Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans

Florian Bénitière, Anamaria Necsulea, Laurent Duret

Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, UMR CNRS 5558, Villeurbanne, France.

Correspondence: Laurent.Duret@univ-lyon1.fr

June 6, 2023

Abstract

Most eukaryotic genes undergo alternative splicing (AS), but the overall functional significance of this process remains a controversial issue. It has been noticed that the complexity of organisms (assayed by the number of distinct cell types) correlates positively with their genome-wide AS rate. This has been interpreted as evidence that AS plays an important role in adaptive evolution by increasing the functional repertoires of genomes. However, this observation also fits with a totally opposite interpretation: given that 'complex' organisms tend to have small effective population sizes (N_e), they are expected to be more affected by genetic drift, and hence more prone to accumulate deleterious mutations that decrease splicing accuracy. Thus, according to this "drift barrier" theory, the elevated AS rate in complex organisms might simply result from a higher splicing error rate. To test this hypothesis, we analyzed 3,496 transcriptome sequencing samples to quantify AS in 53 metazoan species spanning a wide range of N_e values. Our results show a negative correlation between N_e proxies and the genome-wide AS rates among species, consistent with the drift barrier hypothesis. This pattern is dominated by low abundance isoforms, which represent the vast majority of the splice variant repertoire. We show that these low abundance isoforms are depleted in functional AS events, and most likely correspond to errors. Conversely, the AS rate of abundant isoforms, which are relatively enriched in functional AS events, tends to be lower in more complex species. All these observations are consistent with the hypothesis that variation in AS rates across metazoans reflects the limits set by drift on the capacity of selection to prevent gene expression errors.

Keywords Alternative splicing · Random genetic drift · Life history traits · Effective population size · dN/dS · Splice variants · Non-adaptive models · N_e

Introduction

Eukaryotic protein-coding genes are interrupted by introns, which have to be excised from the primary transcript to produce functional mRNAs that can be translated into proteins. The removal of introns from primary transcripts can lead to the production of diverse mRNAs, *via* the differential use of splice sites. This process of alternative splicing (AS) is widespread in eukaryotes (Chen *et al.*, 2014), but its 'raison d'être' (adaptive or not) remains elusive. Numerous studies have shown that some AS events are functional, *i.e.* that they play a beneficial role for the fitness of organisms, either by allowing the production of distinct protein isoforms (Graveley, 2001) or by regulating gene expression post-transcriptionally (McGlincy and Smith, 2008; Hamid and Makeyev, 2014). However, other AS events are undoubtedly not functional. Like any biological machinery, the spliceosome occasionally makes errors, leading to the production of aberrant mRNAs, which represent a waste of resources and are therefore deleterious for the fitness of the organisms (Hsu and Hertel, 2009; Gout *et al.*, 2013). The splicing error rate at a given intron is expected to depend both on the efficiency of the spliceosome and on the intrinsic quality of its splice signals. The information required in cis for the removal of each intron resides in 20 to 40 nucleotide sites, located within the intron or its flanking exons (Lynch, 2006). Besides the two splice sites that are essential for the splicing reaction (almost always GT for the donor and AG for the acceptor), all other signals tolerate some sequence flexibility. Population genetics principles state that the ability of selection to promote beneficial mutations or eliminate deleterious mutations depends on the intensity of selection (s) relative to the power of random genetic drift (defined by the effective population size, N_e): if the selection coefficient is sufficiently weak relative to drift ($|N_\mathrm{e}s| < 1$), alleles behave as if they are effectively neutral. Thus, random drift sets an upper limit on the capacity of selection to prevent the fixation of alleles that are sub-optimal (Kimura *et al.*, 1963; Ohta, 1973). This so-called "drift barrier" (Lynch, 2007) is expected to affect the efficiency of all cellular processes, including splicing. Hence, species with low N_e should be more prone to make splicing errors than species with high N_e.

The extent to which AS events correspond to functional isoforms or to errors is a contentious issue (Bhuiyan *et al.*, 2018; Tress *et al.*, 2017b; Blencowe, 2017; Tress *et al.*, 2017a). In humans, the set of transcripts produced by a given gene generally consists of one major transcript (the 'major isoform'), which encodes a functional protein, and of multiple minor isoforms (splice variants), present in relatively low abundance, and whose coding sequence is frequently interrupted by premature termination codons (PTCs) (Tress *et al.*, 2017a; Gonzàlez-Porta *et al.*, 2013). Ultimately, less than 1% of human splice variants lead to the production of a detectable amount of protein (Abascal *et al.*, 2015). Furthermore, comparison with closely related species showed that AS patterns evolve very rapidly (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012) and that alternative splice sites present little evidence of selective constraints (Pickrell *et al.*, 2010). All these observations are consistent with the hypothesis that a vast majority of splice variants observed in human transcriptomes simply correspond to erroneous transcripts (Pickrell *et al.*, 2010). However, some authors argue that a large fraction of AS events might in fact contribute to regulating gene expression. Indeed, PTC-containing splice variants are recognized and degraded by the non-sense mediated decay (NMD) machinery. Thus, AS can be coupled with NMD to modulate gene expression at the post-transcriptional level (McGlincy and Smith, 2008; Hamid and Makeyev, 2014). This

AN-NMD regulatory process does not involve the production of proteins and does not necessarily imply strong evolutionary constraints on splice sites. Thus, based on these observations, it is difficult to firmly refute selectionist or non-adaptive models.

The analysis of transcriptomes from various eukaryotic species showed substantial variation in AS rates across lineages, with the highest rate in primates (Barbosa-Morais *et al.*, 2012; Chen *et al.*, 2014; Mazin *et al.*, 2021). Interestingly, the genome-wide average AS level was found to correlate positively with the complexity of organisms (approximated by the number of cell types) (Chen *et al.*, 2014). This correlation was considered as evidence that AS contributed to the evolution of complex organisms by increasing the functional repertoire of their genomes (Chen *et al.*, 2014). This pattern is often presented as an argument supporting the importance of AS in adaptation (Verta and Jacobs, 2022; Singh and Ahi, 2022; Wright *et al.*, 2022). However, this correlation is also compatible with a totally opposite hypothesis. Indeed, eukaryotic species with the highest level of complexity correspond to multi-cellular organisms with relatively large body size, which tend to have small effective population sizes (N_e) (Lynch and Conery, 2003; Figuet *et al.*, 2016). Thus, the higher AS rate observed in 'complex' organisms might simply reflect an increased rate of splicing errors, resulting from the effect of the drift barrier on the quality of splice signals (Bush *et al.*, 2017).

To assess this hypothesis and evaluate the impact of genetic drift on alternative splicing patterns, we quantified AS rates in 53 metazoan species, covering a wide range of N_e values, and for which high-depth transcriptome sequencing data were available. We show that the genome-wide average AS rate correlates negatively with N_e, in agreement with the drift barrier hypothesis. This pattern is mainly driven by low abundance isoforms, which represent the vast majority of splice variants and most likely correspond to errors. Conversely, the AS rate of abundant splice variants, which are enriched in functional AS events, show the opposite trend. These results support the hypothesis that the drift barrier sets an upper limit on the capacity of selection to minimize splicing errors.

Results

Genomic and transcriptomic data collection

To analyze variation in AS rates across metazoans, we examined a collection of 69 species for which transcriptome sequencing (RNA-seq) data, genome assemblies, and gene annotations were available in public databases. We focused on vertebrates and insects, the two metazoan clades that were the best represented in public databases when we initiated this project. To be able to compare average AS rates across species, we needed to control for several possible sources of biases. First, given that AS rates vary across genes (Saudemont *et al.*, 2017), we had to analyze a common set of orthologous genes. For this purpose, we extracted from the BUSCO database (Seppey *et al.*, 2019) a reference set of single-copy orthologous genes shared across metazoans (N=978 genes), and searched for their homologues in each species in our dataset. We retained for further analyses those species for which at least 80% of the BUSCO metazoan gene set could be identified (N=67 species; see Materials & Methods). Second, we had to ensure that RNA-seq read coverage was sufficiently high in each species to detect splicing variants. Indeed, to be able to detect AS at a given intron, it is necessary to analyze a minimal number of sequencing reads encompassing this intron (we used a threshold of N=10 reads). To assess the impact of sequencing depth on AS detection, we conducted a pilot analysis with two species (*Homo sapiens* and *Drosophila melanogaster*) for which hundreds of RNA-seq samples are available. This analysis (detailed in Supplementary Fig. 1) revealed that AS rate estimates are very noisy when sequencing depth is limited, but that they converge when sequencing is high enough. We therefore kept for further analysis those species for which the median read coverage across exonic regions of BUSCO genes was above 200 (Supplementary Fig. 1). Our final dataset thus consisted of 53 species (15 vertebrates and 38 insects; Fig. 1A), and of 3,496 RNA-seq samples (66 *per* species on average). In these species, the number of analyzable annotated introns (*i.e.* encompassed by at least 10 reads) among BUSCO genes ranges from 2,032 to 10,981 (which represents 88.6% to 99.6% of their annotated introns; Supplementary Tab. 1). It should be noted that analyzed samples originate from diverse sources; however, they are very homogenous in terms of sequencing technology (99% of RNA-seq samples sequenced with Illumina platforms; refer to `Data10-supp.tab` in the Zenodo data repository).

Proxies for the effective population size (N_e)

Effective population sizes (N_e) can in principle be inferred from levels of genetic polymorphism. However, population genetics data are lacking for most of the species in our dataset. We therefore used two life history traits that were previously proposed as proxies of N_e in metazoans (Waples, 2016; Weyna and Romiguier, 2020; Figuet *et al.*, 2016): body length and longevity (Materials & Methods; Supplementary Tab. 2). An additional proxy for N_e can be obtained by studying the intensity of purifying selection acting on protein sequences, through the dN/dS ratio (Kryazhimskiy and Plotkin, 2008). To evaluate this ratio, we aligned 922 BUSCO genes, reconstructed the phylogenetic tree of the 53 species (Fig. 1A) and computed the dN/dS ratio along each terminal branch (Materials & Methods).

We note that these three proxies provide "inverse" estimates of N_e, meaning that species with high longevity, large body length and/or elevated dN/dS values tend to have low N_e values. As expected, these different proxies of N_e are positively correlated with each other ($p < 1x10^{-3}$, Fig. 1B,C). We note however that these correlations are not very strong. It thus seems likely that none of these proxies provides a perfect estimate of N_e. To take phylogenetic inertia into account, all cross-species correlations presented here were computed using Phylogenetic Generalized Least Squared (PGLS) regression (Freckleton *et al.*, 2002).

Alternative splicing rates are negatively correlated with N_e proxies

To quantify AS rates, we mapped RNA-seq data of each species on the corresponding reference genome assembly. We detected sequencing reads indicative of a splicing event (hereafter termed ‘spliced reads’), and inferred the corresponding intron boundaries. We were thus able to validate the coordinates of annotated introns and to detect new introns, not present in the annotations. For each intron detected in RNA-seq data, we counted the number of spliced reads matching with its two boundaries (N_s) or sharing only one of its boundaries (N_a), as well as the number of unspliced reads covering its boundaries (N_u) (Fig. 2A). We then computed the relative abundance of this spliced isoform compared to other transcripts with alternative splice boundaries ($RAS = \frac{N_s}{N_s + N_a}$) or compared to unspliced transcripts ($RANS = \frac{N_s}{N_s + \frac{N_u}{2}}$).

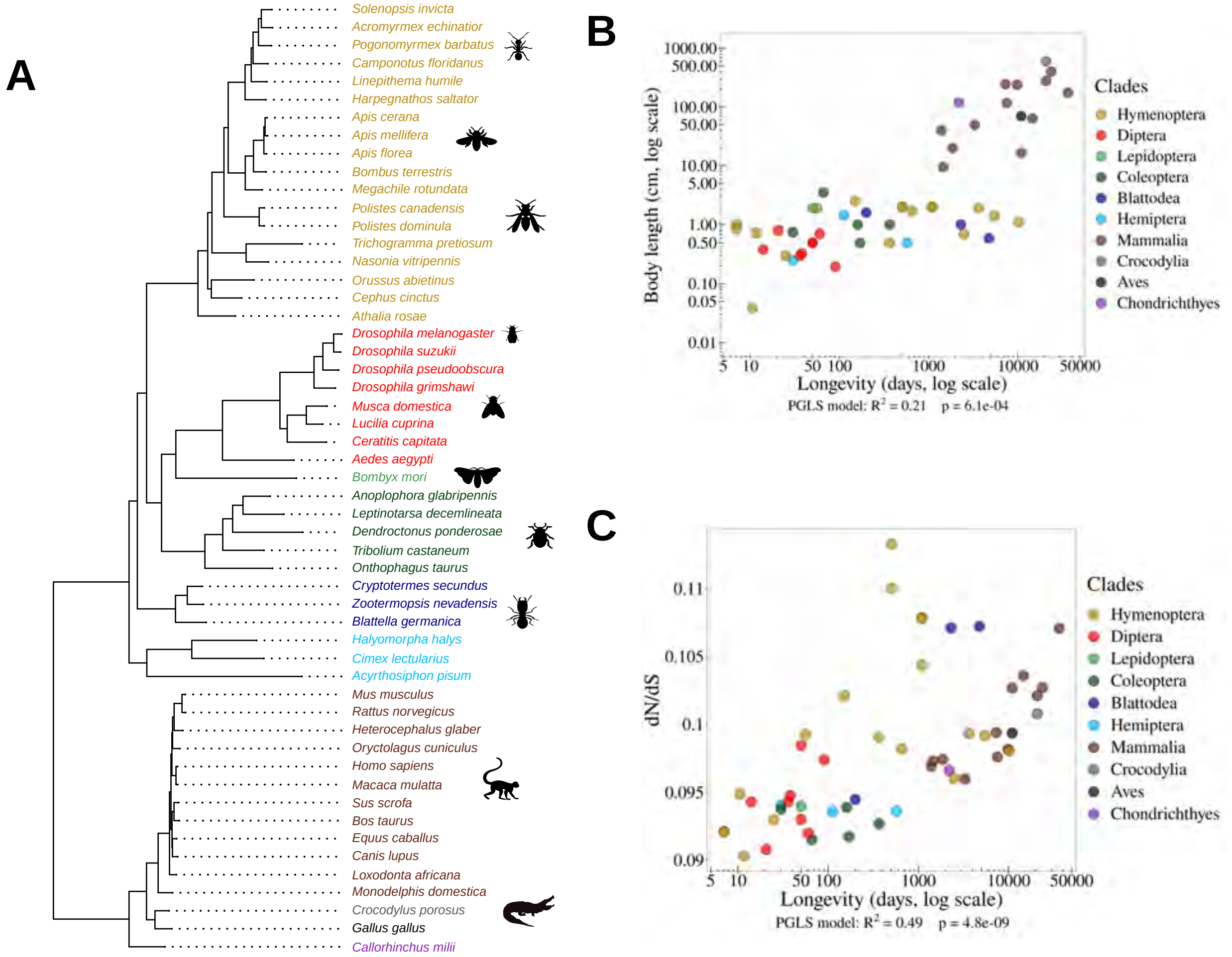

Figure 1: **Species phylogeny and N_e proxies. A**: Phylogenetic tree of the 53 studied species (15 vertebrates and 38 insects). **B**: Relationship between body length (cm, log scale) and longevity (days, log scale) of the organism. Each dot represents one species (colored by clade, as in the species tree in panel A). **C**: Relationship between longevity (days, log scale) and the dN/dS ratio on terminal branches of the phylogenetic tree (Materials & Methods). **B,C**: PGLS stands for Phylogenetic Generalized Least Squared regression, which takes into account phylogenetic inertia (Materials & Methods).

To limit measurement noise, we only considered introns for which both RAS and RANS could be computed based on at least 10 reads (Materials & Methods). In all species, both RAS and RANS metrics show clearly bimodal distributions (Fig. 2B,C): the first peak (mode $< 5\%$) corresponds to 'minor introns', whose splicing occurs only in a minority of transcripts of a given gene, whereas the second one (mode $> 95\%$) corresponds to the introns of major isoforms. It has been previously shown that in humans, for most genes, one single transcript largely dominates over other isoforms (Tress *et al.*, 2017a; Gonzàlez-Porta *et al.*, 2013). Our observations indicate that this pattern is generalized across metazoans. For the rest of our analyses, we

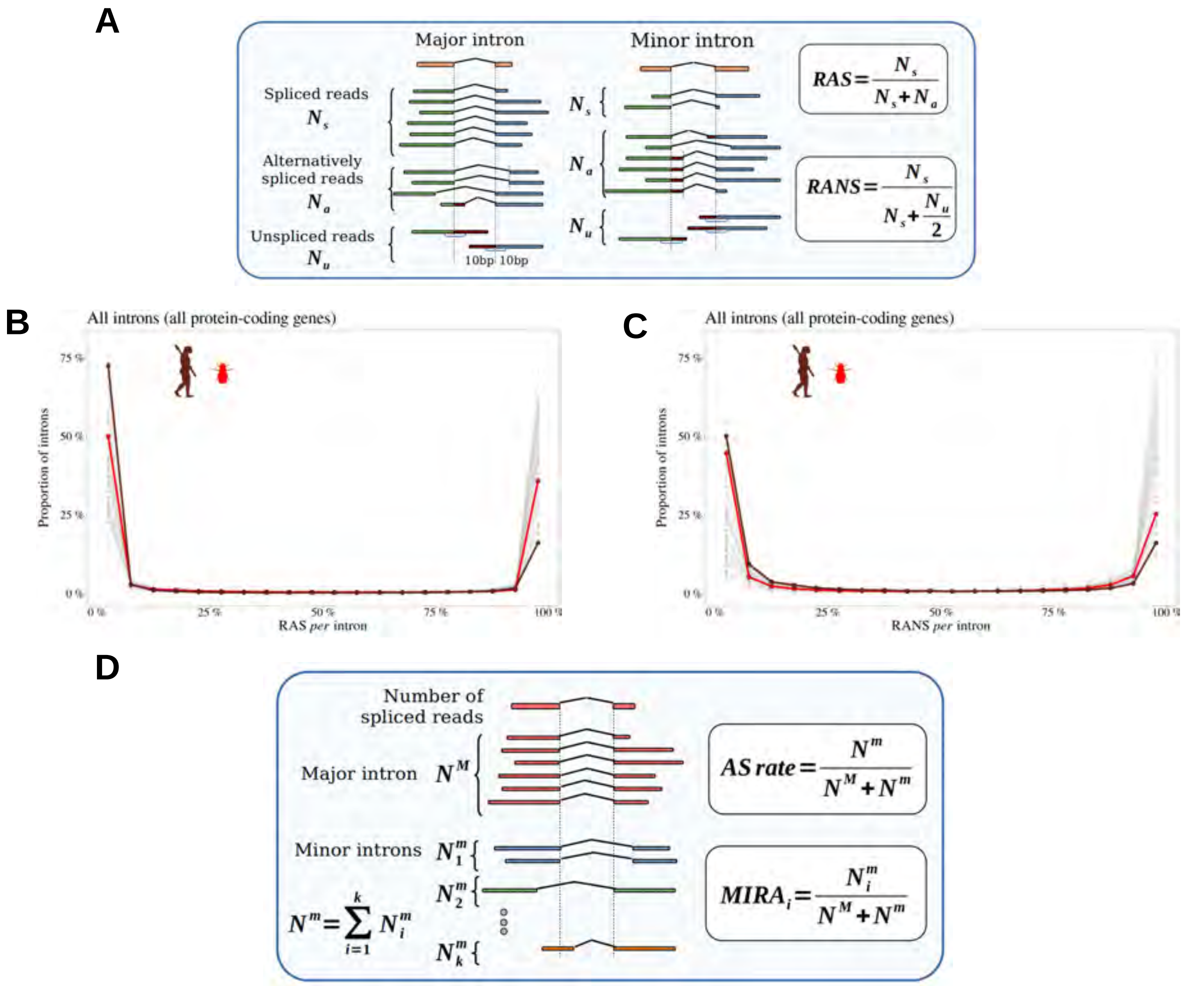

Figure 2: **Distinguishing major and minor introns. A**: Definition of the variables used to compute the relative abundance of the spliced isoform compared to other transcripts with alternative splice boundaries (RAS) or compared to unspliced transcripts (RANS): N_s: number of spliced reads corresponding to the precise excision of the focal intron; N_a: number of reads corresponding to alternative splice variants relative to this intron (*i.e.* sharing only one of the two intron boundaries); N_u: number of unspliced reads, co-linear with the genomic sequence. **B,C** Histograms representing the distribution of RAS and RANS values (divided into 5% bins), for protein-coding gene introns. Each line represents one species. Two representative species are colored: *Drosophila melanogaster* (red), *Homo sapiens* (brown). **D**: Description of the variables used to compute AS rate and minor intron relative abundance (MIRA), given a major intron: N^M: number of spliced reads corresponding to the excision of the major intron; N_i^m: number of spliced reads corresponding to the excision of a minor intron (i); N^m: total number of spliced reads corresponding to the excision of minor introns.

computed the rate of alternative splicing with respect to introns of the major isoform. We will hereafter use the term ‘splice variant’ (SV) to refer to those splicing events that are detected in a minority of transcripts (*i.e.* with RAS ≤ 0.5 or RANS ≤ 0.5) .

We focused our analyses on major introns interrupting protein-coding regions (*i.e.* we excluded introns located within UTRs, Materials & Methods). In vertebrates, each BUSCO gene contains on average 8.4 major introns (Supplementary Tab. 1). The intron density is more variable among insect clades, ranging from 2.8 major introns *per* BUSCO gene in Diptera to 6.1 in Blattodea. As expected, most major introns have GT/AG splice sites (99.1% on average across species), and only a small fraction have non-canonical boundaries (0.8% GC/AG and 0.1% AT/AC). The fraction of non-canonical splice sites is slightly higher among minor introns (2.8% GC/AG and 0.3% AT/AC). This might reflect a true biological difference but might also be caused by the presence of some false positives in the set of minor introns. In any case, the difference in splice signal usage between minor and major introns is small, which indicates that the vast majority of detected minor introns correspond to *bona fide* splicing events.

The proportion of major introns for which AS has been detected (*i.e.* with $\mathrm{N_a} > 0$) ranges from 16.8% to 95.7% depending on the species (Supplementary Tab. 1). This metric is however not very meaningful because it directly reflects differences in sequencing depth across species (the higher the sequencing effort, the higher the probability to detect a rare SV, Supplementary Fig. 2). To allow a comparison across taxa, we computed the AS rate of introns, normalized by sequencing depth ($\mathrm{AS} = \frac{\mathrm{N_a}}{\mathrm{N_s} + \mathrm{N_a}}$, Materials & Methods). The average AS rate for BUSCO genes varies by a factor of 5 among species, from 0.8% in *Drosophila grimshawi* (Diptera) to 3.8% in *Megachile rotundata* (Hymenoptera) (3.4% in humans). Interestingly, the average AS rates of BUSCO gene introns are significantly correlated with the three proxies of N_{e}: species longevity (Fig. 3A), body length and the dN/dS ratio (Supplementary Fig. 3A,B). These correlations are positive, which implies that AS rates tend to increase when N_{e} decreases. It is noteworthy that despite the fact that these proxies are not strongly correlated with each other (Fig. 1B,C), they all show similar relationships with AS rates. Thus, these observations are consistent with the hypothesis that N_{e} has an impact on the evolution of AS rate.

One limitation of our analyses is that we used heterogeneous sources of transcriptomic data. To obtain enough sequencing depth, we combined for each species many RNA-seq samples, irrespective of their origin (whole body, or specific tissues or organs, in adults or embryos, etc.). It is known that genome-wide average AS rates vary according to tissues or developmental stages (Barbosa-Morais *et al.*, 2012; Mazin *et al.*, 2021), and according to environmental conditions (John *et al.*, 2021). To explore how this might have affected our results, we repeated our analyses using a recently published dataset that aimed to compare transcriptomes across seven organs, sampled at several developmental stages in seven species (six mammals, one bird) (Cardoso-Moreira *et al.*, 2019). In agreement with previous reports (Mazin *et al.*, 2021), our analysis of BUSCO genes revealed substantial differences in AS rates among organs, with consistent patterns of variation across species. For instance, in all species, testes and brain tissues show higher AS rates than liver and kidney (Fig. 3B). However, the variation in AS rate among organs in each species is limited compared to differences in AS rate among species. Specifically, in an ANOVA analysis performed on the average AS rate across BUSCO gene introns, with the species and the organ of origin as explanatory variables, the species factor explained 89% of the total variance, while the organ factor explained only 9%. Among insects, we found only one species (*Dendroctonus ponderosae*) for which RNA-seq samples were available from multiple tissues. Here again, the variance in AS rate among tissues was limited compared to inter-species variability (Supplementary Fig. 9). Thus, despite

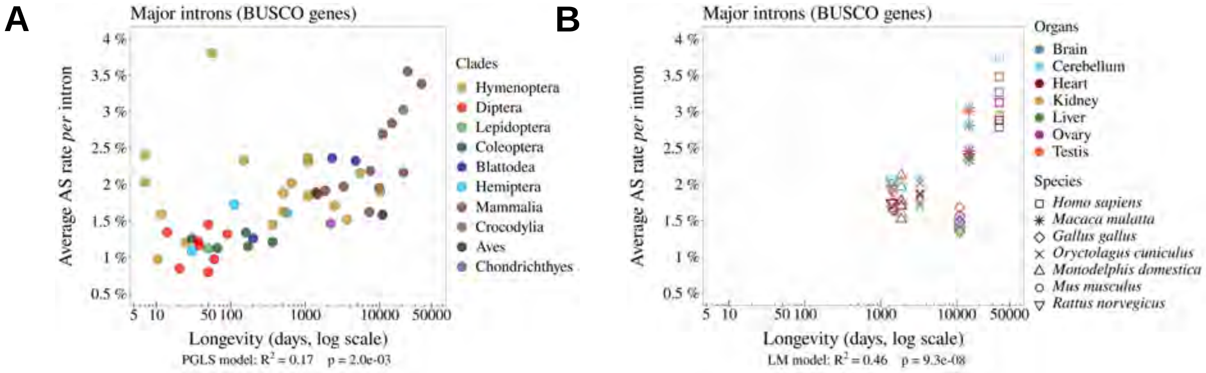

Figure 3: **The rate of alternative splicing correlates with life history traits across metazoans. A**: Relationship between the *per* intron average AS rate of an organism and its longevity (days, log scale). **B**: Variation in average AS rate across seven organs (brain, cerebellum, heart, liver, kidney, testis and ovary) among seven vertebrate species (RNA-seq data from Cardoso-Moreira *et al.* (2019)). AS rates are computed on major introns from BUSCO genes (Materials & Methods).

the variability that can be introduced by the heterogeneity of RNA-seq samples, the relationship between AS rate and longevity remains detectable among these seven species (Fig. 3B).

Functional vs. non-functional alternative splicing

The negative correlation observed between N_e and alternative splicing rates is consistent with the hypothesis that differences in AS rates across species are driven by variation in the rate of splicing errors (drift barrier model). This does not exclude however that functional splicing variants might also contribute to AS rate variation across species. To evaluate this point, we selected a subset of SVs that are enriched in functional AS events. To do this, we reasoned that selective pressure against the waste of resources should maintain splicing errors at a low rate (as low as permitted by the drift barrier), whereas functional SVs are expected to represent a sizeable fraction of the transcripts expressed by a given gene, at least in some specific conditions (cell type, developmental stage. . .). Thus, functional SVs are expected to be enriched among abundant SVs compared to rare SVs.

To assess this prediction, we analyzed the proportion of SVs that preserve the reading frame according to their abundance relative to the major isoform. For this, we focused on minor introns that share a boundary with one major intron and that have their other boundary at less than 30 bp from the major splice site (either in the flanking exon or within the major intron). We determined whether the distance between the minor intron boundary and the major intron boundary was a multiple of 3. We computed the abundance of each minor isoform, relative to the corresponding major isoform, with the following formula: Minor intron relative abundance $\mathrm{MIRA_i} = \frac{\mathrm{N_i^m}}{\mathrm{N^M + N^m}}$ (see Fig. 2D).

We divided minor introns into 5% bins according to their MIRA and computed for each bin the proportion of minor introns that maintain the reading frame of the major isoform (Fig. 4A). In all species, we observe that this proportion varies according to the abundance of splice variants, with two distinct regimes (Fig. 4A). First, for MIRA values above 5%, the proportion of frame-preserving variants correlates positively with

MIRA, reaching up to 60%-70% for the most abundant isoforms. Second, for MIRA values below 1%, the proportion of frame-preserving variants does not covary with MIRA, and fluctuates around 30 to 40%, close to the random expectation (33%). The excess of frame-preserving variants among the most abundant isoforms implies that a substantial fraction of them is under constraint to encode functional protein isoforms. This fraction varies from 0% for MIRA values below 1%, to 50% for isoforms with the highest MIRA values. It should be noted that these estimates correspond to a lower bound, since it is possible that some frame-shifting splice variants are functional. Nevertheless, these observations clearly indicate that the subset of SVs with MIRA values $> 5\%$ (hereafter referred to as 'abundant SVs') is strongly enriched in functional isoforms relative to other SVs (MIRA $\leq 5\%$, hereafter referred to as 'rare SVs'). Of note, the subset of rare SVs represents the vast majority of the SV repertoire (from 62.4% to 96.9% depending on the species; Supplementary Tab. 1).

Investigating selective pressures on minor splice sites

A complementary approach to assess the functionality of AS events consists in investigating signatures of selective constraints on splice sites. For this, we used polymorphism data from *Drosophila melanogaster* and *Homo sapiens* to measure single-nucleotide polymorphism (SNP) density at major and minor splice sites, considering separately rare and abundant SVs. We focused on the first two and last two bases of each intron (consensus sequences GT, AG), which represent the most constrained sites within splice signals. We studied minor introns that share one splice site with a major intron and we measured SNP density at the corresponding major and minor splice sites. To account for constraints acting on coding regions, we considered separately minor splice sites that were located in an exon or in an intron of the major isoform. As negative controls, we selected AG or GT dinucleotides that were unlikely to correspond to alternative splice sites (Fig. 5, Materials & Methods). Furthermore, for *Homo sapiens* we controlled for the presence of hypermutable CpG dinucleotides (Tomso and Bell, 2003) (Supplementary Fig. 4, Materials & Methods).

For both species, the lowest SNP density is observed at major splice signals, which reflects the strong selective constraints on these sites (Fig. 5). In *Drosophila melanogaster*, there is also a strong signature of selection on minor splice signals of abundant SVs: both in introns and in exons, the SNP density at minor splice signals of abundant SVs is much lower than in corresponding controls (from -37% to -74%, Fig. 5A) and than in minor splice signals of rare SVs (from -38% to -71%, Fig. 5B). This observation confirms that abundant SVs are strongly enriched in functional variants compared to rare SVs. In *Homo sapiens*, patterns of SNP density showed little evidence of selective constraints on minor splice sites, irrespective of the abundance of SVs (Fig. 5C,D): minor acceptor splice sites (AG) located within the major intron show a weak but significant SNP deficit relative to corresponding control sites (p-value $< 1\text{x}10^{-5}$), but other categories of minor splice sites do not show any sign of selective constraints. The fact that the signature of selection on minor splice signals is much weaker in humans compared to *Drosophila* is indicative of a lower prevalence of functional variants, even among abundant SVs. This observation is therefore in total contradiction with the adaptive hypothesis (more functional alternative splicing in complex organisms).

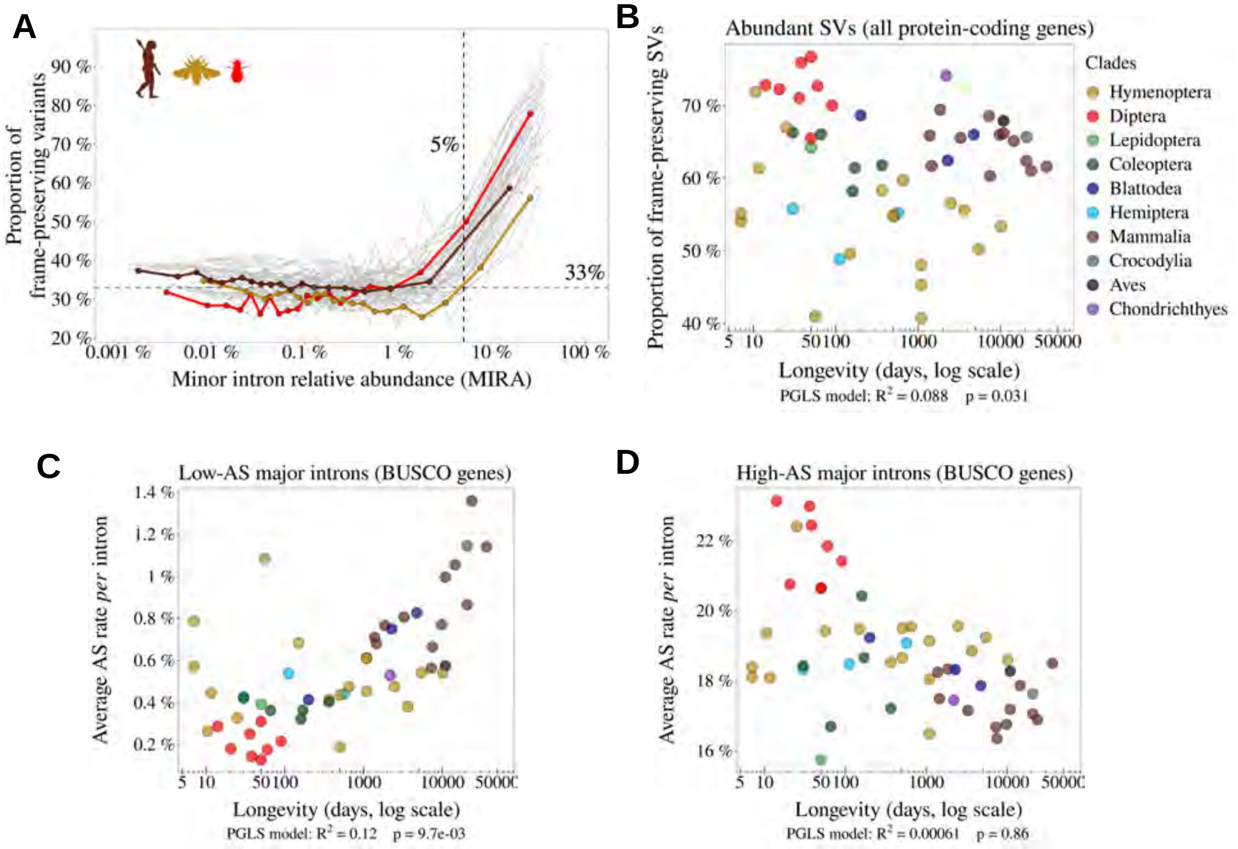

Figure 4: **Variation in AS rate across metazoans: distinguishing abundant splice variants (enriched in functional variants) from rare splice variants. A**: Frame-preserving isoforms are strongly enriched among abundant splice variants (SVs). For each species, SVs were classified into 20 equal-size bins according to their abundance relative to the major isoform (MIRA, see Materials & Methods), and the proportion of frame-preserving SVs was computed for each bin. Each line represents one species. Three representative species are colored: red: *Drosophila melanogaster*, brown: *Homo sapiens*, yellow: *Apis mellifera*. We used a threshold MIRA value of 5% to define 'abundant' vs. 'rare' SVs. **B**: Proportion of frame-preserving SVs among abundant SVs across metazoans. Each dot represents one species. All annotated protein-coding genes are used in the analysis. **C,D**: Relationship between the average *per* intron AS rate of an organism and its longevity (days, log scale). Only BUSCO genes are used in the analysis. **C**: Low-AS major introns (*i.e.* major introns that do not have any abundant SV), **D**: High-AS major introns (*i.e.* major introns having at least one abundant SV).

The splicing rate of rare SVs is negatively correlated with gene expression levels

The above analyses are consistent with the hypothesis that the vast majority of rare SVs correspond to erroneous transcripts, and that changes in N_e contribute to variation in AS rate across taxa by shifting the selection-mutation-drift balance. If true, then this model predicts that the erroneous AS rate should also vary among genes, according to their expression level. Indeed, it has been shown that the selective pressure on splicing accuracy is stronger on highly expressed genes (Saudemont *et al.*, 2017). This reflects the fact that for a given splicing error rate, the waste of resources (both in terms of metabolic cost and of futile mobilization

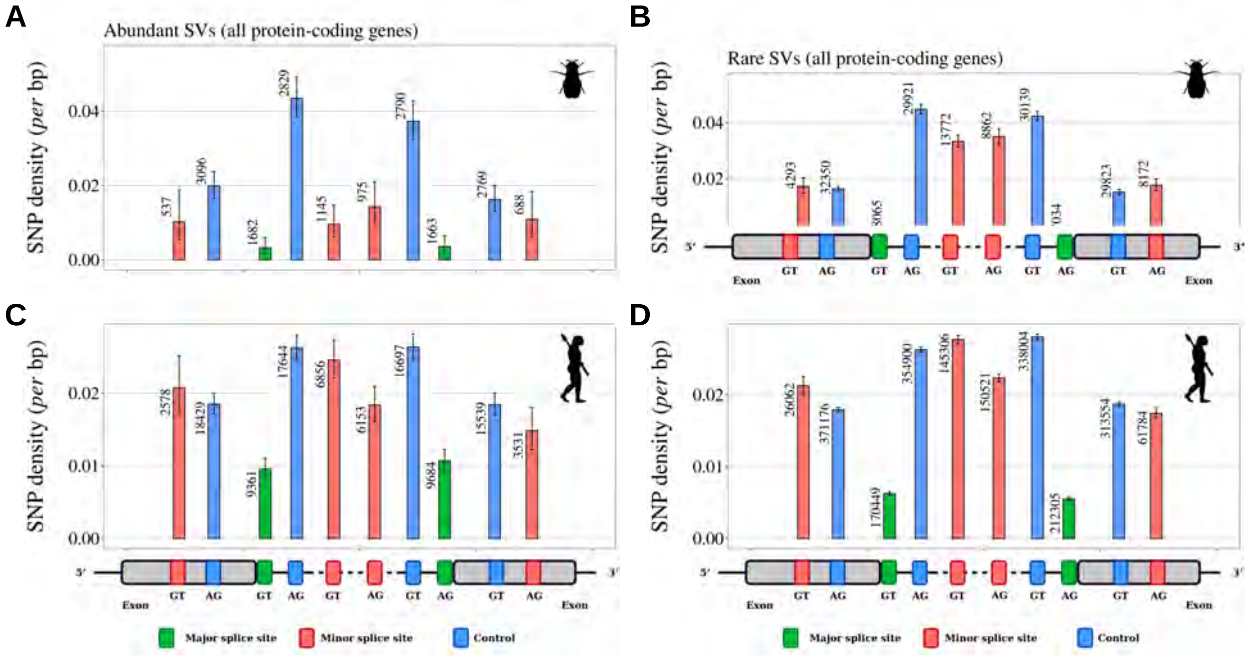

Figure 5: **Variation in selective constraints on alternative splice signals from rare and abundant SVs.** For each minor intron sharing one boundary with a major intron, we measured the SNP density at its minor splice site (red), and at the corresponding major splice site (green). We distinguished minor splice sites that are located in an exon or in an intron of the major isoform. As a control (blue), we selected AG or GT dinucleotides that are unlikely to correspond to alternative splice sites, namely: AG dinucleotides located toward the end of the upstream exon or the beginning of the intron (unlikely to correspond to a genuine acceptor site), and GT dinucleotides located toward the beginning of the downstream exon or the end of the intron (unlikely to correspond to a donor site). To increase the sample size, we analyzed data from all annotated protein-coding genes (and not only the BUSCO gene set). The number of sites studied is shown at the top of each bar. Error bars represent the 95% confidence interval of the proportion of polymorphic sites (proportion test). **A,B**: SNP density in *Drosophila melanogaster* (polymorphism data from 205 inbred lines derived from natural populations, N=3,963,397 SNPs (Huang *et al.*, 2014; Mackay *et al.*, 2012)). **C,D**: SNP density in *Homo sapiens* (polymorphism data from 2,504 individuals, N=80,868,061 SNPs (Auton *et al.*, 2015)). We excluded dinucleotides affected by CpG hypermutability (Materials & Methods, see Supplementary Fig. 4 for CpG sites). **A,C**: Abundant SVs (MIRA > 5%). **B,D**: Rare SVs (MIRA ≤ 5%).

of cellular machineries) increases with gene expression level (Saudemont *et al.*, 2017; Xiong *et al.*, 2017). Thus, the selection-mutation-drift balance should lead to a negative correlation between gene expression level and the rate of splicing errors. To test this prediction, we focused on low-AS major introns, *i.e.* introns that are unlikely to have functional SVs. For each species, we considered all major introns with a sufficient sequencing depth to have a precise measure of their AS rate ($N_s + N_a \geq 100$). The selected subset represents 38.1% to 86.7% of major introns of each species (median=70.9%). Introns were then divided into 20 bins of equal size, according to the expression level of the corresponding genes. For each species, we computed the Pearson correlation between the average AS rate and the average expression level across bins. We observed a negative correlation between AS rates and gene expression levels in 52 out of the 53 species (significant with

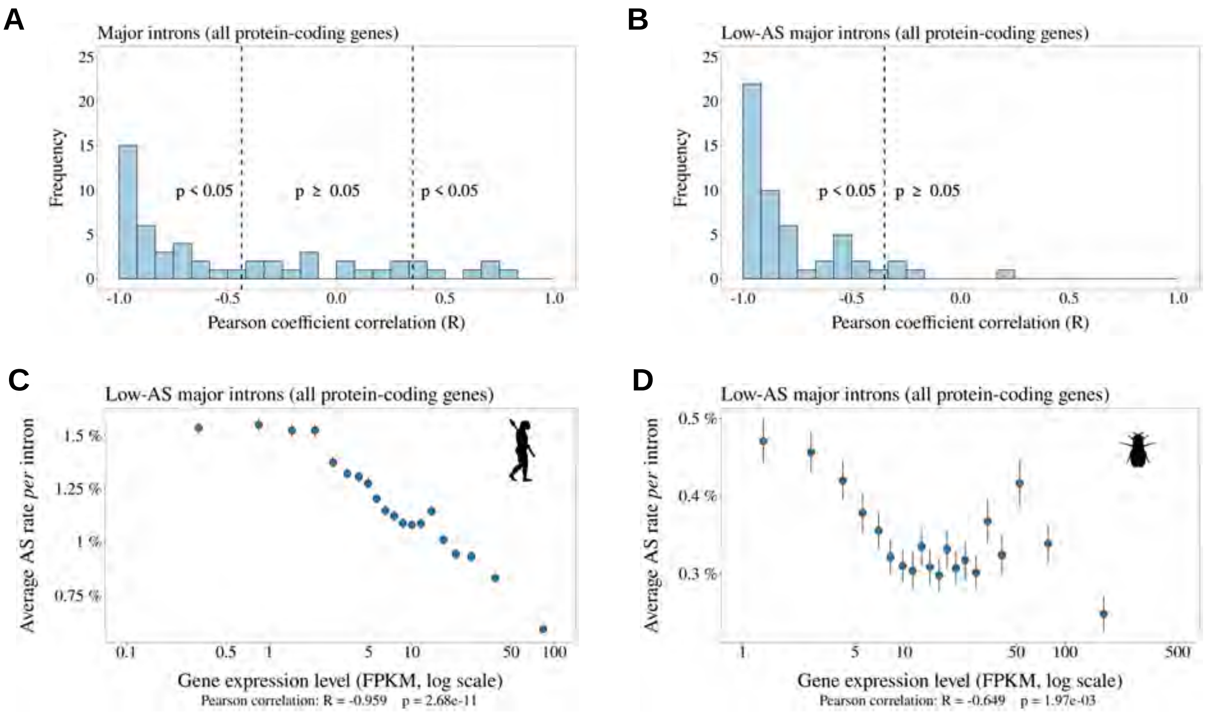

Figure 6: **Relationship between AS rate and gene expression level.** For each species, we selected major introns with a sufficient sequencing depth to have a precise measure of their AS rate ($N_s + N_a \geq 100$). We divided major introns into 5% bins according to their gene expression level and computed the correlation between the average AS rate and median expression level across the 20 bins. To increase sample size, these analyses were based on all annotated protein-coding genes (and not only the BUSCO gene set). **A**: Distribution of Pearson correlation coefficients (R) between the AS rate and expression level observed in the 53 metazoans. The vertical dashed lines indicates the thresholds under and above which correlations are significant (*i.e.* p-value < 0.05). **B**: Distribution of Pearson correlation coefficients computed on the subsets of low-AS major introns (*i.e.* after excluding major introns with abundant SVs). **C,D**: Two representative species illustrating the negative relation between the average AS rate of low-AS major introns and the expression level of their gene. Error bars represent the standard error of the mean. **C**: N=127,599 low-AS major introns from *Homo sapiens*, **D**: N=31,357 low-AS major introns from *Drosophila melanogaster*.

$p < 0.05$, in 48/53 species; Fig. 6B; two representative examples are shown in Fig. 6C and 6D). This pattern indicates that in almost all metazoan species, genes with a higher expression level have a lower AS rate, consistent with the hypothesis the rate of splicing errors is shaped by the selection-mutation-drift balance. It should be noted that this negative correlation between AS rate and gene expression level is not expected for functional SVs (there is *a priori* no reason why the AS rate of functional SVs should be higher in weakly expressed genes than in highly expressed genes). Interestingly, when we performed this analysis on all introns (including those with abundant SVs, which are enriched in functional variants), then most species (31/53) still showed a negative correlation between AS rate and gene expression level (Fig. 6A), but some species, such as *Drosophila melanogaster* showed the opposite pattern (Supplementary Fig. 5). This probably reflects

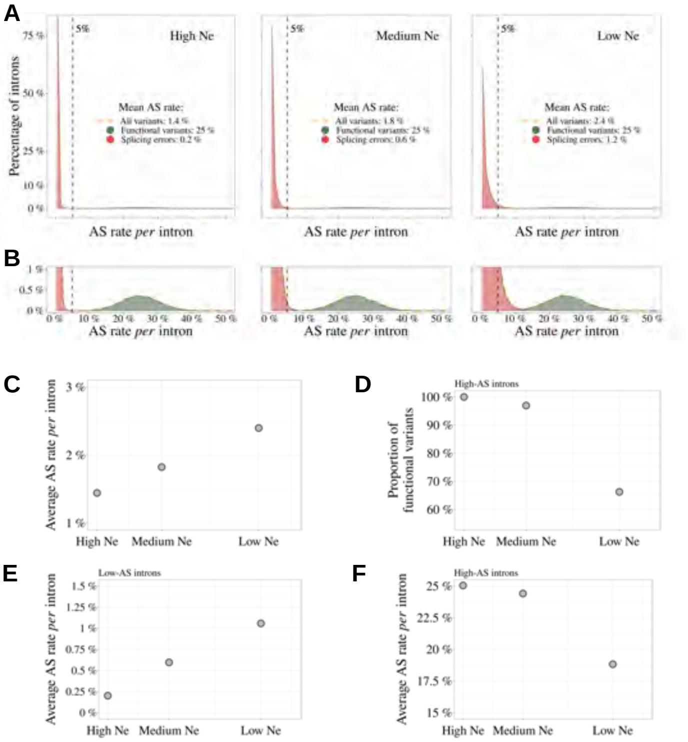

Figure 7: **Impact of the drift-barrier on the genome-wide AS rate: model predictions**. To illustrate the impact of the drift barrier, we sketched a simple model, with three hypothetical species of different N_e. In this model, the repertoire of SVs consists of a mixture of functional variants and splicing errors. We assumed that in all species, only a small fraction of major introns (5%) produce functional SVs, but that these variants have a relatively high AS rate (average=25%, standard deviation=5%; see Materials & Methods for details on model settings). Splicing error rates were assumed to be gamma-distributed, with a low mean value. Owing to the drift barrier effect, the mean error rate was set to vary from 0.2% in species of high N_e to 1.2% in species of low N_e (these parameters were chosen to match approximately the AS rates observed in empirical data for rare SVs). **A** Genome-wide distribution of AS rates in each species (high N_e, medium N_e and low N_e). Each distribution corresponds to a mixture of functional SVs (green) and splicing errors (red). **B**: Zoom on the y-axis to better visualize the contribution of functional SVs to the whole distribution: rare SVs (AS $\leq$ 5%) essentially correspond to splicing errors, while abundant SVs (AS > 5%) correspond to a mixture of functional and spurious variants, whose relative proportion depend on N_e. The following panels show how these different distributions, induced by differences in N_e, impact genome-wide AS patterns. **C**: Relationship between the average AS rate *per* major intron and N_e. **D**: Fraction of frame-preserving splice variants among introns with high AS rates *vs* N_e. Relationship between the average AS rate *per* intron and N_e, for 'low-AS' major introns (MIRA $\leq$ 5%) (**E**), and for 'high-AS' major introns (MIRA > 5%) (**F**).

that fact that, in those species, functional AS events make a significant contribution to the genome-wide average AS rate.

Discussion

To investigate the factors that drive variation in AS rates across species, we analyzed publicly available RNA-seq data across a large set of 53 species, from diverse metazoan clades, covering a wide range of N_e values. To facilitate comparisons across species, we sought to limit the impact of the among-gene variance in AS rates. For this, we primarily based our analyses on a common set of nearly 1,000 orthologous protein-coding genes (BUSCO gene set). We focused our study on introns located within protein-coding regions, because introns from UTRs or lncRNAs are expected to be subject to different functional constraints. We measured AS rates on introns corresponding to a major isoform. When sequencing depth is limited, the set of introns for which AS can be quantified is biased toward the most highly expressed genes. To avoid this bias, we restricted our study to species for which the median sequencing depth of BUSCO exons was above 200. With this setting, on average 96.9% of BUSCO annotated introns could be analyzed in each species (Supplementary Tab. 1).

We observed a 5-fold variation in the average AS rate of BUSCO introns across species from 0.8% in *Drosophila grimshawi* (Diptera) to 3.8% in *Megachile rotundata* (Hymenoptera)(Fig. 3A). In agreement with previous work, we observed that AS rates tend to be high in vertebrates (average=2.3%), and notably in primates (average=3.1%) (Barbosa-Morais *et al.*, 2012; Chen *et al.*, 2014; Mazin *et al.*, 2021). This observation was previously interpreted as an evidence that AS played an important role in the diversification of the functional repertoire necessary for the development of more complex organisms (Chen *et al.*, 2014). However, this pattern is also compatible with the hypothesis that variation in AS rates across species result from differences in splicing error rates, which are expected to be higher in species with low N_e (Bush *et al.*, 2017). Indeed, consistent with this drift barrier hypothesis, we observed significant correlations between AS rates and proxies of N_e (Fig. 3B, Supplementary Fig. 3A,B).

In their original study, (Chen *et al.*, 2014) investigated the hypothesis that variation in AS rates across taxa might be driven by variation in N_e. For this, they focused on 12 species, for which they had measured levels of polymorphism at silent sites (π). They found that the correlation between AS rate and the number of cell types (proxy for organismal complexity) remained significant after controlling for π. They therefore concluded that the association between the cellular diversity and alternative splicing was not a by-product of reduced effective population sizes among more complex species. This conclusion was however based on a very small sample of species. More importantly, it assumed that π could be taken as a proxy for N_e. At mutation-drift equilibrium, π is expected to be proportional to N_eu (where u is the mutation rate *per* bp *per* generation). Thus, if u is constant across taxa, π can be used to estimate variation in N_e. However, the dataset analyzed by Chen et al (2014) included very diverse eukaryotic species, with mutation rates ranging from 1.7x10^{10} mutation *per* bp *per* generation in budding yeast, to 1.1x10^{8} mutation *per* bp *per* generation in humans (Lynch *et al.*, 2016). Hence, at this evolutionary scale, variation in N_e cannot be directly inferred from π without accounting for variation in u. Moreover, the drift barrier hypothesis states that the AS rate of a species should reflect the genome-wide burden of slightly deleterious substitutions, which is expected to depend on the intensity of drift over long evolutionary times (*i.e.* long-term N_e). Conversely, π reflects N_e over a short period of time (of the order of N_e generations), and can be strongly affected by recent population bottlenecks (too recent to have substantially impacted the genome-wide deleterious substitution load). The drift barrier hypothesis therefore predicts that the splicing error rate should correlate more strongly with proxies of long-term N_e (such as dN/dS, life history traits, or organismal complexity) than with π. The fact that AS rates remained significantly correlated to cellular diversity after controlling for π (Chen *et al.*, 2014) is therefore not a conclusive argument against the drift barrier hypothesis.

To contrast the two models (drift barrier vs diversification of the functional repertoire in complex organisms), we sought to distinguish functional splice isoforms from erroneous splicing events. Based on the assumption that splicing errors should occur at a low frequency, we split major introns into two categories, those with abundant SVs (MIRA $> 5\%$), and those without (MIRA $\leq 5\%$). Rare SVs represent the vast majority of the repertoire of splicing isoforms detected in a given transcriptome (from 62.4% to 96.9% according to the species; Supplementary Tab. 1). Two lines of evidence indicate that the small subset of abundant isoforms is strongly enriched in functional transcripts relative to other SVs. First, we observed that in all species, the proportion of SVs that preserve the reading frame is much higher among abundant SVs than among rare SVs (Fig. 4A). Second, the analysis of polymorphism data in *Drosophila* indicates that the average level of purifying selection on alternative splice sites is much stronger for abundant than rare SVs (Fig. 5A,B).

If variation in AS rate across species had been driven by a higher prevalence of functional SVs in more complex organisms, one would have expected the proportion of frame-preserving SVs to be stronger in vertebrates than in insects, in particular for the set of introns with high AS rate (*i.e.* enriched in functional SVs). On the contrary, the highest proportion of frame-preserving SVs is observed in dipterans (Fig. 4B). In fact, the overall higher AS rate of vertebrates (Fig. 3A) is driven by the set of introns with a low AS rate (Fig. 4C), *i.e.* the set of introns in which the prevalence of functional SVs is the lowest. On the contrary, among the set of introns with high AS rate, vertebrates have lower AS rates than insects (Fig. 4D).

These observations are difficult to reconcile with the hypothesis that the higher AS rate in vertebrates results from a higher rate of functional AS. Conversely, these observations fit very well with a model where variation in AS rate across species is entirely driven by variation in the efficacy of selection against splicing errors. To illustrate this model, let us consider three hypothetical species with different N_e, in which a small fraction of major introns (say 5%) is subject to functional alternative splicing. Let us consider that the distribution of AS rates of functional splicing variants is the same for all species (*i.e.* independent of N_e), with a mean of 25% (and a standard deviation of 5%). In addition, we assume that all major introns are potentially affected by splicing errors, with a mean error rate ranging from 0.2% in species of high N_e to 1.2% in species of low N_e, owing to the drift barrier effect (these parameters were set to match approximately the AS rates observed in empirical data for rare SVs). The distributions of AS rate given by this model are presented in Fig. 7A: rare SVs (MIRA $\leq 5\%$) essentially correspond to splicing errors, while abundant SVs (MIRA $> 5\%$) correspond to a mixture of functional and spurious variants, whose relative proportion depend on N_e (Fig. 7B). Interestingly, the predictions of this simple model fit remarkably well with our observations: we observed a positive correlation between AS rate and longevity (*i.e.* a negative correlation with N_e) for the set of low-AS major introns (Fig. 4C), but an opposite trend for high-AS major introns (Fig. 4D), as predicted by the model (Fig. 7D,E). Given that high-AS major introns represent only a small fraction of major introns, this model predicts that, overall, AS rates correlate negatively with N_{e} (Fig. 7), as observed in empirical data (Fig. 3A, Supplementary Fig. 3).

It should be noted that the BUSCO dataset corresponds to genes that are strongly conserved across species, often highly expressed, and hence might not be representative of the entire genome. Notably, AS rates are on average lower in the BUSCO gene set than in other genes, even after accounting for their expression level (Supplementary Fig. 5). However, results remained qualitatively unchanged when we repeated our analyses on the whole set of annotated protein-coding genes for each species: correlations between AS rates and N_{e} proxies are slightly weaker than on the BUSCO subset, but remain significant (Supplementary Fig. 6).

The model also predicts that the proportion of functional SVs among high-AS major introns should vary with N_{e} (Fig. 7C). To assess this point, we measured in each species the enrichment in reading frame-preserving events among abundant SVs compared to rare SVs. As predicted, this estimate of the prevalence of functional SVs tends to decrease with decreasing N_{e} proxies (*e.g.* Fig. 3A, where N_{e} is approximated by longevity). However, these correlations are weak, marginally significant after accounting for phylogenetic inertia with only two of the three N_{e} proxies, and not robust to multiple testing issues (Supplementary Fig. 7). Thus, N_{e} does not appear to be a strong predictor of the prevalence of functional SVs among high-AS major introns.

According to the drift-barrier model, the level of splicing errors is expected to decrease with increasing selective pressure. In all above analyses, we considered AS rates measured *per* intron, and not *per* gene. Yet, the trait under selection is the *per*-gene error rate, which depends not only on the error rate *per* intron, but also on the number of introns *per* gene. Given that intron density varies widely across clades (from 2.8 introns *per* gene in diptera to 8.4 introns *per* gene in vertebrates; Supplementary Tab. 1), the correlations reported above between AS rates and N_{e} may undervalue the predictive power of the drift-barrier model. The RNA-seq datasets that we analyzed consist of short-read sequences, which do not allow a direct quantification of the *per*-gene AS rate. We therefore indirectly estimated the *per*-gene AS rate in each species, based on the *per*-intron AS rate and on the number of introns *per* gene (Materials & Methods). Interestingly, as predicted by the drift-barrier model, N_{e} proxies correlate more strongly with this estimate of the *per*-gene AS than with the *per*-intron AS rates (Supplementary Fig. 8).

One other important prediction of the drift barrier model is that splicing error rate should vary not only across species according to N_{e}, but also among genes, according to their expression level. Indeed, for a given splicing error rate, the waste of resources (and hence the fitness cost) is expected to increase with the level of transcription. Thus, the selective pressure for optimal splice signals is expected to be higher, and hence the error rate to be lower, in highly expressed genes. Consistent with that prediction, nearly all species show a negative correlation between gene expression level and AS rate in low-AS major introns (Fig. 6C).

It should be noted that our analyses suffer from several important limitations. First, the proxies that we considered for N_{e} are quite noisy (Fig. 1). Second, to maximize the number of species in our analyses, we had to use very heterogeneous sources of RNA (whole-body, specific tissues, or organs, at different life stages, in different sexes, different environmental conditions, etc.). Third, we used short-read sequencing data, which allow the quantification of AS rates for individual introns, but do not provide a direct measure of AS rates *per* gene. Hopefully progress of long-read sequencing technologies will soon allow the comparative analysis of

AS rates on full-length transcripts (*e.g.* see Leung *et al.* (2021)). But presently, publicly available long-read transcriptomic data are restricted to a narrow set of model organisms, and their sequencing depth is still too limited to quantify rare splicing events. The fact that we detected significant correlations between AS rate and the three N_e proxies, despite these uncontrolled sources of variability, suggests that we underestimate the effect of N_e on AS rates.

Thus, overall, all observations fit qualitatively well with the predictions of the drift barrier model, according to which most of the variation in AS rate across species reflects differences in splicing error rates. Of course, this model is not in contradiction with the fact, well established, that some AS events play an essential role in various processes. Different criteria can be used to distinguish functional SVs from spurious splicing events. Notably, AS events that are strongly tissue-specific or developmentally dynamic tend to be more conserved across species, which indicates that a substantial fraction of them are evolutionary constrained, and hence functional (Mudge *et al.*, 2011; Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012; Reyes *et al.*, 2013). The abundance of a SV is also an important predictor of its functionality. In particular, we observed that in all species, the proportion of frame-preserving events is much higher among abundant SVs than among rare SVs (Fig. 4A). We note however that the threshold that we used to define abundant SVs is somewhat arbitrary. In fact, according to our model, this class of SVs corresponds to a mixture of functional and spurious events, whose relative proportion is expected to depend on N_e (Fig. 7C). Thus, in low-N_e species, even the subset of abundant SVs includes a substantial fraction of errors. This probably explains why, contrarily to *Drosophila*, we do not detect any signature of purifying selection on alternative splice signals in humans, even for abundant SVs (Fig. 5).

In conclusion, all observations fit with the hypothesis that random genetic drift sets an upper limit on the capacity of selection to prevent splicing errors. It should be noted that this limit on the optimization of genetic systems is expected to affect not only splicing, but all aspects of gene expression. Notably, there is a growing body of evidence that the complexity of transcripts produced by eukaryotic genes (resulting from alternative transcription initiation, polyadenylation, splicing or back-splicing, RNA editing) often does not correspond to fine-tuned adaptations but simply to the accumulation of errors (Pickrell *et al.*, 2010; Saudemont *et al.*, 2017; Xu *et al.*, 2019; Xu and Zhang, 2018; Liu and Zhang, 2018b,a; Xu and Zhang, 2014, 2020; Gout *et al.*, 2013; Zhang and Xu, 2022). It should be noted however that the relationship between the genome-wide error rate and N_e is not expected to be monotonic. Indeed, models predict that in species with very high N_e, selection on each individual gene should favor genotypes that are robust to errors of the gene expression machinery, which in turn, reduces the constraints on the global level of gene expression errors (Rajon and Masel, 2011; Xiong *et al.*, 2017). Thus, paradoxically, species with very large N_e are expected to have gene expression machineries that are more error-prone than species with very small N_e (Rajon and Masel, 2011). This argument was developed by Xiong *et al.* (2017) to account for the fact that transcription error rates had been found to be about 10 times higher in bacteria than in eukaryotes (Traverse and Ochman, 2016; Gout *et al.*, 2013). More recent work indicates that bacterial transcription error rates had been largely overestimated, presumably owing to RNA damages during the preparation of sequencing libraries (Li and Lynch, 2020). Given these uncertainties in the measures of transcription error rates, it seems for now difficult to interpret the differences reported across species. But in any case, it is important to note that it is in principle possible that the drift barrier affects differently the different steps of the gene expression process. It would therefore be important to investigate to which extent each step of gene expression responds (or not) to variation in N_e. As illustrated here by the relationship observed between alternative splicing and N_e, it appears essential to consider the contribution of non-adaptive evolutionary processes when trying to understand the origin of eukaryotic gene expression complexity.

Materials & Methods

Genomic and transcriptomic data collection

To analyze AS rate variation across metazoans, three types of information are required: transcriptome sequencing (RNA-seq) datasets, genome assemblies, and gene annotations. To obtain this data, we first queried the Short Read Archive database (Leinonen *et al.*, 2011) to extract publicly available RNA-seq datasets. We also queried the NCBI Genomes database (NCBI Resource Coordinators, 2018) to retrieve genomic sequences and annotations. When this project was initiated, the vast majority of metazoans represented in this database corresponded to vertebrates or insects. We therefore decided to focus our analyses on these two clades (N=69 species).

Identification of orthologous gene families

To be able to compare average AS rates across species, given that AS rates vary among genes (Saudemont *et al.*, 2017), it is necessary to analyze a common set of orthologous genes. We searched for homologues of the BUSCOv3 (Benchmarking Universal Single Copy Orthologs, (Seppey *et al.*, 2019)) metazoan gene subset (N=978 genes) in each of the 69 genomes. To do this, we used the software BUSCO v.3.1.0 to associate BUSCO genes to annotated protein sequences. For each species, BUSCO genes were removed from the analysis if they were associated to more than one annotated gene or to an annotated gene that was associated to more than one BUSCO gene.

RNA-seq data processing and intron identification

We aligned the RNA-seq reads on the corresponding reference genomes with HISAT2 v.2.1.0 (Kim *et al.*, 2019). We built the genome indexes using annotated introns and exons coordinates in addition to genome sequences, to improve splice junction detection sensitivity. The maximum allowed intron length was fixed to 2,000,000 bp. We then extracted intron coordinates from HISAT2 alignments using an in-house perl script that scanned for CIGAR strings containing N, which indicate regions that are skipped from the reference sequence. For intron detection and quantification we used only uniquely mapping reads that had a maximum mismatch ratio of 0.02. We required a minimum anchor length (that is, the number of bases that align on each flanking exon) of 8 bp for intron detection, and of 5 bp for intron quantification. We kept only those predicted introns that had GT-AG, GC-AG or AT-AC splice signals, and we predicted the strand of the introns based on the splice signal.

We assigned an intron to a gene if at least one of the intron boundaries fell within 1 bp of the annotated exon coordinates of the gene, combined across all annotated isoforms. We excluded introns that could not be unambiguously assigned to a single gene. We distinguish annotated introns (which appear as such in the reference genome annotations) and un-annotated introns, which were detected with RNA-seq data and assigned to previously annotated genes.

We further restricted our analyses to introns located within protein-coding regions. To do this, for each protein-coding gene, we extracted the start codons and the stop codons for all annotated isoforms. We then identified the minimum start codon and the maximum end codon positions and we excluded introns that were upstream or downstream of these extreme coordinates.

The alignment process, which is the most time-consuming step in the pipeline (see Supplementary Fig. 10), can take up to one week when using 16 cores *per* RNA-seq for larger genomes, such as mammals. Additionally, the processed compressed files generated during this process can exceed 7 terabytes in size.

Alternative splicing rate definition

For each intron we noted $\mathrm{N_s}$ the number of reads corresponding to the precise excision of this intron (spliced reads), and $\mathrm{N_a}$ the number of alternatively spliced reads (*i.e.* spliced variant sharing only one of the two intron boundaries). Finally, we note $\mathrm{N_u}$ the number of unspliced reads, co-linear with the genomic sequence, and which overlap with at least 10 bp on each side of an exon-intron boundary. These definitions are illustrated in Fig. 2. We then defined the relative abundance of the focal intron compared to introns with one alternative splice boundary ($\mathrm{RAS} = \frac{\mathrm{N_s}}{\mathrm{N_s} + \mathrm{N_a}}$), as well as relative to unspliced reads ($\mathrm{RANS} = \frac{\mathrm{N_s}}{\mathrm{N_s} + \frac{\mathrm{N_u}}{2}}$).

To compute these ratios we required a minimal number of 10 reads at the denominator. We thus calculated the RAS only if $(\mathrm{N_s} + \mathrm{N_a}) \geq 10$ and the RANS only if $(\mathrm{N_s} + \frac{\mathrm{N_u}}{2}) \geq 10$ (We divided $\mathrm{N_u}$ by 2 because retention is quantified at two sites, which increases the detection power by a factor of 2). If the criteria were not met, the values were labeled as not available (NA). We computed these ratios using reads from all available RNA-seq samples, unless otherwise specified (for example, in sub-sampling analyses). Based on these ratios we defined three categories of introns: major introns, defined as those introns that have RANS > 0.5 and RAS > 0.5; minor introns, defined as those introns that have RANS ≤ 0.5 or RAS ≤ 0.5; unclassified introns, which do not satisfy the above conditions.

For minor introns sharing a boundary with a major intron, we computed the relative abundance of the minor intron (i) with respect to the corresponding major intron, with the following formula: Minor intron relative abundance $\mathrm{MIRA_i} = \frac{\mathrm{N_i^m}}{\mathrm{N^M} + \mathrm{N^m}}$, where $\mathrm{N^M}$ is the number of spliced reads corresponding to the excision of the major intron, $\mathrm{N_i^m}$ is the number of spliced reads corresponding to the excision of a minor intron (i) and $\mathrm{N^m}$ is the total number of spliced reads corresponding to the excision of minor introns (see Fig. 2).

We defined the *per*-gene AS rate as the probability to observe at least one alternative splicing event across all the major introns of a gene. To estimate the per-gene AS rate of a given gene, we assumed that the AS rate is uniform across its major introns, and that AS events occur independently at each intron. We calculated the AS rate for each gene as the number of spliced reads corresponding to the excision of major introns, divided by the number of spliced reads corresponding to minor and major introns ($\frac{\sum \mathrm{N^m}}{\sum \mathrm{N^M} + \mathrm{N^m}}$). The probability for a given gene to produce no splice variant across all its major introns is thus p0=$(1 - \frac{\sum \mathrm{N^m}}{\sum \mathrm{N^M} + \mathrm{N^m}})^{\mathrm{N_i}}$, where $\mathrm{N_i}$ is the number of major introns of the gene. The *per*-gene AS rate (ASg), i.e. the probability to have at least one AS event, is therefore the complement of p0: ASg=1-p0.

Identification of reading frame-preserving splice variants

To determine the proportion of open reading frame-preserving splice variants, we first identified minor introns that had their minor splice site within a maximum distance of 30 bp from the major splice site (either in the flanking exon or within the major intron). Among these introns, we considered that frame-preserving variants are those introns for which the distance between the minor intron boundary and the major intron boundary was a multiple of 3.

Gene expression level

Gene expression levels were calculated with Cufflinks v2.2.1 (Roberts *et al.*, 2011) based on the read alignments obtained with HISAT2, for each RNA-seq sample individually. We estimated FPKM levels (fragments *per* kilobase of exon *per* million mapped reads) for each gene.

The overall gene expression of a gene was computed as the average FPKM across samples, weighted by the sequencing depth of each sample. The sequencing depth of a sample is the median *per*-base read coverage across BUSCO genes.

Phylogenetic tree reconstruction

For each of the 978 BUSCO gene families we collected the longest corresponding proteins identified in each species. We removed proteins for which the amino acid sequence provided with the annotations did not perfectly correspond to the translation of the corresponding coding sequences. We then aligned the resulting sets of protein-coding sequences for each BUSCO gene, using the codon alignment option in PRANK v.170427 (Löytynoja and Goldman, 2008). We translated the codon alignments into protein alignments using the R package seqinr (Charif and Lobry, 2007). To infer the phylogenetic tree rapidly, we sub-sampled the resulting multiple alignments (N=461), selecting alignments with the highest number of species (ranging from 49 to 53 species *per* alignment). We then concatenated these alignments and kept sites that were aligned in at least 30 species. We used RAxML-NG v.0.9.0 (Kozlov *et al.*, 2019) to infer the species phylogeny with a final alignment of 53 taxa and 165,648 sites (amino acids). RAxML was set to perform one model *per* gene with fixed empirical substitution matrix (LG), empirical amino acid frequencies from alignment (F) and 8 discrete GAMMA categories (G8), specified in a partition file with one line *per* multiple alignment. The analysis generated 10 starting trees, 5 starting from a random topology and 5 starting from a tree generated by the parsimony-based randomized stepwise addition algorithm. The best-scoring topology was kept as the final ML tree and 10 bootstrap replicates have been generated.

dN/dS computation

We estimated dN/dS ratios for the BUSCO gene families that were present in at least 45 species (N=922 genes), using the codon alignments obtained with PRANK (see above). We divided the 922 sequence alignments into 18 groups, based on their average GC3 content across species, and concatenated the alignments within each group. We thus obtained concatenated alignments that were 209 kb long on average. We used bio++ v.3.0.0 libraries (Guéguen *et al.*, 2013; Dutheil and Boussau, 2008; Bolívar *et al.*, 2019) to estimate the dN/dS on terminal branches of the phylogenetic tree, for each concatenated alignment. We attributed the dN/dS of the terminal branches to the species that corresponds.

In a first step, we used an homogeneous codon model implemented in bppml to infer the most likely branch lengths, codon frequencies at the root, and substitution model parameters. We used YN98 (F3X4) (Yang and Nielsen, 1998) substitution model, which allows for different nucleotide content dynamics across codon positions. In a second step, we used the MapNH substitution mapping method (Guéguen and Duret, 2018) to count synonymous and non-synonymous substitutions (Dutheil *et al.*, 2012). We defined dN as the total number of non-synonymous substitutions divided by the total number of non-synonymous opportunities, both summed across concatenated alignments, for each branch of the phylogenetic tree. Likewise, we defined dS as the total number of synonymous substitutions divided by the total number of synonymous opportunities, both summed across concatenated alignments. The *per*-species dN/dS corresponds to the ratio between dN and dS, on the terminal branches of the phylogenetic tree.

Life history traits

We used various life history traits to approximate the effective population size of each species. For vertebrates species we considered the maximum lifespan (*i.e.* from birth to death) and body length referenced. For insects we took the maximum lifespan and body length of the *imago*. For eusocial insects and the eusocial mammal *Heterocephalus glaber*, the selected values correspond to the queens. The sources from which the lifespan and the body length information was taken are listed in `data/Data9-supp.pdf` in the Zenodo repository (see Data and code availability).

Analyses of sequence polymorphism

We analyzed the distribution of single nucleotide polymorphisms (SNPs) around splice sites in *Drosophila melanogaster* and *Homo sapiens*.

For *Drosophila melanogaster* we used polymorphism data from the *Drosophila* Genetic Reference Panel (DGRP) (Huang *et al.*, 2014; Mackay *et al.*, 2012), from which we extracted 39,633,97 SNPs that were identified from comparisons across 205 inbred lines. We converted the SNP coordinates from the dm3 genome assembly to the dm6 assembly with the liftOver utility (Hinrichs *et al.*, 2006) of the UCSC genome browser, using a whole genome alignment between the two assemblies downloaded from [https://hgdownload.soe.ucsc.edu/goldenPath/dm3/liftOver/dm3ToDm6.over.chain.gz].

For *Homo sapiens* we used polymorphism data from the 1000 Genomes project, phase 3 release (Auton *et al.*, 2015). This dataset included 80,868,061 SNPs that were genotyped in 2,504 individuals.

For each minor intron sharing one boundary with a major intron, we computed the number of SNPs that occur at their respective splice sites: at their shared boundary, and at the major intron and minor introns specific boundaries.

We focused our study on minor introns that have their specific boundary folding in the exons adjacent to the major intron or in the major intron. As a control, for each minor intron, we searched for one GT and one AG dinucleotides in the interval between 20 and 60 bp with respect to the major splice site, in the neighboring exon and in the major intron, and computed the number of SNPs that occur on these sites. We searched for control AG dinucleotides in the vicinity of the donor splice site of the major intron and for GT dinucleotides in the vicinity of its acceptor splice site, to avoid studying sites that might correspond to unidentified minor splice sites. For *Homo sapiens*, we further divided the splice sites and the control dinucleotides into two groups, depending on whether they were subject to CpG hypermutability or not.

Impact of the drift-barrier on genome-wide AS rates: sketched model

To illustrate the impact of the drift barrier, we sketched a simple model, with three hypothetical species of different N_e (low, medium and high N_e). In each species, the repertoire of SVs consists of two categories: functional variants and spurious variants (which result from errors of the splicing machinery). The rate of splicing error was assumed to be low and to depend on N_e, owing to the drift barrier effect. We considered that in all species, only a small fraction of major introns (5%) produce functional SVs, but that these variants have a relatively high AS rate. The AS rates of functional SVs were modeled by a normal distribution, with a mean of 25% and a standard deviation of 5% (same parameters for the three species). We modeled the distribution of error rates by a gamma distribution, with shape parameter = 1, and with mean values of 0.2%, 0.6% and 1.2% respectively in species of high, medium or low N_e (these parameters were set to match approximately the AS rates observed in empirical data for rare SVs). We then combined the two distributions (functional SVs and splicing errors) to compute the genome-wide average AS rates in each species. We also computed the average AS rate on the subsets of low-AS or high-AS major introns (*i.e.* with AS rates respectively below or above the threshold AS rate of 5%). Finally, we computed the proportion of frame-preserving SVs among high-AS major introns, assuming that two thirds of splicing errors induce frameshifts and that all functional SVs preserve the reading frame.

Acknowledgements

We thank Loïc Guille for his contribution to an initial pilot study, Tristan Lefébure for insightful discussions and Laurent Guéguen for his help on dN/dS analyses. Computational analyses were performed using the computing facilities of the CC LBBE/PRABI and the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013).

Funding

This work was funded by the French National Research Agency (ANR-20-CE02-0008-01 ”NeGA” and ANR-17-CE12-0019-01 ”LncEvoSys”).

Conflict of interest disclosure

The authors declare the following non-financial conflict of interest: Laurent Duret is recommender for PCI Evol Biol.

Data and code availability

All processed data that we generated and used in this study, as well as the scripts that we used to analyze the data and to generate the figures, are available on zenodo DOI: `https://doi.org/10.5281/zenodo.7789408`.

In particular, the sources of transcriptomic data, genome assemblies and annotations are reported in the Zenodo archive in `data/Data1-supp.tab`. The archive includes several directories, including `figure`, which contains the necessary materials to produce the figures of the manuscript. Rmarkdown scripts located in the `table_supp` directory were used to generate supplementary tables, which are also saved in the same directory. The processed data used to generate figures and conduct analyses are stored in the `data` directory in tab-separated text format.

References

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M., Pozo, A. d., Vázquez, J., Valencia, A., and Tress, M. L. 2015. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLOS Computational Biology*, 11(6): e1004325. Publisher: Public Library of Science.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. A., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, E. B. I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and National Eye Institute, N. 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74. Number: 7571 Publisher: Nature Publishing Group.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencowe, B. J. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)*, 338(6114): 1587–1593.

Bhuiyan, S. A., Ly, S., Phan, M., Huntington, B., Hogan, E., Liu, C. C., Liu, J., and Pavlidis, P. 2018. Systematic evaluation of isoform function in literature reports of alternative splicing. *BMC Genomics*, 19(1): 637.

Blencowe, B. J. 2017. The Relationship between Alternative Splicing and Proteomic Complexity. *Trends in Biochemical Sciences*, 42(6): 407–408. Publisher: Elsevier.

Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., and Mugal, C. F. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology*, 20(1): 5.

Bush, S. J., Chen, L., Tovar-Corona, J. M., and Urrutia, A. O. 2017. Alternative splicing and the evolution of phenotypic novelty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713): 20150474. Publisher: Royal Society.

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., Mazin, P. V., Xenarios, I., Harshman, K., Mort, M., Cooper, D. N., Sandi, C., Soares, M. J., Ferreira, P. G., Afonso, S., Carneiro, M., Turner, J. M. A., VandeBerg, J. L., Fallahshahroudi, A., Jensen, P., Behr, R., Lisgo, S., Lindsay, S., Khaitovich, P., Huber, W., Baker, J., Anders, S., Zhang, Y. E., and Kaessmann, H. 2019. Gene expression across mammalian organ development. *Nature*, 571(7766): 505–509.

Charif, D. and Lobry, J. R. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, editors, *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer, Berlin, Heidelberg.

Chen, L., Bush, S. J., Tovar-Corona, J. M., Castillo-Morales, A., and Urrutia, A. O. 2014. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Molecular Biology and Evolution*, 31(6): 1402–1413.

Dutheil, J. and Boussau, B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evolutionary Biology*, 8(1): 255.

Dutheil, J. Y., Galtier, N., Romiguier, J., Douzery, E. J. P., Ranwez, V., and Boussau, B. 2012. Efficient selection of branch-specific models of sequence evolution. *Molecular Biology and Evolution*, 29(7): 1861–1874.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. 2016. Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Molecular Biology and Evolution*, 33(6): 1517–1527.

Freckleton, R., Harvey, P., and Pagel, M. 2002. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. *The American naturalist*, 160: 712–26.

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14(7): 1–11. Number: 7 Publisher: BioMed Central.

Gout, J.-F., Thomas, W. K., Smith, Z., Okamoto, K., and Lynch, M. 2013. Large-scale detection of in vivo transcription errors. *Proceedings of the National Academy of Sciences*, 110(46): 18584–18589. Publisher: Proceedings of the National Academy of Sciences.

Graveley, B. R. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2): 100–107.

Guéguen, L. and Duret, L. 2018. Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition. *Molecular Biology and Evolution*, 35(3): 734–742.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C., Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K., and Dutheil, J. Y. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Molecular Biology and Evolution*, 30(8): 1745–1750.

Hamid, F. M. and Makeyev, E. V. 2014. Emerging functions of alternative splicing coupled with nonsense-mediated decay. *Biochemical Society Transactions*, 42(4): 1168–1173.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D., and Kent, W. J. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(Database issue): D590–D598.

Hsu, S.-N. and Hertel, K. J. 2009. Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA biology*, 6(5): 526–530.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A. M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R. F., Magwire, M. M., Blankenburg, K., Carbone, M. A., Chang, K., Ellis, L. L., Fernandez, S., Han, Y., Highnam, G., Hjelmen, C. E., Jack, J. R., Javaid, M., Jayaseelan, J., Kalra, D., Lee, S., Lewis, L., Munidasa, M., Ongeri, F., Patel, S., Perales, L., Perez, A., Pu, L., Rollmann, S. M., Ruth, R., Saada, N., Warner, C., Williams, A., Wu, Y.-Q., Yamamoto, A., Zhang, Y., Zhu, Y., Anholt, R. R. H., Korbel, J. O., Mittelman, D., Muzny, D. M., Gibbs, R. A., Barbadilla, A., Johnston, J. S., Stone, E. A., Richards, S., Deplancke, B., and Mackay, T. F. C. 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Research*, 24(7): 1193–1208. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

John, S., Olas, J. J., and Mueller-Roeber, B. 2021. Regulation of alternative splicing in response to temperature variation in plants. *Journal of Experimental Botany*, 72(18): 6150–6163.

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8): 907–915. Number: 8 Publisher: Nature Publishing Group.

Kimura, M., Maruyama, T., and Crow, J. F. 1963. The Mutation Load in Small Populations. *Genetics*, 48(10): 1303–1312.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21): 4453–4455.

Kryazhimskiy, S. and Plotkin, J. B. 2008. The Population Genetics of dN/dS. *PLoS Genetics*, 4(12).

Leinonen, R., Sugawara, H., and Shumway, M. 2011. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database issue): D19–D21.

Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., Dempster, E. L., Bray, N. J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D. A., Jeffery, E. D., Prabhakar, S., Schalkwyk, L., Jops, C., Gandal, M. J., Sheynkman, G. M., Hannon, E., and Mill, J. 2021. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports*, 37(7): 110022.

Li, W. and Lynch, M. 2020. Universally high transcript error rates in bacteria. *eLife*, 9: e54898. Publisher: eLife Sciences Publications, Ltd.

Liu, Z. and Zhang, J. 2018a. Human C-to-U Coding RNA Editing Is Largely Nonadaptive. *Molecular Biology and Evolution*, 35(4): 963–969.

Liu, Z. and Zhang, J. 2018b. Most m6A RNA Modifications in Protein-Coding Regions Are Evolutionarily Unconserved and Likely Nonfunctional. *Molecular Biology and Evolution*, 35(3): 666–675.

Lynch, M. 2006. The Origins of Eukaryotic Gene Structure. *Molecular Biology and Evolution*, 23(2): 450–468.

Lynch, M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences*, 104(suppl_1): 8597–8604. Publisher: Proceedings of the National Academy of Sciences.

Lynch, M. and Conery, J. S. 2003. The origins of genome complexity. *Science (New York, N.Y.)*, 302(5649): 1401–1404.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., and Foster, P. L. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11): 704–714. Number: 11 Publisher: Nature Publishing Group.

Löytynoja, A. and Goldman, N. 2008. Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science*, 320(5883): 1632–1635. Publisher: American Association for the Advancement of Science.

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J. C., Jhangiani, S. N., Jordan, K. W., Lara, F., Lawrence, F., Lee, S. L., Librado, P., Linheiro, R. S., Lyman, R. F., Mackey, A. J., Munidasa, M., Muzny, D. M., Nazareth, L., Newsham, I., Perales, L., Pu, L.-L., Qu, C., Ràmia, M., Reid, J. G., Rollmann, S. M., Rozas, J., Saada, N., Turlapati, L., Worley, K. C., Wu, Y.-Q., Yamamoto, A., Zhu, Y., Bergman, C. M., Thornton, K. R., Mittelman, D., and Gibbs, R. A. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384): 173–178. Number: 7384 Publisher: Nature Publishing Group.

Mazin, P. V., Khaitovich, P., Cardoso-Moreira, M., and Kaessmann, H. 2021. Alternative splicing during mammalian organ development. *Nature Genetics*, 53(6): 925–934. Number: 6 Publisher: Nature Publishing Group.

McGlincy, N. J. and Smith, C. W. J. 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences*, 33(8): 385–393.

Merkin, J., Russell, C., Chen, P., and Burge, C. B. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science (New York, N.Y.)*, 338(6114): 1593–1599.

Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., Reymond, A., Guigó, R., Hubbard, T., and Harrow, J. 2011. The Origins, Evolution, and Functional Potential of Alternative Splicing in Vertebrates. *Molecular Biology and Evolution*, 28(10): 2949–2959.

NCBI Resource Coordinators 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1): D8–D13.

Ohta, T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246(5428): 96–98. Number: 5428 Publisher: Nature Publishing Group.

Pickrell, J. K., Pai, A. A., Gilad, Y., and Pritchard, J. K. 2010. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLOS Genetics*, 6(12): e1001236. Publisher: Public Library of Science.

Rajon, E. and Masel, J. 2011. Evolution of molecular error rates and the consequences for evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 108(3): 1082–1087.

Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., and Huber, W. 2013. Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences*, 110(38): 15377–15382. Publisher: Proceedings of the National Academy of Sciences.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17): 2325–2329.

Saudemont, B., Popa, A., Parmley, J. L., Rocher, V., Blugeon, C., Necsulea, A., Meyer, E., and Duret, L. 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biology*, 18.

Seppey, M., Manni, M., and Zdobnov, E. M. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology (Clifton, N.J.)*, 1962: 227–245.

Singh, P. and Ahi, E. P. 2022. The importance of alternative splicing in adaptive evolution. *Molecular Ecology*, 31(7): 1928–1938. Publisher: John Wiley & Sons, Ltd.

Tomso, D. J. and Bell, D. A. 2003. Sequence Context at Human Single Nucleotide Polymorphisms: Overrepresentation of CpG Dinucleotide at Polymorphic Sites and Suppression of Variation in CpG Islands. *Journal of Molecular Biology*, 327(2): 303–308.

Traverse, C. C. and Ochman, H. 2016. From the Cover: Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12): 3311. Publisher: National Academy of Sciences.

Tress, M. L., Abascal, F., and Valencia, A. 2017a. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in Biochemical Sciences*, 42(2): 98–110.

Tress, M. L., Abascal, F., and Valencia, A. 2017b. Most Alternative Isoforms Are Not Functionally Important. *Trends in biochemical sciences*, 42(6): 408–410.

Verta, J.-P. and Jacobs, A. 2022. The role of alternative splicing in adaptation and evolution. *Trends in Ecology & Evolution*, 37(4): 299–308.

Waples, R. S. 2016. Life-history traits and effective population size in species with overlapping generations revisited: the importance of adult mortality. *Heredity*, 117(4): 241–250.

Weyna, A. and Romiguier, J. 2020. Relaxation of purifying selection suggests low effective population size in eusocial Hymenoptera and solitary pollinating bees. *bioRxiv*, page 2020.04.14.038893. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Wright, C. J., Smith, C. W. J., and Jiggins, C. D. 2022. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics*, 23(11): 697–710. Number: 11 Publisher: Nature Publishing Group.

Xiong, K., McEntee, J. P., Porfirio, D. J., and Masel, J. 2017. Drift Barriers to Quality Control When Genes Are Expressed at Different Levels. *Genetics*, 205(1): 397–407.

Xu, C. and Zhang, J. 2018. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. *Cell systems*, 6(6): 734–742.e4.

Xu, C. and Zhang, J. 2020. A different perspective on alternative cleavage and polyadenylation. *Nature Reviews Genetics*, 21(1): 63–63. Number: 1 Publisher: Nature Publishing Group.

Xu, C., Park, J.-K., and Zhang, J. 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biology*, 17(3): e3000197.

Xu, G. and Zhang, J. 2014. Human coding RNA editing is generally nonadaptive. *Proceedings of the National Academy of Sciences*, 111(10): 3769–3774. Publisher: Proceedings of the National Academy of Sciences.

Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46(4): 409–418.

Zhang, J. and Xu, C. 2022. Gene product diversity: adaptive or not? *Trends in Genetics*, 38(11): 1112–1122.

Supplementary Table 1: Description of the main features of the samples analyzed in this study.

	Clade	Number of RNA-seq samples	Sequencing depth (*per*-base read)[a]	Number of annotated introns	Number of analyzable introns[b]	Average number of introns *per* BUSCO gene	Fraction of major introns alternatively spliced[c]	Average AS rate among BUSCO introns	Fraction of rare SVs[d]
Vertebrates									
Callorhinchus milii	Chondrichthyes	11	1068	7700	7467	8.0	0.491	1.47 %	0.831
Gallus gallus	Aves	217	9657	8741	8621	8.4	0.854	1.59 %	0.958
Crocodylus porosus	Crocodylia	12	1819	7867	7668	8.5	0.817	3.02 %	0.908
Monodelphis domestica	Mammalia	269	11371	8538	8407	8.5	0.915	1.91 %	0.957
Heterocephalus glaber	Mammalia	54	2072	9409	9324	8.6	0.803	2.69 %	0.914
Macaca mulatta	Mammalia	177	5571	9328	9261	8.6	0.908	2.84 %	0.948
Oryctolagus cuniculus	Mammalia	338	15503	8036	7885	8.4	0.950	1.97 %	0.969
Rattus norvegicus	Mammalia	362	16611	8469	8196	8.5	0.953	1.89 %	0.965
Mus musculus	Mammalia	317	12245	9327	9080	8.4	0.937	1.87 %	0.958
Bos taurus	Mammalia	26	710	9046	8926	8.5	0.511	1.63 %	0.856
Loxodonta africana	Mammalia	23	3667	9000	8652	8.3	0.896	3.55 %	0.938
Sus scrofa	Mammalia	55	910	8982	8798	8.5	0.644	1.95 %	0.886
Canis lupus	Mammalia	5	348	9279	8628	8.2	0.436	2.18 %	0.764
Homo sapiens	Mammalia	313	10269	11122	10981	8.4	0.957	3.38 %	0.949
Equus caballus	Mammalia	19	998	9190	9072	8.5	0.658	2.16 %	0.884
Insects									
Bombyx mori	Lepidoptera	14	459	5001	4681	5.3	0.393	1.12 %	0.835
Athalia rosae	Hymenoptera	6	359	4772	4701	4.8	0.348	1.6 %	0.782
Cephus cinctus	Hymenoptera	17	2566	5035	5016	4.7	0.744	2.4 %	0.907
Orussus abietinus	Hymenoptera	2	197	4801	4664	4.7	0.370	2.03 %	0.763
Nasonia vitripennis	Hymenoptera	114	4871	4273	4158	4.5	0.648	1.21 %	0.913
Trichogramma pretiosum	Hymenoptera	4	350	3794	3734	4.4	0.268	0.98 %	0.782
Harpegnathos saltator	Hymenoptera	166	1888	4745	4711	4.7	0.565	2.02 %	0.886
Linepithema humile	Hymenoptera	23	1476	4726	4615	4.8	0.570	1.45 %	0.882
Camponotus floridanus	Hymenoptera	37	449	4596	4546	4.7	0.358	1.52 %	0.761
Pogonomyrmex barbatus	Hymenoptera	39	1388	4678	4440	4.5	0.579	1.91 %	0.866
Polistes canadensis	Hymenoptera	14	440	4665	4562	4.8	0.424	1.88 %	0.834
Polistes dominula	Hymenoptera	12	218	4698	4161	4.3	0.180	1.63 %	0.624
Solenopsis invicta	Hymenoptera	23	436	4516	4394	4.6	0.430	1.71 %	0.807
Acromyrmex echinatior	Hymenoptera	42	1470	4716	4638	4.7	0.529	2.15 %	0.835
Megachile rotundata	Hymenoptera	108	3400	5120	5086	4.8	0.898	3.81 %	0.927
Apis mellifera	Hymenoptera	40	1777	4939	4897	4.9	0.673	2.3 %	0.892
Apis florea	Hymenoptera	4	503	4881	4332	4.4	0.318	1.85 %	0.711
Apis cerana	Hymenoptera	12	1401	4508	4439	4.6	0.578	2.36 %	0.839
Bombus terrestris	Hymenoptera	33	2648	4857	4683	4.7	0.763	2.33 %	0.922
Acyrthosiphon pisum	Hemiptera	35	3163	4918	4844	6.0	0.709	1.09 %	0.933
Cimex lectularius	Hemiptera	10	462	5640	5588	6.3	0.431	1.61 %	0.838
Halyomorpha halys	Hemiptera	6	1460	5715	5676	6.5	0.591	1.73 %	0.885
Aedes aegypti	Diptera	27	2469	2369	2290	2.6	0.514	1.35 %	0.870
Drosophila grimshawi	Diptera	30	256	2190	2032	2.7	0.168	0.8 %	0.726
Drosophila pseudoobscura	Diptera	32	3628	2312	2244	2.6	0.433	1.32 %	0.871
Drosophila melanogaster	Diptera	129	4542	2414	2390	2.7	0.551	1.22 %	0.909
Drosophila suzukii	Diptera	23	1979	2187	2052	2.6	0.287	1.17 %	0.810
Ceratitis capitata	Diptera	29	1168	3067	3015	3.3	0.418	1.45 %	0.860
Lucilia cuprina	Diptera	23	2446	2566	2405	2.8	0.268	0.85 %	0.823
Musca domestica	Diptera	12	1056	2545	2401	2.9	0.254	0.98 %	0.795
Onthophagus taurus	Coleoptera	53	644	2836	2753	3.2	0.377	1.34 %	0.810
Tribolium castaneum	Coleoptera	14	2618	3333	3225	3.6	0.556	1.15 %	0.881
Dendroctonus ponderosae	Coleoptera	30	2262	4370	4269	4.9	0.505	1.26 %	0.882
Anoplophora glabripennis	Coleoptera	20	325	3764	3567	4.1	0.299	1.13 %	0.781
Leptinotarsa decemlineata	Coleoptera	21	2071	3372	3132	3.8	0.512	1.21 %	0.883
Blattella germanica	Blattodea	30	943	4911	4454	5.4	0.423	1.26 %	0.827
Cryptotermes secundus	Blattodea	11	481	6471	6391	6.4	0.573	2.32 %	0.832
Zootermopsis nevadensis	Blattodea	53	3944	6727	6613	6.4	0.802	2.36 %	0.927

[a] Median *per*-base read coverage computed on BUSCO gene exons

[b] Number of analyzable introns (*i.e.* with $N_s + N_a \geq 10$) among BUSCO genes

[c] Proportion of major introns for which alternative splicing has been detected (*i.e.* with $N_a > 0$) among BUSCO genes

[d] Fraction of rare spliced variants introns (*i.e.* with MIRA $\leq 5\%$) among all protein-coding genes

Table S1

Supplementary Table 2: Longevity and body lenth across the 53 metazoans studied.

	Clade	Longevity (Days)	Body length (cm)
Vertebrates			
Callorhinchus milii	Chondrichthyes	2190	120.00
Gallus gallus	Aves	10950	70.00
Crocodylus porosus	Crocodylia	20805	600.00
Homo sapiens	Mammalia	36500	175.00
Loxodonta africana	Mammalia	23725	400.00
Equus caballus	Mammalia	20805	280.00
Macaca mulatta	Mammalia	14600	64.00
Heterocephalus glaber	Mammalia	10950	16.50
Sus scrofa	Mammalia	9855	240.00
Canis lupus	Mammalia	7519	117.00
Bos taurus	Mammalia	7300	245.00
Oryctolagus cuniculus	Mammalia	3285	50.00
Monodelphis domestica	Mammalia	1862	20.00
Mus musculus	Mammalia	1460	9.50
Rattus norvegicus	Mammalia	1387	40.00
Insects			
Bombyx mori	Lepidoptera	50	1.90
Pogonomyrmex barbatus	Hymenoptera	10220	1.10
Acromyrmex echinatior	Hymenoptera	5475	1.40
Camponotus floridanus	Hymenoptera	3650	1.90
Solenopsis invicta	Hymenoptera	2482	0.70
Apis mellifera	Hymenoptera	1095	2.00
Apis florea	Hymenoptera	1095	2.00
Apis cerana	Hymenoptera	1095	2.00
Harpegnathos saltator	Hymenoptera	653	1.70
Polistes canadensis	Hymenoptera	506	2.00
Polistes dominula	Hymenoptera	506	2.00
Linepithema humile	Hymenoptera	365	0.50
Bombus terrestris	Hymenoptera	150	2.50
Megachile rotundata	Hymenoptera	56	1.90
Nasonia vitripennis	Hymenoptera	25	0.30
Athalia rosae	Hymenoptera	12	0.73
Trichogramma pretiosum	Hymenoptera	10	0.04
Cephus cinctus	Hymenoptera	7	0.86
Orussus abietinus	Hymenoptera	7	1.00
Cimex lectularius	Hemiptera	572	0.50
Halyomorpha halys	Hemiptera	112	1.44
Acyrthosiphon pisum	Hemiptera	30	0.25
Drosophila pseudoobscura	Diptera	90	0.20
Musca domestica	Diptera	60	0.70
Drosophila grimshawi	Diptera	50	0.50
Ceratitis capitata	Diptera	50	0.50
Drosophila suzukii	Diptera	38	0.33
Drosophila melanogaster	Diptera	36	0.30
Lucilia cuprina	Diptera	21	0.80
Aedes aegypti	Diptera	14	0.38
Leptinotarsa decemlineata	Coleoptera	365	1.00
Tribolium castaneum	Coleoptera	170	0.50
Onthophagus taurus	Coleoptera	160	1.00
Anoplophora glabripennis	Coleoptera	66	3.50
Dendroctonus ponderosae	Coleoptera	30	0.75
Cryptotermes secundus	Blattodea	4745	0.60
Zootermopsis nevadensis	Blattodea	2300	1.00
Blattella germanica	Blattodea	200	1.59

[*] The sources from which the lifespan and the body length information was taken are listed in Data9supp.pdf in the Zenodo data repository (see Data and code availability).

Table S2