HAL
open science

Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs

Diego Javier Zea, Sofya Laskina, Alexis Baudin, Hugues Richard, Elodie Laine

To cite this version:

Diego Javier Zea, Sofya Laskina, Alexis Baudin, Hugues Richard, Elodie Laine. Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs. Genome Research, inPress. hal-03274456

HAL Id: hal-03274456

https://hal.sorbonne-universite.fr/hal-03274456v1

Submitted on 30 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs

Diego Javier Zea[1], Sofya Laskina[2], Alexis Baudin[3], Hugues Richard[1,2*], Elodie Laine[1*]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.
[2] Bioinformatics Unit (MF1), Department for Methods development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany
[3] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France.
[*] corresponding authors: RichardH@rki.de, elodie.laine@sorbonne-universite.fr

Keywords: alternative splicing, evolutionary conservation, transcriptome, multiple sequence alignment, splicing graph, function diversification, protein repeat

Abstract

Understanding how protein function has evolved and diversified is of great importance for human genetics and medicine. Here, we tackle the problem of describing the whole transcript variability observed in several species by generalising the definition of splicing graph. We provide a practical solution to construct parsimonious *evolutionary* splicing graphs where each node is a minimal transcript building block defined across species. We show a clear link between the functional relevance, tissue-regulation and conservation of alternative transcripts on a set of 50 genes. By scaling up to the whole human protein-coding genome, we identify a few thousands of genes where alternative splicing modulates the number and composition of pseudo-repeats. We have implemented our approach in ThorAxe, an efficient, versatile, robust and freely available computational tool.

INTRODUCTION

Eukaryotes have evolved a transcriptional mechanism that can augment the protein repertoire without increasing genome size. A gene can be transcribed, spliced, and matured into several transcripts by choosing different initiation/termination sites or by selecting different exons (Graveley, 2001). Alternative splicing, as well as alternative promoter usage or alternative polyadenylation (hereafter referred as AS), concerns almost all multi-exon genes in vertebrates (Wang et al., 2008), and many organ-specific splicing patterns have diverged rapidly during vertebrate evolution (Barbosa-Morais et al., 2012; Merkin et al., 2012). This mechanism can affect transcript maturation and post-transcriptional regulation, or result in protein isoforms ("proteoforms") with different shapes (Birzele et al., 2008), interaction partners (Yang et al., 2016), and functions (Baralle and Giudice, 2017; Kelemen et al., 2013). Its misregulation is associated with the development of cancer, among other diseases (Climente-González et al., 2017; Scotti and Swanson, 2016; Lim et al., 2011; Ward and Cooper, 2010; Wang and Cooper, 2007). Moreover, the influence of natural isoforms variations between human populations on disease susceptibility is increasingly recognised (Park et al., 2018). Hence, understanding how AS contributes to protein function diversification is of utmost importance for human genetics and medicine.

In recent years, the advent of high-throughput sequencing technologies like RNA-seq has made possible in-depth surveys of transcriptome complexity (Wang et al., 2008; Sultan et al., 2008). However, evaluating how many of the detected transcripts are translated and functional in the cell remains challenging (Wang et al., 2018; Kim et al., 2014). This difficulty has stimulated the development of integrative approaches combining gene annotations, RNA-seq data and also data generated by other high-throughput techniques (Marti-Solano et al., 2020; de la Fuente et al., 2020; Louadi et al., 2020; Ait-hamlat et al., 2020; Agosto et al., 2019; Sterne-Weiler et al., 2018; Denti et al., 2018; Tapial et al., 2017; Tranchevent et al., 2017; Weatheritt et al., 2016; Ezkurdia et al., 2015; Rodriguez et al., 2013; Gonzàlez-Porta et al., 2013; De La Grange et al., 2010) toward a better characterisation of the AS landscape. Recent studies underscore AS functional impact and contribution to protein diversity (Marti-Solano et al., 2020; Agosto et al., 2019).

Evolutionary conservation can arguably serve as a reliable indicator of function. Indeed, we expect splice variants selected over millions of years of evolution to comply with physical and environmental constraints and thus to play a functional role. The classical approach for assessing AS evolutionary conservation first identifies orthologous exons between species and then compares their inclusion/exclusion rates across cell/tissue types. A common practice for orthology detection is the BLAST (Altschul et al., 1990) "all-against-all" methodology (Nichio et al., 2017). When dealing with exons, more specialised

protocols based on pairwise genomic sequence alignments (Mei et al., 2017; Herrero et al., 2016; Abascal et al., 2015; Modrek and Lee, 2003; Xing and Lee, 2005; Barbosa-Morais et al., 2012), or multiple alignments of genomic or protein sequences (Szalkowski, 2012; Christinat and Moret, 2012; Merkin et al., 2012) have been proposed. Challenges associated with this task include correctly handling large indel events, finding plausible matches for highly divergent sequences, and resolving ambiguities arising from highly similar sequences (*e.g.*, resulting from in-gene duplication) or very short sequences. The alternative usage of orthologous exons may then be investigated using compact representations of transcript variability, such as splicing graphs (Heber et al., 2002), where the nodes represent the exons, and the edges denote exon junctions. Hence, a way to assess AS conservation between two genes would be to compare the environment of their orthologous exons in the two corresponding splicing graphs. However, until now, there exists no method combining these two layers of information.

In this work, we addressed the problem of exon orthology detection in the context of AS. Our specific aims were to develop a novel and general method revisiting splicing graph representation to account for the whole transcript variability observed in a set of species, and to apply this method at the protein-coding genome scale to provide, for the first time, granular estimates of AS evolutionary conservation and significantly improve our knowledge on the amount of variations that are functionally relevant.

RESULTS

Evolution-informed model describes transcript variability

Our method maps all transcriptomic information coming from many species on an *evolutionary splicing graph*, where the nodes represent minimal transcript building blocks defined across species (**Fig. 1**). Classically, a splicing graph (SG) contains nodes representing exons, *i.e.* genomic intervals, and edges indicating co-occurences of contiguous exons in a set of transcripts observed for a gene. In the present work, we use a slightly different definition, where the nodes are the genomic intervals supplemented by their reading frames (**Supplemental Methods**). In practice, we work with the corresponding translated amino acid sequences (**Fig. 1A**). Moreover, the nodes may actually represent *sub-exons*, since donor or acceptor sites can be located inside an exon (**Fig. 1A**, n_1 and n_2). We distinguish the edges *induced* by the sub-exon boundaries (**Fig. 1A**, $n_1 \rightarrow n_2$) from the *structural* edges arising from intron boundaries (**Fig. 1A**, $n_1 \rightarrow n_3$).

Our main contribution is to extend the definition of a splicing graph to a set of orthologous genes G. We describe the whole transcript variability of G with an evolutionary splicing graph (ESG) $\mathcal{S} = (\mathcal{V}, \mathcal{E})$ (**Fig. 1B**), where each node $v \in \mathcal{V}$ is a *spliced-exon* (or *s-exon*) and represents a multiple sequence alignment (MSA) of sub-exons or sub-exon parts coming from different species. The edge set $\mathcal{E}$ is comprised of the ensemble of edges linking the nodes in the individual SGs, possibly augmented by some edges induced by the definition of the s-exons (see **Supplemental Methods** for more formal definitions). As a consequence, two nodes in $\mathcal{S}$ may be linked by several edges (at most one per species), and we designate this set as a *multi-edge*. There are many possible ways of grouping the exonic sequences coming from the different genes, and hence of defining the s-exons (**Fig. 1B**). For instance, one may define each s-exon in the ESG by taking at most one sub-exon from each SG (**Fig. 1B**, *solutions 1* and *3*). In that case, the edge set is exactly the set of edges coming from the individual SGs. Alternatively, it may be advantageous to split a sub-exon into two or more subsequences and assign them to different s-exons (**Fig.**

1B, *solution 2*, where "AEI" and "GV" come from the human sub-exon "AEIGV"). This strategy can lead to better MSAs but it increases the complexity of the graph. Ideally, one would like to find a representation as compact as possible and at the same time, conveying meaningful evolutionary information. To estimate both properties, we define the score of the ESG $\mathcal{S}$ as

$$\sigma_{ESG}(\mathcal{S}) = \sum_{v \in \mathcal{V}} \sigma(v) - \sum_{e \in \mathcal{E}} (n_e^{\mathrm{I}} \cdot \sigma_I + n_e^{\mathrm{S}} \cdot \sigma_S) \quad (1)$$

where $\sigma(v)$ is the score of the MSA associated to the node v. σ can be for instance a consensus score or a sum-of-pairs score, and may additionally penalise very short MSAs (< 3 columns). n_e^{I} (resp n_e^{S}) are the numbers of induced (resp. structural) edges in the multi-edge e with associated fixed penalty σ_I (resp. σ_S). In practice, we set $\sigma_I \gg \sigma_S$ to avoid small s-exons induced by ambiguous alignment columns in the MSAs. As an example, with a simple sum-of-pair scoring for σ and an induced edge penalty $\sigma_I = 4 \cdot \sigma_S$, the best-scored ESG in Figure 1B (*solution 1*) comprises the smallest numbers of s-exons, induced edges and gaps. In general, determining the best-scored ESG is a NP-hard problem (see *Methods*).

Here, we provide a practical solution to construct a meaningful parsimonious ESG, given a set of input transcripts (**Fig. 1C** and **Supplemental Fig. S1**). Our heuristic procedure first pre-clusters exons using pairwise alignments. Then, within each cluster, it concatenates the sequences coming from each species in the order of their genomic coordinates, and aligns the obtained sequences using ProGraphMSA (Szalkowski, 2012) (**Supplemental Fig. S2**). The latter allows better handling of AS-induced deletions and insertions than classical progressive alignment methods. Moreover, using the genomic coordinates as ordering contraints helps to disentangle orthology from paralogy relationships between similar sequences (**Supplemental Fig. S2C**). Finally, we locally solve the problem exposed in Eq. 1 by re-aligning some sequences and by maximising the agreement between sub-exon boundaries across different species (**Supplemental Fig. S3**). Controlling the creation of (penalising) induced edges allows us to implicitly evaluate the ESG score. We implemented the heuristic in the fully automated tool ThorAxe.

In the following, we show that ThorAxe allows obtaining simple and meaningful representations for evolution in the context of AS. We primarily rely on gene annotations from Ensembl (Yates et al., 2016), and we complement the computed ESG with RNA-seq data from the Sequence Read Archive (SRA) (Leinonen et al., 2010), together with the tissue annotations compiled from Bgee (Komljenovic et al., 2018) (see **Supplemental Methods**). We focus on one-to-one orthologous genes across 12 species, namely three primates, two rodents, four other mammals, one amphibian, one fish and nematode. Our motivation for this choice was to span different evolutionary distances and to ensure that enough RNA-seq data would be available. We take human as reference for selecting the genes, but ThorAxe ESG construction is reference-free.

ThorAxe recapitulates known functional AS events

We tested ThorAxe on a curated set of 50 genes representing 16 families (**Supplemental Table S1** and **Supplemental Methods**), where several splice variants have been associated with diverse protein functions. ThorAxe detected 448 alternative splicing, initiation and termination events. RNA-seq splice junctions mapping onto the ESGs provided additional support for about one quarter of them and uncovered 101 more events. Detailed information is available on the accompanying website, at `http://www.lcqb.upmc.fr/ThorAxe` (**Supplemental Fig. S4**). We report here the results for a set of 30 documented

events influencing partner binding affinity, selectivity or specificity (**Supplemental Table S2-3**). We observed tissue-regulation patterns well-conserved across mammals for most of them, and as far as amphibians for 7 of them (**Fig. 2A**, see also **Supplemental Table S3**). While the gene annotations and the RNA-seq data show a good overall agreement, a large number of subpaths are contributed solely by RNA-seq in platypus, cow and zebrafish (**Fig. 2B**).

The ESG computed for *CAMK2B* linker region provides an illustrative example where, despite a very high AS-generated complexity, ThorAxe results are interpretable, meaningful and consistent with what has been reported in the literature (**Fig. 3A**). For instance, one can readily see that the shortest isoform lacking the linker (labelled *7* on **Fig. 3B**) has low evolutionary support. This is in line with recent findings emphasising the importance of the linker for regulating the protein activity (Bhattacharyya et al., 2020). Moreover, all the s-exons defined by ThorAxe are conserved at least as far as amphibians. The smallest s-exon (*25_1*) contains only one column of alanines, and corresponds to a well-documented internal splice site (Sloutsky and Stratton, 2020). Finally, the two documented functional AS events are clearly identifiable on the ESG (**Fig. 3A**, grey areas). This observation still holds true when removing the two best-annotated species, namely human and mouse (**Supplemental Fig. S5A**), and when scaling up to about 100 species (**Supplemental Fig. S5B**). Furthermore, RNA-seq mapping revealed evolutionary conserved tissue regulation for both events (**Fig. 3C**). For instance, the alternatively spliced F-actin binding region comprised of the s-exons *15_0* and *15_1* is specifically expressed in the brain and muscles of primates and rodents (**Fig. 3C**, on the left).

ThorAxe summarises within and across-species variations at the human proteome scale

We further assessed ThorAxe on the whole human proteome (18 226 genes, see **Supplemental Methods**). ThorAxe analysis across 12 species completed in less than 20 hours with 15 cores. The genes are well represented in all mammals (**Supplemental Fig. S6**), except in platypus which covers only 40% of the human protein coding genome. Frog, zebrafish and nematode cover about 65, 50 and 14% of the genes, respectively (**Supplemental Fig. S6**). ThorAxe produced ESGs with 26 s-exons, on average, and at most 354 (**Supplemental Table S4**). They are either very lowly or very highly conserved, as measured by the *species fraction* that is the proportion of species where a s-exon is found (**Supplemental Fig. S7-8**).

We distinguish the *species-specific* s-exons detected in only one species and thus containing only one sequence in their MSA, from the s-exons *conserved* in at least two species. The proportion of species-specific s-exons goes from less than 10% in mammals and zebrafish to 23% in frog and 72% in nematode (**Fig. 4A** and **Supplemental Table S5**), emphasizing the high sequence divergence of this organism. These s-exons are often located at the transcript extremities (**Fig. 5A-B**, see nodes in yellow), and tend to be smaller than the conserved ones (**Supplemental Fig. S9**). The vast majority of the latter are well conserved from primates to amphibians, and their species representativity strongly correlates with the evolutionary distance they span (**Fig. 4A**). For instance, almost all the conserved s-exons present in frog also comprise sequences coming from primates, non-primate eutherians and non-eutherian mammals (**Fig. 4A**, pink curve). This correlation is even more evident in the subset of 13 558 genes with one-to-one orthologs in more than 7 species (**Supplemental Fig. S10-11**). Overall, about 40% of the conserved s-exons present in human span an evolutionary time of more than 400 millions year, up to

zebrafish (**Fig. 4A**). The proportion drops down to less than 10% outside of vertebrates. Nevertheless, we identified 295 genes where ThorAxe assigned most of the exonic sequences contributed by nematode to well-conserved s-exons. This set is enriched in genes coding for proteins involved in the transcription or the translation (RNA polymerases, ribosome, spliceosome, chaperones) or in protein degradation (proteasome).

As for the s-exon usage, almost a third is involved in some event (**Fig. 4B**). The most (resp. least) conserved ones, are involved in deletions (resp. insertions) (see green and pink curves). This observation can be explained by the fact that ThorAxe detects events as variations from a reference *canonical* transcript chosen for its high conservation and length (see *Supplemental Methods*). The alternatively expressed s-exons located at the beginning or in the middle of the protein (gold and light blue curves) tend to be more conserved than at the end of the protein (dark blue curve).

The s-exons accurately measure sequence conservation

To evaluate ThorAxe ability to correctly match exonic sequences between more or less distant species, we compared the s-exon *species fractions* (*SF*) with estimates of evolutionary conservation deduced from whole-genome alignments between human and 99 other vertebrates, available as phastCons scores through the UCSC Genome Browser (Siepel and Haussler, 2005; Siepel et al., 2005) (see **Supplemental Methods**). Overall, the two measures agree very well (**Fig. 4C**). For instance, most of ThorAxe species-specific or very lowly conserved s-exons ($SF<0.3$) are not expected to be evolutionary conserved based on genomic alignments (**Fig. 4C**, left column). Nevertheless, ThorAxe seems to underestimate the conservation level of 1 508 s-exons (**Fig. 4C**, top left corner). We investigated whether these s-exons could share significant sequence similarity with some other s-exons defined across distinct species, and we found that only 24 of them may be considered as "false negatives" (**Supplemental Table S6**). Hence, the low conservation estimated by ThorAxe likely reflects the lack of annotated transcripts in certain species rather than errors in the heuristic. Reciprocally, ThorAxe seems to detect more conservation signal than whole-genome alignments for more than 7 000 s-exons (**Fig. 4C**, bottom row, $SF\geq0.3$), without any particular trend in their alternative usage (**Supplemental Fig. S12**). They display high sequence identity (**Supplemental Fig. S12C-D**), suggesting that they are indeed conserved across many species and not "false positives". The collagen type XVIII alpha 1 chain (*COL18A1*) on its own contributes 12 such s-exons, conserved from primates to amphibians according to ThorAxe ($SF>0.8$) but with very low phastCons scores (<0.1). The COL18A1 protein is highly enriched in glycines and prolines and the 12 s-exons fall within regions of low sequence complexity (**Supplemental Fig. S13**). We can hypothesize that this low-complexity context confounds the whole-genome alignments but not ThorAxe heuristic. To get a better view on the s-exon sequence divergence, we computed the sum-of-pair scores of the associated MSAs (see **Supplemental Methods**). Overall, almost half of the conserved s-exons have very high quality MSAs with very few mismatches and gaps (**Fig. 4D**, score > 0.75). This proportion increases up to about 70% on the 50-gene set (**Supplemental Fig. S14**). A very small proportion (about 1%) of s-exons have very poor quality MSAs (**Fig. 4D**, in black), and those are typically short (**Supplemental Fig. S15**). Moreover, the inserted and, to a lesser extent, alternatively expressed s-exons display lower-quality MSAs (**Fig. 4D**, see *Insert* and *Alter-I*). Finally, we checked the relationship between structural order/disorder and s-exon sequence divergence. Structurally disordered s-exons (about 40% of the ensemble) tend to be less conserved and to have lower quality MSAs than well-folded ones (**Supplemental Fig.**

S16). However, the differences between the two groups are rather small.

The comparison of similar s-exons unveils functionally relevant signatures

ThorAxe allows exploring how function diversification may arise through the alternative usage of similar sequences within and across genes. We illustrate the power of the approach on three gene families (**Fig. 5**, and see also **Supplemental Table S2-3**), focusing on a set of events involving two or more highly conserved s-exons with similar consensus sequences. The origin of the events can be traced back to the ancestor common to mammals, amphibians and fishes. The first example is given by the tropomyosin family (**Fig. 5A,C**), whose protein members (TPM1,2,3,4) serve as integral components of the actin filaments forming the cell cytoskeleton. Several conserved events detected in the ESGs have direct implications for actin binding (Pathan-Chhatbar et al., 2018; Wu et al., 2010) (**Fig. 5A**). Among them, the internal mutually exclusive pair displays high sequence similarity and strong sequence conservation across species and between paralogous genes (**Fig. 5C**, on top, see α and β groups). Fourteen specificity-determining sites (SDS) can be identified (**Fig. 5C**). SDS are key positions with specific conservation patterns, and they play a role in diversifying protein function in evolution (Chakraborty and Chakrabarti, 2015). Given two groups, here α and β, type I SDS are conserved in one group and variable in the other one, indicating different functional constraints between the groups. For instance, position 24 is occupied by a glutamate in all the s-exons from the β group, while it is variable in the α group. Type II SDS are conserved in both groups but each group displays a different amino acid. This is the case of positions 14 and 15, where the Thr-Asn couple of the α group is replaced by Asp-Gln in the β group. These SDS may be responsible for the differences observed in actin filaments formation, mobility and myosin recruitment ability between the isoforms (Pathan-Chhatbar et al., 2018). The c-Jun N-terminal kinase family (*MAPK8,9,10*) gives another example with even higher sequence identities (**Fig. 5C**, in the middle). Among the eight identified SDSs, three positively charged residues, His, Lys and Arg, in positions 16, 17 and 23 are specifically conserved in the α group, while the β group is characterised by Lys, Gly and Thr in positions 15, 16 and 23. These observations are in line with our previous study highlighting differences in the dynamical behaviour of these residues (Ait-hamlat et al., 2020) and their potential implication for substrate selectivity (Waetzig and Herdegen, 2005). As a third example, myosin IB comprises a set of consecutive similar s-exons overlapping with calmodulin(CALM1)-binding so-called IQ motifs (**Fig. 5B**). The alternative inclusion of two s-exons, which share almost 50% identity (**Fig. 5C**, at the bottom), results in different binding motifs. Compared to the motif's canonical form (IQXXXRGXXXR) (Houdusse et al., 1996; Bähler and Rhoads, 2002), they all lack the glutamine in the IQ residue pair and the arginine in the RG pair. These differences could explain their lower affinity compared to the constitutive s-exons (Greenberg and Ostap, 2013).

Alternative usage of similar sequences is not a rare phenomenon

At the human genome scale, we identified 2 190 genes (12% of the protein-coding genome) with evidence of evolutionary conserved alternative usage of similar exonic sequences (**Fig. 6**). The corresponding proteins tend to be involved in cell organisation and muscle contraction (cytoskeleton, collagen, fibers...etc), and in inter-cellular communication (**Supplemental Fig. S17**). Our strategy here was to look for similar s-exon pairs involved in some event (see **Supplemental Methods**). We found a total of 31 031 pairs, among which 446 are mutually exclusive (**Fig. 6A-B**, *MEX*). This case scenario highlights

the exclusive usage of one or the other version of a protein region. The 232 concerned proteins are enriched in transporters and channels (**Supplemental Fig. S17**). Another 438 pairs (coming from 134 genes) are alternatively used without mutual exclusivity (**Fig. 6A-B**, *ALT*). In about half of the MEX or ALT s-exon pairs, one of the s-exons is conserved in all studied species, and the other one in more than half of them (**Fig. 6C**). In 3 813 pairs, one s-exon is included in the canonical or alternative subpath of an event, while the other one serves as a "canonical anchor" for the event (**Fig. 6A-B**, *REL*). This highlights the AS-induced modulation of the number of non-identical consecutive copies of a protein region. The remaining 26 334 pairs correspond to cases where one of the s-exons participates in an event (on the canonical or alternative subpath), while the other one is located outside the event in the canonical transcript (*UNREL*). The full lists of s-exons are given in Supplemental Tables 7-9. This resource overlaps well with a previously reported manually curated set of 97 human genes with mutually exclusive homologous exon pairs (Abascal et al., 2015). It extends it by one order of magnitude and represents a more diverse range of AS-mediated relationships between similar protein regions. While most of the identified genes contain only one or a few similar s-exons pair(s), almost 50 genes have several hundreds or thousands of pairs (**Fig. 6D**). Nebulin gives the most extreme example, with 2 380 detected pairs. This giant skeletal muscle protein has evolved through several duplications of nebulin domains, and a definition of pertinent nebulin evolutionary units was proposed (Björklund et al., 2010). These units correspond to parts of exons, in line with ThorAxe s-exons MSAs.

Comparison with other studies

We evaluated the ability of ThorAxe to detect the events reported in two reference studies dealing with the evolution of AS (Barbosa-Morais et al., 2012; Merkin et al., 2012). We could map 40 exons from (Barbosa-Morais et al., 2012) and 323 exons from (Merkin et al., 2012), all displaying conserved tissue-specific splicing patterns, onto our ESGs (see **Supplemental Methods**). For more than 75% of the 41 exons, we found events conserved across mammals and ranked first or second in the ESG (**Supplemental Fig. S18A**, in dark green). The conservation signal extends to amphibians for 18 events, and to teleosts for 5. For the 323-exon set, we detected 277 events, 90% of which are in the top 3 most conserved of ThorAxe ESGs. About 70% are well conserved in mammals, 82 as far as amphibians and 19 as far as teleosts (**Supplemental Fig. S18B**). In particular, we can mention exon 3 from the eukaryotic translation elongation factor 1 delta (*EEF1D-ex3*) and exon 20 from the tight junction protein 1 (*TJP1-ex20*) highlighted in Figures 2 and 4 from (Merkin et al., 2012). In both cases, the deletion of the matching s-exon(s) is the most conserved event of the ESG, and is observed in human, mouse and pig. The deletion of *EEF1D-ex3* (s-exons *1_1* and *1_2*) is also conserved in gorilla and cow, while that of *TJP1-ex20* (s-exon *6_0*) is also conserved in macaque. We can also pinpoint the six exons intersecting with our curated set (**Supplemental Fig. S19**). ThorAxe detected events well-conserved across mammals for all of them (and as far as amphibians for two of them), with RNA-seq evidence of conserved tissue-specific AS patterns for all but one. One of the matching s-exons, *11_1* from *MYO1B*, is part of a couple of alternatively spliced pseudo-repeats (**Fig. 5B**). While we found conserved tissue regulation patterns for *11_3*, the other s-exon in the couple, it was not reported in (Merkin et al., 2012). This example highlights the difficulty of assessing the tissue-specific expression of several instances of (pseudo-)repeated sequences, and showcases ThorAxe's ability to deal with such complexity.

Comparison with other methods

We assessed the pertinence of ThorAxe heuristic by performing an ablation study and by comparing it with two popular exon orthology detection methods. For the ablation study, one strategy was to skip the exon clustering step (see *Methods*, *step b*), and the other was to rely solely on global multiple sequence alignment, which means both the exon clustering step and the s-exon refinement step are skipped (see *Methods*, *steps b,e*). Compared to these two strategies, ThorAxe produces longer and higher-quality s-exon MSAs (**Supplemental Fig. S18C-D**). Specifically, the clustering step helps to improve the s-exon sequence identities (**Supplemental Fig. S18D**, compare blue and orange boxes) by reducing the space of sequences to align. The final local optimization step increases the lengths of the s-exons (**Supplemental Fig. S18C**, compare blue and red boxes) by minimising sub-exon boundaries violations. As popular exon orthology detection methods, we chose the Reciprocal Best Blast Hit (RBBH) method and Ensembl Compara (Herrero et al., 2016). The RBBH method consists in finding the best matching sub-exon pairs across any two species using BLAST. One of the drawback of this method is that many sub-exons remain without any match in other species (**Supplemental Fig. S18E**, orange box). By allowing for one-to-many sub-exon matching between species, ThorAxe covers a much higher proportion of sub-exons (**Supplemental Fig. S18E**, blue box). ThorAxe strategy is justified by the fact that exons may undergo truncation or elongation in the course of evolution, and thus we do not expect a one-to-one relationship between them across a pair of species. Moreover, ThorAxe increased sub-exon coverage is not at the expense of sequence identity (**Supplemental Fig. S20**). Another drawback of RBBH is that defining s-exons from a set of pairwise alignments of sub-exons is a difficult task. Finally, Ensembl Compara relies on whole genome alignments. However, it does not include all the species for which annotated transcripts are available in Ensembl. Moreover, one can expect that working with DNA sequences produces lower-quality alignements compared to working with protein sequences, as is done by ThorAxe.

Assessment of the default parameters' choices

All parameters in ThorAxe are customisable by the user, enabling a rapid adaptation of the method to specific contexts and questions. We investigated the pertinence of some of the default values. For instance, by default, ThorAxe filters out the transcripts flagged in Ensembl as lowly supported (Transcript Support Level, $TSL < 3$). This restriction only concern human and mouse, as the other species do not have any TSL annotations. By varying the TSL value between 1 and 5, we observed that, as expected, the less stringent the TSL criterion, the higher the number of transcripts and of events (**Supplemental Fig. S21A-B**). However, very little change is observed in the definition of the canonical transcript and in the conservation levels of the s-exons, suggesting that the results and their interpretation are robust to this parameter (**Supplemental Fig. S21C-E**). Globally, the biggest changes are observed either when only top-quality transcripts are retained ($TSL \leq 1$), or when transcripts are not filtered at all ($TSL \leq 5$). We thus recommend to use intermediate TSL values (2-4).

The first step of ThorAxe algorithm consists in grouping similar exonic sequences together, with the aim of reducing the complexity of the subsequent construction of the MSAs. By default, ThorAxe applies a sequence identity cutoff of 30% to define the clusters. To assess the suitability of this cutoff value for handling divergent sequences, we looked at the MSA quality with respect to the *species fraction* (**Supplemental Fig. S22**). Although the two measures are correlated, a significant portion of s-exons display high

species fractions (> 0.8) but low MSA scores (≤ 0.5). This observation suggests that ThorAxe is able to cluster together divergent sequences that are difficult to align. The cutoff may be adapted by the user depending on the level of sequence divergence expected in the input data.

To ease interpretability of the results and ensure that the s-exons represent groups of one-to-one orthologous exonic regions, ThorAxe default mode considers only one-to-one orthologous genes, as annotated in Ensembl. This means that organisms with additional round(s) of whole genome duplications and/or separated by long evolutionary distances will likely be excluded from the analysis. We found that taking into account many-to-many gene orthology relationships leads to a better detection of the documented events (**Supplemental Fig. S23**). The improvement is particularly visible for zebrafish where we now have 8 events with both the canonical and alternative subpaths supported by Ensembl annotations (compare panels A and B). The detection in rat, cow and platypus is also improved. Finally, we tested the impact of excluding the two best-annotated species, namely human and mouse. As a result of this exclusion, three documented events are lost. Nevertheless, the conservation profiles of the other events remain almost identical (**Supplemental Fig. S23**, compare panels A and C).

DISCUSSION

We have presented a novel method to describe transcript variability in evolution. Our approach provides a double generalization, by extending the definition of SG to the case of multiple species, and by providing a way to combine MSAs over structures with a partial order. The heuristic is general enough to deal with very different genes (in terms of length, structure, degree of conservation, number of transcripts, etc). Its identification of transcript minimal building blocks (the s-exons) is the first and necessary step for inferring evolutionary scenarios explaining AS-induced protein function diversification (Ait-hamlat et al., 2020). Our data structure is reminiscent of current developments on pangenome graphs. However, pangenome approaches keep track of variations across a population, whereas we highlight conservation across species in the context of AS. To the best of our knowledge, we are the first to do it. Effectively, we consider a *pan-transcriptome* across multiple species. As a consequence, we do not need to rely on a central species and project the transcripts on it. In the analysis conducted here, human was taken as a reference only to find orthologous genes in other species.

To illustrate the potential of the method, we assessed the evolutionary conservation and tissue regulation of a set of documented AS events we compiled from the literature. This set could serve as a reference for future studies. We then scaled up to the human protein-coding genome, and found that AS is conserved across a wide range of evolutionary distances, is not limited to ancient events, and does not generate conserved alternative isoforms in all of the proteins. We have shown that the alternative usage of repeats in protein is not a rare phenomenon in the human proteome and that it is of ancient evolutionary origin. Although we focused on one-to-one orthologs, thereby limiting the contribution of nematode, our analysis can be readily extended to one-to-many orthology relationships to better compare vertebrates with other organisms.

On the one hand, a limitation of the approach is that it mainly relies on gene annotations, which may be partial, incomplete, or erroneous (Salzberg, 2019). To avoid errors, we chose to select only high-quality transcripts, with the risk of biasing the results towards species with more fully annotated alternative splicing landscapes. Another strategy could be to use APPRIS annotations, but we expect a reduction in the overall input transcript

variability, therefore limiting ThorAxe potential to discover AS events. Moreover, APPRIS annotations are derived from an analysis accounting for transcript sequence conservation, which would be somewhat redundant with ThorAxe's own analysis of AS conservation. On the other hand, an important advantage of ThorAxe is its robustness with respect to the presence of highly divergent sequences and the creation of species-specific s-exons. Indeed, the latter simply contribute single nodes to the ESG without preventing the detection of conserved AS events. In a way, detecting too many species-specific s-exons would not be a problem as this would only slightly diminish ThorAxe conservation estimates without hampering the interpretation of the ESGs. Traditional methods may recover genomic conservation at lower levels of sequence similarity, but by disregarding the whole transcript structure, they may not properly evaluate AS conservation.

Future work could benefit from the development of more accurate approximations of the general problem stated here. Another direction is to expand the application field to transcriptomes coming from patients or human populations. In the coming years, we expect a tremendous increase in the available transcriptomic data, including transcriptome annotations generated by long-read sequencing technologies (Byrne et al., 2019). Methods addressing the complexity of these data will become instrumental in understanding the evolution of a disease, *e.g.*, cancer, and the phenotypic variability among human populations and individuals (Park et al., 2018; Lonsdale et al., 2013). ThorAxe could be easily adapted to deal with these data, and, along this line, we have already implemented the possibility to give additional "user-defined" transcripts as input.

METHODS

Complexity of the problem: determining a minimal ESG is NP-Hard

To illustrate the complexity of determining a minimal ESG, let us consider a case example with n input transcripts observed in n species (*i.e.*, 1 transcript per species). Moreover, since the problem is theoretically independent of the penalties σ_I and σ_S (Eq. 1), an algorithm that would solve it in the general case would also be valid for $\sigma_I = \sigma_S = +\infty$. In this scenario, a minimal ESG has no edge and maximises the sum-of-pair-score σ. Thus, the problem of building a minimal ESG is equivalent to solving the problem of multiple sequence alignment with sum-of-pair-score σ on the n input transcripts. Since the n input transcripts can be any string (over the amino acid alphabet), and finding a MSA of any string with sum-of-pair-score is NP-hard (Wang and Jiang, 1994), it follows that finding a minimal ESG is NP-hard.

Description of ThorAxe algorithm and parameters

Given a gene name and a list of species as input, ThorAxe extracts and exploits gene annotations from Ensembl (and, optionally, input transcripts provided by the user) to build an ESG maximising the sequence similarity within each node (or s-exon) and minimising the number of induced edges – which indirectly implies that the number of nodes is minimised. The heuristic approximates the best-scored solution of Eq. 1 by controlling the creation of induced edges, without explicitly computing ESG scores. It unfolds in six main steps (**Fig. 1C** and **Supplemental Fig. S1**).

a- Data acquisition and pre-processing. ThorAxe downloads the gene tree, the transcripts annotated as protein coding and their exons (genomic coordinates, sequences and phases) for the query gene and its (by default one-to-one) orthologs in the selected species

from Ensembl. ThorAxe then removes incomplete or lowly supported ($TSL < 3$, adjustable by the user) transcripts, and translates the retained transcripts' DNA sequences into amino acid sequences using the exon phases. Transcripts or exonic regions leading to the same protein sequence are merged, but the same genomic region may lead to the generation of several protein sequences if it is associated with more than one frame (**Supplemental Methods**). ThorAxe can additionally take as input user-defined transcripts (from any species). The format is similar to the one in Ensembl and includes exons coordinates, their rank, frame and nucleotide sequence.

b- Pairwise-alignment-based exon pre-clustering. ThorAxe clusters the input exons based on their sequence similarity (**Fig. 1C**, 3 clusters colored in red, cyan and blue). This step step provides a coarse-grained partitioning of the sequence space that reduces the complexity of step d (see below). We perform pairwise local alignments using a modified version of the Hobohm I algorithm (**Supplemental Methods**). We use a relatively low default sequence identity threshold of 30% to ensure homology detection across many species. As illustrated in Figure 1C by cluster *2*, pairs of duplicated mutually exclusive exons coming from the same species will likely be grouped in the same cluster (see also **Supplemental Fig. S1**).

c- Redundancy reduction. ThorAxe defines a set of sub-exons for each species. This implies systematically detecting overlapping exons, and replacing them by non-redundant distinct sub-exons. In the example illustrated in Figure 1C, one exon from gorilla leads to the definition of two sub-exons (in red, see also their sequences in **Supplemental Fig. 2A**). This step relies only on the genomic coordinates of the exons and does not require aligning the exonic sequences. It is performed after exon clustering, since dealing with sub-exons at this early stage would add some unnecessary complexity by augmenting the number of comparisons and the ambiguity associated with small sequences.

d- MSA-based s-exon identification. ThorAxe defines a set of s-exons across all species, by aligning the exonic sequences belonging to each cluster defined in step b and identifying blocks in the constructed MSAs (**Fig. 1C**, see the 3 MSAs corresponding to the 3 clusters). The aim of this step is to determine a mapping of exonic sequences between different species (see details in **Supplemental Methods**). To identify the s-exons from each MSA, ThorAxe scans the MSA from left to right and create a new s-exon whenever there is a change of sub-exon in at least one sequence/species (see Algorithm 1). This ensures that the identified s-exons can be used as building blocks to reconstruct any transcript in any species from the input data

e- S-exon refinement. ThorAxe refines the s-exons definition by locally optimising the MSAs built in step d. The aim is two-fold, namely to improve the quality of the MSAs associated to the s-exons, and to reduce the number of s-exons, and hence the number of induced edges in the corresponding ESG. This step thus represents a mean to increase the ESG score expressed in Eq. 1 without explicitly computing it. Specifically, we systematically detect lowly-scored sub-exons and migrate them from one MSA to another, and we minimise the number of very small s-exons, comprising only 1 or 2 columns (**Supplemental Methods**).

f- ESG construction and annotation, and event detection. Once the s-exons have been identified, building the corresponding ESG is straightforward (**Fig. 1C**). ThorAxe

annotates the nodes and the edges of the output graph with evolutionary information and summary statistics (**Supplemental Fig. S2D** and **Supplemental Methods**). Finally, it defines a *canonical* transcript and detects a set of events as variations between this reference transcript and each input transcript. Ideally, the canonical transcript should be well represented across species, and thus, to choose it, we rely on a combination of conservation measures computed over the ESG edges (**Supplemental Methods** and **Supplemental Fig. S24**). By default, the events are detected on a reduced version of the ESG, where the edges supported by only one transcript have been removed (Algorithm 2). We visualised the ESGs with Cytoscape V.3.7.2 (Shannon et al., 2003).

An additional output of ThorAxe is the list of input transcripts described as collections of s-exons (where each s-exon is designated by a symbol) and the gene tree representative of the selected species. These data can directly serve as input for PhyloSofS (Ait-hamlat et al., 2020), toward the reconstruction of transcripts' phylogenetic forests. ThorAxe may also be easily interfaced with other tools requiring the same type of input.

Analysis of ThorAxe results

We give details about the calculation of the MSA scores, the detection of similar pairs of s-exons, the complementation of the ESGs with RNA-seq splice junctions, the characterization of isoforms 3D structures and disorder content, the functional analysis of some genes, the comparison with phastCons scores, other studies, and other methods in the **Supplemental Methods**, **Supplemental Fig. S25-27** and **Supplemental Table S10**.

SOFTWARE AVAILABILITY

ThorAxe is freely available at GitHub (`https://github.com/PhyloSofS-Team/thoraxe`) and as a stand-alone package and Python module as Supplemental Code. All data supporting the findings of this study are available via a supplementary webserver (`http://www.lcqb.upmc.fr/ThorAxe`) and as Supplemental Material.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

A grant of the French national research agency (MASSIV project, ANR-17-CE12-0009) provided a salary to D.J.Z. and S.L.. We thank P. Charpentier and J. Cortés for their help in the systematic detection of the disordered regions, and S. Grudinin for insightful comments. We thank F. Oteri and H. Ripoche for the technical support. We thank the reviewers for their comments which greatly improved the manuscript.

AUTHOR CONTRIBUTIONS

D.J.Z., H.R. and E.L. designed research. D.J.Z. and S.L. carried out the implementation. D.J.Z, S.L., H.R. and E.L. produced and analysed the results. A.B. contributed the proof

of NP-hard complexity. E.L. wrote the manuscript with support and feedback from all authors. H.R. and E.L. supervised the project.

References

Abascal, F., Tress, M. L., and Valencia, A., 2015. The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome biology and evolution*, **7**(6):1392–1403.

Agosto, L. M., Gazzara, M. R., Radens, C. M., Sidoli, S., Baeza, J., Garcia, B. A., and Lynch, K. W., 2019. Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing. *Genome research*, **29**(12):2046–2055.

Ait-hamlat, A., Zea, D. J., Labeeuw, A., Polit, L., Richard, H., and Laine, E., 2020. Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the jnk family. *Journal of Molecular Biology*, .

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990. Basic local alignment search tool. *Journal of molecular biology*, **215**(3):403–410.

Bähler, M. and Rhoads, A., 2002. Calmodulin signaling via the iq motif. *FEBS letters*, **513**(1):107–113.

Baralle, F. E. and Giudice, J., 2017. Alternative splicing as a regulator of development and tissue identity. *Nature reviews Molecular cell biology*, **18**(7):437.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.*, 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**(6114):1587–1593.

Bayer, K.-U., Harbers, K., and Schulman, H., 1998. αkap is an anchoring protein for a novel cam kinase ii isoform in skeletal muscle. *The EMBO journal*, **17**(19):5598–5605.

Bhattacharyya, M., Lee, Y. K., Muratcioglu, S., Qiu, B., Nyayapati, P., Schulman, H., Groves, J. T., and Kuriyan, J., 2020. Flexible linkers in camkii control the balance between activating and inhibitory autophosphorylation. *Elife*, **9**:e53670.

Birzele, F., Csaba, G., and Zimmer, R., 2008. Alternative splicing and protein structure evolution. *Nucleic Acids Res.*, **36**(2):550–558.

Björklund, Å. K., Light, S., Sagit, R., and Elofsson, A., 2010. Nebulin: a study of protein repeat evolution. *Journal of molecular biology*, **402**(1):38–51.

Brocke, L., Srinivasan, M., and Schulman, H., 1995. Developmental and regional expression of multifunctional ca2+/calmodulin-dependent protein kinase isoforms in rat brain. *Journal of Neuroscience*, **15**(10):6797–6808.

Bulleit, R. F., Bennett, M. K., Molloy, S. S., Hurley, J. B., and Kennedy, M. B., 1988. Conserved and variable regions in the subunits of brain type ii ca 2+/calmodulin-dependent protein kinase. *Neuron*, **1**(1):63–72.

Byrne, A., Cole, C., Volden, R., and Vollmers, C., 2019. Realizing the potential of full-length transcriptome sequencing. *Philosophical Transactions of the Royal Society B*, **374**(1786):20190097.

Chakraborty, A. and Chakrabarti, S., 2015. A survey on prediction of specificity-determining sites in proteins. *Briefings in Bioinformatics*, **16**(1):71–88.

Christinat, Y. and Moret, B. M., 2012. Inferring transcript phylogenies. *BMC Bioinformatics*, **13 Suppl 9**:S1.

Climente-González, H., Porta-Pardo, E., Godzik, A., and Eyras, E., 2017. The functional impact of alternative splicing in cancer. *Cell reports*, **20**(9):2215–2226.

Cook, S. G., Bourke, A. M., O'Leary, H., Zaegel, V., Lasda, E., Mize-Berge, J., Quillinan, N., Tucker, C. L., Coultrap, S. J., Herson, P. S., *et al.*, 2018. Analysis of the CaMKIIβ and α splice-variant distribution among brain regions reveals isoform-specific differences in holoenzyme formation. *Sci Rep*, **8**(1):5448.

de la Fuente, L., Arzalluz-Luque, Á., Tardáguila, M., del Risco, H., Martí, C., Tarazona, S., Salguero, P., Scott, R., Lerma, A., Alastrue-Agudo, A., *et al.*, 2020. tappas: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biology*, **21**(1):1–32.

De La Grange, P., Gratadou, L., Delord, M., Dutertre, M., and Auboeuf, D., 2010. Splicing factor and exon profiling across human tissues. *Nucleic acids research*, **38**(9):2825–2838.

Denti, L., Rizzi, R., Beretta, S., Della Vedova, G., Previtali, M., and Bonizzoni, P., 2018. Asgal: aligning rna-seq data to a splicing graph to detect novel alternative splicing events. *BMC bioinformatics*, **19**(1):444.

Ezkurdia, I., Rodriguez, J. M., Carrillo-de Santa Pau, E., Vazquez, J., Valencia, A., and Tress, M. L., 2015. Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, **14**(4):1880–1887.

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.

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

Greenberg, M. J. and Ostap, E. M., 2013. Regulation and control of myosin-i by the motor and light chain-binding domains. *Trends in cell biology*, **23**(2):81–89.

Heber, S., Alekseyev, M., Sze, S.-H., Tang, H., and Pevzner, P. A., 2002. Splicing graphs and est assembly problem. *Bioinformatics*, **18**(suppl_1):S181–S188.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., *et al.*, 2016. Ensembl comparative genomics resources. *Database (Oxford)*, **2016**.

Houdusse, A., Silver, M., and Cohen, C., 1996. A model of ca2+-free calmodulin binding to unconventional myosins reveals how calmodulin acts as a regulatory switch. *Structure*, **4**(12):1475–1490.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S., 2013. Function of alternative splicing. *Gene*, **514**(1):1–30.

Khan, S., Downing, K. H., and Molloy, J. E., 2019. Architectural dynamics of camkii-actin networks. *Biophysical journal*, **116**(1):104–119.

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., *et al.*, 2014. A draft map of the human proteome. *Nature*, **509**(7502):575–581.

Komljenovic, A., Roux, J., and Wollbrett, J., 2018. BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *peer review: 2 approved, 1 approved with reservations*, .

Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I. N. S. D., 2010. The sequence read archive. *Nucleic acids research*, **39**(suppl_1):D19–D21.

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., and Fairbrother, W. G., 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.*, **108**(27):11093–11098.

Liu, J., Taylor, D. W., Krementsova, E. B., Trybus, K. M., and Taylor, K. A., 2006. Three-dimensional structure of the myosin v inhibited state by cryoelectron tomography. *Nature*, **442**(7099):208–211.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.*, 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**(6):580–585.

Louadi, Z., Yuan, K., Gress, A., Tsoy, O., Kalinina, O. V., Baumbach, J., Kacprowski, T., and List, M., 2020. Digger: exploring the functional role of alternative splicing in protein interactions. *Nucleic Acids Research*, .

Marti-Solano, M., Crilly, S. E., Malinverni, D., Munk, C., Harris, M., Pearce, A., Quon, T., Mackenzie, A. E., Wang, X., Peng, J., *et al.*, 2020. Combinatorial expression of GPCR isoforms affects signalling and drug responses. *Nature*, .

Mei, W., Boatwright, L., Feng, G., Schnable, J. C., and Barbazuk, W. B., 2017. Evolutionarily conserved alternative splicing across monocots. *Genetics*, **207**(2):465–480.

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

Modrek, B. and Lee, C. J., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics*, **34**(2):177–180.

Nichio, B. T., Marchaukoski, J. N., and Raittz, R. T., 2017. New tools in orthology analysis: a brief review of promising perspectives. *Frontiers in genetics*, **8**:165.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y., 2018. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics*, **102**(1):11–26.

Pathan-Chhatbar, S., Taft, M. H., Reindl, T., Hundt, N., Latham, S. L., and Manstein, D. J., 2018. Three mammalian tropomyosin isoforms have different regulatory effects on nonmuscle myosin-2b and filamentous β-actin in vitro. *Journal of Biological Chemistry*, **293**(3):863–875.

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G., Valencia, A., and Tress, M. L., 2013. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.*, **41**(Database issue):D110–117.

Salzberg, S. L., 2019. Next-generation genome annotation: we still struggle to get it right.

Scotti, M. M. and Swanson, M. S., 2016. Rna mis-splicing in disease. *Nature Reviews Genetics*, **17**(1):19.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**:2498–2504.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.*, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**(8):1034–1050.

Siepel, A. and Haussler, D., 2005. Phylogenetic hidden markov models. In *Statistical methods in molecular evolution*, pages 325–351. Springer.

Sloutsky, R. and Stratton, M. M., 2020. Functional implications of camkii alternative splicing. *European Journal of Neuroscience*, .

Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C., and Blencowe, B. J., 2018. Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Molecular cell*, **72**(1):187–200.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.*, 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**(5891):956–960.

Szalkowski, A. M., 2012. Fast and robust multiple sequence alignment with phylogeny-aware gap placement. *BMC bioinformatics*, **13**(1):129.

Tapial, J., Ha, K. C. H., Sterne-Weiler, T., Gohr, A., Braunschweig, U., Hermoso-Pulido, A., Quesnel-Valli?res, M., Permanyer, J., Sodaei, R., Marquez, Y., *et al.*, 2017. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res*, **27**(10):1759–1768.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G., 1997. The clustal_x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research*, **25**(24):4876–4882.

Tranchevent, L. C., Aube, F., Dulaurier, L., Benoit-Pilven, C., Rey, A., Poret, A., Chautard, E., Mortada, H., Desmet, F. O., Chakrama, F. Z., *et al.*, 2017. Identification of protein features encoded by alternative exons using Exon Ontology. *Genome Res.*, **27**(6):1087–1097.

Waetzig, V. and Herdegen, T., 2005. Context-specific inhibition of jnks: overcoming the dilemma of protection and damage. *Trends in pharmacological sciences*, **26**(9):455–461.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221):470–476.

Wang, G.-S. and Cooper, T. A., 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, **8**(10):749–761.

Wang, L. and Jiang, T., 1994. On the complexity of multiple sequence alignment. *Journal of computational biology*, **1**(4):337–348.

Wang, P., Wu, Y.-L., Zhou, T.-H., Sun, Y., and Pei, G., 2000. Identification of alternative splicing variants of the β subunit of human ca2+/calmodulin-dependent protein kinase ii with different activities. *FEBS letters*, **475**(2):107–110.

Wang, X., Codreanu, S. G., Wen, B., Li, K., Chambers, M. C., Liebler, D. C., and Zhang, B., 2018. Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Mol Cell Proteomics*, **17**(3):422–430.

Ward, A. J. and Cooper, T. A., 2010. The pathobiology of splicing. *J. Pathol.*, **220**(2):152–163.

Weatheritt, R. J., Sterne-Weiler, T., and Blencowe, B. J., 2016. The ribosome-engaged landscape of alternative splicing. *Nature structural & molecular biology*, **23**(12):1117.

Wu, S., Liu, J., Reedy, M. C., Tregear, R. T., Winkler, H., Franzini-Armstrong, C., Sasaki, H., Lucaveche, C., Goldman, Y. E., Reedy, M. K., *et al.*, 2010. Electron tomography of cryofixed, isometrically contracting insect flight muscle reveals novel actin-myosin interactions. *PLoS One*, **5**(9):e12643.

Xing, Y. and Lee, C., 2005. Assessing the application of Ka/Ks ratio test to alternatively spliced exons. *Bioinformatics*, **21**(19):3701–3703.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., *et al.*, 2016. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**(4):805–817.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., *et al.*, 2016. Ensembl 2016. *Nucleic Acids Res.*, **44**(D1):D710–716.

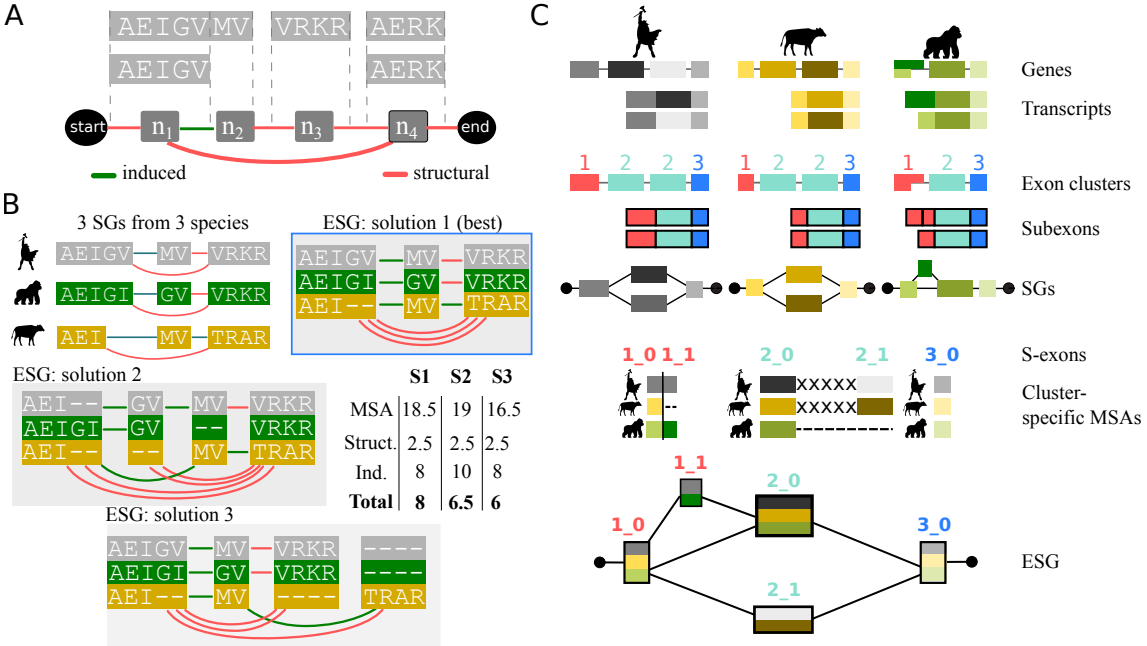

Figure 1: **Principle of the method. A**. Two transcripts are depicted, where each grey box represents a genomic interval and contains the corresponding protein sequence. Below, the minimal SG is shown, with the nodes (n_1, n_2, n_3, n_4) corresponding to sub-exons. The *start* and *end* nodes are added for convenience. Each structural edge in red corresponds to some intron while each induced edge in green corresponds to a junction located inside the initial genomic interval (such as the donor site of exon `AEIGV`). **B**. Close-up view of three SGs corresponding to three orthologous genes coming from human, gorilla and cow, along with three examples of ESGs summarising the same information. The nodes in the ESGs represent *s-exons*, or MSAs of exonic regions. The details of the ESG scores, computed from Eq. 1, are given in the insert table, with $\sigma_{match} = 1$, $\sigma_{mismatch} = -0.5$ and $\sigma_{gap} = 0$ for the MSA scores, and edge penalties $\sigma_S = 0.5$ and $\sigma_I = 2$. The best-scored ESG shows at the same time compactness (parsimony) and good-quality alignments. **C**. Main steps of the ESG construction in ThorAxe. The input genes and transcripts are depicted on top, with exons displayed as boxes. ThorAxe first step consists in grouping similar exons together. Here, three clusters are identified, colored in red (1), cyan (2) and blue (3) – note that cluster 2 groups to multiple exons in human and cow. Then, sub-exons are defined based on intra-species transcript variability. For instance, the first exon from gorilla is split into two sub-exons. The sub-exons would be the nodes in the species-specific minimal SGs, although the latter are not explicitly computed by ThorAxe. The next step consists in aligning the sequences belonging to each cluster (with some padding "X" between mutually exclusive sub-exons) and identifying the s-exons as blocks in the alignment. We keep track of the cluster ids in the s-exon ids, to ease interpretability. Finally, ThorAxe builds an ESG where the nodes are the s-exons. For the sake of clarity, multi-edges are visualized as single edges.

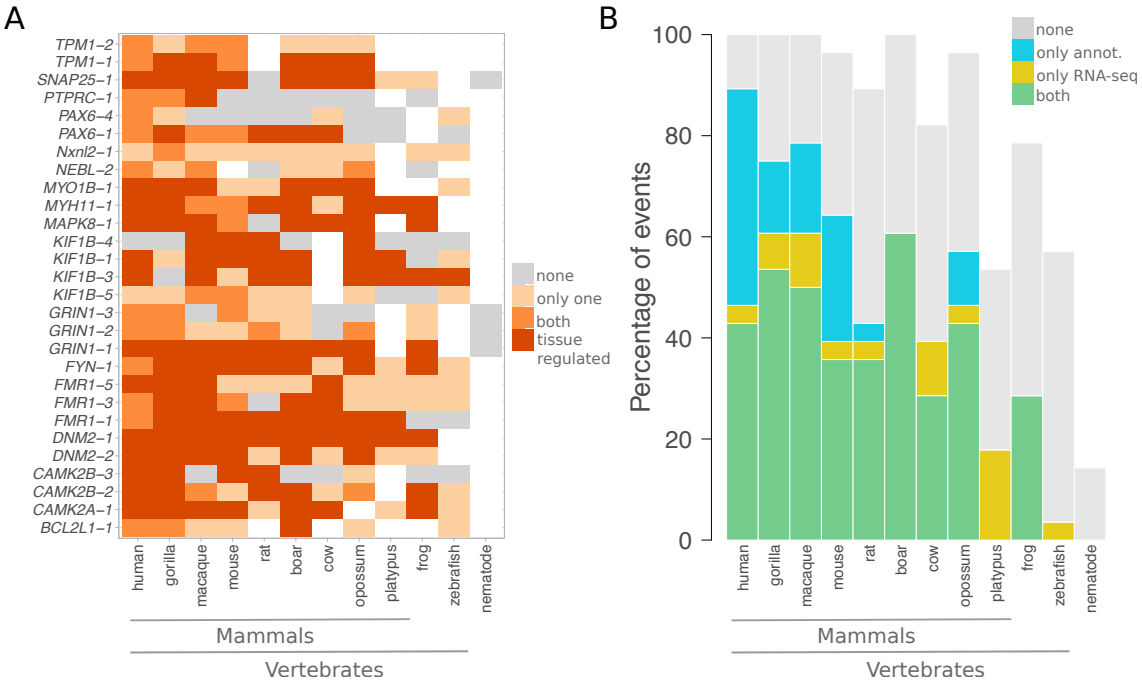

Figure 2: **Conservation and tissue regulation of a set of documented AS events. A.** Each event is designated by the name of the gene where it occurs and its rank in ThorAxe output, the latter reflecting its relative conservation level. In the ESG, an event corresponds to a pair of subpaths, one being canonical and the other alternative. Within each species, either none of the paths are supported by the data (grey), or only one path is supported (light orange), or both paths are supported (orange and dark orange). As data, we consider the gene annotations from Ensembl and the RNA-seq evidence compiled from public databases. When both paths are supported, we highlight the cases where they are differentially expressed in at least one tissue in dark orange. The white cells indicate species where a one-to-one ortholog of the human query gene could not be found. **B.** For each species, the percentages of events supported by both gene annotations and RNA-seq (in green), by only RNA-seq (in yellow), by only gene annotations (in blue), or unsupported (in grey) are reported. Note that an event is considered as supported only if both its canonical and alternative subpaths are detected.

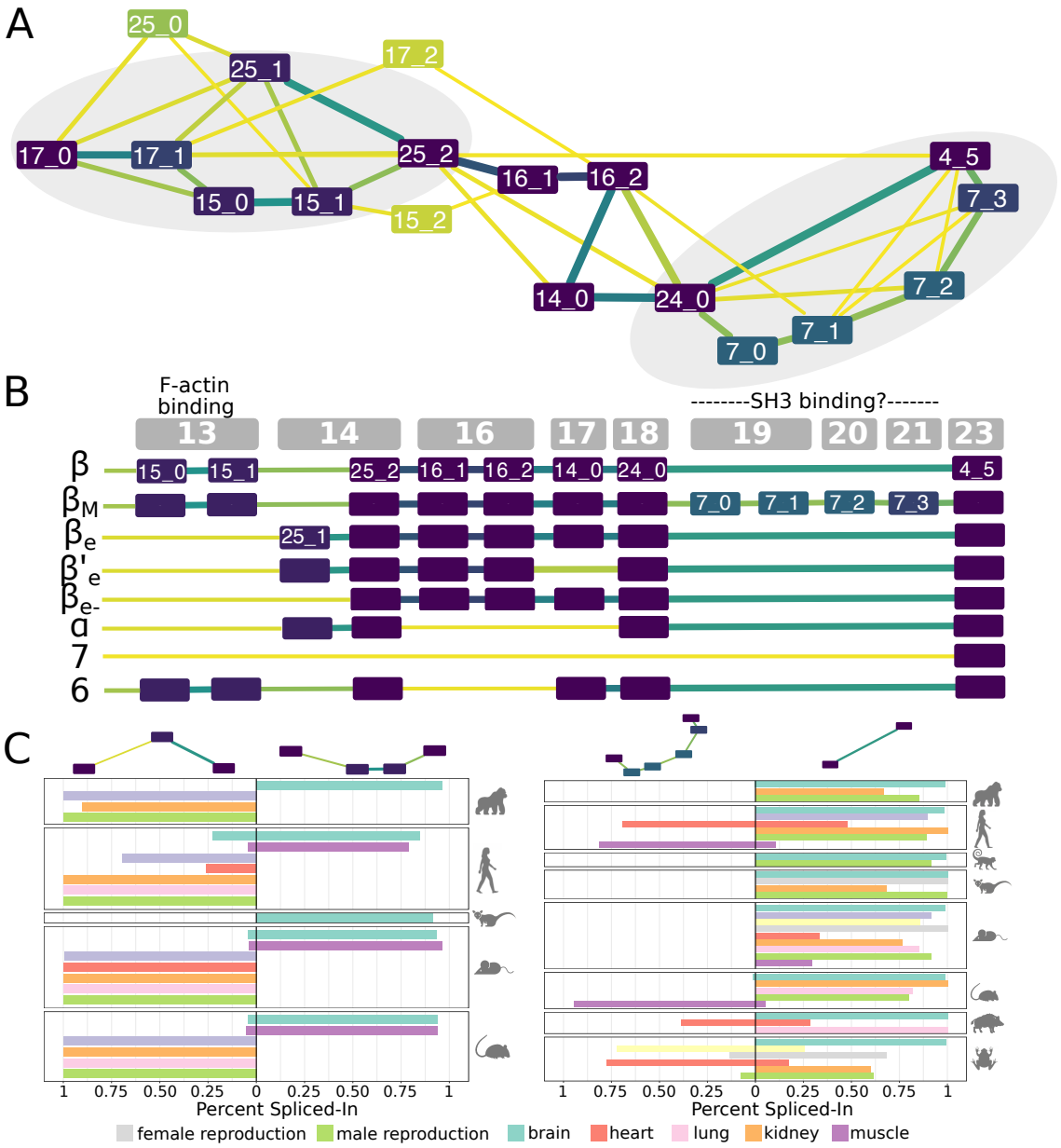

Figure 3: **Transcript variability in the *CAMK2B* linker. A.** Evolutionary splicing subgraph computed by ThorAxe starting from 63 transcripts annotated in 10 species. It corresponds to the region linking the kinase and hub domains of *CAMK2B*. The colours of the nodes and the edges indicate conservation levels, from yellow (low) to dark purple (high). Conservation is measured as the *species fraction* for the nodes (proportion of species where the s-exon is present) and as the *averaged transcript fraction* for the edges (averaged transcript inclusion rate of the s-exon junction). For ease of visualisation, we filtered out the s-exons present in only one species. The events documented in the literature are located in the grey areas. **B.** On top, genomic structure of the human gene. Each grey box corresponds to a genomic exon (nomenclature taken from (Sloutsky and Stratton, 2020)). Below, list of human transcripts. All of them have been described in the literature, referred to as β (Bulleit et al., 1988), β_M (Bayer et al., 1998), β_e (Brocke et al., 1995), β'_e (Brocke et al., 1995), β_e- (Cook et al., 2018), α (Bulleit et al., 1988), 7 (Wang et al., 2000) and 6 (Wang et al., 2000). The functional roles of some exons (Bayer et al., 1998; Khan et al., 2019) are given. **C.** Percent-Spliced In (PSI) computed from RNA-seq splice junctions for the two documented AS events. The two pairs of alternative subpaths depicted on top are also highlighted on panel A.

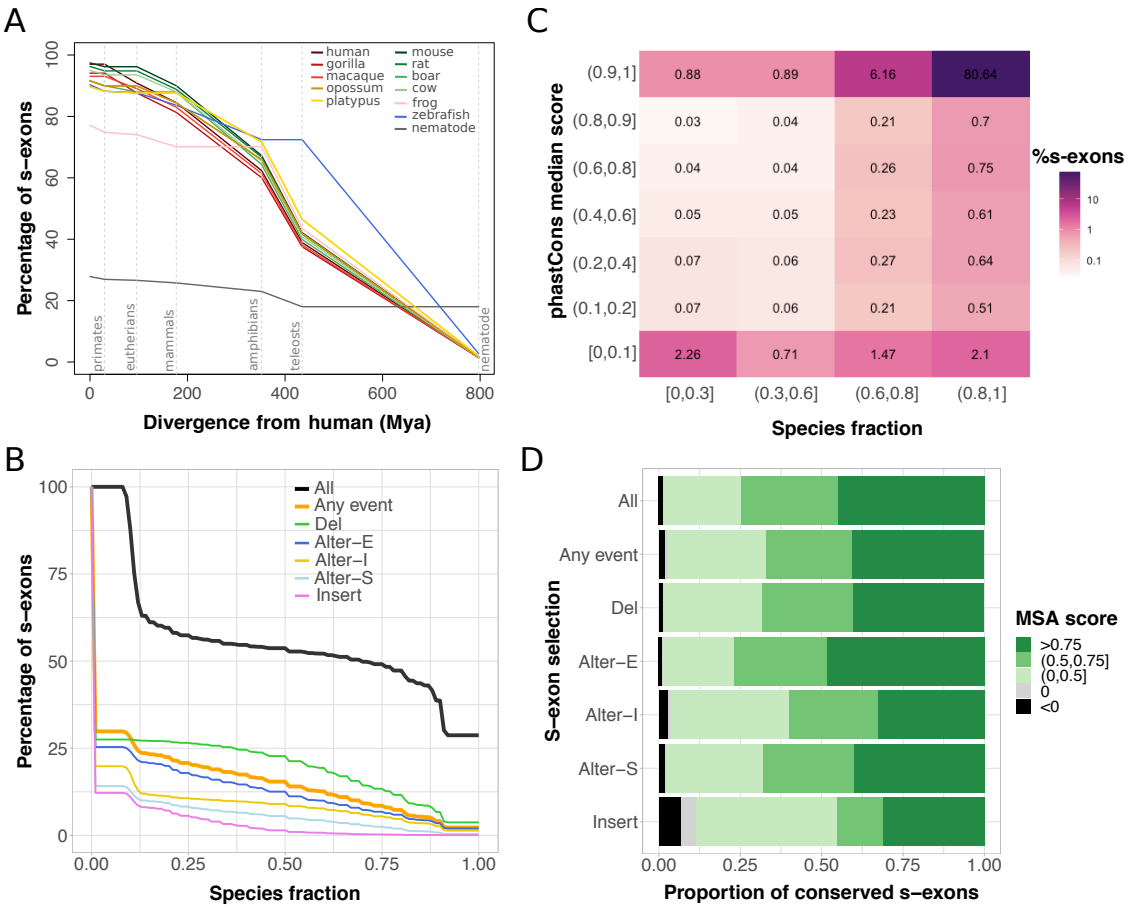

Figure 4: **S-exon evolutionary profiling over the whole protein-coding human genome.** **A**. Percentages of s-exons conserved at different evolutionary distances from human (represented by dashed vertical lines). Each curve is centred on its corresponding species. The values at the origin are the percentages of conserved (*i.e.* not species-specific) s-exons. Conservation is then assessed at each evolutionary distance according to the s-exons possessing at least one representative in each phylogroup. For instance, we report 73-76% of the s-exons of frog (pink curve) as conserved among eutherians (second dashed line) in the sense that they are also conserved in at least one primate (among human, gorilla, macaque) and at least one non-primate eutherian (among rat, mouse, boar, cow). Likely, conservation up to mammals (68-72% for frog) would imply at least one primate, one non-primate eutherian and one non-eutherian mammal. See also Supplemental Fig. S10 for a version of this plot focusing on genes with one-to-one orthologs in more than seven species. **B**. Cumulative distributions of s-exon *species fraction*. On the y-axis we report the percentage of s-exons with a *species fraction* greater than the x-axis value. The different curves correspond to all s-exons (*All*), only those involved in at least an event (*Any event*), or only those involved in a specific type of event. *Alter-S*: alternative start. *Alter-I*: alternative (internal). *Alter-E*: alternative end. *Del*: deletion. *Insert*: insertion. **C**. Heatmap of the s-exon phastCons median scores versus the s-exon *species fractions*. Only the s-exons longer than 10 residues and belonging to genes with one-to-one othologs in more than seven species are shown. **D**. Proportions of conserved s-exons displaying very poor (negative score) to very good (score close to one) alignment quality. The MSA score of a s-exon is computed as a normalised sum of pairs. A score of 1 indicates 100% sequence identity without any gap. The proportions are given for different s-exon selections (same labels as in panel B).

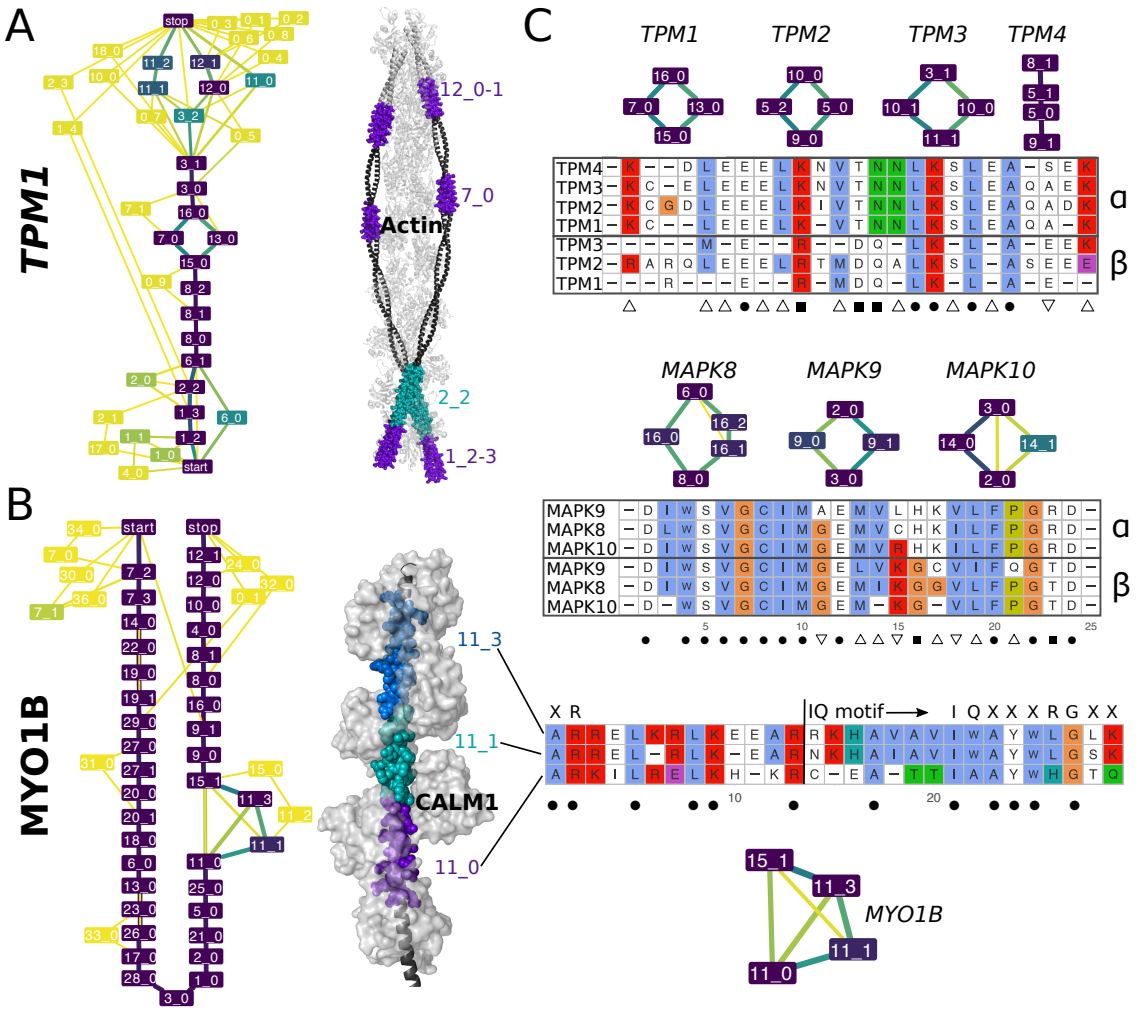

Figure 5: **Examples of evolutionary conserved events with in-gene paralogy. A-B.** ESGs computed by ThorAxe (on the left) and the best 3D templates found by HHblits (on the right, PDB codes 2w49:abuv (Wu et al., 2010) and 2dfs:H (Liu et al., 2006)) for *TPM1* and *MYO1B*. On the ESGs, the colours indicate conservation levels, *species fraction* for the nodes and *averaged transcript fraction* for the edges (see **Supplemental Methods**). The nodes in yellow are species-specific while those in dark purple are present in all species. The 3D structures show complexes between the query proteins (in black) and several copies of their partners (in light grey). The s-exons involved in conserved events are highlighted with coloured spheres. **C.** S-exon consensus sequence alignments within a gene family (*TPM* on top, *MAPK* in the middle) or a gene (*MYO1B*, at the bottom). Each letter reported is the amino acid conserved in all sequences of the corresponding MSA (allowing one substitution). The colour scheme is that of Clustal X (Thompson et al., 1997). The subgraphs show the events in which the s-exons are involved. The symbols α and β on the right indicate groups of s-exons defined across paralogous genes based on sequence similarity (see **Supplemental Methods**). The symbols at the bottom denote highly conserved positions across the gene family. Dot: fully conserved position. Square: position conserved only within each s-exon group. Upward triangle: position conserved in the α group only. Downward triangle: position conserved in the β group only. For *MYO1B*, the start and canonical sequence of the CALM1-binding IQ motif are indicated. The motifs resulting from different combinations of the depicted s-exons are numbered 4, 4/5 and 4/6 in the literature (Greenberg and Ostap, 2013).

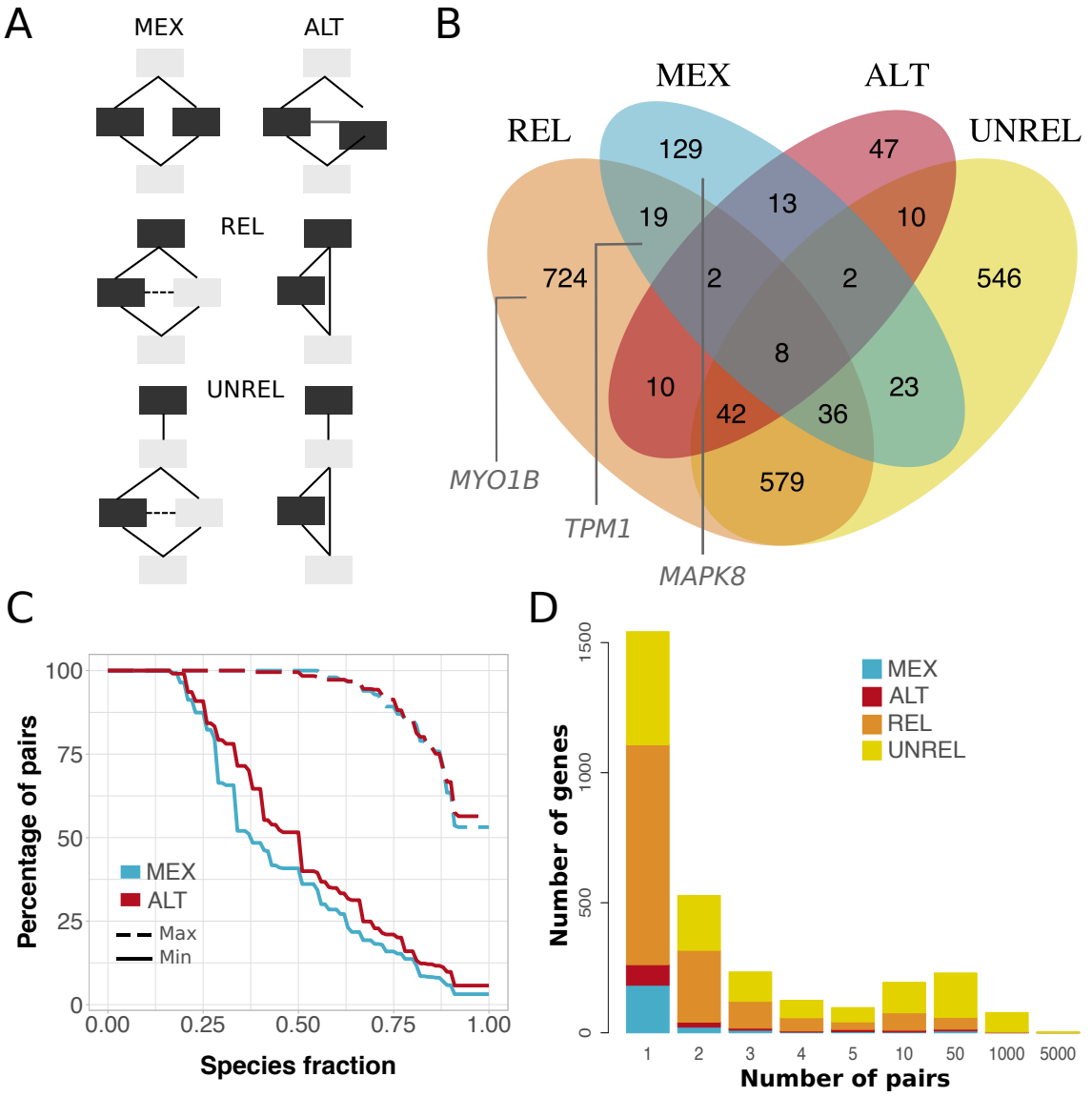

Figure 6: **Alternative usage of similar s-exons. A.** Evolutionary splicing subgraphs depicting different alternative usage scenarios. The detected s-exon pairs are colored in black. MEX: mutual exclusivity. ALT: alternative (non mutually exclusive) usage. REL: one s-exon is in the canonical or alternative subpath of an event (of any type), while the other one serves as a "canonical anchor" for the event. UNREL: one s-exon is in the canonical or alternative subpath of an event (of any type), while the other one is located outside the event in the canonical transcript. Each detected pair is assigned to only one category with the following priority rule: MEX>ALT>REL>UNREL. **B.** Venn diagram of the genes containing similar pairs of s-exons. The genes shown in Figure 5 are highlighted in the corresponding subsets. **C.** Cumulative distributions of s-exon conservation. On the y-axis we report the percentage of s-exon pairs with *species fraction* greater than the x-axis value. The solid (resp. dashed) curve corresponds to the highest (resp. lowest) *species fraction* among the two s-exons in the pair. We report values only for the MEX (in blue) and ALT (in red) categories. **D.** Distribution of per-gene s-exon pair number within each of the four categories. For instance, the yellow rectangle at $x = 50$ gives the number of genes with more than 10 and up to 50 UNREL s-exon pairs.