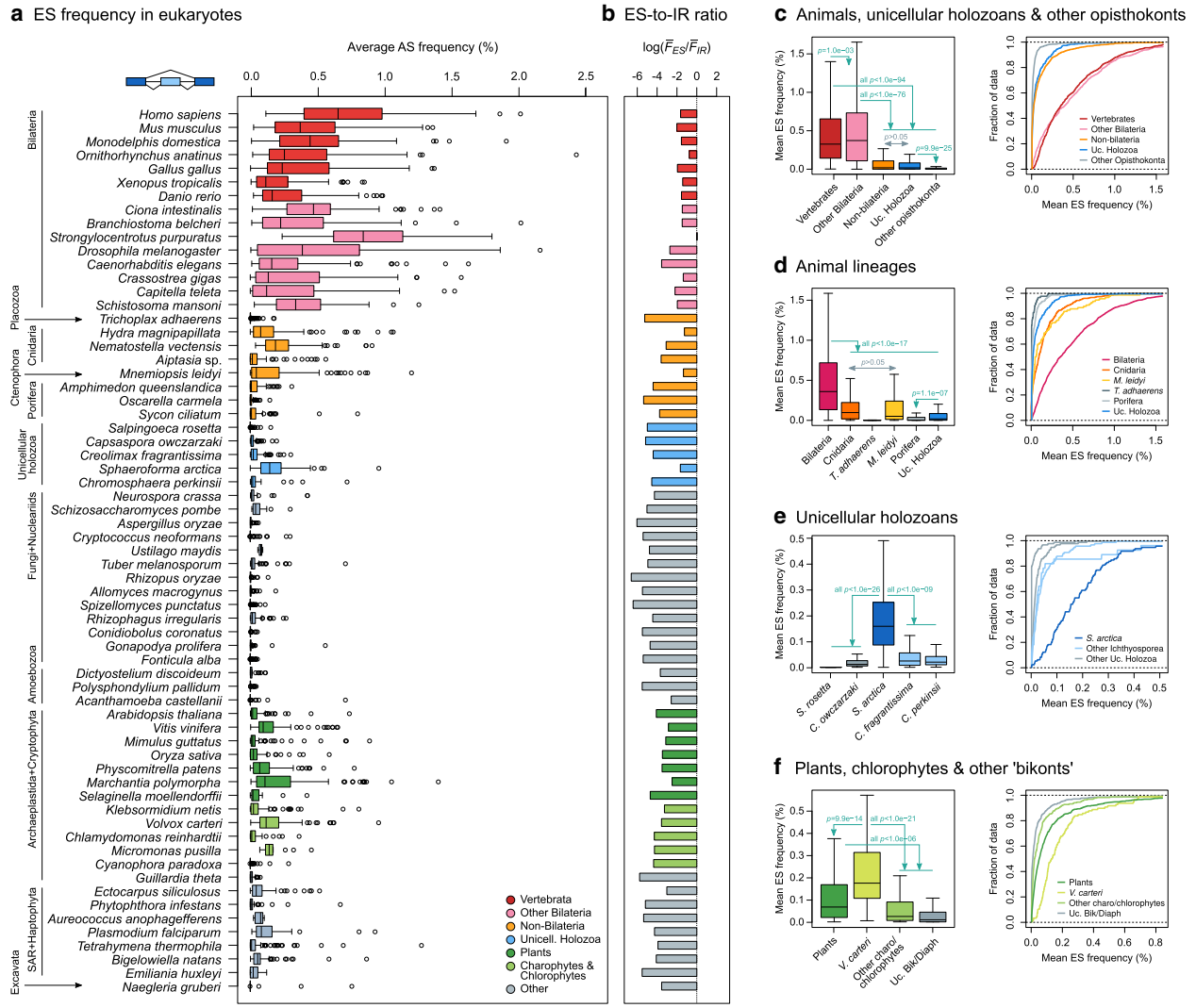

Fig. 2 Frequency of AS in eukaryotes. **a** Distribution of ES frequencies across 65 eukaryotic transcriptomes. Each box plot represents the distribution of ES frequencies ($F_{ES,sp}$) in a given species (as percentage), drawn from mapping up to 10,000 RNA-seq reads to junctions from 100 exon triplets, in up to 100 independent subsets (see "Methods" and Additional file 1: Figure S3). **b** Log-ratio of mean $F_{ES,sp}$ (for ES) over mean $F_{IR,sp}$ (for IR) in each species. $F_{IR,sp}$ distributions in Additional file 1: Figure S6. **c–f** Comparisons of ES frequencies among phylogenetically related eukaryotic groups with a focus on animals, holozoans, and plants. Group-level ES distributions obtained by pooling $F_{ES,sp}$ of multiple species. For each set of comparisons, we report the pooled distributions' box plots (*left*) and empirical cumulative distributions (*right*). Selected *p* values are reported, corresponding to pairwise one-sided Kolmogorov-Smirnov two-sample tests (full report in Additional file 1: Figure S9), and should be interpreted as follows: *green arrows* indicate that the group of origin has significantly higher $F_{ES,sp}$ than others; *grey bidirectional arrows* indicate that no significant differences are observed in any direction at the given threshold

The analysis of two recent transcriptomic datasets [42, 43] showed significantly lower ES frequencies than those reported in the original genome paper [36] (Additional file 1: Figure S7; $p < 1e-16$, Wilcoxon rank-sum test). The reasons behind these differences remain elusive. One possible explanation is that they were caused by differences in environmental conditions, such as abiotic stress, which have been shown to lead to increased levels of both spurious ES and IR in other species [44–46].

Enrichment in frame-preserving ES is common to all animal groups

Next, we examined the potential global impact of AS events on each species' proteome by assessing their effect on the coding capacity of the resulting transcripts. Following previous studies [47], we divided exons or introns between those that had lengths divisible by three (henceforth '3n') and those that did not, under the assumption that 3n sequences would not normally disrupt the open reading frame (ORF) integrity when

alternatively spliced, whereas non-3n sequence lengths would cause frame-shifts, usually resulting in unproductive transcripts. Indeed, maintenance of the ORF integrity is strongly associated with functional and conserved ES events in animals [8, 47, 48].

In the case of IR, virtually no significant biases towards or against 3n divisibility were observed (Fig. 3a and Additional file 1: Figure S11). In contrast, we found that alternatively spliced exons of most animals were significantly enriched in 3n divisible lengths (Fig. 3b, Additional file 1: Figure S12; $p < 0.01$, Fisher's exact test). This includes all vertebrates, most bilaterians, the ctenophore *M. leidyi*, and the cnidarians *Aiptasia* sp. and *H. magnipapillata* ($p < 0.05$). For example, 37.8% of *M. leidyi*'s ES-negative exons were 3n divisible, but this percentage increased up to 76.9% in ES-positive exons. Overall, the fraction of 3n divisible exons was higher in ES-positive exons than in ES-negative ones for 20 out of 23 animal species, and at least half of the ES-positive exons were 3n divisible in 15 out of 23 animals (Fig. 3c), suggesting that exon 3n enrichment is largely an animal feature. This enrichment was also observed when individual, rather than pooled, tissues were analyzed (Additional file 1: Figure S13A) and at different levels of sequencing depth (Additional file 1: Figure S14). Moreover, this pattern was even stronger when we restricted the analysis to highly alternative exons (r_{ES} = 30–70%; Fig. 3d). Most animals showed higher 3n enrichment in this subset, including robust increases in some non-bilaterians that did not exhibit significant 3n biases in the whole transcriptome. For example, the 3n fraction of ES-negative exons of the sponge *Amphimedon queenslandica* is 43.2%, but increases to 73.1% in highly alternative ES-positive ones

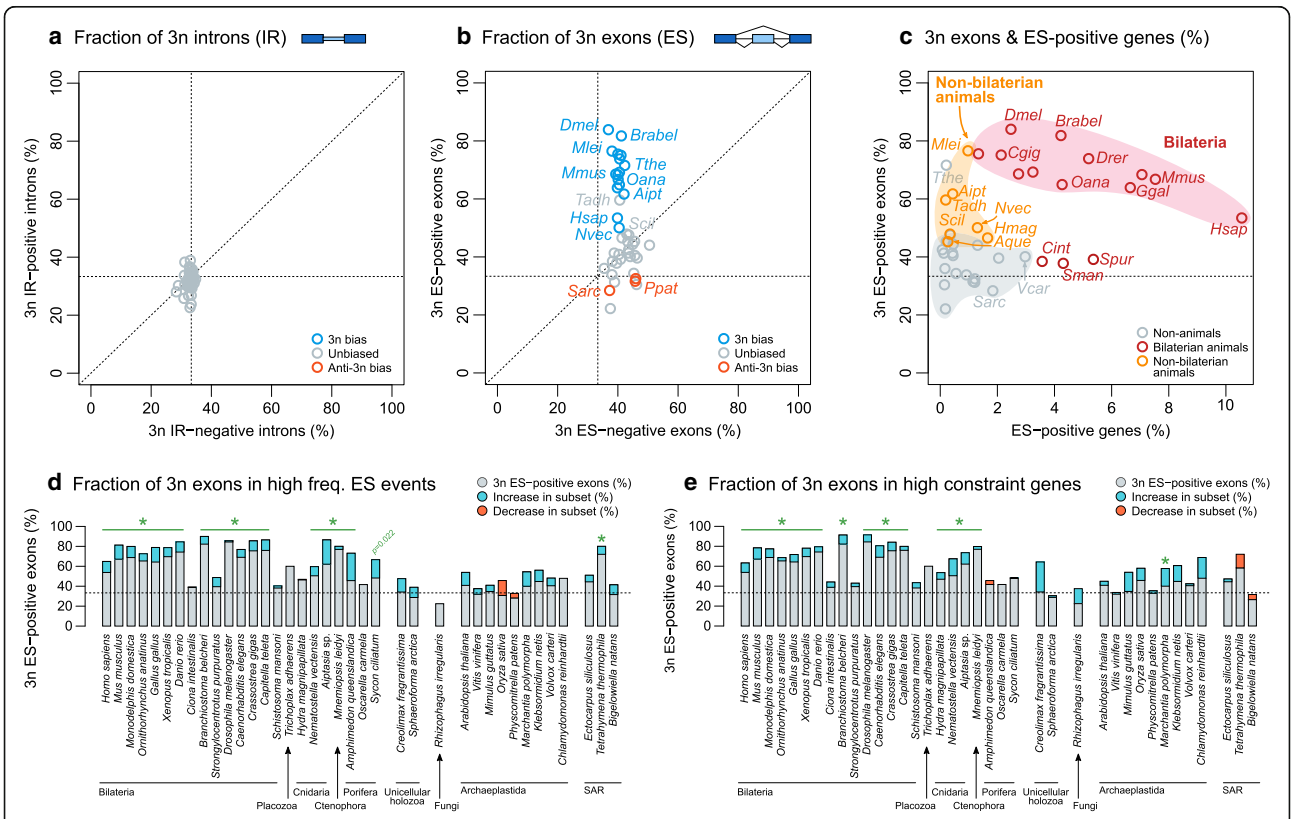

Fig. 3 Fraction of 3n divisible exon and intron lengths. **a** Percentage of introns with 3n divisible lengths that were classified as IR-negative (*x*-axis) and IR-positive (*y*-axis) in each species with more than ten IR-positive introns. **b** Percentage of exons with 3n divisible lengths that were classified as ES-negative (*x*-axis) and ES-positive (*y*-axis) in each species with more than ten ES-positive exons. *Dots* are color-coded according to enrichments in either direction (3n bias in *blue*, anti-3n bias in *red*) as per Fisher's exact test ($p < 0.01$; Additional file 1: Figure S9). **c** Percentage of ES-positive exons that are 3n (*y*-axis) with respect to the percentage of genes with ES-positive exons (*x*-axis), per species. Shading colors grouping the most representative species for certain eukaryotic groups (bilaterians, non-bilaterians, and other eukaryotes) are provided for reference. **d** Fraction of 3n exons in ES events (r_{ES} = 10–90%, *grey bar segment*) compared to a subset of high-frequency events (r_{ES} = 30–70%), highlighting increases (*blue*) or decreases (*red*) of the 3n fraction. In panels **d** and **e**, asterisks indicate significant 3n biases in ES-positive compared to ES-negative exons, as per Fisher's exact test ($p < 0.01$; Additional file 1: Figure S9). In all panels, only species with more than ten exons in all categories are included. **e** Fraction of 3n exons in ES events (*grey bar segment*) compared to a subset of genes with above-median expression (cRPKM) and gene length ('high constraint'), highlighting increases (*blue*) or decreases (*red*) of the 3n fraction

(p = 2.55e − 3, Fisher's exact test), and a mild increase also occurs in another sponge, *Sycon ciliatum* (p = 0.022; Fisher's exact test; Additional file 1: Figure S12).

On the other hand, lack of positive 3n biases was observed in nearly all other eukaryotes, including unicellular holozoans and plants, reaching negative enrichments (*S. arctica* and the plants *Vitis vinifera* and *Physcomitrella patens*) and/or reducing this enrichment in the highly alternative subset in some species (*Oryza sativa* and *P. patens*) (Fig. 3b–d). Only the ciliate *Tetrahymena thermophila* exhibited a 3n bias akin to that of animals (p = 2.18e − 5, Fisher's exact test). The lack of 3n bias in ES has been previously reported (e.g., in *Creolimax fragrantissima* [41] and *B. natans* [36]), and such ES-caused ORF disruptions were proposed to be a consequence of noisy splicing and to produce non-functional isoforms. Consistent with this idea, we found a general robust increase in the fraction of 3n exons within long genes with high expression (Fig. 3e)—a subset of 'high-constraint' genes that are expected to be less prone to splicing errors due to the higher energetic cost of their production [49]. We observed a 3n exon enrichment increase in high constraint genes in animals (except in poriferans and *T. adhaerens*), plants, chlorophytes, and the multicellular phaeophyte *Ectocarpus siliculosus*. However, even if high-constraint genes in non-animals showed higher fractions of 3n exons, these were not significantly biased towards being ES-positive (Fishers' exact test, significant if $p < 0.01$; Additional file 1: Figure S12).

Overall, the 3n bias in ES events recorded in animals suggests that the lengths of alternatively spliced exons are under selective pressure to avoid ORF disruptions, possibly due to an enrichment in functional protein isoform-producing ES events. In other eukaryotes, only high-constraint genes, in which non-3n ES would be more detrimental, showed enrichments in 3n exon lengths.

Sequence and architectural intron–exon traits influence ES and IR frequencies

To investigate how increases in ES frequencies may have taken place during animal evolution, we next studied intra-species associations between ES and IR frequencies and different genomic architectural features. Previous studies have linked the level of ES and IR within a genome to differences in traits such as the length of exons and introns, intron density, sequence composition, splicing site homogeneity, or other *cis* signals [16, 50–53]. Therefore, these associations suggest that the evolutionary processes shaping genome architecture could contribute to the variations in AS frequency across species, including the increase in ES during animal evolution.

To address this possibility, we analyzed the intron/exon structure and sequence composition of the genomic regions associated with the AS events, and compared the gene architecture of ES- and IR-positive and negative exons and introns across our set of 65 eukaryotic species. In particular, we investigated the effect of global length of genes and transcripts; the length of the alternatively spliced exons and its flanking introns (for ES) or vice versa (for IR); intronic and exonic GC content and the differential in GC content between introns and exons; the strength of the 5′ and 3′ splice site definition; the intron density (introns per gene and base pairs of introns per base pairs of coding sequence [CDS]); the relative position of the AS event along the gene (from the start codon); and the mean transcript expression level (using cRPKMs from pooled RNA-seq experiments). See "Methods" for precise definitions of each trait. Our analysis identified consistent relationships between ES, IR, and gene architecture across the eukaryotic tree of life, maintained across genomes from different lineages and robust to tissue pooling and sequencing depth (Figs. 4 and 5 and Additional file 1: Figures S13 and S15–S18).

As expected [54, 55], we identified a widespread relationship between positive cases of ES and IR and weak 5′ and 3′ splice sites (Figs. 4a and 5a; $p < 0.01$, one-sided Kolmogorov-Smirnov test with complementary hypotheses). In the case of ES, this association is significant and consistent for all species with a sufficient number of ES-positive exons, including animals, unicellular holozoans, plants, chlorophytes, the phaeophyte *E. siliculosus*, and *B. natans*. Thus, in most eukaryotes, heterogeneity in the splice sites influences ES frequencies at the intra-species level: exons with more poorly defined intron–exon boundaries are more subject to ES than those closer to the species consensus.

Another consistent association across eukaryotes was found between ES and shorter exon lengths (Fig. 4a), as previously reported for animals [47, 56]. Moreover, ES-positive exons are widely associated with longer flanking introns, both upstream and downstream. Exons with these features are expected to be spliced through 'exon definition', a model that proposes that the recognition of the 5′ and 3′ splice sites occurs across the exonic sequence (as intron ends are more distant); thus, interrupting this process is more likely to result in ES than in IR [50, 53]. The general positive relationship between ES and higher intron-to-exon length ratios also fits this principle, and shows a good correlation with the overall gene architecture at the species level, particularly in bilaterians (Fig. 4c), suggesting that their distinct architectures may have driven their increases in ES frequencies.

Interestingly, we observed similar patterns in the ichthyosporean *S. arctica* (with higher ES rates than other holozoans; Fig. 2e) and the chlorophyte *V. carteri* (also with higher ES rates than other chlorophytes; Fig. 2f). Their ES-positive exons had more heterogeneous splice sites than

Fig. 4 Relationship between gene structure and sequence composition and ES. **a** Heatmap representing the distance between distributions of gene architecture values for ES-positive and ES-negative exons, measured with the *D* statistic of the one-sided Kolmogorov-Smirnov two-sample test (significant if $p < 0.01$; otherwise grey). *D* values are recorded as positive/negative (*blue/red*) according to two one-sided Kolmogorov-Smirnov tests with complementary alternative hypotheses: positive/negative *D* (*blue/red*) reflects a positive/negative relationship between ES and the given feature. **b** Heatmap representing the correlation between intron length and GC content across 65 eukaryotes. Positive correlations (long, GC-rich introns, Spearman's *rho* > 0 and $p < 0.01$) are shown in *blue*, negative correlations in *red*. **c** Heatmap representing genome-level gene architectural traits, for reference: intron density (introns/gene), median exon and intron length (bp), and GC content (percentage). **d** Difference in intron lengths (*y*-axis) between homologous sites of *S. arctica* compared to *C. fragrantissima* or *C. owczarzaki*, ranking introns according to their length differential with *S. arctica* (*x*-axis). *Black arrows* indicate the percentage of introns that are longer in *S. arctica* than in the other species (totals between brackets). **e** Difference in intron lengths (*y*-axis) between homologous sites of *V. carteri* compared to *C. reinhardtii* or *C. variabilis*, ranking introns according to their length differential with *V. carteri* (*x*-axis). *Black arrows* indicate the percentage of introns that are longer in *V. carteri* than in the other species (totals between brackets). **f** GC content in exons (GC_{ex}) and introns (GC_{in}) for each species in the dataset. Note that $GC_{ex} > GC_{in}$ in all species except for the glaucophyte *Cyanophora paradoxa*

ES-negative ones, were shorter, and had higher intron-to-exon length ratios. Furthermore, the majority of *S. arctica* and *V. carteri* introns were longer than their cognates in their close unicellular holozoan (Fig. 4e) and chlorophyte relatives [38] (Fig. 4f), respectively, suggesting recent intron lengthening events. Moreover, both *S. arctica* and *V. carteri* have relatively high intron densities (3.22 and 3.88 introns/CDS kbp; Fig. 1b, c and Additional file 1: Figure S2) that derive from recent, lineage-specific intron gain processes at the root of ichthyophonid Ichthyosporea [27] and Chlorophyceae plus Trebouxiophyceae [26] (Table 1). As the median CDS length is relatively constant across eukaryotes (~ 1400 bp [57]) and is independent of intron content, higher intron densities at the species level usually imply the presence of shorter exons (Fig. 1c, Additional file 1: Figure S19). Thus, *V. carteri* and *S. arctica* seem to have independently acquired ES-conducive genome architectures (higher intron densities, shorter exons flanked by longer introns) that contribute to their ES-richer AS profiles when compared to their closest relatives (Fig. 2e, f). On the other hand, the most notable exceptions to these patterns were the multicellular phaeophyte *E. siliculosus* (which exhibited low ES frequencies despite having unusually long introns; Fig. 4c), the charophyte *Klebsormidium netis*, and *B. natans*. In these species, ES was associated with short exons, but, unusually, also with short introns (Fig. 4a).

Regarding IR, the 'intron definition' splicing model proposes the opposite scenario: impediments to across-intron recognition of splice sites can lead to IR, and this mode of

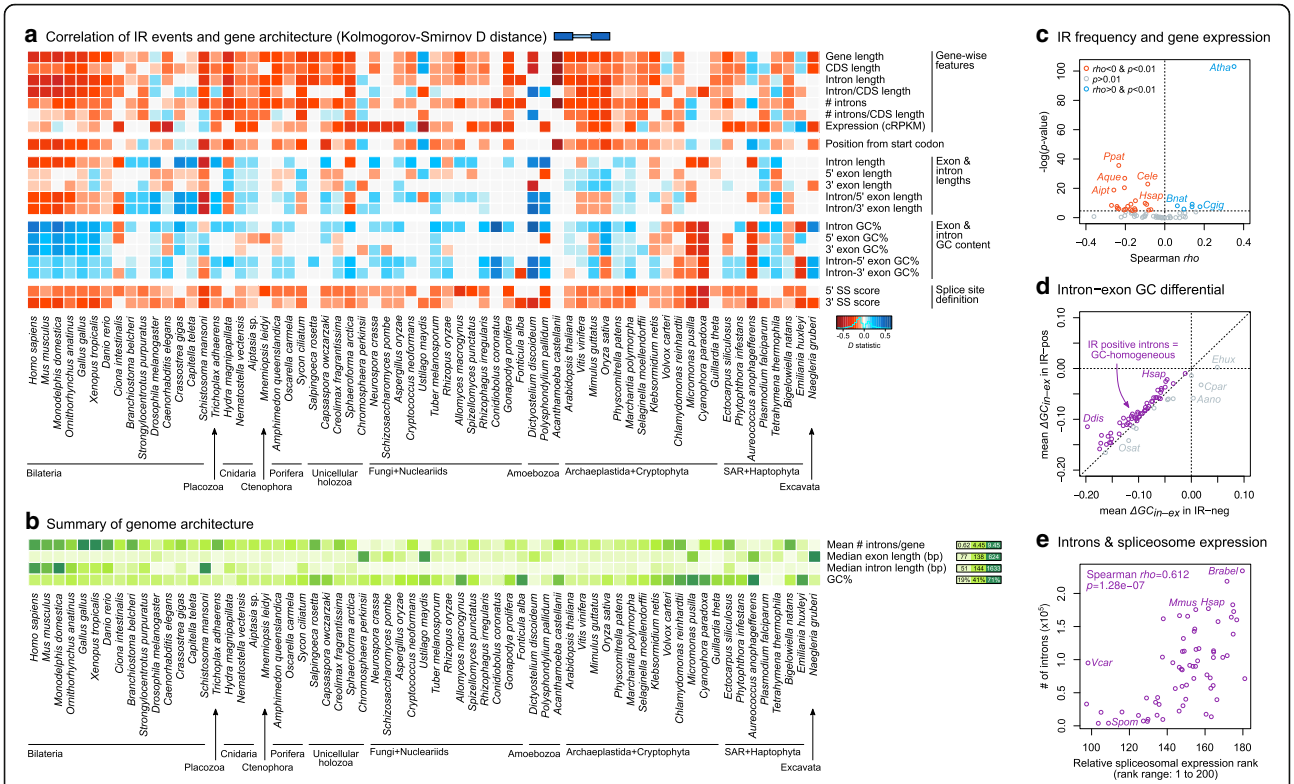

Fig. 5 Relationship between gene structure and sequence composition and ES. **a** Heatmap representing the distance between distributions of gene architecture values for IR-positive and IR-negative introns, measured with the *D* statistic of the one-sided Kolmogorov-Smirnov two-sample test (significant if $p < 0.01$; otherwise grey). *D* values are recorded as positive (*blue*) or negative (*red*) according to two one-sided Kolmogorov-Smirnov tests with complementary alternative hypotheses. **b** Heatmap representing genome-level gene architectural traits, for reference: intron density (introns/gene), median exon and intron length (bp), and GC content (percentage). **c** Volcano plot representing the Spearman's rank correlation coefficient (*rho*, *x*-axis) between IR frequency and transcript expression levels (cRPKM) in IR-positive genes. Significance threshold of $p < 0.01$ (*y*-axis, log-scale). **d** Mean intron–exon GC differential in IR-negative (*x*-axis) and IR-positive (*y*-axis) introns. The majority of eukaryotes lie above the diagonal (*purple*) as they have lower differentials in their IR-positive introns, i.e., IR occurs in GC-homogeneous introns. **e** Correlation between intron content (total number of introns per genome) and the relative rank of expression of the spliceosomal genes, per species (see "Methods"; Additional file 1: Figure S25)

splicing typically happens for short introns flanked by long exons [50, 53]. Surprisingly, however, our analysis revealed that the influence of intron and flanking exons' length on IR is not homogeneous across eukaryotes (Fig. 5a): retained introns are indeed shorter than constitutively excluded ones in chordates (vertebrates and the tunicate *Ciona intestinalis*), *S. arctica*, *Vitis vinifera*, and unicellular algae (*Micromonas pusilla*, *Cyanophora paradoxa*, *Emiliania huxleyi*, *B. natans*, or *Guillardia theta*); but not in most other animals, unicellular holozoans, fungi, or other protists that also exhibit high IR frequencies (Additional file 1: Figure S6). The ratio of intron-to-exon length has a similarly uneven relationship with IR. Across most eukaryotes, however, introns in genes with lower intron densities (introns/gene) were generally more prone to IR, as expected. Overall, the dominance of IR in a given genome does not seem to be determined by a straightforward relationship between intron length and density. Instead, positive or negative associations can be found in a lineage- or species-specific manner.

We also used the GC content of introns (GC_{in}) and exons (GC_{ex}) and its differential ($\Delta GC_{in-ex} = GC_{in} - GC_{ex}$) to examine the effect of global sequence composition in AS within each species. It has been proposed that, in species with extremely long introns (e.g., mammals and *G. gallus*), a differential in GC content between exons (GC-richer) and flanking introns (AT-richer) can act as a compositional mark to assist the recognition of splice sites [52]. Considering that GC_{in} is lower than GC_{ex} in all but one eukaryote in our dataset (Fig. 4f), GC differentials in ES-positive exons could be expected to take negative values ($\Delta GC_{in-ex} << 0$), particularly for long introns. However, we found that this is far from being a general rule, as we observed multiple intricate associations between ΔGC_{in-ex}, ES frequency, and intron length that vary across eukaryotes.

First, we found that the overabundance of ES-positive exons in genic environments where $\Delta GC_{in-ex} << 0$ (i.e., GC-rich exons flanked by AT-richer introns) occurs only in vertebrates and the annelid *C. teleta* (Fig. 4a). In these

Table 1 Intron mean lengths and densities in ancestral eukaryotic genomes

Ancestor	Ancestral mean intron length (bp)	Introns/CDS kbp	Introns/gene	Source intron densities
LECA	–	4.3	6.11	Csűrös et al. 2011 [26]
Uropisthokonta	328	5.1	7.25	Csűrös et al. 2011 [26]
Urholozoa	481	5.52	7.86	Grau-Bové et al. 2017 [27]
Urichthyosporea (*Chromosphaera*, *Creolimax*, *Sphaeroforma*)	263	5.53	7.86	Grau-Bové et al. 2017 [27]
Urichthyophonida (*Creolimax*, *Sphaeroforma*)	393	6.98	9.92	Grau-Bové et al. 2017 [27]
Urmetazoa (animals)	845	8.8/8.7	12.51/12.37	Csűrös et al. 2011 [26]/ Grau-Bové et al. 2017 [27]
Urporifera (sponges)	451	8.63	12.27	Grau-Bové et al. 2017 [27]
Urcnidaria	1016	8.3	11.80	Csűrös et al. 2011 [26]
Urbilateria	1342	7.7/7.7	10.94/10.94	Csűrös et al. 2011 [26]/ Grau-Bové et al. 2017 [27]
Urprotostomia	891	7.4	10.52	Csűrös et al. 2011 [26]
Urecdysozoa	623	7.4	10.52	Csűrös et al. 2011 [26]
Urdeuterostomia	1710	7.7	10.94	Csűrös et al. 2011 [26]
Urvertebrata	3117	7.2	10.23	Csűrös et al. 2011 [26]
Urembryophyta (land plants)	313	6.4	9.10	Csűrös et al. 2011 [26]
Urchlorophyta (all unicellular/colonial green algae)	299	3.4	4.83	Csűrös et al. 2011 [26]
Trebouxiophyceae + Chlorophyceae (*Chlorella*, *Volvox*, *Chlamydomonas*)	–	6.2	8.82	Csűrös et al. 2011 [26]

Mean intron lengths estimated from phylogenetically independent comparisons of descendant species (see "Methods"; Additional file 1: Figure S24). Intron densities (taken from [26, 27]) are reported as introns/CDS kbp and introns/gene (by multiplying by the average CDS length [1422 kbp] of all organisms in our dataset; Additional file 1: Figures S1 and S2)

animals, ES-positive exons were also found to be preferentially flanked by long introns (see above). However, long introns flanking ES-positive exons were only found to be AT-enriched in mammals and *G. gallus* (Spearman $rho < 0$ and $p < 0.01$; Figs. 4b and 6a, b). Second, a number of other eukaryotes (e.g., *S. ciliatum*, *S. arctica*, or *O. sativa*) exhibited a similar correlation between long and AT-rich introns genome-wide (red in Fig. 4b), but not the concomitant association between strong GC differentials ($\Delta GC_{in-ex} << 0$) and ES (Figs. 4a and 6e). Third, ES-positive exons in nearly all non-vertebrate eukaryotes were surprisingly biased towards regions where introns and exons have similar GC content (i.e., $\Delta GC_{in-ex} \sim 0$; Fig. 6b–e). This latter pattern was highly unexpected, as it matches that described for IR events in human [52], which we also observed for IR for most (yet not all) eukaryotes in our dataset (Fig. 5a, d). Altogether, these results reveal complex lineage-specific interplays among GC content, intron length, and ES frequency, which cannot be generalized among eukaryotes (Fig. 6).

We have also examined the effect of whole-transcript expression levels on AS. In the majority of eukaryotes, IR-positive introns are preferentially found in lowly expressed transcripts (Fig. 5a). This result is predicted by two alternative and non-mutually exclusive hypotheses. On the one hand, IR has been widely associated with down-regulation of gene expression via NMD [16]. On the other hand, random splicing errors are more prone to affect lowly expressed genes, given its reduced fitness cost [49]. A closer inspection of IR-positive introns alone also recovered a widespread correlation between high IR rates ($r_{IR,in}$) and lower expression levels in 28 out 37 species where the relationship was significant (Fig. 5c, Additional file 1: Figure S20; $p < 0.01$ and $rho < 0$ in Spearman's rank correlation test). This result had been previously described in mammalian transcriptomes [16] and can thus be extended to all eukaryotes.

Finally, we investigated the relationship between the relative expression of core spliceosomal components and IR and ES frequency genome-wide. Since efficient splicing depends, in principle, on the sufficient expression of spliceosomal components [58, 59], we asked whether low relative expression of core factors correlated with higher ES and/or IR frequencies at the species level (Additional file 1: Figure S21A–C). Using a rank-based score to measure relative spliceosome expression (see "Methods"), we identified a negative association between spliceosome expression and IR frequency (Additional file 1: Figure S21B). This result suggests that species-wide IR levels, which do not follow a phylogenetic pattern (Additional file 1: Figure S6), could be at least partly explained by the competition among unspliced transcripts

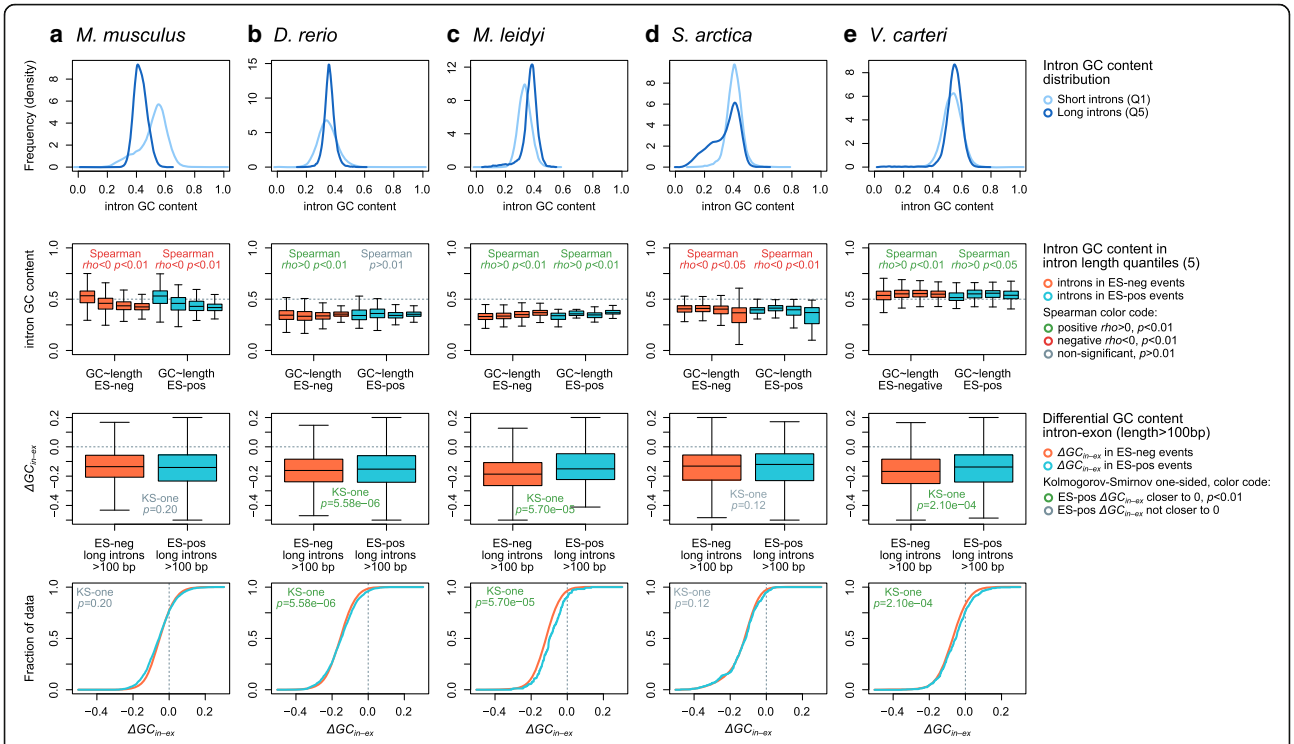

Fig. 6 Representative patterns of the inter-relationships between GC content and AS. For selected eukaryotes (**a**–**e**), we plotted the distribution of GC content (percentage) in long and short introns (*first row*); the distribution of GC content in introns sorted according to their length (five quantiles, Q1–5), and ES status (negative and positive; *second row*); and the differential between intron and exon GC content in long introns (Q5) of ES-positive and ES-negative events, using box plots (*third row*) and empirical cumulative distribution functions (*fourth row*). In the *second row*, we used Spearman's rank correlation coefficient to test the association between intron length and GC content (significant at $p < 0.01$). In the *third* and *fourth rows*, we used Kolmogorov-Smirnov one-sided test to test whether ES-positive introns had higher ΔGC_{in-ex} than ES-negative ones (i.e., $\Delta GC_{in-ex} \sim 0$), and thus GC-homogeneous across introns and exons (significant $p < 0.01$)

for access to the available spliceosomal machinery [58, 59] in a species- and/or sample-dependent manner. On the other hand, we found a mild, unexpectedly positive association of core spliceosomal expression with ES frequency (Additional file 1: Figure S21C). However, this result is likely due to the strong correlation found between relative spliceosomal expression and the total number of introns (Spearman's $rho = 0.612$, $p = 1.28e-7$; Fig. 5e) and intron density (Additional file 1: Figure S21D), which were also positively correlated with ES frequencies (Fig. 4). Thus, relative expression of core spliceosomal components seems adjusted to the number of introns to be spliced in each species.

In summary, ES events across eukaryotes were globally associated with short exons flanked by longer introns, and with weak 5′ and 3′ splice sites. Inasmuch as these features are more common in animals and plants than in most eukaryotes, we can expect higher ES frequencies in these multicellular lineages.

Dating ES transitions in ancestral Holozoa genomes

The relationships between ES and genomic features described above were highly consistent among Holozoa (animals and their closest unicellular relatives), which suggests that these architectural and sequence effects influenced ES frequencies not only in extant organisms but in their extinct ancestors as well. We thus reasoned that these relationships could be used to predict the incidence of ES in ancestral genomes, provided that their genome architectures could be approximated.

For this purpose, we first trained a binomial logistic regression model (see "Methods") that classifies arbitrary exons as either ES-positive or -negative according to their overall gene architecture, assigning them an ES-positive probability (p_{ES}). We used data from a selection of 24 eukaryotes with multiple animals and holozoans (Additional file 1: Figures S22 and S23) and ascertained their sensitivity and specificity by calculating the ROC curve of the classifier (Fig. 7a) and the area underneath (AU-ROC = 0.752, 95% CI = 0.743–0.762). This AU-ROC value indicates a clearly better-than-random classifier and compares with similar AS predictors previously developed for more taxonomically restricted contexts, such as IR in mammals (AU-ROC = 0.79) [16] and ES in vertebrates (AU-ROC = 0.79–0.87 in individual species) [51]. In addition, it shows that gene architecture is a

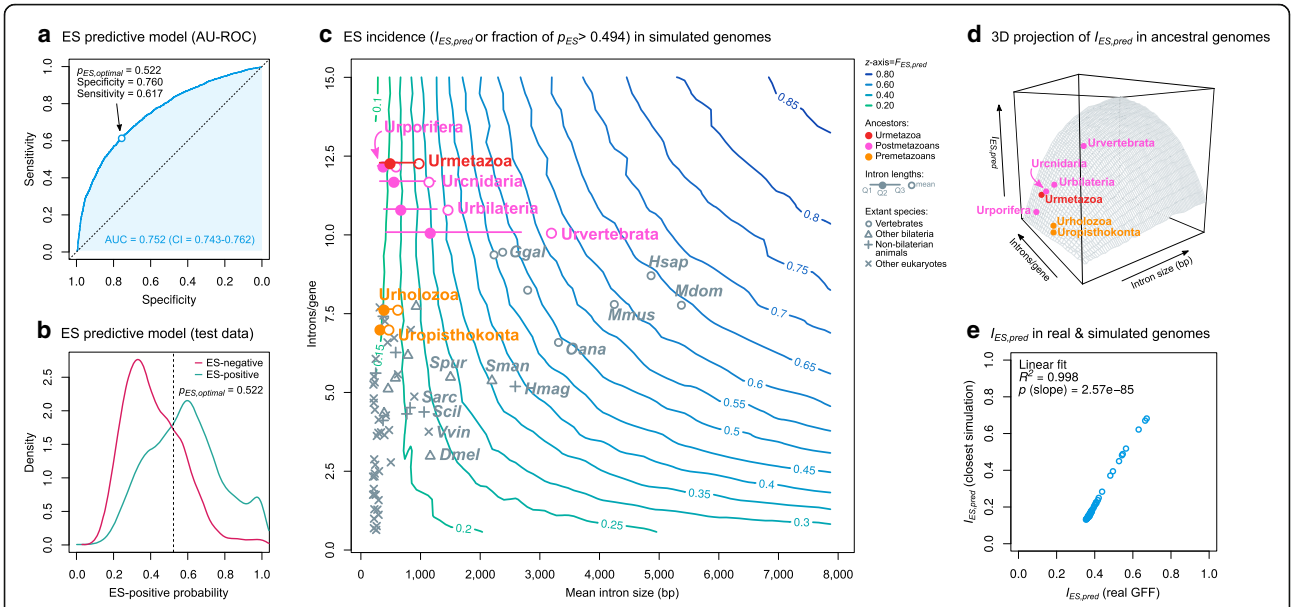

Fig. 7 Prediction of ES incidences in ancestral genomes. **a** Receiver operating characteristic (ROC) curve of a binomial logistic regression model of ES prediction derived from 11 gene architectural features, 22 species, and a dataset of 18,678 exon skipping events for training and testing (see "Methods"). AU-ROC value (0.752) is reported with 95% confidence interval. The non-*pale blue dot* marks the optimal probability threshold (0.522) corresponding to the maximum sum of sensitivity (0.617) and specificity (0.760). **b** ES-positive probability distributions of the test dataset, containing 9318 exons with known ES status (50% positive, in *green*; 50% negative, in *maroon*). The *dotted line* marks the optimal probability threshold (0.522; see above). **c** Contour plot representing ES incidences (I_{ES}, *z* axis) across 1600 simulated genomes of varying mean intron sizes (*horizontal*) and intron densities (*vertical*). I_{ES} is defined as the fraction of exons in a given simulation with $p_{ES} > 0.522$. *Grey marks* represent extant species positions in the intron size and density dimensions. *Red/pink/orange-colored dots* represent estimates of ancestral premetazoan and metazoan genomes. Ancestral intron densities from [26, 27] (Table 1); ancestral intron length distributions are represented by the mean, median, and first and third quartiles (Table 1, Additional file 1: Figure S24; see "Methods"). **d** 3D projection of **c**, with I_{ES}, (*z*-axis), intron sizes (*x*-axis), and intron densities (*y*-axis). Each *dot* represents an estimation of a genome (intron density and mean intron length). **e** Correlation between the I_{ES} values obtained from analyzing real genome architectures with the predictive classifier (*x*-axis) and its closest equivalent from the spectrum of simulated genomes (*y*-axis) in terms of intron density and mean length

consistent predictor of ES events. The highest combined sensitivity and specificity were attained when the ES-positive probability threshold (p_{ES}) was set at 0.522 (Fig. 7b).

Next, we used this classifier to compare ES levels from ancestral genomes, using a topographic map of the predicted ES incidence based on genomic architectural traits (Fig. 7) from 1600 genomes simulated by combining initial assumptions of mean intron density (range 0.5–15 introns/gene) and mean intron length (range 10–8000 bp). For each simulation, we generated 10,000 internal exons with their corresponding gene architectures (see "Methods" for details) that were analyzed with the classifier to assign them an ES-positive probability (p_{ES}). Then, we approximated the incidence of ES in each simulated genome by calculating the fraction of exons with p_{ES} over the optimal threshold ($p_{ES} > 0.522$; henceforth I_{ES}). An examination of ES incidences across the spectrum of simulated ancestral genomes revealed that genomes with higher intron densities and lengths have higher I_{ES} values (Fig. 7c, d), as expected from transcriptomic analyses of extant species (Fig. 4a, b). To assess the accuracy of our simulated genomes, we compared their I_{ES} values with those obtained from the most similar real genome in our dataset (in terms of intron density and mean length), finding a linear correlation (linear fit $R^2 = 0.996$, slope $p = 1.25e-76$; Fig. 7e).

Then, we used this predictive framework to analyze ES transitions in animals by comparing the incidence of ES in reconstructed premetazoan and postmetazoan ancestors (Fig. 7c). To do so, we used previously published estimates of intron density in ancestral genomes obtained by comparative genomic analyses [26, 27], which reported an increase in intron density from 7.85 to 12.37 introns/gene between the origin of Holozoa (Urholozoa) and animals (Urmetazoa) (Table 1). As a proxy for ancestral intron lengths, we used means obtained from extant species' phylogenetically independent contrasts [60] (see "Methods"; Additional file 1: Figure S24). Under these assumptions, the Urmetazoa (12.37 introns/gene of ~850 bp mean length) would have an I_{ES} ~0.30 (Fig. 7c). Conversely, the intron-poorer Urholozoa would reach I_{ES} ~0.30 only if it had an average intron of >1000 bp, which exceeds by an order of magnitude the average intron

lengths of extant unicellular holozoans (120–571 bp; Additional file 1: Figure S2) [27] and our ancestral estimation (~ 481 bp). Therefore, under reasonable assumptions for ancestral genome architecture, the unicellular ancestors of animals likely had lower ES levels than the Urmetazoa (Fig. 7c, d).

This same line of reasoning can be applied within animals. Since the Urmetazoa (12.36 introns/gene, ~ 850 bp), intron loss processes occurred in the ancestral cnidarians, poriferans, bilaterians, and vertebrates (Table 1) [26, 27]. Therefore, to sustain an increase in ES levels relative to the Urmetazoan, these lineages should have undergone intron lengthening processes. This was probably the case already in early bilaterians (I_{ES} ~ 0.40 if mean intron length ~ 1300 bp) and, with higher certainty, in vertebrates (I_{ES} ~ 0.60 if mean intron length ~ 3100 bp; Fig. 7c, d). On the other hand, our results do not support bilaterian-like ES enrichments in the early branching animal ancestors Urcnidaria and Urporifera: although they had high intron densities, they likely had shorter introns compared to the Urbilaterian (Table 1), thus yielding lower I_{ES} values (Fig. 7c, d).

Therefore, we hypothesize that the moderately high ES in certain non-bilaterian eukaryotes appear concomitantly with recent, species-specific changes in their gene architecture. One possible example is *H. magnipapillata*, which has undergone a recent, intra-genus genome size expansion [61, 62] that could help explain its enlarged introns (Fig. 1b). Similarly, ichthyosporeans attained high intron densities independently of animals [27] (Table 1), and the longer introns found in species like *S. arctica* appear to have been recently acquired too (Fig. 4d). Altogether, prediction of ES incidences from ancestral reconstructions and comparison with extant species are consistent with the increase in ES frequency observed in metazoans, particularly from bilaterian ancestors.

Discussion

We performed a comparative survey of AS frequencies in 65 eukaryotes to understand the evolutionary dynamics of this layer of gene regulation. Our analysis revealed that ES and IR events can be found, at varying frequencies, in transcriptomes from all major eukaryotic groups (Opisthokonta, Amoebozoa, Virdiplantae and Cryptophyta, SAR and Haptophyta, and Excavata). This result thus provides further support for an early emergence of dual AS—i.e., involving both ES and IR—in eukaryotes [18, 33, 34]. Given the generally higher prevalence of IR in all eukaryotic super-groups (Fig. 2a, b, Additional file 1: Figure S6), LECA likely exhibited an IR-dominated AS profile as well.

Since its early origin in LECA, the incidence of ES has varied between different eukaryotic groups. Our analysis of ES frequencies across multiple species revealed higher ES frequencies in multicellular animals (particularly vertebrates and other bilaterians) and, to a lesser extent, plants. While most other eukaryotes maintain lower ES levels, a number of punctual exceptions have surfaced—e.g., the ichthyosporean *S. arctica* or the colonial chlorophyte *V. carteri* [38] (Fig. 2). Conversely, other major lineages like fungi or amoebozoans exhibit comparatively low ES levels.

In most ES-rich eukaryotes, we identified a set of gene architectural features that influence the frequency of AS at the gene level: the heterogeneity of 5′ and 3′ splice sites, the length of exons and their flanking introns, and its correlate regarding intron density. These architectural traits were globally coherent across eukaryotic lineages, especially in animals and plants (Fig. 4a). In particular, the mode of splicing by 'exon definition', which posits that short exons flanked by long introns are enriched in ES events [50, 53], is relevant in most eukaryotic transcriptomes with sufficient ES levels analyzed here, except for *E. siliculosus* and *B. natans*. These properties are more common in intron-rich organisms with heterogeneous splice sites [18, 34]. Overall, this result suggests a non-deterministic 'soft code' that influences ES rates across eukaryotic lineages.

The existence of a pan-eukaryotic ES 'soft code' implies that inter-specific changes in ES levels can be associated with the evolutionary histories of their underlying genomic traits. Since the origin of intron-rich genomes in LECA, the evolutionary lineages leading up to ES-rich animals and plants always maintained high intron densities, and later underwent secondary intron gain processes [26, 27, 63] concomitant with ES transitions (Fig. 2). Such intron gain episodes had a direct effect on exon length: given that the mean CDS length is relatively constant across eukaryotes [57] and that intron content is independent of CDS length (Additional file 1: Figure S24), genomes affected by long-term intron gain processes have shorter exons (Fig. 1c, Additional file 1: Figure S24). This is likely a direct consequence of the most common mechanisms of de novo intron creation, which involve the insertion of new intronic sequences splitting pre-existing exons [64–68]. Furthermore, most extant intron-bearing eukaryotic genomes maintain heterogeneous splice sites [18, 34], a conserved core spliceosomal machinery [31] (Additional file 1: Figure S21), and a diverse complement of splicing factors [27, 32], all of which were already present in LECA. Globally, the evolutionary dynamics of these genomic traits are consistent with our results: the early origin of the genetic machinery (core spliceosome) and structure (intron-rich genomes with diverse splice sites) fits the ancestral emergence of ES in LECA and its widespread incidence, and subsequent lineage-specific changes in genome architectures would have paved the way for the evolution of higher ES frequencies.

Furthermore, if the effect of gene architecture on ES is widespread and coherent across eukaryotes, we can infer that it was also relevant in their ancestors. Thus, we took advantage of the pan-eukaryotic 'soft code' of ES determination to investigate the timing of ES transitions in animals and their unicellular ancestry. Specifically, we quantified the relationship between gene architecture and ES on extant eukaryotes and projected these effects into the past with a predictive framework (Fig. 7c, d). We propose an early ES enrichment in the Urmetazoa concomitant with the origin of multicellularity, followed by further enrichment in bilaterians and vertebrates. The ES transition in early animals would come as a consequence of changes in genome architecture, as large genomes with high intron densities and long intronic segments became more common in animals [26, 27, 69] than in their unicellular holozoan relatives [27, 57, 70–72].

Remarkably, most animals showed high fractions of 3n exons among their ES events, but this was not observed in plants or other eukaryotes (Fig. 3b, c). Thus, the pressure to maintain ORFs in the event of ES seems an animal-specific trait. Interestingly, even animals with lower ES frequencies (such as non-bilaterians) or that have secondarily simplified intron–exon architectures (e.g., *C. elegans* and *D. melanogaster*) often have strong 3n biases in their ES profiles. Thus, it is conceivable that a protein isoform-enabling 3n bias already existed in the Urmetazoa, before the major increase in ES frequency occurred in bilaterians (Figs. 2 and 7). If so, these increases were likely more easily recruited to functionally diversify their proteomes, perhaps contributing to the selection of more ES-prone genome architectures, establishing a positive feedback loop. Moreover, it is possible that other regulatory features that are characteristic of bilaterians, such as the high prevalence of long-range enhancer–promoter interactions [73], may have also contributed to longer intron sequences and thus to more ES-prone genome architectures.

On the other hand, neither unicellular holozoans (Fig. 2) nor the unicellular ancestors of animals (Fig. 7) seem to have had ES-rich transcriptomes with 3n exon biases. Therefore, AS-mediated isoform production was a largely irrelevant phenomenon during the unicellular ancestry of animals (since the Urholozoa to the origin of multicellularity), an evolutionary period that was otherwise fecund in other sources of gene innovation [27].

Conclusions

We find that the influence of gene architectural traits in the frequencies of IR and ES is globally conserved across all eukaryotic lineages. Thus, gene architecture (i.e. the lengths of introns and exons, splice site definition, intron density, etc.) is the basis of a 'soft' pan-eukaryotic cis-regulatory code for AS determination that affects both extant and ancestral genomes. This result emphasizes the effect of long-term genome evolutionary patterns in shaping AS, a fast-changing transcriptome regulatory layer. In that regard, we identify multiple ES transitions coinciding with the evolution of ES-favourable genome architectures – e.g. in animals and plants, but also in more restricted taxonomic contexts such as the ichthyosporean *S. arctica*.

Our taxon-rich analysis confirms that animal transcriptomes have a unique AS profile. Quantitatively, extant animals exhibit the highest ES frequencies among eukaryotes as a consequence of cumulative ES enrichments in the Urmetazoa and, above all, the Urbilateria. Furthermore, from a qualitative perspective, the earliest animals became enriched in frame-preserving ES events, which is essential for widespread isoform-mediated proteome diversification. Thus, our observations suggest that the unparalleled increase in ES frequencies of modern bilaterians (including vertebrates) is a consequence of the interplay between the complexification of animal genome architectures, on one hand, and the co-option of ES events for regulated proteome expansion, on the other.

Methods

Sources of genome and transcriptome data

We assembled a dataset consisting of genome assemblies and annotations from 65 eukaryotic species for which high-coverage Illumina RNA-seq data were already available or for which we generated de novo data (Additional file 1: Figure S1 and below). We retrieved the genomic coordinates of genes, transcripts, and exon sequences for each genome from associated GFF annotation files. If more than one isoform per gene was annotated, the longest CDS was considered to be the canonical transcript (a proxy with ~ 90% correspondence with proteomics-driven main isoform selection [74]). In order to homogenize the experimental procedures used to build each RNA-seq library, we used (i) poly(A)-selected libraries only, (ii) either single-end or the forward reads of paired-end data, and (iii) trimmed reads to a length of 50 bp if they were longer (using FASTX Toolkit [75]). In species where RNA-seq experiments included more than one sample (replicates, time series, growth conditions, etc.), all reads were pooled into a single FASTQ file.

In the case of *N. gruberi*, *S. arctica*, and *T. adhaerens*, new RNA-seq datasets were produced and deposited in the European Nucleotide Archive (ENA) under the accession codes PRJEB23822, PRJEB23831, and PRJEB23829, respectively. RNA extractions were performed from confluent axenic cultures of mixed cells (*N. gruberi*, ATCC 1034 medium, modified PYNFH at 30 °C; *S. arctica*, marine broth medium Difco 2216 at 12 °C) or whole

organisms (*T. adhaerens*, artificial sea water at 25 °C). RNA was extracted using Trizol reagent (Life Technologies, Carlsbad, CA, USA) with a further step of Dnase I (Roche) to avoid contamination by genomic DNA, then purified using RNeasy columns (Qiagen). We sequenced paired-end libraries of 50 (*N. gruberi*, 276,095,412 reads; *S. arctica*, 218,701,756 reads) or 125 bp (*T. adhaerens*, 84,583,746 reads) with an insert size of 250 bp. Libraries were constructed using the Trueseq Sequencing Kit v4 (Illumina, San Diego, CA, USA). The libraries were sequenced in one lane of Illumina HiSeq 2000 (*N. gruberi*, *T. adhaerens*) or 2500 (*S. arctica*) at the CRG genomics facility (Barcelona).

Detection and quantification of exon skipping and intron retention events

We adapted and expanded the computational framework previously developed [36, 51] to detect and quantify ES and IR (graphical summary in Additional file 1: Figure S3), as follows.

Exon skipping detection

For each group of three consecutive exons in the genome (exon triplet), we built a composite of exonic junctions consisting of (i) 42 bp from the 5′ end of the first exon and 42 bp from the 3′ end of the second exon (E1–E2 junction); (ii) 42-bp fragments from the 5′ end of the first exon and the 3′ end of the third exon (E1–E3); and (iii) 42-bp fragments from the 5′ end of the second exon and the 3′ end of the third exon (E2–E3). Hence, each triplet consisted of two inclusion junctions (E1–E2 and E2–E3) and one that skipped the middle exon (E1–E3). If any exon was shorter than 42 bp, the entire length of the exon was used, and the resulting junction sequence would be shorter than 84 bp.

Then, we computed the effective mappability of each junction in order to exclude exon–exon boundaries where RNA-seq mapping would be unreliable [76]. Specifically, we (i) built an artificial RNA-seq library consisting of all the possible reads derived from each junction in a 50-bp sliding window; (ii) mapped these reads to the original junctions using *bowtie* v1.1.2, allowing a maximum of two mismatches (*–v 2*) and no multiple alignments (*–m 1*) [77]; and (iii) removed all junction triplets for which at least one triplet had < 20 effectively mappable positions (maximum is 35 for 50 bp reads, and ≥ 8 positions mapped from each exon). Then, we aligned the pooled RNA-seq libraries to the remaining exon triplets with *bowtie* and the same parameters as above. We corrected the number of mapped reads by dividing the read counts by the ratio obtained from dividing the mappable positions of that junction (20–35 bp) and the maximum theoretical mappability (35 bp).

The ES rate of each middle exon (r_{ES}) was then computed as follows:

$$r_{ES} = \frac{m_{E1E3}}{(m_{E1E2} + m_{E2E3})/2 + m_{E1E3}}$$

where m denotes the mappability-corrected number of reads mapping in the E1–E2, E2–E3, and E1–E3 junctions. We classified exon junctions into three categories: (i) ES-positive if $m_{E1E2} + m_{E2E3} + m_{E1E3} > 10$, $m_{E1E2} > 2$, $m_{E2E3} > 2$, $m_{E1E3} > 1$, and $r_{ES} \geq 10\%$ but $< 90\%$; (ii) ES-negative if $m_{E1E2} + m_{E2E3} + m_{E1E3} > 10$, $m_{E1E2} > 2$, $m_{E2E3} > 2$, $m_{E1E3} \geq 0$ but $r_{ES} < 10\%$; and (iii) non-classifiable if any condition was not fulfilled.

Intron retention detection

For each intron of the genome, we built three junction sequences consisting of (i) 42 bp from the 5′ end of the first exon and 42 bp from the 3′ end of adjoining intron (E–I junction); (ii) 42-bp fragments from the 5′ end of the first exon and the 3′ end of the second exon (E–E); and (iii) 42 bp fragments from the 5′ end of the intron and the 3′ end of the second exon (I–E). Hence, each intron triplet had one spliced junction (E–E) and two retention junctions that spanned the intron ends (E–I and I–E), all of them 84 bp long (or less, if any exon or intron was shorter than 42 bp).

The mappability of each exon–intron junction was computed as specified above for ES junctions, also discarding cases with < 20 effectively mappable positions. We then aligned the same pooled RNA-seq data to the remaining exon–intron junctions using *bowtie*, and corrected the number of mapped reads.

The IR rate of each intron (r_{IR}) was computed as follows:

$$r_{IR} = \frac{(m_{IE} + m_{EI})/2}{m_{EE} + (m_{IE} + m_{EI})/2}$$

where m denotes the mappability-corrected number of reads mapping in the I–E, E–I, and E–E junctions. Finally, we classified intron junctions in three categories: (i) IR-positive if $m_{IE} + m_{EI} + m_{EE} > 10$, $m_{IE} > 1$, $m_{EI} > 1$, $m_{EE} > 2$, and $r_{IR} \geq 10\%$ but $< 90\%$; (ii) IR-negative if $m_{IE} + m_{EI} + m_{EE} > 10$, $m_{IE} \geq 0$, $m_{EI} \geq 0$, $m_{EE} > 2$, and $r_{IR.in} < 10\%$; and (iii) non-classifiable if any condition was not fulfilled.

ES and IR detection pipeline quality control

With the aim of identifying biases in our computational pipeline, we assessed the fraction of all potential exon–exon and exon–intron junctions that passed the mappability filters and were used in the ES and IR detection procedures (Additional file 1: Figure S25). Putative systematic biases derived from species-specific genome architectural traits

(paralogy, exon/intron lengths, repetitive genomes) could affect RNA-seq mapping in certain species and thus bias downstream analyses.

We analyzed the fraction of junctions with > 20 effective mappable positions (mappability filter, Additional file 1: Figure S25A, B), finding no evidence of taxonomic biases in any particular eukaryotic group. A few phylogenetically unrelated species exhibited low survival rates (e.g., *Selaginella moellendorffi* and *E. huxleyi*). This general mappability filter was then decomposed in its different components. First, we assessed the fraction of exons and introns from each genome that had the minimal required length to survive the mappability filter (i.e., ≥ 27 bp), finding no bias in any species (Additional file 1: Figure S25C, D). Next, we examined which mapping filters (i.e., *bowtie* multiple mapping [–m 1] or excess of mismatches [–v 2]) caused ES or IR junction removal in each species (Additional file 1: Figure S25E, F), finding that peaks in *S. moellendorffi* and *E. huxleyi* were mostly due to multiple mapping. Multiple mapping is likely caused by abundant recent gene duplications in these two species. Indeed, both *S. moellendorffi* and *E. huxleyi* had a high fraction of intron-bearing genes with recent paralogous sequences (> 99% or > 95% amino acid sequence identity, and > 90% reciprocal alignment coverage, calculated with *diamond* [78]; Additional file 1: Figure S25G, H). Finally, we found that some species also had a relatively high number of uncalled bases (N) in their exon–exon or intron–exon junctions (most notably *O. sativa* but also *S. moellendorffi*; Additional file 1: Figure S25I, J), which can partly explain the lower survival rates in these species after the mappability filter (Additional file 1: Figure S25A, B).

Overall, these quality control analyses show that the compounded effect of genome architectural traits (recent duplication, uncalled bases, and repetitive sequences hindering RNA-seq mapping) only affected individual species and did not systematically affect ES or IR detection in any large group of eukaryotes.

Transcriptome-wide quantification of AS levels

In order to measure the average frequency of ES at the species level, we divided each species' set of classifiable exon triplets into 100 bins of 100 triplets, and calculated the per-triplet frequency of ES from 10,000 randomly chosen reads for each bin (selected among those mapping to the in bin's exon–exon junctions, or a mean sequencing depth of ~ 20×). The average ES rate of each bin i ($r_{ES,i}$) was recorded to obtain a species-level distribution of ES frequencies ($F_{ES,sp}$). An analogous measurement was used to calculate the distribution of species-level IR frequencies ($F_{IR,sp}$) from 100 bins, 100 triplets, and 10,000 reads per bin (mean sequencing depth ~ 20×). This process is summarized in Additional file 1: Figure S3C. It should be noted that comparisons among $F_{IR,sp}$ are more susceptible to technical variability derived from library preparation than those among $F_{ES,sp}$, since differences in the efficiency in poly(A)+ selection are only expected to significantly affect $F_{IR,sp}$ estimates.

Statistical comparisons of poolings of species-level $F_{ES,sp}$ distributions (Fig. 2c–f) were done using one-sided Kolmogorov-Smirnov tests. Differences between intra-species replicates (Additional file 1: Figure S7) were assessed with Wilcoxon rank-sum tests. All statistical contrasts were done with R *stats* [79].

ES frequency analysis was re-assessed by using (i) downsampled RNA-seq experiments (to 2×, 5×, 10×, and 15× sequencing depths, by random read selection; Additional file 1: Figure S10); and (ii) individual tissue-level samples from selected animal (*H. sapiens*, 2× *M. musculus*, *M. domestica*, *O. anatinus*, *G. gallus*, *X. tropicalis*, *D. rerio*, *S. ciliatum*) and plant species (*A. thaliana* and *M. polymorpha*) (Additional file 1: Figure S13A). Original sequencing depths in the ES and IR triplets junctions are available in Additional file 1: Figure S25K, L.

Analysis of gene features: architecture, splice sites and expression levels

For each exon or intron analyzed, we recorded sequence and architectural parameters at the gene and AS event levels. At the gene level, we studied the following parameters: gene length, CDS length (all exons), total number of introns in the gene (intron density), position of the exon/intron within the gene sequence (base pairs from starting codon), and total length of all introns with respect to exons (ratio). At the AS event level, we recorded the length of the individual exon and flanking introns (for ES) or intron and flanking exons (for IR), and the ratio between them (intron/exon lengths); the GC content of exons and flanking introns (for ES) or introns and flanking exons (for IR), and the differential between them ($\Delta GC_{in-ex} = GC_{in} - GC_{ex}$), and a boolean variable describing whether the length of the alternative exon/intron was divisible by three (1 = true, 0 = false). These features were derived from the GFF annotation and genome sequence (data sources in Additional file 1: Figure S1).

In addition, we analyzed the conservation degree of 5′ and 3′ splice sites when compared to species-specific consensus. For each species, we built position-weighted matrices (PWM) from the alignments of all 3′ (23 bp, 20 from the intron and 3 from the exon) and 5′ (9 bp, 3 from the exon and 6 from the intron) splice sites using the consensus matrix function in the *Biostrings* R library [80]. Then, for each individual splice site in the genome, the distance from the PWM consensus was calculated. Splice sites were delimited as in [81].

Finally, we evaluated transcript expression levels using the mappability-corrected RPKM metric (cRPKM),

aligning the pooled RNA-seq data for each species to the genome-predicted transcript sets using *bowtie* (longest transcript per gene only, see above) and calculating transcript-specific effective mappabilities as detailed above [76].

Statistical analysis of AS frequency and gene features

For each species and for each of the quantitative sequence and architectural features listed above, significant differences between the values taken by the IR-/ES-positive triplets and the IR-/ES-negative triplets were evaluated using two independent one-sided Kolmogorov-Smirnov two-sample tests with complementary alternative hypotheses: first, we tested whether the empirical cumulative distribution of the IR-/ES-positive events lied above the IR-/ES-negative events' values (signaling a positive relationship between IR/ES and the given feature); second, we tested whether it lied below (i.e., for a negative relationship). We used the Kolmogorov-Smirnov distance (*D* statistic) to measure the distance between distributions. *D* was recorded as positive if $p < 0.01$ in the first test, negative if $p < 0.01$ in the second one, or as NA if it was not significant in any test or contradictorily significant. The resulting matrix was plotted using the *heatmap.2* function in the R *gplots* v3.0.1 library [82].

To further investigate the relationship between gene expresion levels and IR, we also tested the significance of monotonic correlations between cRPKMs and AS rates using Spearman's rank correlation coefficient (*rho*, significant for $p < 0.01$).

Finally, we tested if the frequency of 3n divisible lengths in IR-/ES-positive events significantly differed from that of IR-/ES-negative (i.e., constitutive) events using Fisher's exact test (significant for $p < 0.01$, except if otherwise stated). All statistical analyses were done with R *stats* library [79].

The relationships between gene architectural traits and ES and IR events were re-assessed by using (i) down-sampled RNA-seq experiments (to 2×, 5×, 10× and 15× sequencing depths, by random read selection; Additional file 1: Figures S17 and S18); and (ii) individual tissue-level samples from selected animal (*H. sapiens*, *M. musculus*, *M. domestica*, *O. anatinus*, *G. gallus*, *X. tropicalis*, *D. rerio*, *S. ciliatum*) and plant species (*A. thaliana* and *M. polymorpha*) (Additional file 1: Figure S13B, C). Original sequencing depths in the ES and IR triplets junctions are available in Additional file 1: Figure S25K, L.

Prediction of ES incidence using gene architectural features

Using our binary classification of positive/negative ES events, we created a binomial logistic regression model for a selection of representative eukaryotes with high ES frequencies. First, we selected 18,678 events with known gene architecture (devised to include an equal number of positive and negative events) from 24 representative eukaryotic species (Additional file 1: Figure S22). We used (i) 12,452 positive and negative ES events (two-thirds of the dataset) as the binary-dependent variable, and (ii) 11 quantitative gene traits and a Boolean factor indicating 3n divisibility as independent predictors (Additional file 1: Figure S22). The binomial logistic regression was built using the generalized linear model function from R *stats* library [79]. The predictive performance of each model was estimated with the area under its corresponding ROC curve (AU-ROC), calculated using an independent test subset (6226 events, one-third of the dataset) with the *pROC* R library [83]. An optimal probability threshold was selected by maximizing the sum of specificity and sensitivity ($p_{ES,optimal} = 0.522$). We assessed the significance of the model's coefficients with the Z-statistic significance according to the Wald test and its corresponding ANOVA deviance table with sequential Chi-square tests (Additional file 1: Figure S22).

The predictive model of ES was applied to a set of 1600 simulated genomes with varying intron densities (0.5–15 introns/gene range, 40 regular intervals) and mean intron sizes (10–8000 bp range, 40 intervals at cubic distances). Each simulated genome contained 20,000 genes of which at least 10,000 were multi-exonic (depending on its input intron density).

For each simulated genome, gene architectures were drawn from the empirical distributions derived from 10,000 randomly selected genes from each of the 30 representative eukaryotes. Specifically, we used log-normal distributions for the lengths of CDS (mean = 1422 bp), 5′ and 3′ introns (mean = input mean intron length), genes (mean = CDS length + input mean intron length × input mean intron density), and exons (mean = CDS length/number of introns per gene); and normal distributions of 5′ splice site (mean = empirical) and 3′ splice site (mean = empirical) scores. For all normal or log-normal distributions, the standard deviations were obtained from the empirical distributions. See Additional file 1: Figure S23 for a complete report of means and standard deviations for each distribution and a list of species employed. All variables were estimated using maximum-likelihood fitting of univariate distributions as implemented in the *fitdistr* utility of the R *MASS* library [84], and each set of simulated gene architectures was built using the normal or log-normal distributions implemented in R *stats* [79].

Then, from each simulated genome, we selected up to 10,000 random internal exons (using R *mosaic* library [85]) without replacement and analyzed them with the

binomial logistic regression model to obtain exon-wise ES-positive probabilities (p_{ES}). In simulated genomes with low intron densities, the number of internal exons (i.e., with two flanking introns) was sometimes less than 10,000 (for the lower bound of 0.5 introns/gene, 7864 internal exons were retrieved on average).

To estimate the incidence of ES across the spectrum of simulated genomes, we calculated the fraction of exons in each simulation with $p_{ES} > 0.522$ (the optimal probability threshold according to the ROC curve with real test data), or I_{ES}. The relationship between I_{ES}, mean intron size, and intron density was investigated using a contour map and its corresponding 3D projection, produced with the R *graphics* [79] and *akima* libraries [86], respectively.

Then, we overlaid the resulting contour map with estimations of ancestral genomes' intron density and length distributions. Ancestral intron densities were obtained from [26, 27] (Table 1). Ancestral intron length distributions were estimated using phylogenetically independent contrasts (PIC) [60, 87] as implemented in the R *ape* library [88]. Specifically, we calculated mean, median, and first and third quartiles of the ancestral intron length distributions, using PIC analysis of the descendant nodes (e.g., Urcnidaria median value corresponds to the phylogeny-controlled medians of the three extant cnidarians in our dataset; available as Additional file 1: Figure S24). In order to account for the phylogenetic relationships as required in PIC analysis [60, 87], we built a phylogenetic tree of the 65 eukaryotes in our dataset with 429 single-copy pan-eukaryotic orthologs from the BUSCO database [89] with IQ-TREE v1.5.1 [90]. For each of the 429 BUSCO orthologs and 65 organisms, we searched the best-matching protein in each predicted proteome with *hmmsearch* [91], which were aligned with MAFFT v7.245 (L-INS-i algorithm with up to 1000 refinement iterations) [92] and trimmed with the trimAL automated algorithm [93]. Then, we concatenated all 429 trimmed alignments in a multi-gene alignment (149,809 amino acid positions) that was analyzed with IQ-TREE using the LG + G4 model and a constrained phylogenetic tree as a reference (manually built from previous phylogenomic analyses [27, 94–97]; see Additional file 1: Figure S24).

Length distribution of homologous introns in holozoans and chlorophytes

To test whether *S. arctica* and *V. carteri* have lengthened or shortened their introns, we compared the length distributions of one-to-one homologous introns between them and their close relatives. First, we built two databases of orthologous genes: for unicellular holozoans (using predicted proteins of *S. arctica* and *C. fragrantissima* plus *C. owczarzaki* as outgroup) and chlorophytes (using *V. carteri* and *C. reinhardtii*, plus *Chlorella variabilis* as outgroup), using in both cases Orthofinder v2.1.2 [98] with MCL inflation = 2.1 [99]. Please note that, although *C. perkinsii* (ichthyosporean) is a closer outgroup for the holozoan analysis than *C. owczarzaki* (filasterean), the former species' intron contents are heavily reduced, whereas the latter shows a remarkable level of intron site conservation despite the high phylogenetic distance with ichthyosporeans [27].

For each database of orthologous genes, we retrieved the transcript sequences of all single-copy orthologs present in the three species, onto which we mapped the in-transcript coordinates of all annotated introns and their length (bp). For each three-gene group, we retrieved 40-bp-long transcript segments around each intron site (20 bp upstream + 20 bp downstream) using Bedtools v2.24.0 [100], and aligned them locally with *blastn* [101] to identify introns inserted in homologous transcript regions (using *-task blastn-short*). We considered as homologous introns those alignments that fulfilled these conditions: (i) alignment length > 8 bp; and (ii) alignment spanned at least 4 ungapped bp up/downstream the intron insertion (20th position). Then, we analyzed the intron length distribution of homologous introns in pairwise species comparisons.

Conservation and expression of the core spliceosomal components

We surveyed the translated proteomes of each eukaryotic species in our dataset to identify bona-fide orthologs of 82 core spliceosomal components. Specifically, we annotated the KEGG orthologous groups [102] of each species using eggNOG mapper [103, 104] and identified the KEGG orthologs (KO) corresponding to the spliceosomal snRNPs, U1, U2, U4/U6-U5 complexes, and the Prp19 complex (Additional file 1: Figure S21). The resulting matrix of KO presence/absence per species was plotted using the *heatmap.2* function of the R *gplots* v3.0.1 library [82], using the eggNOG annotation bitscore values (normalized to the 0–1 range within each KO) as a visual reference of sequence conservation. Then, we analyzed the relative level of expression of the spliceosomal components in each species. Specifically, we used a rank-based score that reflected whether the spliceosomal components were more or less expressed relative to other genes in that species' RNA-seq sample. The species-level relative rank expression of the spliceosome was calculated as follows: (i) we selected 500 random subsets of 199 genes (sampling with replacement); (ii) we sequentially added each of the 82 spliceosomal genes to the subset (totaling 200 genes per subset); (iii) each gene was assigned a rank score ranging from 1 (lowly expressed within the subset) to

200 (highly expressed); (iv) we recorded the rank score of the 82 spliceosomal genes within each random gene subset; and (v) calculated the relative rank of spliceosome expression per species by averaging the ranks of its 82 components (or those present) across all 500 random gene subsets.

Additional files

Additional file 1: Figures S1–25. Data sources, methods overview, complete reporting of statistical analyses, and replicate/technical analyses. (PDF 2453 kb)

Acknowledgements
We thank Scott W. Roy, Arnau Sebé-Pedrós, Alexandre de Mendoza, David López-Escardó, and Nick Brown for the discussion and insightful comments they provided for this study. We also thank Andrej Ondráčka and Meritxell Antó for their material contributions to obtaining RNA samples of *S. arctica* and *N. gruberi*. We also thank Guillaume Fillion and Jon Permanyer for providing and extracting RNA samples of *Trichoplax adhaerens*, respectively.

Funding
This work was supported by the following grants to MI: European Research Council (ERC) Starting Grant (agreement ERC-StG-LS2–637591, under the EU Horizon 2020 Plan) and the Spanish Ministry of Economy and Competitiveness (MINECO, agreement BFU2014–55076-P; part of the Severo Ochoa Excellence Centre plan 2013–2017, SEV-2012-0208). It was also supported by the following research grants to IRT: ERC Consolidator (ERC-2012-Co-616960), the Secretary's Office for Universities and Research of the Generalitat de Catalunya (2014 SGR 619) and MINECO (MINECO BFU2014–57779-P, with European Regional Development Fund support). XGB was supported by a pre-doctoral FPI grant from MINECO and the ERC Consolidator Grant to IRT.

Availability of data and materials
A complete list of previously published genome assemblies, annotations, and transcriptomic datasets [105–176] is available in Additional file 1: Figure S1, including accession numbers from public repositories. This includes new transcriptome sequencing data from *Naegleria gruberi*, *Sphaeroforma arctica*, and *Trichoplax adhaerens* deposited in the ENA repository, under the project codes PRJEB23822 [176], PRJEB23831 [136], and PRJEB23829 [124], respectively.

Authors' contributions
MI conceived the project, designed the experimental approach to AS identification and quantification. XGB designed the study of AS and gene architecture, built the collection of transcriptomes and genomes, and analyzed the data. MI and IRT coordinated the study and provided feedback on the analyses. MI and XGB wrote the manuscript. All authors critically reviewed and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain. [2]Departament de Genètica, Microbiologia i Estadística, Universitat de Barelona, Avinguda Diagonal 643, 08028 Barcelona, Catalonia, Spain. [3]ICREA, Passeig Lluís Companys 23, 08010 Barcelona, Catalonia, Spain. [4]Centre de Regulació Genòmica, Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain. [5]Universitat Pompeu Fabra (UPF), Plaça de la Mercè 10-12, 08002 Barcelona, Catalonia, Spain.

Received: 9 January 2018 Accepted: 1 August 2018
Published online: 17 September 2018

References
1. Breitbart RE, Andreadis A, Nadal-Ginard B. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu Rev Biochem. 1987;56:467–95. https://doi.org/10.1146/annurev.bi.56.070187.002343
2. He F, Jacobson A. Nonsense-mediated mRNA decay: degradation of defective transcripts is only part of the story. Annu Rev Genet. 2015;49:339–66. https://doi.org/10.1146/annurev-genet-112414-054639
3. Boutz PL, Bhutkar A, Sharp PA. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. 2015;29:63–80. https://doi.org/10.1101/gad.247361.114
4. Wong JJ-L, Au AYM, Ritchie W, Rasko JEJ. Intron retention in mRNA: no longer nonsense. BioEssays. 2016;38:41–9. https://doi.org/10.1002/bies.201500117
5. Brogna S, McLeod T, Petric M. The meaning of NMD: translate or perish. Trends Genet. 2016;32:395–407. https://doi.org/10.1016/j.tig.2016.04.007
6. Lykke-Andersen S, Jensen TH. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. Nat Rev Mol Cell Biol. 2015;16:665–77. https://doi.org/10.1038/nrm4063
7. Le Hir H, Nott A, Moore MJ. How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci. 2003;28:215–20. https://doi.org/10.1016/S0968-0004(03)00052-5
8. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463:457–63. https://doi.org/10.1038/nature08909
9. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 2001;17:100–7. https://doi.org/10.1016/S0168-9525(00)02176-4
10. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Philos Trans R Soc B Biol Sci. 2017;372:20150474. https://doi.org/10.1098/rstb.2015.0474
11. Gueroussov S, Gonatopoulos-Pournatzis T, Irimia M, Raj B, Lin Z-Y, Gingras A-C, et al. An alternative splicing event amplifies evolutionary differences between vertebrates. Science. 2015;349:868–73. https://doi.org/10.1126/science.aaa8381
12. Gracheva EO, Cordero-Morales JF, González-Carcacía JA, Ingolia NT, Manno C, Aranguren CI, et al. Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. Nature. 2011;476:88–91. https://doi.org/10.1038/nature10245
13. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11:345–55. https://doi.org/10.1038/nrg2776
14. McGuire AM, Pearson MD, Neafsey DE, Galagan JE. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol. 2008;9:R50. https://doi.org/10.1186/gb-2008-9-3-r50
15. Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 2007;35:125–31. https://doi.org/10.1093/nar/gkl924
16. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24:1774–86. https://doi.org/10.1101/gr.177790.114
17. Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017:1–10. https://doi.org/10.1101/gr.220962.117
18. Irimia M, Roy SW. Origin of Spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol. 2014;6 https://doi.org/10.1101/cshperspect.a016071

19. Blencowe BJ. The relationship between alternative splicing and proteomic complexity. Trends Biochem Sci. 2017;42:407–8. https://doi.org/10.1016/j.tibs.2017.04.001
20. Tress ML, Abascal F, Valencia A. Alternative splicing may not be the key to proteome complexity. Trends Biochem Sci. 2017;42:98–110. https://doi.org/10.1016/j.tibs.2016.08.008
21. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. Gene. 2013;514:1–30. https://doi.org/10.1016/j.gene.2012.07.083
22. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread expansion of protein interaction capabilities by alternative splicing. Cell. 2016;164:805–17. https://doi.org/10.1016/j.cell.2016.01.029
23. Ellis JD, Barrios-Rodiles M, Çolak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell. 2012;46:884–92. https://doi.org/10.1016/j.molcel.2012.05.037
24. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell. 2012;46:871–83. https://doi.org/10.1016/j.molcel.2012.05.039
25. Boothby TC, Zipper RS, Van der Weele CM, Wolniak SM. Removal of retained introns regulates translation in the rapidly developing gametophyte of Marsilea vestita. Dev Cell. 2013;24:517–29. https://doi.org/10.1016/j.devcel.2013.01.015
26. Csűrös M, Rogozin IB, Koonin EV. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput Biol. 2011;7:e1002150. https://doi.org/10.1371/journal.pcbi.1002150
27. Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, et al. Dynamics of genomic innovation in the unicellular ancestry of animals. elife. 2017;6 https://doi.org/10.7554/eLife.26036
28. Irimia M, Penny D, Roy SW. Coevolution of genomic intron number and splice sites. Trends Genet. 2007;23:321–5. https://doi.org/10.1016/j.tig.2007.04.001
29. Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 2008;18:88–103. https://doi.org/10.1101/gr.6818908
30. Irimia M, Roy SW. Evolutionary convergence on highly-conserved 3′ intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. PLoS Genet. 2008;4:e1000148. https://doi.org/10.1371/journal.pgen.1000148
31. Collins L, Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 2005;22:1053–66. https://doi.org/10.1093/molbev/msi091
32. Plass M, Agirre E, Reyes D, Camara F, Eyras E. Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. 2008;24:590–4. https://doi.org/10.1016/j.tig.2008.10.004
33. Roy SW, Irimia M. Splicing in the eukaryotic ancestor: form, function and dysfunction. Trends Ecol Evol. 2009;24:447–55. https://doi.org/10.1016/j.tree.2009.04.005
34. Koonin EV, Csuros M, Rogozin IB. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdiscip Rev RNA. 2013;4:93–105. https://doi.org/10.1002/wrna.1143
35. Sorber K, Dimon MT, DeRisi JL. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res. 2011;39:3820–35. https://doi.org/10.1093/nar/gkq1223
36. Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature. 2012;492:59–65. https://doi.org/10.1038/nature11681
37. Zhang C, Yang H, Yang H. Evolutionary character of alternative splicing in plants. Bioinform Biol Insights. 2015;9:47–52. https://doi.org/10.4137/BBI.S33716
38. Kianianmomeni A, Ong C, Rätsch G, Hallmann A. Genome-wide analysis of alternative splicing in *Volvox carteri*. BMC Genomics. 2014;15:1117. https://doi.org/10.1186/1471-2164-15-1117
39. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol Biol Evol. 2014;31:1402–13. https://doi.org/10.1093/molbev/msu083
40. Sebé-Pedrós A, Irimia M, Del Campo J, Parra-Acero H, Russ C, Nusbaum C, et al. Regulated aggregative multicellularity in a close unicellular relative of metazoa. elife. 2013;2:e01287. https://doi.org/10.7554/eLife.01287
41. de Mendoza A, Suga H, Permanyer J, Irimia M, Ruiz-Trillo I. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. elife. 2015;4:7250–7. https://doi.org/10.7554/eLife.08904
42. Suzuki S, Ishida KI, Hirakawa Y. Diurnal transcriptional regulation of endosymbiotically derived genes in the chlorarachniophyte *Bigelowiella natans*. Genome Biol Evol. 2016;8:2672–82. https://doi.org/10.1093/gbe/evw188
43. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12 https://doi.org/10.1371/journal.pbio.1001889
44. Muñoz MJ, Nieto Moreno N, Giono LE, Cambindo Botto AE, Dujardin G, Bastianello G, et al. Major roles for pyrimidine dimers, nucleotide excision repair, and ATR in the alternative splicing response to UV irradiation. Cell Rep. 2017;18:2868–79. https://doi.org/10.1016/j.celrep.2017.02.066
45. Shalgi R, Hurt JA, Lindquist S, Burge CB. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. Cell Rep. 2014;7:1362–70. https://doi.org/10.1016/j.celrep.2014.04.044
46. Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L. Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in *Arabidopsis*. BMC Genomics. 2014;15:1–14. https://doi.org/10.1186/1471-2164-15-431
47. Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? Trends Genet. 2004;20:68–71. https://doi.org/10.1016/j.tig.2003.12.004
48. Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW. Widespread evolutionary conservation of alternatively spliced exons in *Caenorhabditis*. Mol Biol Evol. 2008;25:375–82. https://doi.org/10.1093/molbev/msm262
49. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, et al. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol. 2017;18:208. https://doi.org/10.1186/s13059-017-1344-6
50. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip Rev RNA. 2013;4:49–60. https://doi.org/10.1002/wrna.1140
51. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338:1587–93. https://doi.org/10.1126/science.1230612
52. Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. 2012;1:543–56. https://doi.org/10.1016/j.celrep.2012.03.013
53. Hollander D, Naftelberg S, Lev-Maor G, Kornblihtt AR, Ast G. How are short exons flanked by long introns defined and committed to splicing? Trends Genet. 2016;32:596–606. https://doi.org/10.1016/j.tig.2016.07.003
54. Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ. An alternative-exon database and its statistical analysis. DNA Cell Biol. 2000;19:739–56. https://doi.org/10.1089/104454900750058107
55. Baek D, Green P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc Natl Acad Sci U S A. 2005;102:12813–8. https://doi.org/10.1073/pnas.0506139102
56. Rukov JL, Irimia M, Mørk S, Lund VK, Vinther J, Arctander P. High qualitative and quantitative conservation of alternative splicing in *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Mol Biol Evol. 2007;24:909–17. https://doi.org/10.1093/molbev/msm023
57. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B Biol Sci. 2015;370:20140331. https://doi.org/10.1098/rstb.2014.0331
58. Munding EM, Shiue L, Katzman S, Donohue JP, Ares M. Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing. Mol Cell. 2013;51:338–48. https://doi.org/10.1016/j.molcel.2013.06.012
59. Fu XD, Ares M. Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet. 2014;15:689–701. https://doi.org/10.1038/nrg3778

60. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985;125:1–15. https://doi.org/10.1086/284325
61. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, et al. The dynamic genome of Hydra. Nature. 2010;464:592–6. https://doi.org/10.1038/nature08830
62. Zacharias H, Anokhin B, Khalturin K, Bosch TCG. Genome sizes and chromosomes in the basal metazoan Hydra. Zoology. 2004;107:219–27. https://doi.org/10.1016/j.zool.2004.04.005
63. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 2007;17:1034–44. https://doi.org/10.1101/gr.6438607
64. Rogozin IB, Carmel L, Csűrös M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct. 2012;7:28. https://doi.org/10.1186/1745-6150-7-11
65. Li W, Kuzoff R, Wong CK, Tucker A, Lynch M. Characterization of newly gained introns in *Daphnia* populations. Genome Biol Evol. 2014;6:2218–34. https://doi.org/10.1093/gbe/evu174
66. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science. 2009;324:268–72. https://doi.org/10.1126/science.1167222
67. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. Nature. 2016;538:533–6. https://doi.org/10.1038/nature20110
68. van der Burgt A, Severing E, de Wit PJGM, Collemare J. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. Curr Biol. 2012;22:1260–5. https://doi.org/10.1016/j.cub.2012.05.011
69. Simakov O, Kawashima T. Independent evolution of genomic characters during major metazoan transitions. Dev Biol. 2016:0–1. https://doi.org/10.1016/j.ydbio.2016.11.012
70. Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, et al. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun. 2013;4:2325. https://doi.org/10.1038/ncomms3325
71. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature. 2008;451:783–8. https://doi.org/10.1038/nature06617
72. Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 2013;14:R15. https://doi.org/10.1186/gb-2013-14-2-r15
73. Irimia M, Tena JJ, Alexis MS, Fernandez-Minan A, Maeso I, Bogdanovic O, et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. 2012;22:2356–67. https://doi.org/10.1101/gr.139725.112
74. Ezkurdia I, Rodriguez JM, Carrillo-De Santa Pau E, Vázquez J, Valencia A, Tress ML. Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res. 2015;14:1880–7. https://doi.org/10.1021/pr501286b
75. Gordon A. FASTX Toolkit. 2017. http://hannonlab.cshl.edu/fastx_toolkit/
76. Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DDR, et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. Stem Cells. 2012;30:1734–45. https://doi.org/10.1002/stem.1144
77. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25. https://doi.org/10.1186/gb-2009-10-3-r25
78. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014;12:59–60. https://doi.org/10.1038/nmeth.3176
79. R Core Team. R: a language and environment for statistical computing. Vienna: R Core Team; 2017. https://www.r-project.org/
80. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: String objects representing biological sequences, and matchingalgorithms. R package version 2.44.2. 2017.
81. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11:377–94. https://doi.org/10.1089/1066527041410418
82. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: various R programming tools for plotting data. 2016.
83. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. https://doi.org/10.1186/1471-2105-12-77
84. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer; 2002. https://doi.org/10.1007/978-0-387-21706-2
85. Pruim R, Kaplan DT, Horton NJ. The mosaic package: helping students to "think with data" using R. R J. 2017;9:77–102.
86. Akima H, Gebhardt A. akima: interpolation of irregularly and regularly spaced data. 2016.
87. Theodore Garland, Paul H. Harvey, Anthony R. Ives; Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts, Systematic Biology. 1992;41(1):18–32. https://doi.org/10.1093/sysbio/41.1.18.
88. Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90. https://doi.org/10.1093/bioinformatics/btg412
89. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351
90. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74. https://doi.org/10.1093/molbev/msu300
91. HMMER. HMMER. 2015. http://hmmer.org/.
92. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010
93. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3. https://doi.org/10.1093/bioinformatics/btp348
94. Torruella G, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, Del Campo J, et al. Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and Fungi. Curr Biol. 2015;25:1–7. https://doi.org/10.1016/j.cub.2015.07.053
95. Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, et al. A large and consistent Phylogenomic dataset supports sponges as the sister group to all other animals. Curr Biol. 2017:1–10. https://doi.org/10.1016/j.cub.2017.02.031
96. Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc R Soc B Biol Sci. 2016;283:20152802. https://doi.org/10.1098/rspb.2015.2802
97. He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SLL. An alternative root for the eukaryote tree of life. Curr Biol. 2014;24:465–70. https://doi.org/10.1016/j.cub.2014.01.036
98. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157. BioMed Central. https://doi.org/10.1186/s13059-015-0721-2
99. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84. https://doi.org/10.1093/nar/30.7.1575
100. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033
101. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
102. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092
103. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through Orthology assignment by eggNOG-mapper. Mol Biol Evol. 2017;34:2115–22. https://doi.org/10.1093/molbev/msx148
104. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93. https://doi.org/10.1093/nar/gkv1248
105. *Homo sapiens* transcriptome (Hsap). NCBI Sequence Read Archive SRP007412. https://www.ncbi.nlm.nih.gov/sra/SRP007412
106. *Homo sapiens* transcriptome (Hsax). NCBI Sequence Read Archive SRP056969. https://www.ncbi.nlm.nih.gov/sra/SRP056969
107. *Mus musculus* transcriptome (Mmus). NCBI Sequence Read Archive SRP007412. https://www.ncbi.nlm.nih.gov/sra/SRP007412

108. *Mus musculus* transcriptome (Mmux). NCBI Sequence Read Archive SRP015997. https://www.ncbi.nlm.nih.gov/sra/SRP015997
109. *Monodelphis domestica* transcriptome (Mdom). NCBI Sequence Read Archive SRP007412. https://www.ncbi.nlm.nih.gov/sra/SRP007412
110. *Ornithorhynchus anatinus* transcriptome (Oana). NCBI Sequence Read Archive SRP007412. https://www.ncbi.nlm.nih.gov/sra/SRP007412
111. *Gallus gallus* transcriptome (Ggal). NCBI Sequence Read Archive SRP007412. https://www.ncbi.nlm.nih.gov/sra/SRP007412
112. *Danio rerio* transcriptome (Drex). NCBI Sequence Read Archive SRP048807. https://www.ncbi.nlm.nih.gov/sra/SRP048807
113. *Xenopus tropicalis* transcriptome (Xtro). NCBI Sequence Read Archive SRP012375. https://www.ncbi.nlm.nih.gov/sra/SRP012375
114. *Xenopus tropicalis* transcriptome (Xtrx). NCBI Sequence Read Archive SRP015997. https://www.ncbi.nlm.nih.gov/sra/SRP015997
115. *Ciona intestinalis* transcriptome (Cint). NCBI Sequence Read Archive SRP042651. https://www.ncbi.nlm.nih.gov/sra/SRP042651
116. *Branchiostoma belcheri* transcriptome (Brabel). NCBI Sequence Read Archive SRP025148. https://www.ncbi.nlm.nih.gov/sra/SRP025148
117. *Strongylocentrotus purpuratus* transcriptome (Spur2). NCBI GEO DataSet GSE97267. https://www.ncbi.nlm.nih.gov/gds/?term=GSE97267
118. *Drosophila melanogaster* transcriptome (Dmel). NCBI Sequence Read Archive SRP001696. https://www.ncbi.nlm.nih.gov/sra/SRP001696
119. *Drosophila melanogaster* transcriptome (Dmel2). NCBI Sequence Read Archive SRP082392. https://www.ncbi.nlm.nih.gov/sra/SRP082392
120. *Caenorhabditis elegans* transcriptome (Cele). NCBI Sequence Read Archive SRP000401. https://www.ncbi.nlm.nih.gov/sra/SRP000401
121. *Crassostrea gigas* transcriptome (Cgig). NCBI Sequence Read Archive SRP014559. https://www.ncbi.nlm.nih.gov/sra/SRP014559
122. *Capitella teleta* transcriptome (Ctel). NCBI Sequence Read Archive SRP102138. https://www.ncbi.nlm.nih.gov/sra/SRP102138
123. *Schistosoma mansoni* transcriptome (Sman). NCBI Sequence Read Archive ERP000427. https://www.ncbi.nlm.nih.gov/sra/ERP000427
124. *Trichoplax adhaerens* transcriptome (Tadh). NCBI Bioproject PRJEB23829. https://www.ncbi.nlm.nih.gov/bioproject/PRJEB23829
125. *Hydra magnipapillata* transcriptome (Hmag). NCBI Sequence Read Archive SRP051110. https://www.ncbi.nlm.nih.gov/sra/SRP051110
126. *Nematostella vectensis* transcriptome (Nvex). NCBI Sequence Read Archive SRP021895. https://www.ncbi.nlm.nih.gov/sra/SRP021895
127. *Nematostella vectensis* transcriptome (Nvec). NCBI Sequence Read Archive SRP018739. https://www.ncbi.nlm.nih.gov/sra/SRP018739
128. *Aiptasia* sp. transcriptome (Aipt). NCBI Sequence Read Archive SRP047443. https://www.ncbi.nlm.nih.gov/sra/SRP047443
129. *Mnemiopsis leidyi* transcriptome (Mlei). NCBI Sequence Read Archive SRP014828. https://www.ncbi.nlm.nih.gov/sra/SRP014828
130. *Amphimedon queenslandica* transcriptome (Aque). NCBI Sequence Read Archive SRR1511618. https://www.ncbi.nlm.nih.gov/sra/SRR1511618
131. *Oscarella carmela* transcriptome (Ocar). NCBI Sequence Read Archive SRR1042012. https://www.ncbi.nlm.nih.gov/sra/SRR1042012
132. *Sycon ciliatum* transcriptome (Scil). NCBI Sequence Read Archive ERP005418. https://www.ncbi.nlm.nih.gov/sra/ERP005418
133. *Salpingoeca rosetta* transcriptome (Sros). NCBI Sequence Read Archive SRP005692. https://www.ncbi.nlm.nih.gov/sra/SRP005692
134. *Capsaspora owczarzaki* transcriptome (Cowc). NCBI Sequence Read Archive SRP022579. https://www.ncbi.nlm.nih.gov/sra/SRP022579
135. *Creolimax fragrantissima* transcriptome (Cfra). NCBI Sequence Read Archive SRP058061. https://www.ncbi.nlm.nih.gov/sra/SRP058061
136. *Sphaeroforma arctica* transcriptome (Sar3). NCBI Bioproject PRJEB23831. https://www.ncbi.nlm.nih.gov/bioproject/PRJEB23831
137. *Chromosphaera perkinsii* transcriptome (Cper). NCBI Sequence Read Archive SRP097609. https://www.ncbi.nlm.nih.gov/sra/SRP097609
138. *Neurospora crassa* transcriptome (Ncrx). NCBI Sequence Read Archive SRP016065. https://www.ncbi.nlm.nih.gov/sra/SRP016065
139. *Schizosaccharomyces pombe* transcriptome (Spom). NCBI Sequence Read Archive ERP001483. https://www.ncbi.nlm.nih.gov/sra/ERP001483
140. *Aspergillus oryzae* transcriptome (Aory). NCBI Sequence Read Archive SRP016952. https://www.ncbi.nlm.nih.gov/sra/SRP016952
141. *Cryptococcus neoformans* transcriptome (Cneo). NCBI Sequence Read Archive SRR847297. https://www.ncbi.nlm.nih.gov/sra/SRR847297
142. *Ustilago maydis* transcriptome (Umay). NCBI Sequence Read Archive ERP001905. https://www.ncbi.nlm.nih.gov/sra/ERP001905
143. *Tuber melanosporum* transcriptome (Tmel). NCBI Sequence Read Archive SRP028655. https://www.ncbi.nlm.nih.gov/sra/SRP028655
144. *Rhizopus oryzae* transcriptome (Rory). NCBI Sequence Read Archive SRP031602. https://www.ncbi.nlm.nih.gov/sra/SRP031602
145. *Allomyces macrogynus* transcriptome (Amac). NCBI Sequence Read Archive SRP022576. https://www.ncbi.nlm.nih.gov/sra/SRP022576
146. *Spizellomyces punctatus* transcriptome (Spun). NCBI Sequence Read Archive SRR343043. https://www.ncbi.nlm.nih.gov/sra/SRR343043
147. *Rhizophagus irregularis* DAOM 181602 transcriptome (Rirr). NCBI Sequence Read Archive DRP002784. https://www.ncbi.nlm.nih.gov/sra/DRP002784
148. *Conidiobolus coronatus* transcriptome (Ccor). NCBI Sequence Read Archive SRR427173. https://www.ncbi.nlm.nih.gov/sra/SRR427173
149. *Gonapodya prolifera* transcriptome (Gpro). NCBI Sequence Read Archive SRR427152. https://www.ncbi.nlm.nih.gov/sra/SRR427152
150. *Fonticula alba* transcriptome (Falb). NCBI Sequence Read Archive SRP022580. https://www.ncbi.nlm.nih.gov/sra/SRP022580
151. *Dictyostelium discoideum* AX4 transcriptome (Ddis2). NCBI Sequence Read Archive . https://www.ncbi.nlm.nih.gov/sra/SRP060392
152. *Polysphondylium pallidum* transcriptome (Ppal). NCBI Sequence Read Archive SRP004023. https://www.ncbi.nlm.nih.gov/sra/SRP004023
153. *Acanthamoeba castellanii* transcriptome (Acas). NCBI Sequence Read Archive SRP028620. https://www.ncbi.nlm.nih.gov/sra/SRP028620
154. *Arabidopsis thaliana* transcriptome (Atha). NCBI Sequence Read Archive SRP052858. https://www.ncbi.nlm.nih.gov/sra/SRP052858
155. *Arabidopsis thaliana* transcriptome (Atha2). NCBI Sequence Read Archive SRP074840. https://www.ncbi.nlm.nih.gov/sra/SRP074840
156. *Vitis vinifera* transcriptome (Vvin). NCBI Sequence Read Archive SRP065417. https://www.ncbi.nlm.nih.gov/sra/SRP065417
157. *Mimulus guttatus* transcriptome (Mgut). NCBI Sequence Read Archive SRP045683. https://www.ncbi.nlm.nih.gov/sra/SRP045683
158. *Oryza sativa* transcriptome (Osat2). NCBI Sequence Read Archive DRP000315. https://www.ncbi.nlm.nih.gov/sra/DRP000315
159. *Physcomitrella patens* transcriptome (Ppat). NCBI Sequence Read Archive SRP011279. https://www.ncbi.nlm.nih.gov/sra/SRP011279
160. *Selaginella moellendorffii* transcriptome (Smoe). NCBI Sequence Read Archive SRP059539. https://www.ncbi.nlm.nih.gov/sra/SRP059539
161. *Klebsormidium netis* (formerly flaccidum) transcriptome (Kfla). NCBI Sequence Read Archive SRP048567. https://www.ncbi.nlm.nih.gov/sra/SRP048567
162. *Volvox carteri* transcriptome (Vcar). NCBI Sequence Read Archive SRP066714. https://www.ncbi.nlm.nih.gov/sra/SRP066714
163. *Chlamydomonas reinhardtii* transcriptome (Crei). NCBI Sequence Read Archive ERP001997. https://www.ncbi.nlm.nih.gov/sra/ERP001997
164. *Micromonas pusilla* transcriptome (Mpus). NCBI Sequence Read Archive SRR847305. https://www.ncbi.nlm.nih.gov/sra/SRR847305
165. *Cyanophora paradoxa* transcriptome (Cpar). NCBI Sequence Read Archive SRR363339. https://www.ncbi.nlm.nih.gov/sra/SRR363339
166. *Ectocarpus siliculosus* transcriptome (Esil). NCBI Sequence Read Archive SRP037532. https://www.ncbi.nlm.nih.gov/sra/SRP037532
167. *Phytophthora infestans* transcriptome (Pinf). NCBI Sequence Read Archive SRR1640225. https://www.ncbi.nlm.nih.gov/sra/SRR1640225
168. *Aureococcus anophagefferens* transcriptome (Aano). NCBI Sequence Read Archive SRP045642. https://www.ncbi.nlm.nih.gov/sra/SRP045642
169. *Plasmodium falciparum* 3D7 transcriptome (Pfal2). NCBI Sequence Read Archive SRP069075. https://www.ncbi.nlm.nih.gov/sra/SRP069075
170. *Tetrahymena thermophila* transcriptome (Tthe). NCBI Sequence Read Archive SRP016619. https://www.ncbi.nlm.nih.gov/sra/SRP016619
171. *Bigelowiella natans* CCMP2755 transcriptome (BnatMEA). NCBI Sequence Read Archive MMETSP. https://www.ncbi.nlm.nih.gov/sra/SRP042159
172. *Bigelowiella natans* CCMP2755 transcriptome (Bnao). NCBI Sequence Read Archive Seq Bioproject PRJNA47111. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA47111
173. *Bigelowiella natans* CCMP2755 transcriptome (Bnat). NCBI Sequence Read Archive DRP003230. https://www.ncbi.nlm.nih.gov/sra/DRP003230
174. *Emiliania huxleyi* transcriptome (Ehux). NCBI Sequence Read Archive SRR847300. https://www.ncbi.nlm.nih.gov/sra/SRR847300
175. *Guillardia theta* transcriptome (Gthe). NCBI Sequence Read Archive SRR747855. https://www.ncbi.nlm.nih.gov/sra/SRR747855
176. *Naegleria gruberi* transcriptome (Ngru). NCBI Bioproject PRJEB23822. https://www.ncbi.nlm.nih.gov/bioproject/PRJEB23822