Assessing conservation of alternative splicing with evolutionary splicing graphs

Supplemental Material

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

Supplemental Methods

Datasets

AS-dedicated gene set

We collected a set of 50 genes, representing 16 families, where AS produces functionally distinct protein isoforms (**Supplemental Table S1**). Within each family, the biochemical activities of several isoforms have been characterised. The set contains single-domain genes as well as multi-domain ones, ranging from less than 200 to 2000 residues. It includes some kinases, some receptors, some RNA-binding proteins, a transcription factor, and a significant portion of proteins involved in the formation of the cell cytoskeleton, in muscle contraction and in membrane trafficking.

Human proteome

We considered the ensemble of 19,976 protein coding genes comprised in the human genome and their one-to-one orthologs from gorilla, macaque, mouse, rat, boar, cow, opossum, platypus, frog, zebrafish and nematode. We downloaded the corresponding gene annotations from Ensembl[1] release 98 (September 2019). The download was successful for 18 241 genes. Among those, 14 genes did not have any good quality transcripts (see below for the criteria) and 1 gene displayed an error in the gene tree, leading to a total of 18 226 valid genes. We should stress that a small fraction of these genes (3%) do not have any valid human transcript.

Definitions

Splicing Graph (SG)

We adopt a peculiar definition of splicing graph which differs from the classical one found in the literature. Specifically, we account for the reading frames in the definition of the nodes and the edges of the graph. Hence, we consider only open reading frames and we exclude untranslated regions from the graph. Given a gene G_i, and its annotated transcripts described as sorted lists of genomic intervals, we define a *splicing graph* (SG) for G as the directed graph $\mathcal{S}_i = (\mathcal{V}_i, \mathcal{E}_i)$ (**Fig. 1A**). Each node $n \in \mathcal{V}_i$ is identified by (n_s, n_e, n_f), where $[n_s, n_e]$ is the genomic interval covered by n and n_f is its reading frame. There is a node n in $\mathcal{S}_i$ if the corresponding coordinates and frame occur in at least one transcript. In practice, we place ourself at the amino acid level and, whenever an amino acid residue is shared between two genomic exons, we assign it to one of the corresponding nodes. There is an edge in $\mathcal{E}_i$ from n to n' if the corresponding coordinates and frames are consecutively transcribed. Edges are classified either as *structural* if n and n' are separated by an intron or *induced* by the nodes' genomic boundaries otherwise. Finally, a *start* and *end* nodes are added and act respectively as the least and the greatest element for the partial order induced on the graph. Note that, due to the partial order on genomic intervals and the colinearity of transcripts in the genome, the graph $\mathcal{S}_i$ is directed acyclic.

In principle, there are many possible splicing graphs, depending on how the nodes are defined. The *minimal splicing graph* is the unique SG with the smallest number of nodes, such that each one of the input transcripts is represented by a path. When not stated explicitly we will always refer to the minimal splicing graph.

Sub-exon

We define sub-exons as nodes in a minimal SG. They are the minimal building blocks for the transcripts observed in a given species. As an illustrative example, let us consider the gene *SNAP25* from gorilla (**Supplemental Fig. S1B**). The two first transcripts, TRX1 and TRX2, start with the same exon (`MAE...ADE`, in yellow) comprised of 24 residues. The third transcript, TRX3, starts with another longer exon (`MAE...GRE`, in yellow and orange) comprised of 33 residues and overlapping with the previous one. In that case, we define two sub-exons, one of 24 residues (`MAE...ADE`, in yellow) and another of 9 residues (`VRS...GRE`, in orange).

Evolutionary Splicing Graph (ESG)

We extend the definition of the SG to a set of orthologous genes $G = \{G_1, G_2, \ldots, G_m\}$ (**Fig. 1B**). Our aim is to construct a graph summarising transcript information from $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ together with evolutionary conservation by means of an evolutionary splicing graph (ESG) $\mathcal{S} = (\mathcal{V}, \mathcal{E})$. Although some cycles may appear in $\mathcal{S}$ due to exon switching, the graph will be directed acyclic in most cases. Each $v \in \mathcal{V}$ is formally described by a list $[(v_s^1, v_e^1, v_f^1), \ldots, (v_s^m, v_e^m, v_f^m)]$ of genomic coordinates and reading frames for each of the genes (some elements can be empty). In practice, we will be interested in the corresponding MSAs of the translated sequences. The edge set $\mathcal{E}$ comprises the ensemble of edges linking the nodes in the gene-specific SGs, namely $\{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_n\}$. Additionally, some edges may be added to link any pair $\left((v_s^i, v_e^i, v_f^i), (u_s^i, u_e^i, u_f^i)\right)$ of genomic coordinates and reading frames from the nodes v and u in $\mathcal{S}$ that were consecutive in the node w defined by (w_s, w_e, w_f) from $\mathcal{S}_i$. This implies that $|v_e^i - u_s^i| = 1$, $[v_s^i, v_e^i] \subset [w_s, w_e]$ and $[u_s^i, u_e^i] \subset [w_s, w_e]$. By definition, any such edge is *induced*. Any two nodes in $\mathcal{S}$ are thus connected by a set of edges, up to m, which we can refer to as a *multi-edge*. Each multi-edge can be either purely structural, purely induced or hybrid, whenever several genomic exons are joined in some of the species.

As for the individual SGs, there are multiple ways of constructing $\mathcal{S}$. Our goal is to minimise the number of nodes while maximising the overall sequence similarity of the MSAs associated to each node. We formalise the problem using a scoring function (Eq. **??**). A *minimal evolutionary splicing graph* is simply a graph with maximal score. Note that there may exist several minimal ESGs for a given gene.

S-exon

We define s-exons as nodes in a minimal ESG. They are the minimal building blocks for the transcripts observed in a set of species. Let us consider again the example of *SNAP25*, studied across eight species (**Supplemental Fig. S1C-D**). Each s-exon is represented by a MSA, where each sequence comes from one species and corresponds to a sub-exon or a part of a sub-exon. The exonic sequences belonging to the same s-exon are supposed to be orthologous. A species-specific s-exon comprises only one sequence (coming from one species), while a conserved s-exon comprises at least two sequences (coming from two species).

Algorithms implemented in ThorAxe

Accounting for exon frame constraints

Since ThorAxe explicitly accounts for the exon frames, two exons in a species starting on a different frame but on the same genomic coordinate will not share a splice junction. Across several species, the existence of different frames does not prevent, in principle, the corresponding exons to be grouped in the same s-exon. However, we expect this scenario to be very unlikely. Indeed, for two splicing edges

with different phases to be considered as conserved/homologous, two criteria need to be met : (1) the same location for the different frames (which could suggest some error in Ensembl annotations), (2) a high similarity between the exonic sequences translated from the different frames such that this signal may confuse the alignment. By analyzing the curated set of 50 genes, we found that among the 39 373 exonic sequences defining the 2 346 s-exons identified by ThorAxe, only one sequence had inconsistent phases with the other sequences of its s-exon. Moreover, this wrong sequence assignment did not have any impact on the estimation of the conservation of the splice junctions associated with this s-exon. Indeed, the splice junctions associated with the "wrong" sequence were observed only in the species from where the sequence originated. We would also like to stress that finding highly similar exonic sequences translated from two different frames at the same location would suggest some error in the way the frames were annotated in Ensembl.

Modified version of the Hobohm I algorithm to cluster exons

We implemented a modified version of the Hobohm I algorithm[2] to cluster the exons sharing some sequence similarity. We first order the exons from the longest to the shortest one. Then, at each step of the algorithm, we locally align the current exon i with the following (smaller) exons using the Smith-Waterman algorithm[3]. If exon j, with $j > i$, shares more than id_{cut}% sequence identity with exon i and is covered at more than cov_{cut}% by the alignment, then we assign the two exons to the same cluster. By default, $id_{cut} = 30$ and $cov_{cut} = 80$, and these parameters can be adjusted by the user. We progressively remove the clustered exons from the initial set, such that they will not be considered in the next iteration of the algorithm for the comparisons. We modified the algorithm to be more stringent on this latter criterion. Specifically, if exon j gets assigned to the same cluster as exon i but shares less than $(id_{cut} + 30)$% sequence identity with i, then it remains in the set. It will be compared against the other exons in the next iteration, and will migrate from one cluster to another if we find a better match. Notice that a cluster may contain only one exon.

Multiple sequence alignment within each cluster

For each exon cluster defined in step b, we generate a MSA comprising n sequences, where n is the number of species with at least one exon in the cluster (**Supplemental Fig. 2B-C**). Each sequence S_i^j is a chimeric construct built by concatenating the sub-exons from cluster i defined in step c for species j. We concatenate the sub-exons in the order of their genomic coordinates, and we introduce a padding sequence of "X" between two sub-exons if they are never observed together in any transcript. Both the constrained genomic order and the padding help disentangle orthology from paralogy relationships. For instance, the similar mutually exclusive exonic sequences from human and gorilla shown in Figure 1C will be separated and assigned to two s-exons (see also **Supplemental Fig. 2C**). To align the sequences, we use the graph-based progressive alignment method ProGraphMSA[4]. We chose this method because its graph-based representation of protein sequences allows recording the whole history of indel events along the guide tree. This framework proved better suited to deal with AS-induced insertions and deletions than classical progressive alignment methods[4].

Algorithm 1: s-exon identification

This algorithm identifies a set s of s-exons as contiguous blocks in an input MSA msa, where the letters indicate to which sub-exon each residue belongs to. For instance, if $msa[i, j] = "a"$, then it means that the residue at position i in the MSA and coming from sequence j belongs to the sub-exon "a". Whenever there is a change in sub-exon, the algorithm define a new s-exon.

$s \leftarrow \{\}$
start $\leftarrow 0$
for $i = 1$ to $L - 1$ **do**
    if $start = 0$ **then**
        $start \leftarrow i$
    end
    $c \leftarrow 0$
    $has2stop \leftarrow$ False
    $j \leftarrow 1$
    while $j \leq n$ and not $has2stop$ **do**
        if $msa[i, j] \neq$ " $-$ " and $msa[i + 1, j] \neq$ " $-$ " **then**
            if $msa[i, j] \neq msa[i + 1, j]$ **then**
                $has2stop \leftarrow$ True
            else
                $c \leftarrow c + 1$
            end
        end
    end
    if $has2stop$ or $c = 0$ **then**
        $stop \leftarrow i$
        $s \leftarrow s||\{[start, stop]\}$
        $start \leftarrow 0$
    end
end
if $start = 0$ **then**
    $start \leftarrow L$
end
$s \leftarrow s||\{[start, L]\}$

where L and n are the numbers of positions and sequences, respectively, in the MSA.

Heuristics to refine the s-exons

By default, we consider that a sub-exon is poorly aligned if it shares less than 30% sequence identity with all the other sequences against which it is aligned. ThorAxe algorithm removes each of these problematic sub-exons from their MSA and aligns it to sequences from the other clusters. If ThorAxe finds a better match, it rescues the sub-exon, otherwise it creates a new species-specific s-exon containing only the sub-exon sequence. Some other problematic s-exons are the very small ones, comprising only 1 or 2 columns. Indeed, they typically arise from inconsistencies between the sub-exon boundaries in the different species of the MSA (**Supplemental Fig. S3**). To minimise such inconsistencies, ThorAxe algorithm shifts the very small sequence stretches isolated by gaps to re-group sub-exons (**Supplemental Fig. S3**, bottom left panel). Moreover, it disintegrates the 1-column s-exons by re-assigning their residues to the neighbouring s-exons, whenever this is consistent with sub-exon boundaries (**Supplemental Fig. S3**, right panels).

Conservation measures computed on the ESGs

For a node v in an ESG, let us denote as $n(v)$ the number of species where v is present, $nt(v)$ the number of transcripts containing v, nt_s the number of transcripts from species s, $nt_s(v)$ the number of transcripts from species s containing v, and nt the total number of input transcripts ($nt = \sum_s nt_s$). We consider three conservation measures:

- the *species fraction*, $F(v) = \frac{n(v)}{n}$
- the *transcript fraction*, $TF(v) = \frac{nt(v)}{nt}$
- *averaged transcript fraction*, $ATF(v) = \frac{1}{n} \sum_s \frac{nt_s(v)}{nt}$.

These measures can be expressed in the same way for a given edge e, or for a given path in the graph. The *species fraction* indicates in how many species a node/edge/path is observed, the *transcript fraction* indicates in how many of the input transcripts the node/edge/path is included and the *averaged transcript fraction* reflects the average transcript usage of the node/edge/path.

Canonical transcript identification

To decide how to choose the canonical transcript isoform, we considered nine measures:

- minimum averaged transcript fraction, $\min_{e \in t} ATF(e)$,
- minimum transcript fraction, $\min_{e \in t} TF(e)$,
- minimum species fraction, $\min_{e \in t} F(e)$,
- averaged transcript fraction sum, $\sum_{e \in t} ATF(e)$,
- transcript fraction sum, $\sum_{e \in t} TF(e)$,
- species fraction sum, $\sum_{e \in t} F(e)$,
- averaged transcript fraction mean, $\frac{1}{ne(t)} \sum_{e \in t} ATF(e)$,
- transcript fraction mean, $\frac{1}{ne(t)} \sum_{e \in t} TF(e)$,
- species fraction mean, $\frac{1}{ne(t)} \sum_{e \in t} F(e)$.

where $ne(t)$ the number of edges in transcript t. We systematically tested all the possible permutations of pairs, triplets and quadruplets of these measures to rank the transcripts. To evaluate the quality of a strategy, we computed the ratios of species fraction and of averaged transcript fraction between the identified canonical transcripts and the other (alternative) transcripts, and we counted the number of times these ratios were above 1. The rationale behind this choice was to minimize the number of alternative transcripts that are more conserved or more representative than the canonical transcript. We found that the best ordered combination was:

1. minimum species fraction, $\min_{e \in t} F(e)$,
2. minimum averaged transcript fraction, $\min_{e \in t} ATF(e)$,
3. averaged transcript fraction sum, $\frac{1}{ne(t)} \sum_{e \in t} ATF(e)$.

It may happen that several transcripts are equally ranked first. In that case, we retain the longest (in residues) and highest-quality (based on TSL) one. The vast majority of the canonical transcripts identified by ThorAxe on the curated set of 50 genes are annotated as "principal" by APPRIS[5] (**Supplemental Fig. S24**). This shows that ThorAxe definition of canonical transcripts agrees well with APPRIS assessment. Nevertheless, some transcripts are not annotated as "principal" or "alternative" by APPRIS although they correspond to very well conserved paths in the ESGs (*e.g.* the transcript ENSDART00000183745 from *TPM3* in zebrafish).

Algorithm 2: event detection

This algorithm detects AS events as variations displayed by a set of input transcripts with respect to a reference canonical transcript c. It enumerates all the pairs of maximal subpaths that do not share any s-exon, where one subpath comes from the canonical transcript and the other one from some input transcript

forEach input transcript t **do**
  $i \leftarrow 2$
  $j \leftarrow 2$
  while $i \leq ne(t)$ **do**
    if $v_t^i \neq v_c^j$ **then**
      find k and l such that $v_t^k = v_c^l$
      $event \leftarrow ([v_c^{(j-1)} : v_c^l],[v_t^{(i-1)} : v_t^k])$
      add $event$ to the list of detected AS events
      $i \leftarrow k + 1$
      $j \leftarrow l + 1$
    end
  end
end

where $ne(t)$ is the number of s-exons in t (or nodes in the path defining t), and v_t^i is the ith s-exon of t (or the ith node in the path defining t). Note that the first and last nodes in each transcript path are the *start* and the *stop*. By default, the events are detected on a reduced version of the ESG, where the edges supported by only one transcript have been removed.

Computational details

ThorAxe v0.6.3 was run on every human protein-coding gene using the following command:

```
thoraxe -i $protein -o $protein -y --plot_chimerics -l $sp
```

where the variable *protein* stores the name of the query gene and *sp* stores the list of species considered. The calculation over the whole proteome completed in 240 hours single-core and about 19 hours using Julia to parallelize the dataset in 15 cores and with the WSL2 of Windows 10, on an Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz. On average, each gene was treated in 56 seconds.
The command used to run ThorAxe without the clustering step was:

```
thoraxe -i $protein -o $protein --coverage 0.0 --identity 0.0 -y --plot_chimerics -l $sp
```

The command used to bypass both the clustering step and the refinement step was:

```
thoraxe -i $protein -o $protein --coverage 0.0 --identity 0.0 -y --plot_chimerics -l $sp\
 --no_movements --no\_disintegration
```

Calculation of MSA scores

To assess the quality of the MSAs associated to the s-exons, we computed a sum-of-pair score, with $\sigma_{match} = 1$, $\sigma_{mismatch} = -0.5$ and $\sigma_{gap} = 0$. This MSA scoring is not part of ThorAxe and we used it *a posteriori* to judge the quality of ThorAxe results. We chose not to penalise gaps such that if there are sub-sequences in the s-exon that are completely missing from some species, it will not reflect badly on the MSA quality. We normalised the raw sum-of-pair scores by dividing them by the maximum expected values. Hence, the final score of the s-exon represented by the node $v \in \mathcal{V}$ is expressed as,

$$\sigma_{rel}(v) = \frac{\sigma(v)}{\binom{n(v)}{2} l(v)}, \tag{1}$$

where $\sigma(v)$ is the raw MSA score, $n(v)$ is the number of species where the s-exon is present and $l(v)$ is the length of the s-exon, computed as,

$$l(v) = \max_{s \in S} \sum_{i=1}^{\ell_s} \mathbb{I}\{s_i \in \mathcal{A}\}, \tag{2}$$

where S is the set of sequences comprised in the MSA, ℓ_s is the length of the aligned sequence s, and $\mathcal{A}$ is the 20-letter amino acid alphabet (*e.g.*, excluding gap characters).

Detection of similar pairs of s-exons

To systematically identify protein sequences sharing some similarity, we performed an all-to-all comparison of the s-exons identified by ThorAxe for each human protein coding gene. Specifically, we created a profile hidden Markov model (HMM) from each s-exon MSA using *hhmake* from the HH-suite[6]. No filtering was applied on the s-exon sequences, and the maximum proportion of gaps in a MSA column to be considered for match states was set to 50%. For each gene, we globally aligned each HMM against all the others using *hhalign*. We considered two s-exons to be similar if the p-value associated to their HMM alignment was lower than 0.001. In total, we detected 150 020 s-exon pairs (2% of all possible pairs) found in 10 814 genes. The median sequence identity between the s-exon consensus sequences is of 36% (**Supplemental Fig. S25**). We further restricted this set to s-exons present in more than one species and involved in at least an event. In addition, we removed the s-exon pairs detected based on HMM alignments smaller than 5 positions, and the pairs where the two s-exons do not have any species in common. Finally, we excluded the pairs where none of the s-exons is included in the canonical transcript, and those where the two s-exons always co-occur in the same subpath (canonical or alternative). Indeed, these pairs do not inform us about the differential usage of similar sequences. These filters reduced the number of s-exon pairs to 31 031 and the number of genes to 2 190 (**Supplemental Fig. S25**). Among those, we retrieved 90% of a previously reported set of 97 genes comprising pairs of mutually exclusive homologous exons[7]. Four of the missing genes (ACSL6, PPAP2A, U2AF1, UGT1A8) were not in the set of 18 226 human protein coding genes treated by ThorAxe. One of them (CYP4F3) did not display any event. In five other genes (H2AFY, HNRNPK, ITGA3, RBM4, SLC39A14), no significant similarity was detected between the mutually exclusive sequences reported in[7].

RNA-seq analysis

We performed a mapping of RNA-seq splice junctions onto the ESGs computed by ThorAxe to complement the annotations from Ensembl and to get insight into AS tissue regulation. We used the Bgee database[8] to select a relevant set of 37 RNA-seq experiments over our 12 species of interest. The

corresponding raw sequences were then downloaded from the Sequence Read Archive (SRA-NCBI)[9] (total: 877 libraries), and aligned to their respective genome versions with the STAR aligner[10]. The resulting BAM files were used to update the reference gene annotation of some species (see below). Whippet[11] was then run on the raw sequences for event quantification.

Data preparation

We assessed read libraries strandedness and quality with the RSeQC package[12]. To obtain the strand information from the BAM files we used the function *infer_experiment*. We determined the mapping quality with the function *bam_stats* and kept only the files containing more than 75% of high quality mapped reads. The selected libraries were then merged for all species and experiments using SAMtools[13].

Whippet annotations

Whippet analysis unfolds in two steps. First, it builds a species-specific index using GTF annotations and, optionally, BAM files. We used both types of files for macaque, rat, cow, platypus, frog, and zebrafish. For human, gorilla, boar, opossum and nematode, we used only GTF annotations. Indeed, the BAM files of gorilla, boar and nematode were of low quality and/or non-stranded. Hence, following Whippet developers' recommandations, we did not use them for indexing. Moreover, another recommandation is that BAM files associated with well-annotated genomes should not be considered. This excluded the BAM files of human and mouse. Finally, the chromosomes of opossum were too long to be handled by SAMtools, and hence its BAM files were also ignored. Secondly, Whippet detects and quantifies the events supported by the reads. Note that it does the quantification without aligning the reads. We used the function *whippet-quant.jl* for computing Percent-Spliced In (PSI) values[14;15]. The function is suitable for treating both single-end and paired-end reads.

Extraction of splice junctions

The splice junctions obtained from Whippet were mapped on the ESGs computed by ThorAxe using genomic coordinates and strand information. For each splice junction, we computed the normalised sum of mapped reads to cope with sequencing depth variations between the different experiments.

Mapping between ThorAxe and Whippet nodes

We created a mapping between the s-exons identified by ThorAxe and the exonic regions represented by the nodes in Whippet. Let us remind that ThorAxe s-exons are defined across species while Whippet nodes are species-specific. Within each species and for each s-exon, we looked for matching Whippet nodes. We identified three main case scenarios (**Supplemental Fig. S26**): (*i*) one perfect match found, (*ii*) partial match(es) found at the 3'- and/or 5'-end, or (*iii*) no match found. Case (*i*) corresponds to a one-to-one non-ambiguous mapping. For case (*ii*), if two partial matches were found, one at each end, we defined a one-to-many mapping for the s-exon. If only one partial match was found, we inferred an overlap between a Whippet node and one of the s-exon extremity. Then, if the matching Whippet node was identified as a partial match for another s-exon at the other extremity, we inferred a many-to-one mapping. Otherwise, it meant a Whippet node corresponded to some s-exon(s) and some contiguous unannotated region. For dealing with case (*iii*), we re-ran the search with relaxed matching constraints. Specifically, if the s-exon was longer than 100 bp, its start and end coordinates were allowed to differ by 25% of its length, compared to the Whippet matching node(s). If the s-exon was shorter, then the allowed variation was set to 33% of the s-exon length.

Finally, the PSI value of a s-exon in a given tissue from a given species was computed as the mean PSI value of its matching Whippet nodes.

Splice graph annotation and events quantification

An edge in the ESG was considered as supported by RNA-seq data in a given species if the normalised mean of mapped reads computed for the corresponding junction was higher than $1e-07$. For the nodes, we set up a PSI threshold of 0.05 (**Supplemental Fig. S27A**). In many cases, Whippet nodes were associated with undefined PSI values. We considered the corresponding s-exons as unsupported by RNA-seq data. The PSI values for the canonical and alternative paths defining the events detected in the curated set and documented in the literature were retrieved from Whippet output using the map we defined between ThorAxe s-exons and Whippet nodes. We considered that an event was tissue-regulated when the PSI difference between the canonical and alternative paths was higher than 0.15 in some tissue in at least two species (**Supplemental Fig. S27B**). If only one of the two paths was present in a tissue, we asked for its PSI value to be above 0.55 (**Supplemental Fig. S27C**). The tissue ontology is described in **Supplemental Table S10**.

3D structural analysis

The 3D structural templates were searched and aligned using our new iterative and s-exon-centred version of PhyloSofS molecular modeling routine[16]. They were visualised using Pymol[17]. We used DisEmbl citeplinding2003protein to detect intrinsically disordered s-exons. Specifically, we considered that a s-exon was disordered if it contained more than 75% of positions predicted as disordered, on average, over the sequences in its MSA.

Gene Ontology analysis

The Gene Ontology (GO) annotations for the human proteome were downloaded from the GO Consortium online ressource[18;19]. We focused on the subset of labels of type "cellular components". For each label and each gene class, MEX, ALT, REL, UNREL or NO (comprised of the protein-coding genes not included in the other classes), we computed a p-value using a two-sided hypergeometric test[20]. We considered that a label was significantly enriched or depleted in a gene class if the p-value was lower than $1e^{-5}$. Generic or vague labels such as "cell" and those containing "organelle" and "part" were excluded from the analysis.

Comparison with phastCons scores

We downloaded the human genome Conservation annotation track from the UCSC Genome Browser (`https://genome.ucsc.edu`). It gives the phastCons score of each base pair in the human genome computed by the PHAST package[21;22] from the multiple alignments with 99 vertebrate genomes. phastCons is a hidden Markov model-based method estimating the probability of each nucleotide to belong to a conserved element. It considers not only the column corresponding to the nucleotide of interest in the alignment but also its flanking columns. The phastCons scores range between 0 and 1. We converted them into residue-based scores by taking the maximum value computed over the 3 nucleotides encoding each residue (<3 for residues overlapping with exon boundaries). Since phastCons is sensitive to "runs" of conserved sites, we restricted the analysis to s-exons longer than 10 residues. In total, we treated 199,916 s-exons covering 10,060,639 residue positions.

Comparison with other studies

The study reported in[23] considered only one-to-one orthologous genes, as annotated in Ensembl, similarly to what we did. The authors detected orthologous exons by converting the genomic coordinates between genomes using LiftOver. They analyzed 3 013 alternatively spliced orthologous exons from human, mouse, chicken and frog. Among those, we extracted the 41 exons reported in Figure S11B from[23] as displaying regulation patterns reflecting organ type. We retrieved their genomic coordinates from the UCSC Genome Browser (`https://genome.ucsc.edu`), and we identified the s-exons defined by ThorAxe that overlapped with these coordinates. This strategy was successful for 34 exons. For the remaining ones, we manually looked for their amino acid sequences in Ensembl and then identified the s-exons defined by ThorAxe that overlapped with these sequences. In total, 40 out of the 41 exons reported in[23] mapped to 49 s-exons and 38 were involved in an event supported by at least 2 transcripts. The study reported in[24] combined multiple genome alignments of 19 amniota retrieved from Ensembl and transcript prediction from Cufflinks to detect orthologous exons. They identified around 48 000 exons with clear orthologs in chicken and at least two mammals, and extracted a set of around 500 well conserved exons with highly tissue-specific splicing patterns. Using LiftOver[25], we converted the genomic coordinates provided in[24] for these exons to match the genome versions we used for macaque, cow, rat and mouse. We could map 323 reported exons (mostly coming from mouse) to 430 s-exons identified by ThorAxe and involved in 277 events.

Comparison with other methods

We bypassed ThorAxe clustering step and/or refinement step by directly modifying the tool's command line arguments. To compare ThorAxe with the RBBH method, we performed an all-to-all comparison between species. Specifically, given two species s_1 and s_2, we ran BLAST for each sub-exon defined by ThorAxe in s_1 against the ensemble of sub-exons from s_2, and reciprocally, and we identified the best reciprocal pairs of s-exons across the two species. Defining s-exons from a set of pairwise alignments of sub-exons is a difficult task, and thus we simply compared the sub-exon pairs identified by RBBH with those implied by ThorAxe s-exon definition.

Supplemental Table S1: **Curated set**

Main gene	Auxiliary genes	#(species)	Function
BCL2L1	-	7	regulates outer mitochondrial membrane channel opening, and hence apoptosis
CAMK2B	*CAMK2A,D,G*	4-10	plays multiple unique roles in actin assembly (organisation, stabilisation, polymerization...)
DNM2	*DNM1,3*	8-10	GTPase involved in membrane remodelling, engaged in many protein-protein interactions
FMR1	*FXR1,2*	9-11	multifunctional polyribosome-associated RNA-binding protein
FYN	*FGR, SRC, YES1*	10-11	tyrosine kinase involved in T-cell and neuronal signaling
GRIN1	*GRIN2A,2B,2C,2D,3A,3B*	6-11	glutamate and ion channel protein receptor activated when glycine and glutamate bind to it
KIF1B	*KIF1A,C*	10	microtubule-dependent motor protein involved in cellular trafficking
MAPK8	*MAPK9,10*	9-11	serine/threonine kinase involved in many essential signaling pathways
MYH11	*MYL6,9,12B*	6-10	major contractile protein involved in muscle contraction
MYO1B	*MYO1C,D,E,F,G,H*	7-12	links lipid membrane to the actin cytoskeleton, plays roles in membrane trafficking and dynamics
NEBL	-	8	Binds to actin and plays an important role in the assembly of the Z-disk
NXNL2	*NXNL1*	9-10	may be involved in the viability of sensory neurons
PAX6	-	10	key transcription factor involved in eye development
PTPRC	-	9	protein tyrosine phosphatase receptor regulating lymphocytes signalling
SNAP25	*SNAP23*	10-11	Part of the SNARE complex, which is involved in vesicle fusion
TPM1	*TPM2,3,4*	5-7	actin-binding protein, involved in muscle contraction and cytoskeleton formation

For each family, we defined a *main* gene and focused on retrieving information from the literature for that gene. The third column indicates the number of species range where one-to-one orthologs were found, for each family. The function descriptions were taken from Uniprot (`https://www.uniprot.org`) and Wikipedia (`https://www.wikipedia.org`).

Supplemental Table S2: **Documented AS events from the curated set**

	Gene	Event	Function
1	*BCL2L1*	63-aa deletion	function reversion[26] & partner loss[27]
2	*CAMK2A*	11-aa insertion	cellular localisation[28]
3	*CAMK2B*	insertion of 43-aa proline-rich repeats	partner acquisition (?)[29]
4		25-aa deletion	protein partner loss[30]
5	*DNM2*	46-aa mutually exclusive homologous exons	partner specificity[31]
6		4-aa deletion	cellular localisation[31]
7	*FMR1*	21-aa deletion (ex. 12)	RNA-binding affinity[32]
8		17-aa deletion (ex. 14)	cellular localisation & RNA binding[32]
9		13-aa deletion (alt. acc. in ex. 15)	PTM sites & RNA-binding affinity[32]
10		25-aa deletion (alt. acc. in ex. 15)	PTM sites & RNA-binding affinity[32]
11		17-aa deletion (alt. acc in ex. 17)	cellular localisation[32]
12	*FYN*	50-aa mutually exclusive homologous exons	domain organisation (linker)[33]
13	*GRIN1*	21-aa deletion (ex. 5)	ligand binding affinity & regulation[34;35]
14		alternative end	ligand binding affinity & regulation[34;35]
15		37-aa deletion (ex. 21)	ligand binding affinity & regulation[34;35]
16	*KIF1B*	490 aas (1 ex.) replaced by 1100 aas (27 ex.)	partner binding specificity[36]
17		83-aa deletion	partner binding specificity[36]
18		6-aa deletion in the N-domain	binding affinity & dimerisation[36]
19		40-aa deletion in the N-domain	binding affinity & dimerisation[36]
20	*MAPK8*	24-aa mutually exclusive homologous exons	substrate selectivity[37]
21	*MYH11*	7-aa insertion	function regulation[38]
22	*MYO1B*	25-aa insertions and variations	protein binding affinity[39]
23	*NEBL*	12 nebulin repeats replaced by a LIM domain	partner selectivity[40]
24	*NXNL2*	thioredoxin domain present/absent	novel function in signaling[41]
25	*PAX6*	14-aa insertion	DNA-binding specificity[42]
26		paired-domain deletion	DNA-binding specificity[42]
27	*PTPRC*	up to 3 exon-deletions	immunological recognition[43]
28	*SNAP25*	34-aa mutually exclusive homologous exons	partner specificity (?)[44]
29	*TPM1*	26-aa mutually exclusive homologous exons	partner affinity and dissociation[45]
30		3 ex. deletion coupled with replacement	partner affinity and dissociation[45]

The symbol "?" in the last column indicates that the functional annotation of the ASE remains speculative. The row colors indicate the level of evidence from gene annotations and RNA-seq splice junctions. Green: supported by both the annotations and the RNA-seq data, and tissue-regulated. Light green: supported by both the annotations and the RNA-seq data, without any evidence of tissue regulation. Light blue: detected only in the annotations. Grey: not detected. See Supplemental Table S3 for more detailed information.

Supplemental Table S3: **Detection of the documented AS events by ThorAxe**

	Gene	**Number of species**	**Event rank**	**Number of s-exons**	SF		ATF		SF^{RNASeq}		tissue regulation
					can	*alt*	*can*	*alt*	*can*	*alt*	
1	*BCL2L1*	7	1	2	100	29	69	12	29	29	
2	*CAMK2A*	10	1	1	100	70	51	28	70	70	✓
3	*CAMK2B*	10	2	4	100	70	46	22	80	50	✓
4			3	2	50	40	25	14	50	40	✓
5	*DNM2*	10	2	2/2	100	60	77	17	100	60	✓
6			1	1	100	80	57	32	100	100	✓
7	*FMR1*	11	1	1	82	64	42	24	73	73	✓
8			-	-	-	-	-	-	-	-	
9			-	-	-	-	-	-	-	-	
10			3	2	91	27	66	8	82	45	✓
11			5	1	73	45	45	30	36	36	✓
12	*FYN*	11	1	1/1	100	73	52	32	91	73	✓
13	*GRIN1*	10	1	1	90	70	62	28	90	90	✓
14			2	2/1	60	60	27	21	20	0	
15			3	3	50	40	23	15	10	0	
16	*KIF1B*	10	5	37/1	40	60	15	18	0	30	
17			3	1	100	40	61	9	80	80	✓
18			1	1	80	70	51	32	70	70	✓
19			4	2	40	40	17	14	40	40	✓
20	*MAPK8*	9	1	1/2	89	89	38	36	89	78	✓
21	*MYH11*	10	1	1	90	90	44	40	80	70	✓
22	*MYO1B*	9	1	2	89	89	38	23	78	89	✓
23	*NEBL*	8	2	24/4	63	50	19	13	0	0	
24	*NXNL2*	10	1	1/1	89	22	80	15	10	0	
25	*PAX6*	10	1	1	70	70	25	31	50	50	✓
26			4	4	30	20	12	13	0	0	
27	*PTPRC*	9	1	3	33	33	11	7	11	22	
28	*SNAP25*	11	1	1/1	82	64	48	34	82	64	✓
29	*TPM1*	7	1	1/1	100	100	52	40	86	71	✓
30			2	3/1	86	57	45	26	0	0	

The rank of an event reflects its conservation level with respect to the other events detected for the gene. SF and ATF are the species and averaged transcript fractions (given in percentages) computed by ThorAxe for the canonical (*can*) and alternative (*alt*) paths. SF^{RNASeq} gives the proportion of species where the presence of the canonical or alternative subpath is supported by RNA-seq data. Tissue regulation must be observed for at least one tissue in at least two species to be considered.

Supplemental Table S4: **Per-gene statistics computed over all human protein coding genes**

Variable	Mean	Std	Min	q25	Median	q75	Max
Species	8.5	2.6	1	7	9	10	12
Transcripts	17.1	10.3	1	10	15	22	113
Exons	103.1	102.2	1	31	74	142	1677
Sub-exons	102.3	102.2	1	31	73	140	1685
S-exons	25.8	25.2	1	9	18	35	354

Supplemental Table S5: **Per-species s-exon conservation statistics.** The conservation is measured as the species fraction, *i.e.* the proportion of species contributing sequences to the s-exon.

Species	Species fraction Median	Species fraction Mean	Species fraction Std	Percentage of species-specific s-exons
Human	1.00	0.88	0.20	2.98
Gorilla	0.92	0.85	0.24	5.92
Macaque	0.92	0.85	0.24	6.94
Mouse	1.00	0.90	0.18	2.57
Rat	1.00	0.89	0.20	3.74
Boar	0.92	0.84	0.26	8.46
Cow	1.00	0.88	0.22	5.01
Opossum	1.00	0.85	0.26	8.38
Platypus	1.00	0.85	0.28	10.35
Xenopus	0.91	0.73	0.37	22.88
Zebrafish	0.92	0.85	0.27	9.71
Nematode	0.10	0.33	0.39	72.15

Supplemental Table S6: **List of potential "false negative" s-exons**. This subset was extracted from the 1508 s-exons with low *species fraction* (≤0.3) but high phastCons scores (>0.9) (see top left corner in Fig. 4C). Each s-exon in the subset shares significant sequence similarity with another s-exon coming from the same gene and defined across distinct species. Sequence similarity is measured through an HMM alignment of the s-exons. We ask that the p-value is lower than 0.001, the query s-exon a coverage is higher than 75%, and the *species fraction* (*SF*) computed from the union of the two similar s-exons is higher than 0.3.

Gene	S-exon	phastCons score	Old *SF*	New *SF*
TEAD3	17_1	1	0.25	0.375
FRY	12_0	1	0.25	0.375
KIF9	33_0	0.968	0.22	0.44
SRPK1	1_3	1	0.11	0.33
MIOX	14_0	1	0.09	0.54
KCTD17	12_0	0.996	0.25	0.75
CDIPT	11_0	1	0.27	0.36
ZMIZ1	33_0	1	0.09	0.91
CAMTA2	21_0	1	0.3	0.4
ADGRB3	29_0	1	0.3	0.4
GCC2	3_0	0.995	0.09	0.82
ESCO1	21_0	1	0.27	0.73
TARS2	21_0	1	0.25	0.375
CRELD1	1_1	1	0.3	0.9
KIAA1958	8_0	1	0.27	0.91
PLEKHG5	25_0	0.912	0.22	0.33
ANO5	28_0	0.912	0.22	0.55
FAM120C	14_1	1	0.27	0.36
PCDH9	1_2	1	0.3	0.4
BTBD8	36_0	1	0.11	1
MIER1	23_0	1	0.3	0.4
TBKBP1	12_0	1	0.22	0.33
ERCC6	7_0	1	0.3	1
MRPL46	2_0	1	0.27	0.91

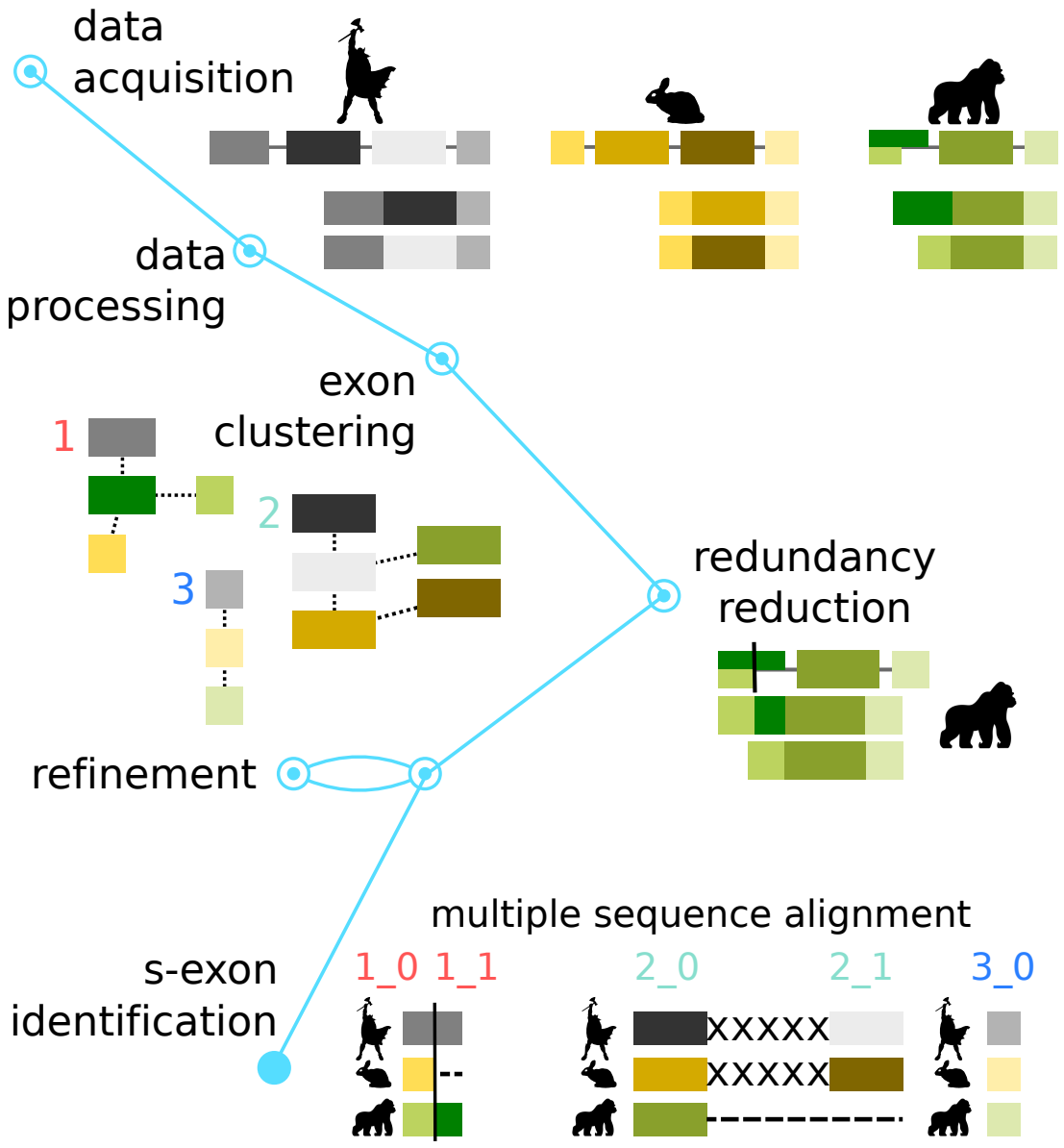
Supplemental Table S7: **List of MEX and ALT s-exons.** See the supplemental data *Supplemental_Table_S7.csv*. The s-exons are grouped by ASE.

Supplemental Table S8: **List of REL s-exons.** See the supplemental data *Supplemental_Table_S8.csv*. The s-exons are grouped by ASE.

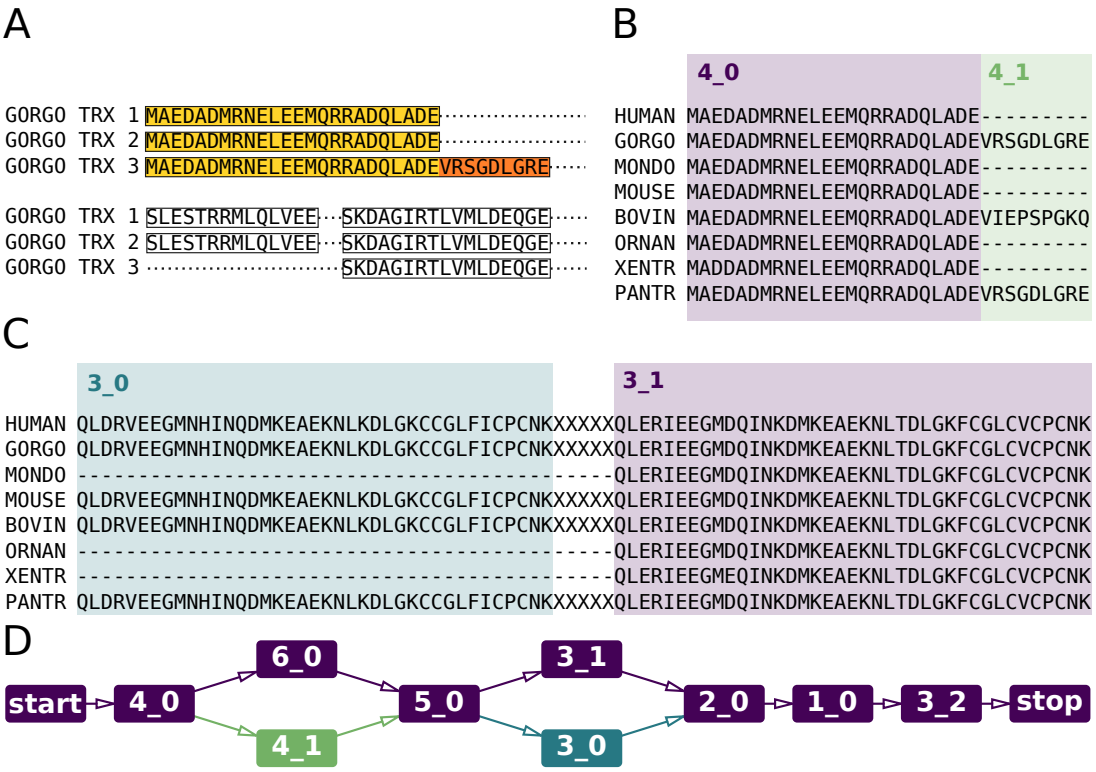
Supplemental Table S9: **List of UNREL s-exons.** See the supplemental data *Supplemental_Table_S9.csv*. The s-exons are grouped by ASE.

Supplemental Table S10: **Tissue ontology**

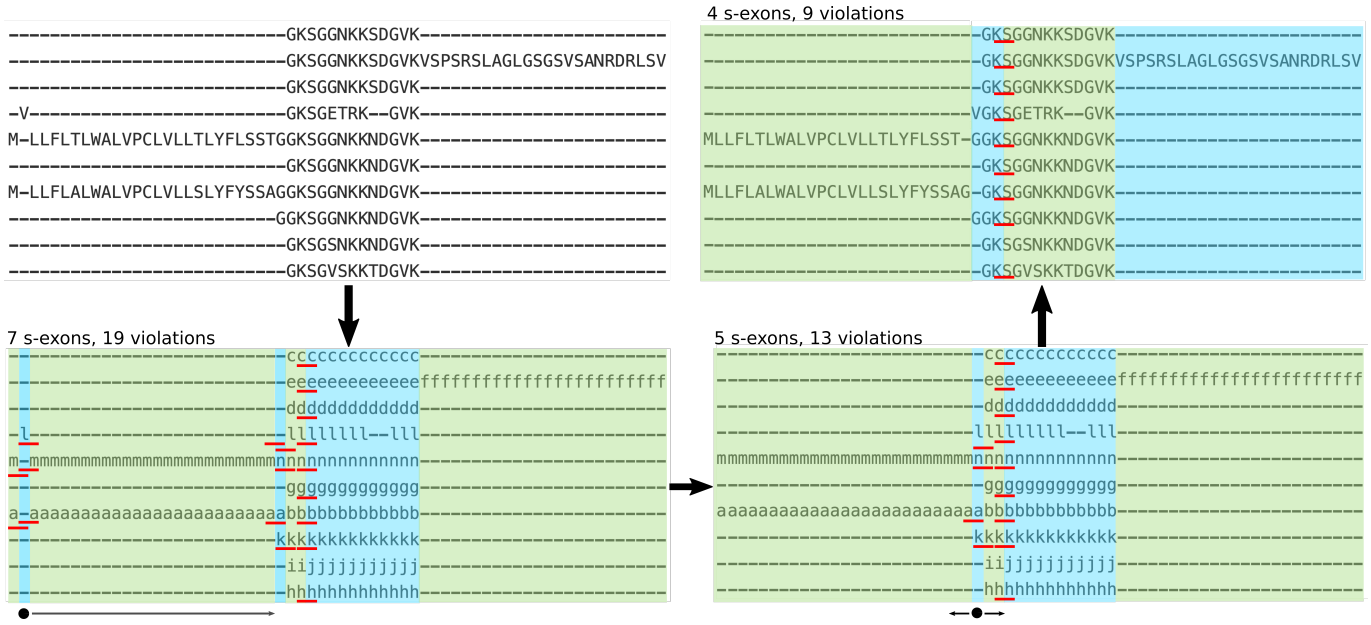
Tissue Group	**Tissue**
Ammon's horn	Ammon's horn
immune	CD4-positive helper T cell leukocyte spleen
immune/thyroid	head kidney lymph node thymus
kidney	adult mammalian kidney
bone	bone tissue
brain	brain cerebellum frontal cortex head prefrontal cortex
digestive	colon intestine liver stomach
embryo	blastula embryo gastrula placenta
eye	eye
female reproduction	female gonad mature ovarian follicle
heart	heart heart left ventricle
kidney	kidney mesonephros
lung	lung
male reproduction	prostate gland testis
multi-cellular organism	multi-cellular organism
muscle	muscle of leg muscle tissue skeletal muscle tissue
pharyngeal gill	pharyngeal gill
thyroid gland	thyroid gland
zone of skin	zone of skin



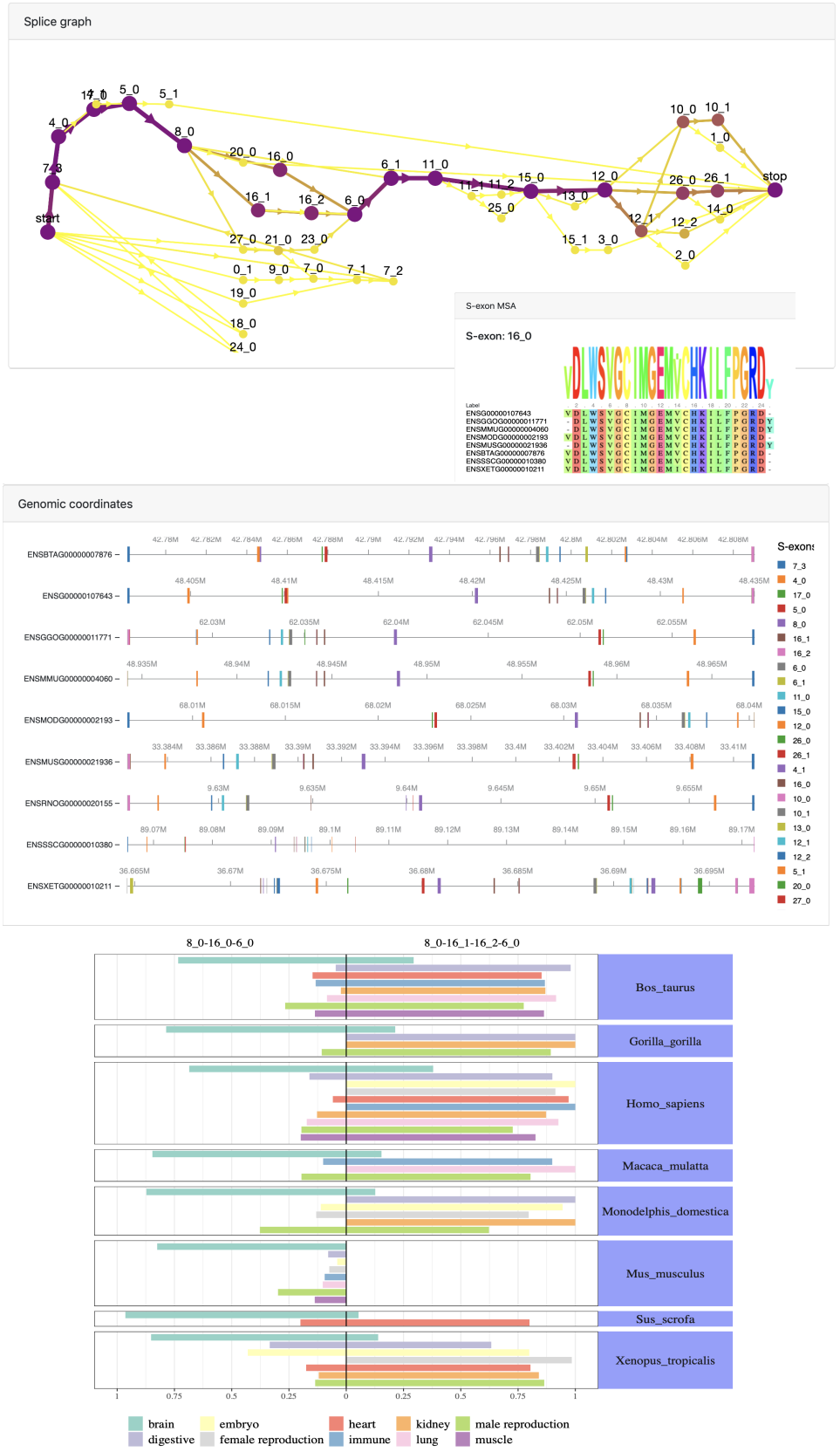
Supplemental Figure S1: **Schematic workflow of ThorAxe pipeline.** On top, the input genes and transcripts are displayed. The exons (grey and colored boxes) are first clustered based on their similarities, then split into sub-exons to account for intra-species variability (redundancy reduction step). Finally, the sequences belonging to each cluster are aligned and blocks in the alignments are identified to output a set of s-exons (*1_0*, *1_1*...).



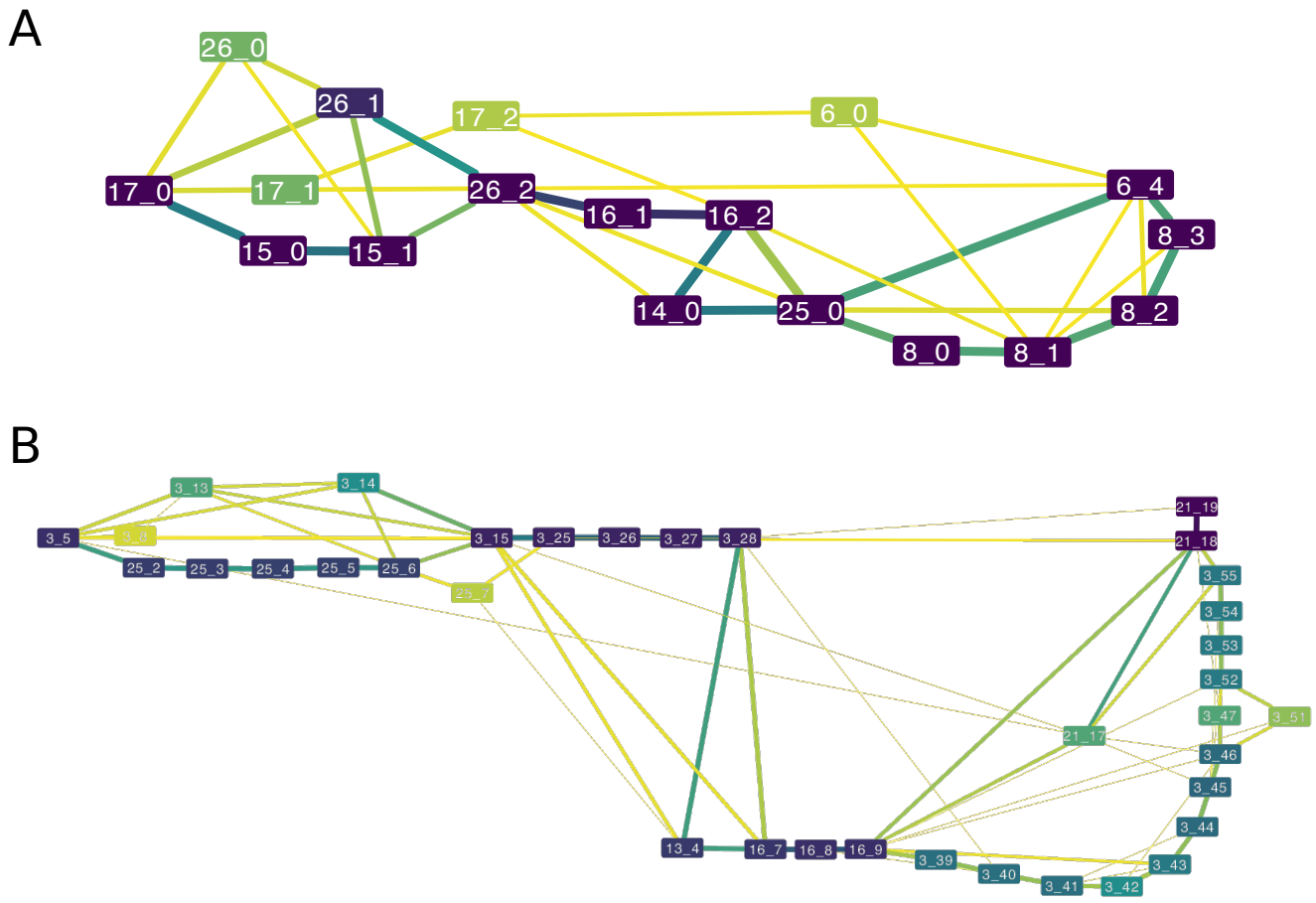
Supplemental Figure S2: **Examples of ThorAxe intermediate and final outputs.** The results were produced for the *SNAP25* gene across 8 species. **A**. Redundancy reduction step illustrated with three transcripts observed in gorilla. From two overlapping exons, one shorter and one longer, we define two sub-exons, highlighted in yellow and orange. **B-C**. Examples of MSAs built to identify the s-exons. The s-exons are highlighted by colored contiguous blocks and labelled. The colors reflect the s-exon conservation, from light green to dark purple. In panel B, the s-exons *4_0* is defined across 8 species while the s-exons *4_1* is defined across 3 species. In panel C, the two s-exons are separated by a padding sequence of "X", indicating that they are mutually exclusive. Note that their sequences are highly similar. **D**. Evolutionary splicing graph. Each node represents a s-exon, and two nodes are linked by an edge if they are consecutive in at least one input transcript. The edges and the nodes are colored according to their conservation level. For ease of visualisation, the species-specific s-exons (*i.e.* defined in only one species) were filtered out.



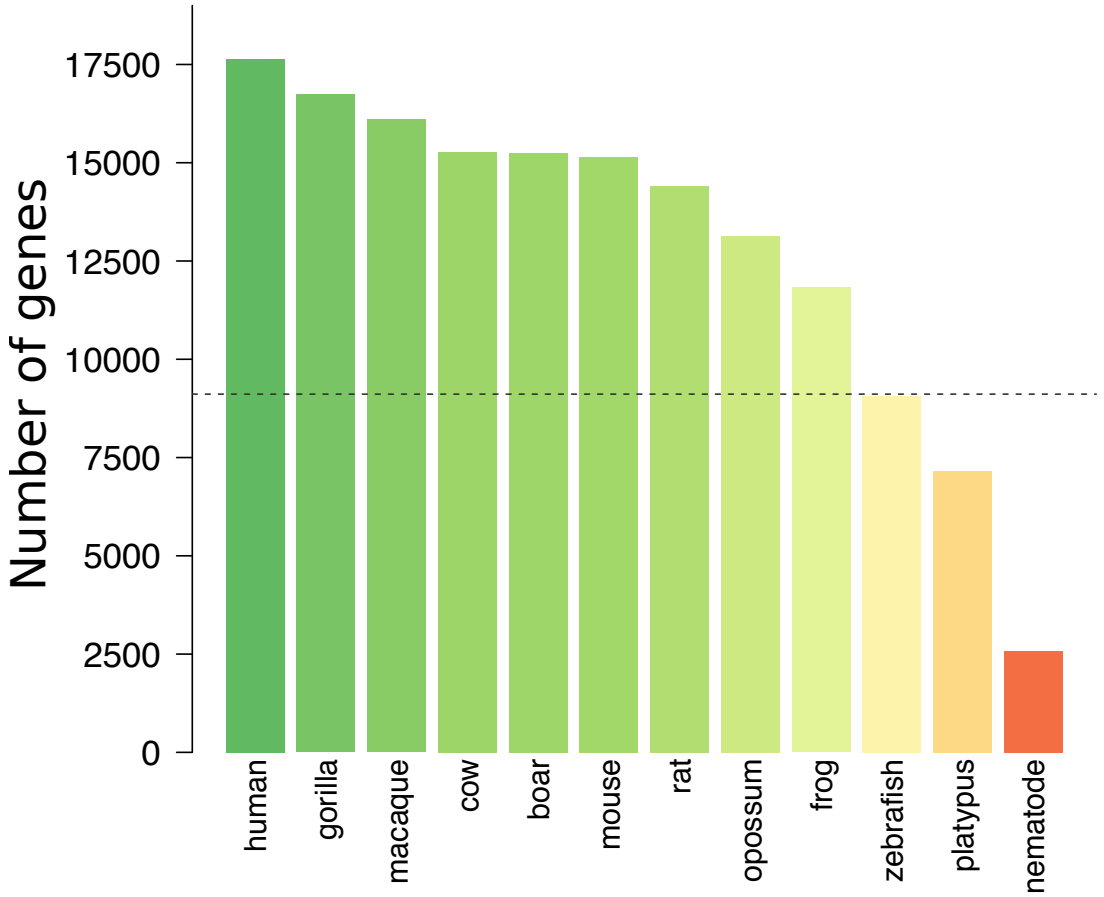
Supplemental Figure S3: **Refinement procedure.** The top left panel shows a chimeric MSA, where each letter corresponds to an amino acid residue. The bottom left panel shows the translation of this MSA into the sub-exon alphabet. In other words, each letter indicates to which sub-exon the residue at this position belongs to. The coloured blocks correspond to the s-exons defined from the MSA. There are 7 s-exons, among which 3 very small ones (1- or 2-column long) arising from inconsistencies between sub-exon boundaries. This is highlighted by the red lines, which indicate the s-exon boundaries splitting the sub-exons. We refer to those cases as *violations*, and there are 19 of them in this MSA. ThorAxe refinement procedure locally optimises the MSA by first moving the 1-residue long stretch "V" (indicated by a black dot) to the right in order to group it with the rest of its sub-exon, denoted "l". The bottom right panel shows that this operation reduces the number of s-exons to 5 and the number of violations to 13. Second, the remaining 1-residue long s-exon (indicated by a black dot) is disintegrated by separating sub-exons "l", "n" and "k" from sub-exon "a". In practice, the "m"- and the "a"- stretches are shifted to the left in the MSA. This leads to the solution displayed in the top right panel, with only 4 s-exons and 9 violations.



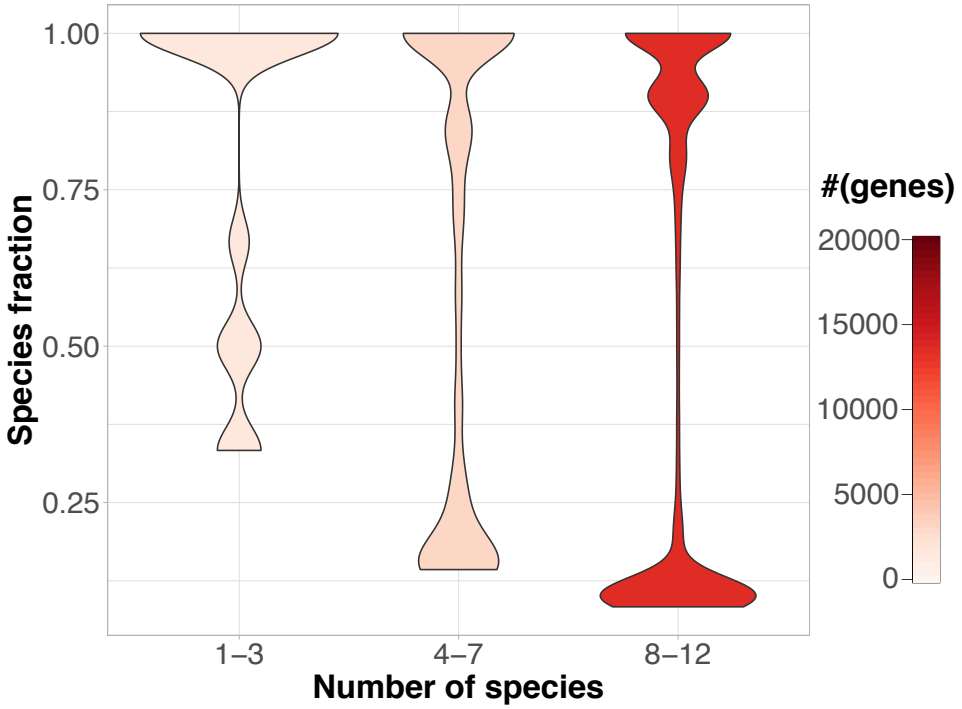
Supplemental Figure S4: **Snapshots from the accompanying website at `http://www.lcqb.upmc.fr/ThorAxe`.** The ESG is on top, with an insert displaying the MSA of a s-exon. The gene structure is in the middle, and a barplot reporting RNA-seq data is at the bottom.



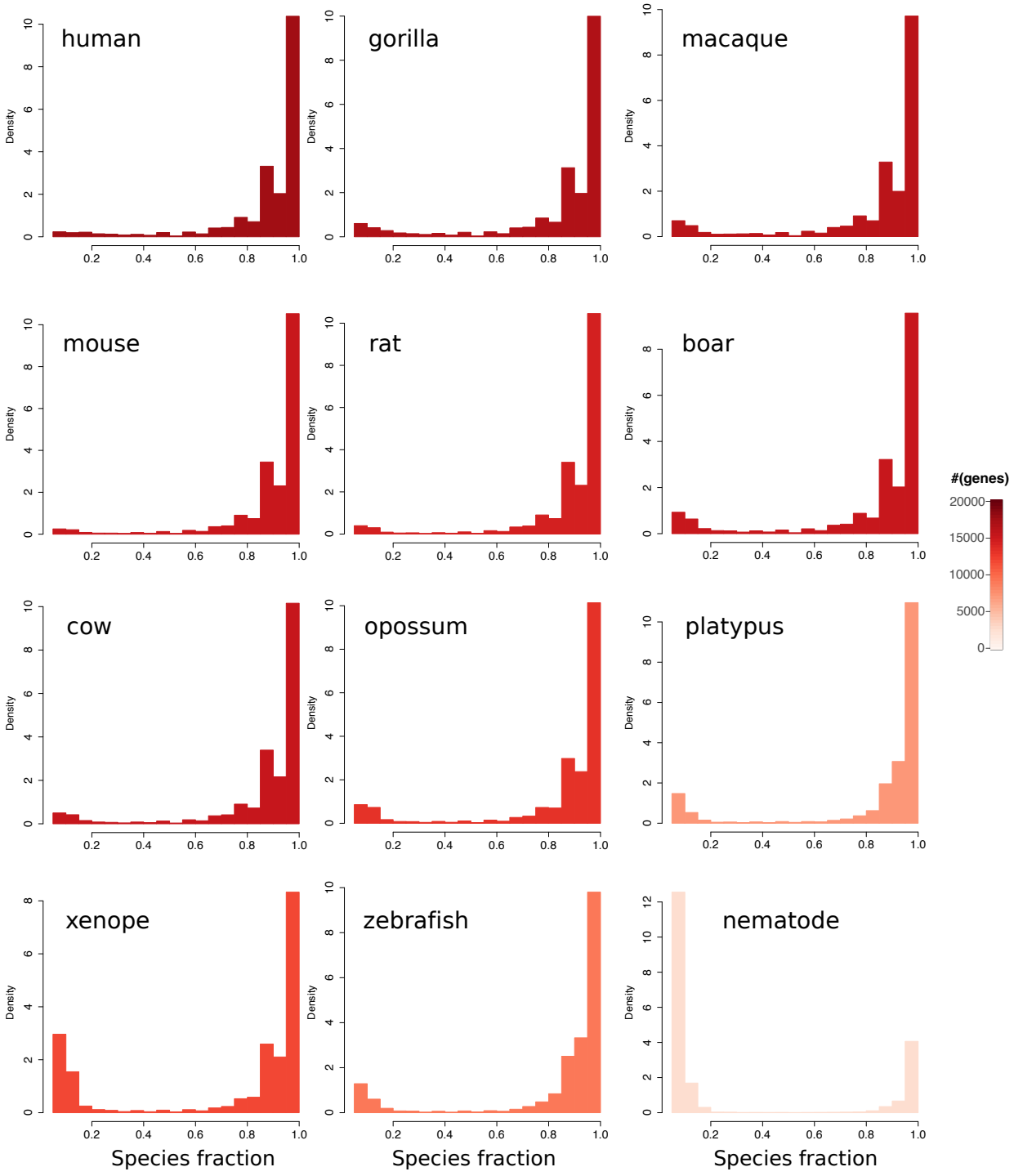
Supplemental Figure S5: *CAMK2B* **linker transcript variability across different species sets.** **A.** For this calculation, we took the initial set of twelve species minus human and mouse. For ease of visualisation, we removed the s-exons present in only one species. **B.** ThorAxe computed this ESG starting from 499 transcripts annotated in 93 species. These are all the species annotated in Ensembl where one-to-one orthologs could be found. For ease of visualisation, we removed the s-exons present in less than 10% of the species. Both ESGs are to be compared with that shown in Figure 3A.



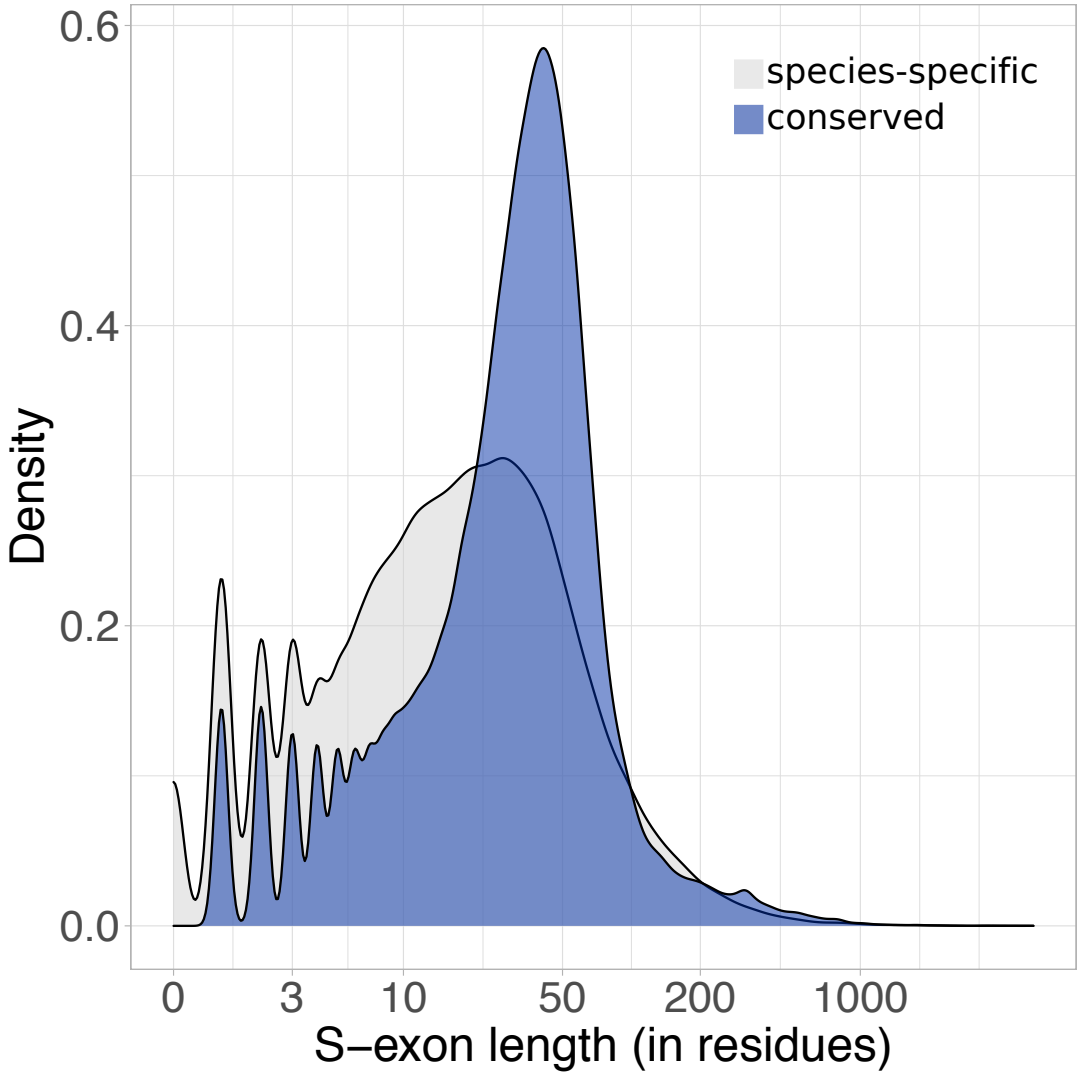
Supplemental Figure S6: **Number of genes within each considered species.** The color code goes from green (many genes) through yellow to red (few genes).



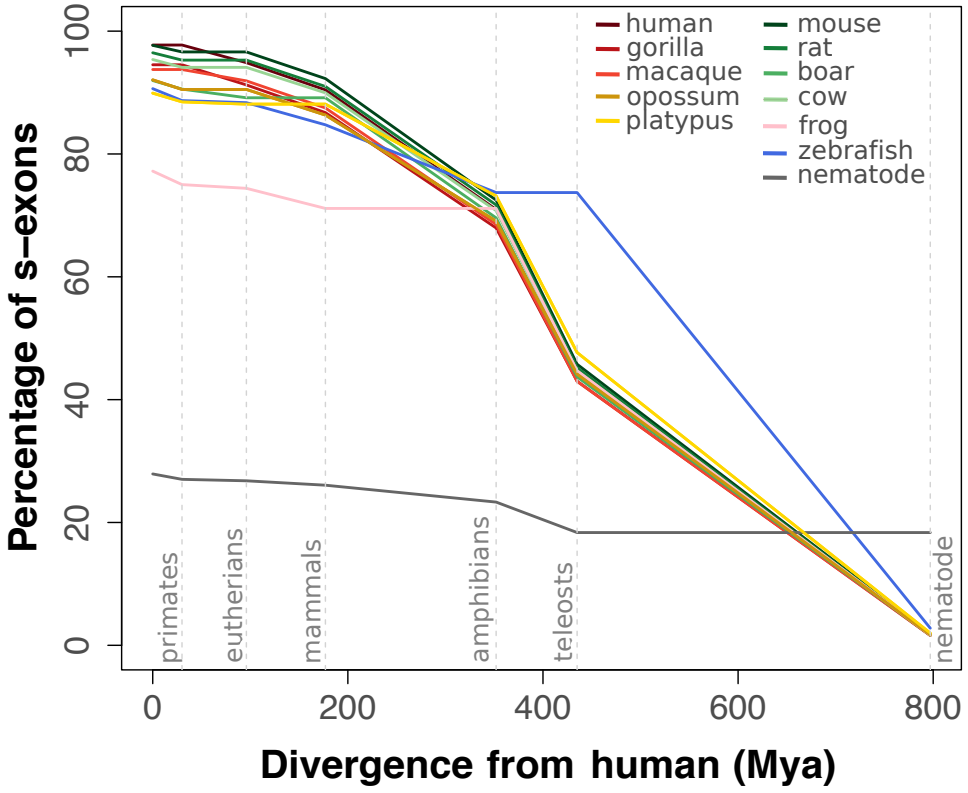
Supplemental Figure S7: **S-exon conservation.** Distributions of the s-exon species fractions depending on the number of species where one-to-one orthologs were found. The red tones indicate the number of genes comprised in each distribution.



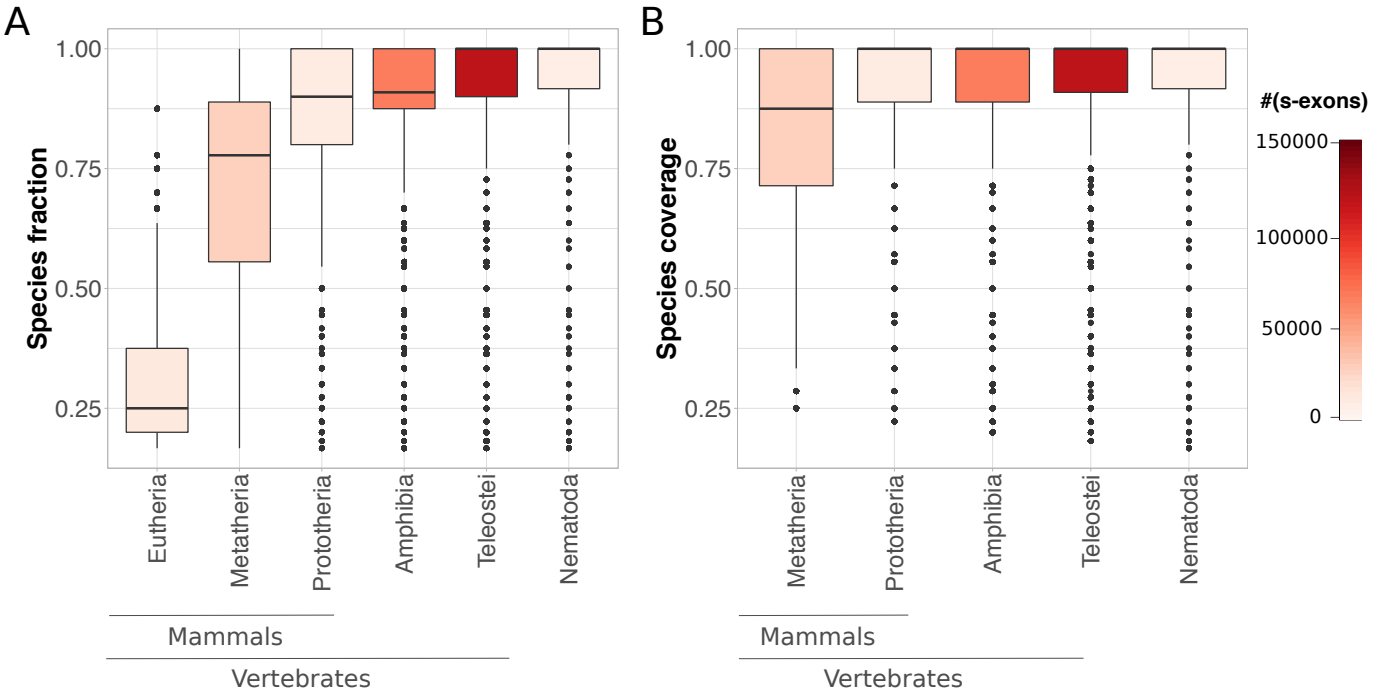
Supplemental Figure S8: **S-exon conservation distributions.** Conservation is measured by the species fraction, *i.e.* the proportion of species contributing sequences to the s-exon. The colors indicate the number of genes within each considered species.



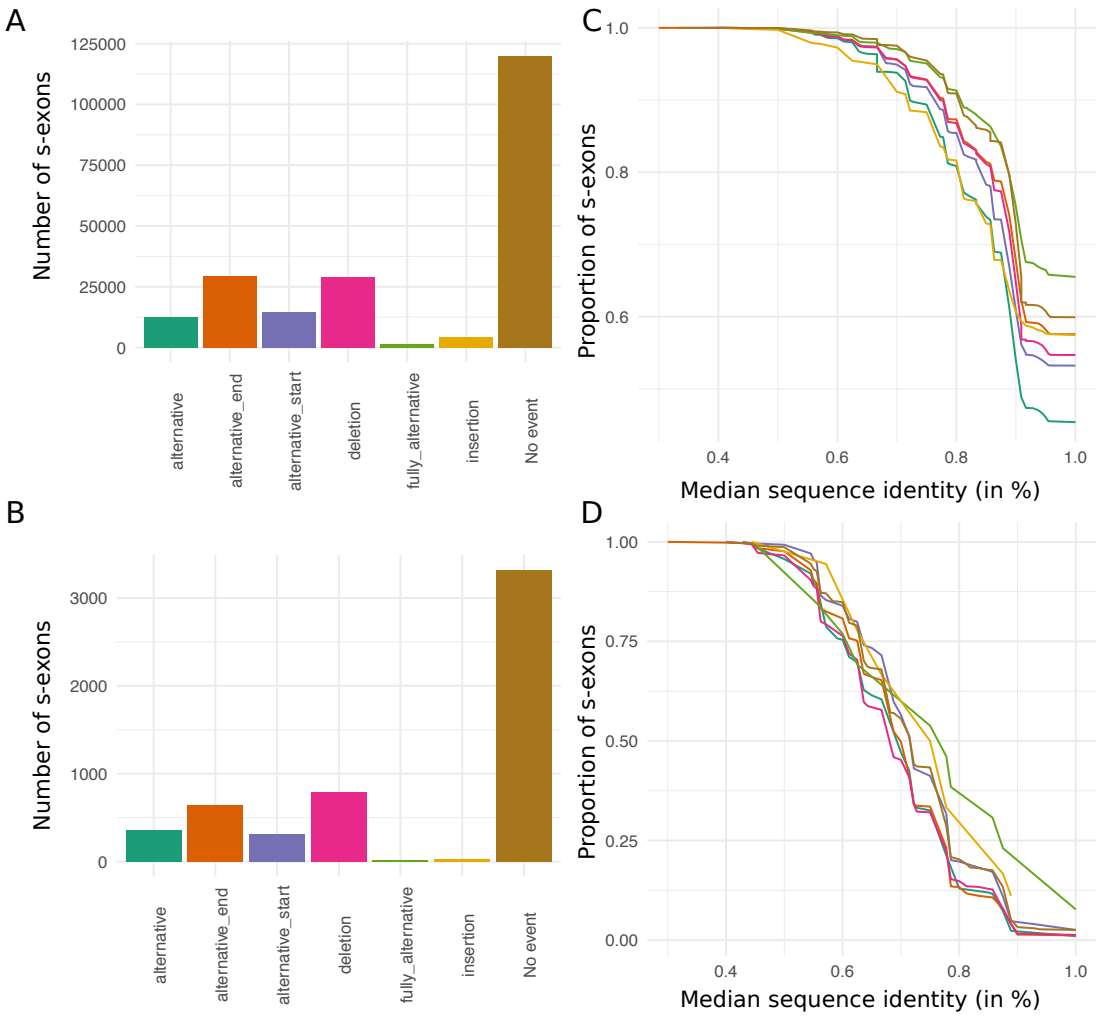
Supplemental Figure S9: **S-exon length.** Distribution densities of the s-exon lengths. The length of a s-exon is the maximum number of amino acid residues determined over the sequences comprised in the MSA (see *Materials and Methods*). The s-exons are classified as either species-specific (only one sequence in the MSA) or conserved (more than one sequence).



Supplemental Figure S10: **S-exon evolutionary conservation.** Percentages of s-exons (y-axis) conserved up to different evolutionary distances (x-axis) from human. For instance, the y-value for the crossing point between the pink curve and the dashed vertical line labelled *eutherians* gives the percentage of s-exons present in frog that are also conserved in at least one primate (among human, gorilla, macaque) and at least one non-primate eutherians (among rat, mouse, boar, cow). The y-values at $x = 0$ give the percentages of s-exons conserved in at least another species. We report values only for the genes with one-to-one orthologs in more than seven species (class *8-12* in Supplemental Fig. S7.

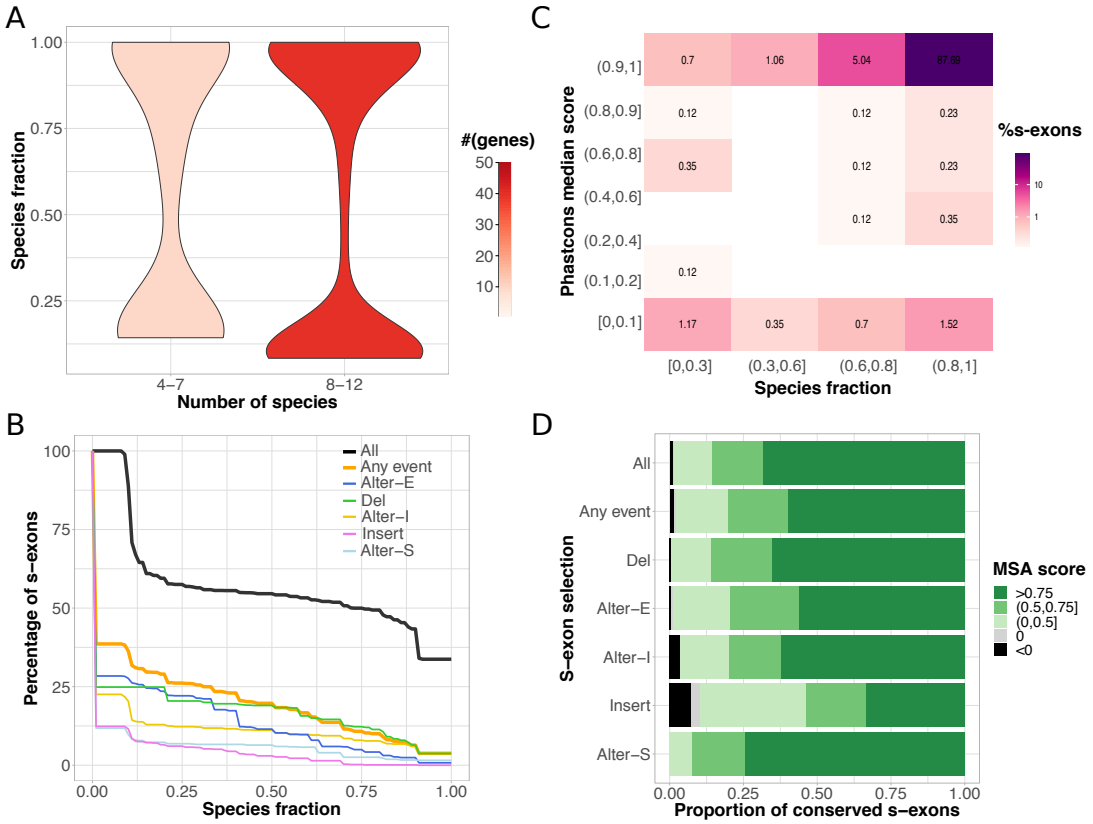


Supplemental Figure S11: **S-exon evolutionary conservation.** We focus on the s-exons conserved in more than one species, and coming from the genes for which one-to-one orthologs could be found in more than seven species (class *8-12* in Supplemental Fig. S7). They are classified in six groups, defined by the evolutionary distances they span. For instance, the s-exons in the "Eutheria" class are observed only in eutherians (human, gorilla, macaque, mouse, rat, boar, cow). Those labeled as "Metatheria" are observed up to oppossum, and not in more distant species, etc... The red tones code for the number of s-exons in each distribution. **(A)** Species fraction distributions. The species fraction gives the overall proportion of species where the s-exon is present. **(B)** Species coverage distributions. Given a s-exon observed in n species, s_n being the most distant one from human, its species coverage is computed as the ratio between $n-1$ and the total number of species less distant than s_n. For instance, a value of one in the class "Teleostei" indicates that the s-exon is present in all mammals, in xenopus (amphibian) and in zebrafish (teleost).

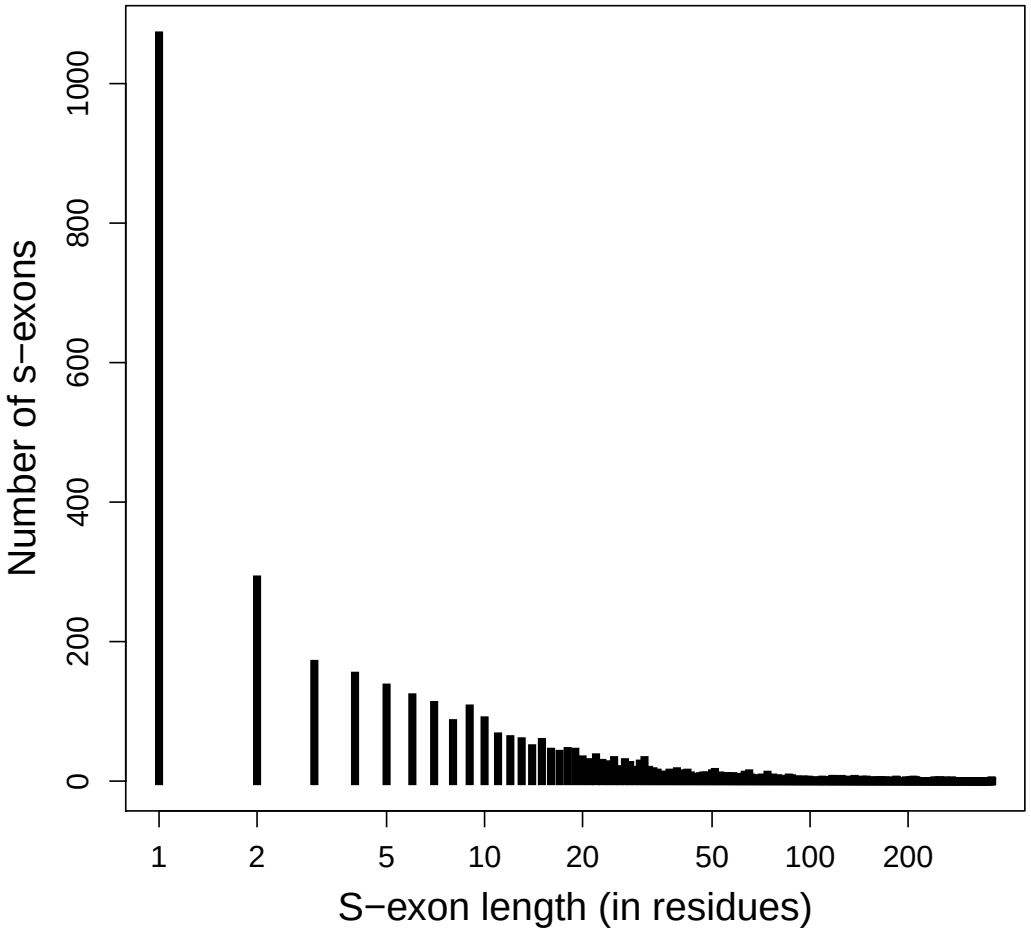


Supplemental Figure S12: **S-exon involvement in ASEs and phastCons scores. (A-B)** Barplots of the numbers of s-exons involved in the different types of ASEs, and those not involved in any ASE. **(C-D)** Cumulative distributions of MSA sequence identity percentage. On the y-axis we report the percentage of s-exons with a median column identity greater than the x-axis value. **(A,C)** All s-exons longer than 10 residues and belonging to genes with one-to-one orthologs in more than seven species. **(B,D)** Selection of s-exons displaying high *species fractions* (>0.8) but low phastCons median scores (<0.1).

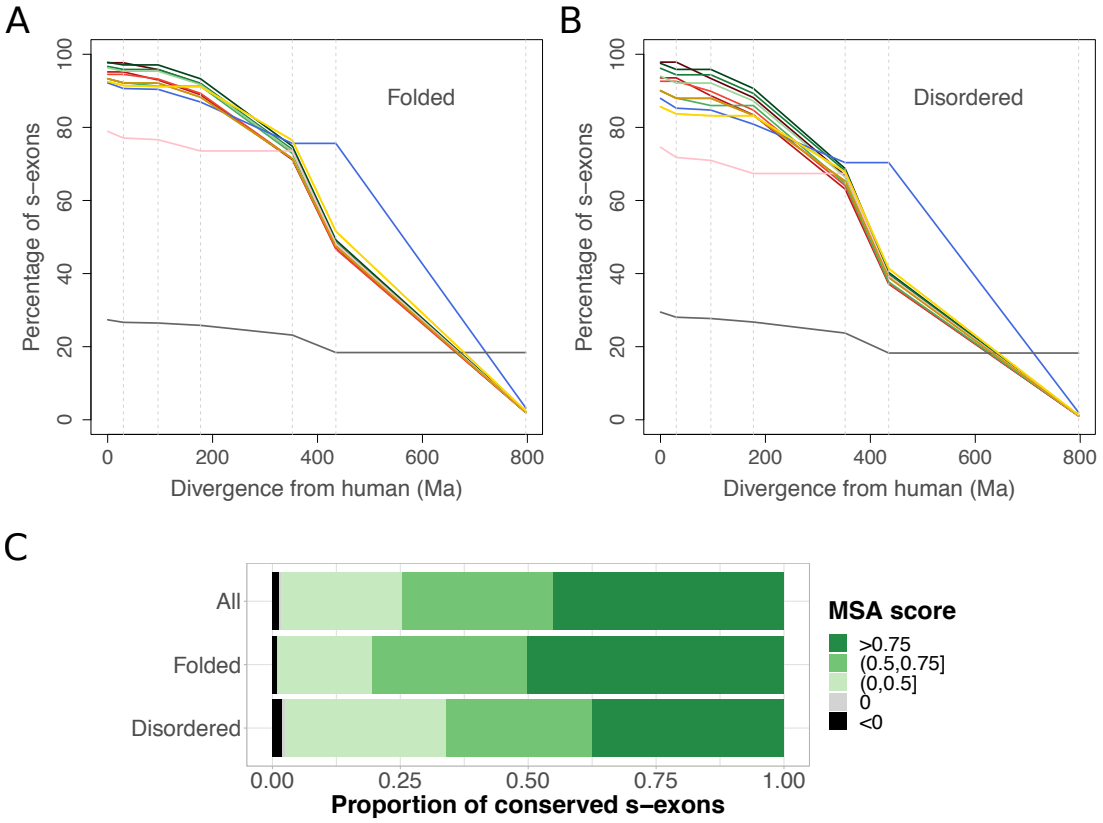
Supplemental Figure S13: **Low-complexity regions detected in** *COL18A1***.** On top, summary of the predictions produced by different methods and integrated in the PlaToLoCo webserver[46] (`http://platoloco.aei.polsl.pl`). The light grey rectangles indicate the positions of the 12 s-exons displaying high *species fractions* (>0.8) but low phastCons scores (<0.1, see bottom right corner on Fig. 4C). At the bottom, frequencies of occurrences of amino acids in the protein sequence (yellow) compared with those computed over popular databases (other colors). The input given to the server was the UNIPROT identifier of the human *COL18A1* (P39060).



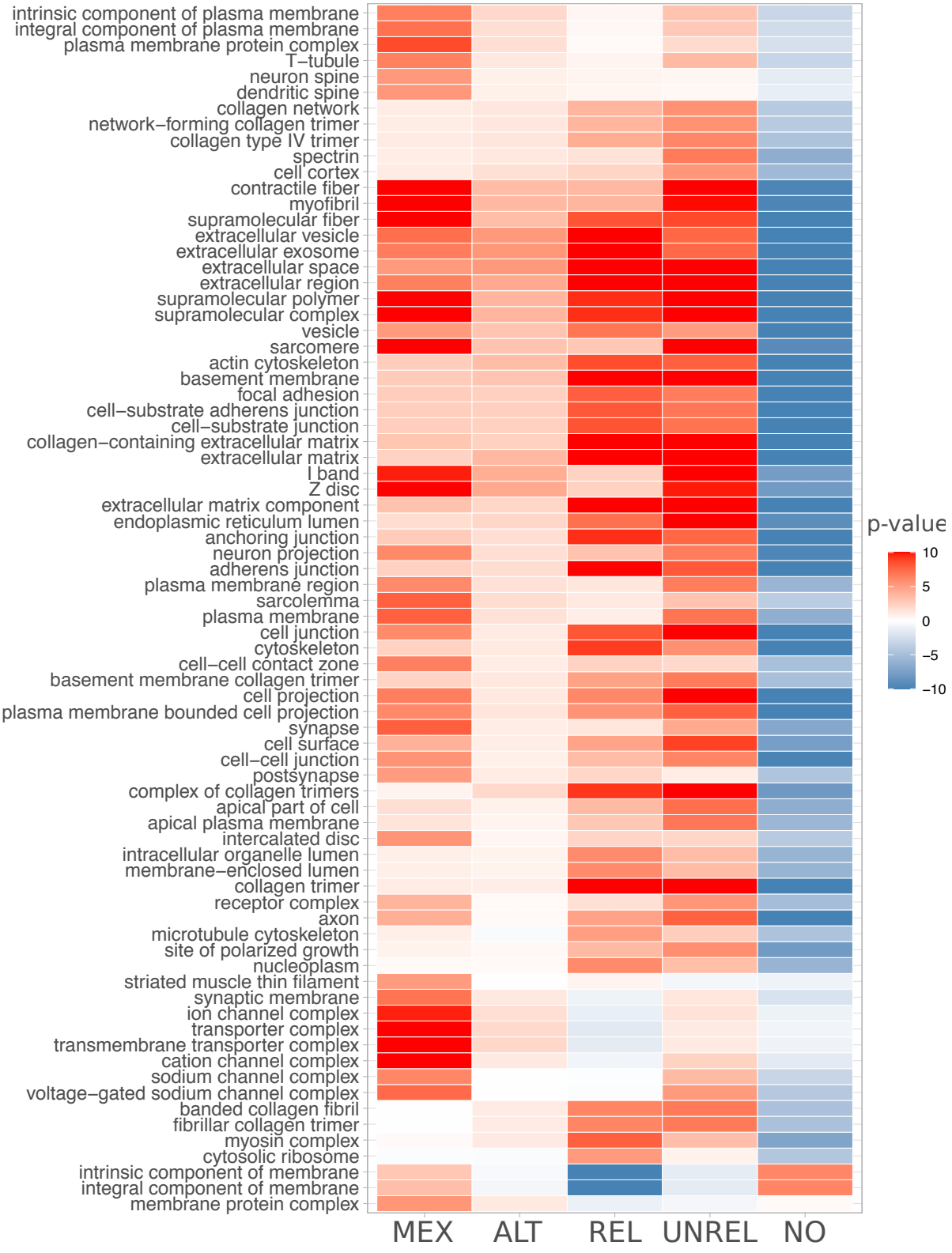
Supplemental Figure S14: **Properties of the s-exons and AS events detected over our curated set of 50 genes. A**. Distributions of the s-exon species fractions depending on the number of species. The red tones indicate the number of genes comprised in each distribution. **B**. Cumulative distributions 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 ASE (*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 at least 8 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).



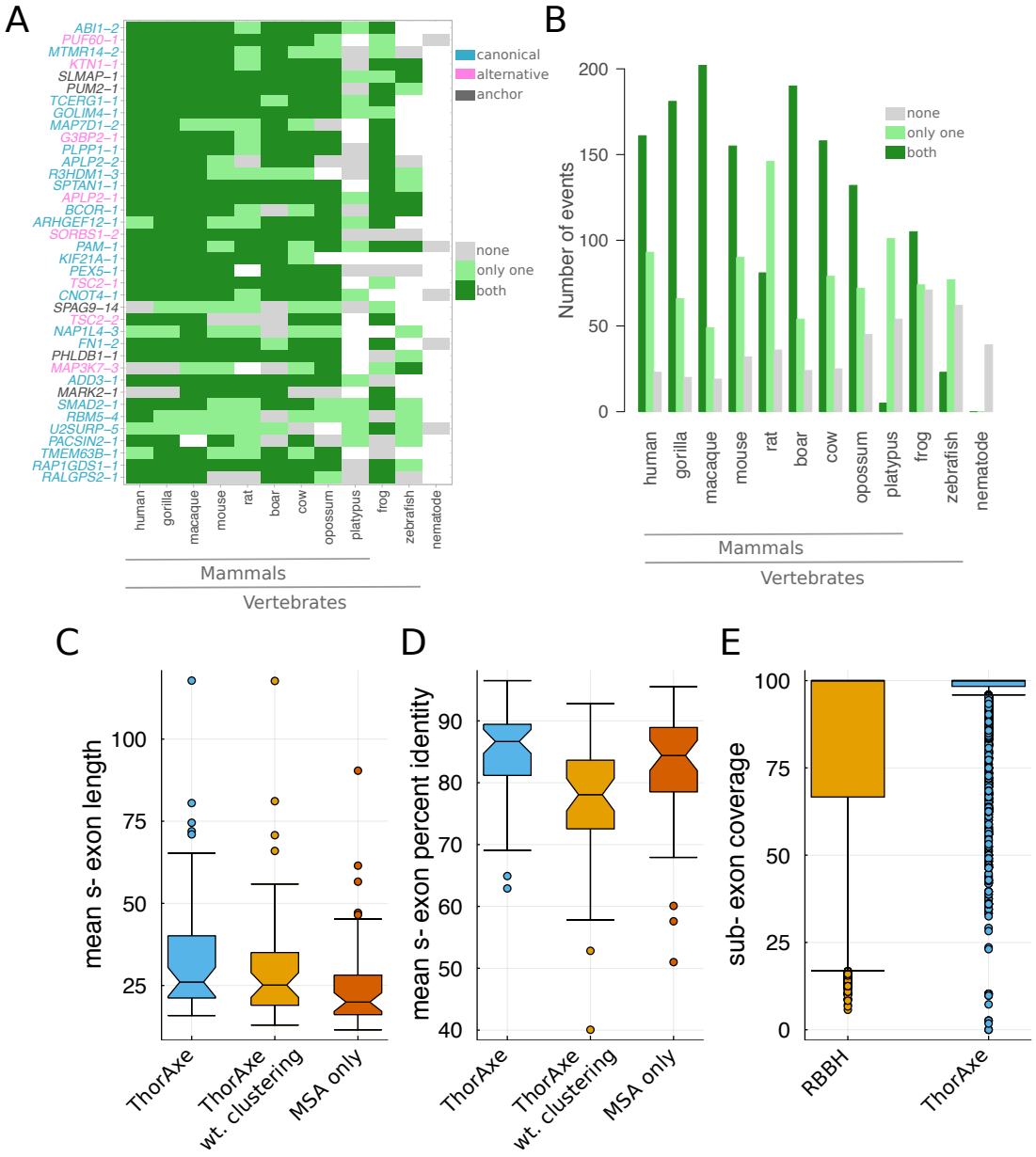
Supplemental Figure S15: **Length distribution for the s-exons defined by very poor quality MSAs (negative score).**



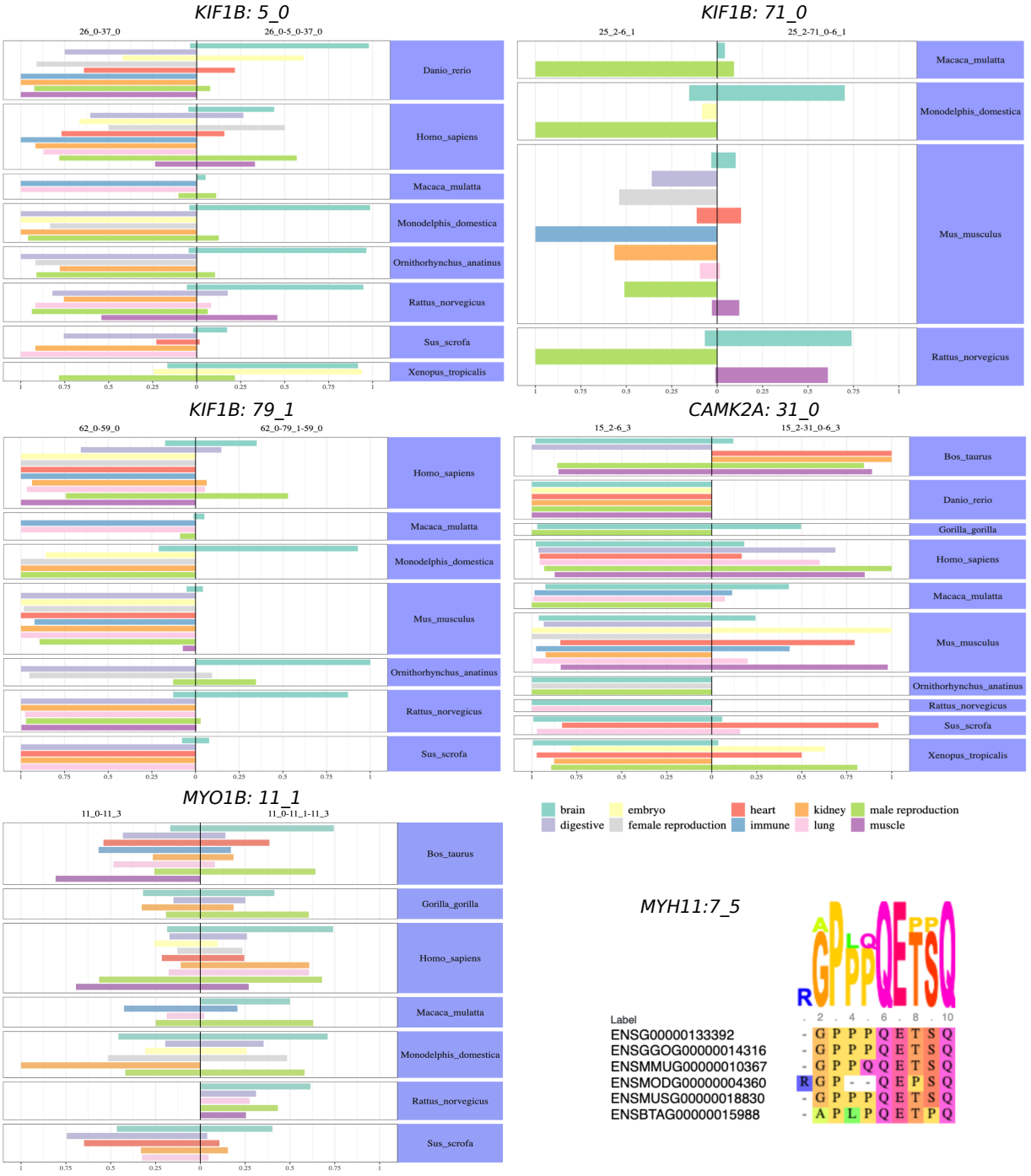
Supplemental Figure S16: **S-exon conservation of folded versus disordered s-exons. (A-B)** Percentages of s-exons (y-axis) conserved up to different evolutionary distances (x-axis) from human. The y-values at $x = 0$ give the percentages of s-exons conserved in at least another species. We report values only for the genes with one-to-one orthologs in more than seven species (class *8-12* in Supplemental Fig. S7). The results are reported for the folded s-exons (panel A, 23 7584 s-exons) and the disordered ones (panel B, 17 3018 s-exons) as predicted by DisEmbl[47] (see *Materials and Methods*). **(C)** Proportions of conserved s-exons displaying very poor (negative score) to very good (score close to one) alignment quality. We consider the whole set, and two subsets comprised of the folded s-exons (27 3077) and the disordered s-exons (19 7985 s-exons). 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.



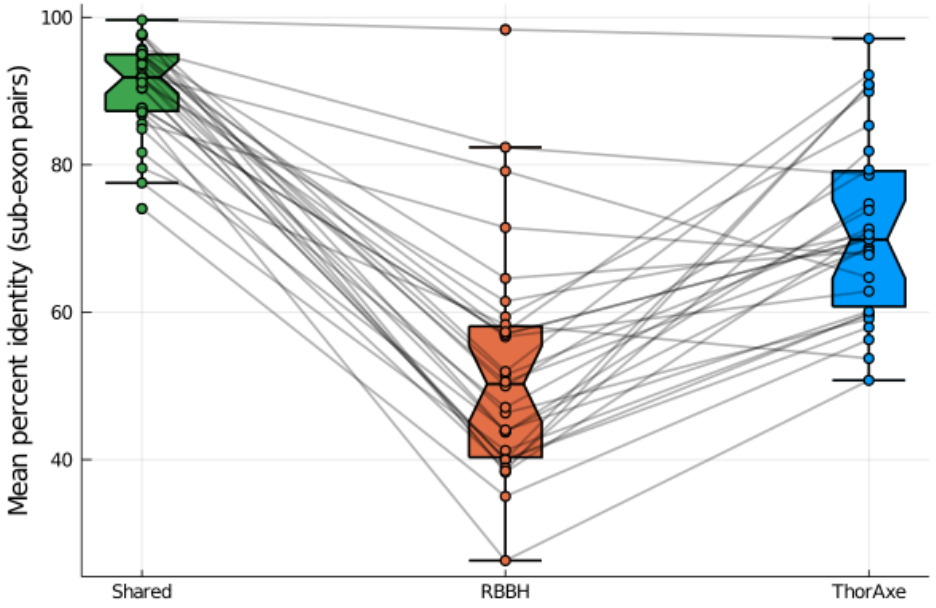
Supplemental Figure S17: **Cellular localisation of the genes where similar sequences are alternatively used.** The genes are classified according to the role of the detected similar s-exons in ASEs. MEX: mutually exclusive s-exons. ALT: alternative (non mutually exclusive) s-exons. REL: one s-exon is 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. NO: no s-exon pair detected. We report the set of Gene Ontology labels of type "cellular component" that are significantly enriched in at least one of the considered gene classes.



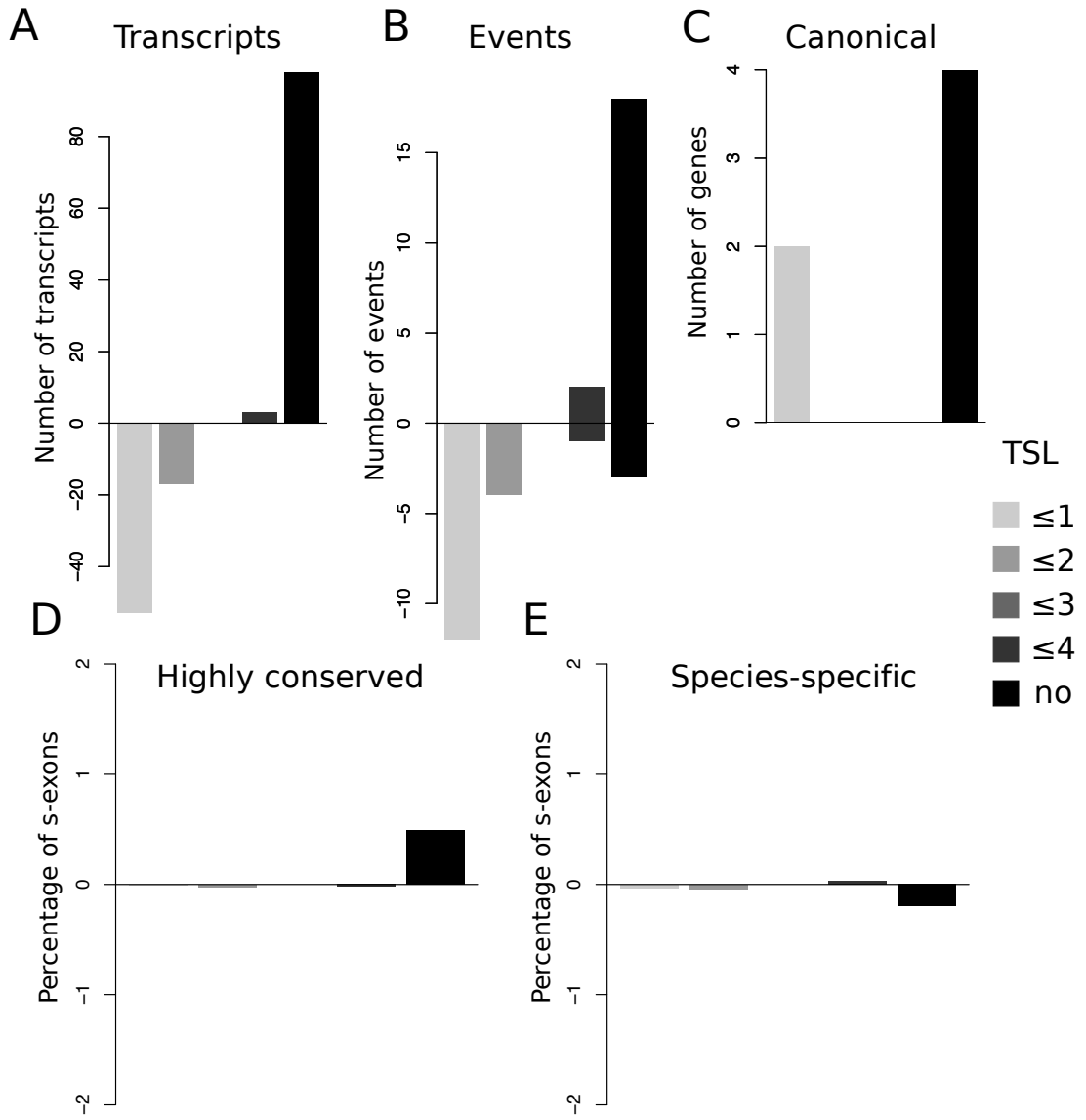
Supplemental Figure S18: **Comparison with other studies and methods. A-B.** ThorAxe conservation assessment of two sets of AS events reported in[23] (A) and[24] (B). For each event and within each species, either none of the paths are supported by the Ensembl annotations data (grey), or only one path is supported (light green), or both paths are supported (dark green). **A.** Each event is designated by its gene name and its rank in ThorAxe output, which reflects its relative conservation level. The color of the label indicates the status of the matching s-exon(s) in the event. A few of the reported exons were mapped to some s-exon(s) located immediately upstream or downstream an event, thus labeled as *anchor* (in grey). This discrepancy likely comes from changes in the gene annotations that occurred in the past 10 years. **B.** Number of events detected within each species, out of a total of 277 events. **C-D.** ThorAxe ablation study assessed on the s-exon lengths (C) and s-exon MSA percent identities (D). The values reported in the distributions are per-gene averages computed over the curated set. **E.** Distributions of sub-exon coverages obtained from RBBH (in orange) and ThorAxe (in blue).



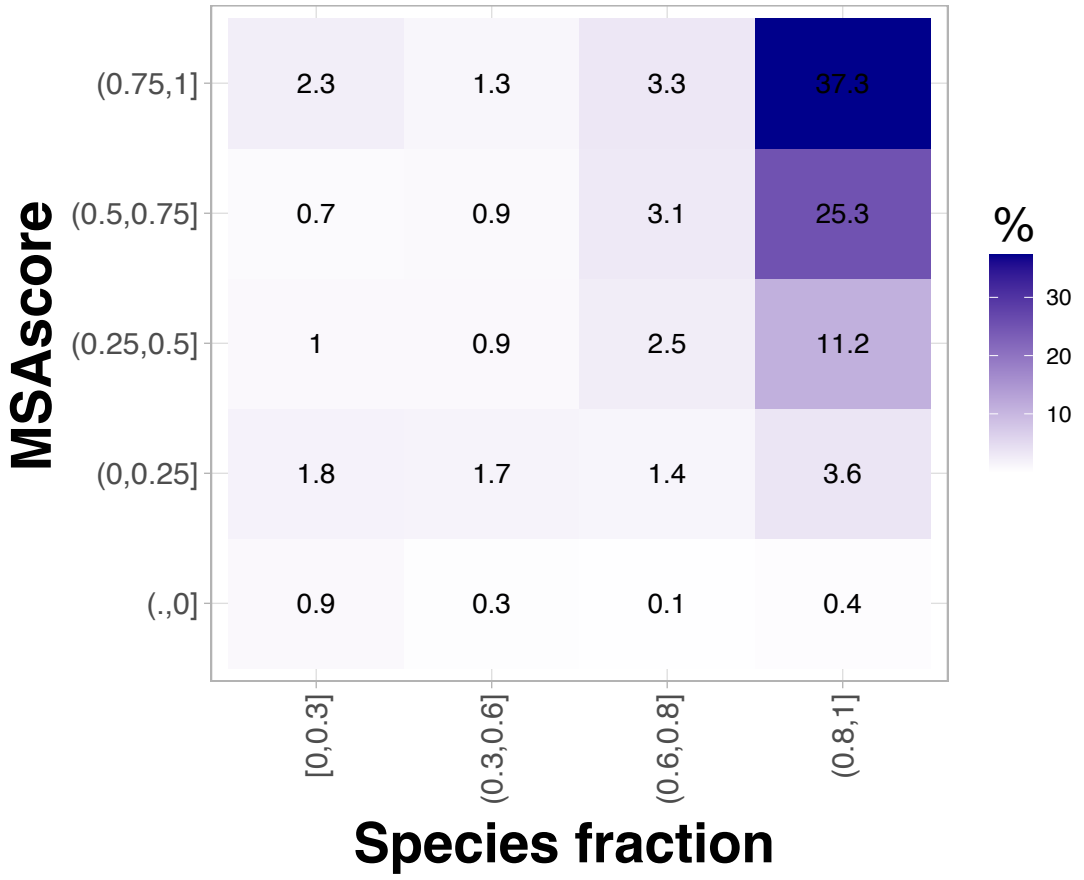
Supplemental Figure S19: **Across-species tissue regulation of six s-exons from the curated set.** The s-exons were selected because they instersect with the set of exons reported in[24]. The colored barplots report the Percent-Spliced In (PSI) computed from RNA-seq splice junctions. We show the MSA associated with the s-exon *7_5* from *MYH11* instead of a PSI barplot. All the plots are accessible at: `http://www.lcqb.upmc.fr/ThorAxe`.



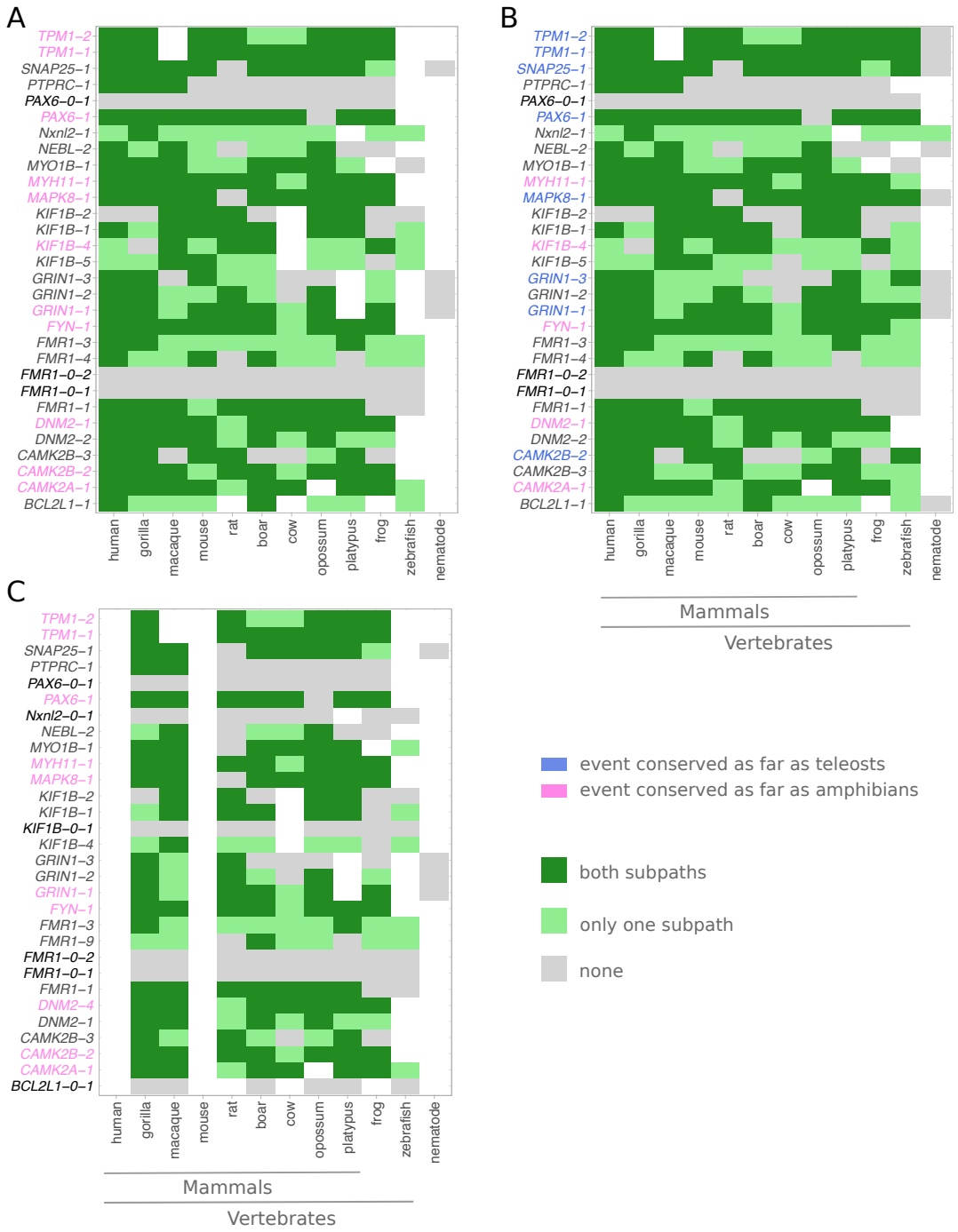
Supplemental Figure S20: **Sequence identity of the sub-exon pairs detected by ThorAxe and RBBH.** The identity percentages are computed for all the sub-exon pairs detected in the curated set and averaged by gene. 86.98% of the pairs are shared (detected by both approaches), 11.51% are discovered only by ThorAxe, and only 1.51% are RBBH specific.



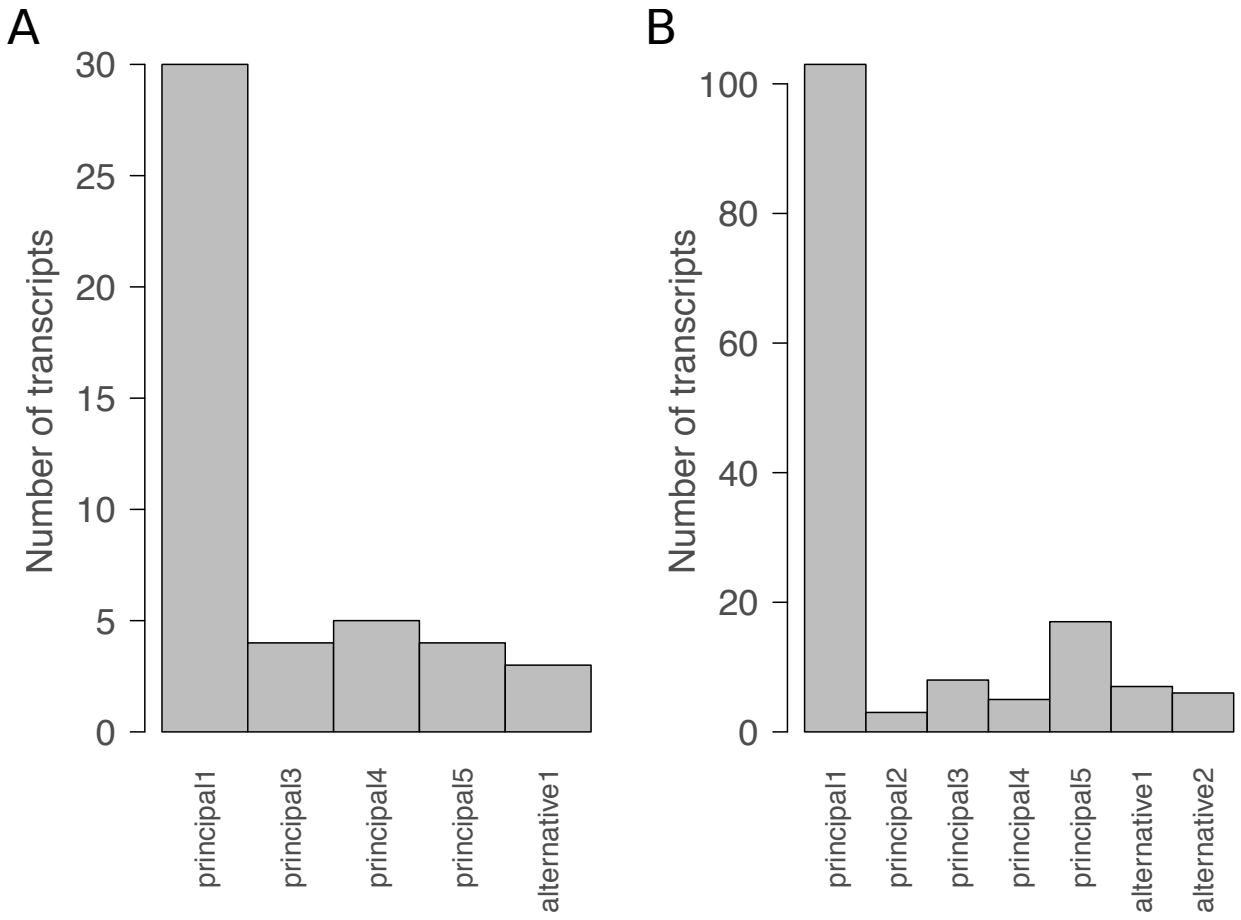
Supplemental Figure S21: **Influence of the Transcript Support Level (TSL) on the curated set of 50 genes.** We report the differences between the results obtained using different TSL thresholds and the reference results, obtained using the default TSL threshold of 3. The TSL threshold is controlled by the *maxtsl* option in ThorAxe, from 1 to 5. A threshold of x indicates that the transcripts with TSL higher than x are filtered out. The "no" label indicates that the transcripts are not filtered based on TSL (*maxtsl=5*). **(A)** Number of gained (>0) or lost (<0) transcripts. **(B)** Number of gained (>0) or lost (<0) events. **(C)** Number of genes where the canonical isoform differs. **(D)** Difference in the percentage of s-exons with *species fraction* higher than 0.8. (E) Difference in the percentage of species-specific s-exons.



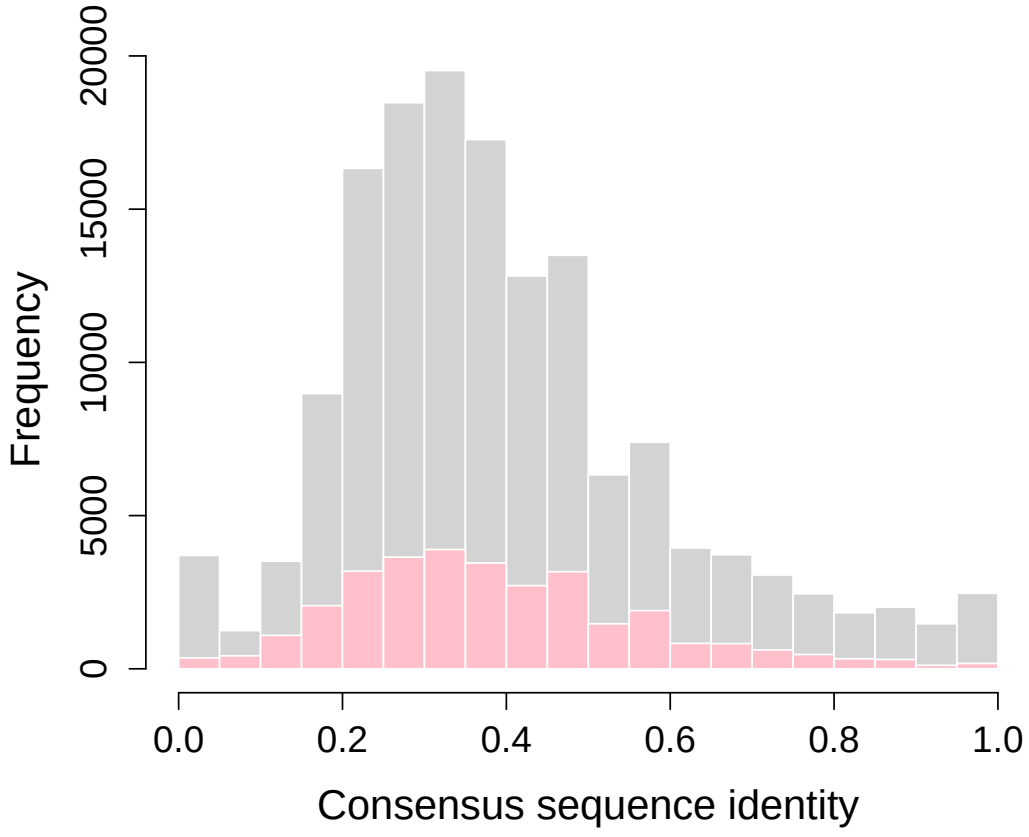
Supplemental Figure S22: **MSA quality score versus *species fraction*.** We consider the set of conserved s-exons identified in the genes with one-to-one orthologs in more than seven species.



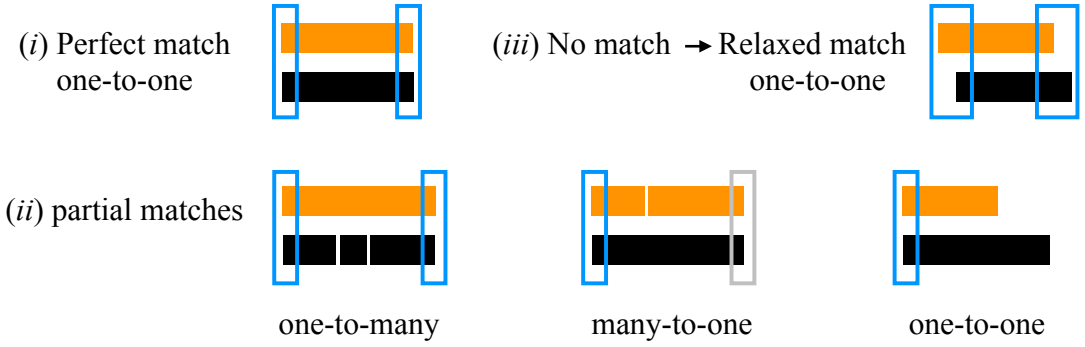
Supplemental Figure S23: **Influence of the choice of genes and species.** Each event is designated by the name of the gene where it occurs and its rank in ThorAxe output, the latter reflecting its relative conservation level. The color of each label indicates the evolutionary distance across which the event is conserved. The color of each cell indicates whether both of the (canonical and alternative) subpaths defining the event are supported by the Ensembl annotations (dark green), or only one path (light green), or none (grey) in a given species. **A.** Default analysis: only species with one-to-one (*1-to-1*) orthologs of the query human genes are considered. **B.** Species with many-to-many (*m-to-n*) orthologs are also considered. **C.** Same as default analysis but excluding human and mouse.



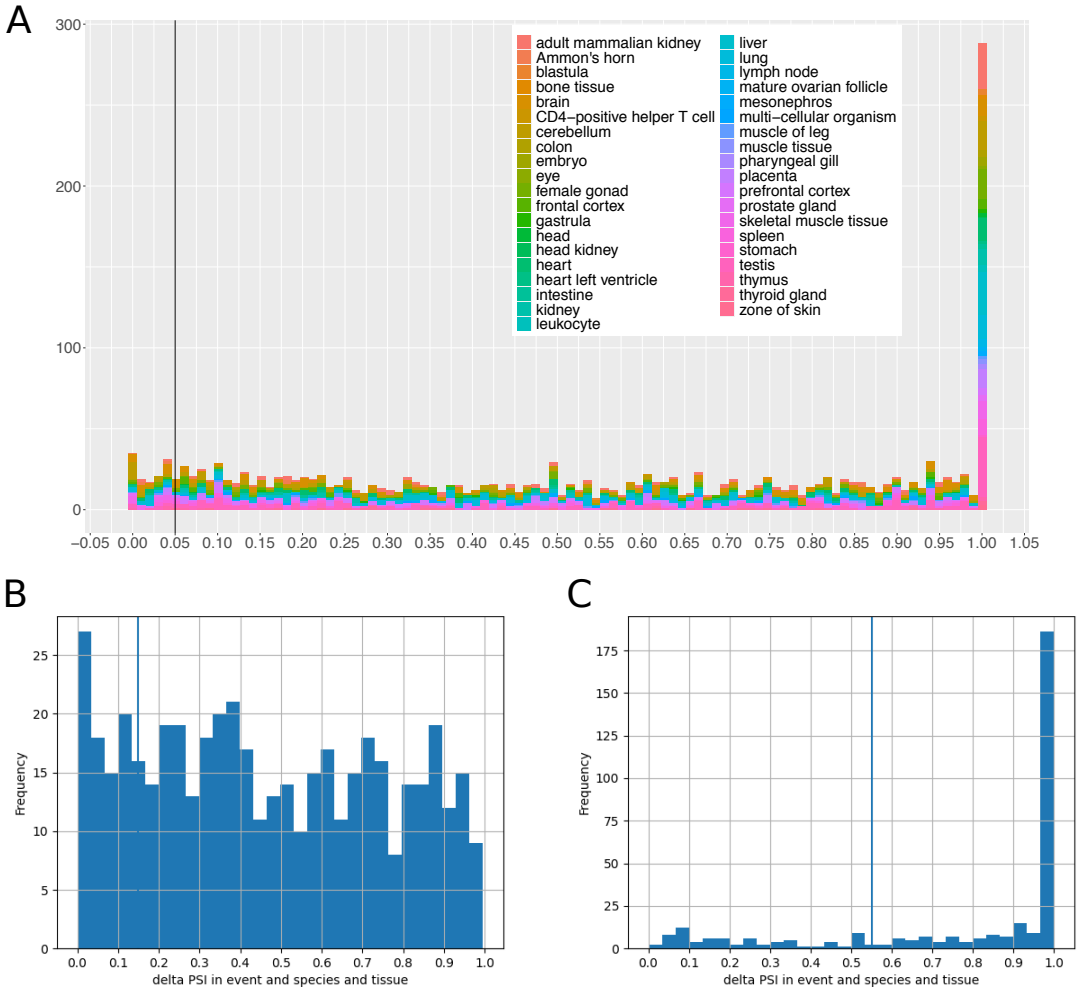
Supplemental Figure S24: **APPRIS annotations for the canonical transcripts defined by ThorAxe on the curated set of 50 genes. A.** Annotations for the 50 canonical transcripts. **B.** Annotations for the 50 canonical paths, meaning that all the transcripts, typically coming from different species, represented by the same path as the canonical transcript in the ESG are considered.



Supplemental Figure S25: **Sequence identity for the similar pairs of s-exons.** For each pair, we counted the number of aligned positions where the consensus sequences of the two s-exons displayed the same amino acid. Variable or highly gapped positions were not considered ("∼" symbol in *hhalign* output). The grey distribution represents all 150 020 s-exon pairs with a p-value lower than 0.001. The pink distribution represents the 31 031 finally selected pairs.



Supplemental Figure S26: **Types of mapping between ThorAxe s-exons and Whippet nodes.** The orange rectangles represent species-specific sequences extracted from s-exons. In each panel, the query s-exon is the first one. The black rectangles represent Whippet matching nodes sequences. The empty rectangles indicate matches, in blue for the query s-exon and in grey for some other s-exon.



Supplemental Figure S27: **RNA-seq Percent Spliced In statistics.** **A**. Distribution of the PSI values computed for the canonical and/or alternative subpaths defining the ASEs detected in the curated set and documented in the literature. The colors indicate the tissues. **B**. Distribution of PSI absolute differences between the canonical and alternative paths, when both are present in the same tissue and species. **C**. Distribution of PSI values for either the canonical or the alternative path, when only one of the two is present in a tissue from a species.

References

[1] Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **2016**, *44*, D710–716.

[2] Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Selection of representative protein data sets. *Protein Sci.* **1992**, *1*, 409–417.

[3] Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **1981**, *147*, 195–197.

[4] Szalkowski, A. M. Fast and robust multiple sequence alignment with phylogeny-aware gap placement. *BMC bioinformatics* **2012**, *13*, 129.

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

[6] Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* **2019**, *20*, 1–15.

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

[8] Komljenovic, A.; Roux, J.; Wollbrett, J. 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* **2018**,

[9] Leinonen, R.; Sugawara, H.; Shumway, M.; Collaboration, I. N. S. D. The sequence read archive. *Nucleic acids research* **2010**, *39*, D19–D21.

[10] Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2012**, *29*, 15–21.

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

[12] Wang, L.; Wang, S.; Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **2012**, *28*, 2184–2185.

[13] Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.

[14] Venables, J. P. et al. Identification of alternative splicing markers for breast cancer. *Cancer Res* **2008**, *68*, 9525–9531.

[15] Katz, Y.; Wang, E. T.; Airoldi, E. M.; Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **2010**, *7*, 1009–1015.

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

[17] DeLano, W. The PyMOL Molecular Graphics System. 2002; http://www.pymol.org.

[18] Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000**, *25*, 25–29.

[19] Consortium, G. O. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research* **2019**, *47*, D330–D338.

[20] Rivals, I.; Personnaz, L.; Taing, L.; Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **2007**, *23*, 401–407.

[21] Siepel, A.; Haussler, D. *Statistical methods in molecular evolution*; Springer, 2005; pp 325–351.

[22] Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **2005**, *15*, 1034–1050.

[23] Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **2012**, *338*, 1587–1593.

[24] Merkin, J.; Russell, C.; Chen, P.; Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **2012**, *338*, 1593–1599.

[25] others„ et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research* **2021**, *49*, D1046–D1057.

[26] Schwerk, C.; Schulze-Osthoff, K. Regulation of apoptosis by alternative pre-mRNA splicing. *Molecular cell* **2005**, *19*, 1–13.

[27] Yang, X. et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **2016**, *164*, 805–817.

[28] Zalcman, G.; Federman, N.; Romano, A. CaMKII isoforms in learning and memory: Localization and function. *Frontiers in molecular neuroscience* **2018**, *11*, 445.

[29] Bayer, K.-U.; Harbers, K.; Schulman, H. αKAP is an anchoring protein for a novel CaM kinase II isoform in skeletal muscle. *The EMBO journal* **1998**, *17*, 5598–5605.

[30] Khan, S.; Downing, K. H.; Molloy, J. E. Architectural dynamics of CaMKII-actin networks. *Biophysical journal* **2019**, *116*, 104–119.

[31] Durieux, A.-C.; Prudhon, B.; Guicheney, P.; Bitoun, M. Dynamin 2 and human diseases. *Journal of molecular medicine* **2010**, *88*, 339–350.

[32] Pretto, D. I.; Eid, J. S.; Yrigollen, C. M.; Tang, H.-T.; Loomis, E. W.; Raske, C.; Durbin-Johnson, B.; Hagerman, P. J.; Tassone, F. Differential increases of specific FMR1 mRNA isoforms in premutation carriers. *Journal of medical genetics* **2015**, *52*, 42–52.

[33] Saito, Y. D.; Jensen, A. R.; Salgia, R.; Posadas, E. M. Fyn: a novel molecular target in cancer. *Cancer: Interdisciplinary International Journal of the American Cancer Society* **2010**, *116*, 1629–1637.

[34] Cull-Candy, S.; Brickley, S.; Farrant, M. NMDA receptor subunits: diversity, development and disease. *Current opinion in neurobiology* **2001**, *11*, 327–335.

[35] Chandrasekar, R. Alcohol and NMDA receptor: current research and future direction. *Frontiers in molecular neuroscience* **2013**, *6*, 14.

[36] Matsushita, M.; Yamamoto, R.; Mitsui, K.; Kanazawa, H. Altered Motor Activity of Alternative Splice Variants of the Mammalian Kinesin-3 Protein KIF1B. *Traffic* **2009**, *10*, 1647–1654.

[37] Waetzig, V.; Herdegen, T. Context-specific inhibition of JNKs: overcoming the dilemma of protection and damage. *Trends in pharmacological sciences* **2005**, *26*, 455–461.

[38] Eddinger, T. J.; Meer, D. P. Myosin II isoforms in smooth muscle: heterogeneity and function. *American Journal of Physiology-Cell Physiology* **2007**, *293*, C493–C508.

[39] Greenberg, M. J.; Ostap, E. M. Regulation and control of myosin-I by the motor and light chain-binding domains. *Trends in cell biology* **2013**, *23*, 81–89.

[40] Li, B.; Zhuang, L.; Trueb, B. Zyxin interacts with the SH3 domains of the cytoskeletal proteins LIM-nebulette and Lasp-1. *Journal of biological chemistry* **2004**, *279*, 20401–20410.

[41] Jaillard, C. et al. Nxnl2 splicing results in dual functions in neuronal cell survival and maintenance of cell integrity. *Hum Mol Genet* **2012**, *21*, 2298–2311.

[42] Sasamoto, Y.; Hayashi, R.; Park, S.-J.; Saito-Adachi, M.; Suzuki, Y.; Kawasaki, S.; Quantock, A. J.; Nakai, K.; Tsujikawa, M.; Nishida, K. PAX6 isoforms, along with reprogramming factors, differentially regulate the induction of cornea-specific genes. *Scientific reports* **2016**, *6*, 20807.

[43] Tchilian, E. Z.; Beverley, P. C. Altered CD45 expression and disease. *Trends in immunology* **2006**, *27*, 146–153.

[44] Nagy, G.; Milosevic, I.; Fasshauer, D.; Muller, E. M.; de Groot, B. L.; Lang, T.; Wilson, M. C.; Sørensen, J. B. Alternative splicing of SNAP-25 regulates secretion through nonconservative substitutions in the SNARE domain. *Molecular biology of the cell* **2005**, *16*, 5675–5685.

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

[46] others,, et al. PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic acids research* **2020**, *48*, W77–W84.

[47] Linding, R.; Jensen, L. J.; Diella, F.; Bork, P.; Gibson, T. J.; Russell, R. B. Protein disorder prediction: implications for structural proteomics. *Structure* **2003**, *11*, 1453–1459.