Master's Thesis
M2 BIM-info 2021-2022

Modulating protein structure and function through the alternative usage of pseudo-repeats

Antoine Szatkownik

Supervisors :
Hugues Richard
Elodie Laine

Acknowledgments

I would like to deeply thank Hugues Richard and Elodie Laine for offering me the opportunity to work with them. My work with Hugues and Elodie always happened in good humor filled with enthusiasm. Their expertise, excitement and passion for scientific research is inspiring. I also would like to thank Diego Javier Zea without whom this work would not have been possible. I would like to thank the Robert Koch Institute for letting me work in their great facility. I would like to thank the Franco-German Youth Office (l'Office franco-allemand pour la Jeunesse, OFAJ) for awarding me a grant.

Finally I would like to thank Milène without whom the poster presented at JOBIM 2022 would not have seen the light of day, and obviously my family who always supported me in what I was doing despite not understanding anything about it.

Figure 2. **OFAJ logo.** Office Franco-Allemand pour la Jeunesse.

Contents

1. **Work environment**

The Robert Koch Institute is a German federal government agency and research institute located in Berlin, Germany, next to the Charite hospital campus. It is one of the most important bodies for the safeguarding of public health in Germany. The RKI is specialized in the identification, surveillance and prevention of diseases, especially infectious diseases; monitoring and analysing long-term public health trends in Germany; providing a scientific basis for health-related political decision-making. More specifically, bioinformatics unit at the RKI consists of around 15 persons addressing the needs of the institute and also working on various topics such as the monitoring of the evolution of SARS-CoV-2 mutations in Germany and worldwide, the detection of horizontal gene transfer in bacterial communities with a view to monitoring their resistance to antibiotics (antimicrobial resistance, AMR), and the identification of proteins using high-throughput sequencing and deep learning technologies. I lived in Berlin for the duration of my internship, i.e. 6 months, and worked mainly with Hugues Richard and Elodie Laine. Elodie followed remotely my work with some face-to-face meetings. At some point, I used the LCQB CPUs cluster to perform extensive calculations. Francesco Oteri and Hugues Ripoche also helped me regarding technical issues regarding the LCQB's server and cluster.
During the 6 months of internship, I had the occasion to present the poster of my work at JOBIM 2022 in Rennes (see appendix for the poster).

Figure 3. The Robert Koch Institute building at Wedding, Nordufer 20 in Berlin, Germany. (wikipedia)

2. Introduction

2.1. Background

In this work, I have focused on the protein diversity arising from the duplication of genetic material, from the alternative initiation/termination of transcription, and from the alternative splicing of pre-messenger RNA (mRNA) transcripts. These mechanisms play important role in creating new biological functions. The duplication of a gene into several copies, called paralogs, can lead to sub- or neo-functionaliation through subsequent sequence divergence [1]. Alternative splicing (AS), by retaining different combinations of exons, can produce multiple mature mRNA transcripts from a single gene [2]. Some of these transcripts will lead to different protein isoforms, or proteoforms [3], which may adopt different shapes [4], interact with distinct cellular partners [5], and perform unrelated functions [6-8]. Hence, there is a growing interest in assessing the functional impact of AS at the protein level, and in determining the evolutionary origin of the proteoforms. Moreover there is no standard method to characterise the different proteoforms experimentally and it remains unclear which proportion of AS has functional implication.
The interplay between genetic duplication and alternative splicing may result in a variety of scenarios [9-11]. One of the most documented cases is the duplication of a gene's protein-coding region, followed by the mutually exclusive inclusion of one or the other copy in transcripts [12]. We are interested in studying the evolutionary conservation of the alternative usage of duplicated regions (pseudo-repeats). Beyond the conservation of pseudo-repeats themselves across protein families and species, we ask whether they are alternatively used in the same way. An archetypal example is a pair of paralogous mutually exclusive exons. If both members of the pair are conserved across the paralogs of the family, then the most parsimonious explanation is that the exon duplication event occurred before the gene duplication (GD) event. Some paralogs may have kept the alternative usage of the pair, and thus express proteoforms containing one or the other exon. Others may have sub-functionalized by retaining only one exon [13]. The former scenario is coined as the non-interchangeable model, the latter as the interchangeable model or function-sharing model. A wide variety of cases of those scenarios for the interplay of AS and GD has been described in the literature, thus establishing the functional importance of these pseudo-repeats. [14-16]

2.2. Preliminary work

Zea et al. [17-18] recently extended the notion of splicing graph [19] to a set of orthologous genes/species, and developed an efficient method, ThorAxe, to build **Evolutionary Splicing Graphs** (ESG) and assess the evolutionary conservation of exons, transcripts and AS events. The whole transcript variability observed in a set of genes/species generated by alternative splicing, alternative promoter usage and alternative poly-adenylation is represented under the form of a parsimonious graph, the ESG (figure 4). It is parsimonious since it minimizes the number of nodes while maximising the overall sequence similarity of the MSAs associated to each node. Nodes in an ESG called spliced-exons (s-exons), are the minimal building blocks of transcripts conserved across a set of species, i.e. the exonic sequences belonging to the same s-exon are supposed to be orthologous.
At the scale of the human protein-coding genome, they identified 2190 genes with evidence of evolutionary conserved alternative usage of pseudo-repeats. These AS events have implications for the establishment of functional protein interactions.

During my M1 project I began to develop an automatic method to detect these alternative usage of pseudo-repeats from the set of 2190 genes, which had been compared with a previous study based on a manual curation [12], hence we a priori had a validation of the preliminary results. I had done some post-processing work on these preliminary data to identify false positives. The limitation of this previous work was that I wasn't able to identify similarities between pseudo-repeats coming from distinct genes without doing extensive computations. I thus changed my strategy from the start by working on raw/more upstream data.

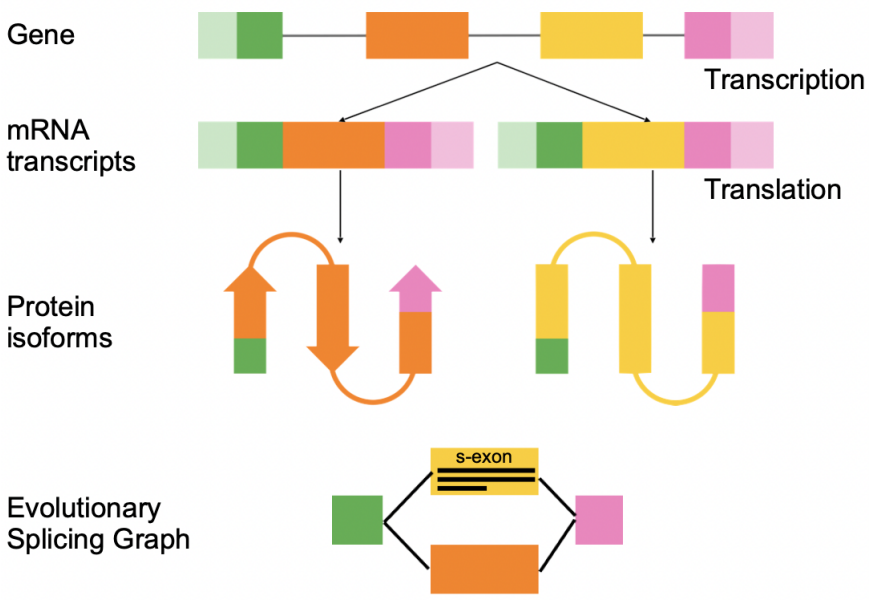


Figure 4. Schematic representation of the alternative splicing of a gene. From a gene to an evolution-informed representation of its proteoforms.

2.3. **Motivations**

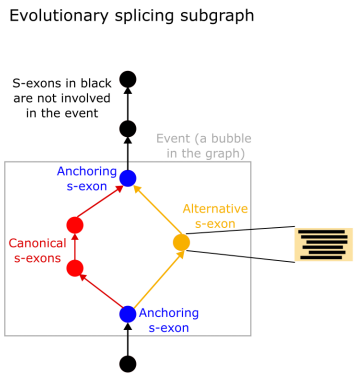
The main motivation of this work is the investigation of how AS modulates the use of similar protein regions (pseudo-repeats), both within a protein and across proteins, during evolution. The repertoire of conserved pseudo-repeats across the proteoforms coming from paralogous members of a protein family will enable us to infer scenarios about the evolutive origin of these pseudo-repeats and their conservation levels will be indicative of their functional implication. It will also help us understand what the constraints/uses we can see in the modulation of the pseudo-repeats are. These similar protein regions do not necessarily correspond to annotated repeats, nonetheless we have confirmation of a fraction of those pseudo-repeats from RepeatsDB [20], and other studies [21-23].
The starting hypothesis that serves as a guideline is that AS rewire the interactions [24], more precisely, the differences in amino acids observed between proteoforms and conserved across paralogs and species are important for the specificity of the interactions. Indeed, these sequence signatures which can have an implication for the interactions, consists in differences in the amino acid sequences of the alternatively used pseudo-repeats or there can be a different number of these pseudo-repeats due to insertion, deletion events. This variability introduced by AS modulates the stoichiometry, kinetics, specificity of the interactions accordingly. Hence we want to systematically and automatically exploit the tension between evolutionary divergence, gene duplication, and alternative splicing toward identifying molecular determinants of protein interaction specificity.

3. Materials and Methods

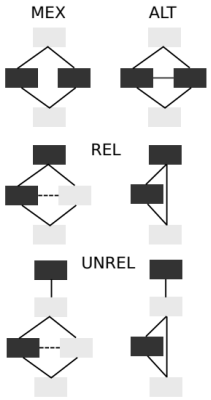
3.1. Definitions

An **ESG** is a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_i \in \mathcal{V}$ is a **spliced-exon** (**s-exon**) and represents a multiple sequence alignment (MSA), $\tilde{s_i} = \{s_A^i, s_B^i, ...\}$, of translated exonic regions coming from a set of orthologous genes $\{G_A, G_B, G_C, ...\}$, where $A, B, ...$ are different species. Two nodes : "start" and "stop" without sequences are added (they represent the beginning and end). There exists a directed edge $e_{i\to j} \in \mathcal{E}$ from node v_i to node v_j if for at least one species K there is a transcript t such that $s_K^i s_K^j \in t$. A transcript observed in a gene corresponds to a path in the ESG from start to stop.

In order to characterize alternative splicing we first define a canonical transcript. The rationale behind the **canonical transcript** is that it should be well-represented across species. An **alternative splicing event** (**ASE**) is a variation between a pair of maximal subpaths that do not share any s-exon, where one subpath necessarily comes from the canonical transcript and the other one from some input transcript. We restrict the space of events by discarding variations between transcripts where none is a canonical transcript. Formally we define an ASE as a bubble in the ESG defined from the edge set $\mathcal{E}$, see figure 5 (a). The event b can be described as the tuple $((v_s, v_e), (v_1^c, v_2^c, ..., v_l^c), (v_1^a, v_2^a, ..., v_m^a))$, where each element is a list of nodes. The nodes v_s and v_e are the starting and ending anchors respectively. The list $(v_1^c, v_2^c, ..., v_l^c)$ defines the canonical subpath bounded by the anchors, while the list $(v_1^a, v_2^a, ..., v_m^a)$ defines the alternative subpath bounded by the anchors. Note that $l \geq 0$ and $m \geq 0$.



(a) **A subgraph extracted from the evolutionary splicing graph.** Each node stands for a s-exon and the directed edges between the nodes indicate pairs of consecutive s-exons observed in the input transcripts. The colors indicate the status of the s-exons with respect to the event (bubble in the graph) highlighted by the grey square. Here, the event involves two mutually exclusive groups of s-exons, in red and orange respectively. One subpath serves as a reference and is thus defined as *canonical* while the other is *alternative*. The event starts and terminates by two anchoring nodes (in blue), present in both the canonical and the alternative transcripts.



(b) Evolutionary splicing subgraphs depicting different alternative usage scenarios. The pairs of ASPRs linked by an IPR are colored in black.

Figure 5

We now define pseudo-repeated elements in the context of the ESG. An **In-gene Paralogy Relationship** (**IPR**) is a similarity between two s-exons sequences within a gene/ESG. An **Alternatively Spliced Pseudo Repeat** (**ASPR**) is a pseudo-repeat conserved across species and with some evidence of alternative usage. Thus we define it as a sequence of k consecutive s-exons ($k \geq 1$) in a transcripts that is similar to at least another sequence of s-exons in the gene (in-gene paralogy relationship). We additionally require the list of k s-exons to be compatible with the splicing structure of the ESG (no breaking criterion): the k s-exons are always expressed together in the observed transcripts, i.e. their junction is not broken by any alternative splicing events of the ESG. The bounds of an ASPR are necessarily defined through the splicing events, i.e. the topology of the ESG. Formally it is a subpath in the ESG.
An IPR is assigned a class depending on their alternative usage relationship in the ESG (see figure 5.b), either MEX, ALT, REL, UNREL and NO, in that order of priority (see figure 5.b). While **MEX** ASPRs are mutually exclusive, **ALT** ASPRs are used alternatively but without mutual exclusivity. The **REL** relationship states that there exists at least an event where one ASPR is included in the canonical or alternative subpath, while the other one serves as an anchor. The **UNREL** relationship states that one ASPR belongs to the canonical or alternative subpath of the event while the other one is located outside of the event. **NO** means the ASPRs aren't related by an ASE.

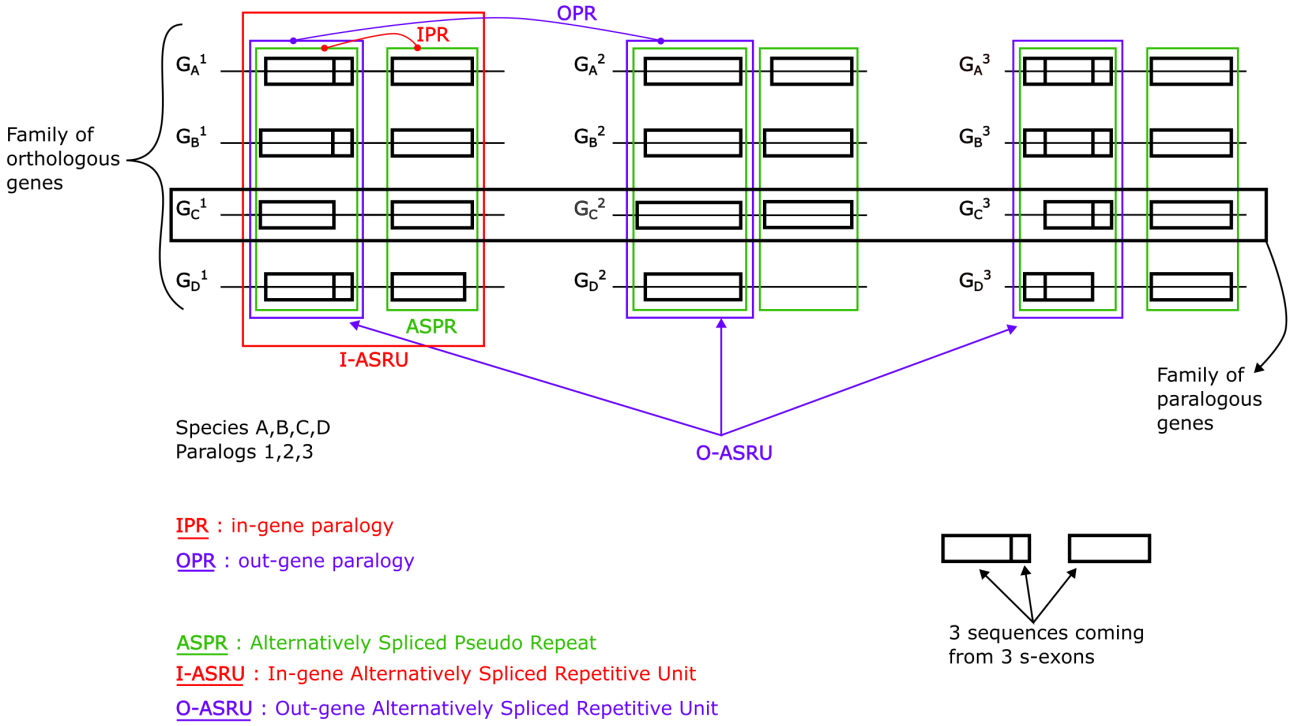


Figure 6. Overview of the concepts/terminology. Note that for each G_X^i we do not represent the ESG but rather the pair of ASPRs and their IPRs exploded across species.

We now need to extend our nomenclature to be able to analyse a sequence conservation/divergence signal across three dimensions, namely across species, paralogs in a species and proteoforms in a gene towards the identification of specificity-determining sites [25]:

- **In-gene Alternatively Spliced Repetitive Unit** (**I-ASRU**) as a set of ASPRs linked one to another by an IPR of the type MEX, ALT, REL or UNREL.

- **Out-gene Paralogy Relationship** (**OPR**) as a relationship that is defined on the basis of strong and unambiguous similarity between ASPRs from two distinct genes/ESGs.
- **Out-gene Alternatively Spliced Repetitive Unit** (**O-ASRU**) as a set of ASPRs coming from N paralogous genes, linked one to another by an OPR.

I-ASRUs are minimal units to analyse the direct role the direct role of AS for functional diversity, while O-ASRU inform on their evolutionary history over the gene family. Thereafter we will abusively talk about a gene, to refer about a group of orthologous genes, whose transcript variability is represented by an ESG.

3.2. **Data**

ThorAxe was used to build ESGs for the ensemble of 18 226 human protein-coding genes (due to computational errors, the ensemble amounted to 17 920 genes) and their one-to-one orthologs across 12 species. Namely three primates (Homo sapiens, Gorilla gorilla, Macaca mulatta), two rodents (Rattus norvegicus, Mus musculus), four other mammals (Bos taurus, Sus crofa, Ornithorhynchus anatinus, Monodelphis domestica) , one amphibian (Xenopus tropicalis), one fish (Danio rerio), and one nematode (Caenorhabditis elegans). We downloaded the corresponding gene annotations from Ensembl release 105 (December 2021). The parameters used for running ThorAxe are the following: one-to-one orthologs, no filter based on TSL.

3.3. **ASPRine: a tool to automatically identify I-ASRUs**

I have been dedicating a large part of my internship work to develop a fully automated computational tool, ASPRine for identifying I-ASRUs starting from a set of transcripts observed in a set of species/orthologous genes, figure 7. ASPRine workflows takes as input the ESG of a queried gene and outputs tables summarizing the information about the I-ASRUs and its ASPRs, it unfold in four main steps:

(1) **All-to-all comparison of s-exons.** The algorithm starts with the ESG of a queried gene. Using $reformat$ function from $HH-Suite3$ [26], we convert all the s-exons of the query gene, except those s-exons that have a length inferior to 5 aa, from FASTA format to A2M format. The next phase consists in building probabilistic models, namely hidden Markov Models (HMM) profiles, from the MSAs associated to each s-exon with $hhmake$ function from $HH-Suite3$. For each gene, we globally aligned each HMM against all the others using $hhalign$, thus resulting in $\frac{n(n-1)}{2}$ alignments, where n is the number of profiles HMM of a given gene. The parsing of those raw HMM alignments enabled us to construct an intra-gene similarity graph.

(2) **Identification of similar s-exon pairs.** The identification of similar s-exon pairs consists in filtering the edges of the intra-gene similarity graph. We considered two s-exons to be similar if:
   - the p-value associated to the profile HMM-HMM alignment was lower than 0.01
   - the percentage of consensus sequence identity was higher than 45%
   - the coverage of the alignment is at least 80% for the query or the target
   - both s-exons are at least conserved in two species

   Those critera were carefully chosen and fine-tuned by considering corner-case false positives such as shown in appendix, figure 14. We set the percentage of consensus sequence identity to 45%, below that the matching positions are were sparse. Since we are interested in repeated regions, we want the alignment to cover almost all of the exonic region. The criterion over the number of species provide us with a notion of conservation. After filtering the edges with the above criteria, we continued further by

retaining those pairs of s-exons that are linked by an ASE, that is pairs to which were assigned a class, either MEX, ALT, REL and UNREL, in that order of priority.

(3) **Clustering of similar s-exons.** Next, we cluster pairs of similar s-exons (the final edges in the intra-gene similarity graph), resulting in connected components that are recognized as I-ASRUs. Each s-exon in the connected component corresponds to a seed ASPR.

(4) **Refinement of the ASPRs.** The final phase is to refine the ASPRs definition to make the I-ASRU more homogeneous (in length, in aa). That is we seek to extend, as much as possible, the seed ASPRs with adjacent s-exons with respect to the ESG so as to minimise the difference in length (in aa) of the ASPRs belonging to the same I-ASRU, i.e. minimise the standard deviation over the length (in aa) of the ASPRs of an I-ASRU. For an extension to be valid between an initial ASPR called seed and an adjacent s-exon it must respect the constraint that no alternative splicing event conserved in at least 2 species break the junction between the seed ASPR and the extension (in figure 7, 17_1 is the seed ASPR and is extended in the N-terminal direction with ASPR 17_2). This extension process is implemented in an iterative way, and is carried out as long as the margins computed from the alignments is strictly positive (see fig. 7).

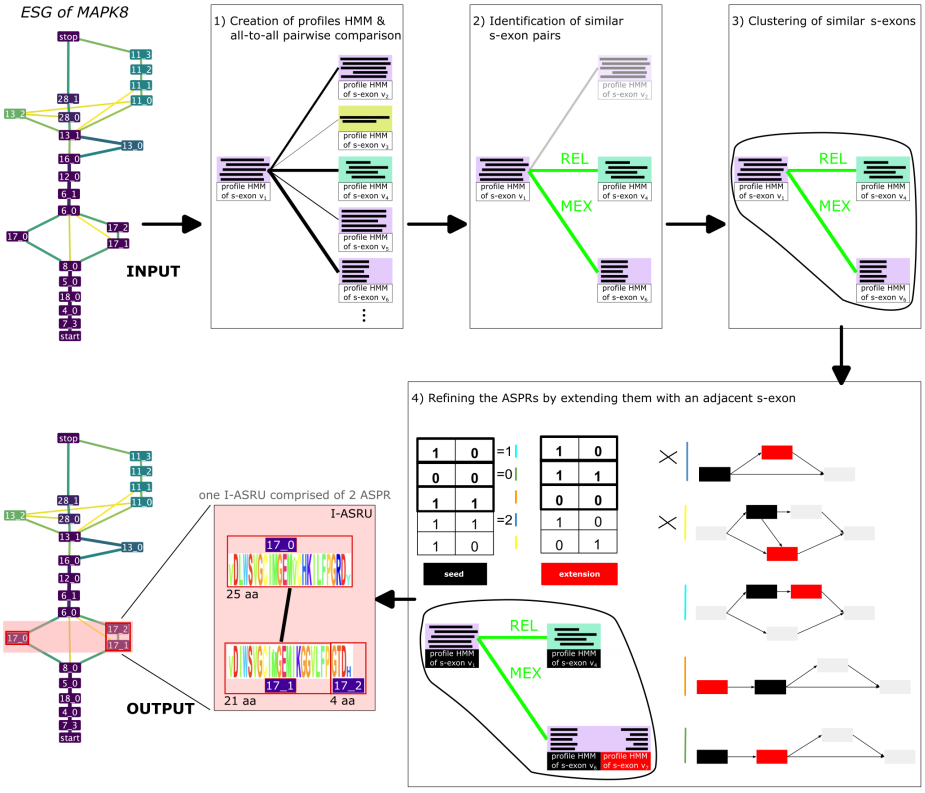


Figure 7. Overview of the method: it takes as input a gene (Ensembl ID), creates the profile HMM for each s-exon that are at least 5 amino acid long, perform all-to-all pairwise profile HMM-HMM alignment, filter the alignments with user-defined parameters values and finally output tables summarizing informations about the I-ASRUs, the ASPRs, the similar pairs of s-exons and the raw alignments. The Ensembl ID of MAPK8 is ENSG00000107643.

3.4. **Extending the analysis to entire protein families**

To extend the analysis of I-ASRUs to entire protein families, we seek to cluster together ASPRs detected across members of a family of paralogous genes with respect to a similarity distance. Beside having a notion of distance between the corresponding ASPRs inside a family (thus ensuring the detection of cases of the non-interchangeable model), we also needed to rescue paralogs with s-exons homologous to these ASPRs but not alternatively used, i.e. case of the interchangeable model, hence the construction of an inter-gene similarity graph is necessary. In fig ?B left, gene W has a s-exon similar to the ASPRs in the I-ASRUs of gene U and V, but it is not involved alternatively used, we say that the s-exon is fixed in the protein. In continuity with the line of work we propose, we are not dealing with single genes, but with groups of orthologous genes, which is an additional dimensionality. Let's give some numbers to get an idea. Let's say we have a family of 4 paralogs in human, we are studying these genes across 12 species, so we may be effectively looking at 4× 6 = 24 genes at once (there are on average, 6 one-to-one ortholog for each ESG/gene across the families, see appendix fig.16).

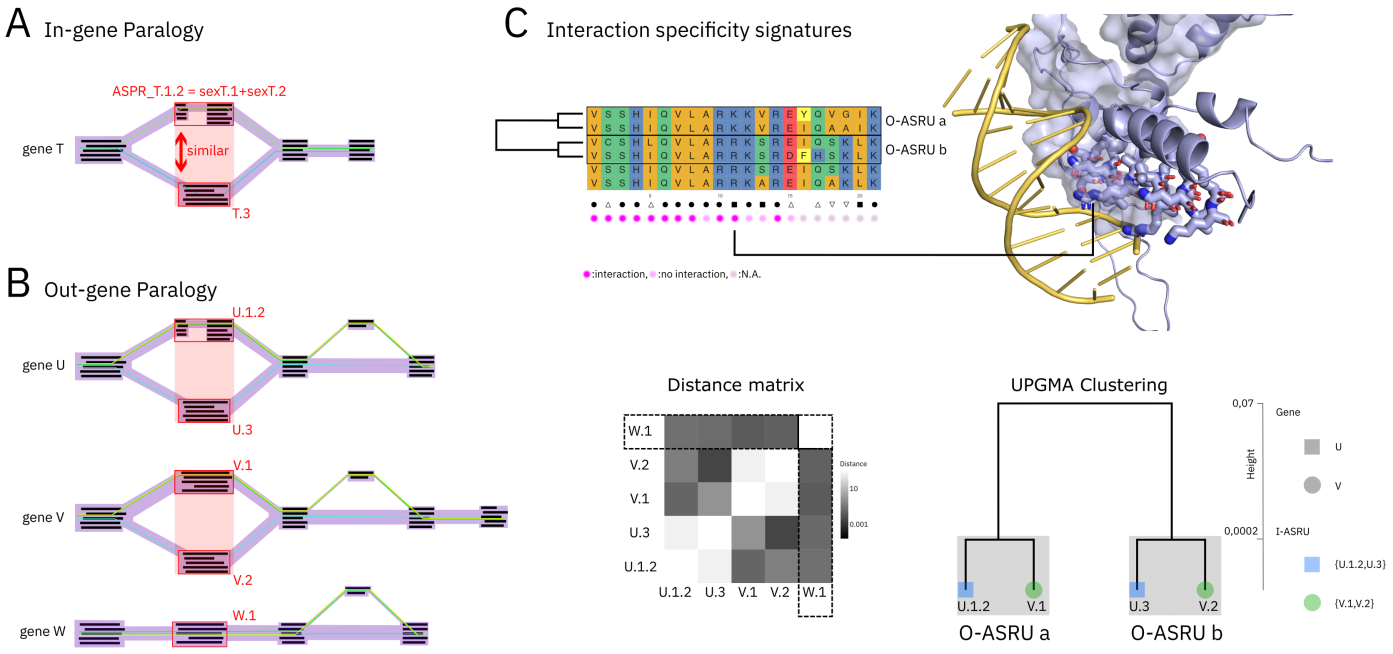


Figure 8. Overview of the analysis. (panel A) A schematic representation of an IPR, that relates the ASPR_T.3 and the ASPR_T.1.2 which consists of the subpath s-exon_T.1 + s-exon_T.2. (panel B) Gene U,V,W are three paralogous genes, each having respectively one I-ASRU (of two ASPRs) and one fixed s-exon (i.e. a s-exon similar to the ASPRs of the paralogs that is not alternatively used). The distance matrix is restricted to ASPRs only, the fixed s-exons are ignored. The UPGMA clustering from the distance matrix enables us to detect O-ASRUs. (panel C) MSA of the O-ASRUs of TEAD[1,2,3,4]+fixed s-exons and the mapping of the residues on the AlphaFold models of the canonical and alternative proteoform of TEAD3. The signatures (square, circle, upward/downward triangles) are computed with respect to the ASPRs only not the fixed s-exons.

3.4.1. *Creation of families of paralogous genes.*

To create families of paralogous genes, I retrieved information about human gene families from Ensembl Biomart version 105 into a table, where the rows are pairs of paralogous genes (concretely, the Ensembl gene ID and its human paralogue Ensembl gene ID) and sequence identity levels (target paralogous gene to query gene and query gene to target paralogous gene). We filtered this table with a minimum threshold of 60% target to query **and** query to target, i.e. they are reciprocal hits. With this high level of sequence identity, we expect comparable ESG topologies and conservation signals across paralogs to be detectable at the amino acid level. From these pairs of paralogous genes we constructed a graph and considered a connected component in that graph as a family. We finally retained the families that contained at least two genes with at least one I-ASRU.

3.4.2. *Creation of a s-exons database.*

To quantify distances between ASPRs inside a family and to retrieve the s-exons similar to these ASPRs that comes from a paralog, we created a profile HMM s-exons database, called thereafter *sexDB*, covering the whole human proteome and their one-to-one orthologs across 12 species, with the help of the LCQB CPUs cluster (i.e. HPC) and by using $HHSuite3$ []. Only profiles HMM of s-exons of length > 5 aa were made. The length of an s-exon v, is computed as,

$$l(v) = \max_{s \in S} \sum_{n=1}^{l_s} \mathbb{1}_{\{s_i \in \mathscr{A}\}} \tag{1}$$

where S is the set of sequences comprised in the MSA, l_s is the length of the aligned sequence s, and $\mathscr{A}$ is the 20-letter amino acid alphabet (e.g., excluding gap characters).
Then we performed global profile HMM-HMM alignments with $hhsearch$ for each s-exon (its profile HMM) against *sexDB*. For each search of a s-exon against *sexDB*, a list of hits was proposed comprising many-to-many alignments, that is a s-exon could be found multiple times inside another one (one-to-many), and the other way around. Those alignments enabled us to construct a directed similarity multigraph, where the nodes are the s-exons and the directed edges are the alignments (from query to target and target to query). We trimmed the edges between two s-exons to retain only one direction (query to target or target to query), and reduced further the edges in the remaining direction to retain the best hit. Since the s-exon is a minimal transcript building block, retaining the best hit among the one-to-many alignments does not affect our analysis. The resulting similarity graph contained intra-gene and inter-gene similarities. With the help of the intra-gene similarity graph, we further clustered similar s-exons for each genes and refined the ASPRs, as in 3.3.(3-4).

3.4.3. *Distance matrix.*

We have for a family of N paralogous genes, N ESGs (fig ?B, left). Each member of the family has at least one I-ASRU, represented with the light red rectangle comprising two red boxes. From the similarity graph of intra-gene and inter-gene similarities we are able to recover the information about the all-to-all comparisons of the ASPRs and about the fixed s-exons inside the rescued paralogs.
We then seek to design a mapping from the raw alignments/edges to a quantity that will allow us to cluster, via a distance matrix, homologous ASPRs under the constraint that they do not belong to the same gene, and that they are located at the same position along the protein. We thus distinguish the case where the ASPRs are inside the same gene or not and avoid intra-gene

clustering before inter-gene clustering. We have a criterion that is derived from a property of the set of genes and not from each gene individually to correct for cases where the intra and inter distances have the same order of magnitude and for cases of scenario imbalance from one gene to another creating a disparity, i.e. one gene of the family possess a fully connected I-SRU, thus the max distance between ASPRs in this gene is small.

$$d : \{(I_1, I_2) \text{ is alignment between ASPR } I_1 \text{ and ASPR } I_2 \,|I_1 \in \text{gene } G_1,\, I_2 \in \text{gene } G_2\} \to \mathbb{R}_+^*$$

(2)
$$d(I_1, I_2) = \mathbb{1}_{\{G1 \neq G2\}} \cdot d_{sim}(I_1, I_2)^2 + \mathbb{1}_{\{G1=G2\}} \cdot \left[d_{sim}(I_1, I_2) + \alpha_{para}^G \cdot d_{max}^{global} + |r_{I_1} - r_{I_2}| \cdot \alpha_{rang} \right]$$

where

- $d_{sim}(I_1, I_2) = -\log_{10} \mathbb{P}(I_1 \text{ homologous to } I_2)$
- α_{para}^G : Number of ASPRs per gene.
- $d_{max}^{global} = \max_G d_{max}^G \quad \text{with} \quad d_{max}^G = \max_{(I_1, I_2)} d_{sim}(I_1, I_2)$ for $I_1, I_2 \in G$.
- $r_I =$ rank of ASPR I in the gene.
- α_{rang} : Weighting between two different ASPR in the same gene. Taken $= d_{max}^{global}$

$d_{sim}(I_1, I_2)$ is the $-\log_{10}$ of the probability that profile HMM of ASPR I_1 is homologous to profile HMM of ASPR I_2, as computed by $hhalign/hhsearch$. We take the $-\log_{10}$ so that the higher is the probability the lower is the distance between the ASPRs. d_{max}^G is a weighting on the max distance between ASPRs of the same gene.
The rank of an ASPR I in a gene is intended to reflect the position of the ASPR in the protein sequence, it is computed as follow : we pick one species that contains all the ASPRs, sorted in ascending order a list of tuples containing the ASPRs and the start of their genomic coordinates, then replace the genomic coordinates by an increasing sequence of integers. This method for computing the rank of an ASPR in a gene, based on the genomic coordinates of the exons, could sometimes lead to errors. For example in the case of Exon shuffling, the order of the exons of a gene in two distinct genomes/species could be different, this situation is rare.

3.4.4. *Clustering of ASPRs with UPGMA.*
Once the distance matrix is created for the set of N ESGs (fig.8.B, center), we employed UPGMA to cluster their ASPRs. The rationale of this step is to avoid intra-gene clustering before inter-gene clustering. Neighbor-Joining was not suitable because for each ASPR it involves a sum over all the other ASPRs. Since we put a penalty on intra-gene pairs, it would yield negative branch length. The resulting grouping consists of O-ASRUs, that is ASPRs related by OPRs (i.e. similar ASPRs, each belonging to distincts genes), see fig.8.B, right. We ignored the fixed s-exons in this clustering step, and added them a posteriori to the O-ASRU to which they are most similar.

3.4.5. *Identification of Specificity-Determining-Sites.*

Then, we build the multiple sequence alignments for the different groups of O-ASRUs. Since it consists of similar sequences we can identify sites/AS-signatures potentially involved in the modulation of molecular recognition specificity, i.e. sites that vary between ASPRs but remain

conserved across species. We compute the consensus sequence of each ASPRs, by setting an exception parameter specifying the number of sequences allowed to differ from the consensus to 2. Positions exceeding this threshold are filled with a dot ".". We group the sequences according to the tree and align them with *Clustal O* [27].The alignment might produce gaps "-", that are artefacts to be distinguished from ".". The color scheme is that of GGMSA [28]. The symbols at the bottom denote highly conserved positions across the gene family: (dot) fully conserved position; (square) position conserved only within each O-ASRU; (upward triangle) position conserved in O-ASRU *a* only; (downward triangle) position conserved in O-ASRU *b* only.

3.4.6. *Structural Assessment.*
Finally, we carried out an analysis of structural models for the families where we identified AS duplications. We retrieved the ensemble of PDB chains that could be identified in the PDB (as of June 2022) for the genes of interest. It puts into correspondence the Ensembl Gene ID , the Uniprot ID, the PDB chain, and the matching intervals in Uniprot sequence and PDB chain. Then we clustered the PDB chains using MMseq2 at >80% seq. id. and >80% seq. coverage. When looking for interactions with partners, we consider, for each gene of interest, all PDB chains associated to the gene in Uniprot and we run the following analysis:

- identify the cluster to which the PDB chain belongs.
- identify interface residues in all PDB complexes (bio assemblies only, note that the NMR structures are not considered) involving the PDB chain or a member of its cluster (close homolog).
- map the interface residues onto the PDB chain.

The results of this analysis for all genes of interest led us for each PDB chain to get the labels for each residue, either 1 if non-interacting, 2 if interacting and 0 when no information was available. We recovered for a query PDB chain, its homolog and the partner, the ensemble of interacting sites identified in the associated PDB complexes. Sites are detected as the ensemble of residues closer then 5A to the partner.

3.4.7. *Structure of the output.*

In the end we produced MSAs for each trees/subtrees for all families. The output is as follow :

```
>ENSG00000007866@10_1-a
VSSHIQVLARKKVREYQVGIK
>ENSG00000187079@11_0-a
VSSHIQVLARKKVREIQAAIK
>ENSG00000007866@10_3-b
VCSHLQVLARRKSREIQSKLK
>ENSG00000187079@11_2-b
VSSHIQVLARRKSRDFHSKLK
>ENSG00000074219@2_1-b
..SHIQVLARRKSREIQSKLK
>ENSG00000197905@17_0-b
VSSHIQVLARRKAREIQAKLK
>signatures
+(++(+++++%+%+(7())%+
>ENSG00000187079@11_0-a$4z8e_A
21122112100000000000
>ENSG00000187079@11_0-a$5nnx_A
12222221122112000000
>ENSG00000187079@11_2-b$4z8e_A
21122112100000000000
>ENSG00000187079@11_2-b$5nnx_A
12222221122112000000
>ENSG00000197905@17_0-b$5gzb_A
122222211221121111111
>ENSG00000197905@17_0-b$5no6_I
12222221122112000000
>union
222222221221211111111
```

Figure 9. Output for the TEAD[1,2,3,4] family. The symbols in the sequence signatures denote highly conserved positions across the gene family: "+" fully conserved position; "%" position conserved only within each O-ASRU; "(" position conserved in O-ASRU *a* only; ")" position conserved in O-ASRU *b* only. For the interface mapping : "2" in interaction; "1" present but not interacting; "0" N/A.

4. Results

4.1. Statistical analysis of the I-ASRUs over the whole human proteome.

Since we wanted to scale the identification of I-ASRUs over the whole human proteome and their one-to-one orthologs across 12 species with the aim of analyzing I-ASRUs among members of families of paralogous genes, we used *sexDB* and the resulting intra/inter-gene similarity graph. In the end, we identified 1261 I-ASRUs spanned over 869 genes and comprising 4258 ASPRs. These I-ASRUs were defined starting from the detection of similar pairs of s-exons coming from the same gene, i.e. via the intra-gene similarity graph. Note that an I-ASRU necessarily has at least two ASPRs. We obtained 123 330 pairs of intra-gene s-exons and applied the filters described in 3.3.(2) on this set. The application of these filters reduced the number of s-exons pairs to 18 672, including 14 084 for which we have evidence of an alternative usage of the two s-exons in the ESG. Over the 14 084 pairs of similar s-exons involved in an ASE, 606 of them are MEX, 399 are ALT, 997 are REL and 12 082 are UNREL.

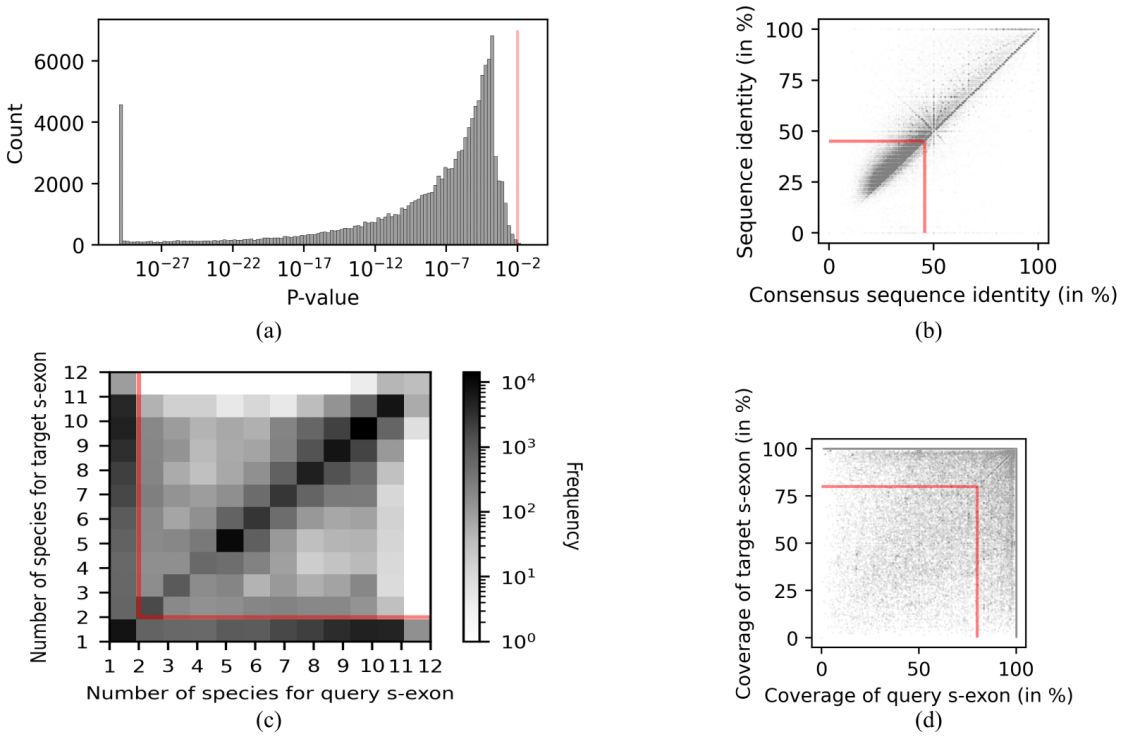


Figure 10. Statistical analysis over 123 330 unfiltered pairs of intra-gene s-exons. (a) **Histogram of the p-value over the unfiltered pairs of s-exons.** The s-exons in the pair comes from the same gene. There is 57 pairs with a p-value > 0.01, which is the cut-off value we used (represented with the red bar), the max has p-value of 0.47. The x-axis has log scale. (b) **Scatterplot of the consensus sequence identity against sequence identity over the unfiltered pairs.** The red bars represents the cut-off value which was set to 45% id. (c) **Heatmap of the number of species contained in the MSAs of both s-exons of the pair.** The red bars represents the cut-off value which was set to two species for both s-exons. (d) **Scatterplot of the coverage of the alignment for both s-exons of the pair.** The red bars represents the cut-off value which was set to 80%.

To understand the passage from 123 330 pairs of s-exons to 18 672, we plotted independently the distribution of these unfiltered pairs with respect to the p-value of the alignment (cut-off = 0.01), the consensus sequence identity and the sequence identity (cut-off = 45%) which is the identity between the alignment of the first sequences in the MSA of the two s-exons, the number of species in the MSAs of both s-exons of the pair (cut-off = 2 species) and the coverage of the alignment for both s-exons of the pair (cut-off = 80%). Each time the cut-off value was represented with red bars that partition the space into two sub-space, each corresponding to the accepted pairs and the rejected ones. In fig.10.a, 57 pairs are above the cut-off value. In fig.10.b 79 001 pairs have consensus sequency identity < 45% or sequence identity < 45% (negation of the corresponding criterion). In fig.10.c 48 228 pairs are such that one of the s-exon of the pair is species-specific, that is it comprises only one sequence (coming from one species). In fig.10.d 12 709 s-exons pairs are such that the coverage of the alignment for both s-exons of the pair are < 80 %. Those quantities were obtained by applying each filter independently on the unfiltered pairs, hence adding the (independently) lost pairs doesn't amounts to 123 330 - 18 672 = 104 658 lost pairs, this mismatch is due to an effect of applying the filters simultaneously (some pairs lost with the identity criterion might contain pairs where one of the s-exon is species-specific etc) . We can observe that it is with the criterion over the consensus sequence identity and the number of species that we lost most pairs.

We can see in fig.10.a that a bit more than half of the I-ASRUs are homogeneous in length (in aa) over the ASPRs. The bar at the left i.e. cases of heterogeneity might contain cases where we have a very long ASPR and the alternative one is found multiple times inside the long one. In fig?.(b) we can observe that most genes have a single I-ASRU. There are some extreme cases, with at most 36 I-ASRUs for MYH6 (myosin heavy chain), MGAM (Maltase-glucoamylase) with 22 I-ASRUs, ATP1A2 with 13 I-ASRUs, Carboxylesterase 1 with 12 I-ASRUs. These cases suggests that we are dealing with a duplicated domain.

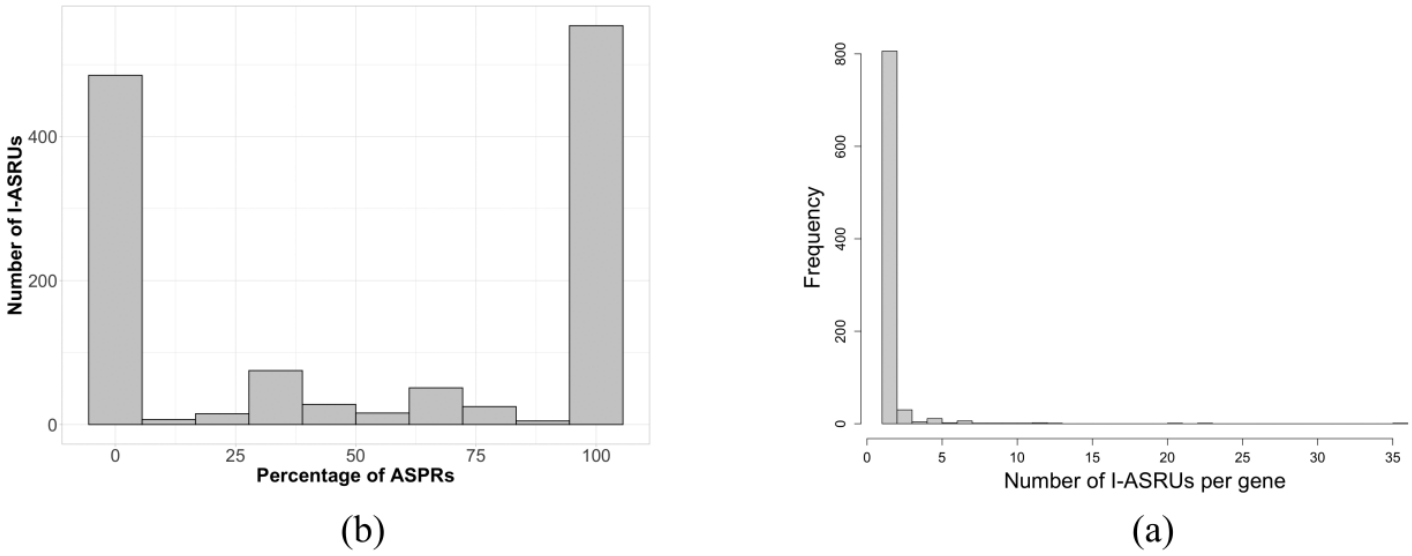


Figure 11. (a) Number of I-ASRUs where the distance of the ASPRs to the median is at most 5 amino acids. (b) Number of I-ASRUs per gene.

Most genes have a single I-ASRU containing two ASPRs that are related by a single ASE, Nebulin (NEB) is the gene that has the I-ASRU with the highest number of ASPRs, namely 84 ASPRs, and such that the number of ASE linking them is less, namely 40, see fig.12.left. We recover scenarios of genes having lots of I-ASRUs populated with lots of ASPRs (HSPG2)

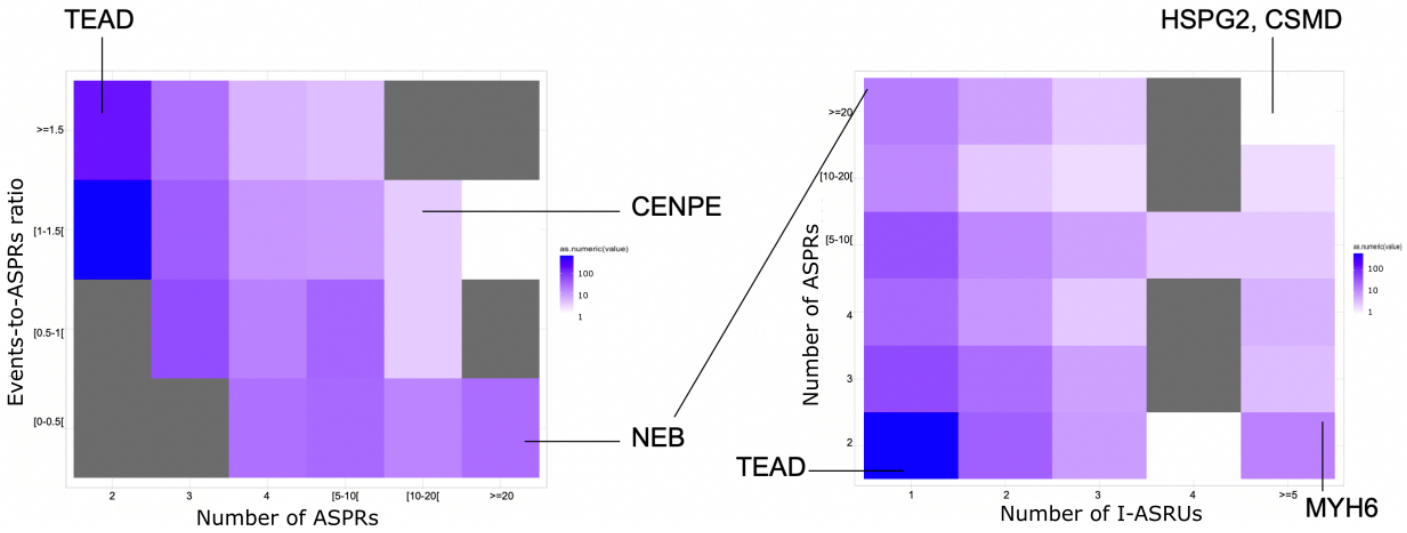


Figure 12. Left.Heatmap of the number of instances versus the number of I-ASRUs per gene. Right. Heatmap of the ratio events-to-ASPRs against the number of instances of an ASRU. This ratio is calculated by dividing the number of events involving an I-ASRU by the number of ASPRs of that I-ASRU minus 1. (For example, 2 ASPRs linked by 1 event results in a ratio of 1). The grey cells correspond to zero values. The color scale is logarithmic.

Figure 13 is concordant with fig.11.a, where the two violet rows located at <1 aa and [10-23[aas in standard deviation corresponds to a large fraction of the right most bar and left most bar in fig?.(a) respectively. This reflects a variability in the length of ASPRs connected by an IPR, suggesting that the pseudo-repeat structure is not preserved by splicing.

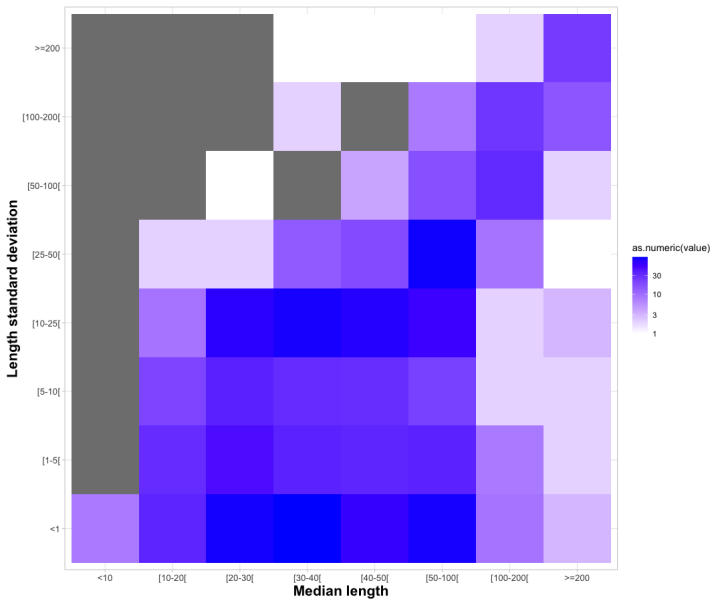


Figure 13. Heatmap of standard deviation on lengths versus median over the sizes of the ASPRs in an I-ASRU (in aas)

4.2. **Analysis of entire protein families**

4.2.1. *Families of paralogous genes.*

Were found using a filtering value of 60% sequence identity, 1933 families spanned over 9640 genes. From those 9640 genes, 3146 of them are in common with the 17 920 genes treated by ThorAxe. By further retaining families that contained at least two genes with at least one I-ASRU, we reduced the number of families to 33, spanned over 82 genes.

4.2.2. *sexDB.*

sexDB contains 416 485 s-exons spread over 17 920 genes. Each gene/ESG has on average 23 s-exons (see fig.15.a). The average length of an s-exon is 45 amino acid (see fig.15.b). 750 genes have a single s-exon with a mean length over those 750 s-exons of 305 aa.
sexDB enabled the construction of an intra/inter-gene similarity graphs. With this intra-gene similarity graph, we applied our method of identification of I-ASRUs to the 33 families. Over these 33 families, we retrieved 165 I-ASRUs and 925 ASPRs. An additional set of 23 genes that do not have any I-ASRU, but that contains fixed s-exons were rescued by means of the inter-gene similarity graph. Indeed we looked at adjacent s-exons to the 925 ASPRs, filtered the edges with respect to the above criteria, and rescued those that were inside a paralog.

4.2.3. *Detection of O-ASRU and visualisation.*

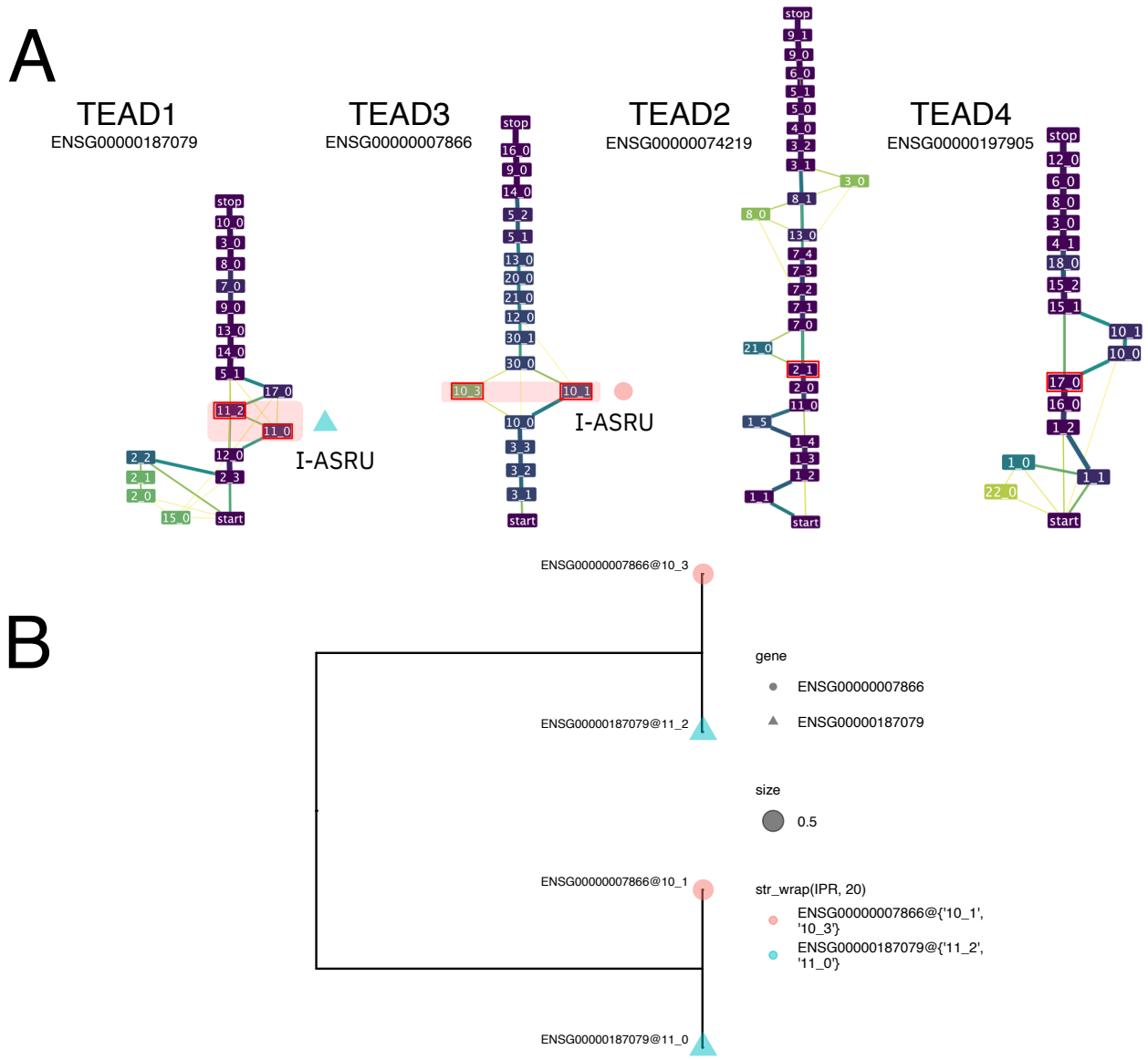
For families of two paralogs comprising two I-ASRUs, the analysis can be easily automated. A simple tree consisting in two subtrees, where the leaves of each subtrees are ASPRs coming from distinct genes is constructed and the identification of O-ASRU is straightforward. For more complex trees such as the one of the CACNA1[C,D,F] family, the detailed analysis had to be done manually.

We first identify the I-ASRUs in each ESG by giving them a colored symbol (panel A). We then perform the tree clustering on the ASPRs to define O-ASRUs (panel B). Finally we can compare sequence signatures of the O-ASRUs with information from interaction and structure (panel C and D). For panel C, the number on the right of each ASPR's consensus sequence in the MSAs are the number of species in the MSAs of each ASPRs.

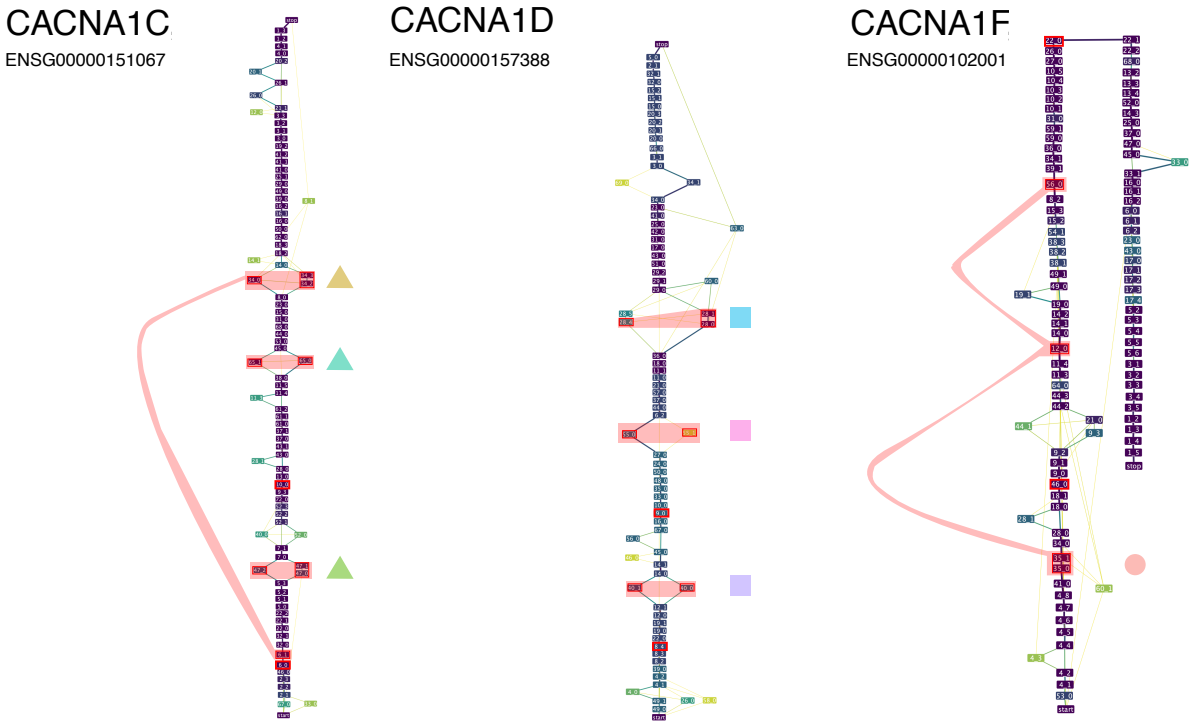
In the analysis of the TEAD family (Transcriptional enhancer factor), TEAD[1,2,3,4], which is a transcription factor, we detected a different alternative usage of pseudo-repeats, REL in TEAD1 (insertion or not of 11_0 in the transcript), MEX in TEAD3, and a sub-functionalization by the fixation of one of the s-exon in TEAD2 and TEAD4, (panel A). The 3D structure are AlphaFold Models of the canonical and alternative proteoform of TEAD3 (panel D). It is represented as a cartoon: the non interacting surface, as a surface: residues interacting with at least one partner, based on experimental structures of the protein itself or its close homologs at >80% seq. id, as sticks: s-exon 10_1. We can see that the s-exon 10_1 is located in the region at the DNA binding interface. Moreover we can see through the MSA that the AS-induced variation involving the transition between a valine and a serine is conserved across species and across paralogs (square, panel C). These AS-signatures (square, upper/lower triangle) inform us about the potential modulation of the specificity of the interactions. Indeed for TEAD we successfully provided a structural assessment of the pseudo-repeat of interest, and we can observe that variable residues are interacting residues.

For the CACNA1[C,D,F] family (Voltage-dependent L-type calcium channel subunit alpha-1[C,D,F]), our results are concordant with those of (Abascal et al. 2015), with some additional findings. CACNA1C has three I-ASRUs scattered along the protein and one fixed s-exon. CACNA1D has three I-ASRUs and two fixed s-exons. CACNA1F has one I-ASRUs and two fixed s-exons (panel A). The topology of CACNA1F is different than the other paralogs, the IPRs which makes up the I-ASRU are UNREL thus these ASE status are not strong enough compared to the other paralogs. The ASPR 47_2 in CACNA1C is divergent in its O-ASRU (panel C), if it is not under-representated in terms of species it might suggest a modulation of the specificity of the interactions. Indeed in the column "YFYRW", they are all aromatic, except the arginine ("R" residue) which is positively charged. It can be an indicator of a change of interacting partners for the proteoform of this paralog.

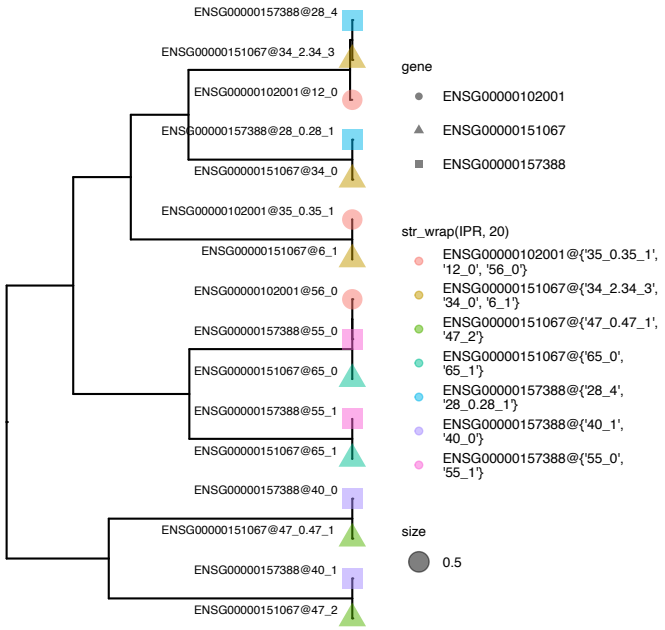
The SCN[1,2,3,4,5,8,9]A family (Sodium channel protein type [1,2,3,4,5,8,9] subunit alpha), is the family with the highest number of paralogs. Both members of the I-ASRUs are conserved across the paralogs. Except for SCN4A where the ASPR is fixed in the protein (i.e. not alternatively used).


A
TEAD1
ENSG00000187079
TEAD3
ENSG00000007866
TEAD2
ENSG00000074219
TEAD4
ENSG00000197905
I-ASRU
I-ASRU
B
ENSG00000007866@10_3
ENSG00000187079@11_2
ENSG00000007866@10_1
ENSG00000187079@11_0
gene
ENSG00000007866
ENSG00000187079
size
0.5
str_wrap(IPR, 20)
ENSG00000007866@{'10_1', '10_3'}
ENSG00000187079@{'11_2', '11_0'}
C
O-ASRU α
O-ASRU β
:interaction, :no interaction, :N.A.
D

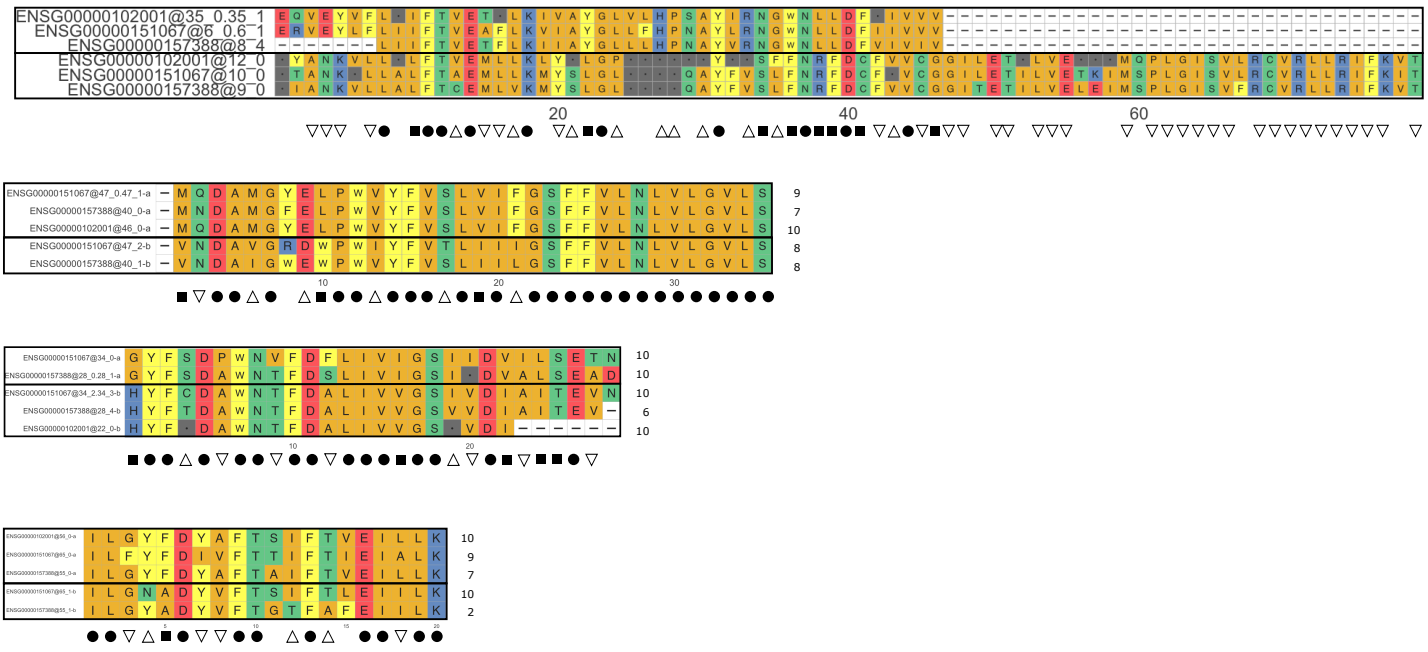
A


CACNA1C
ENSG00000151067
CACNA1D
ENSG00000157388
CACNA1F
ENSG00000102001

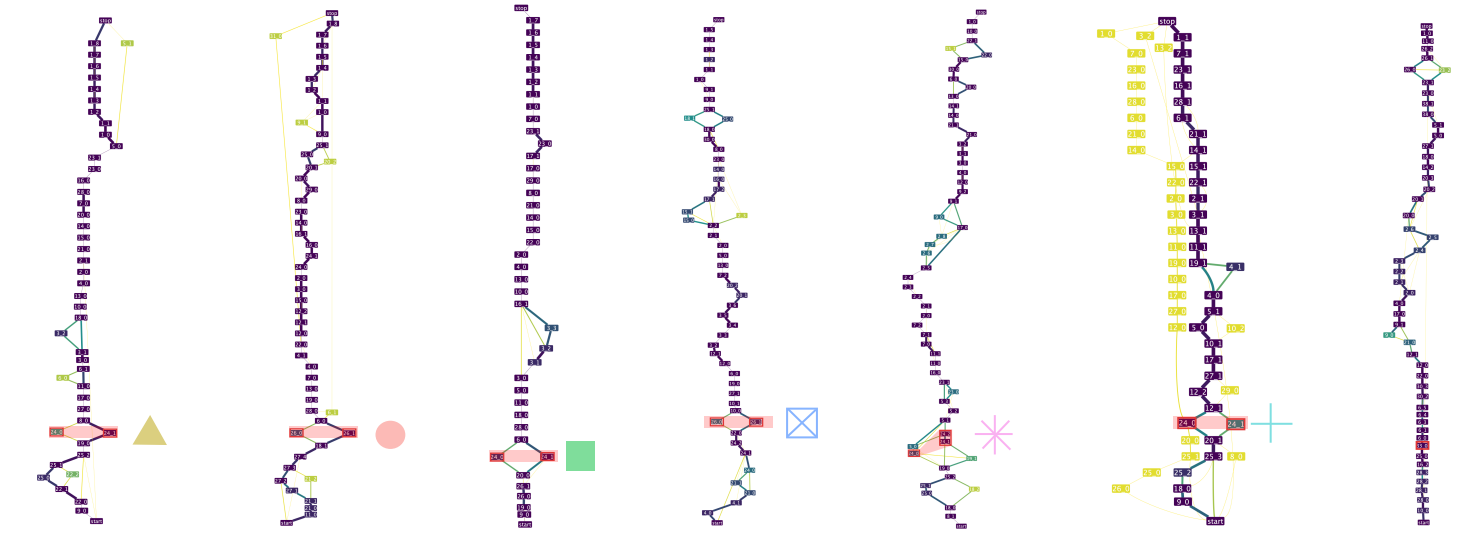
B


ENSG00000157388@28_4
ENSG00000151067@34_2.34_3
ENSG00000102001@12_0
ENSG00000157388@28_0.28_1
ENSG00000151067@34_0
ENSG00000102001@35_0.35_1
ENSG00000151067@6_1
ENSG00000102001@56_0
ENSG00000157388@55_0
ENSG00000151067@65_0
ENSG00000157388@55_1
ENSG00000151067@65_1
ENSG00000157388@40_0
ENSG00000151067@47_0.47_1
ENSG00000157388@40_1
ENSG00000151067@47_2
gene
ENSG00000102001
ENSG00000151067
ENSG00000157388
str_wrap(IPR, 20)
ENSG00000102001@{'35_0.35_1', '12_0', '56_0'}
ENSG00000151067@{'34_2.34_3', '34_0', '6_1'}
ENSG00000151067@{'47_0.47_1', '47_2'}
ENSG00000151067@{'65_0', '65_1'}
ENSG00000157388@{'28_4', '28_0.28_1'}
ENSG00000157388@{'40_1', '40_0'}
ENSG00000157388@{'55_0', '55_1'}
size
0.5

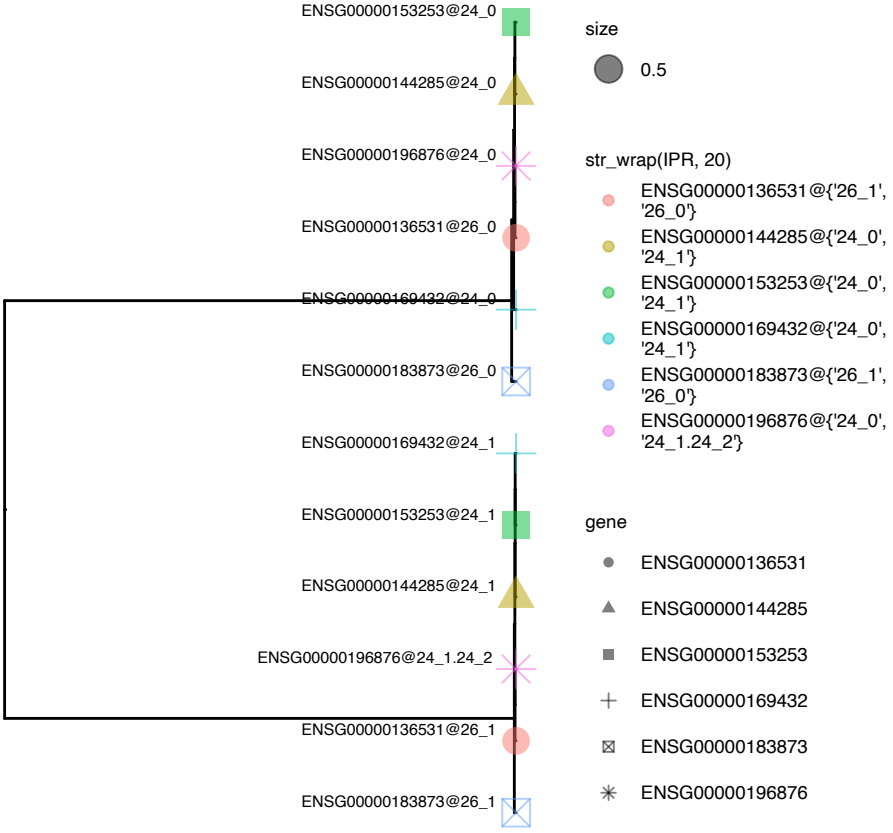
C


ENSG00000102001@35_0.35_1
ENSG00000151067@6_1
ENSG00000157388@8_4
ENSG00000102001@12_0
ENSG00000151067@10_0
ENSG00000157388@9_0
ENSG00000151067@47_0.47_1-a
ENSG00000157388@40_0-a
ENSG00000102001@46_0-a
ENSG00000151067@47_2-b
ENSG00000157388@40_1-b
ENSG00000151067@34_0-a
ENSG00000157388@28_0.28_1-a
ENSG00000151067@34_2.34_3-b
ENSG00000157388@28_4-b
ENSG00000102001@22_0-b

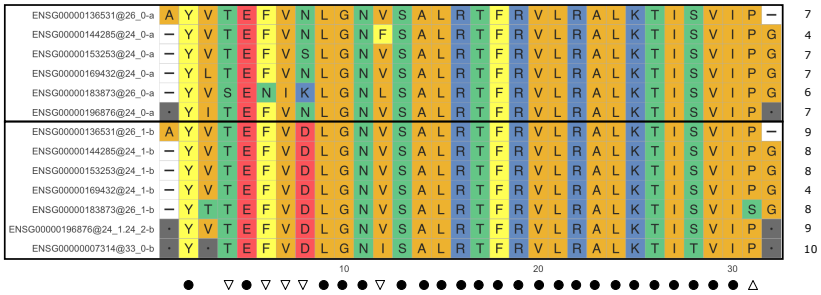
A


SCN1A
ENSG00000144285
SCN2A
ENSG00000136531
SCN3A
ENSG00000153253
SCN5A
ENSG00000183873
SCN8A
ENSG00000196876
SCN9A
ENSG00000169432
SCN4A
ENSG00000007314

B


ENSG00000153253@24_0
ENSG00000144285@24_0
ENSG00000196876@24_0
ENSG00000136531@26_0
ENSG00000169432@24_0
ENSG00000183873@26_0
ENSG00000169432@24_1
ENSG00000153253@24_1
ENSG00000144285@24_1
ENSG00000196876@24_1.24_2
ENSG00000136531@26_1
ENSG00000183873@26_1
size
0.5
str_wrap(IPR, 20)
ENSG00000136531@{'26_1', '26_0'}
ENSG00000144285@{'24_0', '24_1'}
ENSG00000153253@{'24_0', '24_1'}
ENSG00000169432@{'24_0', '24_1'}
ENSG00000183873@{'26_1', '26_0'}
ENSG00000196876@{'24_0', '24_1.24_2'}
gene
ENSG00000136531
ENSG00000144285
ENSG00000153253
ENSG00000169432
ENSG00000183873
ENSG00000196876

C


ENSG00000136531@26_0-a
ENSG00000144285@24_0-a
ENSG00000153253@24_0-a
ENSG00000169432@24_0-a
ENSG00000183873@26_0-a
ENSG00000196876@24_0-a
ENSG00000136531@26_1-b
ENSG00000144285@24_1-b
ENSG00000153253@24_1-b
ENSG00000169432@24_1-b
ENSG00000183873@26_1-b
ENSG00000196876@24_1.24_2-b
ENSG00000007314@33_0-b
10
20
30

5. **Limits of the computational approach**

There are multiple limitations we could point out : Thoraxe relies on gene annotations, which may be partial, incomplete, or erroneous. There is a lack of structural data regarding complexes that prevent us from doing the mapping of the residues lying at the interface of the interactions. The values for the filtering threshold prevent us from retrieving all the relevant detections, indeed some similarities are not detected because the identity percentage is equal to 44, some families are not recovered because their identity percentage is slightly below 60. The task of finding the right threshold for determining O-ASRUs from the UPGMA trees could not be automated in the allowed time.

6. **Comparison with existing approaches**

Compared to other approaches [29-31], ThorAxe is applicable at a much larger scale due to its higher computational efficiency and its ability to deal with many species. Moreover, it is the first and only to effectively reconstruct a pan-transcriptome across multiple species. Zea et al. did an upstream validation on a benchmark set of 97 curated genes where they had a good overlap with (Abascal et al. 2015) [9] study. The current work retrieve those results. I performed a manual check on a well-known set of 10 genes with the aim of identifying empirically filtering criteria. Moreover I did a manual comparison with the study of (Abascal et al. 2015) [12] of our results and theirs for the analysis of the paralogous families. We have in common the following genes : GRIA[1,2,3,4], DNM[1,2], MAPK[8,9,10], ACTN[2,4], ACSL[1,6], SCN[1,2,3,5,8,9]A and CACNA1[C,D,F], where for example for the latter we have the following mapping between their exons and our ASPRs : CACNA1C (E8a) = ASPR 47_1, CACNA1C (E8b) = ASPR 47_2, CACNA1D (E8a) = ASPR 40_0, CACNA1D (E8b) = ASPR 40_1, etc. ThorAxe and ASPRine, which are versatile, user-friendly and fully automated, produce results which agree with and expand results obtained from years of efforts combining semi-automated pipelines and manual curation. We also have another confirmation for two O-ASRUs of CACNA1C from (Garcia, A.B. et al 2020, [32]).

7. **Conclusion**

We have presented a novel method to analyse a sequence conservation/divergence signal across three dimensions, namely across species, paralogs and proteoforms towards identifying molecular determinants of protein interaction specificity. This is not an easy problem as it involves a variety of objects, duplicates and similar objects, that needed to be disentangled before analysing them. Our approach provides a nomenclature that guides the detection of cases of the interchangeable and non-interchangeable model. We provided a tool that takes as input a gene and outputs information about its I-ASRUs and their ASPRs. To illustrate the potential of the method, we scaled the analysis of the I-ASRUs to the whole human proteome and their one-to-one orthologs across 12 species. Finally this analysis was extended to paralogous families, the comparison of similar ASPRs unveiled functionally relevant AS-signatures.

Appendix A. **Parameter of the developed tool and an example**

An example of command line for launching the script is :

```
python3 gene2ASRU/scripts/withoutDB/main.py \
--path_data ~/lab_bench --path_hhsuite ~/hhsuite \
--gene ENSG00000107643 --len 5 --id 100 --norealign 1 \
--glo_loc 1 --mact 0 --id_pair 45 --idCons_pair 45 \
--pval 0.001 --nbSpe 2 --cov 0.8
```

where the required arguments are :

- [**path_data**] is path to folder containing Thoraxe's output
- [**path_hhsuite**] is path to hhsuite directory
- [**gene**] is Ensembl Gene ID
- [**len**] is minimum length of an MSA in order to create its HMM profile (def=5)
- [**id**] is maximum pairwise sequence identity (%) in MSA (def=100)
- [**norealign**] bool, 1 if norealign else 0, do NOT realign displayed hits with Maximum Accuracy algorithm (MAC)
- [**glo_loc**] bool, 1 if global else 0, use global/local alignment mode for searching/ranking (def=local)
- [**mact**] [0,1[posterior prob threshold for MAC realignment controlling greediness at alignment ends. 0:global >0.1:local (default=0.35) [id_pair] Identity percentage threshold between first sequence in msa of s-exon for each s-exon in a pair'
- [**idCons_pair**] Identity percentage threshold between consensus sequence of msa of s-exon for each s-exon in a pair (should be equal to id_pair)
- [**pval**] p-value threshold for HMM-HMM alignment of a s-exons pair
- [**nbSpe**] minimum number of species in msa for s-exons in the pair
- [**cov**] Threshold for coverage of s-exon A and B in alignment of A and B

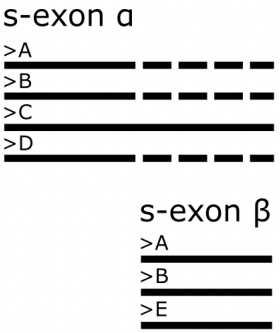
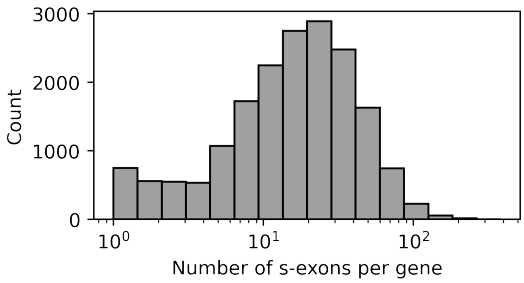


Figure 14. **False positive alignment.**The sequence detected as a repeat exists only in a species, *C*, that does not appear in the MSA of the other s-exon of the pair. The dashed lines are gaps.

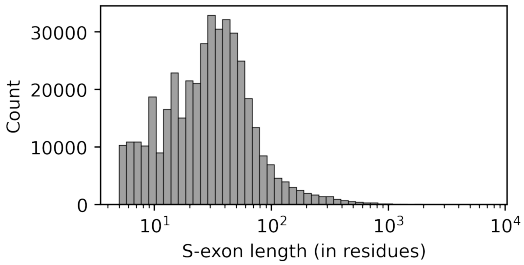
Appendix B. **Creation of** *sexDB*

```
scripts/reformat.pl {}.fasta {}.a2m
bin/hhmake -i {}.a2m -o {}.hhm -v 0 -name {} -id 100 -M a2m
```

where the **id** filtering option is the maximum pairwise sequence identity (%) in the MSA of a s-exon.



(a) **Number of s-exons per gene for s-exon in *sexDB*.** The mean is at 23 s-exons. Titin is the gene having the largest number of s-exons, namely 385 s-exons, and also the longest one (in aa). The x-axis has log scale.



(b) **S-exon length for s-exon in *sexDB*.** The mean is at 45aa, the min is at 5aa. Mucin 16 is the gene having the longest s-exon which is 7231 aa long. The x-axis has log scale.

Figure 15

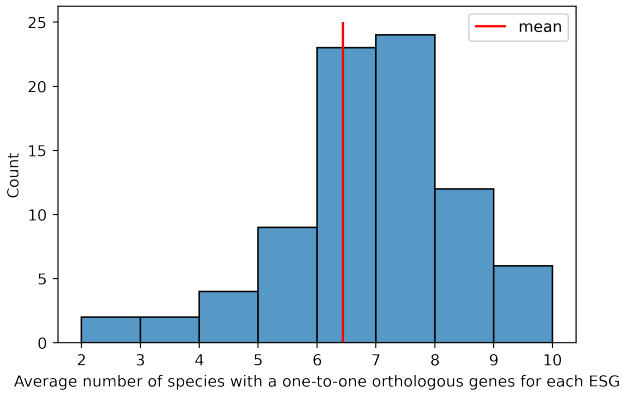


Figure 16. Average number of species with a one-to-one orthologous genes over the 82 ESGs. The mean is located around 6,7 species.

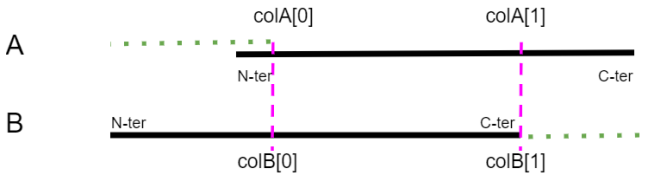


Figure 17. General case where for a pair of s-exons (A,B) aligned between the pink dotted lines, one has the possibility to extend A in N-ter and B in C-ter, the margins are represented in green. In this case, $margin_B^{Nter}$ and $margin_A^{Cter}$ are positive integers (representing a length in aa). The black bars represent the consensus sequences of the s-exons (=MSA). Further conditions for the extension process is that an adjacent should have a degree ≥ 1 to be a candidate. This choice was made so that we stay in the context of repeated regions. When an adjacent s-exon is inferior to 5 aa in length we automatically extend the seed with it.

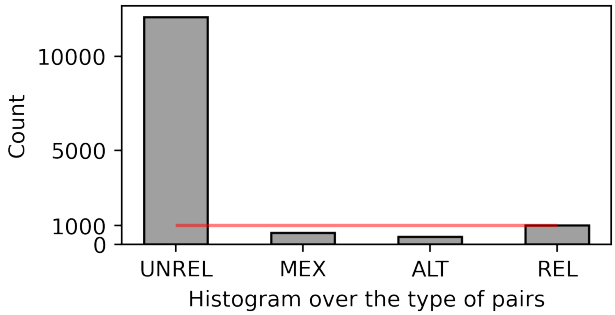


Figure 18. Histogram over the type of pairs. From the 14 084 pairs of similar s-exons involved in an ASE, 606 of them are MEX, 399 are ALT, 997 are REL and 12 082 are UNREL.

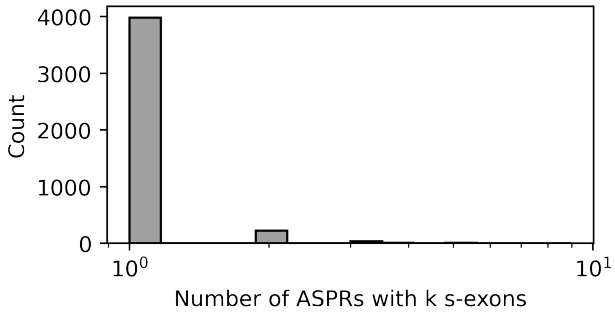


Figure 19. Number of ASPRs containing k s-exons. Among the 4258 ASPRs, 3981 remained composed of a single s-exon, 221 of two s-exons, 33 of three s-exons, 8 of four s-exons, 7 of five s-exons, 3 of six s-exons, 2 of seven s-exons, 2 of eight s-exons and one of 9 s-exons.

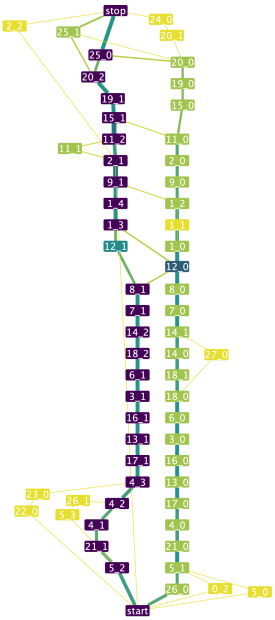


Figure 20. **ESG of ATP1A4.** The topology of the ESG is split into two looks like a global error of annotation on the gene. This is also the case for SCN9A.

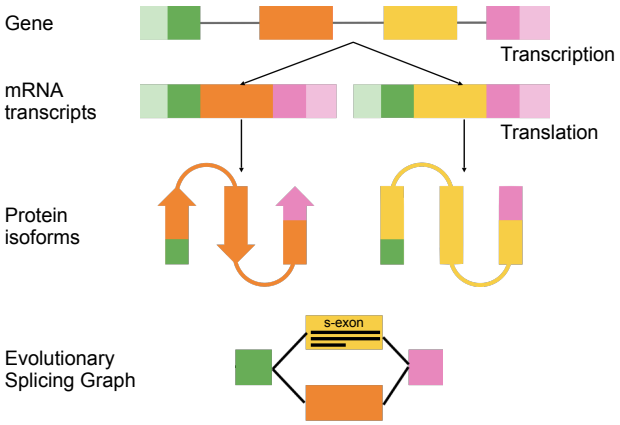
Alternative splicing modulates the number and composition of similar exonic regions

Antoine Szatkownik[1,2], 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

1. Challenges

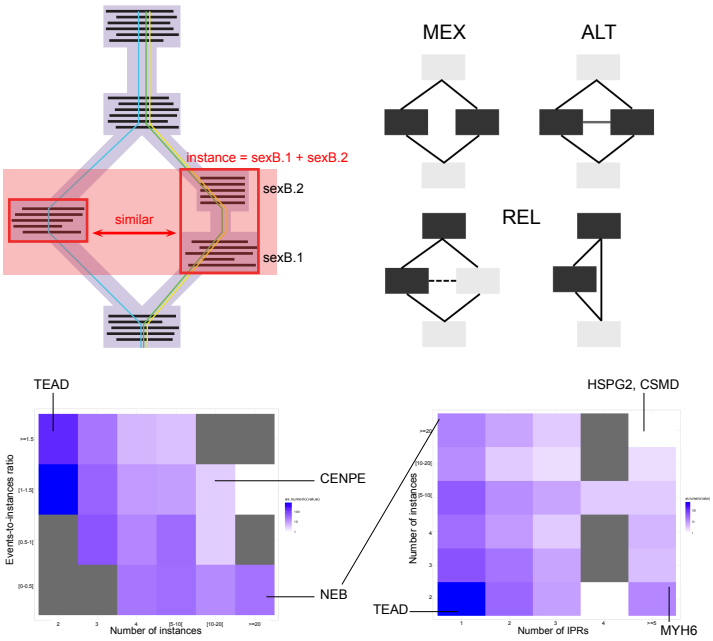
- Define meaningful protein «themes» alternatively used in evolution.
- Identifiying signatures for molecular recognition specificity?

2. Evolution-informed representation of proteoforms



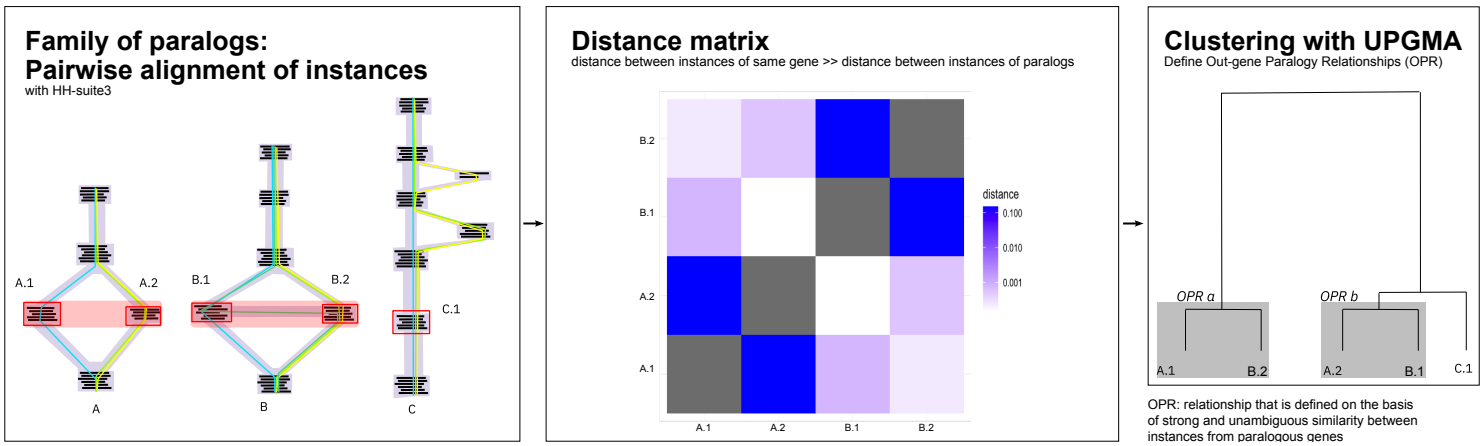
46 000 evolutionary conserved alternative splicing events coming from 8 000 genes, shared among a dozen of species, from human to nematode

3. In-Gene Duplication



14 000 pairs of similar s-exons spanned over 869 genes

4. Methodology

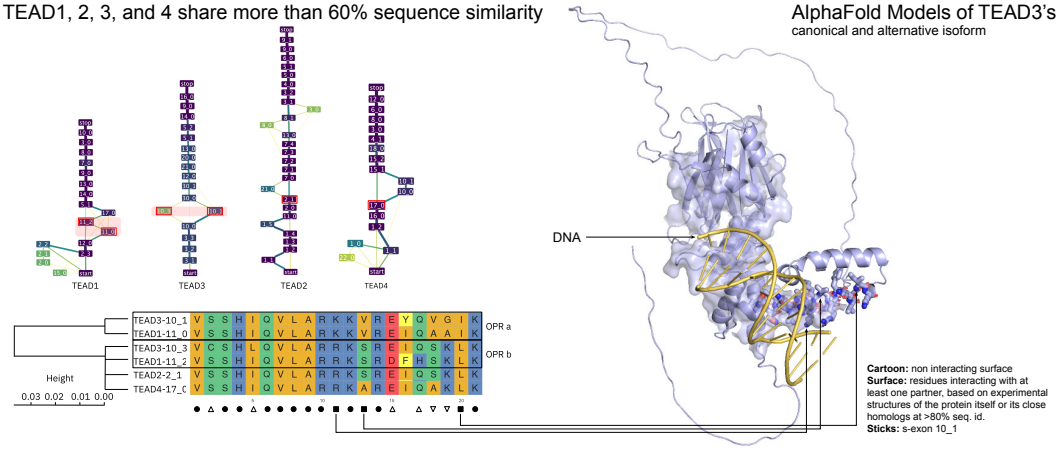


OPR: relationship that is defined on the basis of strong and unambiguous similarity between instances from paralogous genes

5. Case Study: TEAD [1,2,3,4]

TEAD1, 2, 3, and 4 share more than 60% sequence similarity

AlphaFold Models of TEAD3's canonical and alternative isoform



The comparison of similar s-exons unveils functionally relevant signatures

6. Perspectives

- Structural assessment of alternative isoforms/s-exons
- Alternative Splicing-inspired protein design

References

Zea et al Genome Research 2021
Zea et al Bioinformatics 2022
www.lcqb.upmc.fr/Ases

SCIENCES SORBONNE UNIVERSITÉ
ROBERT KOCH INSTITUT

References

[1] Ohno, S; Evolution by Gene Duplication; *Springer Berlin Heidelberg*: Berlin, Heidelberg, **1970**.

[2] Ast, G.; How Did Alternative Splicing Evolve? *Nat. Rev. Genet.*, **2004**, 5 (10), 773–782, doi: 10.1038/nrg1451.

[3] Smith, L. M.; Kelleher, N. L.; Proteoforms as the next Proteomics Currency. *Science*, **2018**, 359 (6380), 1106–1107, doi: 10.1126/science.aat1884.

[4] Birzele, F.; Csaba, G.; Zimmer, R.; Alternative Splicing and Protein Structure Evolution. *Nucleic Acids Res.*, **2008**, 36 (2), 550–558, doi: 10.1093/nar/gkm1054.

[5] Yang, X.; Coulombe-Huntington, J.; Kang, S.; Sheynkman, G. M.; Hao, T.; Richardson, A.; Sun, S.; Yang, F.; Shen, Y. A.; Murray, R. R.; Spirohn, K.; Begg, B. E.; Duran-Frigola, M.; MacWilliams, A.; Pevzner, S. J.; Zhong, Q.; Trigg, S. A.; Tam, S.; Ghamsari, L.; Sahni, N.; Yi, S.; Rodriguez, M. D.; Balcha, D.; Tan, G.; Costanzo, M.; Andrews, B.; Boone, C.; Zhou, X. J.; Salehi-Ashtiani, K.; Charloteaux, B.; Chen, A. A.; Calderwood, M. A.; Aloy, P.; Roth, F. P.; Hill, D. E.; Iakoucheva, L. M.; Xia, Y.; Vidal, M.; Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, **2016**, 164 (4), 805–817, doi: 10.1016/j.cell.2016.01.029.

[6] Kelemen, O.; Convertini, P.; Zhang, Z.; Wen, Y.; Shen, M.; Falaleeva, M.; Stamm, S.; Function of Alternative Splicing. *Gene*, **2013**, 514 (1), 1–30, doi: 10.1016/j.gene.2012.07.083.

[7] Tapial, J.; Ha, K. C. H.; Sterne-Weiler, T.; Gohr, A.; Braunschweig, U.; Hermoso-Pulido, A.; Quesnel-Vallières, M.; Permanyer, J.; Sodaei, R.; Marquez, Y.; Cozzuto, L.; Wang, X.; Gómez-Velázquez, M.; Rayon, T.; Manzanares, M.; Ponomarenko, J.; Blencowe, B. J.; Irimia, M.; An Atlas of Alternative Splicing Profiles and Functional Associations Reveals New Regulatory Programs and Genes That Simultaneously Express Multiple Major Isoforms. Genome Res., **2017**, 27 (10), 1759–1768. doi: 10.1101/gr.220962.117.

[8] Baralle, F. E.; Giudice, J.; Alternative Splicing as a Regulator of Development and Tissue Identity. *Nat. Rev. Mol. Cell Biol.*, **2017**, 18 (7), 437–451, doi: 10.1038/nrm.2017.27.

[9] Roux J; Robinson-Rechavi M; Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res.*, **2011**, doi: 10.1101/gr.113803.110.

[10] Iñiguez LP; Hernández G; The Evolutionary Relationship between Alternative Splicing and Gene Duplication. *Front Genet.*, **2017**, doi: 10.3389/fgene.2017.00014.

[11] Talavera D; Vogel C; Orozco M; Teichmann SA; de la Cruz X; The (in)dependence of alternative splicing and gene duplication. *PLoS Comput Biol.*, **2007**, doi: 10.1371/journal.pcbi.0030033.

[12] Abascal, F.; Valencia, A.; Tress, M. L.; The Evolutionary Fate of Alternatively Spliced Homologous Exons after Gene Duplication. *Genome Biology and Evolution*, **June 2015**, Volume 7, Issue 6, Pages 1392–1403, doi: 10.1093/gbe/evv076.

[13] Lambert MJ; Cochran WO; Wilde B;, Olsen KG; Cooper CD; Evidence for widespread subfunctionalization of splice forms in vertebrate genomes. *Genome Res.*, **2015**, doi: 10.1101/gr.184473.114.

[14] Abascal, F.; Ezkurdia, I.; Rodriguez-Rivas, J.; Rodriguez, J. M.; del Pozo, A.; Vázquez, J.; Valencia, A.; Tress, M. L.; Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. *PLoS Comput. Biol.*, **2015**, 11 (6), e1004325, doi: 10.1371/journal.pcbi.1004325.

[15] Rodriguez, J. M.; Pozo, F.; di Domenico, T.; Vazquez, J.; Tress, M. L.; An Analysis of Tissue-Specific Alternative Splicing at the Protein Level. PLoS Comput. Biol., **2020**, 16 (10), e1008287, doi: 10.1371/journal.pcbi.1008287.

[16] Martinez Gomez, L.; Pozo, F.; Walsh, T. A.; Abascal, F.; Tress, M. L.; The Clinical Importance of Tandem Exon Duplication-Derived Substitutions. *Nucleic Acids Res.*, **2021**, 49 (14), 8232–8246, doi: 10.1093/nar/gkab623.

[17] Diego Javier Zea; Sofya Laskin; Hugues Richard; Elodie Laine; Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs. *Biorxiv* **2020**, doi: 10.1101/2020.11.14.382820.

[18] Zea, D. J.; Richard, H.; Laine, E; ASES: visualizing evolutionary conservation of alternative splicing in proteins. *Bioinformatics*, **2022**, doi: 10.1093/bioinformatics/btac105.

[19] Steffen Heber; Max Alekseyev; Sing-Hoi Sze; Haixu Tang; Pavel A. Pevzner; Splicing graphs and EST assembly problem. *Bioinformatics*, Volume 18, Issue suppl_1, **July 2002**, Pages S181–S188, doi: 10.1093/bioinformatics/18.suppl_1.s181.

[20] Paladin L; Bevilacqua M; Errigo S; Piovesan D; Mičetić I; Necci M; Monzon A; Fabre ML; Lopez JL; Nilsson J; Rios J; Lorenzano Menna P; Cabrera M; Gonzalez Buitron M; Gonçalves Kulik M; Fernandez-Alberti S; Silvina M; Parisi G; Lagares A; Hirsh L; Andrade-Navarro MA; Kajava AV; Tosatto SCE; RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Research*, **2020**, doi: 10.1093/nar/gkaa1097.

[21] Kolodny Rachel; Nepomnyachiy, Sergey; Ben-Tal, Nir; Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *PNAS*, **2017**, doi: 10.1073/pnas.1707642114

[22] Kolodny Rachel; Nepomnyachiy, Sergey; Tawfik, Dan S; Ben-Tal, Nir; Bridging Themes: Short Protein Segments Found in Different Architectures. *Mol Biol Evol*, **2021**, doi: 10.1093/molbev/msab017

[23] Qiu K; Ben-Tal N; Kolodny R.; Similar protein segments shared between domains of different evolutionary lineages. *Protein Sci.*, **2022**,doi: 10.1002/pro.4407.

[24] Yang X; Coulombe-Huntington J; Kang S; Sheynkman GM; Hao T; Richardson A; Sun S; Yang F; Shen YA; Murray RR; Spirohn K; Begg BE; Duran-Frigola M; MacWilliams A; Pevzner SJ; Zhong Q; Trigg SA; Tam S; Ghamsari L; Sahni N; Yi S; Rodriguez MD; Balcha D; Tan G; Costanzo M; Andrews B; Boone C; Zhou XJ; Salehi-Ashtiani K; Charloteaux B; Chen AA; Calderwood MA; Aloy P; Roth FP; Hill DE; Iakoucheva LM; Xia Y; Vidal M.; Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, **2016**, Feb 11;164(4):805-17, doi: 10.1016/j.cell.2016.01.029.

[25] Abhijit Chakraborty; Saikat Chakrabarti; A survey on prediction of specificity-determining sites in proteins, *Briefings in Bioinformatics*, Volume 16, Issue 1, **January 2015**, doi: 10.1093/bib/bbt092.

[26] 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**, doi: 10.1186/s12859-019-3019-7.

[27] Sievers F; Wilm A; Dineen DG; Gibson TJ; Karplus K; Li W; Lopez R; McWilliam H; Remmert M; Söding J; Thompson JD; Higgins DG; Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **2011**, doi: 10.1038/msb.2011.75.

[28] Zhou, Lang;Feng, Tingze; Xu, Shuangbin; Gao, Fangluan; Lam, Tommy T; Wang, Qianwen; Wu, Tianzhi; Huang, Huina; Zhan, Li; Li, Lin; Guan, Yi; Dai, Zehan; Yu, Guangchuang; ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Briefings in Bioinformatics*, **2022**, doi: 10.1093/bib/bbac222

[29] Blanquart, S.; Varré, J.-S.; Guertin, P.; Perrin, A.; Bergeron, A.; Swenson, K. M.; Assisted Transcriptome Reconstruction and Splicing Orthology. *BMC Genomics*, **2016**, doi: 10.1186/s12864-016-3103-6.

[30] Jammali, S.; Aguilar, J.-D.; Kuitche, E.; Ouangraoua, A. SplicedFamAlign: CDS-to-Gene Spliced Alignment and Identification of Transcript Orthology Groups. *BMC Bioinformatics*, **2019**, doi: 10.1186/s12859-019-2647-2.

[31] Márquez, Y.; Mantica, F.; Cozzuto, L.; Burguera, D.; Hermoso-Pulido, A.; Ponomarenko, J.; Roy, S. W.; Irimia, M. ExOrthist: A Tool to Infer Exon Orthologies at Any Evolutionary Distance. *Genome Biol.*, **2021**, doi: 10.1186/s13059-021-02441-9.

[32] Clark, M.B.; Wrzesinski, T.; Garcia, A.B. et al; Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. Mol Psychiatry,**2020**, doi: 10.1038/s41380-019-0583-1.