Alternative Splicing Across the Tree of Life: A Comparative Study

Rebeca de la Fuente[1*], **Wladimiro Díaz-Villanueva**[2†], **Vicente Arnau**[2†], **Andrés Moya**[1,2,3†]

***For correspondence:** rebeca.fuentes@fisabio.es

[†]These authors contributed equally to this work

[1]Foundation for the Promotion of Sanitary and Biomedical Research of the Valencian Community (FISABIO), 46020 Valencia, Spain; [2]Institute of Integrative Systems Biology (I2Sysbio), University of Valencia and Spanish National Research Council (CSIC), 46980 Valencia, Spain; [3]Center for Biomedical Research in Epidemiology and Public Health Network (CIBEResp), 28029 Madrid, Spain

Abstract There is a growing understanding of how alternative splicing contributes to functional specialization and adaptation, especially in well-studied model organisms. However, its large-scale evolutionary dynamics remain poorly understood. Through a comparative analysis of alternative splicing across 1,494 species spanning the entire tree of life, this study integrates numerous lines of prior evidence to provide a unified view of alternative splicing. We propose a novel genome-scale metric designed to support cross-species comparison. Our findings indicate that alternative splicing is highly variable across lineages. While unicellular eukaryotes and prokaryotes display minimal splicing, mammals and birds exhibit the highest levels of alternative splicing. Despite sharing a conserved intron-rich genomic architecture, mammals and birds show considerable interspecies divergence in splicing activity. In contrast, plants display moderate levels of alternative splicing but exhibit high variability of genomic composition. Furthermore, a strong negative correlation is observed between alternative splicing and the proportion of coding content in genes, with the highest levels of alternative splicing observed in genomes containing approximately 50% intergenic DNA.

Introduction

Alternative splicing is the post-transcriptional mechanism by which different protein isoforms are generated from a single gene by the selective combination of exons and introns (***Rogalska et al., 2024***; ***Bush et al., 2017***; ***Chen et al., 2012***). Thus, it allows functional diversification of a single gene without increasing the number of genes (***Chen et al., 2014***, ***2012***), though it relies on a highly sophisticated set of splicing factors and regulatory pathways to maintain accuracy. This process is central to organismal complexity, regulating gene expression across different tissues, developmental stages, and environmental conditions. On one side, it is involved in essential biological functions, such as tissue differentiation (***Chen et al., 2014***; ***Wright et al., 2022***), neural development, and immune response (***Kim et al., 2006***; ***Rogozin et al., 2012***; ***Wolf and Koonin, 2013***), where it is highly conserved to maintain functional integrity and precise gene regulation. However, its weak splice sites allow for greater plasticity, enabling adaptation to different conditions (***Caron et al., 2017***). As a result, it also serves as a major driver of biological innovation, promoting functional diversification and adaptability by expanding transcriptomic and proteomic diversity (***Su et al., 2006***). It has been described as a form of "evolutionary tinkering", where gradual accumulation of mutations can introduce or remove splice sites, leading to the production of alternative protein isoforms

with modified structural and regulatory properties. Thus, it enables functional adaptations while maintaining essential biological processes (***Mattick, 2009***).

Over the past 1,400 million years, alternative splicing rates have steadily increased, particularly within the metazoan lineage, coinciding with the emergence of specialized cell types and tissue-specific gene regulation. Prokaryotes and unicellular eukaryotes exhibit significantly lower alternative splicing rates, supporting the idea that it represents an advanced regulatory feature associated with multicellularity (***Rogozin et al., 2012***). On the other side, vertebrates exhibit significantly higher levels of alternative splicing than invertebrates (***Hinman and Cary, 2017***), and different taxonomic groups have developed distinct splicing strategies. Animals, in particular mammals, birds, and arthropods, display the highest levels of alternative splicing, whereas plants show intermediate splicing levels, suggesting that it evolved independently in this lineage. Notably, ***Chen et al.*** (***2014***) found no significant difference in alternative splicing levels between birds and mammals, contradicting earlier reports. In contrast to animals, several studies have shown that plants exhibit lower alternative splicing rates but compensate through gene duplication and genome expansion via transposable elements (***Koralewski and Krutovsky, 2011***). In ***Koonin*** (***2006***), authors observed that some large plant genomes, such as maize, exhibit higher alternative splicing rates than smaller genomes like Arabidopsis, suggesting a positive correlation between genome complexity and transcriptomic flexibility. In plants, genome expansion frequently occurs through whole-genome duplications and repetitive element accumulation, both of which create additional opportunities for splice site recognition and regulation. Polyploidy, the presence of multiple chromosome sets, is a widespread feature in plant evolution, with approximately 70% of flowering plants having undergone whole-genome duplications. These duplications lead to subfunctionalization, where duplicated genes evolve different splicing isoforms to fulfill distinct functional roles, thereby increasing alternative splicing diversity. Another major factor influencing alternative splicing in plants is the expansion of transposable elements, particularly retrotransposons, which significantly contribute to genome size and structural variation. As a consequence, plant genomes accumulate large non-coding intergenic regions due to expansion of transposable elements, further influencing alternative splicing regulation.

The expansion of non-coding genomic regions has played a key role in the evolutionary development of alternative splicing. The evolution of coding sequences is not uniform across genes; essential genes involved in transcription, translation, and metabolic processes tend to evolve more slowly, while genes related to immune responses and environmental adaptation undergo faster rates of sequence change (***Cavalier-Smith, 2005***). It has been revealed that highly expressed genes evolve at slower rates due to their structural stability and efficient folding (***Koonin, 2009***). Selection pressure against protein misfolding constrains sequence variation, particularly in essential genes, such as those encoding core cellular machinery. As a consequence, the constraints imposed on coding sequences also vary across the major groups, with prokaryotes displaying higher sequence conservation due to compact genomes and high selective pressures. In contrast, multicellular eukaryotes exhibit higher plasticity, reflecting the impact of relaxed selection and functional specialization (***Koonin and Wolf, 2010***). Comparative genomic studies have revealed that while coding sequences are subject to strong functional constraints, non-coding elements exhibit greater evolutionary plasticity, often expanding through genome duplications and transposable element activity (***Ahnert et al., 2008***; ***Silva et al., 2020***). Non-coding sequences introduce new regulatory elements that influence alternative splicing and gene expression patterns (***Veena S. Patil and Rana, 2014***; ***Koonin and Wolf, 2010***; ***Wolf and Koonin, 2013***; ***Li et al., 2024***). In particular, increased intron length correlates with greater transcriptomic complexity, as it is involved in the fine-tuning of gene expression through alternative splicing mechanisms (***Oesterreich et al., 2016***).

Together, these observations highlight the complex interplay between non-coding sequence

expansion and splicing regulation, although the precise relationship between genome architecture and alternative splicing dynamics remains to be fully elucidated. This gap motivates the need for standardized, genome-wide metrics to better understand splicing patterns across species. Advances in high-throughput sequencing technologies have revealed an unprecedented level of transcript diversity, uncovering previously unrecognized splicing events across multiple species and contributing to our knowledge of its regulatory mechanisms (***Shapiro, 2017***). Several studies have developed databases for alternative splicing analysis (***Kim et al., 2006***; ***Itoh et al., 2004***; ***Holste et al., 2006***; ***Kim et al., 2006***). However, while these studies have uncovered important gaps, they remain isolated findings. Most of these databases focus on model organisms, such as human and mouse genomes, and primarily focus on conserved splicing events. In addition, most comparative studies rely on orthology-based methods and multiple sequence alignment. While these techniques are effective for detecting evolutionary conservation in alternative splicing patterns, these methodologies restrict their applicability in large-scale comparative analyses across distantly related taxa. To date, no study has integrated the vast body of knowledge generated. There is still no standardized metric available for cross-species comparisons of alternative splicing. Here, we propose a whole-genome measure, the *Alternative Splicing Ratio*, which can be computed from high-quality annotation files such as those provided by the NCBI (***Pruitt et al., 2012***). Additionally, we examine methodological biases in the NCBI annotation files, evaluating how factors such as RNA-Seq depth, computational approaches, and tissue diversity influence alternative splicing detection. Thus, we provide a standarized measure that minimizes these biases, providing a reliable framework for cross-species comparisons. This study represents the first large-scale comparative genomic analysis that integrates data from alternative splicing events across more than a thousand species, encompassing the entire tree of life, and offering a comprehensive perspective on their evolutionary relationship.

Results

Genome variables

The compositional structure of genomes is organized into hierarchical domains, reflecting distinct levels of DNA sequence organization. To characterize this structure, we consider three key genomic variables: genome size, gene content, and coding content. Within the genome, the gene content represents the total amount of DNA that forms genes, encompassing both coding and non-coding regions. Coding content, in turn, refers to the amount of DNA within genes that is transcribed into mRNA. Unlike untranslated regions (UTRs) or intronic sequences, Coding DNA Sequences (CDSs) are composed of sequences that contribute directly to mRNA synthesis.

Alternative splicing is the process by which different combinations of exons and introns are selectively retained or excluded during pre-mRNA processing, resulting in the production of distinct mRNA isoforms. Here we propose a novel genome-wide metric, the *Alternative Splicing Ratio* (ASR), which quantifies the average number of distinct transcripts generated per coding sequence. The ASR is mathematically defined in Methods and schematically illustrated in ***Figure 1***. Although it is computable for single genes, it is extended to the entire genome by mapping every transcribed CDS detected in a species onto its genomic coordinates. Thus, it provides a global quantification of alternative splicing, capturing the extent to which coding sequences are reused across the entire transcriptome. By summarizing this complexity into a single numerical value, it enables large-scale comparative analyses across species, facilitating the study of transcriptomic diversity and evolutionary patterns of splicing regulation. As described in Methods, ASR was computed for 1,494 species spanning the entire tree of life. Additionally, to correct for annotation-related biases, we computed a normalized value, ASR^*, that accounts for inconsistencies in genome annotations across species. This normalization was necessary to mitigate biases introduced by differences in

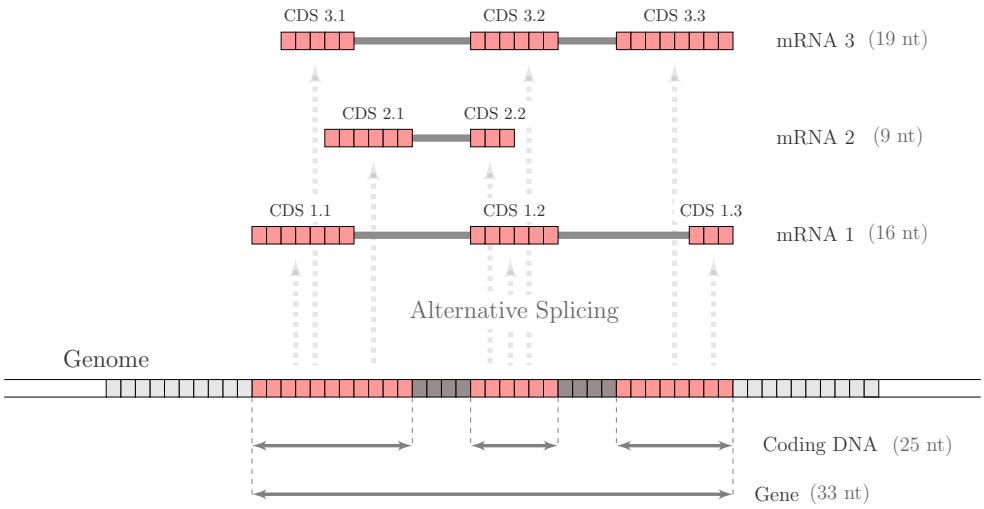

Figure 1. Schematic representation of the compositional structure of a gene and alternative splicing. The selective combination of exons and introns in a gene of 33 nucleotides gives rise to three distinct mRNA isoforms: mRNA1 (16 nucleotides), mRNA2 (9 nucleotides), and mRNA3 (19 nucleotides). When these CDSs are mapped onto the genome, the coding DNA—defined as the DNA sequences that are transcribed into a mRNA—is found to be composed of 25 nucleotides. The ASR is then computed as the ratio of the total number of nucleotides in mRNA isoforms to the number of nucleotides composing coding DNA: $ASR = (16 + 9 + 19)/25 = 1.76$.

sequencing depth, tissue diversity, assembly quality, and computational gene prediction, ensuring more accurate cross-species comparisons.

Variation in Alternative Splicing Among Major Clades

We performed a comparative analysis to identify variations in alternative splicing across different taxonomic groups. We conducted comparisons of mean values using Welch's ANOVA with Bonferroni correction to control for multiple testing. To further validate our findings, we also conducted a Monte Carlo permutation test, which supported the conclusions (see Methods). Due to the presence of extreme outliers, we conducted the permutation test by comparing both the mean and the median. These two measures of central tendency provide complementary insights into the data distributions, which are non-normal in most cases. The mean, while sensitive to extreme values, gives an overall measure of centrality, whereas the median offers a more robust estimate that is less influenced by outliers. Quantitative results of the Monte Carlo permutation test are presented in Table 1 of Supplementary File 1.

To explore how alternative splicing varies across the tree of life, we first compared the distribution of ASR and ASR* values across a broad range of taxonomic groups (***Figure 2***). The results revealed significant differences for most comparisons, with only few exceptions. Overall, unicellular organisms—including archaea, bacteria, fungi, and unicellular eukaryotes—tend to exhibit comparable levels of alternative splicing, with particular similarity observed between unicellular eukaryotes, fungi, and bacteria. Among multicellular taxa, mammals and birds display similar ASR values, as do arthropods and fish. Slight differences were observed in the comparisons when using the normalized ASR value (ASR*). In particular, we found that plants exhibit values comparable to those of fish, despite their distant evolutionary relationship. Similarly, arthropods and birds also display similar ASR* median values. As shown in Table 3 of Supplementary File 1, the average ASR values indicate that mammals exhibit the highest alternative splicing activity (3.1), followed by birds

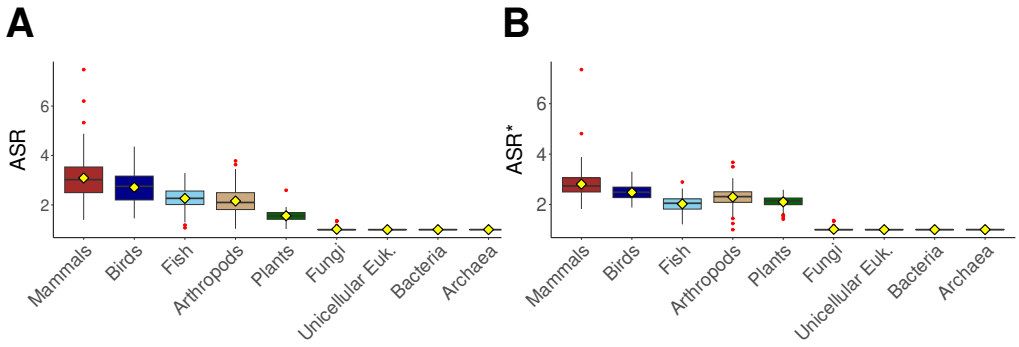

Figure 2. Comparative analysis of (A) alternative splicing ratio (ASR) and (B) normalized alternative splicing ratio (ASR*) distributions across taxonomic groups, including mammals, birds, fish, arthropods, plants, fungi, unicellular eukaryotes, bacteria, and archaea. Box plots represent the median (horizontal line), interquartile range (IQR, box), and whiskers extending to $1.5\times$ IQR. A yellow diamond-shaped point represents the mean, and outliers are shown as individual red points. Colors in the box plots correspond to taxonomic classifications.

(2.7), fish and arthropods (2.2), plants (1.6), and unicellular organisms (1), the latter indicating minimal or no alternative splicing. When normalized ASR* values are considered, the ranking among multicellular taxa is maintained, with mammals showing the highest values (2.8), followed by birds (2.5), arthropods (2.3), and both fish and plants with values around 2.

In addition to alternative splicing, we also examined whether certain genomic composition variables differ across major taxonomic groups. Specifically, we focused on the proportion of the genome composed of genes, the proportion of coding sequences within genes, and the overall proportion of coding sequences within the genome. By assessing whether these genomic variables also distinguish taxonomic groups, we can evaluate to what extent variation in alternative splicing values may be influenced by genomic composition. Consistent with this, we found significant differences across most taxonomic groups in both mean and median values of the genomic variables (see Table 2 in Supplementary File 1). As reported in Table 3 of Supplementary File 1, bacteria and archaea exhibit the highest gene content proportions, with genes accounting for approximately 86% of their genomes. This percentage is followed by fish, arthropods, fungi, and unicellular eukaryotes, all of which show similar percentages of approximately 61%, with no significant differences among them. Birds (53%) and mammals (45%) show moderate gene content, whereas plants have the lowest proportion among all groups, at around 25%. On the other side, the coding-to-gene proportion increases progressively across taxa as follows: mammals (3%), birds (5%), fish (9%), arthropods (11%), plants (27%), fungi and unicellular eukaryotes (88%), and finally, bacteria and archaea, each with approximately 99%.

Evolutionary Patterns of Alternative Splicing

Understanding genomic expansion patterns and its relationship with alternative splicing across taxa is key to determining whether genome evolution follows universal constraints or taxon-specific trajectories. To systematically address this question, we conducted two complementary analyses. On one side, we performed a Phylogenetic Generalized Least Squares (PGLS) to identify possible relationships among genomic percentages and alternative splicing while accounting for phylogenetic dependence (see Table 4 of Supplementary File 1). On the other side, we analyzed the relative variability of each genomic variable across taxonomic groups using the coefficient of variation (Tables 5 and 6 of Supplementary File 1). First, we computed the relative variability of each genomic feature x, (CV_x), providing a means to quantify the extent of variation within each taxonomic group.

Second, we computed variability ratios between pairs of genomic features, (CV_x/CV_y), allowing us to compare the variability of one genomic trait to another within and across taxa. A detailed explanation of both methods can be found in the Methods section.

PGLS analyses revealed no significant correlation between alternative splicing and genome size across any taxonomic group for both ASR and ASR* values. However, in mammals and birds, alternative splicing showed a strong association with both gene and coding content. This association further extends to genomic proportions, particularly the ratio of coding to gene content. These correlations remain consistent for both ASR and ASR*, except in the case of coding content, which does not correlate with ASR* in any taxonomic group. These findings indicate that, for mammals and birds, genomes with larger amounts of gene content and coding sequences exhibit higher levels of alternative splicing. When analyzing relative proportions rather than absolute content, we found that alternative splicing (ASR and ASR*) is strongly associated with two key genomic ratios: the genomic fraction occupied by genes and the proportion of coding sequence within genes. In the latter case, we observe a strong negative correlation, supporting the hypothesis that genes with a higher intronic fraction are more likely to give rise to multiple isoforms through alternative splicing.

Notably, in contrast to mammals and birds, neither gene content nor coding content shows an association with alternative splicing (ASR or ASR*) across the other taxonomic groups examined. The only exception is observed in plants, where gene content shows a positive association with ASR*, albeit considerably weaker than the correlation observed in mammals and birds. When considering genomic proportions, however, the analysis reveals correlations in some lineages, although differences are observed between ASR and ASR*. Nonetheless, these associations are generally weaker than those observed in mammals and birds, suggesting that the association between genomic architecture and alternative splicing is stronger in these two groups. In fish, the proportion of genes within the genome is positively associated with both ASR and ASR*. In addition, gene proportion is positively associated with ASR in arthropods and with ASR* in plants. The coding-to-gene ratio appears as a relevant predictor of alternative splicing ratio, correlating with both ASR and ASR* in plants and fungi. Furthermore, its correlation with ASR* is also significant in arthropods and archaea. Another genomic proportion indicative of the genome structure is the coding-to-genome ratio. While ASR shows no significant correlation with this ratio in any taxonomic group, ASR* exhibits a positive correlation in both arthropods and plants. Overall, these findings indicate that the genomic architecture of coding and gene content, and not only their abundance, may play a central role in shaping splicing complexity across most multicellular lineages, particularly in mammals and birds. In these groups, alternative splicing is strongly associated with intron-rich genes, and the results may reflect a shared, lineage-specific trajectory in the evolution of transcriptomic complexity.

Beyond lineage-specific associations—which may contribute to broader patterns—an overarching trend is observed when examining variability across taxa. We first quantified the relative variability of alternative splicing using the coefficient of variation, which expresses dispersion normalized by the mean, making it independent of scale. Unicellular organisms exhibit negligible variability in alternative splicing, as expected given their minimal or absent splicing activity. Within these groups, only fungi show a slight degree of variation, though still close to zero. In contrast, multicellular taxa display considerable variability in alternative splicing values, with clear differences among groups. Among them, plants consistently exhibit the lowest alternative splicing variability, both in ASR and normalized ASR* measures. Arthropods, fish, and birds show higher variability than plants in ASR; however, after normalization in ASR*, their variation is only slightly higher than plants. Finally, mammals exhibit the highest levels of variability in both ASR and ASR*, indicating a high degree of heterogeneity in splicing activity. Also, as a measure of relative variability, we used the ratio of coefficients of variation to compare how much alternative splicing values vary in

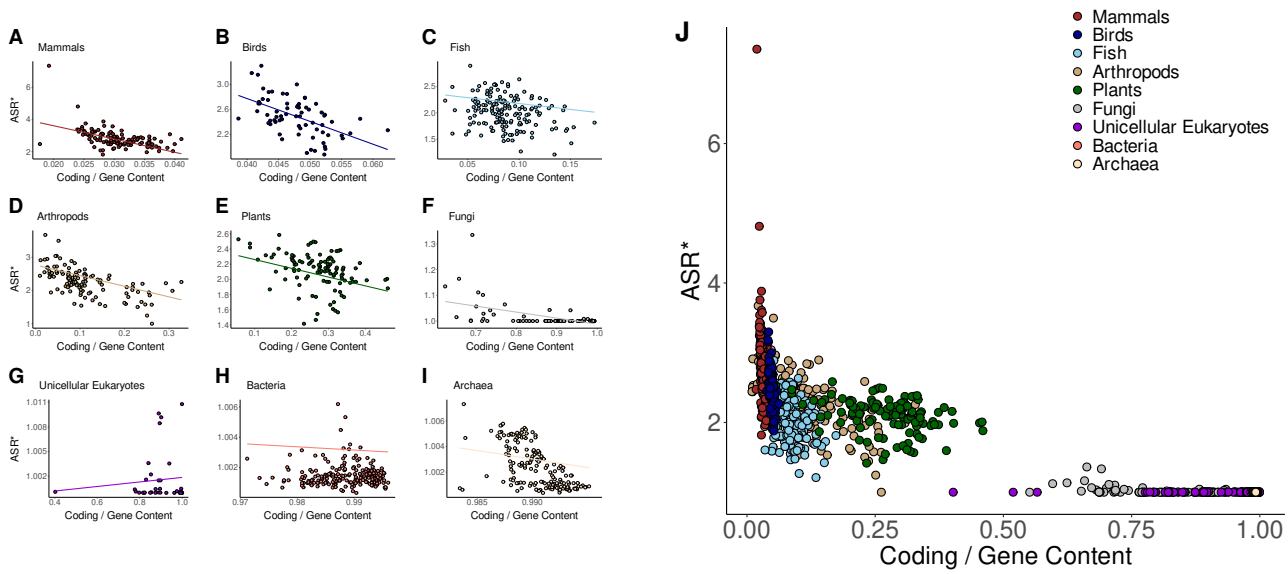

Figure 3. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the proportion of coding relative to gene content and the normalized ASR* across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

relation to genomic features. Specifically, the variability ratio between ASR and the percentage of coding within genes, $\hat{CV}_{ASR}/\hat{CV}_{Coding/Gene}$, follows a progressive trend from unicellular organisms to animals, with mammals and birds exhibiting the highest ratios. These relations are visually reflected in ***Figure 3*** and Fig. 5 of Supplementary File 2. A variability ratio above 1 indicates that alternative splicing is more variable across species within a taxonomic group than the proportion of coding. Notably, only birds and mammals display values exceeding 1 for both ASR and its normalized value ASR*, highlighting a greater degree of interspecific heterogeneity in splicing relative to the proportion of coding composing genes. As a consequence, these lineage-specific patterns of variability in splicing activity, and their strong association with gene architecture, may reflect specific regulatory strategies in mammals and birds. At the opposite end of the spectrum, unicellular organisms exhibit uniformly low variability ratios, indicating that the percentage of coding varies more across species than alternative splicing levels. This pattern is consistent with the near absence of alternative splicing in these groups, which show little to no variation. Among them, unicellular eukaryotes display the lowest variability ratio, approaching zero, reflecting an almost complete lack of variability in splicing relative to coding content. Between these two opposite trends, fish, plants, and arthropods exhibit intermediate variability ratios, all around 0.5 for both ASR and ASR* values. Although slight differences exist, the decreasing trend from fish to plants and arthropods remains consistent, indicating that these groups display a balanced degree of variation in alternative splicing compared to coding content. This result suggests a continuum between two distinct strategies: one characterized by coding-rich genomes with minimal alternative splicing, as observed in prokaryotes, and another in which small changes in intron-rich gene architectures are coupled with high variations in alternative splicing levels—as seen in mammals and birds.

A phase-like transition in alternative splicing activity is observed at approximately 20 Mb of coding DNA, a value that appears to be a threshold separating unicellular and multicellular organisms. As shown in Figure 3 of Supplementary File 2, species below this threshold exhibit negligible alternative splicing values. Beyond this threshold, which is exceeded in all multicellular taxa, alternative

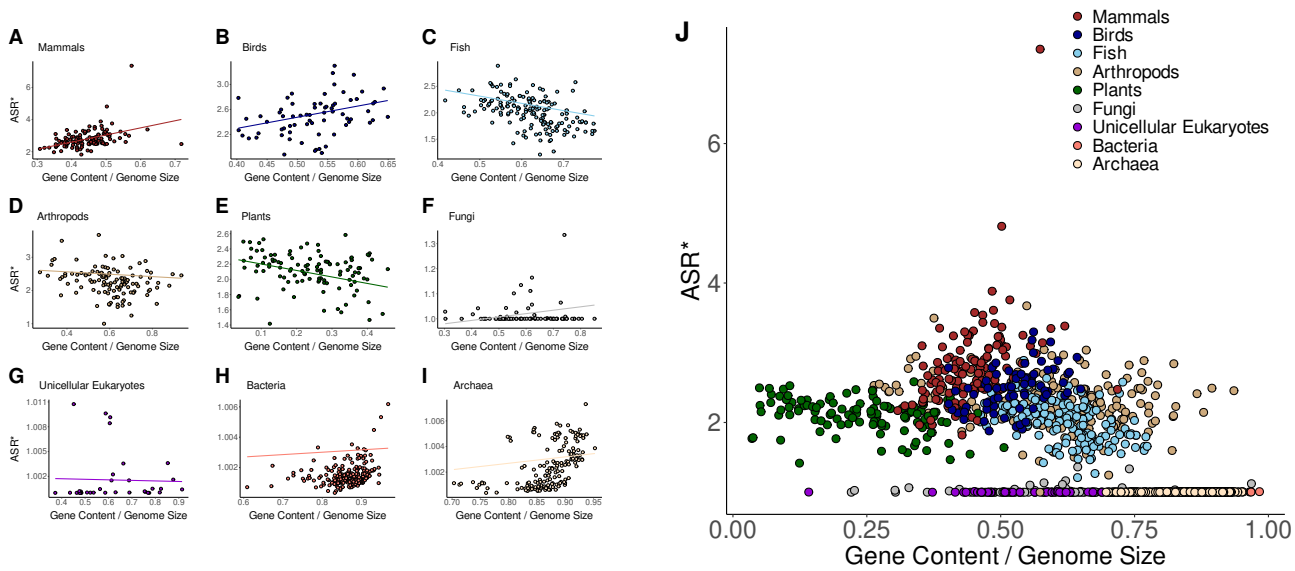

Figure 4. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the proportion of gene content relative to genome size and the normalized ASR* across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

splicing levels increase. Still, alternative splicing values vary substantially across species, indicating that while a shift toward more complex post-transcriptional regulatory strategies is common among multicellular organisms, the degree of splicing activity is modulated by different factors, including fine-scale differences in gene architecture.

Finally, as previously observed, a significant relationship exists between alternative splicing and the proportion of gene content in most multicellular groups, suggesting that intergenic regions may play a relevant role in shaping splicing dynamics. Otherwise, it may reflect an indirect association with the regulatory architecture of alternative splicing. These patterns are visually represented in ***Figure 4*** and Figure 6 of Supplemetary File 1. The results further indicate that both ASR and ASR* reach their highest values in genomes with approximately 50% intergenic content, supporting the hypothesis that this genomic configuration may favor optimal conditions for alternative splicing. Notably, while mammals and birds exhibit a positive association between ASR* and the proportion of gene content, fish and plants display an inverse pattern, with regression slopes shifting from positive to negative. In mammals, where genes represent less than 50% of the genome, ASR* increases with the percentage of genic content, peaking at 50%. In contrast, fish—whose genomes often exceed this threshold—further gene expansion is no longer positively associated with splicing activity, potentially indicating an evolutionary constraint on genome organization. After examining the correlation between genic proportion and ASR*, we next explored the relationship between the number of genes and ASR. This analysis revealed strong positive correlations in mammals, birds, fish, and arthropods. Altogether, these results suggest the existence of an optimal intergenic composition that enhances alternative splicing.

Discussion

A Whole-Genome Approach for Quantifying Alternative Splicing

In this study, we propose a whole-genome measure to quantify alternative splicing, which can be computed using genome annotation files such as those provided by the NCBI Genome Annotation

Pipeline. This measure is robust, as we exclusively consider high-quality genomic data from assemblies at the chromosome and whole-genome levels. Since genomes with poor assembly quality are not considered, potential biases in the annotation of alternative isoforms due to low-quality data are minimized. Furthermore, the NCBI pipeline applies the same computational model to annotate both animal and plant genomes, so we also minimize potential biases arising from differences in annotation methodologies between these groups. Our evaluation of methodology-associated biases introduced by the NCBI annotation pipeline reveals that the only significant bias is related to the level of experimental support, with better-supported annotations showing higher alternative splicing estimates. Specifically, the number of CDSs introduced in the pipeline is positively correlated with higher alternative splicing values. As a consequence, this factor represents a potential source of bias. To address this, we apply a normalization to ASR values, effectively correcting for discrepancies introduced by differences in annotation confidence and RNA-Seq coverage across taxa. Furthermore, taxonomic comparisons and lineage-specific patterns in alternative splicing yielded similar results for both ASR and the normalized metric ASR*, reinforcing the validity of ASR as a robust metric for large-scale comparative analyses.

While alternative splicing represents a prominent mechanism of transcriptomic diversification, it should be considered that it constitutes only one component of the broader landscape of gene regulation. Structural and behavioral complexity in organisms arises from a combination of regulatory processes—including transcriptional control, chromatin remodeling, epigenetic modifications, and RNA editing—whose combined interactions ultimately shape phenotypic diversity (***Levine and Tjian, 2003***; ***Shlyueva et al., 2014***; ***Feschotte, 2008***). In this study, we focus specifically on alternative splicing as a measurable and comparable proxy for regulatory complexity at the genome level. However, we recognize that a more comprehensive analysis of organismal complexity would require the integration of additional layers of regulation and functional data across lineages.

Evolutionary Strategies of Transcriptomic Diversification

While previous studies have highlighted the role of splice site strength, exon length, and regulatory element density in shaping splicing outcomes (***Bortfeldt et al., 2008***; ***Keren et al., 2010***; ***Itoh et al., 2004***), our results provide a genomic-level perspective that contextualizes alternative splicing within large-scale patterns. Specifically, we find that alternative splicing activity is not randomly distributed among taxa but follows lineage-specific associations with genome composition, particularly with respect to the proportion of coding and non-coding DNA. In mammals and birds—groups that exhibit the highest alternative splicing values—we observe a strong negative correlation between alternative splicing and the proportion of coding DNA within genes. This indicates that intron-rich gene architectures may promote isoform diversity, consistent with models in which exon skipping predominates and weak splice sites are compensated by dense regulatory motifs (***Kalsotra and Cooper, 2011***). Interestingly, these groups exhibit a highly conserved gene structure, with coding regions representing less than 5% of gene length, while alternative splicing levels display considerable variability across species. Thus, within a narrow coding-to-gene compositional range, alternative splicing may be finely regulated, with minor variations in gene structure associated with substantial changes in isoform diversity. This pattern suggests that transcriptomic modularity could be achieved through subtle adjustments in gene architecture, potentially reflecting an evolutionary strategy favoring regulatory refinement over gene duplication. However, future research should investigate the mechanistic basis of this relationship and test whether such fine-tuned structural control indeed underlies shifts in splicing dynamics. In line with models describing alternative splicing as a modular system of transcript generation, this genomic strategy allows functional diversification without the need for coding expansion (***Modrek and Lee, 2002***; ***Nilsen and Graveley, 2010***). These findings also align with recent theoretical work proposing that alternative splicing and regulatory complexity increase as coding accumulation slows down, acting as an evolutionary response to structural constraints (***Koonin, 2011***).

Our results reveal a threshold at approximately 20 Mb of coding DNA, separating unicellular from multicellular organisms. Below this coding size, species display a lack of alternative splicing, whereas above it, splicing activity becomes increasingly significant. This observation supports the hypothesis that coding expansion reaches a critical point, beyond which additional genomic innovation depends less on gene duplication and more on regulatory diversification, particularly through alternative splicing (***Lynch, 2007***). Importantly, while coding content increases in multicellular organisms, the relative proportion of coding sequences within genes decreases sharply, especially in mammals and birds. As illustrated in ***Figure 5***, once the proportion of coding drops below 50%, alternative splicing emerges significantly, reaching its highest levels in coding-poor organisms. This result is in agreement with previous studies showing that a decrease in coding density is strongly associated with increased splicing activity (***Lynch and Conery, 2000***; ***Holland et al., 2017***; ***Hinman and Cary, 2017***). As coding expansion slows down, the reliance on alternative splicing intensifies, indicating a shift toward post-transcriptional plasticity as a dominant evolutionary strategy (***Ast, 2004***; ***Rogozin et al., 2012***). This shift may be key for shapping broader evolutionary patterns in the relationship between genome composition and transcriptomic complexity across lineages. These findings also support the hypothesis that the slowdown in coding sequence accumulation is intrinsically linked to the emergence of post-transcriptional strategies, such as alternative splicing.

There is a large-scale progressive trend in the variability relationship between coding-to-gene composition and alternative splicing across evolutionary lineages. At one end of the spectrum, prokaryotes exhibit highly compact genomes, with minimal non-coding DNA and negligible alternative splicing activity. In the absence of alternative splicing, these organisms rely on gene duplication as the primary mechanism for functional innovation (***Lynch, 2007***). At the opposite end, mammals and birds show the highest levels of alternative splicing (***Barbosa-Morais et al., 2012***), despite having highly constrained genomic architectures. Between these extremes, plants, arthropods, and fish occupy intermediate positions, with high variability in both coding percentage and splicing levels. This finding supports the existence of two contrasting strategies: one that favors expansion through gene duplication and coding sequence accumulation (as in prokaryotes and some unicellular eukaryotes), and another that maximizes regulatory reuse and complexity through alternative splicing in genomically constrained organisms with high functional plasticity, such as mammals and birds. As a consequence, differences in variability relationships across taxa may be attributed to fundamental shifts in genome expansion strategies. These results support the view that alternative splicing is not merely a means of increasing functional complexity, but also an evolutionary adaptation to structural constraints within the genome.

Our findings also reveal that alternative splicing activity peaks in genomes with approximately 50% of intergenic content, supporting the hypothesis that there exists an optimal genic-to-intergenic balance that maximizes regulatory efficiency. While intergenic regions may contribute to alternative splicing activity, their excessive expansion, as observed in plants, does not result in increased alternative splicing levels. Thus, a threshold of approximately 50% of intergenic content may represent a point beyond which additional non-coding DNA becomes functionally redundant. Thus, rather than promoting continued increases in isoform diversity, expansive intergenic growth may impose architectural limitations that reduce the utility of additional regulatory space. This pattern is consistent with previous work indicating that transcriptomic complexity does not scale linearly with genome size(***Hahn and Wray, 2002***; ***McShea, 1996***; ***Claverie, 2001***; ***Choi et al., 2020***) and may instead be constrained by saturation effects in regulatory architecture (***Muro et al., 2025***; ***Luiselli et al., 2025***; ***Shapiro, 2017***). Future research should explore the mechanisms underlying this apparent intergenic threshold, testing whether regulatory saturation or spatial constraints limit the contribution of additional non-coding regions to alternative splicing complexity.

Plants exhibit a distinct evolutionary strategy compared to animals, particularly in the way non-coding regions have expanded. In contrast to prokaryotes, unicellular eukaryotes, and most animals, where genes constitute over 40% of the genome, plant genomes show a significantly lower gene content, ranging from as little as 7% to a maximum of 40%. This reduction is accompanied by a pronounced expansion of intergenic regions, which varies widely across plant species. Despite this extensive non-coding accumulation, plants display only moderate levels of alternative splicing, suggesting that increased intergenic space does not necessarily enhance transcriptomic complexity (***Chamala et al., 2015***). Intron retention is the predominant mechanism of alternative splicing in plants, whereas exon skipping dominates in animals. Splice sites in plants are also generally weaker and lack the compensatory enrichment of regulatory sequences, such as exonic splicing enhancers, commonly observed in metazoans. These differences imply that alternative splicing in plants may be governed by distinct regulatory constraints, likely shaped by evolutionary processes such as polyploidy, genome duplication, and differential selection on non-coding expansion (***Reddy et al., 2013***; ***Filichkin et al., 2010***).

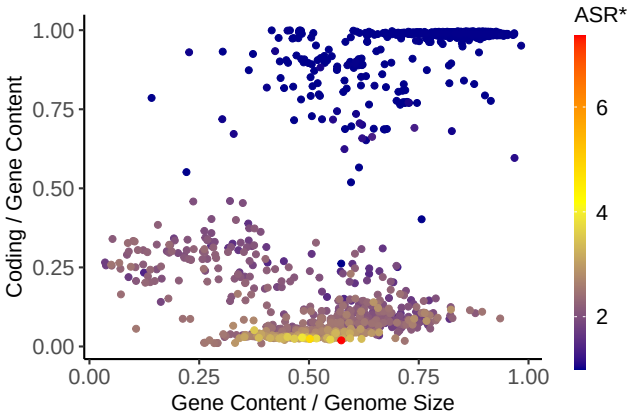

Figure 5. Normalized Alternative Splicing Ratio, ASR*, is represented as a color gradient across different genomic profiles. The x-axis spans the gene-to-genome proportion, while the y-axis spans the coding-to-gene proportion.

Evidence for the Adaptive Role of Alternative Splicing

One of the major debates regarding alternative splicing is whether it arises as an adaptive trait or as a byproduct of reduced selection efficiency (***Lynch, 2007***). Species with smaller effective population size experience weaker purifying selection, allowing the accumulation of slightly deleterious mutations, including those affecting alternative splicing. This "non-adaptive model" suggests that alternative splicing increases in complex organisms not due to selective advantages but as a consequence of weaker selection pressure in species with small population sizes. However, genomic analyses contradict this view, revealing that alternative splicing is not randomly accumulated but is rather highly conserved in some genes, making it functionally relevant across taxa (***Rogalska et al., 2024***). Advances in long-read RNA sequencing technologies have revealed that while many alternative splicing events may be non-functional or represent splicing errors, a significant subset leads to functionally relevant protein isoforms that are preferentially expressed in specific tissues (***Wright et al., 2022***). Several studies have revealed the conservation of alternative splicing patterns across different animal genomes, such as human, mouse, rat, chicken, fish, and insect genomes, suggesting that splicing regulation plays a fundamental role in evolutionary adaptation (***Kim et al., 2006***; ***Holste et al., 2006***). These events are particularly enriched in genes associated with essential biological processes and functionally critical organs, such as the brain, heart, muscles, and testes. This highlights their role in preserving key physiological functions while also contributing to evolu-

tionary adaptation. On the other hand, beyond tissue specialization, alternative splicing is increasingly recognized as a key factor in rapid environmental adaptation, allowing organisms to fine-tune gene expression and modulate protein isoform production in response to external stimuli (***Koonin, 2006***). Stress conditions, such as temperature fluctuations and hypoxia, have been observed to regulate alternative splicing patterns, helping organisms adjust their physiological responses dynamically (***Shapiro, 2017***; ***Aravind et al., 2009***). Another example is found in genes associated with the immune system, where alternative splicing modulates immune responses to pathogens, further emphasizing its role in adaptive flexibility. Beyond environmental adaptation, it also plays a major role in biological differentiation 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***).

The findings presented in this study support the view that alternative splicing is not simply a byproduct of relaxed purifying selection, but instead reflects a finely regulated mechanism linked to genome architecture. While the non-adaptive model states that splicing diversity increases stochastically due to weaker selection in species with small effective population sizes, our results show that alternative splicing levels are strongly associated with genomic architecture, particularly with the proportion of coding sequences within genes. Mammals and birds exhibit highly conserved, intron-rich gene structures, in which the proportion of coding sequences within genes remains low and stable across species. Despite this structural conservation, these groups display substantial interspecific variability in alternative splicing levels, but this variability is strongly and negatively correlated with the proportion of coding DNA within genes, indicating that even small differences in coding compression are associated with pronounced differences in splicing activity. These results suggest that alternative splicing is highly sensitive to structural variation, rather than accumulating stochastically. It may act as a modular regulatory system that enables transcriptomic diversification without requiring further expansion of coding sequences. Together with prior evidence of tissue-specific conservation and adaptive responses mediated by alternative splicing, our results reinforce its role as a key evolutionary mechanism that integrates structural features of the genome with regulatory plasticity. However, more comprehensive research is needed to determine whether these patterns reflect causal relationships. Future studies should integrate high-quality annotations, functional assays, and comparative analyses to fully disentangle the relative contributions of structural constraints and regulatory complexity.

Methods

Alternative Splicing Ratio: A Genome-Scale Measure

The genome is defined as the sequential arrangement of nucleotides, where the ith position of a nucleotide is denoted as $G(i)$. We define $f(i,j)$ as the number of times the nucleotides $G(i)$ and $G(j)$ appear together in a CDS. By way of example, a value of $f(i,j) = 3$ indicates that the nucleotides at positions i and j co-occur in three distinct isoforms. Accordingly, $f(i,i)$ corresponds to the number of different CDSs where nucleotide $G(i)$ is inserted. The *transcription matrix* M, with $M_{ij} = f(i,j)$, is a symmetric matrix that encodes the co-occurence patterns of nucleotides within multiple mRNAs, providing a genotype-phenotype mapping framework for transcriptomic and proteomic diversity. The binary version of the transcription matrix is defined as:

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

The diagonal of this matrix represents the projection of CDSs onto the genome, which maps the nucleotides that constitute the coding DNA. Thus, the *Alternative Splicing Ratio* is mathematically

defined as:

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

where $tr(\cdot)$ denotes the trace of the matrix, i.e., the sum of the diagonal elements. It quantifies the extent to which coding sequences can be reutilized through alternative splicing to generate multiple mRNA isoforms. It provides a measure of transcriptomic modularity, reflecting the parallelization of genetic information. Here, we extend the concept of the Alternative Splicing Ratio from individual genes to the whole genome by computing the transcription matrix using all nucleotides that compose genes.

NCBI RefSeq Dataset

Genomic features are computed from high-quality, whole-genome assemblies available in the RefSeq database provided by the NCBI (***Database, 2024***). These assemblies are annotated by the NCBI genome annotation pipeline, which integrates diverse experimental evidence, including RNA-Seq data, ESTs, and protein alignments (***Kitts et al., 2015***). The annotation files are in GFF3 format, and the data is provided in tab-delimited tables, where each row represents an entity (e.g., gene, mRNA or CDS) and columns contain attributes like unique identifiers, genomic coordinates and functional annotations. The annotations also include additional attributes such as start and stop codons, and untranslated regions (UTRs). Genome size, gene content, and coding region coverage are estimated by looking at the start and end coordinates of each genomic element. Gene content is calculated as the total number of nucleotides falling within annotated gene regions. In contrast, the coding region coverage is determined by summing all nucleotides that belong to at least one CDS.

Each file is organized hierarchically in a parent-child structure that represents the relationship between genes, their transcripts (mRNA), and coding sequences (CDS). At the top level, each gene is uniquely identified and associated with multiple mRNA transcripts, which represent the alternative splicing variants of that gene. These mRNA transcripts, in turn, are linked to one or more CDS, which correspond to the portions of the mRNA that are translated into proteins. The set of CDSs associated with a given mRNA is retained in the spliced mRNA, defining a distinct protein isoform. Hierarchical relationships are preserved through shared identifiers across levels: a gene ID links to its transcripts, and a transcript ID links to its associated CDS entries. This structured format allows for efficient navigation and analysis of the dataset, supporting downstream applications like isoform quantification. Note that we have only considered CDSs with a well-defined parent-child relationship. Pseudogenes and duplicated genes have been excluded for this study.

In this study we propose a novel genome-level measure of alternative splicing. In annotation files, each isoform is identified by looking at the set of CDSs associated with a given mRNA transcript. Since each gene can produce multiple protein isoforms, the corresponding mRNA transcripts often contain regions with overlapping CDSs. Thus, the metric is calculated as the ratio of the total cumulative length of all annotated CDSs to their length when projected onto the genome. This approach provides novel quantitative insights into estimating alternative splicing activity at the species level.

First, the assembly summary of RefSeq was downloaded, which includes detailed information about the genome annotations available in the database (***Reports, 2024***). Second, the dataset was subsequently filtered based on the following criteria: (i) only genomes annotated by the NCBI RefSeq annotation pipeline were selected, ensuring consistent and reliable annotations; (ii) assemblies were restricted to those with an assembly level of 'chromosome' or 'complete genome,' which guarantees high-quality data; and (iii) genomes were further filtered to include only those belonging to specific taxonomic groups of interest (see Section Taxonomy assignment). Finally, the filtered an-

notation files were downloaded in November of 2024, and the genomic features were computed for each species.

Taxonomy assignment

We collected annotation files for organisms representing the entire tree of life. However, the sampling process was constrained by the availability of whole-genome annotation files, resulting in a dataset of 694 eukaryotic species. Taxonomy tools from the NCBI (***Sayers et al., 2023***) were utilized to map the phylogenetic relationships among species that meet conditions (i) and (ii) described in the previous section. Next, iTOL was employed for the visualization and schematic representation of these relationships (***Letunic and Bork, 2021***). Taxonomic groups at the levels of kingdom, phylum, and class containing at least 20 distinct species were selected. This approach resulted in the identification of the following major groups: 133 mammals, 77 birds, 169 fish, 187 arthropods, 128 plants, 130 fungi, and 53 unicellular eukaryotes. For the other biological domains, given that the NCBI provides annotations for thousands of species, we randomly selected 400 archaea and 400 bacteria (see ***Figure 6***). This approach ensures that the taxonomic groups are well-defined and adequately represented to provide statistical significance for subsequent analyses.

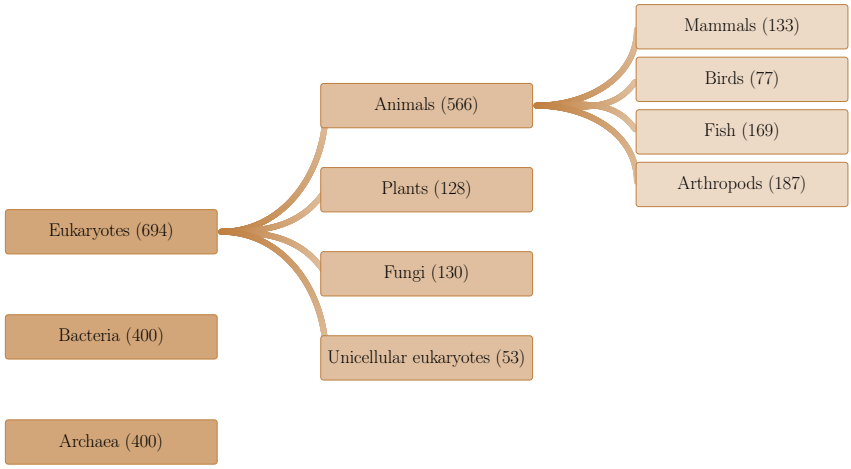

Figure 6. Organisms were classified into taxonomic groups based on the hierarchical structure provided by the NCBI Taxonomy Database, using annotation data from NCBI that meet conditions (i) and (ii) described in Section NCBI RefSeq Dataset.

Methodology-associated biases in the data

In this study, we used assembled genomes annotated by the NCBI genome annotation pipelines. The NCBI Genome Annotation Pipeline is designed for the annotation of both eukaryotic and prokaryotic genomes. On one side, the NCBI Eukaryotic Genome Annotation Pipeline (EGAP) is an evidence-based pipeline for the high-quality annotation of genes, transcripts, and proteins of eukaryotes (***Pruitt et al., 2012***). Each RefSeq assembly generated from the EGAP includes a very detailed annotation report that provides information about the supporting sequencing data and variations in methodology. All eukaryotic genomes annotated to date from the EGAP can be found in the following repository: https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/all/. On the other side, the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) is specifically designed to annotate bacterial and archaeal genomes, focusing on the unique genomic features of these domains (***Tatusova et al., 2016***). Annotations for prokaryotes rely on curated data from the NCBI RefSeq database, including known gene sequences and protein information. It processes assembled genomes, identifying coding sequences (CDSs), ribosomal RNA (rRNA), transfer RNA (tRNA), small RNAs, and other ge-

netic elements. Unlike the eukaryotic pipeline (EGAP), which incorporates complex RNA-Seq alignments and isoform modeling, PGAP emphasizes the detection of essential genes, operons, and conserved elements, as splicing is rare in prokaryotes. The prokaryotic pipeline is optimized for high-throughput processing, ensuring consistent and accurate annotation of thousands of bacterial and archaeal genomes.

While archaeal and bacterial genomes are annotated through the Prokaryotic Genome Annotation Pipeline (PGAP), annotation files for plants and animals are derived through the Eukaryotic Genome Annotation Pipeline (EGAP). However, not all RefSeq genomes have been processed through the EGAP pipeline. RefSeq annotations of fungi and unicellular eukaryotes use the annotations as submitted to the International Nucleotide Sequence Database Collaboration (INSDC), so there is variation in methodology. These organism groups are generally not annotated using RNA-Seq data and have few, if any, alternative isoforms annotated. Nevertheless, this lack of annotation does not imply that these organisms have no alternative splicing. Instead, it reflects limitations in the data and annotation methodologies.

There can be various sources of noise in the annotation of alternative isoforms. Regarding the annotation methodology, we can expect—based on previous evidence—that alternative splicing is a significant mechanism in multicellular organisms. Since the same annotation pipeline is applied to both animals and plants, it is reasonable to assume that genome annotation is not a source of noise for these eukaryotic groups and does not introduce significant biases in the number of alternative isoforms. Moreover, due to the cross-referenced and high-quality data provided by the RefSeq database, we ensure that the annotated features are reliable and comparable across diverse taxa within multicellular eukaryotic groups.

Methodology-associated biases in NCBI annotation files primarily stem from the reliance on the quality and availability of input data. Additionally, the use of reference-based annotation methods can introduce biases by propagating errors or omissions from reference genomes. Variability in sequencing technologies may also impact the accuracy of annotations, potentially leading to biased estimates of alternative splicing events. On the other hand, poor assembly quality can negatively impact the annotation of alternative splicing. However, in this study, we focus exclusively on genomes assembled at the chromosome or complete genome level, effectively reducing biases caused by assembly quality. Biases in annotation genomes also include the overrepresentation of model organisms, which leads to more accurate annotations for well-studied species. For less-studied species, genome annotations often rely on automated predictions and may lack experimental validation. RNA-Seq depth influences the annotation of alternative splicing events and lncRNAs. Consequently, the amount of annotated alternative isoforms depends on the organism, but that's primarily for humans, mice, and rats, for which we can expect some biases. We have thousands of genome annotations from a variety of species that are not model organisms, so similar biases may exist across many organisms. The amount and diversity of RNA-Seq data used for annotations are highly variable, which can also lead to significant biases. While many organisms are annotated using substantial amounts of RNA-Seq data (billions of reads), this varies considerably across organisms, and tissue diversity is often limited by data availability. Such limitations can affect downstream analyses, reducing the generalizability and reliability of conclusions drawn from these data.

For this study, we have implemented two filtering criteria to reduce methodological biases in the dataset: first, all multicellular organisms were annotated using the EGAP pipeline, which integrates evidence-based computational models; second, only genomes assembled at the chromosome level were included. Despite these measures, potential biases remain, particularly those related to RNA-Seq depth and tissue diversity. To account for potential biases introduced by the

annotation process in eukaryotic genomes, we evaluated how the experimental evidence of CDSs affects estimates of alternative splicing. We found that the proportion of coding sequences that are fully supported by experimental evidence is strongly associated with ASR values, whereas annotations mainly based on computational models systematically underestimate splicing complexity. A multivariate regression analysis identified three key predictors of ASR variation: fully supported CDSs, known CDSs, and model-derived CDSs, which are related to each other. Among them, the proportion of fully supported CDSs is the dominant factor. Based on this, we implemented a normalization procedure that adjusts ASR values using a polynomial regression model, effectively removing annotation-related biases. This normalization, which we denote as ASR*, enables more accurate cross-species comparisons while preserving relative differences in splicing levels. However, normalized values should be interpreted as standardized estimates rather than absolute measures, as the correction is empirical and designed to improve comparability across diverse taxonomic groups.

Monte Carlo permutation tests

Pairwise statistical analysis using Monte Carlo permutation tests was performed to assess differences in genomic variables (e.g., genome size, gene content, coding size, and alternative splicing ratio) across taxonomic groups. Given that some variables present a high number of outliers and, in certain cases, deviate from normality and homoscedasticity–both critical assumptions for many parametric analyses–this approach was employed as a complementary method to Welch's ANOVA, which accounts for heteroscedasticity.

This non-parametric approach evaluates statistical significance by generating a null distribution through random resampling. Specifically, we tested differences using two statistical measures: the mean and the median. First, we computed the observed difference between groups and then compared it to a null distribution generated from 10,000 random permutations, in which group labels were shuffled to simulate the null hypothesis of no difference between groups. The empirical p-value was calculated as the proportion of permuted test statistics greater than or equal to the observed difference. To account for multiple comparisons, we applied the Bonferroni correction, ensuring a stringent control of Type I error across all pairwise tests. All analyses in this study confirm that 10,000 permutations provide a robust estimation of the exact test, as increasing the number of permutations to 20,000 yielded statistically equivalent results.

Comparative Genomic Analysis

We used the TimeTree platform (timetree.org, ***Kumar et al.*** (***2022***)) to generate a phylogenetic tree, which provides evolutionary relationships based on common ancestry. The tree was saved in Newick format (.nwk). Since not all species in our dataset (for which we have annotation files) are documented on this platform, the resulting phylogenetic tree includes a subset of species. Specifically, our tree consists of 170 archaea, 213 bacteria, 35 unicellular eukaryotes, 79 fungi, 110 plants, 116 arthropods, 163 fish, 70 birds, and 122 mammals.

We conducted a Phylogenetic Generalized Least Squares (PGLS) regression to assess the relationships between genomic variables and alternative splicing ratio while accounting for phylogenetic non-independence among species. This method incorporates evolutionary relationships into the statistical model by adjusting for shared ancestry, ensuring that trait correlations are not confounded by phylogenetic structure. To quantify these relationships, we showed the slope (β), which indicates the direction of the relationship. Positive values indicate an increase in the dependent variable as the independent variable increases, and negative values indicate the opposite. We also showed the p-value, which assesses the statistical significance of this relationship, with lower values ($p < 0.05$) indicating that the observed pattern is unlikely to be due to chance. The adjusted R^2, (R^2_{adj}), measured the proportion of variance explained by the model while correcting for the

number of predictors. Finally, λ quantified the phylogenetic signal, with values close to 1 indicating strong phylogenetic dependence and values near 0 reflecting evolutionary independence. These metrics collectively allow us to determine genomic relationships and evaluate whether they are driven by taxon-specific patterns.

We also computed the coefficient of variation ($\hat{CV}$) for each variable, a scale-independent measure of dispersion. The $\hat{CV}$ is defined as the ratio of the standard deviation s to the sample mean $\bar{x}$, expressed as:

$$\hat{CV} = \frac{s}{\bar{x}}. \tag{3}$$

This metric quantifies the relative variability of a trait, making it particularly useful for comparing datasets with different units or magnitudes. A higher $\hat{CV}$ indicates greater dispersion relative to the mean, suggesting higher heterogeneity, whereas a lower value implies more constrained variability. By analyzing the $\hat{CV}$ of genomic variables and alternative splicing, we assess the extent of evolutionary variation within and across taxonomic groups. This approach allows us to detect regions of genomic composition that display higher heterogeneity, providing insights into which aspects of genome evolution are more flexible and are subject to stronger evolutionary constraints. Moreover, by comparing these patterns across taxonomic groups, we can assess whether the distribution of genomic heterogeneity follows similar trends across lineages or if taxon-specific factors shape distinct variability regimes. We introduce a metric to quantify these variability relationships, defined as the ratio of the coefficient of variation between different variables, calculated as:

$$\hat{CV}_x / \hat{CV}_y, \tag{4}$$

where $\hat{CV} = s/\bar{x}$ is the coefficient of variation of x. These ratios provide a scale-independent measure of how the variability of one genomic trait compares to another within and across taxonomic groups.

Additional files

Supplementary files

- **Supplementary File 1.** Tables summarizing the results of statistical comparisons, including pairwise tests, summary statistics, and phylogenetic regressions across taxonomic groups.
- **Supplementary File 2.** Graphical representation of local PGLS analyses and global relationships between genomic features and alternative splicing ratio across taxonomic groups.

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://doi.org/10.5281/zenodo.15189630.

Author's contribution

Repasar: Conceptualization, R.d.l.F., W.D.-V., V.A. and A.M.; methodology, R.d.l.F.; writing, R.d.l.F.; review and editing, R.d.l.F., W.D.-V., V.A. and A.M.; supervision, A.M. All authors have read and agreed to the actual version of the manuscript.

References

Ahnert SE, Fink TMA, Zinovyev A. How much non-coding DNA do eukaryotes require? Journal of Theoretical Biology. 2008; 252(4):587–592. https://www.sciencedirect.com/science/article/pii/S002251930800057X, doi: https://doi.org/10.1016/j.jtbi.2008.02.005.

Aravind L, Anantharaman V, Venancio TM. Apprehending multicellularity: Regulatory networks, genomics, and evolution. Birth Defects Research Part C: Embryo Today: Reviews. 2009; 87(2):143–164. https://onlinelibrary.wiley.com/doi/abs/10.1002/bdrc.20153, doi: https://doi.org/10.1002/bdrc.20153.

Ast G. How did alternative splicing evolve? Nature Reviews Genetics. 2004; 5(10):773–782. https://doi.org/10.1038/nrg1451, doi: 10.1038/nrg1451.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science. 2012; 338(6114):1587–1593. https://www.science.org/doi/abs/10.1126/science.1230612, doi: 10.1126/science.1230612.

Bortfeldt R, Schindler S, Szafranski K, Schuster S, Holste D. Comparative analysis of sequence features involved in the recognition of tandem splice sites. BMC genomics. 2008; 9(1):1–25. https://doi.org/10.1186/1471-2164-9-202, doi: 10.1186/1471-2164-9-202.

Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017; 372(1713):20150474. https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2015.0474, doi: 10.1098/rstb.2015.0474.

Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, Bell CJ, Bharti A, Dyhrman ST, Guida SM, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nature Reviews Microbiology. 2017; 15(1):6–20. doi: 10.1038/nrmicro.2016.160.

Cavalier-Smith T. Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. Annals of Botany. 2005 01; 95(1):147–175. https://doi.org/10.1093/aob/mci010, doi: 10.1093/aob/mci010.

Chamala S, Feng G, Chavarro C, Barbazuk WB. Genome-wide comparative analysis of alternative splicing in plants: opportunities and challenges. Genome Biology. 2015; 16(1):1–16. https://doi.org/10.1186/s13059-015-0584-2, doi: 10.1186/s13059-015-0584-2.

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. Molecular Biology and Evolution. 2014 03; 31(6):1402–1413. doi: 10.1093/molbev/msu083.

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

Choi IY, Kwon EC, Kim NS. The C-and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy. Genes & genomics. 2020; 42(7):699–714. https://doi.org/10.1007/s13258-020-00941-9, doi: 10.1007/s13258-020-00941-9.

Claverie JM. What If There Are Only 30,000 Human Genes? Science. 2001; 291(5507):1255–1257. https://www.science.org/doi/abs/10.1126/science.1058969, doi: 10.1126/science.1058969.

Database RRS, National Center for Biotechnology Information (NCBI); 2024. Accessed: 2024-10-01. https://www.ncbi.nlm.nih.gov/refseq/.

Feschotte C. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics. 2008; 9(5):397–405. doi: 10.1038/nrg2337.

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Research. 2010; 20(1):45–58. https://doi.org/10.1101/gr.093302.109, doi: 10.1101/gr.093302.109.

Fiszbein A, Kornblihtt AR. Alternative splicing switches: Important players in cell differentiation. BioEssays. 2017; 39(6):1600157. https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201600157, doi: https://doi.org/10.1002/bies.201600157.

Fu RH, Liu SP, Ou CW, Yu HH, Li KW, Tsai CH, Shyu WC, Lin SZ. Alternative Splicing Modulates Stem Cell Differentiation. Cell Transplantation. 2009; 18(9):1029–1038. https://doi.org/10.3727/096368909X471260, doi: 10.3727/096368909X471260, pMID: 19523332.

Grantham ME, Brisson JA. Extensive Differential Splicing Underlies Phenotypically Plastic Aphid Morphs. Molecular Biology and Evolution. 2018; 35(8):1934–1946. https://doi.org/10.1093/molbev/msy095, doi: 10.1093/molbev/msy095.

Hahn MW, Wray GA. The g-value paradox. Evolution & Development. 2002; 4(2):73–75. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1525-142X.2002.01069.x, doi: 10.1046/j.1525-142X.2002.01069.x.

Harr B, Turner LM. Genome-wide analysis of alternative splicing evolution among Mus subspecies. Molecular Ecology. 2010; 19(s1):228–239. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2009.04490.x, doi: https://doi.org/10.1111/j.1365-294X.2009.04490.x.

Hinman V, Cary G. Multicellularity: The evolution of gene regulation. eLife. 2017 may; 6:e27291. https://doi.org/10.7554/eLife.27291, doi: 10.7554/eLife.27291.

Holland PWH, Marletaz F, Maeso I, Dunwell TL, Paps J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017; 372(1713):20150480. https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2015.0480, doi: 10.1098/rstb.2015.0480.

Holste D, Huo G, Tung V, Burge CB. Hollywood: a comparative relational database of alternative splicing. Nucleic Acids Research. 2006 01; 34(1):56–62. https://doi.org/10.1093/nar/gkj048, doi: 10.1093/nar/gkj048.

Itoh H, Washio T, Tomita M. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. Rna. 2004; 10(7):1005–1018. doi: 10.1261/rna.5221604.

Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nature Reviews Genetics. 2011; 12(10):715–729. https://doi.org/10.1038/nrg3052, doi: 10.1038/nrg3052.

Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics. 2010; 11(5):345–355. https://doi.org/10.1038/nrg2776, doi: 10.1038/nrg2776.

Kim N, Alekseyenko AV, Roy M, Lee C. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Research. 2006 11; 35(1). doi: 10.1093/nar/gkl884.

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Research. 2015 11; 44(D1):D73–D80. doi: 10.1093/nar/gkv1226.

Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biology direct. 2006; 1(22):1–23. doi: 10.1186/1745-6150-1-22.

Koonin EV. Darwinian evolution in the light of genomics. Nucleic Acids Research. 2009; 37(4):1011–1034. doi: 10.1093/nar/gkp089.

Koonin EV. The Logic of Chance: The Nature and Origin of Biological Evolution. FT Press; 2011.

Koonin EV, Wolf YI. Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics. 2010; 11(7):487–498. doi: 10.1038/nrg2810.

Koralewski TE, Krutovsky KV. Evolution of Exon-Intron Structure and Alternative Splicing. PLOS ONE. 2011 03; 6(3):1–10. https://doi.org/10.1371/journal.pone.0018055, doi: 10.1371/journal.pone.0018055.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. TimeTree 5: An Expanded Resource for Species Divergence Times. Molecular Biology and Evolution. 2022 08; 39(8):msac174. https://doi.org/10.1093/molbev/msac174, doi: 10.1093/molbev/msac174.

Letunic I, Bork P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. 2021; 49(W1):W293–W296. doi: 10.1093/nar/gkab301.

Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003; 424(6945):147–151. doi: 10.1038/nature01763.

Li J, Ma H, Qin Y, Zhao Z, Niu Y, Lian J, Li J, Noor Z, Guo S, Yu Z, et al. Chromosome-level genome assembly and annotation of rare and endangered tropical bivalve, Tridacna crocea. Scientific Data. 2024; 11(1):186. doi: 10.1038/s41597-024-03014-8.

Luiselli J, Banse P, Mazet O, Lartillot N, Beslon G. Structural mutations set an equilibrium non-coding genome fraction. bioRxiv. 2025; https://www.biorxiv.org/content/early/2025/02/05/2025.02.03.636187, doi: 10.1101/2025.02.03.636187.

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. Correction: The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers. PLOS Biology. 2011 01; 9(1):null. https://doi.org/10.1371/annotation/2db9ee19-faa4-43f2-af7a-c8aeacca8037, doi: 10.1371/annotation/2db9ee19-faa4-43f2-af7a-c8aeacca8037.

Lynch M. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates; 2007.

Lynch M, Conery JS. The Evolutionary Fate and Consequences of Duplicate Genes. Science. 2000; 290(5494):1151–1155. https://www.science.org/doi/abs/10.1126/science.290.5494.1151, doi: 10.1126/science.290.5494.1151.

Mattick JS. Deconstructing the Dogma. Annals of the New York Academy of Sciences. 2009; 1178(1):29–46. https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2009.04991.x, doi: https://doi.org/10.1111/j.1749-6632.2009.04991.x.

McShea DW. Perspective metazoan complexity and evolution: is there a trend? Evolution. 1996 04; 50(2):477–492. https://doi.org/10.1111/j.1558-5646.1996.tb03861.x, doi: 10.1111/j.1558-5646.1996.tb03861.x.

Modrek B, Lee C. A genomic view of alternative splicing. Nature genetics. 2002; 30(1):13–19.

Muro EM, Ballesteros FJ, Luque B, Bascompte J. The emergence of eukaryotes as an evolutionary algorithmic phase transition. Proceedings of the National Academy of Sciences. 2025; 122(13):e2422968122. https://www.pnas.org/doi/abs/10.1073/pnas.2422968122, doi: 10.1073/pnas.2422968122.

Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010; 463(7280):457–463. https://doi.org/10.1038/nature08909, doi: 10.1038/nature08909.

Oesterreich FC, Bowne-Anderson H, Howard J. The contribution of alternative splicing probability to the coding expansion of the genome. bioRxiv. 2016; https://www.biorxiv.org/content/early/2016/04/11/048124, doi: 10.1101/048124.

Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 2012; 40(Database issue):D130–D135. doi: 10.1093/nar/gkr1079.

Reddy ASN, Marquez Y, Kalyna M, Barta A. Complexity of the alternative splicing landscape in plants. The Plant Cell. 2013; 25(10):3657–3683. https://doi.org/10.1105/tpc.113.117523, doi: 10.1105/tpc.113.117523.

Reports A, National Center for Biotechnology Information (NCBI); 2024. Accessed: 2024-10-01. https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS.

Rogalska ME, Mancini E, Bonnal S, Gohr A, Dunyak BM, Arecco N, Smith PG, Vaillancourt FH, Valcárcel J. Transcriptome-wide splicing network reveals specialized regulatory functions of the core spliceosome. Science. 2024; 386(6721):551–560. https://www.science.org/doi/abs/10.1126/science.adn8105, doi: 10.1126/science.adn8105.

Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biology Direct. 2012; 7:1–28. doi: 10.1186/1745-6150-7-11.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2023; 51(D1):D29–D38. https://www.ncbi.nlm.nih.gov/taxonomy.

Shapiro JA. Living Organisms Author Their Read-Write Genomes in Evolution. Biology. 2017; 6(4). https://www.mdpi.com/2079-7737/6/4/42, doi: 10.3390/biology6040042.

Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics. 2014; 15(4):272–286. doi: 10.1038/nrg3682.

Silva R, Silva JRP, Anselmo DHAL, Alcaniz JS, da Silva WJC, Costa MO. An alternative description of power law correlations in DNA sequences. Physica A: Statistical Mechanics and its Applications. 2020; 545:123735. doi: https://doi.org/10.1016/j.physa.2019.123735.

Steward RA, de Jong MA, Oostra V, Wheat CW. Alternative splicing in seasonal plasticity and the potential for adaptation to environmental change. Nature Communications. 2022; 13(1):755. https://doi.org/10.1038/s41467-022-28306-8, doi: 10.1038/s41467-022-28306-8.

Su Z, Wang J, Yu J, Huang X, Gu X. Evolution of alternative splicing after gene duplication. Genome Research. 2006; 16(2):182–189. http://genome.cshlp.org/content/16/2/182.abstract, doi: 10.1101/gr.4197006.

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Research. 2016; 44(14):6614–6624. doi: 10.1093/nar/gkw569.

Veena S Patil RZ, Rana TM. Gene regulation by non-coding RNAs. Critical Reviews in Biochemistry and Molecular Biology. 2014; 49(1):16–32. https://doi.org/10.3109/10409238.2013.844092, doi: 10.3109/10409238.2013.844092.

Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. BioEssays. 2013; 35(9):829–837. https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201300037, doi: https://doi.org/10.1002/bies.201300037.

Wright CJ, Smith CW, Jiggins CD. Alternative splicing as a source of phenotypic diversity. Nature Reviews Genetics. 2022; 23(11):697–710.

Supplementary File 1

Table 1. Pairwise comparisons of ASR and ASR* across taxonomic groups, showing differences in statistical measures (mean and median). Statistical significance was assessed using Monte Carlo permutation tests, and adjusted p-values were calculated using the Bonferroni correction (see Methods). Colors represent significance levels: red for $p \leq 0.001$ (***), green for $0.001 < p \leq 0.05$ (**), and blue for $0.05 < p \leq 0.1$ (*). Values with $p > 0.1$ have no color.

Differences in Mean Alternative Splicing Ratio (ASR)

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.38 *	0.83 ***	0.93 ***	1.53 ***	2.08 ***	2.09 ***	2.09 ***	2.09 ***
Birds		0	0.45 ***	0.56 ***	1.15 ***	1.70 ***	1.71 ***	1.71 ***	1.71 ***
Fish			0	0.11	0.70 ***	1.25 ***	1.26 ***	1.26 ***	1.26 ***
Arthropods				0	0.60 ***	1.14 ***	1.16 ***	1.15 ***	1.15 ***
Plants					0	0.55 ***	0.56 ***	0.56 ***	0.56 ***
Fungi						0	0.01	0.01 ***	0.01 ***
Uni. Euk.							0	0.00	0.00 ***
Bacteria								0	0.00 ***
Archaea									0

Differences in Median Alternative Splicing Ratio (ASR)

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.27	0.75 ***	0.92 ***	1.47 ***	2.02 ***	2.02 ***	2.02 ***	2.02 ***
Birds		0	0.49 ***	0.65 ***	1.20 ***	1.75 ***	1.75 ***	1.75 ***	1.75 ***
Fish			0	0.17	0.72 ***	1.27 ***	1.27 ***	1.27 ***	1.27 ***
Arthropods				0	0.55 ***	1.10 ***	1.10 ***	1.10 ***	1.10 ***
Plants					0	0.55 ***	0.55 ***	0.55 ***	0.55 ***
Fungi						0	0.00	0.00 ***	0.00 ***
Uni. Euk.							0	0.00 ***	0.00 *
Bacteria								0	0.00 ***
Archaea									0

Differences in Mean Normalized Alternative Splicing Ratio (ASR*)

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.33 ***	0.78 ***	0.51 ***	0.70 ***	1.79 ***	1.81 ***	1.81 ***	1.80 ***
Birds		0	0.46 ***	0.18 **	0.37 ***	1.46 ***	1.48 ***	1.48 ***	1.48 ***
Fish			0	0.27 ***	0.08	1.01 ***	1.02 ***	1.02 ***	1.02 ***
Arthropods				0	0.19 ***	1.28 ***	1.29 ***	1.29 ***	1.29 ***
Plants					0	1.09 ***	1.10 ***	1.10 ***	1.10 ***
Fungi						0	0.01	0.01 ***	0.01 ***
Uni. Euk.							0	0.00	0.00 ***
Bacteria								0	0.00 ***
Archaea									0

Differences in Median Normalized Alternative Splicing Ratio (ASR*)

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.25 ***	0.69 ***	0.42 ***	0.60 ***	1.74 ***	1.74 ***	1.73 ***	1.73 ***
Birds		0	0.44 ***	0.17	0.35 ***	1.48 ***	1.48 ***	1.48 ***	1.48 ***
Fish			0	0.26 ***	0.09	1.05 ***	1.05 ***	1.05 ***	1.05 ***
Arthropods				0	0.17 ***	1.31 ***	1.31 ***	1.31 ***	1.31 ***
Plants					0	1.14 ***	1.14 ***	1.14 ***	1.14 ***
Fungi						0	0.00	0.00 ***	0.00 ***
Uni. Euk.							0	0.00 ***	0.00 *
Bacteria								0	0.00 ***
Archaea									0

Table 2. Pairwise comparisons of genomic variables across taxonomic groups, showing differences in statistical measures (mean and median) for gene content relative to genome size, coding size relative to gene content, and coding relative to genome size. Statistical significance was assessed using Monte Carlo permutation tests, and adjusted p-values were calculated using the Bonferroni correction (see Methods). Colors represent significance levels: red for $p \leq 0.001$ (***), green for $0.001 < p \leq 0.05$ (**), and blue for $0.05 < p \leq 0.1$ (*). Values with $p > 0.1$ have no color.

Differences in Mean Gene Content / Genome Size Ratio

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.08 ***	0.17 ***	0.16 ***	0.20 ***	0.19 ***	0.16 ***	0.41 ***	0.41 ***
Birds		0	0.09 ***	0.08 ***	0.28 ***	0.11 ***	0.08 **	0.33 ***	0.34 ***
Fish			0	0.02	0.37 ***	0.01	0.01	0.24 ***	0.24 ***
Arthropods				0	0.36 ***	0.03	0.01	0.25 ***	0.26 ***
Plants					0	0.39 ***	0.36 ***	0.61 ***	0.61 ***
Fungi						0	0.02	0.23 ***	0.23 ***
Uni. Euk.							0	0.25 ***	0.25 ***
Bacteria								0	0.00
Archaea									0

Differences in Median Gene Content / Genome Size Ratio

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.08 ***	0.18 ***	0.17 ***	0.19 ***	0.19 ***	0.16 ***	0.42 ***	0.42 ***
Birds		0	0.10 ***	0.09 ***	0.27 ***	0.11 ***	0.08 **	0.34 ***	0.34 ***
Fish			0	0.01	0.36 ***	0.01	0.02	0.24 ***	0.25 ***
Arthropods				0	0.36 ***	0.02	0.01	0.25 ***	0.25 ***
Plants					0	0.38 ***	0.34 ***	0.61 ***	0.61 ***
Fungi						0	0.03	0.23 ***	0.23 ***
Uni. Euk.							0	0.26 ***	0.26 ***
Bacteria								0	0.00
Archaea									0

Differences in Mean Coding Size / Gene Content Ratio

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.02 ***	0.06 ***	0.08 ***	0.24 ***	0.84 ***	0.86 ***	0.96 ***	0.96 ***
Birds		0	0.04 ***	0.06 ***	0.23 ***	0.82 ***	0.84 ***	0.94 ***	0.94 ***
Fish			0	0.03 **	0.19 ***	0.78 ***	0.80 ***	0.90 ***	0.90 ***
Arthropods				0	0.16 ***	0.76 ***	0.77 ***	0.88 ***	0.88 ***
Plants					0	0.60 ***	0.61 ***	0.71 ***	0.72 ***
Fungi						0	0.02	0.12 ***	0.12 ***
Uni. Euk.							0	0.10 ***	0.10 ***
Bacteria								0	0.00 ***
Archaea									0

Differences in Median Coding Size / Gene Content Ratio

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.02 ***	0.05 ***	0.07 ***	0.25 ***	0.87 ***	0.87 ***	0.96 ***	0.96 ***
Birds		0	0.04 ***	0.05 ***	0.23 ***	0.85 ***	0.85 ***	0.94 ***	0.94 ***
Fish			0	0.02 **	0.20 ***	0.82 ***	0.81 ***	0.90 ***	0.91 ***
Arthropods				0	0.18 ***	0.80 ***	0.80 ***	0.89 ***	0.89 ***
Plants					0	0.62 ***	0.61 ***	0.71 ***	0.71 ***
Fungi						0	0.01	0.08 ***	0.09 ***
Uni. Euk.							0	0.09 ***	0.10 ***
Bacteria								0	0.00 ***
Archaea									0

Differences in Mean Coding Content / Genome Size Ratio

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.01 ***	0.04 ***	0.06 ***	0.05 ***	0.54 ***	0.52 ***	0.83 ***	0.84 ***
Birds		0	0.03 ***	0.04 ***	0.04 ***	0.53 ***	0.51 ***	0.82 ***	0.83 ***
Fish			0	0.02 ***	0.01 **	0.50 ***	0.48 ***	0.79 ***	0.80 ***
Arthropods				0	0.00	0.48 ***	0.47 ***	0.78 ***	0.78 ***
Plants					0	0.49 ***	0.47 ***	0.78 ***	0.78 ***
Fungi						0	0.02	0.29 ***	0.30 ***
Uni. Euk.							0	0.31 ***	0.32 ***
Bacteria								0	0.01
Archaea									0

Differences in Median Coding Content / Genome Size Ratio

	Mammals	Birds	Fish	Arthropods	Plants	Fungi	Uni. Euk.	Bacteria	Archaea
Mammals	0	0.01 ***	0.04 ***	0.05 ***	0.05 ***	0.54 ***	0.51 ***	0.84 ***	0.85 ***
Birds		0	0.03 ***	0.04 ***	0.04 ***	0.52 ***	0.50 ***	0.83 ***	0.83 ***
Fish			0	0.01 ***	0.01 ***	0.50 ***	0.47 ***	0.80 ***	0.81 ***
Arthropods				0	0.00	0.49 ***	0.46 ***	0.79 ***	0.80 ***
Plants					0	0.49 ***	0.46 ***	0.79 ***	0.80 ***
Fungi						0	0.03	0.30 ***	0.31 ***
Uni. Euk.							0	0.33 ***	0.34 ***
Bacteria								0	0.01
Archaea									0

Table 3. Summary statistics for the percentage of gene content relative to genome size (Gene Content / Genome Size (%)), the percentage of coding relative to gene size (Coding Size / Gene Content (%)), the percentage of coding relative to genome size (Coding Size / Genome Size (%)), the alternative splicing ratio (ASR), and the normalized alternative splicing ratio (ASR*) across different taxonomic groups. The table includes the mean ($\bar{x}$), the interpercentile range between the 5th and 95th percentiles ($[Q_{0.05}, Q_{0.95}]$), and standard deviation (σ) for each group.

	Gene Content / Genome Size (%)			Coding Size / Gene Content (%)			Coding Size / Genome Size (%)			ASR			ASR*		
Group	$\bar{x}$	$[Q_{0.05}, Q_{0.95}]$	σ	$\bar{x}$	$[Q_{0.05}, Q_{0.95}]$	σ	$\bar{x}$	$[Q_{0.05}, Q_{0.95}]$	σ	$\bar{x}$	$[Q_{0.05}, Q_{0.95}]$	σ	$\bar{x}$	$[Q_{0.05}, Q_{0.95}]$	σ
Mammals	44.7	[35.5,55.6]	6.27	3.05	[2.47,3.76]	0.42	1.35	[1.09,1.60]	0.15	3.09	[1.78,4.60]	0.92	2.81	[2.11,3.50]	0.59
Birds	52.6	[41.8,62.5]	6.05	4.79	[4.18,5.53]	0.45	2.50	[2.21,2.83]	0.20	2.71	[1.76,3.62]	0.64	2.48	[1.96,2.96]	0.31
Fish	61.9	[49.4,74.3]	7.50	8.65	[5.23,13.4]	2.52	5.39	[2.69,8.57]	1.77	2.26	[1.55,2.92]	0.41	2.02	[1.57,2.50]	0.29
Arthropods	60.3	[33.1,79.9]	14.0	11.3	[2.15,25.1]	7.15	6.92	[0.910,15.1]	4.36	2.16	[1.34,3.14]	0.54	2.30	[1.59,2.90]	0.39
Plants	24.7	[7.04,40.4]	11.0	27.3	[14.3,38.5]	7.50	6.80	[1.69,13.6]	3.68	1.56	[1.30,1.87]	0.20	2.10	[1.61,2.46]	0.24
Fungi	63.2	[41.2,85.9]	13.8	86.9	[67.6,99.2]	11.5	55.3	[33.2,82.7]	15.4	1.01	[1,1.09]	0.05	1.01	[1,1.09]	0.05
Unicellular Eukaryotes	60.9	[42.5,88.1]	16.2	88.6	[69.2,99.9]	12.0	53.6	[33.0,76.3]	14.9	1.00	[1,1.01]	2.56×10^{-3}	1.00	[1,1.01]	2.56×10^{-3}
Bacteria	85.7	[76.8,92.2]	5.06	98.8	[97.7,99.5]	0.57	84.6	[75.6,91.2]	5.07	1.00	[1.00,1.00]	8.21×10^{-4}	1.00	[1.00,1.00]	8.21×10^{-4}
Archaea	86.1	[74.5,92.4]	5.21	99.1	[98.7,99.4]	0.22	85.3	[73.9,91.5]	5.10	1.00	[1.00,1.01]	1.72×10^{-3}	1.00	[1.00,1.01]	1.72×10^{-3}

Table 4. Phylogenetic generalized least squares (PGLS) analysis was fitted to the data using a linear model $Y \sim \beta X$, with β the regression coefficient estimating the effect of X on Y. The Table is structured into two parts: the first part presents the results of the model fitted to ASR vs. genome size, ASR vs. gene content, ASR vs. coding content, ASR vs. the percentage of gene content within genome size, ASR vs. the percentage of coding within genes, and ASR vs. the percentage of coding within genome size. The second part reports the model fitted to the same relationships, but with ASR values normalized, ASR*. The model was fitted separately for each taxonomic group. The table presents the regression coefficients (β), the adjusted coefficient of determination (R^2_{adj}), and the importance of the phylogenetic structure (λ). Statistical significance is indicated with asterisks and colors: red for $p \leq 0.001$ (***), green for $0.001 < p \leq 0.05$ (**), and blue for $0.05 < p \leq 0.1$ (*). Values with $p > 0.1$ have no color.

	ASR vs. Genome Size			ASR vs. Gene Content			ASR vs. Coding			ASR vs. Gene / Genome			ASR vs. Coding / Gene			ASR vs. Coding / Genome		
	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ
Mammals	5.0×10^{-4}	0.01	0.35	4.0×10^{-3}***	0.55	0.41	0.28***	0.16	0.41	9.95***	0.43	0.49	-184.05***	0.52	0.52	-8.25	-0.01	0.39
Birds	5.0×10^{-4}	-0.01	0.75	7.0×10^{-3}***	0.70	0.97	0.23***	0.27	0.25	8.11***	0.58	0.62	-152.85***	1.00	1.00	87.09**	0.05	0.36
Fish	-2.0×10^{-4}*	0.02	0.59	7.6×10^{-5}	-4.6×10^{-3}	0.50	5.0×10^{-3}	8.0×10^{-4}	0.47	4.51***	0.50	0.75	-1.39	-8.4×10^{-4}	0.6	6.34**	0.04	0.61
Arthropods	-1.0×10^{-4}	0.01	0.59	-5.3×10^{-5}	-0.01	0.59	-0.01	2.5×10^{3}	0.59	2.24***	0.27	0.71	-1.24	0.01	0.66	1.29	3.2×10^{-4}	0.58
Plants	-9.2×10^{-6}	-0.01	0.90	4.8×10^{-4}**	0.04	0.93	-4.5×10^{-4}	-0.01	0.90	0.33**	0.03	0.89	-1.55***	0.26	0.93	-0.36	-4.4×10^{-3}	0.91
Fungi	2.0×10^{-4}	-0.01	0.50	3.8×10^{-3}**	0.11	0.50	1.7×10^{-3}	-8.5×10^{-4}	0.50	0.14**	0.06	0.20	-0.35***	0.26	0.50	-0.1	-0.01	0.5
Unicellular Eukaryotes	-9.2×10^{-6}	-0.02	0.25	-2.0×10^{-5}	-0.2	0.22	-1.2×10^{-5}	-0.03	0.25	-1.4×10^{-4}	-0.03	0.25	0.01	-4.6×10^{-3}	0.25	3.3×10^{-3}	-0.02	0.50
Bacteria	-7.2×10^{-5}**	0.03	0.91	-7.2×10^{-5}**	0.02	0.91	-7.3×10^{-5}	0.02	0.91	1.6×10^{-3}**	0.02	0.91	-0.02**	0.02	0.92	1.5×10^{-3}**	0.01	0.91
Archaea	-8.3×10^{-5}	-3.0×10^{-4}	0.97	-6.4×10^{-5}	-4.0×10^{-3}	0.97	-6.7×10^{-5}	-3.8×10^{-3}	0.97	0.01***	0.07	0.97	-0.13	0.13	0.96	4.9×10^{-3}***	0.06	0.97

	ASR* vs. Genome Size			ASR* vs. Gene Content			ASR* vs. Coding			ASR* vs. Gene / Genome			ASR* vs. Coding / Gene			ASR* vs. Coding / Genome		
	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ	β	R^2_{adj}	λ
Mammals	3.7×10^{-4}**	0.03	0.25	1.9×10^{-3}***	0.32	0.25	0.12**	0.06	0.25	4.44***	0.18	0.25	-82.74***	0.27	0.18	-37.77	-3.0×10^{-3}	0.50
Birds	6.68×10^{-4}**	0.06	0.07	2.2×10^{-3}***	0.29	0.25	0.08**	0.13	0.10	1.79**	0.11	0.08	-35.67***	0.25	0.25	-14.96	-0.01	0.11
Fish	1.42×10^{-4}**	0.03	0.70	1.48×10^{-4}	0.01	0.69	4.2×10^{-4}	-0.01	0.69	-1.37***	0.12	0.66	-2.23**	0.03	0.68	-5.56***	0.09	0.67
Arthropods	7.29×10^{-5}**	0.02	0.74	3.23×10^{-4}**	0.07	0.72	0.02**	0.05	0.75	-0.36	4.4×10^{-3}	0.76	-3.19***	0.23	0.59	-5.29***	0.26	0.56
Plants	8.23×10^{-5}***	0.15	0.88	1.15×10^{-3}***	0.18	0.84	3.8×10^{-3}**	0.07	0.87	-0.85***	0.15	0.91	-1.16***	0.10	0.80	-3.24***	0.23	0.87
Fungi	2.4×10^{-4}	-0.01	0.25	3.8×10^{-3}**	0.11	0.50	1.7×10^{-3}	-8.5×10^{-4}	0.50	0.14**	0.06	0.20	-0.35***	0.26	0.50	-0.03	-0.01	0.25
Unicellular Eukaryotes	-9.2×10^{-6}	-0.02	0.25	-2.40×10^{-5}	-0.02	0.25	-1.2×10^{-5}	-0.03	0.25	-1.4×10^{-4}	-0.03	0.25	0.01	-4.6×10^{-3}	0.25	1.6×10^{-3}	-0.03	0.25
Bacteria	-7.30×10^{-5}**	0.03	0.91	-7.16×10^{-5}**	0.02	0.91	-7.28×10^{-5}**	0.02	0.91	1.6×10^{-3}**	0.02	0.91	-0.02**	0.02	0.92	1.5×10^{-3}**	0.02	0.91
Archaea	-8.31×10^{-5}	-3.0×10^{-4}	0.97	-6.38×10^{-5}	-4.0×10^{-3}	0.97	-6.67×10^{-5}	-3.8×10^{-3}	0.97	0.01***	0.07	0.97	-0.13***	0.13	0.96	4.8×10^{-3}***	0.07	0.97

Table 5. Coefficient of variation ($\hat{CV}$) for each genomic variable and taxonomic group. The coefficient of variation is calculated as $\hat{CV} = s/\bar{x}$, where s is the standard deviation and $\bar{x}$ is the mean of each variable within a taxonomic group. Each column represents a different variable, with the coefficient of variation computed for genome size, gene content, coding content, the percentage of gene content within genomes, the percentage of coding within genes, the percentage of coding within genomes, the alternative splicing ratio (ASR), and its normalized ratio, ASR*.

	$\hat{CV}_{Genome}$	$\hat{CV}_{Gene}$	$\hat{CV}_{Coding}$	$\hat{CV}_{Gene/Genome}$	$\hat{CV}_{Coding/Gene}$	$\hat{CV}_{Coding/Genome}$	$\hat{CV}_{ASR}$	$\hat{CV}_{ASR*}$
Mammals	0.13	0.16	0.04	0.14	0.14	0.11	0.30	0.21
Birds	0.10	0.12	0.05	0.12	0.09	0.08	0.24	0.12
Fish	0.57	0.47	0.23	0.12	0.29	0.33	0.18	0.14
Arthropods	1.87	1.28	0.19	0.23	0.63	0.63	0.25	0.17
Plants	1.11	0.58	0.38	0.45	0.27	0.54	0.13	0.11
Fungi	0.75	0.51	0.43	0.22	0.13	0.28	0.05	0.05
Uni. Euk.	1.02	0.79	0.72	0.27	0.14	0.28	0.00	0.00
Bacteria	0.53	0.53	0.54	0.06	0.01	0.06	0.00	0.00
Archaea	0.38	0.34	0.34	0.06	0.00	0.06	0.00	0.00

Table 6. Relative variability among genomic features, computed as the ratio of the coefficient of variation $\hat{CV}$ between different genomic variables, $\hat{CV}_x/\hat{CV}_y$, where $\hat{CV}_x = s/\bar{x}$ (standard deviation divided by the sample mean). It is computed separately for each taxonomic group. The first table presents variability relationships among variables related to genome composition, whereas the second and third tables show the relationships of these variables with the alternative splicing ratio and its normalized value, respectively.

	$\frac{\hat{CV}_{Gene}}{\hat{CV}_{Genome}}$	$\frac{\hat{CV}_{Coding}}{\hat{CV}_{Gene}}$	$\frac{\hat{CV}_{Coding}}{\hat{CV}_{Genome}}$	$\frac{\hat{CV}_{Gene/Genome}}{\hat{CV}_{Genome}}$	$\frac{\hat{CV}_{Coding/Gene}}{\hat{CV}_{Gene}}$	$\frac{\hat{CV}_{Coding/Genome}}{\hat{CV}_{Genome}}$
Mammals	1.22	0.25	0.30	1.04	0.85	0.83
Birds	1.19	0.40	0.48	1.11	0.77	0.80
Fish	0.83	0.48	0.40	0.21	0.61	0.58
Arthropods	0.69	0.15	0.10	0.12	0.49	0.34
Plants	0.52	0.65	0.34	0.40	0.47	0.49
Fungi	0.68	0.84	0.57	0.29	0.26	0.37
Uni. Euk.	0.77	0.91	0.70	0.26	0.17	0.27
Bacteria	1.00	1.01	1.01	0.11	0.01	0.11
Archaea	0.89	1.00	0.90	0.16	0.01	0.16

	$\frac{\hat{CV}_{ASR}}{\hat{CV}_{Genome}}$	$\frac{\hat{CV}_{ASR}}{\hat{CV}_{Gene}}$	$\frac{\hat{CV}_{ASR}}{\hat{CV}_{Coding}}$	$\frac{\hat{CV}_{ASR}}{\hat{CV}_{Gene/Genome}}$	$\frac{\hat{CV}_{ASR}}{\hat{CV}_{Coding/Gene}}$	$\frac{\hat{CV}_{ASR}}{\hat{CV}_{Coding/Genome}}$
Mammals	2.22	1.82	7.34	2.13	2.15	2.67
Birds	2.30	1.94	4.79	2.06	2.50	2.89
Fish	0.32	0.38	0.79	1.50	0.62	0.55
Arthropods	0.13	0.20	1.32	1.08	0.40	0.40
Plants	0.12	0.22	0.34	0.29	0.47	0.24
Fungi	0.07	0.10	0.11	0.23	0.37	0.18
Uni. Euk.	0.00	0.00	0.00	0.01	0.02	0.01
Bacteria	0.00	0.00	0.00	0.01	0.14	0.01
Archaea	0.00	0.01	0.01	0.03	0.77	0.03

	$\frac{\hat{CV}_{ASR^*}}{\hat{CV}_{Genome}}$	$\frac{\hat{CV}_{ASR^*}}{\hat{CV}_{Gene}}$	$\frac{\hat{CV}_{ASR^*}}{\hat{CV}_{Coding}}$	$\frac{\hat{CV}_{ASR^*}}{\hat{CV}_{Gene/Genome}}$	$\frac{\hat{CV}_{ASR^*}}{\hat{CV}_{Coding/Gene}}$	$\frac{\hat{CV}_{ASR^*}}{\hat{CV}_{Coding/Genome}}$
Mammals	1.57	1.28	5.19	1.51	1.52	1.89
Birds	1.21	1.02	2.51	1.08	1.31	1.51
Fish	0.25	0.30	0.63	1.19	0.50	0.44
Arthropods	0.09	0.13	0.89	0.73	0.27	0.27
Plants	0.10	0.19	0.30	0.25	0.41	0.21
Fungi	0.07	0.10	0.11	0.23	0.37	0.18
Uni. Euk.	0.00	0.00	0.00	0.01	0.02	0.01
Bacteria	0.00	0.00	0.00	0.01	0.14	0.01
Archaea	0.00	0.01	0.01	0.03	0.77	0.03

Supplementary File 2

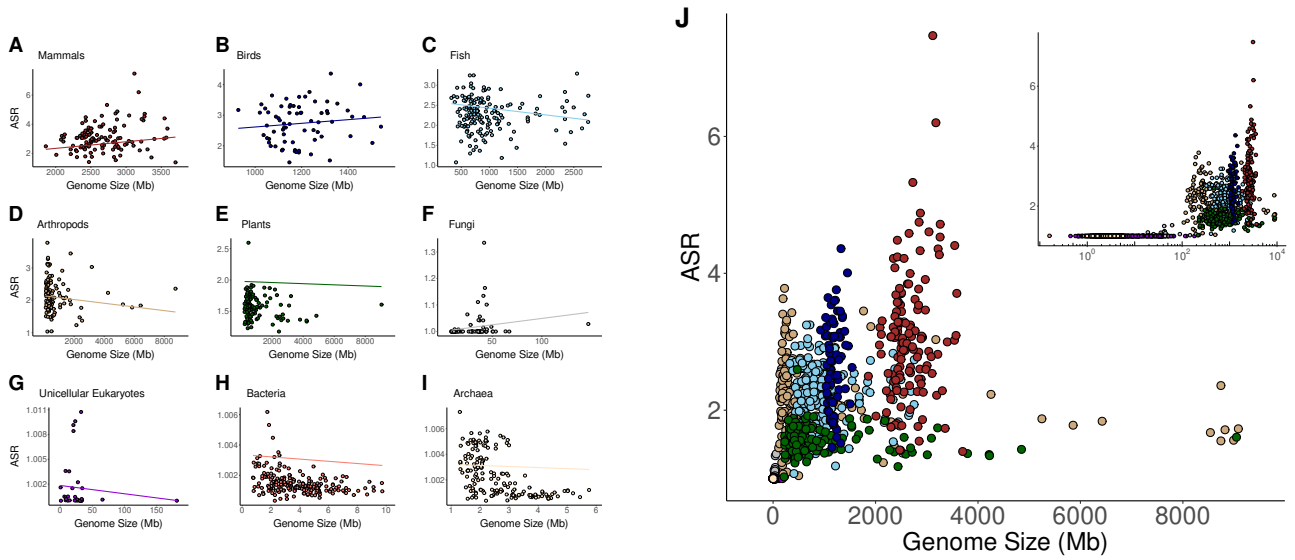

Figure 1. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between genome size and ASR across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups. The inset provides a logarithmic representation of the x-axis.

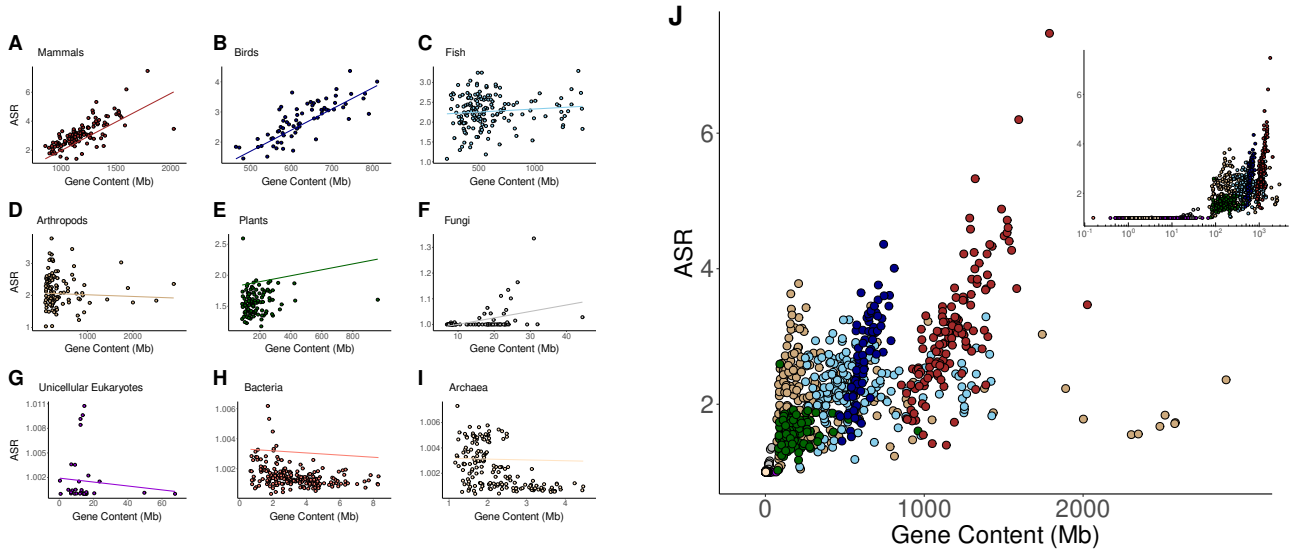

Figure 2. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between gene content and ASR across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups. The inset provides a logarithmic representation of the x-axis.

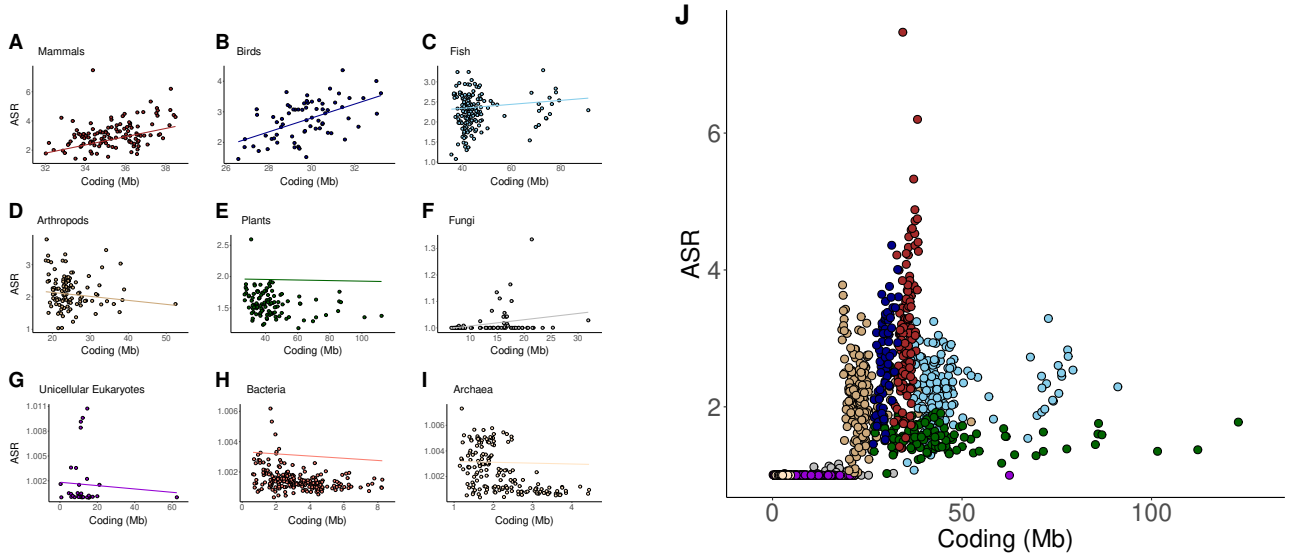

Figure 3. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the amount of coding and ASR across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

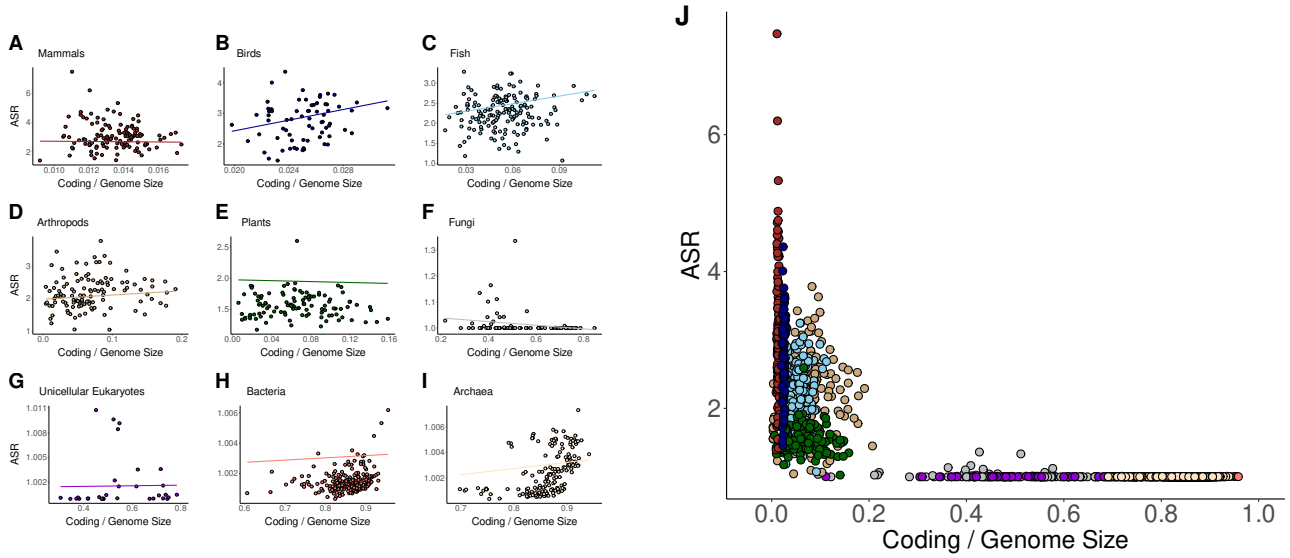

Figure 4. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the proportion of coding relative to genome size and ASR across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

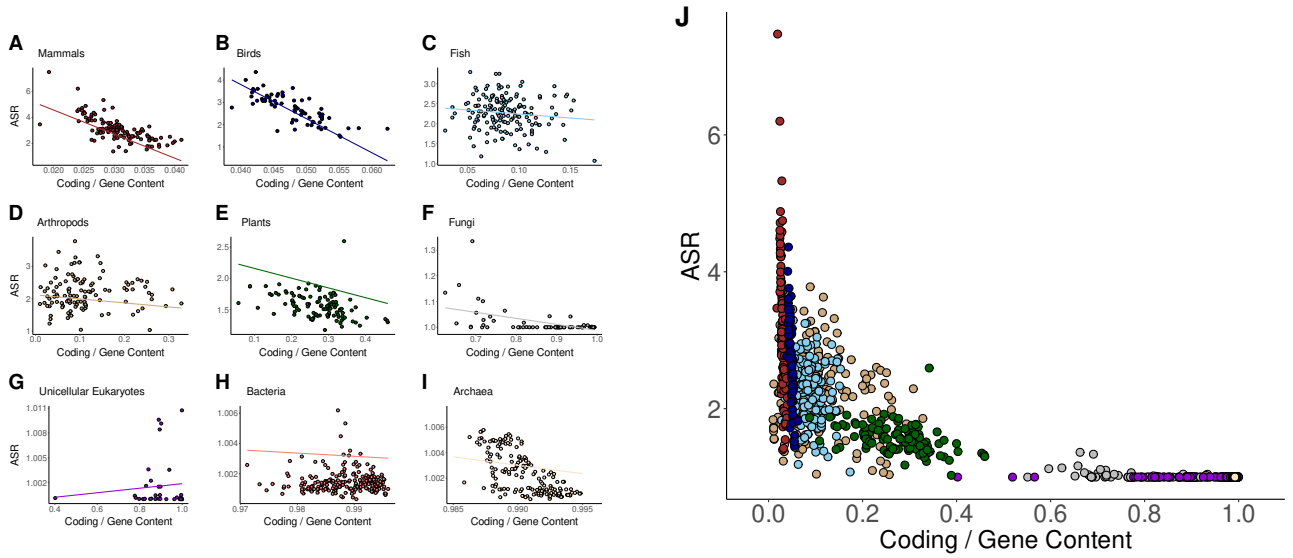

Figure 5. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the proportion of coding relative to gene content and ASR across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

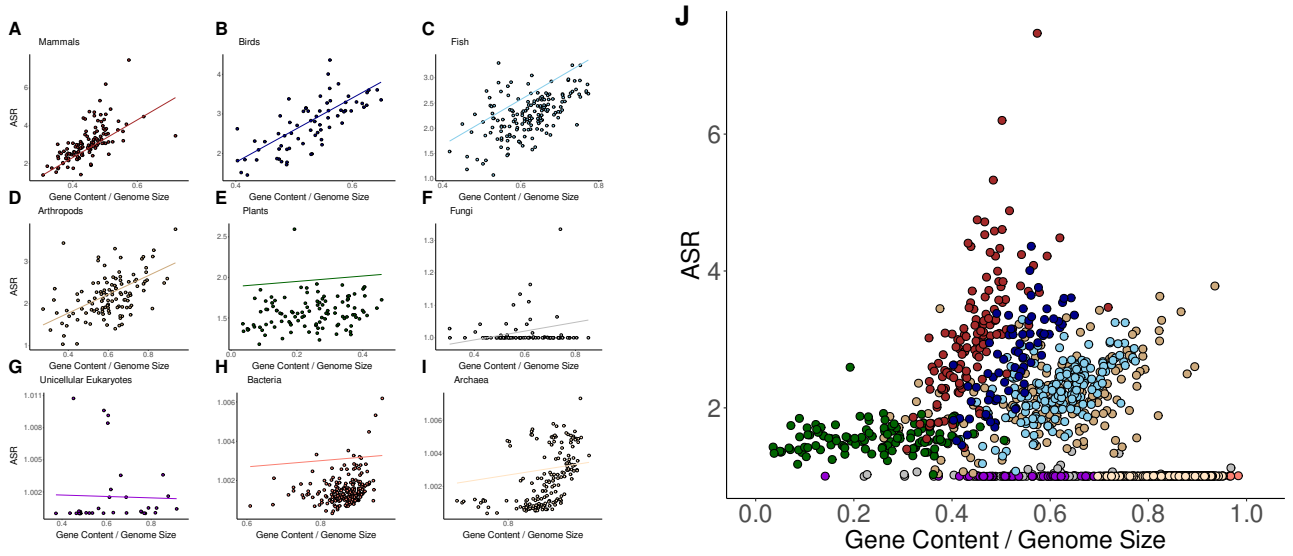

Figure 6. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the proportion of gene content relative to genome size and ASR across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

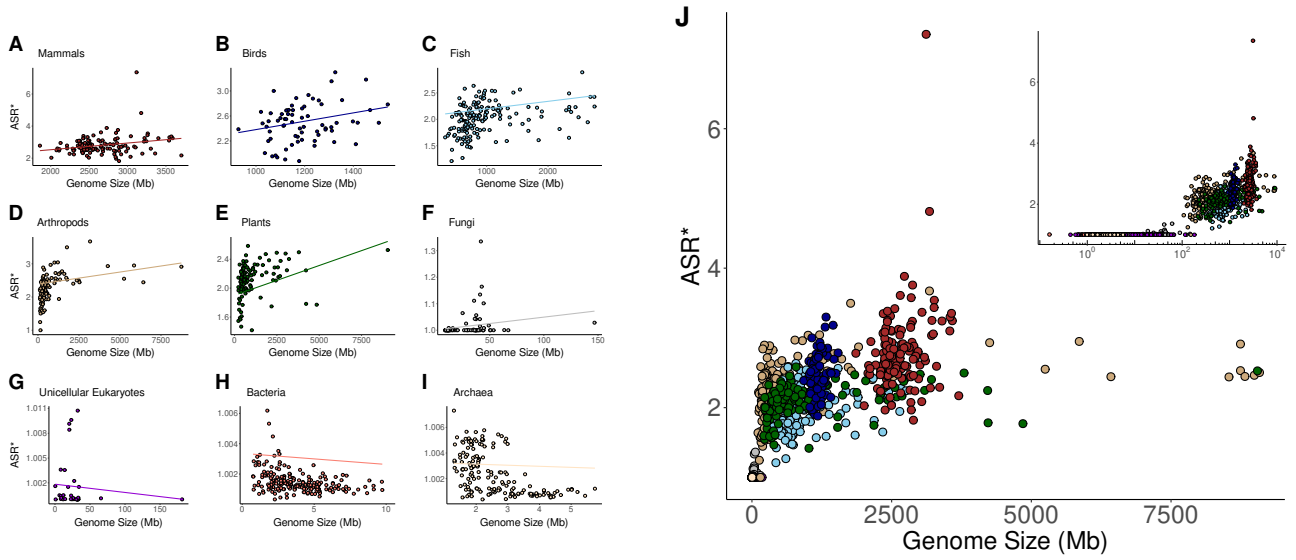

Figure 7. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between genome size and the normalized ASR* across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups. The inset provides a logarithmic representation of the x-axis.

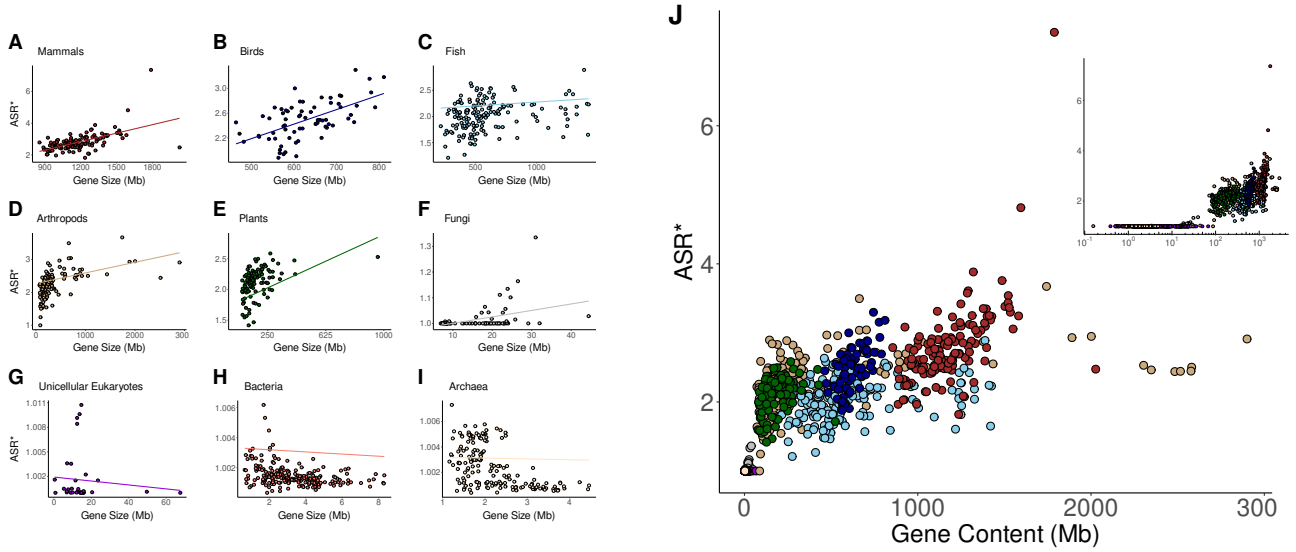

Figure 8. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between gene content and the normalized ASR* across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups. The inset provides a logarithmic representation of the x-axis.

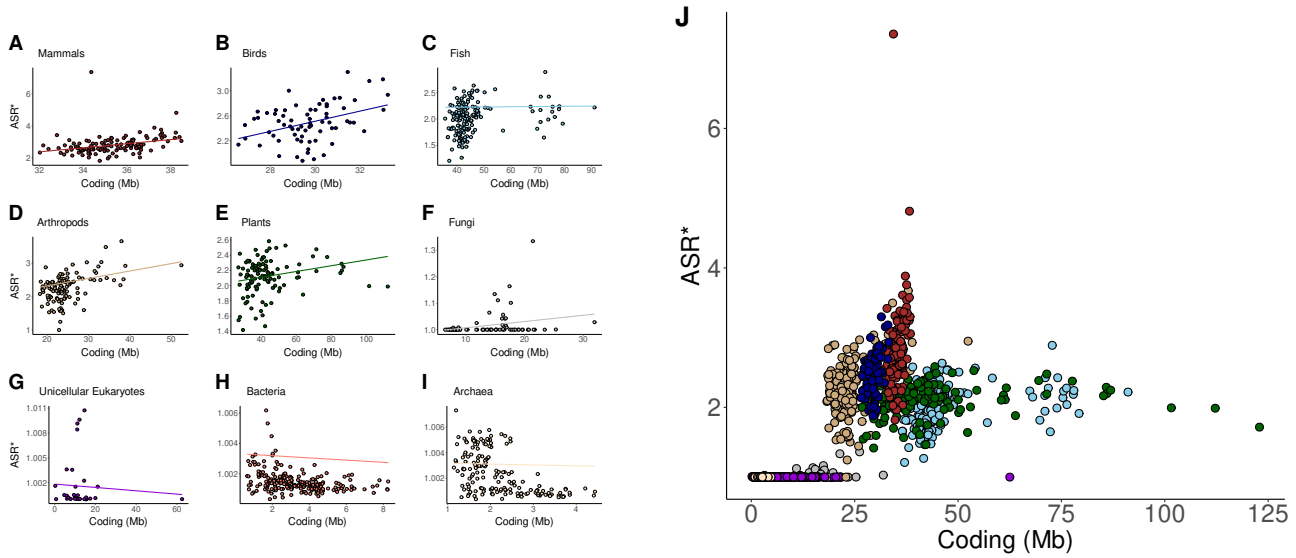

Figure 9. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the amount of coding and the normalized ASR* across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.

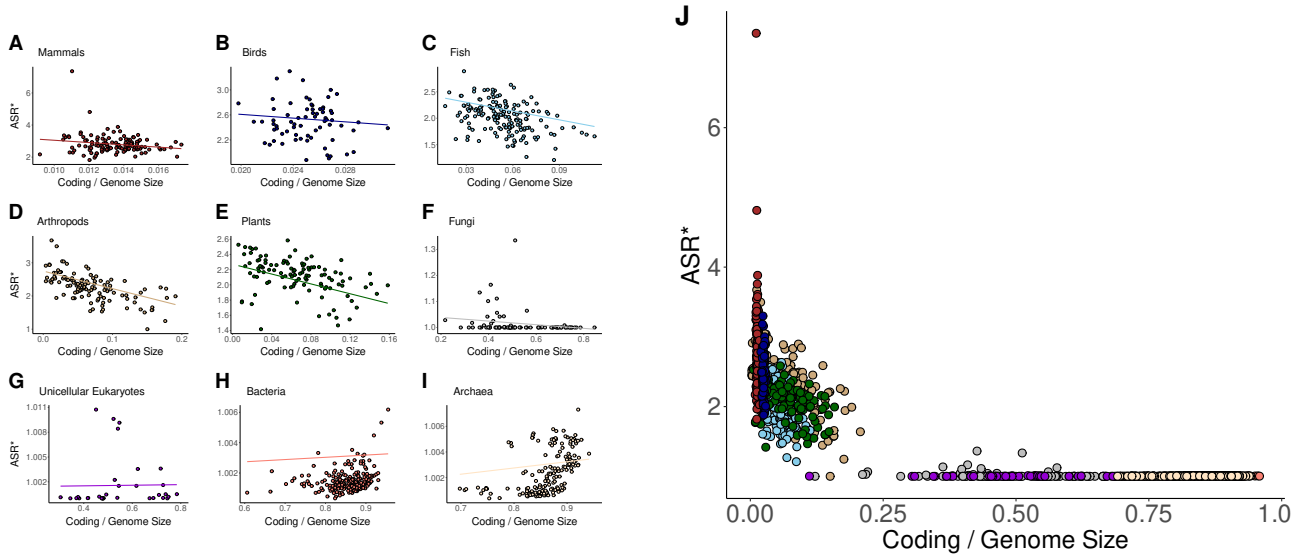

Figure 10. (A-I) Phylogenetic Generalized Least Squares (PGLS) regression between the proportion of coding relative to genome size and the normalized ASR* across different taxonomic groups. Each panel represents a distinct taxonomic group. The regression lines represent the estimated evolutionary relationship between the two variables while accounting for phylogenetic dependence. (J) Global relationship across all taxonomic groups.