Computational and Systems Biology
Evolutionary Biology

Evolutionary trends of alternative splicing

Reviewed Preprint
v1 • June 12, 2024
Not revised

Rebeca de la Fuente ✉, **Wladimiro Díaz-Villanueva, Vicente Arnau, Andrés Moya**

Austral Center for Scientific Research (CADIC-CONICET), 9410 Ushuaia, Argentina; • Institute of Integrative Systems Biology (I2Sysbio), University of Valencia and Spanish National Research Council (CSIC), 46980 Valencia, Spain • Foundation for the Promotion of Sanitary and Biomedical Research of the Valencian Community (FISABIO), 46020 Valencia, Spain • Center for Biomedical Research in Epidemiology and Public Health Network (CIBEResp), 28029 Madrid, Spain

Abstract

Alternative splicing is the post-transcriptional mechanism by which many different isoforms are generated from a single gene, contributing to increasing spatio-temporal transcriptome complexity. We propose a novel genome-level measure of alternative splicing, which associates it with a single value for each species. Thus, a comparative analysis of species spanning the whole tree of life has revealed certain evolutionary trends in alternative splicing, prevalence in specific lineages, and relation to genome compositional structures.

eLife assessment

The authors examined whether the frequency of alternative splicing across entire genomes correlates with measures of complexities across prokaryotes and eukaryotes. Although the question is very interesting and **important** for our general understanding of the evolution of life forms, the work is **inadequate**: the methods, data, and analyses do not support the primary claims. The measure of alternative splicing frequency used by the authors is problematic; the method is inappropriate; the observed correlations may also be explained by known population genetics principles; and parts of the manuscript are difficult to understand.

https://doi.org/10.7554/eLife.94802.1.sa2

Introduction

In evolutionary biology, the classification of organisms in terms of biological complexity is a fundamental question (***McShea, 1996***). Genome size and the amount of protein-coding DNA were initially thought to reflect complexity. However, cumulative evidence of the complex relationships in the genotype-phenotype mapping would indicate that the *one gene-one protein* paradigm is no longer supported and, therefore, genome composition alone is insufficient to quantify organismal complexity (***Claverie, 2001***; ***Choi et al., 2020***). On one hand, genome sizes

display high variability in eukaryotes, even for closely related species (***Bennett and Leitch, 2005***; ***Gregory, 2005***). On the other hand, a lower-than-expected number of protein-coding genes has been observed in some eukaryotic lineages, a problem known as the G-value paradox (***Hahn and Wray, 2002***). In humans, it is estimated that 24,000 genes encode a total of 100,000 different proteins (***Lander et al., 2001***).

Transcript diversification is driven by two main mechanisms: gene duplication (***Lynch and Conery, 2000***; ***Holland et al., 2017***) and alternative splicing (***Bush et al., 2017***). Although both mechanisms contribute to phenotype innovation (***Su et al., 2006***), alternative splicing has been cited as the primary candidate to resolve the G-value paradox (***Chen et al., 2014***, ***2012***). It consists of a post-transcriptional mechanism by which different protein isoforms are generated from a single gene by the selective combination of exons and introns (***Modrek and Lee, 2002***; ***Keren et al., 2010***). Thus, it contributes to increasing biological complexity by the differentiation of entities at different levels of the biological organization, including transcript diversification (***Chen et al., 2012***; ***Nilsen and Graveley, 2010***), cell differentiation (***Fiszbein and Kornblihtt, 2017***; ***Fu et al., 2009***) and speciation events, such as morphs (***Steward et al., 2022***; ***Grantham and Brisson, 2018***), castes (***Lyko et al., 2011***) or even subspecies (***Harr and Turner, 2010***).

Alternative splicing profiles are conserved in a species-specific way (***Barbosa-Morais et al., 2012***). The percentage of alternatively spliced genes is higher in vertebrates than in invertebrates, and only a few examples have been found in bacteria and archaea (***Anantharaman et al., 2002***; ***Kim et al., 2006***; ***Schad et al., 2011***). By contrast, at least 70% of multi-exon genes undergo alternative splicing in humans, mainly regulated in a tissue-specific way (***Pan et al., 2008***). However, there are still some unanswered questions regarding the roles played by alternative splicing during genome evolution, its prevalence in the different lineages, and its relation to genome composition. In the present work, we propose a novel measure, namely, the alternative splicing ratio, which quantifies the number of different isoforms that can be transcribed using the same amount of information encoded in the genome and compute it for species spanning the whole tree of life. Accordingly, we perform a comparative analysis of the alternative splicing ratio and identify relations to the genome composition.

Results

Genome composition

A comparison of the genome composition for species spanning the whole tree of life has enabled us to distinguish certain genomic traits characterizing the major evolutionary transitions. This differentiates prokaryotes, unicellular eukaryotes, and the high taxonomic groups of multicellular life forms. Genome, genes, and coding sizes increase proportionally to each other at a lineage-specific level (see ***Figure 1—figure Supplement 1***,***Figure 1—figure Supplement 2***, and ***Figure 1—figure Supplement 3***). Specifically, we observe a significant correlation between each pair of genome variables in all cases, except for birds, where the genome size is weakly correlated to gene content and coding DNA. Furthermore, the slopes given by the linear regressions differ for each taxonomic group, which suggests key differences in their genome evolution. ***Figure 1*** A-C shows the relationship between each pair of genome variables, differentiated by each taxonomic group. Because genome variables are related to each other in an embedding structure, we show in ***Figure 1*** D the relation between the percentage of coding composing genes and the respective percentage of gene content in genomes. Complementary to this information, we quantify some statistics characterizing the genome composition found in our data. These results are illustrated in ***Table 1*** and Supplementary file 1.

Figure 1.

Relation between each pair of genome variables: (A) Genome size and coding, (B) Genome size and gene content, (C) Gene content and coding. Note the logarithmic scale on both axes. Insets show the linear relation for all multicellular groups. (D) Relation between the percentage of coding within genes and genes within genomes. Colors represent the different taxonomic groups under study.

Figure 1—figure supplement 1. Relation between genome size and gene content for each taxonomical group.

Figure 1—figure supplement 2. Relation between genome size and coding DNA for each taxonomical group.

Figure 1—figure supplement 3. Relation between gene content and coding DNA for each taxonomical group.

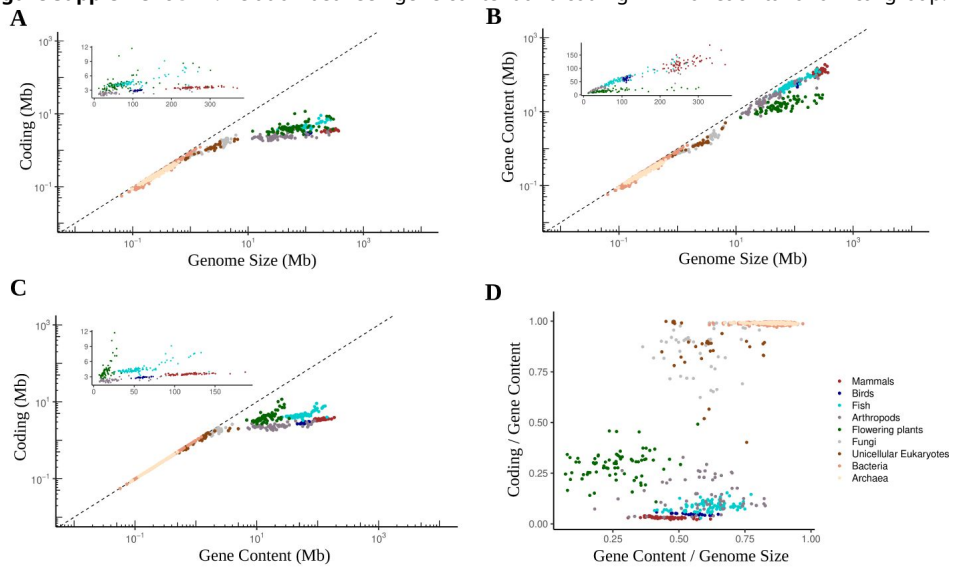


Table 1.

Average, standard deviation and the range (i.e, the minimum and maximum values) of the relative amounts of gene content and coding DNA within genomes, and coding DNA within genes. Each row corresponds to a given taxonomical group: prokaryotes (n=684), unicellular eukaryotes (n=34), fungi (n=69), flowering plants (n=75), arthropods (n=61), fish (n=88), birds (n=26), and mammals (n=70).

	Gene / Genome (%)			Coding / Genome (%)			Coding / Genes (%)		
Group	$\overline{x}$	s_x	range	$\overline{x}$	s_x	range	$\overline{x}$	s_x	range
Prokaryotes	85.8	5.0	[61.5,96.9]	84.9	5.0	[60.4,95.7]	98.9	0.4	[97.0,99.7]
Unicellular Euk.	60.9	11.8	[43.5,82.4]	52.2	11.8	[30.4,74.9]	86.4	13.4	[40.2,99.9]
Fungi	64.7	13.6	[36.3,90.8]	58.6	15.9	[31.7,89.6]	90.0	11.2	[62.4,99.5]
Flowering Plants	26.4	10.7	[7.0,57.1]	7.8	4.3	[1.3,28.1]	29.0	7.0	[10.8,49.2]
Arthropods	58.4	14.0	[18.1,82.7]	8.3	4.7	[1.0,17.8]	14.1	7.7	[3.2,32.8]
Fish	61.1	8.1	[40.7,77.3]	5.3	1.7	[1.6,11.3]	8.6	2.2	[2.8,15.3]
Birds	55.3	5.6	[41.5,64.9]	2.6	0.2	[2.4,3.1]	4.8	0.4	[4.1,5.8]
Mammals	45.2	5.7	[30.8,62.0]	1.3	0.1	[0.9,1.7]	3.0	0.4	[2.0,4.0]

In the case of prokaryotes we observe a one-to-one relation among all three genome factors, where at least 97% of the gene content is composed of protein-coding sequences. Thus, genomes of prokaryotes are formed by consecutive single-exon genes, i.e., genes containing only one exon, which encode a unique protein. Despite this result, ***Table 1*** highlights the starting accumulation of non-coding DNA in prokaryotes as intergenic DNA. In particular, we have estimated that genes compose between 61% and 97% of genomes. However, this amount of non-coding DNA is still negligible if compared to eukaryotes, which start to accumulate large amounts of non-coding DNA, both as intergenic DNA and as intronic sequences. In particular, the drastic accumulation of non-coding DNA in eukaryotes is reflected in the deviations from equality observed in ***Figure 1*** A-C, which concurs with a genome structure diversification.

Figure 1 D shows that the percentage of coding composing genes is the factor that best differentiates single-celled organisms from multicellular ones. It represents between 40% and 99.9% of gene content in unicellular eukaryotes, with typical percentages of about 86.4%. This percentage is much higher than in most multicellular life forms, where coding comprises less than 50% of genes. Specifically, flowering plants have a typical percentage of about 29%, and it decreases gradually for arthropods (14.1%), fish (8.6%), birds (4.8%), and finally mammals, where it comprises only 3%. This result indicates that gene composition of unicellular eukaryotes is closer to that of prokaryotes than multicellular life forms. On the other hand, the percentage of intergenic DNA differs considerably among plants and animals. While it accounts for between 43% and 97% of flowering plants, it drops dramatically in animals, with a range comparable to that of unicellular eukaryotes. Specifically, intergenic regions compose between 17% and 82% of arthropods, 23% and 60% of fish, 35% and 59% of birds, and 38% and 69% of mammals. This result highlights the high accumulation of intergenic DNA in plants, while it maintains at moderate values in animals.

Correlations to Alternative Splicing

With some exceptions, alternative splicing is unrelated to genome factors (see ***Figure 2—figure Supplement 1***, ***Figure 2—figure Supplement 2***, and ***Figure 2—figure Supplement 3***). In the case of genome sizes, only arthropods show a significant correlation, but it is still too weak to draw meaningful conclusions. Conversely, mammals and birds have alternative splicing values positively correlated with the amount of coding and gene content. In particular, we find that alternative splicing and the amount of coding have a correlation coefficient of $r = 0.531$ in mammals and $r = 0.541$ in birds. In contrast, the correlation with gene content is $r = 0.711$ in mammals and $r = 0.786$ in birds. As we can observe, this correlation becomes very strong for the latter case; therefore, as there is more gene content in the genome of land animals, higher ratios of alternative splicing are detected. According with the definition of the alternative splicing ratio, an increase in the amount of gene composition is not associated to a greater production of protein isoforms, but rather with an increase of transcript diversification. Furthermore, the percentage of coding composing genes appears to be highly correlated to alternative splicing values in mammals ($r = -0.631$) and birds ($r = -0.791$), with a negative correlation in both cases (see ***Figure 2—figure Supplement 5***), so these results are in concordance with a key role of intronic sequences in the generation of novel protein isoforms. It is important to recall that these associations are only significant in mammals and birds, which suggests that only sophisticated organisms optimize alternative splicing by increasing intronic sequences within the intron-exon structure. Also, mammals are a unique taxonomic group with alternative splicing highly correlated to both, the number of genes and coding. This fact may indicate a highly conserved gene structure coupled to alternative splicing in highly evolved groups.

The percentage of coding relative to genome sizes does not correlate to alternative splicing for any taxonomical group (See ***Figure 2—figure Supplement 4***). Regarding the proportion of genes composing genomes, we find a lack of correlation with alternative splicing ratios in plants but a high correlation in all groups of animals. As we observe in ***Figure 2***, the corresponding

correlation coefficients are r = 0.716 for mammals, r = 0.823 for birds, r = 0.604 for fish, and r = 0.712 for arthropods. This result suggests that alternative splicing may be coupled to a genome-level structure in animals, with increasing activity for gene-rich genomes.

Evolutionary trends

The average values of the alternative splicing ratio for each taxonomic group are shown in ***Table 2***, together with the standard deviation, the coefficient of variation and the range. As we can observe, the alternative splicing ratio is close to one in all single-celled organisms and increases gradually in flowering plants, arthropods, fish, birds, and mammals. In concordance with its role in the differentiation process, this result suggests that it may only be a relevant mechanism in the genome evolution of multicellularity, especially in highly evolved groups such as mammals. In contrast, the variation coefficients in animals increase with increasing alternative splicing ratios, which suggests that alternative splicing may be coupled to the expected variability appearing throughout the tree-like pattern of the evolution of life. This is also in concordance with the appearance of different modes of alternative splicing with evolutionary time. In the case of plants, we observe a low coefficient of variation, which suggests there is a more constrained mechanism of alternative splicing in plants.

The relation of alternative splicing to genome composition is also analyzed. First, the percentage of intergenic DNA and its relation to alternative splicing for the different taxonomic groups is shown in ***Figure 3***A. While prokaryotes contain less than 30% of intergenic DNA, unicellular eukaryotes and fungi show a percentage that falls within a similar range to that of animals, from 17% to 57%. The absence of alternative splicing characterizes both single-celled groups. At the other end of the spectrum we find plants with a high prevalence of intergenic DNA, in between 50% and 96% of genomes, and with alternative splicing ratios of about 1.554. This result suggests that, while alternative splicing is maintained at moderated values in plants, genome evolution occurs by increasing the intergenic content. A tendency to increase alternative splicing values is only observed in animals, and it is optimized in organisms where the proportion of intergenic DNA represents about 50% of the genome. This result concurs with the constraints imposed by the eukaryotic cell, where the energetic cost of having large genomes is high.

A clear tendency of genome evolution is observed in ***Figure 3***B, where alternative splicing increases for the taxonomic groups having high intron-rich genes. Specifically, the trend starts from prokaryotes, where alternative splicing is almost absent, followed by unicellular eukaryotes, fungi, plants, arthropods, fish, birds, and finally mammals, where alternative splicing reaches maximum values. This result is also in concordance with the prevalence of intron retention in plants and the evolution of alternative splicing towards more complex modes of splicing, such as exon skipping, found in mammals, where genes usually have a highly conserved intron-rich structure. Interestingly, we observe an opposite tendency between prokaryotes and the most complex lineages, such as mammals, while the other groups distribute as intermediate stages of evolution according to their divergence from mammals. Thus, this may indicate an evolutionary tendency whereby alternative splicing ratios increase in the evolution of complex organisms, shaped by an intron-rich gene structure.

Figure 4 shows the contribution of the alternative splicing ratio to genome size, gene content, coding, and the percentage of coding composing the genome. A qualitative comparison shows that the different taxonomic groups are segregated into different spectrum regions and shape certain evolutionary trends. In all figures, the slopes correspond to variability relations, computed as $m_s = s_x/s_y$, where s_x denotes the standard deviation of alternative splicing and s_y to the corresponding genome variable. Furthermore, to quantify potential trends we compare the variability of each pair of variables regardless of scale. Thus, we have first computed the variation coefficient for each variable, defined as $d_x = s_x/\overline{x}$, and obtained the relation of variability $r = d_x/d_y$ for each taxonomic group, as shown in ***Table 3***.

Figure 2.

Pearson correlation between the relative amount of genes within genomes and alternative splicing ratio for (A) *n* = 70 mammals (r=0.716,p-value=3.181e-12), (B) *n* = 26 birds (r=0.823,p-value=2.419e-07), (C) *n* = 88 fish (r=0.604,p-value=4.646e-10), (D) *n* = 61 arthropods (r=0.712,p-value=1.234e-10), and (E) *n* = 75 flowering plants (r=0.151,p-value=0.193). Slopes correspond to linear regressions.

Figure 2—figure supplement 1. Relation between genome size and alternative splicing ratio for each taxonomical group.

Figure 2—figure supplement 2. Relation between the gene content and alternative splicing ratio for each taxonomical group.

Figure 2—figure supplement 3. Relation between the amount of coding DNA and alternative splicing ratio for each taxonomical group.

Figure 2—figure supplement 4. Relation between the relative amount of coding DNA within genomes and alternative splicing ratio for each taxonomical group.

Figure 2—figure supplement 5. Relation between the relative amount of coding DNA within genes and alternative splicing ratio for each taxonomical group.

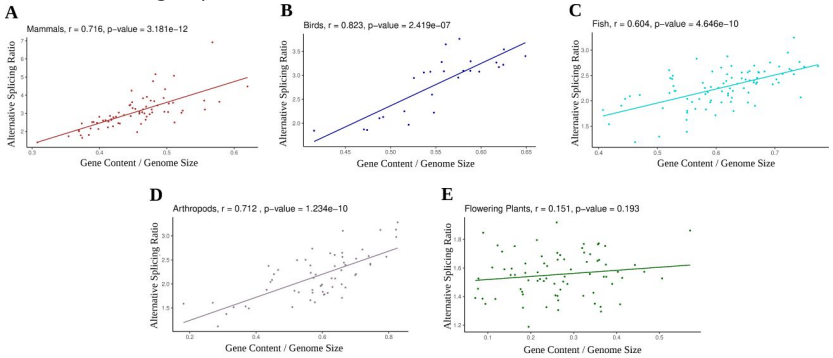


Table 2.

Average, standard deviation, coefficient of variation, and range (i.e, minimum and maximum values) of alternative splicing ratio for each taxonomic group: prokaryotes (n=684), unicellular eukaryotes (n=34), fungi (n=69), flowering plants (n=75), arthropods (n=61), fish (n=88), birds (n=26), and mammals (n=70).

Group	$\overline{x}$	s_x	$s_x/\overline{x}$	$[x_{min}, x_{max}]$
Prokaryotes	1.001	0.001	0.001	[1,1.007]
Unicellular Euk.	1.001	0.003	0.003	[1,1.01]
Fungi	1.017	0.064	0.062	[1,1.362]
Flowering Plants	1.554	0.152	0.097	[1.188,1.918]
Arthropods	2.163	0.470	0.217	[1.116,3.279]
Fish	2.264	0.370	0.163	[1.185,3.247]
Birds	2.837	0.599	0.211	[1.845,3.759]
Mammals	3.056	0.906	0.296	[1.396,6.919]

Figure 3.

Contribution of alternative splicing ratio to (A) the percentage of intergenic DNA and (B) the percentage of coding composing genes. The inset shows the relation in logarithmic scale and its associated pearson correlation (r=-0.305). Colors represent the different taxonomical groups.

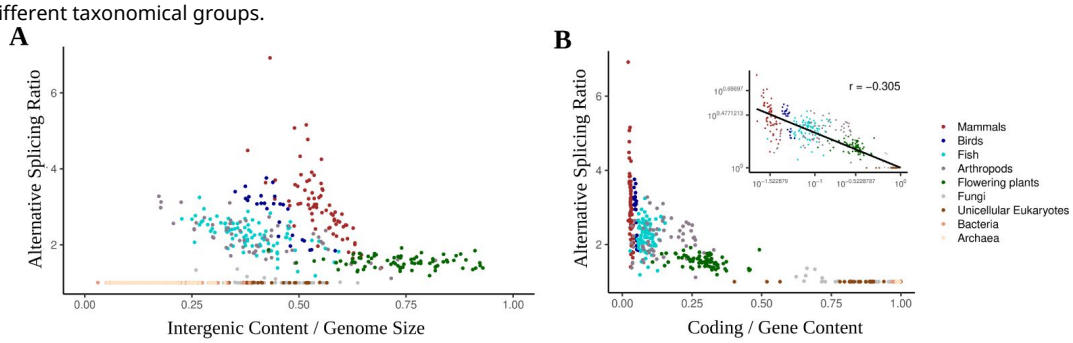


Figure 4.

Contribution of alternative splicing ratio to (A) genome size, (B) gene content, (C) protein-coding DNA, and (D) the percentage of coding relative to genome sizes. Slopes correspond to variability relations (i.e, the ratio between their standard deviations).

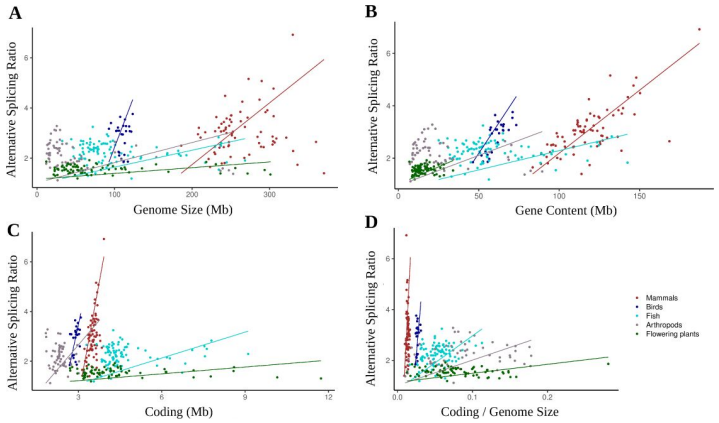


Table 3.

Relation of variability, r, between alternative splicing ratio (x) and genome size, gene content, coding, and percentage of coding relative to genome sizes (y) for each taxonomic group: mammals (n=70), birds (n=26), fish (n=88), arthropods (n=61), and flowering plants (n=75)

Group	$r(y = Genome)$	$r(y = Genes)$	$r(y = Coding)$	$r(y = Coding(\%))$
Mammals	2.119	1.862	7.066	2.550
Birds	3.150	1.992	5.296	3.191
Fish	0.301	0.382	0.717	0.503
Arthropods	0.189	0.268	1.447	0.382
Flowering plants	0.069	0.258	0.253	0.174

As we can observe in ***Figure 4*** A, plants assume values in a restricted range of alternative splicing ratios at low values while showing a high variability of genome sizes. Indeed, while a small group of plants have huge genomes, the distribution is skewed to small genomes. Conversely, mammals typically have higher genome sizes than plants but span different alternative splicing values. The same pattern observed in mammals appears in birds, with smaller genome sizes and a shorter range of alternative splicing values. There is no clear tendency in fish and arthropods, in which values span both spectrum ranges. The variability relation is $r > 1$ for mammals and birds, whereas for the other groups it is $r < 1$, with plants reaching a minimum value close to zero. A similar pattern shaped by the different taxonomic groups is observed for gene content (***Figure 4*** B). We also compare alternative splicing values to the amount of coding in ***Figure 4*** C. On the one hand, mammals, birds and arthropods conserve the relative positions found in the genome size and gene content, whereas fish accumulate higher amounts of coding. Considering the results in ***Figure 3***, we discern that marine animals differ from land animals; they have shorter genomes, higher amounts of coding and, in some cases, a lower percentage of intergenic DNA. Conversely, while plants typically have smaller genome sizes than most animals, their genomes accumulate large amounts of coding DNA, which forms exon-rich genes, and large quantities of intergenic DNA.

Figure 4 D shows the percentage of coding DNA composing genomes and its relation to the alternative splicing ratio. In mammals, the coding percentage is between 0.92% and 1.73%. Only about 1% of mammalian genomes comprise protein-coding DNA, while their alternative splicing ratios range from $\rho_{min} = 1.396$ to $\rho_{max} = 6.919$. Birds show a similar pattern, where the coding percentages are also low, between 2.4% and 3.12%, while splicing ratios span from $\rho_{min} = 1.845$ to $\rho_{max} = 3.759$. By contrast, fish show a decrease in the maximum value of alternative splicing, which spans the range of $\rho_{min} = 1.185$ to $\rho_{max} = 3.247$, while showing a higher coding variability. In particular, it composes between the 1.63% and 11.26% of their genomes. Similarly, arthropods have a coding percentage of between 1% and 17.83%, while their splicing ratios are low, ranging from $\rho_{min} = 1.116$ to $\rho_{max} = 3.279$. Finally, flowering plants are the taxonomic group exhibiting the maximum coding variability, which forms in between 1.28% and 28.11% of genomes but with a maximal reduction in the range of splicing ratios, going from $\rho_{min} = 1.188$ to $\rho_{max} = 1.918$. As we can observe, the corresponding variability coefficients are $r = 0.174$ for plants, $r = 0.382$ for arthropods, $r = 0.503$ for fish, $r = 3.191$ for birds, and $r = 2.550$ for mammals, forming a progressive evolutionary pattern. In plants, alternative splicing does not seem to play a relevant role as compared to mammals, which show an opposite pattern. Moreover, we have found that the other taxonomic groups are between these two spectrums, highlighting the relative roles of genome composition and alternative splicing in the evolution of each taxonomic group. As observed in ***Table 3***, the variability relations have values of $r > 1$ for mammals and birds in all cases, which indicates a higher variability of alternative splicing compared to the other genome variables. On the other hand, we obtain values of less than one for the other taxonomic groups, with a minimum corresponding to plants. These results suggest that genome evolution follows a progressive trend that goes from prokaryotes, unicellular eukaryotes, fungi, flowering plants, arthropods, fish, birds and finally mammals, where the relative roles of introns, exons and intergenic regions become more constrained in highly evolved organisms.

Discussion

Genomes of prokaryotes are mainly composed of single-exon genes; therefore, increasing genome sizes reflects an increase in phenotype innovation. However, this scenario is very different in eukaryotes. Despite the increased genome size, the amount of protein-coding DNA seems to be limited to a size of about 10 MB, which may be related to biophysical constraints (***Choi et al., 2020***). Thus, the limitation of forming new proteins may be overcome by the selective combination of exons through alternative splicing. This has contributed to the transition from a one-to-one relationship between the genotypic and phenotypic spaces in prokaryotes to a complex

relationship in eukaryotes. Genome architecture has allowed this transition in eukaryotes, which differ from bacteria and archaea in different aspects. The organization of DNA into chromosomes and the presence of the nuclear membrane have constrained genome evolution of eukaryotes (**Archibald, 2015**; **Thomas, 1971**; **Volff and Altenbuchner, 2000**). In particular, the compartmentalization of the eukaryotic cell has provided physical support for the accumulation of non-coding DNA and the development of new modes of alternative splicing (**Ros-Rocher et al., 2021**; **Dey et al., 2014**; **Mattick, 2003**). Various authors highlight the role of alternative splicing in the differentiation process, even at varying levels of the biological organization (**Chen et al., 2012**; **Nilsen and Graveley, 2010**). In concordance with ancestral reconstructions (**Anantharaman et al., 2002**) and comparative studies (**Chaudhary et al., 2019**), our results suggest that alternative splicing is almost exclusive to multicellular life forms.

Genome composition in eukaryotes, however, differs among the different taxonomic groups. As summarized in **Table 4**, we observe the starting accumulation of non-coding DNA in the eukaryotic cell, reaching very high levels in multicellular organisms. In particular, the coding DNA of unicellular eukaryotes makes up about 50% of genomes, but it decreases drastically for all multicellular organisms, comprising less than 2% in mammals. In general, we observe that coding, gene, and genome size increase proportionally to each other at the lineage-specific level. Although this relationship is linear in almost all cases, the slopes differ for each taxonomic group. On the one hand, plants accumulate large amounts of intergenic DNA, composing about 74% of genomes, and have high exon-rich genes compared to animals. On the other hand, intergenic DNA in animals is much lower, about 50%, and coding reduces drastically, making up less than 14% of genes for all animal groups.

The relative roles of sequence duplication and alternative splicing show lineage-specific patterns and shape a trend where the role of alternative splicing is relatively low in plants, followed by arthropods, fish, birds, and finally mammals, where alternative splicing plays a key role and coding is maintained at low values. The high variability of genome factors characterizing eukaryotic genomes and its coupling to alternative splicing can be partially explained by the limitations imposed by cell physiology. Genome sizes fluctuate through the expansion and reduction of DNA sequences coupled with cell activity (**Cavalier-Smith, 1978**). Whole-genome duplication and poly-ploidization are recurrent mechanisms in plants and are the mainly responsible for huge genomes (**Clark and Donoghue, 2018**; **Bennett and Leitch, 2005**). About 70% of plants are polyploids (**Soltis et al., 2003**), and most species in the plant kingdom have undergone at least one round of whole-genome duplication (**Clark and Donoghue, 2018**). This coupling of genome sizes to cell activity may explain why plants, which are less restricted in terms of storing DNA sequences compared to animals, do not prioritize alternative splicing as a mechanism of transcript diversification. In particular, we observe that the alternative splicing ratio is fixed at a moderate value in all flowering plants, a much lower value than in most animals. By contrast, we observe a high variability in their genome composition, suggesting that plants' primary mechanism of genome evolution is by expanding their genomes. By contrast, animals are more constrained in terms of energy consumption, so maintaining excessive amounts of repeat DNA requires a high energy cost to the cell (**Choi et al., 2020**). As a counterbalancing response to the expansion of genomes, repetitive sequences are reduced through a process termed down-sizing (**Doyle and Coate, 2019**), which prevents the uncontrolled expansion of genomes and controls the proper functioning of the cell (**Francis et al., 2008**; **Bennett and Leitch, 2005**; **Gregory, 2005**). In animals, it has been observed that genome sizes are continuously adjusted to cell activity (**Cavalier-Smith, 1978**; **Knight and Beaulieu, 2008**), which may explain why alternative splicing is the most relevant mechanism of transcript diversification. These results show that animals and plants follow very different evolutionary pathways and suggest that alternative splicing reflects some organismal complexity in animals alone.

Group	Gene / Genome (%)	Coding / Genome (%)	Coding / Genes (%)	AS
Prokaryotes	85.8	84.9	98.9	1.001
Unicellular Euk.	60.9	52.2	86.5	1.001
Fungi	64.7	58.6	90.0	1.017
Flowering Plants	26.4	7.8	29.0	1.554
Arthropods	58.4	8.3	14.1	2.163
Fish	61.1	5.3	8.6	2.264
Birds	55.3	2.6	4.8	2.837
Mammals	45.2	1.3	3.0	3.056

Table 4.

Percentages of genome composition and the alternative splicing ratio for each taxonomic group: prokaryotes (n=684), unicellular eukaryotes (n=34), fungi (n=69), flowering plants (n=75), arthropods (n=61), fish (n=88), birds (n=26), and mammals (n=70).

The different mechanisms of transcript diversification shaping a trend that goes from plants, arthropods, fish, birds, and mammals is also in concordance with the fact that alternative splicing has evolved and diversified by different modes. The molecular processes behind alternative splicing can be described in terms of how introns and exons are defined, i.e., the constraints imposed by the splicing recognition machinery (***Xing and Lee, 2006***; ***Keren et al., 2010***). In particular, it is observed during the evolution of a shift from intron to exon recognition, associated with increased intron size for highly evolved organisms (***Rogozin et al., 2012***; ***Schwartz et al., 2009***, ***2008***). Among the most important modes of alternative splicing we find exon skipping, a mechanism by which some exons are spliced out of the transcript to generate mature mRNA. This is the most frequent mode in higher eukaryotes, and its prevalence gradually decreases further down the eukaryotic tree, being very rare in lower eukaryotes (***Kim et al., 2008***). Intron retention is other mechanism that occurs when an intron is retained in the mature mRNA transcript. While it only represents about 5% of events in animals, it is the prevalent mechanism in plants, fungi, and protozoa (***Jacob and Smith, 2017***; ***Alekseyenko et al., 2007***). In accordance with previous results, we observe a gradual increase in alternative splicing ratios coupled with a rise in intron-rich gene composition, ranging from unicellular eukaryotes, where genes are mainly composed by single-exons and alternative splicing does not play a key role, to flowering plants, arthropods, fish, birds, and finally mammals. A comparison of these values with the relative amounts of genome composition, summarized in ***Table 4***, indicates that alternative splicing is closely linked to gene composition structures at each specific lineage level. Thus, this progressive trend differentiates each of the taxonomic groups in terms of their genome compositional structure, where alternative splicing is maximized for mammalian species with high intron-rich genes and an intergenic DNA composition of about 50% of genomes. Because exons flanked by long introns are more prone to exon skipping, this result concurs with a prevalence of exon skipping in mammals and its association with an intron-rich gene structure, which becomes increasingly conserved in highly evolved organisms. Finally, the variability of alternative splicing values increases for increasing average values, suggesting lineage-specific evolutionary trends.

In summary, transcript diversification by the selective combination of intron and exon sequences is reflected in the alternative splicing ratios. So, we propose it as a potential candidate to quantify organismal complexity. We observe a gradual increase of alternative splicing ratios toward highly evolved taxonomical groups. In particular, we find low values for flowering plants (ρ = 1.55), and an increase for arthropods (ρ = 2.16), fish (ρ = 2.26), amphibians (ρ = 2.38), birds (ρ = 2.83) and finally mammals (ρ = 3.05). In the case of single-celled organisms, alternative splicing does not appear to play a crucial role in genome evolution, thus we observe values close to one in most cases. The genome compositional trait that best differentiates multicellular life forms to single-celled organisms is the presence of multi-exon, intron-rich genes, which seems to occur coupled to increasing transcript diversification through alternative splicing. This result suggests that alternative splicing is only a relevant mechanism in multicelular life forms, where differentiation regulated at the genome level becomes crucial.

We find a strong negative correlation between alternative splicing ratios and the percentage of coding composing genes in mammals and birds, which highlights the key role of intronic sequences in highly evolved organisms. In general, alternative splicing is maximized within a genome structure that maintains about 50% of intergenic DNA and with intron-rich genes. We also observe an interesting pattern, whereby plants maintain alternative splicing values at moderate values while showing high variability in their genome composition, specifically in the percentage of coding. The opposite trend is observed for mammals, which show a highly conserved genome compositional structure, with a percentage of coding around 1.35%, while presenting a high variation in alternative splicing values. Finally, the tendencies for other taxonomic groups lie between these two spectrums, highlighting the relative roles of sequence duplication and alternative splicing. Indeed, the different modes of alternative splicing and their association to genome compositional structure are in concordance with the energy contraints imposed by the cell. This may explain why sophisticated groups, such as mammals, evolve towards an efficient

organization of information processing, while other groups play with the relative roles of storage capacity and information-processing efficiency. Indeed, plants, which have less energy constraints if compared to animals, seem to play with large accumulations of intergenic DNA.

Methods

Genome variables

We consider three genome variables characterizing DNA sequence composition: genome size, gene content, and coding. While gene content refers to the amount of DNA composing all genes, coding represents the amount of DNA within genes that codes for a protein. Furthermore, we defined a novel measure, the *alternative splicing ratio*, which quantifies the ratio of the different protein isoforms (CDSs) that can build up through the alternative combination of protein-coding sequences. ***Figure 5*** shows a representative example of how it is computed.

Mathematically, it is defined as follows. We consider the genome as the sequential order of nucleotides, where the ith position of a genome is denoted as $G(i)$. We define $f(i,j)$ as the number of times the nucleotides i and j appear together in a CDS. Thus, $f(i,i)$ corresponds to the number of different CDSs where nucleotide $G(i)$ is inserted. The symmetric matrix $\boldsymbol{M}_{ij} = f(i,j)$, the *transcription matrix*, captures the connectivity among nucleotides shaping the genotype-phenotype map. The binary version of the transcription matrix is:

$$A_{ij} = \begin{cases} 1, & \text{if } M_{ij} > 0 \\ 0, & \text{if } M_{ij} = 0. \end{cases} \quad (1)$$

Thus, the alternative splicing ratio is defined as:

$$\rho = \frac{tr(M)}{tr(A)}, \quad (2)$$

where $tr(\cdot)$ denotes the trace of a matrix, i.e., the sum of the diagonal elements.

Data source

All genome features are computed from high-quality and whole-genome assembly annotated files from the NCBI database (***Kitts et al., 2015***). All files have been downloaded the December of 2021, except for the human genome. Due to the recent advances in the genome sequencing of *Homo sapiens*, we have downloaded the associated annotated file the December of 2023. We have only considered files assembled at the genome level in order to reduce noisy data. All files are in GFF3 format, which provides well-structured data in tab-delimited columns. In particular, information is provided about the starting and ending positions of the following genome attributes: regions, genes, mRNAs, and CDSs, which are all organized in a tree-like pattern. Thus, the genome variables have been computed by counting the number of positions associated with each attribute. Note that we have only considered CDSs with a well-defined parent-child relationship. Pseudogenes and duplicated genes have been excluded from this study. It is important to recall that estimation of alternative splicing ratios can be highly biased. On the one hand, various methods exist in genome sequencing and data assembly, which can be a source of noise. On the other hand, there is a significant bias causing higher alternative splicing ratios for the more frequently studied species.

We have collected annotation files for organisms spanning the whole tree of life. However, the sampling process was limited by the availability of whole-genome annotation files, which reduced the number of organisms to 449 in the case of eukaryotes. Regarding bacteria and archaea, we randomly collected 396 and 288 species, respectively. We also excluded the organisms with extreme values from the study, such as *Triticum dicoccoides*, which has a genome size larger than

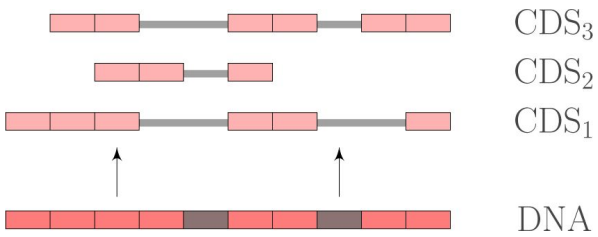


Figure 5.

Representation of the alternative splicing ratio. In this figure, coding DNA has eight nucleotides, which build three different protein isoforms composed of 6, 3, and 6 nucleotides, respectively. Thus, the alternative splicing ratio is computed as (6 + 3 + 6)⁄8 = 1.875.

900 Mb. Also, in order to get statistical significance in the different analysis, we need a high number of species representing each taxonomical group. We had only eight organisms representing Amphibians, so we have excluded them from the study. For the same reason we also restricted the plant kingdom to flowering plants. This procedure has resulted in a total of 423 eukaryotes, which are classified as follows, according to the NCBI Taxonomy Database (***Federhen, 2011***, ***2014***): 70 mammals, 26 birds, 88 fish, 61 arthropods, 75 flowering plants, 69 Fungi, and 34 unicellular eukaryotes (see ***Figure 6***).

Additional files

Supplementary files

- Supplementary file 1. Supplementary tables. Range, average, variance, and coeficient of variation of genome sizes (Table S1), gene content (Table S2), coding DNA (Table S3), and alternative splicing (Table S4), computed for each taxonomical group.

Data availability

Availability of data and materials: Data has been downloaded from the NCBI database (***Kitts et al., 2015***). The list of species under study, code, and the datasets generated in this study are available in Zenodo, at *https://zenodo.org/doi/10.5281/zenodo.10064095* (***de la Fuente, 2023***).

Figure 6.

Classification of the organisms into taxonomical groups according to the NCBI Taxonomy Database (***Federhen, 2011***, ***2014***)

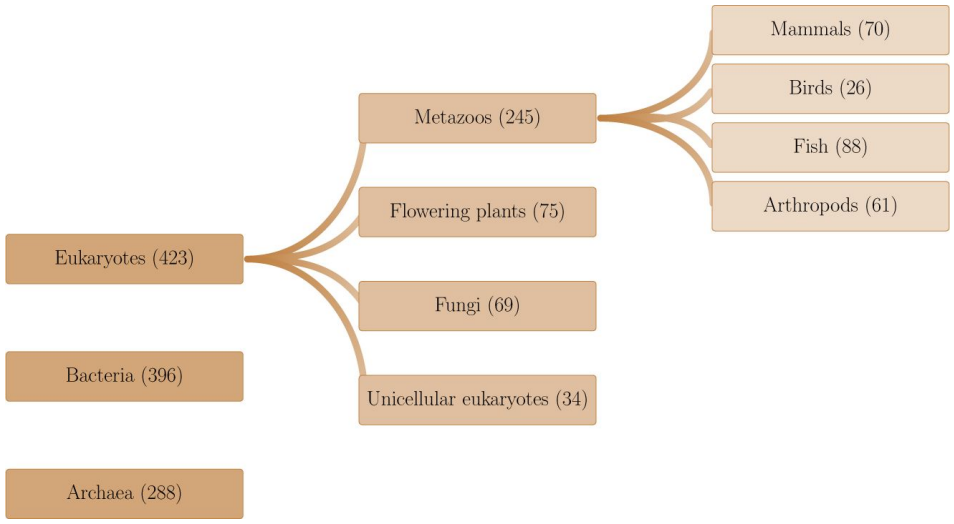


Figure 1—figure supplement 1.

Pearson correlation between genome size and gene content for (A) *n* = 70 mammals (r=0.629,p-value=5.25e-09), (B) *n* = 26 birds (r=0.401,p-value=0.04228), (C) *n* = 88 fish (r=0.962,p-value<2.2e-16), (D) *n* = 61 arthropods (r=0.893,p-value<2.2e-16), (E) *n* = 75 flowering plants (r=0.669,p-value=5.135e-11), (F) *n* = 69 fungi (r=0.951,p-value<2.2e-16), (G) *n* = 34 unicellular eukaryotes (r=0.946,p-value<2.2e-16), (H) *n* = 396 bacteria (r=0.994,p-value=2.2e-16), and (I) *n* = 288 archaea (r=0.986,p-value<2.2e-16). Slopes correspond to linear regressions.

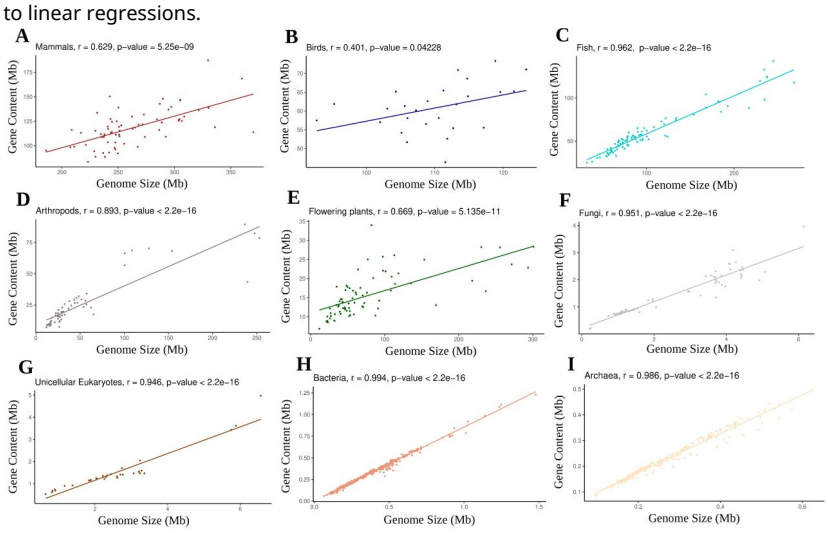


Figure 1—figure supplement 2.

Pearson correlation between genome size and coding DNA for (A) *n* = 70 mammals (r=0.481,p-value=2.405e-05), (B) *n* = 26 birds (r=0.425,p-value=0.03039), (C) *n* = 88 fish (r=0.8,p-value<2.2e-16), (D) *n* = 61 arthropods (r=0.531,p-value=1.045e-05), (E) *n* = 75 flowering plants (r=0.473,p-value=1.799e-05), (F) *n* = 69 fungi (r=0.951,p-value<2.2e-16), (G) *n* = 34 unicellular eukaryotes (r=0.929,p-value=2.306e-15), (H) *n* = 396 bacteria (r=0.994,p-value<2.2e-16), and (I) *n* = 288 archaea (r=0.985,p-value<2.2e-16). Slopes correspond to linear regressions.

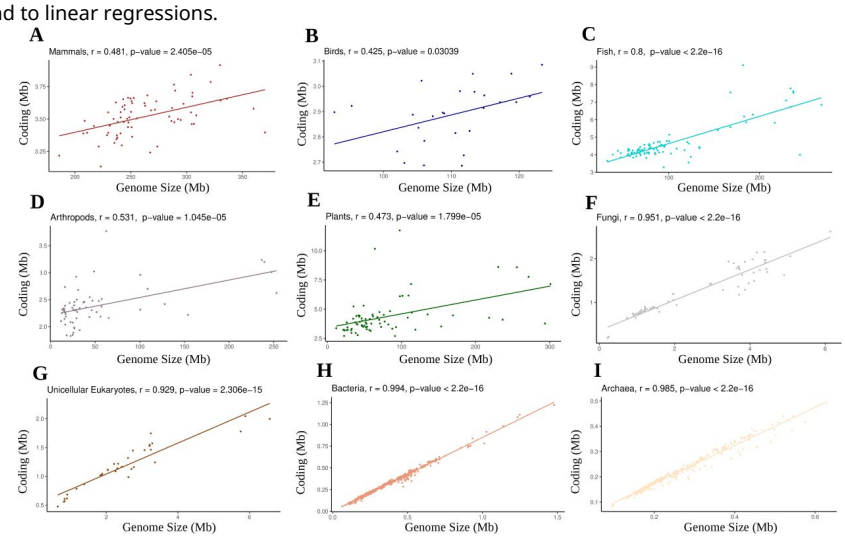


Figure 1—figure supplement 3.

Pearson correlation between gene content and coding DNA for (A) *n* = 70 mammals (r=0.617,p-value=1.292e-08), (B) *n* = 26 birds (r=0.762,p-value=0.6.067e-06), (C) *n* = 88 fish (r=0.774,p-value<2.2e-16), (D) *n* = 61 arthropods (r=0.457,p-value=0.00021), (E) *n* = 75 flowering plants (r=0.649,p-value=2.853e-10), (F) *n* = 69 fungi (r=0.962,p-value<2.2e-16), (G) *n* = 34 unicellular eukaryotes (r=0.832,p-value=1.032e-09), (H) *n* = 396 bacteria (r=0.999,p-value<2.2e-16), and (I) *n* = 288 archaea (r=0.999,p-value<2.2e-16). Slopes correspond to linear regressions.

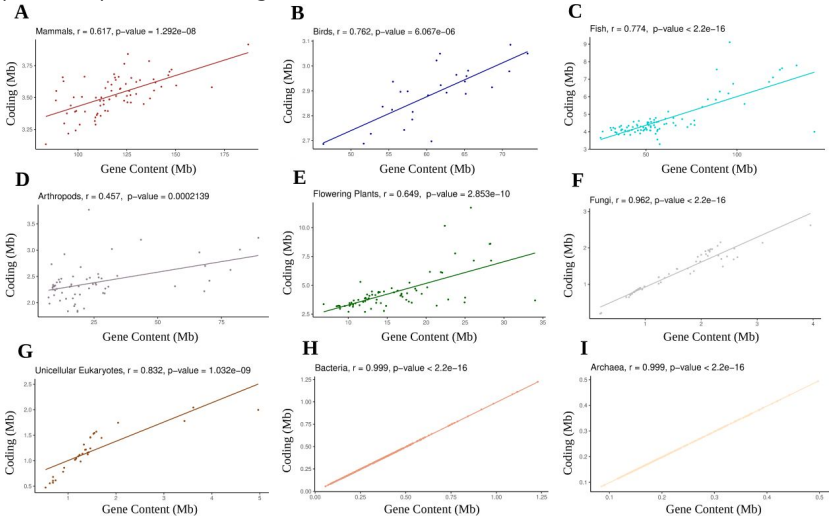


Figure 2—figure supplement 1.

Pearson correlation between genome size and alternative splicing ratio for (A) *n* = 70 mammals (r=0.107,p-value=0.376), (B) *n* = 26 birds (r=0.002,p-value=0.99), (C) *n* = 88 fish (r=-0.162,p-value=0.129), (D) *n* = 61 arthropods (r=-0.4,p-value=0.001), and (E) *n* = 75 flowering plants (r=-0.023,p-value=0.844). Slopes correspond to linear regressions.

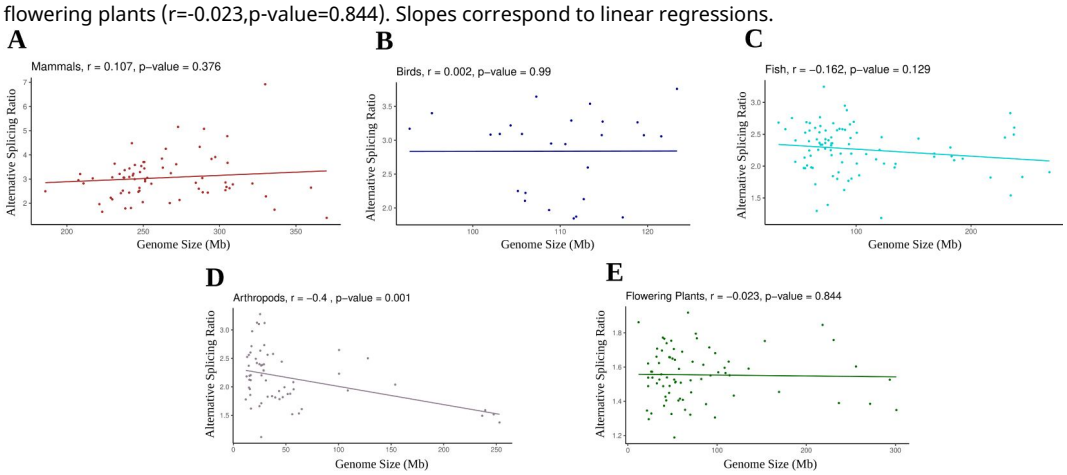


Figure 2—figure supplement 2.

Pearson correlation between the gene content and alternative splicing ratio for (A) *n* = 70 mammals (r=0.711,p-value=5.407e-12), (B) *n* = 26 birds (r=0.786,p-value=1.887e-06), (C) *n* = 88 fish (r=-0.012,p-value=0.911), (D) *n* = 61 arthropods (r=-0.238,p-value=0.064), and (E) *n* = 75 flowering plants (r=0.149,p-value=0.202). Slopes correspond to linear regressions.

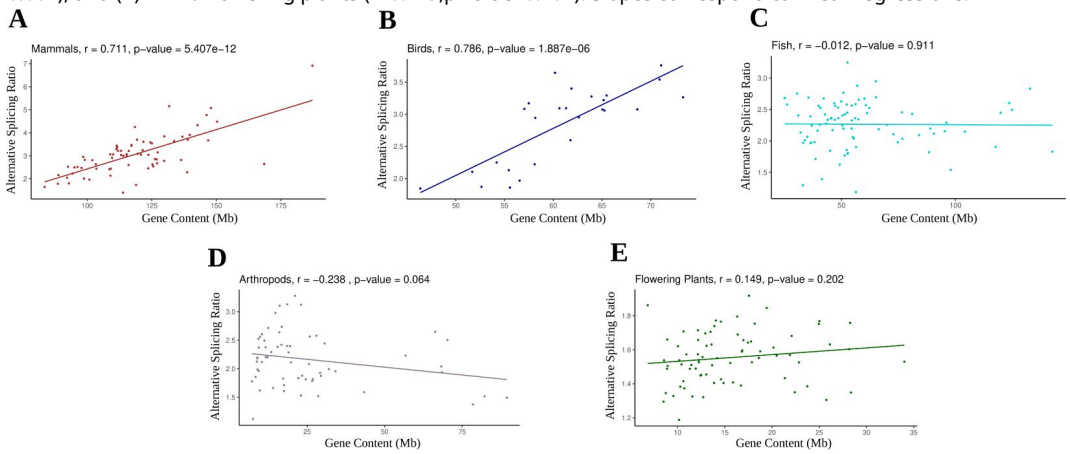


Figure 2—figure supplement 3.

Pearson correlation between the amount of coding DNA and alternative splicing ratio for (A) *n* = 70 mammals (r=0.531,p-value=2.272e-06), (B) *n* = 26 birds (r=0.541,p-value=0.004), (C) *n* = 88 fish (r=0.055,p-value=0.607), (D) *n* = 61 arthropods (r=-0.314,p-value=0.013), and (E) *n* = 75 flowering plants (r=-0.139,p-value=0.231). Slopes correspond to linear regressions.

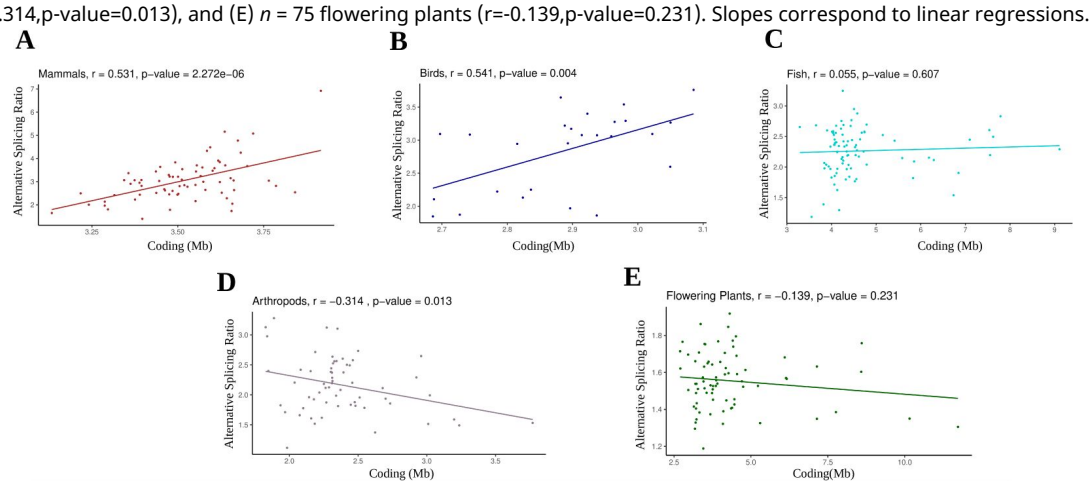
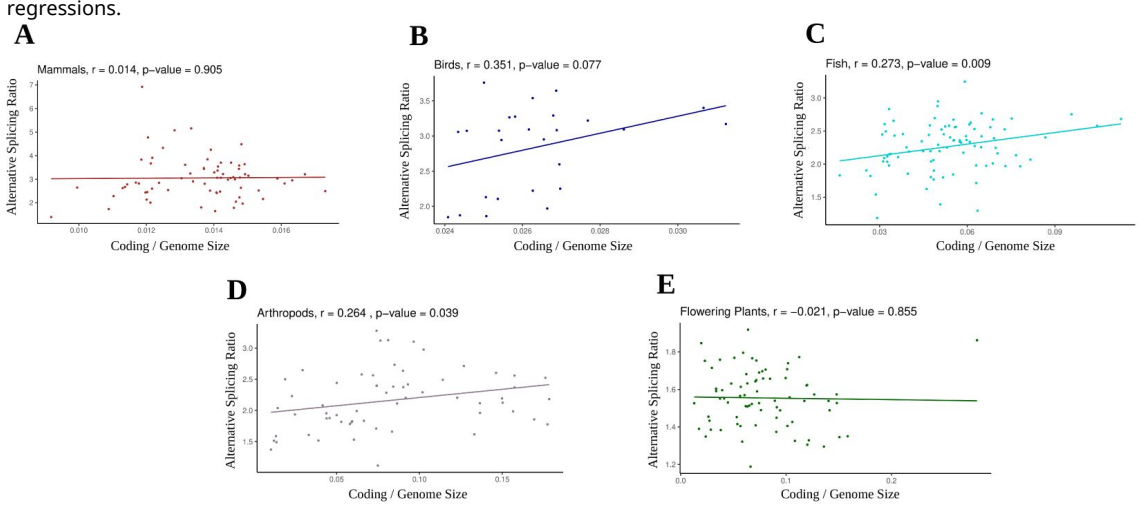


Figure 2—figure supplement 4.

Pearson correlation between the relative amount of coding DNA within genomes and alternative splicing ratio for (A) *n* = 70 mammals (r=0.014,p-value=0.905), (B) *n* = 26 birds (r=0.351,p-value=0.077), (C) *n* = 88 fish (r=0.273,p-value=0.009), (D) *n* = 61 arthropods (r=0.264,p-value=0.039), and (E) *n* = 75 flowering plants (r=-0.021,p-value=0.855). Slopes correspond to linear regressions.



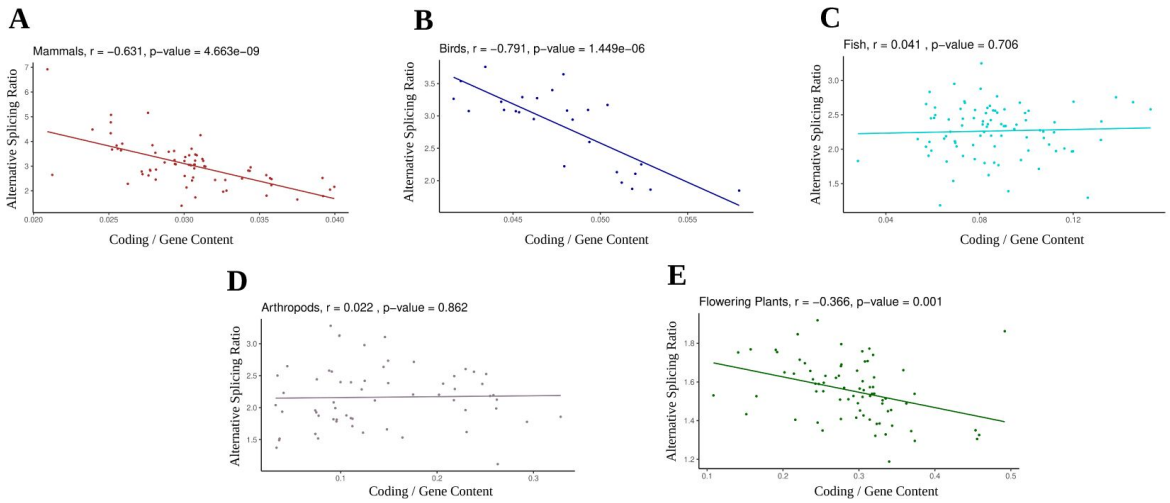


Figure 2—figure supplement 5.

Pearson correlation between the relative amount of coding DNA within genes and alternative splicing ratio for (A) *n* = 70 mammals (r=-0.631,p-value=4.663e-09), (B) *n* = 26 birds (r=-0.791,p-value=1.449e-06), (C) *n* = 88 fish (r=0.041,p-value=0.706), (D) *n* = 61 arthropods (r=0.022,p-value=0.862), and (E) *n* = 75 flowering plants (r=-0.366,p-value=0.001). Slopes correspond to linear regressions.

References

Alekseyenko AV, Kim N, Lee CJ (2007) **Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes** *RNA* **13**:661–670 https://doi.org/10.1261/rna.325107

Anantharaman V, Koonin EV, Aravind L (2002) **Comparative genomics and evolution of proteins involved in RNA metabolism** *Nucleic Acids Research* **30**:1427–1464 https://doi.org/10.1093/nar/30.7.1427

Archibald J. (2015) **Endosymbiosis and Eukaryotic Cell Evolution** *Current Biology* **25**:R911–R921 https://doi.org/10.1016/j.cub.2015.07.055

Barbosa-Morais NL *et al.* (2012) **The Evolutionary Landscape of Alternative Splicing in Vertebrate Species** *Science* **338**:1587–1593 https://doi.org/10.1126/science.1230612

Bennett MD, Leitch IJ (2005) **Genome Size Evolution in Plants** *In: The Evolution of the Genome Academic Press* :89–162 https://doi.org/10.1016/B978-012301463-4/50004-8

Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO (2017) **Alternative splicing and the evolution of phenotypic novelty** *Philosophical Transactions of the Royal Society B: Biological Sciences* **372** https://doi.org/10.1098/rstb.2015.0474

Cavalier-Smith T (1978) **Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox** *Journal of Cell Science* **34**:247–278 https://doi.org/10.1242/jcs.34.1.247

Chaudhary S, Khokhar W, Jabre I, Reddy ASN, Byrne LJ, Wilson CM, Syed NH (2019) **Alternative Splicing and Protein Diversity: Plants Versus Animals** *Frontiers in Plant Science* **10** https://doi.org/10.3389/fpls.2019.00708

Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO (2014) **Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity** *Molecular Biology and Evolution* **31**:1402–1413 https://doi.org/10.1093/mol-bev/msu083

Chen L, Tovar-Corona JM, Urrutia AO (2012) **Alternative splicing: a potential source of functional innovation in the eukaryotic genome** *International journal of evolutionary biology* https://doi.org/10.1155/2012/596274

Choi IY, Kwon EC, Kim NS (2020) **The C-and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy** *Genes & genomics* **42**:699–714 https://doi.org/10.1007/s13258-020-00941-9

Clark JW, Donoghue PCJ (2018) **Whole-Genome Duplication and Plant Macroevolution** *Trends in Plant Science* **23**:933–945 https://doi.org/10.1016/j.tplants.2018.07.006

Claverie JM (2001) **What If There Are Only 30,000 Human Genes?** *Science* **291**:1255–1257 https://doi.org/10.1126/science.1058969

Dey BK, Mueller AC, Dutta A (2014) **Long non-coding RNAs as emerging regulators of differentiation, development, and disease** *Transcription* **5** https://doi.org/10.4161/21541272.2014.944014

Doyle JJ, Coate JE (2019) **Polyploidy, the Nucleotype, and Novelty: The Impact of Genome Doubling on the Biology of the Cell** *International Journal of Plant Sciences* **180**:1–52 https://doi.org/10.1086/700636

Federhen S (2011) **The NCBI Taxonomy database** *Nucleic Acids Research* **40**:D136–D143 https://doi.org/10.1093/nar/gkr1178

Federhen S (2014) **Type material in the NCBI Taxonomy Database** *Nucleic Acids Research* **43**:D1086–D1098 https://doi.org/10.1093/nar/gku1127

Fiszbein A, Kornblihtt AR (2017) **Alternative splicing switches: Important players in cell differentiation** *BioEssays* **39** https://doi.org/10.1002/bies.201600157

Francis D, Davies MS, Barlow PW (2008) **A Strong Nucleotypic Effect on the Cell Cycle Regardless of Ploidy Level** *Annals of Botany* **101**:747–757 https://doi.org/10.1093/aob/mcn038

Fu RH, Liu SP, Ou CW, Yu HH, Li KW, Tsai CH, Shyu WC, Lin SZ (2009) **Alternative Splicing Modulates Stem Cell Differentiation** *Cell Transplantation* **18**:1029–1038 https://doi.org/10.3727/096368909X471260

de la Fuente R (2023) **AltSpCode** *Zenodo* https://doi.org/10.5281/zenodo.10064096

Grantham ME, Brisson JA (2018) **Extensive Differential Splicing Underlies Phenotypically Plastic Aphid Morphs** *Molecular Biology and Evolution* **35**:1934–1946 https://doi.org/10.1093/mol-bev/msy095

Gregory TR (2005) **Genome Size Evolution in Animals** *In: The Evolution of the Genome Academic Press* :3–87 https://doi.org/10.1016/B978-012301463-4/50003-6

Hahn MW, Wray GA (2002) **The g-value paradox** *Evolution & Development* **4**:73–75 https://doi.org/10.1046/j.1525-142X.2002.01069.x

Harr B, Turner LM (2010) **Genome-wide analysis of alternative splicing evolution among Mus subspecies** *Molecular Ecology* **19**:228–239 https://doi.org/10.1111/j.1365-294X.2009.04490.x

Holland PWH, Marletaz F, Maeso I, Dunwell TL, Paps J (2017) **New genes from old: asymmetric divergence of gene duplicates and the evolution of development** *Philosophical Transactions of the Royal Society B: Biological Sciences* **372** https://doi.org/10.1098/rstb.2015.0480

Jacob AG, Smith CWJ (2017) **Intron retention as a component of regulated gene expression programs** *Human Genetics* **136**:1043–1057 https://doi.org/10.1007/s00439-017-1791-x

Keren H, Lev-Maor G, Ast G (2010) **Alternative splicing and evolution: diversification, exon definition and function** *Nature Reviews Genetics* **11**:345–355 https://doi.org/10.1038/nrg2776

Kim E, Goren A, Ast G (2008) **Alternative splicing: current perspectives** *BioEssays* **30**:38–47 https://doi.org/10.1002/bies.20692

Kim E, Magen A, Ast G (2006) **Different levels of alternative splicing among eukaryotes** *Nucleic Acids Research* **35**:125–131 https://doi.org/10.1093/nar/gkl924

Kitts PA *et al.* (2015) **Assembly: a resource for assembled genomes at NCBI** *Nucleic Acids Research* **44**:D73–D80 https://doi.org/10.1093/nar/gkv1226

Knight CA, Beaulieu JM (2008) **Genome Size Scaling through Phenotype Space** *Annals of Botany* **101**:759–766 https://doi.org/10.1093/aob/mcm321

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W (2001) **Initial sequencing and analysis of the human genome** *Nature* **409**:860–921 https://doi.org/10.1038/35057062

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R (2011) **Correction: The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers** *PLOS Biology* **9** https://doi.org/10.1371/annotation/2db9ee19-faa4-43f2-af7a-c8aeacca8037

Lynch M, Conery JS (2000) **The Evolutionary Fate and Consequences of Duplicate Genes** *Science* **290**:1151–1155 https://doi.org/10.1126/sci-ence.290.5494.1151

Mattick JS (2003) **Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms** *BioEssays* **25**:930–939 https://doi.org/10.1002/bies.10332

McShea DW (1996) **Perspective metazoan complexity and evolution: is there a trend?** *Evolution* **50**:477–492 https://doi.org/10.1111/j.1558-5646.1996.tb03861.x

Modrek B, Lee C (2002) **A genomic view of alternative splicing** *Nature genetics* **30**:13–19

Nilsen TW, Graveley BR (2010) **Expansion of the eukaryotic proteome by alternative splicing** *Nature* **463**:457–463 https://doi.org/10.1038/nature08909

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing** *Nature genetics* **40**:1413–1415 https://doi.org/10.1038/ng.259

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

Ros-Rocher N, Pérez-Posada A, Leger MM, Ruiz-Trillo I (2021) **The origin of animals: an ancestral reconstruction of the unicellular-to-multicellular transition** *Open Biology* **11** https://doi.org/10.1098/rsob.200359

Schad E, Tompa P, Hegyi H (2011) **The relationship between proteome size, structural disorder and organism complexity** *Genome biology* **12**:1–13 https://doi.org/10.1186/gb-2011-12-12-r120

Schwartz S, Meshorer E, Ast G (2009) **Chromatin organization marks exon-intron structure** *Nature structural & molecular biology* **16**:990–995 https://doi.org/10.1038/nsmb.1659

Schwartz S, Silva J, Burstein D, Pupko T, Eyras E, Ast G (2008) **Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes** *Genome Research* **18**:88–103 https://doi.org/10.1101/gr.6818908

Soltis DE, Soltis PS, Bennett MD, Leitch IJ (2003) **Evolution of genome size in the angiosperms** *American Journal of Botany* **90**:1596–1603 https://doi.org/10.3732/ajb.90.11.1596

Steward RA, de Jong MA, Oostra V, Wheat CW (2022) **Alternative splicing in seasonal plasticity and the potential for adaptation to environmental change** *Nature Communications* **13** https://doi.org/10.1038/s41467-022-28306-8

Su Z, Wang J, Yu J, Huang X, Gu X (2006) **Evolution of alternative splicing after gene duplication** *Genome Research* **16**:182–189 https://doi.org/10.1101/gr.4197006

Thomas CA (1971) **The genetic organization of chromosomes** *Annual Review of Genetics* **5**:237–256 https://doi.org/10.1146/annurev.ge.05.120171.001321

Volff JN, Altenbuchner J (2000) **A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution** *FEMS Microbiology Letters* **186**:143–150 https://doi.org/10.1111/j.1574-6968.2000.tb09095.x

Xing Y, Lee C (2006) **Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes** *Nature Reviews Genetics* **7**:499–509 https://doi.org/10.1038/nrg1896

Editors

Reviewing Editor
Virginie Courtier-Orgogozo
CNRS - Université Paris Cité, Paris, France

Senior Editor
Alan Moses
University of Toronto, Toronto, Canada

Reviewer #1 (Public Review):

Summary:

The authors collected genomic information from public sources covering 423 eukaryote genomes and around 650 prokaryote genomes. Based on pre-computed CDS annotation, they estimated the frequency of alternative splicing (AS) as a single average measure for each genome and computed correlations with this measure and other genomic properties such as genome size, percentage of coding DNA, gene and intergenic span, etc. They conclude that AS frequency increases with genome complexity in a somewhat directional trend from "lower" organisms to "higher" organisms.

Strengths:

The study covers a wide range of taxonomic groups, both in prokaryotes and eukaryotes.

Weaknesses:

The study is weak both methodologically and conceptually. Current high throughput sequencing technologies, coupled with highly heterogeneous annotation methods, can observe cases of AS with great sensitivity, and one should be extremely cautious of the biases and rates of false positives associated with these methods. These issues are not addressed in the manuscript. Here, AS measures seem to be derived directly from CDS annotations downloaded from public databases, and do not account for differing annotation methods or RNA sequencing depth and tissue sample diversity.

There is no mention of the possibility that AS could be largely caused by random splicing errors, a possibility that could very well fit with the manuscript's data. Instead, the authors adopt early on the view that AS is regulated and functional, generally citing outdated literature.

There is no question that some AS events are functional, as evidenced by strongly supported studies. However, whether all AS events are functional is questionable, and the relative fractions of functional and non-functional AS are unknown. With this in mind, the authors should be more cautious in interpreting their data. The "complexity" of organisms also correlates well (negatively) with effective population size. The power of selection to eliminate (slightly) deleterious mutations or errors decreases with effective population size. The correlation observed by the authors could thus easily be explained by a non-adaptive interpretation based on simple population genetics principles.

The manuscript contains evidence that the authors might benefit from adopting a more modern view of how evolution proceeds. Sentences such as "... suggests that only sophisticated organisms optimize alternative splicing by increasing..." (L113), or "especially in highly evolved groups such as mammals" (L130), or the repeated use of "higher" and "lower" organisms need revising.

Because of the lack of controls mentioned above, and because of the absence of discussion regarding an alternative non-adaptive interpretation, the analyses presented in the manuscript are of very limited use to other researchers in the field. In conclusion, the study does not present solid conclusions.

https://doi.org/10.7554/eLife.94802.1.sa1

Reviewer #2 (Public Review):

Summary:

In this contribution, the authors investigate the degree of alternative splicing across the evolutionary tree and identify a trend of increasing alternative splicing as you move from the base of the tree (here, only prokaryotes are considered) towards the tips of the tree. In particular, the authors investigate how the degree of alternative splicing (roughly speaking, the number of different proteins made from a single ORF (open reading frame) via alternative splicing) relates to three genomic variables: the genome size, the gene content (meaning the fraction of the genome composed of ORFs), and finally, the coding percentage of ORFs, meaning the ratio between exons and total DNA in the ORF. When correlating the degree of alternative splicing with these three variables, they find that the different taxonomic groups have a different correlation coefficient, and identify a "progressive pattern" among metazoan groups, namely that the correlation coefficient mostly increases when moving from flowering plants to arthropods, fish, birds, and finally mammals. They conclude that therefore the amount of splicing that is performed by an organismal group could be used as a measure of its complexity.

Weaknesses:

While I find the analysis of alternative splicing interesting, I also find that it is a very imperfect measure of organismal complexity and that the manuscript as a whole is filled with unsupported statements. First, I think it is clear to anyone studying evolution over the tree of life that it is the complexity of gene regulation that is at the origin of much of organismal structural and behavioral complexity. Arguably, creating different isoforms out of a single ORF is just one example of complex gene regulation. However, the complexity of gene regulation is barely mentioned by the authors. Further, it is clear that none of their

correlation coefficients actually show a simple trend (see Table 3). According to these coefficients, birds are more complex than mammals for 3 out of 4 measures. It is also not clear why the correlation coefficient between alternative splicing ratio and genome length, gene content, and coding percentage should display such a trend, rather than the absolute value. There are only vague mechanistic arguments.

Much more troubling, however, is the statement that the data supports "lineage-specific trends" (lines 299-300). Either this is just an ambiguous formulation, or the authors claim that you can see trends *within* lineages. The latter is clearly not the case. In fact, within each lineage, there is a tremendous amount of variation, to such an extent that many of the coefficients given in Table 3 are close to meaningless. Note that no error bars or p-values are presented for the values shown in Table 3. Figure 2 shows the actual correlation, and the coefficient for flowering plants there is given as 0.151, with a p-value of 0.193. Table 3 seems to quote r=0.174 instead. It should be clear that a correlation within a lineage or species is not a sign of a trend.

There are several wrong or unsupported statements in the manuscript. Early on, the authors state that the alternative splicing ratio (a number greater or equal to one that can be roughly understood as the number of different isoforms per ORF) "quantifies the number of different isoforms that can be transcribed using the same amount of information" (lines 51-52). But in many cases, this is incorrect, because the same sequence can represent different amounts of information depending on the context. So, if a changed context gives rise to a different alternative splice, it is because the genetic sequence has a different meaning in the changed context: the information has changed. In line 149, the authors state that "the energetic cost of having large genomes is high". No citation is given, and while such a statement seems logical, it does not have very solid support. If there was indeed a strong selective force to reduce genome size, we would not see the stunning diversity of genome sizes even within lineages. This statement is repeated (without support) several times in the manuscript, apparently in support of the idea that mammals had "no choice" to increase complexity via alternative splicing because they can't increase it by having longer genomes. I don't think this reasoning can be supported. Even more problematic is the statement that "the amount of protein-coding DNA seems to be limited to a size of about 10MB" (line 219). There is no evidence whatsoever for this statement. The reference that is cited (Choi et al 2020) suggests that there is a maximum of 150GB in total genome size due to physiological constraints. In lines 257-258, the authors write that "plants are less restricted in terms of storing DNA sequences compared to animals" (without providing evidence or a citation). I believe this statement is made due to the observation that plants tend to have large intergenic regions. But without examining the functionality of these interagency regions (they might host long non-coding RNA stretches that are used to regulate the expression of other genes, for example) it is quite adventurous to use such a simple measure as being evidence that plants "are less restricted in terms of storing DNA sequences", whatever that even means. I do not think the authors mean that plants have better access to -80 freezers. The authors conclude that "plant's primary mechanism of genome evolution is by expanding their genome". This statement itself is empty: we know that plants are prone to whole genome duplication, but this duplication is not, as far as we understand, contributing to complexity. It is not a "primary mechanism of genome evolution". In lines 293-294, the authors claim that "alternative splicing is maximized in mammalian genomes". There is no evidence that this ratio cannot be increased. So, to conclude (on lines 302-303) that alternative splicing ratios are "a potential candidate to quantify organismal complexity" seems, based on this evidence, both far-fetched and weak at the same time.

I am also not very comfortable with the data analysis. The authors, for example, say that they have eliminated from their analysis a number of "outlier species". They mention one: Emmer wheat because it has a genome size of 900 Mb (line 367). Since 900MB does not appear to be extreme, perhaps the authors meant to write 900 Gb. When I consulted the paper that

sequenced Triticum dicoccoides, they noted that 14 chromosomes are about 10GB. Even a tetraploid species would then not be near 900Gb. But more importantly, such a study needs to state precisely which species were left out, and what the criteria are for leaving out data, lest they be accused of selecting data to fit their hypothesis.

I understand that Methods are often put at the end of a manuscript, but the measures discussed here are so fundamental to the analysis that a brief description of what the different measures are (in particular, the "alternative splicing ratio") should be in the main text, even when the mathematical definition can remain in the Methods.

Finally, a few words on presentation. I understand that the following comments might read differently after the authors change their presentation. This manuscript was at the border of being comprehensible. In many cases, I could discern the meaning of words and sentences in contexts but sometimes even that failed (as an example above, about "species-specific trends", illustrates). The authors introduced jargon that does not have any meaning in the English language, and they do this over and over again.

Note that I completely agree with all the comments by the other reviewer, who alerted me to problems I did not catch, including the possible correlation with effective population size: a possible non-adaptive explanation for the results.

https://doi.org/10.7554/eLife.94802.1.sa0

Author response:

> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *The authors collected genomic information from public sources covering 423 eukaryote genomes and around 650 prokaryote genomes. Based on pre-computed CDS annotation, they estimated the frequency of alternative splicing (AS) as a single average measure for each genome and computed correlations with this measure and other genomic properties such as genome size, percentage of coding DNA, gene and intergenic span, etc. They conclude that AS frequency increases with genome complexity in a somewhat directional trend from "lower" organisms to "higher" organisms.*
>
> *Strengths:*
>
> *The study covers a wide range of taxonomic groups, both in prokaryotes and eukaryotes.*
>
> *Weaknesses:*
>
> *The study is weak both methodologically and conceptually. Current high throughput sequencing technologies, coupled with highly heterogeneous annotation methods, can observe cases of AS with great sensitivity, and one should be extremely cautious of the biases and rates of false positives associated with these methods. These issues are not addressed in the manuscript. Here, AS measures seem to be derived directly from CDS annotations downloaded from public databases, and do not account for differing annotation methods or RNA sequencing depth and tissue sample diversity.*

We are aware of the bias that may exist in annotation files. Since the source of noise can be highly variable, we have assumed that most of the data has a similar bias. However, we agree with the reviewer that we could perform some analysis to test for these biases and their association to different methodologies. Thus, we will measure the uncertainty present in the data. From one side, we will be more explicit about the data limitations and the biases it can

generate in the results. On the other side, while analyzing the false positives in the data is out of our scope, we will perform a statistical test to detect possible biases regarding different methods of sequencing and annotation, and types of organisms (model or non-model organisms). If positive, we will proceed, as far as possible, to normalize the data or to estimate a confidence interval.

> *Here, AS measures seem to be derived directly from CDS annotations downloaded from public databases, and do not account for differing annotation methods or RNA sequencing depth and tissue sample diversity.*

Beyond taking into account the differential bias that may exist in the data, we do not consider that our AS measure is problematic. The NCBI database is one of the most reliable databases that we have to date and is continuously updated from all scientific community. So, the use of this data and the corresponding procedures for deriving the AS measure are perfectly acceptable for a comparative analysis on such a huge global scale. Furthermore, the proposal of a new genome-level measure of AS that allows to compare species spanning the whole tree of life is part of the novelty of the study. We understand that small-scale studies require a high specificity about the molecular processes involved in the study. However, this is not the case, where we are dealing with a large-scale problem. On the other side, as we have previously mention, we agree with the reviewer to analyze the degree of uncertainty in the data to better interpret the results.

> *There is no mention of the possibility that AS could be largely caused by random splicing errors, a possibility that could very well fit with the manuscript's data. Instead, the authors adopt early on the view that AS is regulated and functional, generally citing outdated literature.*
>
> *There is no question that some AS events are functional, as evidenced by strongly supported studies. However, whether all AS events are functional is questionable, and the relative fractions of functional and non-functional AS are unknown. With this in mind, the authors should be more cautious in interpreting their data.*

Many studies suggest that most of the AS events observed are the result of splicing errors and are therefore neither functional nor conserved. However, we still have limited knowledge about the functionality of AS. Just because we don’t have a complete understanding of its functionality, doesn’t mean there isn’t a fundamental cause behind these events. AS is a highly dynamic process that can be associated with processes of a stochastic nature that are fundamental for phenotypic diversity and innovation. This is one of the reasons why we do not get into a discussion about the functionality of AS and consider it as a potential measure of biological innovation. Nevertheless, we agree with the reviewer’s comments, so we will add a discussion about this issue with updated literature and look at any possible misinterpretation of the results.

> *The "complexity" of organisms also correlates well (negatively) with effective population size. The power of selection to eliminate (slightly) deleterious mutations or errors decreases with effective population size. The correlation observed by the authors could thus easily be explained by a non-adaptive interpretation based on simple population genetics principles.*

We appreciate the observation of the reviewer. We know well the M. Lynch’s theory on the role of the effective population size and its eventual correlation with genomic parameters, but we want to emphasize that our objective is not to find an adaptive or non-adaptive explanation of the evolution of AS, but rather to reveal it. Nevertheless, as the reviewer

suggests, we will look at the correlation between the AS and the effective population size and discuss about a possible non-adaptive interpretation.

> *The manuscript contains evidence that the authors might benefit from adopting a more modern view of how evolution proceeds. Sentences such as "... suggests that only sophisticated organisms optimize alternative splicing by increasing..." (L113), or "especially in highly evolved groups such as mammals" (L130), or the repeated use of "higher" and "lower" organisms need revising.*

As the reviewer suggests, we will proceed with the corresponding linguistic corrections.

> *Because of the lack of controls mentioned above, and because of the absence of discussion regarding an alternative non-adaptive interpretation, the analyses presented in the manuscript are of very limited use to other researchers in the field. In conclusion, the study does not present solid conclusions.*
>
> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *In this contribution, the authors investigate the degree of alternative splicing across the evolutionary tree and identify a trend of increasing alternative splicing as you move from the base of the tree (here, only prokaryotes are considered) towards the tips of the tree. In particular, the authors investigate how the degree of alternative splicing (roughly speaking, the number of different proteins made from a single ORF (open reading frame) via alternative splicing) relates to three genomic variables: the genome size, the gene content (meaning the fraction of the genome composed of ORFs), and finally, the coding percentage of ORFs, meaning the ratio between exons and total DNA in the ORF. When correlating the degree of alternative splicing with these three variables, they find that the different taxonomic groups have a different correlation coefficient, and identify a "progressive pattern" among metazoan groups, namely that the correlation coefficient mostly increases when moving from flowering plants to arthropods, fish, birds, and finally mammals. They conclude that therefore the amount of splicing that is performed by an organismal group could be used as a measure of its complexity.*
>
> *Weaknesses:*
>
> *While I find the analysis of alternative splicing interesting, I also find that it is a very imperfect measure of organismal complexity and that the manuscript as a whole is filled with unsupported statements. First, I think it is clear to anyone studying evolution over the tree of life that it is the complexity of gene regulation that is at the origin of much of organismal structural and behavioral complexity. Arguably, creating different isoforms out of a single ORF is just one example of complex gene regulation. However, the complexity of gene regulation is barely mentioned by the authors.*

We disagree with the reviewer with that our measure of AS is imperfect. Just as we responded to the first reviewer, we will quantify the uncertainty in the data and correct for differential biases caused by annotation and sequencing methods. Thus, beyond correcting relevant biases in the data, we consider that our measure is adequate for a comparative analysis at a global scale. A novelty of our study is the proposal of a genome-level measure of AS that takes into account data from the entire scientific community.

We want also to emphasize that we assume from the beginning that AS may reflect some kind of biological complexity, it is not a conclusion from the results. An argument in favor of such an assumption is that AS is associated with stochastic processes that are fundamental for phenotypic diversity and innovation. Of course, we agree with the reviewer that it is not the

only mechanism behind biological complexity, so we will emphasize it in the manuscript. On the other side, we will be more explicit about the assumptions and objectives, and will correct any unsupported statement.

> *Further, it is clear that none of their correlation coefficients actually show a simple trend (see Table 3). According to these coefficients, birds are more complex than mammals for 3 out of 4 measures.*

An evolutionary trend is broadly defined as the gradual change in some characteristic of organisms as they evolve or adapt to a specific environment. Under our context, we define an evolutionary trend as the gradual change in genome composition and its association with AS across the main taxonomic groups. If we look at Figure 4 and Table 3 we can conclude that there is a progressive trend. We will be more precise about how we define an evolutionary trend and correct any possible misinterpretation of the results. On the other side, we do not assume that mammals should be more complex than birds. First, we will emphasize that our results show that birds have the highest values of such a trend. Second, after reading the reviewer's comments, we have decided that we will perform an additional analysis to correct for differences in the taxonomic group sizes, which will allow us to have more confidence in the results.

> *It is also not clear why the correlation coefficient between alternative splicing ratio and genome length, gene content, and coding percentage should display such a trend, rather than the absolute value. There are only vague mechanistic arguments.*

The study analyzes the relationship of AS with genomic composition for the large taxonomic groups. We assume that significant differences in these relationships are indicators of the presence of different mechanisms of genome evolution. However, we agree with the reviewer that a correlation does not imply a causal relation, so we will be more cautious when interpreting the results.

To quantify the relationships we use correlation coefficients, the slopes of such correlations, and the relation of variability. Although the absolute values of AS are also illustrated in Table 4, we consider that they are less informative than if we include how it relates to the genomic composition. For example, we observe that plants have a different genome composition and relation with AS if compared to animals, which suggest that they follow different mechanisms of genome evolution. On the other hand, we observe a trend in animals, where high values of AS are associated to a large percentage of introns and a percentage of intergenic DNA of about the 50% of genomes.

> *Much more troubling, however, is the statement that the data supports "lineage-specific trends" (lines 299-300). Either this is just an ambiguous formulation, or the authors claim that you can see trends *within* lineages.*

We agree with the reviewer that this statement is not correct, so we will proceed to correct it.

> *The latter is clearly not the case. In fact, within each lineage, there is a tremendous amount of variation, to such an extent that many of the coefficients given in Table 3 are close to meaningless. Note that no error bars or p-values are presented for the values shown in Table 3. Figure 2 shows the actual correlation, and the coefficient for flowering plants there is given as 0.151, with a p-value of 0.193. Table 3 seems to quote r=0.174 instead. It should be clear that a correlation within a lineage or species is not a sign of a trend.*

The reviewer is not understanding correctly the results in Table 3. It is precisely the variation of the genome variables what we are measuring. Given the standardization of these values by the mean values, we have proceeded to compare the variability between groups, which is the result shown in Table 3. In this case there are no error bars or p-values associated. On the other hand, we agree that a correlation is not a sign of a trend. But the relations of variability, together with the results obtained in Figure 3, are indicators of a trend. As we mentioned before, we will proceed to analyze whether the variation in the group sizes is causing a bias in the results.

> *There are several wrong or unsupported statements in the manuscript. Early on, the authors state that the alternative splicing ratio (a number greater or equal to one that can be roughly understood as the number of different isoforms per ORF) "quantifies the number of different isoforms that can be transcribed using the same amount of information" (lines 51-52). But in many cases, this is incorrect, because the same sequence can represent different amounts of information depending on the context. So, if a changed context gives rise to a different alternative splice, it is because the genetic sequence has a different meaning in the changed context: the information has changed.*

We agree that there are not well supported statements, so we will proceed to revise them.

> *In line 149, the authors state that "the energetic cost of having large genomes is high". No citation is given, and while such a statement seems logical, it does not have very solid support.*

We will also revise the bibliography and support our statements with updated references.

> *If there was indeed a strong selective force to reduce genome size, we would not see the stunning diversity of genome sizes even within lineages. This statement is repeated (without support) several times in the manuscript, apparently in support of the idea that mammals had "no choice" to increase complexity via alternative splicing because they can't increase it by having longer genomes. I don't think this reasoning can be supported.*

We agree with the reviewer in this issue, so we will carefully revise the statements that indirectly (or directly) assume the action of selective forces on the genome composition.

> *Even more problematic is the statement that "the amount of protein-coding DNA seems to be limited to a size of about 10MB" (line 219). There is no evidence whatsoever for this statement.*

In Figure 1A we observe a one-to-one relationship between the genome size and the amount of coding. However, in multicellular organisms, although the genome size increases we observe that the amount of coding does not increase by more than 10Mb, which suggest the presence of some genomic limitation. Of course, this is not an absolute or general statement, but rather a suggestion. We are only describing our results.

> *The reference that is cited (Choi et al 2020) suggests that there is a maximum of 150GB in total genome size due to physiological constraints. In lines 257-258, the authors write that "plants are less restricted in terms of storing DNA sequences compared to animals" (without providing evidence or a citation).*

We will revise the bibliography and add updated references.

> *I believe this statement is made due to the observation that plants tend to have large intergenic regions. But without examining the functionality of these interagency regions (they might host long non-coding RNA stretches that are used to regulate the expression of other genes, for example) it is quite adventurous to use such a simple measure as being evidence that plants "are less restricted in terms of storing DNA sequences", whatever that even means. I do not think the authors mean that plants have better access to -80 freezers. The authors conclude that "plant's primary mechanism of genome evolution is by expanding their genome". This statement itself is empty: we know that plants are prone to whole genome duplication, but this duplication is not, as far as we understand, contributing to complexity. It is not a "primary mechanism of genome evolution".*

We will revise these statements.

> *In lines 293-294, the authors claim that "alternative splicing is maximized in mammalian genomes". There is no evidence that this ratio cannot be increased. So, to conclude (on lines 302-303) that alternative splicing ratios are "a potential candidate to quantify organismal complexity" seems, based on this evidence, both far-fetched and weak at the same time.*

Our results show the highest values of AS in mammals, but we understand that the results are limited to the availability and accuracy of data, which we will emphasize in the manuscript. As we previously mention, we will also proceed to analyze the uncertainty in data and carry out the appropriate corrections.

> *I am also not very comfortable with the data analysis. The authors, for example, say that they have eliminated from their analysis a number of "outlier species". They mention one: Emmer wheat because it has a genome size of 900 Mb (line 367). Since 900MB does not appear to be extreme, perhaps the authors meant to write 900 Gb. When I consulted the paper that sequenced Triticum dicoccoides, they noted that 14 chromosomes are about 10GB. Even a tetraploid species would then not be near 900Gb. But more importantly, such a study needs to state precisely which species were left out, and what the criteria are for leaving out data, lest they be accused of selecting data to fit their hypothesis.*

The reviewer is right, we wanted to say 900Mb, which is approximately 7.2Gb. We had a mistake of nomenclature. This value is extreme compared to the typical values, so it generates large deviations when applying measures of central tendency and dispersion. We want to obtain mean values that are representative of the most species composing the taxonomic groups, so we find appropriate to exclude all outlier values in the study. Nevertheless, we will specify the criteria that we have used to select the data in a rigorous way.

> *I understand that Methods are often put at the end of a manuscript, but the measures discussed here are so fundamental to the analysis that a brief description of what the different measures are (in particular, the "alternative splicing ratio") should be in the main text, even when the mathematical definition can remain in the Methods.*

We agree with the reviewer, so we will add a brief description of the genomic variables at the beginning of the Results section.

> *Finally, a few words on presentation. I understand that the following comments might read differently after the authors change their presentation. This manuscript was at the border of being comprehensible. In many cases, I could discern the meaning of words*

and sentences in contexts but sometimes even that failed (as an example above, about "species-specific trends", illustrates). The authors introduced jargon that does not have any meaning in the English language, and they do this over and over again.

Note that I completely agree with all the comments by the other reviewer, who alerted me to problems I did not catch, including the possible correlation with effective population size: a possible non-adaptive explanation for the results.

https://doi.org/10.7554/eLife.94802.1.sa3