Alternative splicing: regulation, function and evolution

by

Carlos Martí Gómez-Aldaraví

Licenciado en Biotecnología

PhD Program in Molecular Biosciences
Departamento de Bioquímica, Facultad de Ciencias
Universidad Autonoma de Madrid (UAM)
2016-2020

Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC)

Thesis supervisors:

Enrique Lara Pezzi
Fátima Sánchez Cabo

Abstract

Introns populate eukaryotic genes to a variable extent across species, being widespread in vertebrates and mammals. While the evolutionary advantages, if any, of introns, remain unclear, their expansion has provided the opportunity to splice genes in more than a single way, allowing the production of different mRNAs from a single gene through Alternative splicing (AS). AS patterns change during the development of complex organisms and diverge across different tissues and experimental conditions. These highly reproducible changes evidences the existence of a regulatory network that ensures repeatable responses to certain stimuli and suggest that, at least some of them, play a role in the overall physiological response or adaptation. Not surprisingly, perturbation of some elements of this network is often associated with pathological conditions. However, not only we are far from a complete characterization of the molecular mechanisms that drive AS changes in most pathologies like those affecting the heart, but the computational tools that are currently used to study these regulatory networks are limiting our ability to extract all the information that is hidden in the data.

It has been long hypothesized that AS contributes to a great expansion of the proteome and facilitates the evolution of new functions from pre-existing ones without gene duplication. While there are very well known examples of how AS enables the production of different functional proteins or mRNAs, the proportion of AS isoforms that are actually functional remains large unknown. Indeed, recent studies from different perspectives, including both transcriptomic, proteomics and sequence evolutionary analysis suggest that this percentage may be rather small and that much of the observed transcriptomic diversity is driven by non-functional noise in the splicing process.

In this thesis, we have studied global AS patterns through computational analysis of large RNA-seq datasets to characterize the causes and consequences of AS changes from different perspectives. First, we have analyzed how AS global patterns change during heart development and disease using data from a variety of mouse models. We found that AS changes modulate different biological processes than gene expression ones and are associated to isoform specific protein-protein interactions. Disease patterns partially recapitulate developmental patterns probably through the upregulation of PTBP1, which is sufficient to induce pathological changes in the heart. Second, in an attempt to improve computational tools for identification of regulatory elements, we have developed dSreg. This tool leverages the power of bayesian inference and hierarchical models to pool information across the whole transcriptome to infer, not only the changes in the activities of the underlying regulatory elements, but also the changes in inclusion rates, outperforming competing methods and tools made for both purposes separately. Finally, we have studied the evolutionary process driving AS divergence during mammalian evolution using models of phenotypic evolution in a phylogenetic framework. We found that AS patterns have evolved under weak stabilizing selection that allows widespread variability in AS patterns across species, with only about 5% of the genes probably encoding AS isoforms with different functions. Rates of neutral evolution are high, preventing the identification of adaptive changes at this long evolutionary scale. In summary, this thesis provides new computational tools and knowledge about the evolution and regulation of AS in different biological conditions and helps to better understand its relevance from different persepectives.

Resumen

Una gran parte de los genes eucariotas están compuestos por exones e intrones. Aunque las ventajas evolutivas de la fragmentación de los genes, en caso de haberlas, no se conocen del todo bien, su expansión ha permitido procesar los transcritos primarios de varias diferentes maneras, y por tanto, de producir diferentes ARN mensajeros maduros. Potencialmente, esto permite la producción de proteínas con diferentes funciones mediante splicing alternativo (SA) a partir de un único gen. Los patrones de SA cambian de forma dinámica durante el desarrollo de los organismos complejos y se diferencian en distintos tejidos y condiciones experimentales. Esto sugiere la existencia de un programa de regulación específico para la reproducción de estos cambios en diferentes individuos y en respuesta a estímulos concretos, que difícilmente se habrá generado en ausencia de función y selección. De hecho, la perturbación de algunos de los elementos que forman parte de esta red de regulación se asocia frecuentemente con el desarrollo de diversas patologías. No obstante, estamos lejos de una completa caracterización de los mecanismos de regulación que dirigen los cambios en los patrones de SA en la mayuría de las patologías, como las cardiacas. Además, las herramientas computacionales disponibles para estudiar estas redes de regulación presentan una serie de limitaciones que reducen nuestra capacidad para extraer información completamente fiable de los datos de transcriptómica.

Desde el descubrimiento de los intrones, se ha hipotetizado que el SA permite la expansión del proteoma y facilita la evolución de nuevas funciones moleculares a partir de algunas ya existentes sin que haya duplicación génica. Aunque se ha caracterizado una importante cantidad de genes que producen isoformas con diferentes funciones mediante SA, aún se desconoce cómo de general es este mecanismo a nivel genómico. De hecho, estudios recientes desde diferentes persepectivas, incluyendo tanto transcriptómica como proteómica y análisis evolutivo de las secuencias implicadas, parecen indicar que el porcentaje de genes que generan diferentes isoformas funcionales mediante SA es más bien pequeño, y que por tanto gran parte de la diversidad transcripcional se debe a errores en el proceso de splicing.

En esta tesis doctoral, se han estudiado los patrones globales de SA mediante análisis computacional de grandes conjuntos de datos de RNA-seq. Esto ha permitido la caracterización tanto de las causas como de las consecuencias del SA y sus cambios desde diferentes perspectivas. En primer lugar, se ha analizado cómo estos patrones cambian durante el desarrollo y la enfermedad cardiaca, empleando datos de ratón como modelo animal. Se ha visto que los cambios en los patrones de SA afectan a diferentes funciones biológicas comparados con la modulación de la expresión génica. Estos cambios cuantitatives suelen afectar a isoformas de SA con distintos patrones de interacción proteína-proteína. Los patrones de SA en enfermedad recapitulan parcialmente los observados durante el desarrollo, posiblemente mediante la re-expresión de la proteína reguladora PTBP1, cuya sobre-expresión en corazones de ratones sanos es suficiente para inducir cambios patológicos. En segundo lugar, en un intento de mejorar las herramientas computacionales disponibles para la identificación de las proteínas reguladoras actuando en diferentes contextos, se ha desarrollado dSreg. Esta herramienta aprovecha el poder la inferencia bayesiana y los modelos jerárquicos para aunar la información contenida a lo largo del transcriptoma e inferir, no sólo los cambios cuantitativos en los niveles de inclusión de eventos concretos, sino también los cambios en la

actividad de las proteínas reguladoras subyacentes a esos cambios de SA. dSreg funciona mejor en ambos aspectos que los métodos previamente empleados para cada aplicación por separado. Finalmente, se han estudiado las fuerzas genéticas que dirigen la evolución del SA como carácter cuantitativo empleando, por primera vez, modelos de evoluci ón fenotípica a lo largo de la historia evolutiva de diferentes especies de mamíferos. Estos resultados indican que los patrones de SA evolucionan bajo una selección estabilizadora débil, que permite una gran variación en los patrones de SA entre diferentes especies. De este modo, apenas un 5% de los genes codifican isoformas de SA con diferentes funciones. La tasa de evolución neutral es tan alta que impide distinguir cambios aleatorios mediados por la mutación y la deriva genética de aquellos que suponen una adaptación, al menos en esta escala evolutiva tan grande. En resumen, esta tesis proporciona nuevas herramientas computacionales y conocimiento sobre la evolución y la regulación del SA en diferentes condiciones biológicas, y ayuda a entender mejor su relevancia desde diferentes perspectivas.

Acknowledgements

En primer lugar, me gustaría agradecer a mis directores Enrique y Fátima por haberme dado la oportunidad de hacer la tesis bajo su supervisión, por haberme dado libertad y apoyo para desarrollar mis propias ideas e intereses científicos más allá de los planteados inicialmente, y por haberme acompañado a lo largo de este periodo de crecimiento personal y, posiblemente, hasta científico. Gracias por vuestro optimismo y buen humor ante mi a veces reinante negatividad. Me gustará dar las gracias también a Claus, mi supervisor en Viena, por el tiempo que ha dedicado a ayudarme y enseñarme sobre biología evolutiva, no sólo durante mi corta estancia allí, sino también posteriormente, siempre con su buen sentido del humor. A mi tío Enrique por el diseño y elaboración de la portada.

A mis compañeros de laboratorio, tanto pasados como presentes, con los que he compartido diferentes etapas de tesis, además de preocupaciones, alegrías y penas: A Alberto, Girolamo y Enda, por acompañarme duranto mis inicios en la bioinformática del splicing y la práctica del inglés. A María y Eli, por compartir sus ideas conmigo y aguantar mis chapas sobre estadística, incluso cediendo a algunas de mis demandas en ese aspecto. A Javi y Miriam, por esas conversaciones de desahogo que se alargaban en los pasillos del CNIC. A Marta, por su alegría y charlas sobre intereses frikis comunes. A las Lauras, Paula, Fernando y Marina, por su paciencia ante mis largas y poco intelegibles presentaciones y mis frecuentes interrupciones y preguntas en los lab meetings.

A los miembros de la unidad de bioinformática y otros refugiados que tan inocentemente me acogieron cuando parecía que no iba a molestar mucho, y acabaron sufriendo largas discusiones sobre estadística bayesiana, politica y otros temas de actualidad. A Jorge, por animarme cuando el ánimo estaba bajo cediendo a las presiones de la ensaladilla y llenando el cluster de core-dumps, y, en general, por el bullying recíproco intercambiado durante estos años. A Fernando Martínez, por el constante apoyo en los misteriosos problemas de la informática y la torpeza de una mente ingenua, así como por sus magníficas recomendaciones gastronómicas. A mi infatigable rival Manuel, por sus mortíferos saques de padel, a.k.a. manoletinas, y sus enérgicas estocadas en esgrima, que han mantenido mis reflejos en forma más allá de la pantalla del ordenador. A Carlos, por las discusiones científicas y por dejar siempre lo que estuviera haciendo para intentar resolver cualquier duda que tuviera, aunque no supiera la respuesta en el momento. A Jazberna, por inspirarnos a todos con sus carreras alrededor de la rotonda y mantenernos al día de lo que se comenta en las mañanas de Federico. A Fernando Benito, por mantenernos actualizados en el mundo de las construcciones y las tendencias ecológicas y una participación activa en los debates sobre el estado de la nación. A Juan Carlos, por su incesante buen humor y excelentes dotes de anfitrión. A Jose Luis, por su inestimable ayuda en la diaria lucha contra el Ciudadano de la Barrera. Finalmente, a Felipe, por su buen hacer y su risa contagiosa.

A Ana y Lola, componentes del hogar de los soñadores, donde habité muy gustosamente durante la mayor parte de la realización de esta tesis. Gracias por ayudarme a manterme un poco más atado al mundo real, a animarme a la diversificación de mis actividades y por formar parte de mi exclusivo club de fans como ukelelista a partes casi iguales, que no es cosa menor. Gracias por llevar petos, chalecos, sombreros y tobillos al aire para mantenerme al día en las últimas tendencias de la moda. A Mr Foodie,

por sus críticas gastronómicas y por traer siempre doritos con guacamole cuando hacía falta.

A Jose, Rebeca, Isaac e Irene por lo todos los buenos momentos que hemos compartido durante estos años. Desde las largas meriendas en el CNIC a la luz de las chocobons, las cenas gourmet de macarrones con philadelphia, las papillas de desayuno y hasta por venir a pedir dinero para un café. Gracias por todos los viajes y las poco realistas imitaciones mutuas que hemos compartido y disfrutado, por aguantar mi mal humor y acertadas y desacertadas críticas, y por apoyarme en los momentos difíciles.

A los integrantes del ya conocido como grupo de Cactus, por los planes improvisados, los miércoles de cine, viajes y escapadas ocasionales para abstraerse de la tesis y otras preocupaciones. A Marta, por su capacidad de no organizar pero sacar adelante grandes planes, por sus sopas de ajos y patatas a la lumbre, así como la asistencia en la gestión de dramitas. A Alba, por animarme a salir y a pasarlo bien más de lo que habría hecho por mi cuenta, con buenos resultados las más veces. A Juan Luis, por su inquebrantable optimismo y radiante felicidad. A Víctor Fanjul, por su carácter integrador y hablar sin prisa pero sin pausa. A Austin, por su motivación, entusiasmo y ganas de hacer cosas. A Mercedes, por su maravillosa inocencia y honestidad.

A Pedro y Santos, por camuflar con planes culturales nuestras comilonas y divagaciones varias, por preguntarme qué tal va la tesis puntualmente, por dejarme vislumbrar como discuten los verdaderos científicos sobre los fundamentos del universo en el que habitamos y los misterios de la tierra sigilata, y las risas que nos echamos a nuestra propia costa.

A los miembros del viejo y nuevo Ateneo, por proporcionarme un hogar a la altura del anterior durante los últimos años de la tesis, por las largas discusiones y reducción al absurdo de cualquier cosa y contagiarme vuestra disfrutonería. A Víctor, por las feroces críticas que ponen a prueba cualquier dialéctica mínimamente frágil y llevan a uno a reconsiderar sus principios más básicos; y por el incesante flujo de nuevas ideas y temas sobre los que conversar. A Álvaro, por reirse de hasta mis peores chistes y referencias más viejunas, por compartir su sabiduría gastronómica y iluminarnos con una cultura popular desbordante. A Junior, por sus teorías, justa forma de hacer negocios en los juegos y su visión tan enriquecedora de la vida y las relaciones personales.

A Alejandra, por tu perseverancia. Por tu creciente flexibilidad y haber compartido alegrías y preocupaciones en estos tiempos inciertos, por llevarme al monte y cocinar mejor que yo mis propias recetas.

Por último, me gustaría agradecer a mis padres y hermanos por sobrellevar tanto mi ausencia como mis ocasionales presencias, por estar y seguir ahí ante viento y marea.

Contents

Acronyms

Ψ Percent Spliced In. 24, 28, 29, 36, 37, 45, 46, 49, 52, 53, 64, 65, 73, 74, 90

$\hat{R}$ Gelman-Rubin R. 43, 63

AA alternative acceptor. 18, 46

AAV9 Adeno-Associated virus type 9. 75, 77–79

AD alternative donor. 18, 46

AS Alternative splicing. 2, 7, 18–20, 22–25, 28–41, 44–46, 59–61, 64–66, 68–76, 78–81, 84, 87, 96, 99, 102–105, 108–122

AUROC area under the Receiver Operating Characteristic curve. 28, 38, 44, 74, 76, 81, 85, 86, 100, 101

BH Benjamini-Hochberg. 36–38, 43, 44

BM Brownian motion. 56–58, 87, 89–95, 116–118

BP branch point. 13, 25

CI 95% Posterior credible interval. 83, 84, 93

CLiP-seq cross-linking and immunoprecipitation followed by sequencing. 72, 78, 84, 87, 103, 109, 110

CV Cross-Validation. 28

DAG Directed Acyclic Graph. 41

DCM dilated cardiomyopathy. 33, 34

DDI domain-domain interaction. 37, 68

DEG Differentially expressed genes. 36, 68

DM myotonic distrophy. 34, 112

DMS deep mutational screening. 26

dN/dS ratio of non-synonymous to synonymous substitutions. 23

ED embryonic development. 19, 38, 64, 66, 68, 70, 73–76

ES exon skipping. 18–20, 24, 29, 46, 52, 89

ESS effective sample size. 63

FDR False discovery rate. 37, 81

GE Gene expression. 36, 46, 68–71, 75, 78, 79, 114, 116, 122

GLM Generalized Linear Model. 37, 38, 43, 44, 70, 80, 81, 83

GLMM Generalized Linear Mixed Model. 36, 64

GO Gene Ontology. 38, 70–72, 78, 96

GSEA Gene Set Enrichment Analysis. 30, 44, 45, 81, 83, 84, 86–88, 108, 109, 120

HF heart failure. 33

HMC Hamiltonian Monte Carlo. 63

ID intron definition. 18

IR intron retention. 18, 19, 21, 26, 29, 32

KD knock-down. 73, 74

KO knock-out. 73

LAD left anterior descending. 64, 73

LECA last eukaryotic common ancestor. 15, 16, 18, 19

LM Linear Model. 37

LOF loss of function. 74

LUCA last universal common ancestor. 13, 14

MCL Markov clustering. 54, 55

MCMC Markov Chain Monte Carlo. 41, 43, 62, 63, 80, 92

MI myocardial infarction. 34, 36, 38, 64–66, 68, 70, 73–76, 78, 105, 114

miRNA micro RNA. 21, 27

MLE Maximum Likelihood Estimation. 52, 83

mRNA messenger RNA. 13–15, 18–23

MSA Multiple sequence alignment. 54, 55

MSE mean squared error. 85

MXE mutually exclusive exons. 18, 20, 32

Ne effective population size. 16, 18, 22

NMD non-sense mediated decay. 15, 16, 18, 21–24, 31–33, 66, 115, 118

NUTS No-U Turn Sampler. 43, 63

ORA Over-representation Analysis. 30, 44, 45, 81, 83, 84, 86–88, 108, 109, 120

ORF open reading frame. 66

OU Ornstein-Uhlenbeck. 57–59, 61, 62, 89, 92–101, 114, 116–118, 120

PCA Principal Component Analysis. 37, 65, 66, 75, 78

PD post-natal development. 38, 64, 66, 68, 70, 73–76

PE paired-end. 49

PPI protein-protein interaction. 37, 67–69, 114

PPT poly-pyrimidine trait. 25, 31

PWM Position Weight Matrix. 110

RBP RNA binding protein. 21, 26–28, 30, 32, 34, 38–45, 70, 72, 74–76, 80, 81, 83, 84, 86, 87, 103, 105, 108–113, 115, 120, 121

RFP reading frame preservation. 67

ROC Receiver Operating Characteristic. 81, 82, 100, 101

RPKM reads per kilobasepair and million. 36

rRNA ribosomic RNA. 13, 14

RUST regulated unproductive splicing and translation. 21, 22, 66, 105, 115, 118

SJ Splice Junctions. 45–53, 102

SMA spinar muscular atrophy. 28

snRNA small nucleolar RNA. 14, 18, 25

SVM Support Vector Machine. 28

TAC trans-aortic constriction. 34, 36, 38, 64–66, 68, 70, 73–76, 78, 103–105, 114

TF transcription factor. 27

tRNA transference RNA. 13

UTR untranslated region. 21, 27, 34

WT wild-type. 75

1. Introduction

1.1 Splicing: from prokaryotes to higher eukaryotes

1.1.1 Discovery of introns and splicing

Until the late 70's, protein coding genes were thought to be contiguous fragments of DNA, also called cistrons, that underwent transcription and translation to produce a single polypeptide. Thus, a gene could encode only a single protein sequence. This idea was based almost entirely on data from prokaryotic species, mainly *Escherichia coli*, and there was no reason to think eukaryotic genes would behave differently [65]. This notion, however, was first challenged by the finding that primary nuclear RNA transcripts appeared to be longer than messenger RNA (mRNA)s [61]; and second, by studies aiming to map protein coding fragments to the genome of the lytic adenovirus type 2. They used electron microscopy to characterize DNA-RNA hybrids and that pairing in one of the moleculres was interrupted by extra unpaired sequence, absent in the complementary chain. This suggested that some fragments of viral RNA were removed and the remaining ones were joined together to form the final mRNA [55, 29]. Studies on other eukaryotic species suggested that not only coding sequences tended to be interrupted in the genome by silent DNA, but also non-coding genes such as the transference RNA (tRNA) and ribosomic RNA (rRNA)s. These silent DNA fragments were then called introns (intragenic regions) and the flanking sequences that were kept in the mature mRNA were named exons (expressed) [90]. Years later, it became widely accepted that coding or expressed regions are often interrupted in the genome by introns across all eukaryotic lineages. At the same time, molecular biologists unveiled how these organisms can remove introns before protein synthesis or final RNA maturation: first, both exons and introns are transcribed into a primary o precursor mRNA; second, intron sequences are recognized and removed in a complex process called splicing to finally produce functional mature mRNA [65]. Most eukaryotic introns are spliced out by a large molecular complex known as spliceosome. The human spliceosome is a large and highly conserved ribo-nucleoprotein complex including 5 different small RNAs and over 200 different proteins. This complex is sequentially assembled around 3 main signals in the pre-mRNA (splice donor and acceptor sites and branch point (BP)), and catalyzes the intron removal process through 2 trans-esterification reactions [113].

1.1.2 Origin and evolution of introns

Some of the main questions raised by the discovery of introns regarded their origin and evolution, now thought to be highly linked with the origin of eukaryotic cells [140]. Were introns present already in the last universal common ancestor (LUCA)? To what extent? Have they been subsequently lost in prokaryotes and *archaea*, or inserted and expanded in eukaryotes? During the early 80's, the prevailing notion was that eukaryotes emerged from endosymbiosis within prokaryotes. This, together with the absence of spliceosomal introns across prokaryotic genes, suggested that they were introduced during or

after evolution of eukaryotic cells (*Introns-late view*) [65, 140].

The introduction and expansion of introns into coding sequences in eukaryotic cells may have increased the chances of recombination between different pre-existing functional protein elements to build more complex protein functions [90]. Thus, the presence of introns would provide a way to greatly accelerate protein evolution by allowing combination of pre-exisiting functions rather than by generating them *de novo* by mutation of non-functional sequences. This idea, known as the *domain shuffling hypothesis*, predicted that exons would encode for different isolated functional domains [33]. Such association, and even evidence of exon shuffling, was found across a reduced number of proteins [65], including the LDL receptor and EGF precursor [89], but the relative contribution of this mechanism to protein evolution at a global scale has remained unclear for some time [263]. Latest studies using 88 complete eukaryotic genomes already available found alignment of domains and exon boundaries more often than expected by chance, particularly in chordates. However, not only this association was weak, but most domains remain encoded by several exons, suggesting that the benefit of fragmentation may come from a further division beyond functional domains into smaller functional elements [256].

The exon theory of genes

Despite the initial support for the *introls-late* hypothesis, as divergence of eukaryotic and prokaryotic lineages became clear to be older than previously thought, the presence of introns in the LUCA and their subsequent lost in procaryotes became a plausible hypothesis. This *introns-early* hypothesis was supported by the idea that the introns-derived inefficiency could be a sign of archaic genome organization that was negatively selected in prokaryotes and, to a minor extent, in early branching eukaryotes such as *S. cerevisiae* (average of 0.05 introns per gene) [239].

The discovery of the replicating and autocatalyic introns in prokaryotic, *archaeal* and eukaryotic cytoplasmic genomes, such as that from ribosomic RNA (rRNA) 28S of *Tetrahymena thermophila*, suggested that introns may pre-date, not only the origin of eukaryotes, but even perhaps the origin of cellular life [140]. This type of introns, later known as group II introns [113], form very similar structures to those of the hybrids between snRNA and splice sites [140], suggesting a potential common origin. Type II introns inspired the theory of a primitive RNA world, in which the replicating ability of catalytic RNAs would be under natural selection. With the appearance of the translation machinery, the exonic regions of these RNAs could encode for proteins that might help the replicating activity and survival of the RNA, leading to the differentiation between genomic and messenger RNAs. Intronic sequences would then be present in the RNA ancestor and incorporated also to the first DNA genomes. Introns would be maintained in the DNA genome, as this molecule lacks self-excision ability, and only removed when transcribed to produce a functional mRNA. Under this theory, introns may just be relics of a pre-cellular genome assembly process but they "were not introduced in response to any selection pressure for long-term evolutionary flexibility experienced by individual organisms within population" [65]. Thus, introns could be considered exaptations, this is, characters evolved from selection for some function different from their current use, rather than adaptations for cellular species.

This primitive RNA world was the seed for the *Exon theory of genes*, which proposes that first genes were assembled by recombination of sequences randomly encoding for short peptides ($\sim$20 aminoacids) that had evolved some minor structure and function. Hence, genes would naturally have arisen with exons and introns. Introns could be occasionally removed from the gene by recombination with processed mRNA, whereas more complex exons might arise by retro-transposition. These new longer exons may encode for larger functional structures, e.g. protein domains, and serve again as substrate for recombination to form new and more complex proteins [89].

If exon shuffling was key to generate the basic protein repertoire of living cells but no longer required,

as exon-intron structure is reorganized and modified through other mechanisms, the association between exons and functional domains will become weaker with time. Thus, one would expect to find such association mainly in highly conserved proteins with relatively unchanged exon-intron structure. On the other hand, if exon shuffling has remained an active mechanism for protein evolution, new genes assembled by exon shuffling will show more clear evidence of exon-domain association. Indeed, recent studies support the ongoing *domain shuffling* idea, as the domain-exon association was stronger in more recently re-arranged genes [256].

However, there are other processes that can explain, at least partially, the association of exon and domain boundaries. If protein sequences are more constrained in the functional domains than in the regions separating them, even with completely random insertion of introns, a higher proportion of these insertions will be fixed in these inter-domain regions of the proteins, since potential changes in the protein sequence derived from intron insertion may be less deleterious. Additionally, intron insertion may be not completely random, but favored at some nucleotides contexts, specially if they carry specific machinery of insertion. These regions with higher intron insertion probabilities may encode aminoacids with higher chances of forming inter-domain regions, and thus provide an alternative explanation for the exon-domain association [89, 65]. Thus, as our understanding about the molecular mechanisms for intron insertion and their fitness effects increases, we may be able to tease appart the relative contribution of each process to the exon-domain association.

Intron origin and expansion during eukaryogenesis

The study of the origin of introns is naturally linked to how they have evolved to the present patterns. If introns were not present in the ancestral genes, then, they have been continuously acquired and spread during eukaryotes diversification. On the other hand, if introns were ancestral to all live beings, they may have been present only in a few genes as type II introns, and expanded in eukaryotes; or, alternatively, introns were very abundant and were subsequently lost, not only in prokaryotes and *archaea*, but also in many eukaryotic species, which show a wide variation in intron content [140]. Thus, the study of intron evolution, not only is interesting *per se*, but also adds information about their origin. There have been several attempts to fit evolutionary models to intron data, showing initially controversial results: some models supported an scenario with massive intron loss, whereas others supported massive intron gain, with the subsequent consequences in the estimated abundance of introns in the common ancestor. There are, however, more known mechanisms of intron loss than for intron gain, suggesting that they may actually be more common, at least a the mutational level [140].

These ideas about intron evolution challenged the *Exon theory of genes*, and links the origin of introns with some of the key innovations of the eukaryotic cell. This modified *introns-late* theory proposes that pre-existing type II introns in the α-proteobacteria endosymbiont invaded the *archaeal* host's genome. The host cell was highly unprotected from these retroelements, leading to an intron-rich last eukaryotic common ancestor (LECA). Then, the self-splicing machinery was fragmented into different elements and started to work in trans over the intron sequences. Introns lost the ability to replicate and self-splice and kept only the cis-regulatory elements for being recognized by these new trans elements. This idea is supported by the great biochemical and structural similarity between type II introns and spliceosomal introns, and by the relative abundance of these retroelements in the α-proteobacteria compared with *archaea* (closer to the protoeukaryotic host), in which they are nearly absent [113]. Thus, to prevent the translation of the transcripts massively invaded by intronic sequences, there would be a great pressure for compartmentalization of the splicing reaction within the nucleous or/and a more effective trans mechanism for their catalysis. These mechanisms may not have been enough, and further mechanisms for controlling mRNA and protein fidelity, like the non-sense mediated decay (NMD) and ubiquitin mediated proteolysis,

could have been selected as a consequence. Whereas NMD possibly derived from a bacterial toxin-antitoxin system containing nuclease domains similar to those employed in NMD; ubiquitin signaling might derive from the biosynthesis pathways of molibdopterin and thyamin. The great movement and expansion of these retroelements may have also driven the linearization and fragmentation of the DNA, and even provided the solution for telomere shortening driven by replication: the catalytic subunit of the telomerase seems to have evolved from the retrotranscriptase encoded by the invader group II introns [140].

Type II introns are estimated to have invaded up to 70% of the host genome from the relatively low abundance in α-proteobacteria (< 30 copies per genome) [139]. This expansion of retroelements is much larger than any registered in prokaryotes, by orders of magnitude [113], and arguably possible in an asexually reproducing host, raising doubts about the order of events in eukaryogenesis and about the expansion of introns as main driver of every major eukaryotic innovation. This, together with the absence of *archaea* carrying endosymbionts at the present, suggests that some of the eukaryotic features, such as the nucleus, meiosis and linear chromosomes, may have pre-existed in the host cell, and have allowed the expansion of introns rather than been driven by it [220].

An additional requirement for the massive expansion of type II introns, which at the time were likely to be deleterious or, at most, neutral, is a small effective population size (Ne). One of the consequences of the development of eukaryotic cells is the increase their cell size, and the subsequent slow-down on the metabolism due to the limited exchange rates of nutrients and compounds with the external compartment. As a consequence, not only higher eukaryotes, but also unicellular yeasts, show a reduction in Ne of several orders of magnitude compared to prokaryotes and *archaea* [176], from $N_e \sim 10^9$ to about $N_e \sim 10^6$ in LECA, which could have allowed the expansion of introns after the endosymbiosis in the LECA [139].

Differential retention and expansion of introns during eukaryotic diversification

Phylogenetic reconstruction of ancestral states suggests that LECA was relatively intron rich, comprising up to 70% of the genomic DNA and possessed a complex spliceosome, even if there is a wide variability in intron content in present eukaryotic species [139, 113]. Great differences can be observed between unicellular and multicellular organisms (*S.cerevisiae* has a total of 253 introns in only 3% of its genes compared with human genes, with about ∼7.8 introns per gene), but also within unicellular organisms (43 % of *S.pombe*'s genes have introns) [14]. Thus, current hypotheses point towards a continuous net intron loss since the intron-rich LECA, with punctual events of high intron gain. Interestingly, both rates seem to have decelerated over time [139, 231, 113]. There have been, at least, 3 mayor episodes of massive intron gain during the evolution of eukaryotes: at the root of the Metazoa, at the shared last common ancestor between Metazoa and Choanoflagellata, and at the root of Ichthyosporea [231, 98] (See Figure 1.1)

The great variability in the rates of intron loss and gain during the evolution of eukaryotes has puzzled scientist for a long time, who have come up with new molecular models able to explain such complex mutations, including intronization of exonic sequences, intron transposition, insertion mediated by transposon, insertion of group II introns, reverse splicing and retrotranscription and template switching [116, 237, 238, 307, 306]. Among them, intronization of exonic sequences has been of particular interest, as it provides a completely intrinsic and simple mechanism: single point mutations can easily create new splice sites with low affinity, which can be maintained in the population as it has a very low fitness cost, possibly invisible to an eukaryotic organism with low Ne. At the same time, a mutation creating a new stop codon or frameshift in the coding sequence may take place. If the alternative splice site allows skipping this new modification, it will be selected to avoid the potential deleterious effect of the resulting truncated protein, by just removing a small exonic sequence. This model predicts some of

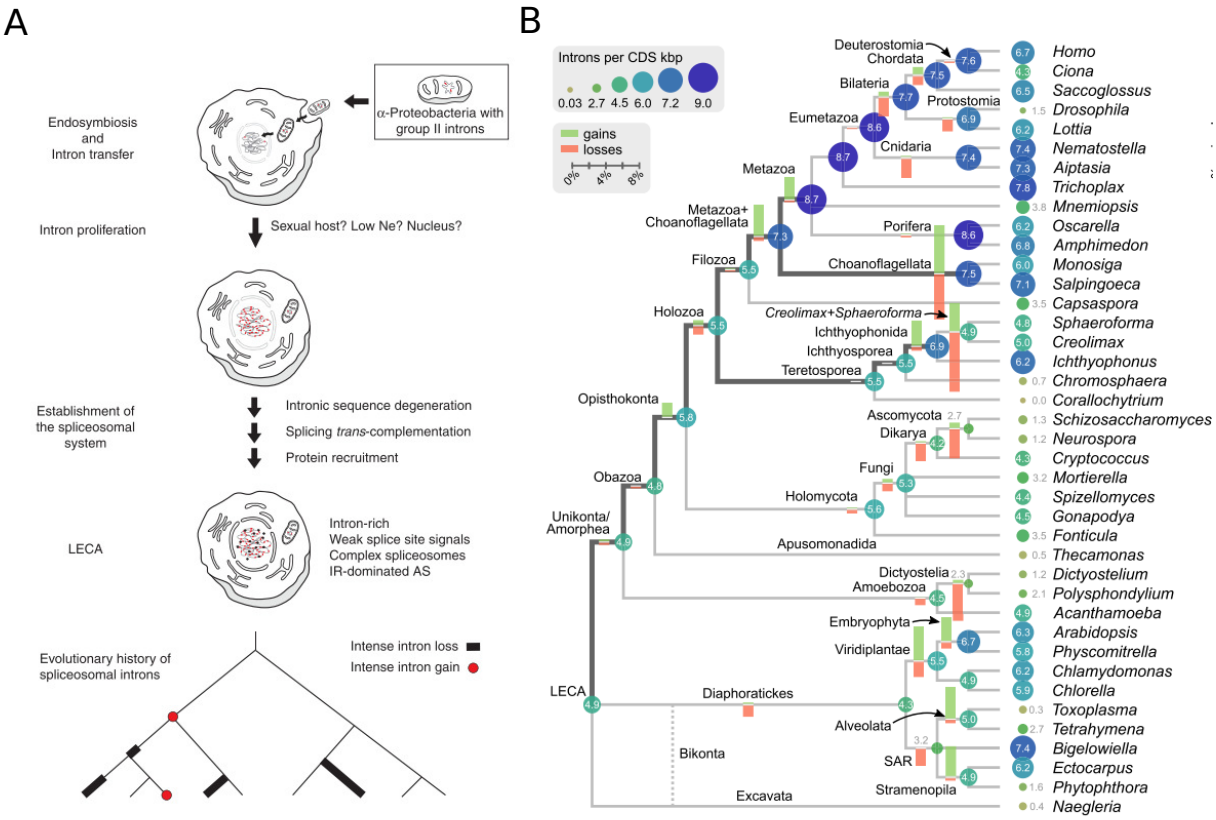

Figure 1.1: Intron evolution. **A** Hypothesis of invasion of type II introns accompanying the endosymbiosis as a key step in eukaryogenesis (from [113]). **B** Reconstructed evolutionary history of intron content along eukaryotic diversification [98]

the observed biases in the gene structure such as the accumulation of new and shorter introns in the 5' end of mRNAs, where protein truncation may have stronger fitness consequences and thus intronization of short sequences may be more strongly selected [48, 49]. Cryptic splice sites play an important role in the position in which introns appear in a gene, not only because they may allow production of a minor alternative isoform after the appearance of premature stop codons; but also may favor spliceosome dependent intron insertion [240]. However, these mechanisms fail to explain the origin of recently created introns. As more and more genomes become available, the power to identify patterns that may contribute to explain intron creation are expected to increase [307, 306, 316]. Recently, a great expansion of introns mediated by non-autonomous DNA transposons has been reported in *Micromonas pusilla* and *Aureococcus anophagefferens*. These transposable elements carry one splice site, whereas the other is co-opted after transposon insertion allowing perfect splicing of newly created introns. Episodic expansion of this type of transposable elements may explain the punctual expansions of introns during eukaryotes diversification and some of the observed biases [108].

Population bottlenecks may have played a major role, not only for the early expansion of introns in LECA, as previously proposed [140], but also for the selective maintenance and expansion of introns during vertebrates diversification, since intron content is highly associated with small Ne in extant species [176, 178, 180, 139, 231]. These events of large intron expansions, although potentially deleterious or neutral in the first place, may have contributed to increase the genetic variability available to selection to an extent unreachable by species with large Ne [176, 178, 180].

1.1.3 Origin and evolution of alternative splicing

Accompanying the first expansion of type II introns in the LECA, current hypotheses suggest that trans-complementation of the splicing reaction into a major complex spliceosome composed by 5 small nucleolar RNA (snRNA) and about 75 different proteins was required to relieve the selective pressure from individual introns to this large and highly conserved molecular complex and allowed an increase in the speed and fidelity of the splicing reaction [113]. Still, variation in the splicing process may occur, allowing different types of Alternative splicing (AS) events, including intron retention (IR), exon skipping (ES), alternative acceptor (AA) and alternative donor (AD) selection, mutually exclusive exons (MXE), and many more complex combinations [99, 202].

The type and frequency of each type of event is naturally constrained by the gene architecture in each species, e.g. very few genes have more than a single intron in *S.cerevisiae*, yielding ES events virtually impossible. The simplest and most prevalent event across eukaryotes is IR, particularly in plants [97, 280, 226, 50], as it may happen with as few as 1 intron per gene. Despite the few hundred introns populating *S.cerevisiae* genes, these are enriched in ribosomic proteins. Their retention increases upon aminoacid depletion and, coupled with the NMD, mediates the downregulation of the translation machinery to compensate the lack of nutrients in the media as a way to modulate the translation machinery [239]. Indeed, intron removal show little to no fitness effects in normal conditions, but introns provided resistance to starvation by repressing ribosomal genes through TORC1 signaling [212]. These introns may even have a role *per se* after being spliced out from the transcript, as their removal regulates cell growth in yeast through TORC1 signaling [197]. *S.cerevisiae* has not only few, but short introns. In contrast, *S.pombe*, diverged from *S.cerevisiae* 370 million years ago, has many more introns, with splicing signals more similar to those of *metazoa*, but they remain as short as in *S.cerevisiae*. Splicing in this species takes place through intron definition (ID), this is, by recognizing the signals defining an intron and splicing it out. As *S.pombe* only shows IR events, not only intron number, but also intron length is thought to be a key determinant of the splicing mechanism and of the type and prevalence of AS events that can be produced [14].

More complex eukaryotes, especially *bilateria*, and vertebrates within them, have more abundant and longer introns. Recognition of splicing signals in such a long sequence may be more difficult, leading these species to use a different splicing mechanism focused on the recognition of exons, known as embryonic development (ED). This mechanism is associated with other types of AS events beyond IR; i.e. ES, alternative splice donor and acceptor sites; now derived from variation in the recognition of whole exons or exon boundaries by the splicing machinery [97]. These types of events are associated with weaker splice sites and more conserved intronic sequences, to which regulatory proteins potentially bind to modulate their recognition by the spliceosome [14]. Conservation of AS across highly diverged eukaryotic lineages, together with the high intron density and relatively weak splice sites expected in the LECA, suggest that alternative splicing may be ancestral to all eukaryotes, and that fast growing species, like *S.cerevisiae*, have lost alternative splicing together with introns and the optimization of splicing signals for greater mRNA processing efficiency [112, 115, 130, 113].

Regardless of whether AS was ancestral to all eukaryotes or not, alternative processing events have been continuously evolving, leading to the observed patters in extant species. Thus, comparative studies have allowed identification of mechanisms for the generation of alternative exons in a gene architecture characterized by long introns and short exons [130]. One of such mechanisms is the *alternativization* of constitutive exons. Accumulation of mutations weakening the splice sites (mainly the 5' splice site) lead to suboptimal recognition of the exons and partial ES. These mutations may or may not be maintained during evolution depending on the fitness consequences and the extent of partial ES. A different mechanism, favored by long introns, is the exonization of intronic sequences *de novo*. Mutation may create, by chance, over a long evolutionary times and sequence, the minimal splicing signals required for partial recognition and splicing by the spliceosome. Whereas the first mechanism creates a new isoform by removing part of the original protein, the former does so by adding a completely new coding sequence [130]. Exonization of intronic sequences was thought to be the main mechanism generating new alternative exons across mammals, since most new exons have an homologous sequence in the intronic species of closely related species [5]. These new alternative exons are mostly included at low rates [313], in a tissue regulated manner, and are associated to changes in tissue regulated expression, upstream intronic deletions and increased nucleosome occupancy [193]. Of them, an increase in nucleosome occupancy predates the exonization event, suggesting that transcription slowdown of the RNA pol II by nucleosomes increases the chances of *de novo* exon formation [163]. Interestingly, exonization events are also associated to repetitive sequences: up to 45% of new human exons are associated to known repeats, mostly of the *Alu* family [5]. *Alu* elements are primate-specific retroelements that have expanded to amount more than 10% of the human genome. This element carries multiple sequences that are very similar to splice sites, facilitating the creation of splice sites by random mutation, even if that may happen millions of years after the insertion of the transposable element in the genome. These newly created exons are also usually spliced at low rates [259]. Similar associations have been found in other lineages: 65% of new exons in rodents also derive from repetitive elements [259]. Thus, transposable elements may have played a key role in shaping intron insertions during eukaryotes evolution [108], but also for exonization of intronic sequences into alternatively spliced exons [259, 5, 130].

1.1.4 Gene duplication and alternative splicing: interconnected mechanisms generating new functions

Gene duplication has been known for a long time to be an important mechanism in evolution, as it provides an opportunity to build new functions from an already existing one without losing it or having to start from scratch. Thus, it is not surprising that gene duplicates are found across all domains of life, and that a large percentage of genes in any genome are originated by gene duplication: estimates

range from 17 to 65% depending on the species, suggesting rates of gene duplication similar to those of nucleotide substitution. Gene duplicates, at the time of duplication, evolve neutrally and pseudogenize most of the times; a few duplicates, however, can either maintain the original function thanks to an increased gene dosage, evolve a new function, or subfunctionalize i.e. differential retention of ancestral gene functions into the separate gene duplicates [205, 310, 177, 179].

Soon after the discovery of introns, it became clear that alternative splicing could be a complementary mechanism, similar to gene duplication, for the generation of new protein functions. Gilbert already proposed that the chances of retention of gene duplicates may be higher when the gene had two already pre-existing functions e.g. generated by alternative splicing, to subfunctionalize into the different gene copies [90]. Comparative studies, mainly performed on mouse and human genomes, showed an association between the two mechanisms: the size of a gene family is inversely correlated with the frequency of alternative splicing, specially in recent duplication events. At the same time, there is a positive correlation between duplicates age and alternative splicing, suggesting that AS is slowly gained after gene duplication [141, 53, 145]. Very recent duplicates, however, showed higher levels of AS, suggesting an early expansion of splice isoforms after gene duplication, possibly due to relaxed selection, and thus increased ground for subfunctionalization of splice isoforms into the different gene copies [120]. When restricted to genes with highly alternative and ancestral MXE, some of them show clear subfunctionalization patterns, whereas other seem to maintain both exons within the paralogs, suggesting that evolution of alternative gene duplicates may be affected by gene properties or function [1]. A potential way of subfunctionalization is through partitioning of expression patterns across tissues. In this sense, alternative duplicates tend to show more specialized tissue-regulated gene expression patterns, as do gene duplicates with divergent exon structure [266, 145]. Similar trends have been described in *C.elegans*, but not in plants, where the opposite is actually found. This pattern suggests potential differences in the impact of AS on the retention of duplicates depending on the type of AS, as ES is prevalent in animals but not in plants [111].

1.1.5 Function and fitness consequences of alternative splicing

The finding of introns and their distribution across the genomes of species populating the earth have puzzled scientists since their discovery. One of the fundamental questions in the field, which is not quite clear yet, is: what is the function, if any, of introns and splicing? In other words, what are the fitness consequences of alternative splicing and its contribution to species evolution? Ever since their discovery in the late 70's [55, 29], scientists have proposed a plethora of reasons and theories of all the possible benefits that may derive from dividing genes in pieces along the genome [90]. The first hypothesis, as previously discussed, was the *domain shuffling* hypothesis and its developed version as the *Exon theory of genes* to explain the early origin of life and the assembly of the first proteins from smaller functional elements [90, 89]. This compartmentalization of protein information into different units called exons may not only facilitate the recombination between different variants within the same gene and re-organization of protein domains or functions into new proteins, but also accelerate protein sequence evolution: mutations at splice sites can easily result in the partial or complete exclusion of exon sequences from the mature mRNA and therefore in the removal of a stretch of aminoacids in a mutant protein. This may increase the potential fitness effect or relax selection in the partially skipped protein region [65]. A very special type of alternative splice acceptor involves the selection of splice sites that are separated by only 3 nucleotides at the frequently observed NAGNAG sequences, which gave name to this type of event. They have been associated to rapid changes in exon size during evolution and to contribute to the biased codon composition at the beginning of exons. Thus, these highly variable AS events may contribute to accelerate protein evolution at the beginning of the exons. These events allow the introduction or removal of a single aminoacid into the protein, which is highly biased by the phase

e.g. phase 2 exons are strongly enriched in adding serine and arginine residues to the protein, potential phosphorylation targets [37].

Despite the early focus on the impact of introns in protein sequence evolution, Gilbert already noticed that variation in the splicing process may allow one gene to simultaneously encode different functions by selectively including in the final mRNA different functional elements [90]. Indeed, the first cases of alternative splicing were characterized very soon after the discovery of introns. For instance, alternative splicing of the μ chain of IgM produced two different mRNAs encoding functionally different proteins by selectively including a domain that determines the IgM to be secreted to the media or kept as a membrane protein [66, 7]. In the case of pyruvate kinase gene (PKM), alternative splicing enables the production of a second isoform PKM2, which can specifically translocate to the nucleus and interact with transcription factors to regulate changes in gene expression promoting the Warburg effect in cancer [304]. However, one of the classical and most famous examples can be found in the sex determination system in *Drosophila melanogaster*. In this species, sex is controlled by the switch gene *Sex lethal* (Sxl), which drives female development. It is inactivated in males by including an exon whose inclusion produces an alternative mRNA with an in frame stop codon that is subsequently degraded by NMD [28]. Sxl is itself an RNA binding protein (RBP) and regulates its own splicing to ensure that the frame shifting exon is skipped in females once Sxl is active. Although this exact mechanism is not widely conserved, different insects have independently evolved splicing-regulated sex determination systems, including *Apis mellifera* and *Musca domestica* [241].

Thus, alternative splicing enables one gene to encode different protein isoforms, but may work also as a post-transcriptional regulatory mechanism to modulate mRNA levels in coordination with the NMD pathway[151]. This process, named regulated unproductive splicing and translation (RUST), is used by many RBPs to modulate their own and other RBPs' and regulators, like CDC-like kinases mRNA levels e.g. SRSF1, SRSF2, PTBP1/2, MBNL1, TRA2B [36, 151, 262]. RUST is widespread and conserved among RBPs and has evolved independently for many of them, even within the same gene family. This high prevalence of an apparently inefficient regulatory mechanism may be explained by the benefits of self-regulation (protein levels act simultaneously as intrinsic sensor and regulator) and high evolutionary accessibility: there are many possible splicing variants and at many regions of a gene leading to aberrant mRNAs targeted by NMD [150]. Although RUST appears fundamental to form very interconnected alternative splicing regulatory networks, it is not limited to RBPs self-regulation. Quite the opposite, it appears to be relatively general: a good part of the products derived from alternative splicing events in humans are predicted to be targeted by NMD [151]. In particular, IR has been associated with decreased expression levels across tissues in mammals and appears to be a relatively conserved mechanism for fine-tuning gene expression [280]. This regulatory mechanism requires the alternative splicing event to be located in a protein coding region. However, a good part of alternative splicing events are found in the UTRs of the mRNAs. These events may however retain regulatory potential, by including sequences with subcelullar localization signals for mRNAs, or modulating stability (directly or through miRNA targets inclusion) or translational efficiency [151]. For instance, regulation by alternative splicing of binding sites in 3'UTR of SR proteins, which function as molecular adapters of the export machinery to mRNAs, may affect their transport rate to the cytoplasm and, indirectly, the abundance of the encoded protein [198].

1.1.6 Debating functional alternative splicing and splicing noise

Since the discovery of introns and splicing, other scientist have been arguing against them playing a fundamental role in evolution, at least, in general terms: "Thus, in terms of beneficial effects on the fitness of organisms, we almost certainly cannot account for the presence of the majority of individual introns, nor for the propensity to have introns at all, even though introns may on the average represent

as much as 90% of the length of a gene and perhaps as much as half of the total DNA in some complex eukaryotes such as human" [65]. Thus, introns, rather than being kept and expanded in "highly evolved" or more complex organisms, may actually have negative or no fitness consequences whatsoever, and more freely expand in those species with small effective population size (Ne), where selection is not sufficiently efficient to remove them from the population [178, 175]. Indeed, the expansion of both introns and exons associated to transposable elements [108, 259, 5, 130] suggest that, at least at their initial expansion, have no major fitness consequences.

In the human genome, the percentage of genes with evidence of undergoing alternative splicing has been increasing with the development of new technologies: from first estimates of about 74% based on exon junction microarrays [122] to about 95% of genes based on RNA-seq data [209, 202]. Nowadays, it is widely accepted that there is widespread variation in the splicing process leading to a plethora of splicing variants that can be expressed in mammalian cells. During the last 40 years, the number of characterized splicing isoforms with different functions has been increasing [99], but much more slowly. Thus, it raises the question of how much of this transcriptome diversity is actually translated into protein or has any other functional role, a topic under intense debate over the last years [202, 276, 277, 70, 34].

First transcriptome-wide studies struggled to consistently find gene categories associated to alternative splicing, going from enzymatic to immune and neural functions [151]. This may have been limited by data availability, as later studies on regulated alternative splicing seem to coincide in genes related with cytoskeleton, ion channels and vesicle secretion [92, 221, 125]. Whereas associating up or down-regulation of the expression of genes with certain categories provide information about the function of the gene expression regulation in a particular scenario, it is more difficult for alternative splicing, as the impact would greatly depend on the actual part of the protein that is affected and on the specifics of its molecular function. Thus, it is more useful to associate alternative splicing events, not necessarily with genes, but with functional domains. Such studies show an association with KRAB, ankyrin repeat and tubulin-binding domains, which mainly mediate protein interactions [227]. Following this idea, several protein properties or features have been associated with alternatively spliced exons: they are mainly associated to predicted phosphorylation sites, disordered regions linking different functional domains, which tend to remain mostly unaffected by AS, and small linear motifs that are potentially involved in protein-protein interactions [194, 41, 296, 114, 87, 305]. Altogether, this suggests a non-random distribution of alternative exons along protein regions and functions. Whether it is beneficial to have alternative exons encoding protein sequences with certain properties, or it is detrimental to have them in sequences with opposite properties remains an unsolved issue.

As any biochemical reaction, the splicing reaction is not 100% efficient [99]. This inefficiency could explain, at least partially, the great variety of observed mRNA isoforms. Several sources of evidence point in that direction, although counter-arguments have been made for some of them.

- Most of these isoforms are predicted to be targeted by the NMD pathway and therefore degraded [151]. Although this suggests that few of these transcripts are actually translated into protein, and most of them will be degraded by the NMD, they may still have an important regulatory role through RUST, as previously described [150].

- Most genes have a single major isoform expressed at greater levels [70]. Although low abundance of the alternative products is more parsimoniously explained by a relatively low error rate in the splicing process, it does not necessarily imply that the low abundant transcripts are not important [34].

- If splicing errors are a byproduct of optimizing global splicing efficiency, stronger selective pressure will be placed on highly expressed genes to reduce error rates. Even if highly expressed genes

provide higher power to detect more splicing variants, they showed, in average, lower transcript diversity than lowly expressed genes [218]

- Another gene property that is expected to be negatively correlated with the number of transcript isoforms derived from a single gene is the number of introns: the more introns, the more likely is that at least one of the splicing reaction fails and generates a different isoform. A simple stochastic noise model accounting for number of introns and expression predicts a great deal of the observed transcript diversity [192]

- Proteomics fail to detect most of the predicted protein isoforms [276]. Nonetheless, high throughput proteomics is known to suffer from technical limitations, and has many difficulties detecting low abundant proteins. Although the authors estimated the proportion of estimated protein isoforms that are detectable taking into account these limitations and was still much higher than the observed ones, ribosome profiling experiments suggest that most of these isoforms are actually bound by the ribosome, and thus potentially translated [297]. This is not incompatible with proteomics results if most of these peptides are unstable or degraded, since NMD requires a pioneer translation round for detecting premature stop codons [170].

- Alternative exons are under weaker purifying selection at the sequence level in human populations: they show ratio of non-synonymous to synonymous substitutions (dN/dS) ratios and allele frequencies similar to neutral expectation [277]. This is a direct consequence of the first argument: if a good amount of them are not even protein coding, there would be no purifying selection operating on the protein sequence, but they may still have regulatory potential. Moreover, there may be heterogeneity within alternative exons: while most alternative exons may evolve under weaker stabilizing selection than constitutive ones in average, the same pattern may derive from a small percentage of positively selected alternative exons within a larger set of negatively selected background. Indeed, previous studies show that about 27% of aminoacids substitutions between human and chimp in alternative exons have been fixed by positive selection [224]. Thus, alternative splicing may contribute to lineage specific adaptations.

One, maybe the most definitive, way to settle the debate is the characterization of the function the alternative splicing isoforms, either at the regulatory mRNA level or at the protein function level, and unify this information in a carefully curated database. Some efforts in this direction have already been made, although mainly focused on human, mouse and other vertebrate species [268, 32]. This exhaustive work, although necessary, will take a very long time to yield more accurate estimates of the proportion of functional splicing variants in a given species.

1.1.7 Approaching AS function through comparative studies

The mere existence of many alternative splicing isoforms for most genes does not necessarily imply that they are functional, nor does it its association with expression changes, which may not have any impact themselves or changes in protein sequence with uncertain impact on the protein function most of the times. Inferring functionality is therefore a rather difficult task. A way to approach this question indirectly is to study the underlying evolutionary forces driving its evolution: functional features are more likely to be maintained during evolution, whereas non-functional traits are more likely to diverge. Alternative splicing conservation can be studied at different levels.

The first level regards the conservation of the alternative splicing event itself, this is, whether the two or more ways to process a transcript is maintained in orthologous genes of different species, or more precisely, if we have detected the different splice variants across these species. First studies during

the early 2000's seemed to obtain very variable estimates, from 8% to 98% conservation of alternative splicing events were reported to be conserved between mouse and humans [151, 208]. These studies were mainly based on EST information with uneven coverage across the different genomes, resulting in different statistical power to detect such AS events and hindering the interpretation of the results. What this type of approach allowed, however, was the discovery that most often the ancestral isoform was the longest, suggesting that most of the alternative protein isoform derive from alternativization of constitutive exons rather than exonization of intronic sequences. *De novo* exon formation seemed to be a minor but still relevant mechanism for generation of new protein isoforms [137]. Although it is unlikely that newly arisen exons provide novel functional sequence to a protein, it is more likely to produce a target of NMD, which could be subject to selection if such regulation becomes beneficial [151].

Alternative splicing can be studied also quantitatively: instead of assessing whether a certain isoform or just one more than one isoform can be produced in a certain species or scenario, one could aim to quantify or estimate the proportion of each isoform Ψ. Microarray technologies using exon junction sequences allowed for the first time to quantitatively study alternative splicing globally in a particular sample. We obtained the first estimates of tissue-specificity of alternative splicing patterns and found evidence for 74% of human genes to be alternatively spliced [122]. Using this technology in a comparative setting, it was shown that Ψ patterns are dynamically regulated during *C.elegans* development, and highly conserved when compared with *C.briggsae*. Similar trends were found when comparing Ψ values between human and chimp in brain and heart tissues [117, 43]. However, the fact that microarrays had to be designed *a priori* led to only examining a small amount of already known and conserved ES events in these studies. With the development of RNA-seq techniques, it became possible to interrogate all splice variants, known and unknown, across the whole transcriptome of several vertebrates species [194, 21]. These seminal studies showed an overall poor conservation of exon Ψs across the different species, down to 15% of conserved events between human and frog. In contrast to gene expression patterns, which are more similar within tissues than within species, AS patters are highly species-specific, except for a reduced set of conserved alternative exons across all species, which are regulated in a tissue-dependent manner. Moreover, Ψ values are more conserved in brain, which is also the tissue with more different splicing profiles when compared with other tissues in the same species [194, 21]. Interestingly, introduction of the human chromosome 21 into a mouse cell lines shows splicing patterns similar to those in human, suggesting that divergence of AS patterns are mostly driven by changes in cis-regulatory elements [21]. Similar conclusions have been obtained in comparative studies with different sets of species, like *Drosophila* and cichlids [88, 254].

Conservation *per se*, however, is insufficient to support the functionality of splicing rates, because we do not know how splicing rates evolve under neutral evolution, and thus can not calculate the deviation from this pattern. Although some have argued that these highly species-specific patterns suggest that AS patterns may drive or be associated to phenotypic diversification [21, 88], these patterns are more parsimoniously explained by an scenario of mostly neutral evolution of splicing rates Ψ in the absence of more clear evidence. When studying selection at the sequence level, we compare the observed substitution rate in a region of interest with some sequences that we expect to evolve neutrally, such as synonymous sites. Unfortunately, there is not such a clear way to define AS variants evolving neutrally to approach the problem with the same strategy. Alternatively, one can study the genetic forces shaping character evolution using of models of phenotypic evolution in a phylogenetic framework [73, 42, 82, 58]. These methods have been broadly applied to study macroscopic quantitative characters for a long time, and started to be applied to model the evolution of molecular traits such as gene expression over the last years [38, 26, 124, 233, 232, 47, 52]. These models allow, not only to estimate the strength of selection constraining a particular trait in a certain clade, but also to infer changes in phenotypic optima along the phylogeny. Using these models, it was shown that gene expression patterns evolve under a relatively

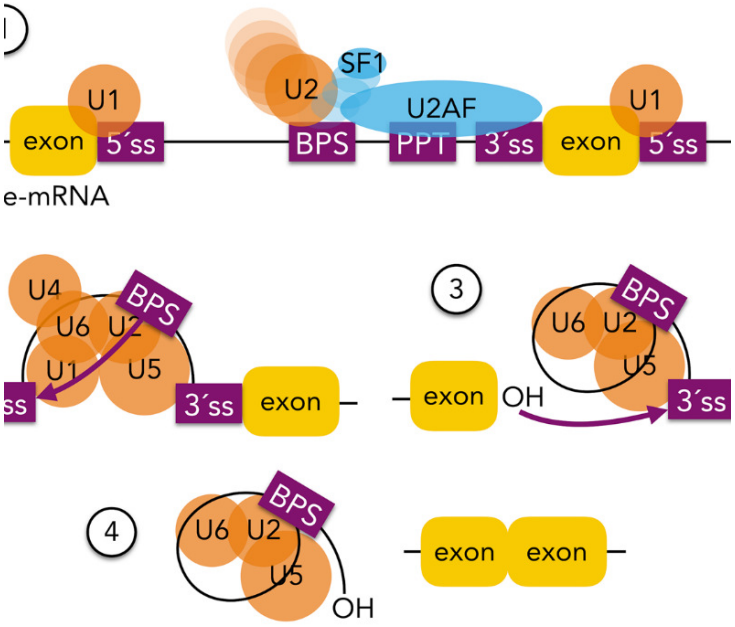

Figure 1.2: Spliceosome assembly on the intron and steps through which the splicing reaction takes place. Extracted from [80]

weak but clear stabilizing selection, and have proposed some interesting adaptive changes linked to diet changes in humans [233]. This approach, however, has not been applied to study alternative splicing evolution, in part due to the inherent difficulties introduced by the nature of AS data from RNA-seq experiments as compared to gene expression.

1.2 Regulation of alternative splicing

The splicing machinery, a large molecular complex, assembles around the intron to remove it by recognizing 4 main elements across the introns. In the first place, U1 snRNA binds to the 5' splice site or splice donor by base pairing of a few nucleotides. The SF1 and U2AF proteins bind to the BP and PPT, together with the U2 snRNA. The interaction between U1 and U2 bound to the boundaries of an exon allows the recruitment of the U4/5/6 complex and mediates the nucleophilic attack of the BP to the 5' splice donor. The splice donor is then free to attack the 3' splice acceptor to bind the two exons together and release the intron lariat [80] (Figure 1.2). In exon definition, SR proteins bind to the exon body and allow the interaction between the U2 complex and the U1 bound to the 5' splice site of the exon to be spliced. This interaction enhances the recognition of the exons within the long intronic sequences [14]. From now on, we will focus on the regulatory mechanisms as described.

In this reaction framework, one can imagine that the splicing reaction efficiency can be modulated by interference with any step of the process, which can be classified in to major classes depending on their nature: cis and trans. Cis-elements are located nearby the splice sites and are genetically linked to the splicing outcome. Cis-elements in the RNA sequence are bound by trans-regulatory and mediate their interaction with the spliceosomal machinery for modification of the splicing outcome, either enhancing it or inhibiting it.

1.2.1 Cis-regulatory mechanisms and the genetic architecture of splicing rates

Besides the 4 key regulatory elements present in every intron, there are other cis-acting elements on the RNA sequence, known as splicing enhancers and silencers, to which different regulatory proteins, including SR and hnRNP proteins, can bind to facilitate or block the recognition of exons or introns by the spliceosome [51]. Naturally, cis-regulatory elements remain constant within an individual along the developmental and differentiation trajectories and thus cannot account for differences between splicing patters within a given individual in different environmental conditions. However, between individual genetic variation in these cis-regulatory elements lead to differences in the splicing patterns and can provide information about the genetic architecture of splicing rates i.e. how the DNA sequence determines the phenotype.

Large studies genetic association studies in human populations, like GTex, have identified genetic variants associated to changes in splicing rates. These variants are strongly associated with the splice sites, but are found all along the transcripts, suggesting a complex genetic architecture. Moreover, the effect of these variants was found to be mostly the same across tissues even if there was a relatively high tissue-specific non-genetic variation in splicing rates [196, 3]. These association studies are an indirect way of assessing the effect of genetic variants on the different traits. However, both the environment and the genetic context, together with population structure, hinder the inference of causality. An alternative approach to better understand how genetic variation controls splicing rates is deep mutational screening (DMS). These assays test the effect of large number of mutations and their interactions on the splicing rates of mini-gene constructs derived from real exon sequences in a more controlled biological context. Mutations affecting Ψ are distributed all along the exon, most of them have a negative effect on splicing rates and show extensive pairwise epistatic interactions between different genetic variants and gene contexts [123, 299]. This highly epistatic landscape, in which the effect of a mutation depends on the Ψ of the background sequence, can be easily explained by a competition model between splice sites with different affinities for the spliceosome, which are affected by mutations or variants [16].

1.2.2 Splicing regulation by trans-acting factors

Cis-regulatory elements are bound by trans-acting factors through a RNA binding domain and modulate the assembly or activity of some components of the spliceosome to enhance or prevent the splicing reaction. This regulatory effect is mediated by protein-protein interaction domains that are also present in these RNA binding protein (RBP)s. These elements can act at every step of the spliceosome assembly and even during the conformation changes that mediate the catalysis of the trans-esterification reactions resulting in intron removal [51]. Thus, changes in the activity of RBPs typically results in dynamic regulation of splicing patterns. Splicing patterns dynamically change during embryonic development and across different tissues like liver, brain, skeletal and cardiac muscles; suggesting the existence of an underlying regulatory network driving them in a coordinated manner [18]. It is relatively well studied that regulated alternative splicing requires relatively weak splice sites, such that modulation of the activity of splicing regulators can effectively modify the splicing rates [14, 113, 19]. Interestingly, no association between splicing and expression changes is generally found, suggesting the uncoupling of the two mechanisms, except for IR events [19, 280].

Over the last years, much effort has focused on the identification of the binding targets of the RBPs enconded in our genome, hoping to improve our understanding of the regulatory mechanisms behind these dynamic changes in splicing rates. First large scale systematic assays of *in vitro* RBPs binding affinities to different RNA sequences found a great conservation of these binding affinities of homologous proteins across *metazoan*. Most RBPs showed binding to single stranded RNA with no apparent preference for RNA secondary structures. Although they showed *in vitro* binding was a relatively good predictor of

in vivo binding [225], several high-throughput techniques have been developed to examine the binding specificity *in vivo* of some RBPs [265, 236, 281, 199, 13]. With time, the results of these experiments have been collected in different databases, like DoRiNA, starBase2, CLIPdb or CLIPZ [301, 35, 131, 158]. Later, as part of ENCODE consortium, a much larger set of human RBPs have been characterized in cell lines, showing a larger context and RNA structure dependent *in vivo* binding for most RBP than previously anticipated. This greater dependency on RNA structure could explain most of the differences from *in vitro* predictions [64, 267], leading to the development of complex tools that incorporate 2D and 3D RNA structure information for inference and prediction of binding preferences [311, 78, 187, 210]. Thus, RNA structure can play an important role in regulation by exposing or masking cis-regulatory elements. However, it can also act directly by directly blocking the recognition of splice sites [251] or whole exons e.g. exon 6b of chicken β-tropomyosin can form stem and loop secondary structures that prevent its inclusion in the mature transcript [51].

RBPs, by definition, bind to RNA, but they can perform a wide variety of functions besides splicing regulation, like mRNA transport, stabilization and degradation [198] e.g. by interacting and binding with micro RNA (miRNA) to their targets [204]. These functions are probably not mutually exclusive, given the conservation of motifs in different regions of the transcripts that are relevant to different processes [225] e.g. CELF4 has been reported to regulate both splicing and translation [142, 143, 289], whereas MBNL1 can regulate both splicing and mRNA decay by binding to exons boundaries and 3'UTR, respectively. To add more complexity, regulators may simultaneously act as inhibitors or activators depending on the binding location [51] e.g. MBNL1 blocks splicing when binding to the upstream intronic flank, but promotes it if binding to the downstream intron [186]. Competition for binding sites with other splicing factors may explain the dual regulatory role of many RBPs, working both as activators and inhibitors [51].

Although most of the research on splicing regulators has been naturally focused on RBPs, recent work suggest that a good amount of transcription factor (TF)s can actually bind RNA and modulate splicing rates [103]. Additionally, splicing regulators have been mostly studied in isolation, but there are well known cases of interaction between the binding of different regulators e.g. SRSF7 and hnRNPF compete for binding to the sample cis-regulatory element in exon 2 of α-tropomyosin [51]. Thus, little is known about how different RBPs interact, coordinate and compete to create a more complex and rich regulatory network modulating and maintaining splicing rates.

1.2.3 Kinetic model of co-transcriptional splicing and derived regulatory features

Splicing takes place simultaneously to transcription by RNA polymerase II, 5' capping and 3' polyadenylation, suggesting a potential interaction between these processes. A good number of proteins interacting with the RNA polymerase II complex are known splicing regulators. These interactions are thought to increase overall efficiency by enhancing binding of the spliceosome to the newly synthesized pre-mRNA before its diffusion. Mutants of the RNA polymerase II with reduced elongation rates showed altered the splicing rates, both *in vivo* and *in vitro*. This regulation can be easily explained by a competition model between splice sites: high elongation rates will shorten the time a splice acceptor is not competing with following ones, increasing the chances of alternative splicing [16, 251, 200]. During exon definition, slow elongation may also favor the recruitment of factors to the upstream exon before the recognition of the second one. However, not every exon is dependent on transcription elongation rates which can only be explained by a mathematical models with 3, rather than 1, limiting steps in the splicing kinetics i.e. transcription, assembly and catalysis [200].

There is a strong association between exons and nucleosomes across a wide range of animals, including

C.elegans, *D.melanogaster* and *M,musculus*, possibly driven by GC content bias [243]. This association appears to be not only at the positional level, as human exons have an average length of ∼150 nucleotides, very similar to the amount of DNA that can be wrapped around nucleosomes [271, 243]. Interestingly, changes in nucleosome occupancy are associated to changes in splicing rates, suggesting a potential regulation of splicing rates [109, 110]. Indeed, nucleosome occupancy is associated to the repression of pseudo-exons and its increase creation of new exons during evolution [271, 193, 163]. Current models explain these observations as consequence of the slowdown of the transcription at nucleosomes, which facilitates the recognition of the exon and delays the appearance of other competing splice sites [271, 110]. This association may be bidirectional, as modification of splicing patters seem to also induce changes in nucleosome occupancy, potentially through interactions between RBPs and histone modifying enzymes e.g. Hu splicing factors interact with HDAC2 and inhibit its acetylase activity [129, 200].

The kinetic model predicts that any local variation of the transcription rate by RNA polymerase II or their interaction with splicing factors, including chromatin marks, DNA methylation or chromatin remodelling proteins and complexes; can serve as indirect regulatory mechanisms for AS e.g. H3K36me3, a modification in actively transcribed genes, can recruit PTBP1 through MRG15, which then binds to their targets in pre-mRNA and modulate splice site selection [200].

1.2.4 Towards the splicing code

As more and more factors involved in regulation or associated with variation in splicing rates were discovered, the idea of integrating all this knowledge in what came to be known as *splicing code* became one of the most challenging and important tasks in the field. The *splicing code* would allow to interpret the information stored in the genome and present in the pre-mRNA and their impact on relative isoform abundances. It would imply the identification of cis-regulatory elements and how the context-dependent trans-regulators interact with them to regulate the observed splicing patterns across different conditions e.g. tissues, disease, mutations, etc; together with other features affecting splicing rates. First attempts to decipher such code used only sequence information and Support Vector Machine (SVM) to identify splicing enhancers and silencers, which, independently of splice sites, could explain differences in splicing rates using pseudoexons i.e. intronic sequences with splice sites and branch points that are not recognized as exons, as negative controls [76, 312]. Next steps required the identification of trans-factors binding to these cis-regulatory elements to start building the regulatory networks driving splicing changes within an individual e.g. during development, between tissues or different environmental conditions, and aiming to integrate it with RNA polymerase II elongation speed [294]. The next version of the *splicing code* aimed to predict tissue-specific AS patterns, and incorporated RNA structure and sequence conservation information, as well as gene structure. This model was very accurate distinguishing tissue-regulated exons from constitutive exons (area under the Receiver Operating Characteristic curve (AUROC)=0.94 in 5-fold Cross-Validation (CV)) in a set of mouse tissues[20]. These results were further improved using deep learning to predict, not only tissue-regulated splicing, but quantitative changes in the Ψ, being particularly good at predicting intermediate splicing rates. The improved performance of deep learning over linear models suggests that there are widespread non-linearities in the regulatory system [156]. A similar approach was taken with human samples from healthy donors across different tissues, showing an overall $r^2 = 0.65$ for prediction of Ψs, reaching $r^2 = 0.94$ if training only with low variability exons. Interestingly, this model allowed predicting the effect of genetic variants on splicing rates for known mendelian diseases like spinar muscular atrophy (SMA) [300]. These models, however, provide an static picture of tissue regulated splicing and ignore the potential interactions between cis and trans regulatory elements. To tackle this issue, a new deep learning method was trained adding the information about the expression of trans-regulators and data from knock-down experiments. This new model can now make

predictions for a combination of cis and trans-regulatory elements specific for any condition, and even serve as prior distribution for inference of splicing rates from RNA-seq data [315].

We have seen how, over the years, more known regulatory features are being incorporated into these computational models for the *splicing code*. Although there are still some regulatory features that may keep improving the results, like nucleosome positioning or transcriptional data, these increasingly complex deep learning models are already relatively good at making predictions on splicing rates or differences between tissues or different conditions. However, the interpretation of the results and the understanding of the underlying regulatory mechanisms in more diverse environmental conditions and how the different layers are coordinated remain a challenging task.

1.2.5 Computational methods for studying alternative splicing

The development of RNA-seq technologies was key to improve our understanding of the extent of alternative processing that takes places in a human cell [202]. It allowed us to estimate the Ψ across different types of AS events in an unbiased manner, without having to define *a priori* a set of known events and the sequences to quantify them. This new type of data, however, required new methods for estimation of Ψs, which have been under continuous development since their birth.

The output of a RNA-seq experiment is a large collection of short reads corresponding to one or the 2 ends of fragments of a cDNA library. Thus, the first requirement is to guess the most likely position in the genome that gave rise to each of the reads in the sequencing output. Optimal alignment techniques are very computationally expensive for such a large amount of sequences. Given that one of the sequences to align was very long and always the same (the genome), clever strategies to index the genome to accelerate sequence search, like the Burrows-Wheeler transformation, were developed [274]. However, the presence of introns in the eukaryotic genomes imposed further difficulties for mapping RNA-seq reads back to the genome, as the observed reads may not be contiguous in the reference sequence but interrupted by introns. With this aim, TopHat was the first developed program allowing alignment of reads across the splice junctions [273], followed by other tools, like MapSplice, RUM, TopHat2 and STAR, whose performance has been systematically evaluated over time [291, 96, 132, 63]. Latest benchmarking studies point towards Hisat2 [217] and STAR as best RNA-seq aligners [**Teng2016**, 84, 23, 183].

Once the genomic position from which each read is known, one can start to build models that allow inference of gene expression and relative transcript abundance based on the number of reads mapping to each position in the transcriptome. MISO was the first program specifically designed to estimate splicing rates. It models how reads can be generated from different splice isoforms or AS events taking into account some of the known biases and the nature of RNA-seq data, and performs bayesian inference on the underlying Ψ [128]. To check for differences in splicing rates across different samples, it uses Bayes Factors to compare a model in which both samples show the same Ψ with a second model with a different Ψ per sample. This approach not only limits the potential experimental designs that can be analyzed, but more importantly, it does not take into account biological replicates and potential environmental and biological variability within groups. Since then, a wide variety of tools have been developed, being MATS and DEXseq some of the most popular [12, 248, 249] (reviewed by [4]). Some methods have been developed to focus on different aspects of types of AS, e.g. vast-tools focuses on detection of very small ES events [114, 268]; IRFinder focuses on IR events [195]. More recent tools like whippet and MAJIQ aimed to identify and quantify more complex AS events [261, 287], even if interpretation of those complex events remains challenging. Other methods have focused on improving estimations by building an approximate informative prior distribution using event information: BRIE uses k-mer composition as proxy of cis-regulatory elements to pool information across all events and be able to estimate Ψs in single cell RNA-seq experiments [107]. DARTS, as previously mentioned, uses information about

both cis and trans-regulatory elements to build a prior distribution that can be combined with RNA-seq data to produce a posterior probability distribution of the inclusion rates and the differences between experimental conditions [315].

1.2.6 Computational approaches to study alternative splicing regulation

When studying AS changes among different conditions e.g. different tissues, developmental stages, differentiation processes, disease conditions, etc. it is generally assumed that observed changes are regulated by variation in the activity of trans-regulatory elements that allow a common response among individuals subjected to the same stimuli. Hence, one of the most relevant questions when studying AS dynamics is: what are the regulatory proteins driving these changes?

If at least some of the identified AS changes are under the control of a regulatory network, one expects a non-random association between the AS changes and the cis-regulatory elements bound by the trans-acting factors with varying activities. As many of these cis elements are well characterized for a number of proteins, as previously explained, and this information is available in different databases [91, 301, 225, 64], one can count the number of AS events with and without the regulatory feature and test if significantly changed AS have an over-representation of such features using a Fisher test (Over-representation Analysis (ORA)). A sufficiently large set of significantly changed events is required to reach enough statistical power to reliably detect enrichment of regulatory features. Therefore, as ORA requires the categorization of splicing changes into different groups e.g. included or skipped, or clusters with a given temporal trend, it ignores quantitative information about AS changes.

Several approaches have been developed to make use of quantitative information in the enrichment procedure, including the widely known Gene Set Enrichment Analysis (GSEA) [264, 252]. Although these tools were designed for functional analysis, they have been used to perform enrichment of known targets of regulatory elements as the rationale behind the analysis is exactly the same [278, 246]. However, the inherently noisier nature of the estimation of differences in AS compared to those of differential gene expression may limit the applicability of GSEA-like methods. Moreover, an additional limitation affecting both ORA and GSEA approaches lies on the high number of different features or binding sites that we usually want to explore and on the potential co-linearities among them. Similar binding profiles introduce confounding effects: RBPs with similar binding profiles to the actively regulatory factor would also appear to be affected. Therefore, this issue is expected to introduce a high number of false positive associations. Despite all these theoretical limitations, to our knowledge, there is no systematic evaluation of the performance of these tools for the identification of trans-regulatory elements driving AS changes

1.3 Function and regulation of alternative splicing in health and disease

Of particular interest is understanding the specific consequences and regulation of splicing across human tissues and conditions. This is often approximated using animal models that facilitate the investigation and study of different biological processes of interest, such as embryonic development, differentiation of cell types or diseases conditions. Knowing when and where splicing changes take place may also help to understand the potential function of those AS events and their implications for human physiology and disease. Indeed, recent genetic studies suggest that splicing variation may be a key mediator between genetic variation and disease [162].

First transcriptome-wide studies showed that a large amount of splicing transitions along different developmental and differentiation processes take place simultaneously. The same has been described for RBPs, potential candidates for splicing regulation. This coordination is highly suggestive of an underlying

regulatory system driving most of AS changes [18]. Moreover, if such regulatory system has evolved during mammalian diversification to control specific splicing changes, one can argue that these splicing changes, individually or in coordination, must play a role in organ and tissue physiology. Malfunctioning of those systems, on the other hand, may lead to pathological conditions.

1.3.1 Alternative splicing in neuron function and disease

Comparative studies showed that alternative splicing patterns are particularly conserved and specific in a subset of tissues, including brain, muscle and heart [21, 194]. Thus, brain has been the preferred tissue to study the global impact of AS over the last years. A large number of genes are regulated by AS during neuron differentiation and across different subtypes of neurons [18]. These coordinated changes in AS patterns take place in two main waves: an early switch, taking place at birth; and a late wave, after 2 weeks of life. These two switches, associated with overall increased exon inclusion, seem to be associated with different biological processes and driven by different regulatory elements [298]. A subset of exons that are particularly enriched in brain tissues are microexons i.e. very short exons, down to 3 nucleotides [114, 161, 303, 242].

Members of the Nova, Celf, Mbnl, Rbfox and Ptb families have been characterized as important regulators of neuronal differentiation [18]. Ptb proteins in particular inhibit, at least partially, the neuron AS program in undifferentiated cells by binding to the upstream intronic flank and blocking the binding of U2AF to the PPT [234]. As it is downregulated during differentiation, all the repressed exons and start to be included in the corresponding transcripts, which is enough to drive neuron differentation from fibroblasts [80]. Different Ptb genes show an interesting interplay in their regulatory program: Ptbp1 promotes the inclusion of exon 10, a cassette exon with an in frame stop codon, in Ptbp2 gene (also known as nPTB) [260]. In presence of Ptbp1, Ptbp2 transcripts are degraded by the NMD pathway. However, as Ptbp1 is downregulated during differentiation, Ptbp2 starts to be produced. Ptbp2 binds to similar targets as its paralog, suggesting that Ptbp2 may be repressing Ptbp1 targets that are to be kept repressed in later stages of differentiation [**Vuong2016**, 165, 166]. Indeed, Ptbp2 is required for proper neuron maturation, by regulating genes related with neurite growth and synaptic assembly and transmission [160]. One interesting target of Ptbp1 is the transcription factor Pbx1. As Ptbp1 is downregulated, exon 7 in Pbx1 is included to produce a new isoform, which activates the transcription of known neuronal genes [166]. Other important AS regulators for neuron function belong to the Rbfox family. Rbfox1 is upregulated during neuronal differentiation and promotes exon inclusion by binding to their downstream intronic flank [75]. Central nervous system specific knock-out of Rbfox1 in mice drives multiple splicing changes in genes associated to synaptic transmission and epilepsy [85, 298]. Whereas Ptbp1 and Rbfox1 participate in AS regulation throughout neuronal differentation, the late AS switch in neuron differentiation is probably mediated specifically by Mbnl [298]. Within neuronal AS regulatory networks, microexons are regulated by a specific regulator called nSR100/Srmm4, essential for differentiation, besides more general neural regulators Ptbp1 and Rbofx1. nSR100 binds is able to outcompete Ptbp1 and instead promote the inclusion of their target exons [221, 114, 161, 223].

Both specific splicing changes that are regulated during neuronal differentiation and global changes induced by changes and mutations in the AS neural regulatory program have been recurrently associated with neurological disorders such as autism, epilepsy, Parkinson or Alzheimer [114, 161, 80, 18, 152]

1.3.2 Alternative splicing in cancer

Another widely studied human condition in which splicing is particularly altered is cancer. Mutation of key cis-regulatory elements of the splicing reaction, like splice sites, is a common mechanism for gene loss of function. As loss of function mutations in tumor suppressor genes lead to cancer development, one

can expect that splicing alterations are relatively common causes of cancer. In this line, APC gene shows recurrent mutation in splice sites in familial and sporadic cases of colon cancer, whereas a part of Brca1 loss of function mutations predisposing or causing cancer happen on splice sites [269]. Numa1 alternative splicing is able to remove a potential phosphorylation site and lead to an increase in proliferation in normal cells, and has been often characterized in cancer types, even if no clear mechanism as cancer driver is known [246].

Thus, a part of cancer research has been focused on the characterization of global AS patters across different cancer types. There is a subset of AS events that are recurrently altered across a wide variety of cancers. These events tend to be differentially altered across cancer types, and can be used to predict patient survival [279]. Following this idea, several studies have tried to build AS signatures to predict survival across different cancer types: breast, prostate, melanoma, pancreatic ductal adenocarcinoma or endometrial cancer [288, 44, 181, 302, 293]. Interestingly, AS patterns are reported to predict better patient survival than do gene expression patterns across several cancer types [250]. To better understand its predictive effect, several studies have focused on different aspects of the functional characterization of the cancer related AS changes. In general terms, cancer associated splicing changes are associated with protein domains that are frequently mutated in cancer and potentially disrupt protein-protein interactions, but negatively correlated with the number of mutations in those genes. This observation suggests that alternative splicing may provide an additional mutational path to loss of interactions leading cancer development [57]. AS can not only have an important role at modulating and changing protein sequence, but also at a regulatory level by coupling with the NMD pathway [150]. Indeed, cancer mutations associated to IR are enriched in tumor suppressor genes [80].

Cancer associated AS changes can be driven by cis or trans regulatory somatic variation. Although mutations affecting cis-regulatory elements, such as splice sites, are the most commonly studied ones, there is increasing evidence of important contribution of trans-regulators to shape cancer AS patterns. Alterations of specific RBPs, like Tra2b1 and Yb1, have been previously associated with cancer development [269]. Their expression is often altered due to widespread variation in copy number variants and the coordinated downregulation of a number of RBPs (including Mbnl1 Rbm20 and Rbfox1 and 2) is speculated to be driven by mutations in common enhancers [246]. Srsf1, often upregulated in cancer, is associated with a general increase in the complexity of splicing patterns in cancer, such that its overexpression in cell lines partially recapitulates cancer patterns [261]. Overall, cancer splicing patterns are more similar to those of undifferentiated cells, potentially driven by Mbnl1 [246], suggesting that cancer can also make use of the existing regulatory networks to modulate their AS patterns globally. Moreover, if there are splicing regulators modulating pro or anti-cancer AS patterns, cancer can evolve not only by targeting the regulator, but also their cis-regulatory elements. Indeed, there are known RBPs, like Srsf10 or Pcbp1, whose targets are particularly enriched in cancer mutations [**Singh2017a**].

1.3.3 Alternative splicing in heart function and disease

Skeletal muscle and heart are, besides brain, the tissues with more specific and conserved AS patterns, which suggests a particular implication of this regulatory mechanism in the physiology of the tissues [21, 194]. There are very well cases of key genes for cardiac contraction that undergo AS: inclusion of exon 5 of the cardiac troponin T is decreased in adult cardiac muscle compared with the embryonic one [80]. Another sarcomeric protein undergoing AS is Tropomyosin. There are 4 genes encoding different tropomyosin isoforms, which, at the same time, can be expanded through MXE for tuning the interaction between actin and myosin during development [149]. But one of the most classical examples involves the largest protein in the human genome: Titin. Titin is a protein formed by 363 exons, some of which encode repeated Ig-like domains that fold and unfold in response to mechanical forces as a molecular spring. This

protein is inserted in the sarcomere and is thought to provide mechanical resistance and stiffness to the tissue. AS can be used to modulate the number of domains and therefore the passive stiffness of the tissue. And this is what happens during post-natal heart development: AS changes produce shorter isoforms and increase the passive stiffness of the tissue to better respond to the increased mechanical stress after birth [144, 295]. Despite these and other very well known examples (reviewed by [149] and [285]), it was not until the development of transcriptomics techniques that we started to have a glimpse of the global impact of AS in heart physiology.

As with neurons, much of the research effort on transcriptome-wide AS dynamics has focused on the study of the differentiation and development of the highly specialized cells composing most of the tissue: the cardiac myocytes. First microarray results on developmental patterns also showed a biphasic response, with a switch during late embryonic development, and a second late response after birth. This program was conserved between mouse and chicken, suggesting that the AS changes are important for the development of functional heart, and proposed to be under control of Celf and Mbnl gene families [126]. These studies, later expanded with RNA-seq to analyze many more AS events, show that genes involved in vesicular trafficking are particularly regulated by AS during post-natal development, particularly affecting cardiac myocytes. They hypothesize that AS changes are key for proper organization of ion channels and distribution of cellular components in an increasingly hypetrophic cardiomyocyte as with their maturation process. Indeed, Celf1 downregulation was shown to be required for proper establishment of t-tubules [92]. Thus, as with neurons, the perturbation of the AS regulatory network, in this case Celf1, leads to impaired tissue function. Loss of function models for Rbfox1 or Srsf3 led to the development of cardiac hypertrophy, and severe cardiac disfunction, respectively [81, 207]. Rbfox1, in particular, modulates the splicing of Mef2d, a key transcription factor for the activation of the late expression program during myogenesis [80]. Ptbp1 downregulation promoted the trans-differentiation of fibroblast to cardiomyocytes, and represents a critical barrier to acquire the cardiomyocyte-specific AS patterns [169].

Defects in the splicing of particular sets of proteins, including Tnnt-2 TnnI3, Myh7 and Flnc, have been previously associated with ischemic cardiomyopathy in humans, and even predictive of the disease [138]. This is not limited to sarcomeric genes, as it also affects genes related to in ion handling, like Scn5a; and cardiomyocytes identity, such as Gata4 or Tbx5. Modification of the splicing patterns of these genes have been associated with cardiac diseases [285, 149], mostly by disrupting constitutive splicing and inducing degradation by NMD. On the other hand, changes in the relative isoform abundance of genes known to be alternatively spliced, have also been associated to cardiac diseases: patients with congestive heart failure (HF) and dilated cardiomyopathy (DCM) showed changes in the ratios of Titin protein isoforms [80]. Similarly, re-expression of the embryonic splicing isoform of EH-myomesin in adult hearts has been described as a marker for HF and is strongly induced in patients with DCM. Mouse models lacking AS isoforms for key genes involved in Ca^{2+} hanling i.e. Serca2 and RyR2 showed indeed contractility and relaxation problems [149]. These examples suggest that re-expression of fetal patterns, besides disruption of constitutive splicing, might drive cardiac disease. Even if transcriptome-wide analysis revealed a partial re-expression of the fetal transcript and gene expression patterns [11], whether this happens also for relative mRNA isoform abundances or at the AS event level remains unclear.

In this scenario, it is reasonable to think that, as with neurons, perturbation of the AS regulatory networks involved in heart development and cardiomyocyte differentiation, controlling known events to be associated to cardiac diseases, may have an impact on cardiac function and cause the disease. It is the case of Rbm20, which modulates the AS patterns of Titin during development, and is itself associated with increased risk of HF and sudden death in humans. Animal models lacking Rbm20 show defective splicing patterns in Titin. Moreover, Rbm20 also modulates genes involved in Ca^{2+} handling, such as CamkIId and RyR2, which activate a specific Ca^{2+} current in the mouse model and possibly explain the association with arrhythmias observed in humans [286, 100]. KO mice for Srsf2, another regulator of

RyR2 AS, also develop DCM, even if other downstream genes and isoforms may be involved [149]. An interesting disease caused by the indirect loss of function of a splicing regulator is myotonic distrophy (DM): even if the trans-regulator Mbnl1 remains unaltered itself, an expansion of CTG repeats in the 3'UTR of the Dmpk gene is thought to sequester this regulator and prevent the regulation of the AS patters of other genes, potentially leading to cardiac dysfunction [149].

Most of the current knowledge on how AS patterns change in the heart and its underlying regulation are derived human genetic studies of specific mutations or from mouse models defective in the expression of a single RBP. While the latter allow the investigation of whether a particular AS regulator is involved in the development, maintenance and proper functioning of the heart, few studies have profoundly characterized the AS changes taking place in animal models of general heart diseases like myocardial infarction (MI) or trans-aortic constriction (TAC) and how they are regulated in an unbiased manner.

2. Objectives

In this thesis, we aimed to study transcriptome-wide AS from different perspectives to explore, not only how specific AS events may impact the physiology of the heart, but also to understand what are the mechanisms that underlie global dynamical changes across development and disease and the rules that govern the quantitative evolution of exon inclusion rates. Under this general purpose, we specified the following objectives:

1. To study the functional impact of the AS changes that take place during heart disease and compare them to developmental transitions
2. To investigate the molecular mechanisms underlying these dynamic changes in both sets of conditions, focusing on the identification of the trans-regulatory elements that modulate AS patterns upon modification of their activity
3. To develop new computational tools to study AS regulatory patterns from RNA-seq data
4. To characterize the evolutionary process driving divergence of AS patterns during mammalian evolution using models of phenotypic evolution.
5. To estimate the relative contribution of different evolutionary forces to this process and the degree to which AS has contributed to lineage specific adaptations.

3. Materials and methods

3.1 Functional impact and regulation of alternative splicing in heart development and disease

3.1.1 Dataset

We collected a series of 21 RNA-Seq experiments related to the mouse heart. Table S1 summarizes the main biological and technical characteristics of the samples used: run, experiment, paired-end, sample type, condition, sample ID, sequencer and genetic background. These experiments included samples from isolated cardiomyocytes, left ventricle, both ventricles, and the full heart at different developmental stages (embryonic, neonatal, and adult). They also included border and remote-area samples from mouse models of myocardial infarction (MI) and myocardial samples from the TAC model of pressure-overload-induced cardiac hypertrophy (Table S1). We filtered out samples from tissues or cell types different from those listed above as well as those from KO mice. Samples used as controls in the collected KO, TAC, or MI experiments were added to the corresponding pool of adult, neonatal, or embryonic heart samples. We additionally included data generated by our lab from infarcted mice at 7 days post-infarction (MI7d), performed as previously described [72], resulting in a total of 136 samples.

3.1.2 Gene expression and alternative splicing analysis

GE and AS were quantified using vast-tools [113]. First, reads mapping to the genome were removed, and unmapped reads were later mapped to a library of exon-exon and exon-intron junction sequences to quantify inclusion levels for different AS event types. For Gene expression (GE) analysis, we filtered out samples with less than 1M reads and selected as expressed genes those genes with at least 1 reads per kilobasepair and million (RPKM) in at least 5 samples. We then used *limma* [229] with *voom* normalization to find differentially expressed genes using the experiment as random effect in the linear regression model. Since only one random variable is allowed, in most cases, the experiment included simultaneously both tissue type and batch effect. Differentially expressed genes (DEG) were defined as those with an adjusted p-value <0.01 and an absolute log(FoldChange)>1.

For AS analysis, we aimed to estimate the probability of inclusion of a particular event Ψ and to find significant differences in inclusion probabilities $\Delta\Psi$ between conditions. We used corrected inclusion and skipping reads from vast-tools results, and selected those events supporting alternative usage (defined as having at least one read mapping to the alternative event) in at least 20% of the samples. We then used a Generalized Linear Mixed Model (GLMM) in *lme4* [25] with binomial likelihood and logit link function to find differentially spliced events. As random effects, we added as covariates the experiment ID, sample type, and individual. In this way, we added biological variability to the binomial variance in the model. We used the adult stage as baseline and stored the p-value of coefficients corresponding to the different conditions under study. Multiple test correction was applied using the Benjamini-Hochberg

(BH) method. We considered as differentially spliced those events with a False discovery rate (FDR) <0.01 and an estimated absolute $|\Delta\Psi| > 0.1$.

3.1.3 Principal component analysis

Principal Component Analysis (PCA) was performed using the *prcomp* function in R on the log transformed normalized counts (adding a pseudocount) for expression. For AS, sample Ψ for each event was estimated by dividing the number of reads supporting inclusion by the total number of reads mapping to the event. We then used these estimated Ψ values as the input for PCA. Genes or exons with missing entries were removed for this analysis. All PCAs were carried out without performing batch correction to check whether the batch showed an important contribution to global variability.

3.1.4 Analysis of effect of alternative splicing changes on protein-protein interaction networks

As protein-protein interaction (PPI) can not be directly inferred from protein sequences alone, to analyze the impact of AS changes in the regulation of PPI networks, we took two different approaches. First, we focused on interactions mediated by specic protein domains or domain-domain interaction (DDI). DDI data were downloaded from [87], and only pairs with corresponding human-mouse one to one orthologs were used. Mouse-human orthologs were downloaded from Ensembl-Biomart and interactions were transformed to mouse reference assuming that they will be mostly conserved. Exon domain data was downloaded from VASTDB [268], and an interaction was considered to be potentially regulated by Alternative splicing (AS) whenever it was mediated by a domain overlapping, at least partially, with an alternative exon. We then tested for an enrichment of exons mapping to domains involved in PPIs in exons that were either skipped or included in each comparison independently. This was done using a GLM with logit link function: we defined as outcome whether the inclusion or skipping of an exon was affecting a domain involved in a PPI, which can be modelled as a *Bernoulli* distribution, whose underlying parameter depended on the exon group (Included, Skipped or No-change, taking the last one as baseline). We then extracted the estimates for each coefficient to estimate the probabilities for each group and their p-values to test whether exons with increased or decreased inclusion rates were more or less likely to be affecting PPIs than those exons without significant inclusion rate changes in each particular transition. Secondly, as many PPIs are not necessarily mediated by structured protein domains but by linear disordered regions of the protein structure, we performed a second analysis based on experimental data on different protein isoforms. Isoform specific interactions were downloaded from [305]. Data was collected at the isoform level, whereas our analysis was done at the exon level. To combine both, we performed the analysis at the gene level, and could not identify exon inclusion or skipping with gain or loss of interaction. Therefore, we could only test whether genes that were changing at the AS level had also isoforms showing differential protein interaction patterns. Statistical analysis was now performed at the interaction level: we modeled the probability of an interaction to be isoform specific as a function of whether the tested gene was found to have significantly altered exons in the heart and whether it showed changes in a particular contrast using again a GLM with logit link function. Estimations and significance were calculated as previously described.

Finally, each dataset was used to build an undirected graph representing AS-dependent protein-protein interaction networks. We then used Networkx python library to estimate edge betweenness for each interaction. The betweenness measures how many of the shortest paths between pairs of nodes in a graph go through a particular edge. Thus, edges connecting different modules will have a high betweenness, whereas edges that do not affect network connectivity will show low betweenness. To test for differences in betweenness, we used a Linear Model (LM) in R, assuming that the edge betweenness

is normally distributed and is a function of whether the interaction is potentially regulated by AS (if applicable) and whether significant changes were observed in each comparison performed.

3.1.5 Gene ontology category analysis

An L1-regularized logistic regression with Gene Ontology (GO) categories as group predictors was used to select meaningful and independent categories in Scikit-learn [216]. Standard logistic regression in statsmodels [245] was then used to find categories with an increased probability of being represented in the selected gene set. BH multiple test correction was performed. We then selected the top 10 categories for each comparison and calculated pairwise semantic similarity using GOSemSim [308]) for the Biological Process ontology. A heatmap with hierarchical clustering was calculated using these distances and python seaborn clustermap function.

3.1.6 CLiP-seq enrichment analysis

We extracted sequences corresponding to the 250 bp closest to the alternatively spliced exon in the flanking introns and 50 bp at both ends of the alternative and flanking exons. We then downloaded data on experimentally determined binding sites for mouse RBPs from doRiNA, Starbase, CLiPdb and ENCODE [35, 158, 301, 64]. For each database, we considered any binding site detected in any sample. Overlapping binding sites for the same protein were merged using bedtools [222]. CLiPdb data, provided for reference mm10, were then transformed to mm9 coordinates using the liftOver utility. Since results from the ENCODE database were obtained from human cell lines, coordinates were transformed to mm9 coordinates using liftOver. BED files were indexed with Tabix and used to find overlaps with selected regions. We then used the one-tailed Fisher test to look for features that are over-represented in either included or skipped exons compared with those with no significant change. RBPs binding to specific regions showing significant enrichment (p-value < 0.01) in any of the groups of exons analyzed were subsequently used in a multiple regression analysis using a GLM with logit link function and corrected for co-linearities in binding profiles.

3.1.7 Analysis of interactions between pairs of RBPs binding sites

RBPs and regions selected for the regression analysis were subsequently tested for pairwise regulatory interactions or synergistic effects i.e. whether co-binding of a pair of RBPs had an effect different from the sum of the individual effects as if they were independent. To do so, we extended the set of independent variables by adding all possible pairwise combinations RBPs pairs. We then used L1 regularized logistic regression using scikit-learn [245] to predict belonging to a certain group (Included or Skipped for each comparison under study: ED, PD, TAC, MI). To select the optimal regularizing constant we first performed 10-fold cross validation, by splitting the total number of exons in two sets, one to fit the model and one to evaluate its predictive power, over a range of regularizing constants (from 10^{-3} to 10^{3}). We used the AUROC to evaluate the predictive power for each model fitting and select the value of the regularizing constant that provided better predictions on unseen data. This value was used to fit a L1-logistic regression models will the full dataset. Coefficient values were extracted for further analysis and classified in two types according to whether they represented single RNA binding protein (RBP)s or combinations of them (also named as interactions).

3.1.8 Analysis of correlation among RBPs expression levels

To calculate condition-specific gene expression correlations among pairs of RBPs, we selected the samples involved in each comparison and calculated the Pearson coefficient on the log transformation of the nor-

malized counts. Since the number and nature of samples included in each comparison is different, they were not directly comparable: correlation coefficients estimated from a reduced number of samples are expected to be noisier. We then took samples of 5000 pairs of randomly selected genes to estimate the expected mean and standard deviation of randomly selected genes, which we used to standardize correlation coefficients among RBPs into z-scores. Therefore, we evaluated whether RBPs were more correlated than expected from random genes. As a positive control, we calculated the z-score of the correlation for genes encoding proteins that are known to physically interact from Intact database previously used.

3.2 dSreg: A Bayesian model to integrate changes in AS and RBP activity

3.2.1 dSreg: a mechanistic probability model for differential splicing

dSreg models the AS changes between two different conditions, a and b, as a function of changes in the activity of a few of the existing RBPs acting through their known binding sites. Given K AS events detected across N samples, we observe $I_{k,i}$ reads supporting exon inclusion out of a total of $T_{k,i}$ reads mapping to the k^{th} exon skipping event in sample i, which depends on the unknown probability of inclusion $\Psi_{k,i}$. The conditional probability of observing $I_{k,i}$ reads given $T_{k,i}$ and $\Psi_{k,i}$ is given by the binomial distribution.

$$p(I_{k,i} \mid T_{k,i}, \Psi_{k,i}) = Binomial(I_{k,i} \mid T_{k,i},\ \Psi_{k,i}) \tag{3.1}$$

$\Psi_{k,i}$ is therefore different for each sample i, but depends on the condition or group to which it belongs. Since probabilities are bound between 0 and 1, to model this dependency, we take the logit transformation $X_{k,i}$,

$$X_{k,i} = log\left(\frac{\Psi_{k,i}}{1-\Psi_{k,i}}\right) \tag{3.2}$$

We assume that $X_{k,i}$ is drawn from a normal distribution with a common standard deviation σ_k and different means per condition: α_k for condition a; and $\alpha_k + \beta_k$ for condition b, such that β_k represents the difference between the two conditions. For simplicity, we assume here that the standard deviation is the same across all K AS events ($\sigma_k = \sigma$) .

$$p(X_{k,i} \mid D_i, \alpha_k, \beta_k, \sigma) = Normal\left(X_{k,i} \mid \alpha_k + D_i\beta_{k,i},\ \sigma\right) \tag{3.3}$$

where D_i is a constant that takes the value 1 when the sample belongs to condition b, and 0 when it belongs to condition a:

$$D_i = \begin{cases} 1 & \text{if sample } i \text{ in group } a \\ 0 & \text{if sample } i \text{ in group } b \end{cases}$$

So far, this model is a simple logistic regression for each event with the only assumption that the within group variance is common across events and conditions. However, the changes in the probability of inclusion of exon k between two conditions, indirectly modeled by β_k, should depend on the change in the activity θ_j of a particular regulatory RBP j and on whether it can bind to exon k. The binding information is encoded in a matrix $\mathbf{S}_{K\times J}$, with value 1 whenever the RBP j binds to the exon k and 0 otherwise. Position dependent effects can be easily included by considering RBP j binding to different relative locations as different and independent RBPs. At the same time, the matrix $\mathbf{S}$ could also contain continuous values such as the probabilities of binding, affinities or scores given by Position Weighted

Matrices (PWMs) [225] or any other predictive tool [187, 6].

$$S_{k,j} = \begin{cases} 1 & \text{if the combination of RBP-region } j \text{ is present in event } k \\ 0 & \text{otherwise} \end{cases}$$

Now we can model β_k, the change in the logit-transformed inclusion rate of exon k, as a normal distribution centered at a linear combination of regulatory effects $\vec{\theta}$ and $\vec{S_k}$ (the binding profile of exon k) with a given standard deviation ν. Adding variance ν to the distribution of β_k some changes in AS not to be explained by the regulatory features included in the model.

$$p(\beta_k|\vec{\theta}, \vec{S_k}, \nu) = Normal\left(\beta_k \mid \sum_{j=0}^{j=J} S_{k,j}\theta_j,\ \nu\right) \tag{3.4}$$

In this type of exploratory analysis, large numbers of regulatory proteins are usually tested. However, we expect that AS changes are driven by only a few RBPs. We formalize this prior belief setting a horseshoe prior for the change in the activity of regulator j θ_j[46]. The horseshoe prior, a member of the family of hierarchical shrinkage priors, specifies a normal prior for θ_j with mean 0 and a standard deviation τ_j, where τ_j is not a fixed value, but drawn from a common half Cauchy distribution with mean 0 and ρ standard deviation. τ_j represents a local shrinkage parameter, as it only affects protein j, whereas ρ can be understood as a global shrinkage parameter. We further set a half Cauchy prior in ρ with mean 0 and standard deviation 1 as recommended [46]. Note that this prior can be adapted according to the expected number of non-zero parameters [219].

$$p(\theta_j \mid \tau_j) = Normal(\theta_j \mid 0,\ \tau_j) \tag{3.5}$$

$$p(\tau_j \mid \rho) = Cauchy^{+}(\tau_j \mid 0,\ \rho) \tag{3.6}$$

$$p(\rho) = Cauchy^{+}(\rho \mid 0,\ 1) \tag{3.7}$$

Finally, we need to specify prior distributions for the remaining parameters α_k and σ. Since we expect most of the exons to be included most of the times ($\Psi \sim 1$) and α_k is the logit transformation of the inclusion rate in condition a, we set a normal prior centered at 3 (which reflects an expected $\Psi = 0.95$), with standard deviation 3 for each exon k to enable some deviation from this expectation. Moreover, as we expect little variation among samples, we set a half Cauchy prior distribution with 0 mean and standard deviation 1 on σ.

$$p(\alpha_k) = Normal(\alpha \mid 3,\ 3) \tag{3.8}$$

$$p(\sigma) = Cauchy^{+}(\sigma \mid 0,\ 1) \tag{3.9}$$

The joint posterior probability of the parameters Θ given the data (I) is proportional to the joint probability distribution of the data and Θ, since the marginal probability of obtaining the data $p(\mathbf{I})$ is constant for any Θ.

$$p(\Theta|\mathbf{I}) = \frac{p(\Theta, \mathbf{I})}{p(\mathbf{I})} \propto p(\Theta, \mathbf{I}) \tag{3.10}$$

Using the conditional probabilities and prior distributions that we have defined for each variable, we

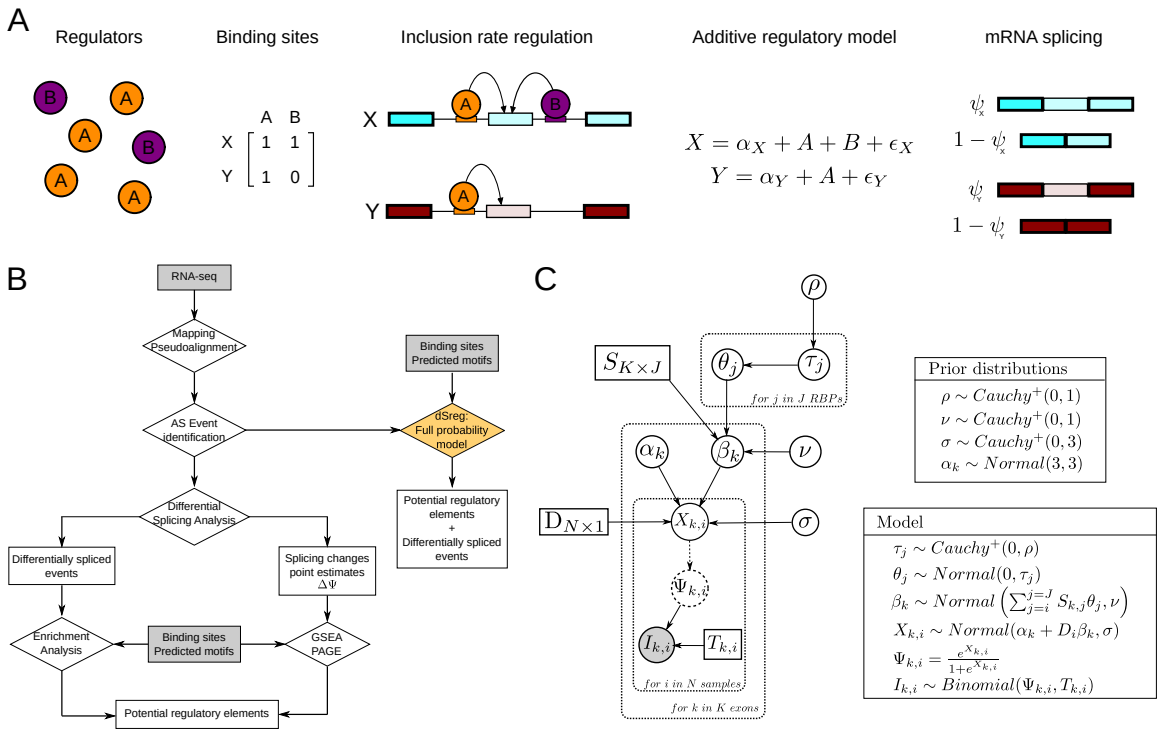

Figure 3.1: General and proposed work-flows for AS regulation analysis. **A** Schematic representation of idealized model for the regulation of splicing rates by RBPs through direct binding to their binding motifs in the pre-mRNA **B**. Diagram representing the different steps required for a classical analysis of regulation of alternative splicing using RNA-seq data and the proposed model in dSreg. **C**. Directed Acyclic Graph (DAG) representing the full probabilistic model integrating both differential AS analysis with binding sites presence and changes in the activity of RBPs.

can calculate this joint probability distribution applying the chain rule.

$$p(\Theta, \mathbf{I}) = p(\mathbf{I},\ \mathbf{T},\ X, \alpha, \beta, \nu, \theta, \tau, \rho, D, S) = \tag{3.11}$$

$$= p(\Theta, \mathbf{I}) = p(\sigma)p(\nu)p(\rho)\prod_{j}^{J}[p(\theta_j|\tau_j)p(\tau_k|\rho)]\prod_{k}^{K}[p(\beta_k|\mathbf{S}, \theta, \nu)p(\alpha_k)P(I_k)]$$

where,

$$P(I_k) = \prod_{i}^{N}(p(I_{k,i}|T_{k,i}, X_{k,i})p(X_{k,i}|\alpha_k, \beta_k, \sigma, D_i)) \tag{3.12}$$

Once the full posterior distribution is completely specified, it can be explored using Markov Chain Monte Carlo (MCMC) algorithms. We implemented this model in stan [45], using a non-centered parametrization whenever possible to alleviate sampling difficulties from hierarchical models [30]. The full model is represented as a Directed Acyclic Graph (DAG) to show dependencies among parameters in Fig. 3.1B.

3.2.2 Data simulation

Data can be simulated by setting fixed values on the parent nodes of the DAG (σ, $\vec{\alpha}$, $\vec{\theta}$, $\mathbf{T}$ and $\mathbf{S}$) representing the probabilistic model (Fig. 1C) and drawing samples from the corresponding distributions for each parameter. We simulated 20 datasets for each initial set of conditions, all with K=2000 events,

3 samples per condition (N=6) and J=50 potential regulatory elements with correlated binding profiles, of which only 5 showed non-zero effects on splicing changes between the two conditions.

To simulate realistic values of inclusion rates for the condition a ($\Psi_{k,a}$) across the K=2000 exons, we assumed that 20% of the exons are alternative, with inclusion rates following a uniform distribution between 0 and 1; and 80% are consitutive, with inclusion rates drawn from a Beta(10, 1), to promote generally high inclusion rates.

$$u_k \sim Uniform(0,1) \tag{3.13}$$

$$\Psi_{k,a} \sim \begin{cases} Beta(10,1)\ if\ u_k > 0.2 \\ Uniform(0,1)\ if\ u_k < 0.2 \end{cases} \tag{3.14}$$

$$\alpha_k = logit(\Psi_{k,a}) = log\left(\frac{\Psi_{k,a}}{1-\Psi_{k,a}}\right) \tag{3.15}$$

We aimed to simulate matrices of correlated binding profiles to take into account that certain groups of RBPs often bind to similar regions in the exons. To do so, we first simulated a covariance matrix Σ of size J sampling from an inverse Wishart distribution,

$$\Sigma_{J\times J} \sim InvWishart\left(J+1, \frac{1}{J} I_J\right) \tag{3.16}$$

and used it to simulate K samples from a multivariate normal distribution using a mean of -2.5. This value represents an expected 7.5% of events bound by a particular RBP.

$$\vec{M_k} \sim MvNormal(-2.5, \Sigma) \tag{3.17}$$

Then, we took the inverse logit to transform $\mathbf{M}$ matrix into the probability matrix $\mathbf{T}$ and use these probabilities to simulate binary binding profiles across exons ($\mathbf{S}_{K\times J}$ matrix) by sampling from a Bernoulli distribution for each element in the $\mathbf{T}_{K\times J}$ matrix.

$$P_{k,j} = InvLogit(M_{k,j}) = \frac{e^{M_{k,j}}}{1+e^{M_{k,j}}} \tag{3.18}$$

$$S_{k,j} \sim Bernoulli(P_{k,j}) \tag{3.19}$$

We randomly drew a set $A = \{A_1, A_2, A_3, A_4, A_5\}$ of 5 active regulatory proteins (with non-zero effects on changes in the inclusion rates) from the whole set of regulatory proteins $R = \{1, 2, ..., J\}$. The regulatory effect for RBP j θ_j was then drawn from a uniform distribution between -2.5 and 2.5 if j belonged to the set of active regulatory elements A and set to zero otherwise. These values of θ_j represent the mean increase in the log(odds ratio) of exons having a binding site for that protein compared with those without a binding site.

$$\theta_j \sim \begin{cases} Uniform(-2.5, 2.5)\ if\ j \in A \\ 0\ otherwise \end{cases} \tag{3.20}$$

Once the parent nodes of the DAG were simulated, we could easily simulate the final data by sampling parameter values along the graph according to our model. First, we drew changes in the logit-transformed inclusion rates β_k from a normal distribution with mean obtained from a linear combination of effects $\vec{\theta}$ and binding sites $\vec{S_k}$ and standard deviation $\nu = 0.1$. This way we introduced noise with small random

changes in inclusion rates of exons that were not targets of any of the differentially active RBP.

$$\beta_k \sim Normal\left(\sum_{j=1}^{j=J} S_{k,j}\theta_j,\ \nu\right) \tag{3.21}$$

We then combined α_k and β_k to obtain the mean $logit(\Psi)$ for condition b, and sample 3 samples from each mean using $\sigma = 0.2$ to introduce some inter-individual variability. Being D_i a variable that takes value 1 when sample i belongs to condition b and 0 otherwise,

$$X_{k,i} \sim Normal(\alpha_k + D_{k,i}\beta_k, \sigma) \tag{3.22}$$

The total number of reads mapping to each event $T_{k,i}$ were drawn from a Poisson distribution with $log(\lambda) = 2$ by default,

$$T_{k,i} \sim Poisson(\lambda) \tag{3.23}$$

They were subsequently used to sample the corresponding reads supporting inclusion $I_{k,i}$ from the binomial distribution with $p = \Psi_{k,i}$, obtained from the inverse logit transformation of $X_{k,i}$.

$$I_{k,i} \sim Binomial\left(T_{k,i}, InvLogit(X_{k,i})\right) \tag{3.24}$$

Using these default parameter values, we additionally simulated data for increasing sequencing depths (from $log(\lambda) = 1$ to $log(\lambda) = 5.5$) and with an increasing number of total regulatory proteins (from J=50 to J=250), maintaining a total of 5 differentially active RBPs to evaluate the effect of this variables on the methods performance.

3.2.3 Bayesian inference

The probabilistic models were implemented in Stan [45] using non-centered parametrization, whenever it was possible, to improve sampling efficiency [30]. The joint posterior distributions of the parameters were approximated using No-U Turn Sampler (NUTS) as implemented in Stan [106], running 4 chains along 4000 iterations, being 2000 of them for warming up. Convergence of the Markov Chain Monte Carlo (MCMC) algorithm was checked in each case by means of the split Gelman-Rubin R ($\hat{R}$) [86].

3.2.4 Differential splicing analysis

In order to identify exons with significant changes in inclusion rates, a GLM with binomial likelihood was used to model the probability of inclusion of a particular exon using the sample condition D_i as only predictor. After fitting the model, we extracted the estimate and p-value for the coefficient representing the condition of interest. We then obtained adjusted p-values by means of Benjamini-Hochberg (BH) multiple test correction.

3.2.5 Over-representation analysis

We tested over-representation of binding sites for a particular RBP on the set of significantly changed exons using a Generalized Linear Model (GLM) with binomial likelihood to model the probability of being significantly changed as a function of the presence of a binding site for a particular RBP. We then extracted the p-value for the coefficient for each RBP and applied BH multiple test correction.

3.2.6 Gene set enrichment analysis

We implemented an in-house algorithm for GSEA in python following [264]. We sorted exons according to the estimated coefficient representing log-transformation of change in exon inclusion odds between the two conditions under study. We then used the matrix with binding sites for each exon and RBP and subtracted the mean for each column. This way, we give weight to each binding site depending on the number of binding sites present for a particular RBP. We then calculated the cumulative sum and took the maximum and minimum values as enrichment scores. We permuted 10000 times the list of exons to calculate a null distribution of enrichment scores, estimated p-values as the proportion of permutations with bigger enrichment scores and performed BH multiple test correction.

3.2.7 Regulatory features: CLiP derived RBPs binding sites

CLiP-seq binding sites were collected from several databases and merged in a single BED file [35, 158, 301, 64]. Human binding sites and mouse binding sites in mm10 were transformed to mm9 coordinates using liftOver tool for compatibility with vast-tools. For simplicity, only binding sites mapping to the 250bp upstream or downstream the alternative exons were included in the analyses.

3.2.8 Bench-marking of differential splicing methods using real data

In order to assess the performance of dSreg in real biological data, we used the GSE112037 dataset, which contained an independent quantification of exon inclusion rates using RASL-seq for the quantitative evaluation of the performance of different methods [315]. We also evaluated the impact of sequencing depth on the performance of the different methods by serial down-sampling of sequencing up to 1/512 times the original depths (120M reads). dSreg was run using processed event counts as provided by DARTS, which is itself based on rMATS [249, 315]. GLM analysis was also performed using the same event counts. MISO and BRIE were run using their own event annotation, corresponding to hg19 genome version and Ensembl annotation release 75 for all methods [128, 107]. An additional $Nullmodel$ for dSreg without regulatory information, as in the simulations, was run to test the improvement in detection of splicing changes by including regulatory features. For evaluation, we selected events with at least 50 total reads in the RASL-seq experiment, and calculated the real inclusion rates as the proportion of reads supporting exon inclusion. Real AS changes were defined as those with a $|\Delta\Psi| > 0.05$ and $FDR < 0.05$ using a basic GLM in R. Then, performance was evaluated by comparing the estimation of the $\Delta\Psi$ in the down-sampled RNA-seq experiments and the ones derived from RASL-seq. We assessed the quantitative estimation of inclusion rates by calculating the Pearson coefficient with the real $\Delta\Psi$. AUROC was used to asses the ability to identify differentially spliced. The scoring function for AUROC calculation were: i) Bayes Factors for BRIE and MISO: 1 - FDR for GLM; ii) and $P(|\Delta\Psi| > 0.05|data)$ for DARTS and dSreg; and iii) MISO and BRIE were evaluated using only the subset of events that were also represented in the RASL-seq experiments.

3.2.9 Assessment of the ability of dSreg to identify AS regulatory drivers using ENCODE knock-down experiments

In order to evaluate the performance of dSreg in detecting the RBPs that drive AS changes between two conditions, we used the data from systematic knock-down experiments of 206 RBPs in two different human cell lines from the ENCODE project and their corresponding binding profiles [203, 64]. We downloaded the rMATS processed files available from the website and analyzed their regulatory patterns using GLM+ORA, GLM+GSEA and dSreg. Regulators were defined as differentially active if FDR<0.05

for both ORA and GSEA; or if the posterior probability of the θ_j being different from 0 was higher than 95% ($P(|\theta_j| > 0|data) > 0.95$) for dSreg. The performance was evaluated with 3 different measures. First, we analyzed the number of times the RBP that was down-regulated was found among the driver regulators of Alternative splicing (AS). This measure was normalized by the expected proportion of matches if the regulatory elements were selected randomly from the set of available regulators. Second, we measured the proportion of RBPs defined as differentially active were differentially expressed in in the knock-down experiment. Third, we sorted by absolute differential activity (or FDR for ORA and GSEA) the RBPs and calculated of RBP that was knocked-down.

3.2.10 Real data analysis

GSE59383 fastq data were downloaded and mapped using vast-tools 0.2.0 [114] to identify AS events. We restricted our analysis to exon cassette events that showed at least 1 inclusion and skipping read in at least one sample. Once extracted the number of inclusion and total counts for each event and sample, we used all the methods described here (ORA, GSEA and dSreg) to analyze regulatory patterns using a compendium of CLiP-seq binding sites.

3.3 Quantitative evolution of exon inclusion rates in mammals

Data

We used the European Nucleotide Archive browser to look for Paired-end RNA-seq data from livers of mammalian species in adult stage without any particular treatment, reaching a total of 132 samples from 76 different mammalian species with known phylogenetic relationships [282] (see Figure 4.19) and characterized gene orthologs in OrthoMaM.v10b database [244]. Liver was selected because of the abundance of RNA-seq data across different species and its relatively low cell type complexity, as most of the liver is formed by hepatocytes. This way, we should limit the influence of changes in cell populations on the overall tissue AS patterns. Sample information can be found in Table S2. Samples from species missing in OrthoMaM or unavailable genomic references were mapped against the closest genome, as indicated in Table S2.

Reference genomes, annotation and indexes

Reference genomes, transcriptomes, and annotations were retrieved from Ensembl and GenBank, as shown in Table S3 in detail for each species. Transcriptome fasta files were indexed with kallisto v0.43.0, using a k-mer size of 27. To build genome indexes with inserted Splice Junctions (SJ) for mapping and estimating AS events, we tried to correct at least some of the differential annotation biases by first extracting all possible Splice Junctions (SJ) across the longest transcript for each of the genes, reaching about ∼2 million SJ per genome. These SJ were used to build a genome index with STAR v2.6.1 [63]. Thus, we mainly assume that gene structure is well annotated and that main annotation biases between species will be due to differential transcript and isoform coverage. Moreover, this way, we focus on studying the Ψ of already annotated exons in a quantitative manner, rather than on a complete description and enumeration of all AS events taking place in each of the species.

3.3.1 Counting reads supporting exon inclusion and skipping

Sequencing data were mapped with STARv2.6.1b [63] on the genome index built with customized sets of SJs. Then, for each exon, we considered as inclusion reads all those mapping to annotated SJs going

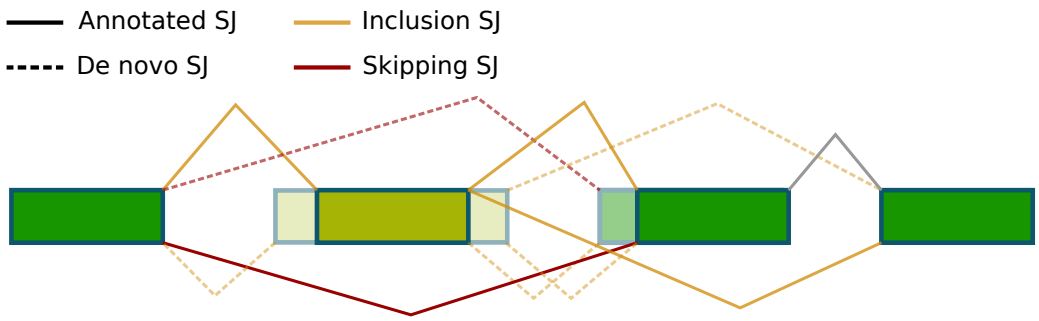

Figure 3.2: Schematic representation of the process to extend the sets of SJs supporting exon inclusion or skipping overlapping with alternative splice site selection events AD, AA.

from or into the exon boundaries or the exon splice sites; and as skipping reads, all those mapping SJs that join pairs of upstream and downstream splice sites.

Moreover, there may be simultaneous alternative splice site selection events (AD and AA)) and ES events. If we only count the reads mapping to the predefined exon boundaries in the annotation, we are missing some of the reads supporting inclusion, obtaining biased Ψ estimates. Thus, we assume that the alternative splice site selection has no or only very minor impact on the function of the exon, e.g. for instance NAGNAG events only modify 1 or 2 aminoacids of the protein sequence. In other words, there may be equally valid splice sites within a certain neighborhood for the inclusion of a particular exon. Thus, we tackle the issue of co-existence of these AS types by expanding sets of SJs supporting inclusion and skipping. As inclusion SJs, we also considered SJs located within a window of 50bp of each splice site. Skipping SJs were expanded to included any SJ mapping within the gene boundaries and going over the target exon (Figure 3.2).

3.3.2 Systematic biases in estimation of exon inclusion rates

Even if we were able to fully identify the SJs that characterize each AS event and provide information about the relative splicing rates Ψs, there are known technical biases that modify our ability to obtain or detect reads from a particular region in the transcriptome, often introduced during the preparation of the library e.g. fragmentation bias due to selection of cDNA fragment of certain sizes, GC and fragment length amplification biases; or by the sequencing conditions or mapping procedure. Whereas the most widely used tools for estimating GE explicitly model some of these biases [215, 39], most methods for AS do not take some of them into account [248, 249, 114, 128]. In small scale experiments in which there is a very homogeneous processing and handling of the samples in a single species, these biases may be rather similar across all samples and thus, likely not to affect relative differences. However, biases that depend at the same time on the genome properties, such as GC-dependent amplification biases or fragment length selection, become worrying in a comparative study. As these genomic properties diverge, even if library preparation was similar across samples, these differential properties would introduce artificial differences between species, and hinder the study of the evolution of the quantitative trait of interest. If, in addition, library preparation protocols are not homogeneous across samples included in the study, as happens when collecting data from different comparative datasets, technical biases may even amplify differences between species and thus hinder the interpretation of the results.

Fortunately, as we know how some of these biases arise, we can try to correct them by taking into account the systematic deviations that they introduce from the real value. In other words, we need to define a function that specifies the the expected observations from a given real value. In this specific case, we need to define the expected number of reads from each SJ to correct Ψ estimations.

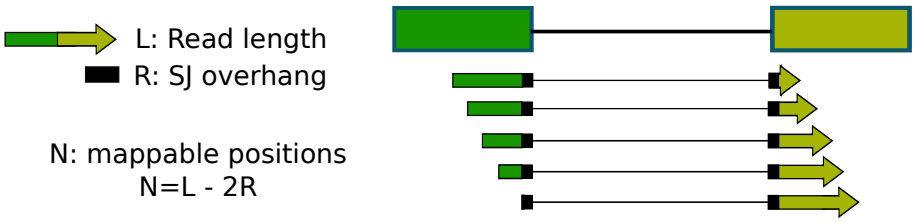

Figure 3.3: Influence of read length and overhang on the number of mappable positions spanning a single SJ

Read length

As we are going to base our analysis only on the reads that span the exon-exon junctions, one of the first issues that need to be taken into account are the requirements for mapping reads across the SJ and count them. Thus, we need to know the expected number of reads that can be generated from each SJ. We assume that the total number of reads spanning a SJ is proportional to the number of mappable positions across that SJ, which depends mainly on 2 parameters:

- Read length (L). Naturally, the number of different positions a read can map spanning the SJ cannot exceed the length of the sequencing read, which must be located within the read.
- Minimum read overhang (R). It is the number of bases required for counting a read as spanning the SJ. This parameter can be manually set in mapping tools like STAR [63], and is key to filter out spuriously mapping reads. It is well known that mapping errors are more common when the read overhang is small: if we require as low as a single base to allow mapping through an SJ, there is about 1/4 probability to have a match just by chance. However, as the read overhang increases, the probability of random matching decreases exponentially (4^{-R}). Sequencing errors, which accumulate at the end of the reads, nucleotide variants or just finding this sequence somewhere else in the genome can increase the mapping error at positions with small R. We could try to model this phenomenon explicitly allowing some error rate in these positions, to extract as much information as possible from the data [84, 183], but it is simpler to just count only reads with a relatively large minimum read overhang e.g. 8 nucleotides. Thus, the number of different reads N_{SJ} that can align to a SJ is

$$N_{SJ} = L - 2R \tag{3.25}$$

This effect is, in principle, independent the particular SJ. Thus, although it affects the absolute number of reads that we will map to each SJ, as it is expected to affect all of them equally, the relative number of reads supporting exon inclusion or skipping should remain mostly unaffected.

Transcript structure

However, this is not always the case, specially as the read length L becomes larger. If the SJ is close to the transcript start or end, then the possible number of reads that span the SJ may be reduced. Assuming that transcripts are always longer than the reads, we can only observe this effect on one side of the SJ. Thus, the number of mappable positions N will be limited by the read length L, except if the distance to either the end d_e or start d_s of the transcript is shorter than L:

$$N_{SJ} = min(L - R, d_s, d_e) - R \tag{3.26}$$

Thus, for most cases $L >> d_s$ and $L >> d_e$, such that $N = L - 2R$, as previously derived. However, if, for instance, $d_e = 50$, and $L = 100$, we would only have at most, 50 mappable positions, to which we

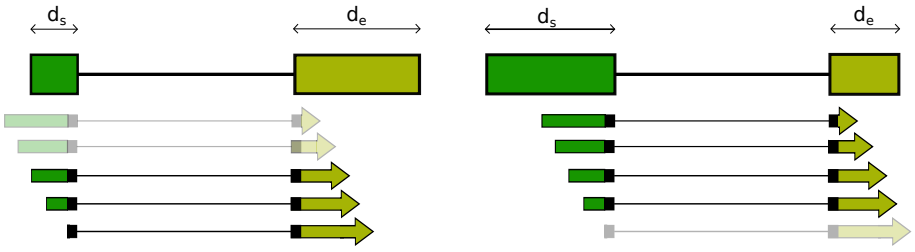

Figure 3.4: Influence of read length and overhang depending on the distance to the transcript start and end

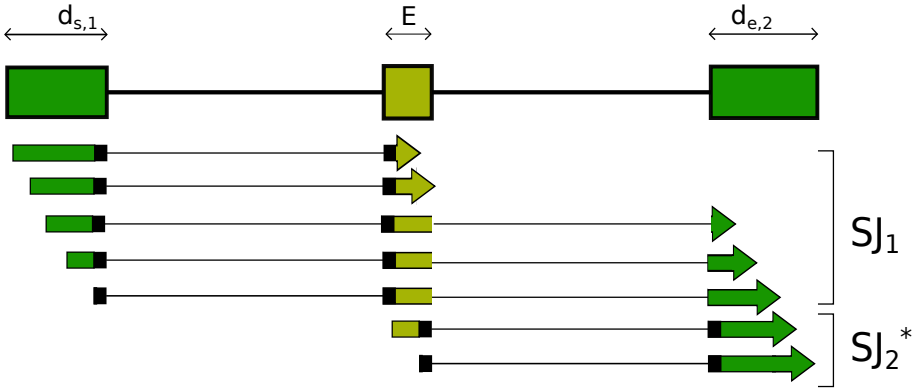

Figure 3.5: Influence of read length and overhang in depending on the distance to the transcript start and end and exon size and exon length

need to remove the R positions that are not taken into account (Figure 3.4).

Taking into account that we are interested in studying exons rather than isolated SJ, a similar scenario arises depending on exon length, as some reads may overlap the two SJ at the same time. If we use the SJ counts as directly provided by STAR, we should take into account that we are counting some reads more than once, as if the SJ were independent. This is simpler, as we can directly use the independent counts from each SJ and does not bias the estimations, but it does underestimate the uncertainty.

Exon size

However, the best approximation is to directly count each read only once and calculate the number of mappable positions as a function of the exon length. For this, we can adjust the N for the second SJ by counting only the reads starting in the same exon. Formally, this can be done using the exon length E as distance to the d_s. Thus, the number of mappable position for an exon of size E is

$$N_{exon} = N_{SJ_1} + N_{SJ_2^*} = min(L - R, d_{s,1}, E + d_{e,2}) + min(L - R, E, d_{e,2}) - 2R \tag{3.27}$$

Fragment size distribution across a single SJ

Sequencing reads, however, are only at the ends of larger nucleotide fragments derived from the process of fragmentation and size selection during the library preparation protocol. Again, depending on the relative position of an SJ in the transcript, the selection of fragments with certain size may affect different SJs to different extents (Figure 3.6).

If we assume that fragmentation is a random process, depending on the nature of the fragmentation process, we may expect different size distributions e.g. a completely uniform fragmentation leads to a Weibull distribution of fragment sizes given the the starting fragment size [270]. Thus, if the different

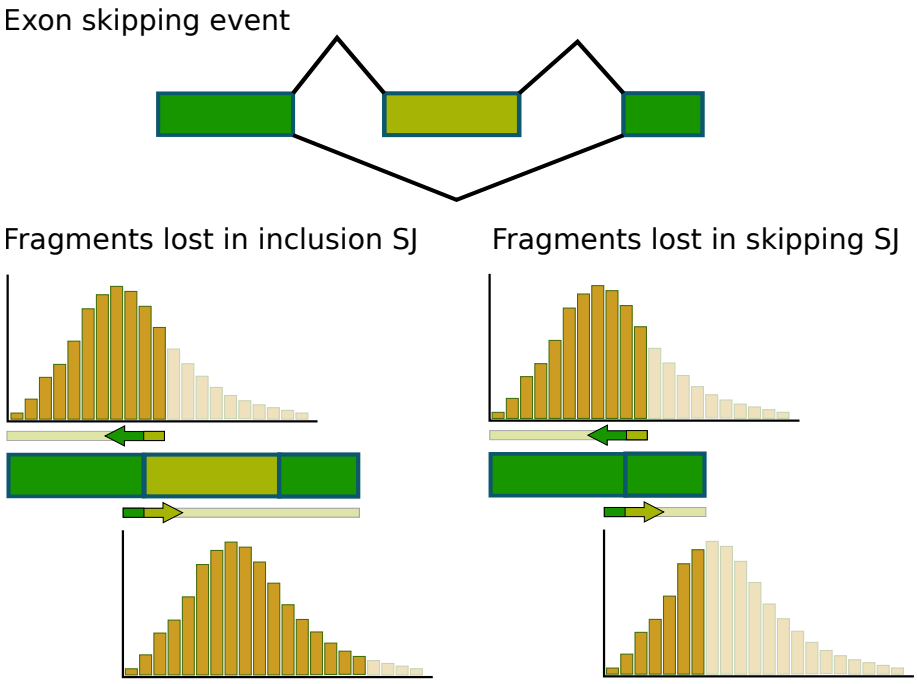

Figure 3.6: Selection of fragment sizes influence the proportion of observable fragments with reads overlapping SJs resulting in biased estimations of exon Ψ. In this case, given the gene structure, more fragments with reads mapping to the skipping SJ are lost during fragment size selection, introducing an apparent increase in the exon Ψ.

positions along the transcript have the same probability of breaking, then we would not expect each position to produce a different fragment size distribution. This fact is important, as it means that by performing size selection of the resulting fragments, we are not enriching in particular breaking points. In other words, the relative representation of the different positions along the transcript remains unchanged after fragment size selection. More importantly, this allows us to calculate the proportion of fragments that can be generated from each position depending on their size.

The correction that we need to make follows the same logic as for the read length: if the selected fragments are longer that the ones we need for their ends to map across the SJ, then we will not be able to recover those positions i.e. they are not mappable. However, in contrast to read length, which is fixed in the sequencing conditions, we have a distribution of fragment sizes. Thus, in probability terms, we previously had a probability 0 or 1 of observing a read from a position at a certain distance i of the SJ depending on read length and overhang.

$$p(obs|L,R,i,d_e,d_s) = \begin{cases} 1 & \text{if } (d_s+i) > L \text{ or } (d_e-i) > L \\ 0 & \text{otherwise} \end{cases}$$

For fragments, however, as the length is not fixed, but follows a certain probability distribution, we now have to calculate the probability of obtaining a read at each distance i from the SJ, which depends on the fragment size distribution $p(f)$. Le ts assume that we know this fragment size distribution, either from the library preparation stage, or the empirical distribution derived from mapping PE into long exons. Now, a read spanning a SJ can actually come from two different types of fragments: when the read is at the beginning or at the end of the fragment (relative to the transcript in 5' to 3' sense). Reads mapping at distance i from the SJ may come from two types of fragments, depending on which end of the fragment they are located, with different observation probabilities:

- For reads located at the beginning of the fragments, we will observe only fragments that are shorter

than the distance from position i to the end of the transcript

$$p(obs|start, d_e, i) = p(f < d_e - i)$$

- For reads located at the end of the fragments, we will observe only those fragments that are shorter than the distance from the SJ to the transcript start d_s

$$p(obs|end, d_s, i) = p(f < d_s + i)$$

Generally, we can assume that we have an equal probability of sequencing each end of a fragment, i.e. $p(start) = p(end) = 0.5$, which allow us to calculate the probability of observing a particular read at distance i from the SJ given the fragment size distribution $p(f)$

$$p(obs|d_e, d_s, i) = p(obs|start, d_e, i)p(start) + p(obs|end, d_s, i)p(end) = \frac{p(f < d_e - i) + p(f < d_s + i)}{2} \tag{3.28}$$

Thus, if we want to calculate the full probability of observing a read across the whole SJ, independently of the distance i, we can apply again the *Total probability theorem* to sum the probabilities across all possible distances from the SJ, as previously derived (3.26). If we assume that each distance i has the same probability of being sampled: $p(i) = \frac{1}{N_{SJ}}$:

$$p(obs|d_e, d_s) = \sum_{i=0}^{N_{SJ}} p(obs|d_e, d_s, i)p(i) = \frac{1}{N_{SJ}} \sum_{i=0}^{N_{SJ}} \frac{p(f < d_e - i) + p(f < d_s + i)}{2} \tag{3.29}$$

Marginalizing across fragment lengths

Alternatively, we can reformulate this probability as a function of fragment lengths:

$$p(obs|d_e, d_s) = \sum_{f=L}^{\infty} p(obs|d_e, d_s, f)p(f) \tag{3.30}$$

Now, instead of calculating the probability of observing reads starting at a certain distance from the SJ i, we need to calculate the probability of observing reads given a specific fragment size f, which corresponds to the proportion of mappable positions across the SJ that can generate a fragment of size f, taking into account the equi-probable relative location of a read in a fragment.

$$p(obs|d_e, d_s, f) = p(obs|d_e, d_s, f, start)p(start) + p(obs|d_e, d_s, f, end)p(end)$$

We can decompose these probabilities by summing over all possible distances from the read to the SJ i, which go from R to $R + N_{SJ}$.

$$p(obs|d_e, d_s, f, start) = \sum_{i=R}^{R+N_{SJ}} p(obs|d_e, d_s, f, start, i)p(i)$$

$$p(obs|d_e, d_s, f, end) = \sum_{i=R}^{R+N_{SJ}} p(obs|d_e, d_s, f, end, i)p(i)$$

for which we actually know whether a fragment of size f can or can not be generated given the specific conditions:

$$p(obs|d_e, d_s, f, start, i) = \begin{cases} 1 & \text{if } (d_e - i) > f \\ 0 & \text{otherwise} \end{cases}$$

$$p(obs|d_e, d_s, f, end, i) = \begin{cases} 1 & \text{if } (d_s + i) > f \\ 0 & \text{otherwise} \end{cases}$$

Thus, the number of positions that follow these conditions in each case can be directly counted depending on f, d_e, d_s, being limited by the number of mappable positions N_{SJ} and the number of positions that allow a fragment of size f with certain read length L and overhang R, given by $d_e - i$ and $d_s - i$ for reads at the beginning and end of the fragment, respectively. Moreover, again, we can assume that the probability of sampling any possible distance from the SJ is the same ($p(i) = \frac{1}{N_{SJ}}$).

$$p(obs|d_e, d_s, L, R, f, start) = \frac{max(min(N_{SJ}, d_e + L - R - f), 0)}{N_{SJ}}$$

$$p(obs|d_e, d_s, L, R, f, end) = \frac{max(min(N_{SJ}, d_s + L - R - f), 0)}{N_{SJ}}$$

$$p(obs|d_e, d_s, L, R, f) = \frac{max(min(N_{SJ}, d_e + L - R - f), 0) + max(min(N_{SJ}, d_s + L - R - f), 0)}{2N_{SJ}} \tag{3.31}$$

In practice, we can set a maximal value for the fragment length f_{max}, e.g. the maximal observed value or the 99% percentile, to easily approximate $p(obs|d_e, d_s)$ and reduce computational cost.

Fragment size distribution across an exon

Again, if we have exons that are shorter than the read length, we need to re-adjust the probability of observing a read derived from a fragment of particular size f $p(obs|d_e, d_s, f)$ to avoid double counting the same reads that map across the 2 SJ supporting exon inclusion. We obtain a modified number of mappable positions across SJ_2, which is limited by the exon length E, since reads mapping beyond the exon will have been counted already in SJ_1:

$$N^*_{SJ_2} = min(E, N_{SJ_2}) \tag{3.32}$$

Thus, for a given fragment size f, the probability of observing reads supporting exon inclusion depends also on exon length E.

$$p(obs|d_{s,1}, d_{e,1}, E, d_{s,2}, d_{e,2}, f) = \sum_{i=R}^{R+N_{SJ_1}} p(obs|d_{s,1}, d_{e,1}, f, i)p(i) + \sum_{i=R}^{R+N^*_{SJ_2}} p(obs|d_{s,2}, d_{e,2}, f, i)p(i) \tag{3.33}$$

where again we assume a common probability for observing reads across all mappable positions across the 2 SJ.

$$p(i) = \frac{1}{N_{SJ_1} + N^*_{SJ_2}} = \frac{1}{N_{SJ_1} + min(E, N_{SJ_2})}$$

Fragment size distribution across an exon with multiple SJs

In the previous section, we have assumed that there is a single splicing path giving rise to the inclusion of a particular exon, this is, with a unique combination of upstream and downstream exons. If exon skipping is very rare, then we would not expect to have multiple skipping events taking place simultaneously in the same transcript. However, as exon skipping becomes more frequent in a gene, or as the number of exons per gene increases, so does the chance of having multiple exon skipping events in a single transcript and this potentially different combinations of upstream and downstream exons. Thus, we need to update the

probability of generating reads supporting exon inclusion to account for all these possible combinations of splicing reactions.

Lets assume that for a given exon, there are U possible upstream exons and D possible downstream exons. If we assume that they are independently selected during a splicing reaction with certain probabilities, we can calculate easily the probability of a certain combination of SJ as the product of the probabilities of selecting each upstream donor u and downstream acceptor d:

$$p(u,d) = p(u)p(d) \tag{3.34}$$

Knowing their corresponding distances to transcript start $d_{s,1,u}$, $d_{s,2,u}$, and a common distance to transcript end $d_{e,1,d}$, $d_{e,2,d}$, we can use the expression for a single combination previously derived (equation 3.33) to reconstruct again the probability of observing reads supporting exon inclusion given a certain fragment size f.

$$p(obs|\vec{d_{s,1}}, \vec{d_{e,1}}, E, \vec{d_{s,2}}, \vec{d_{e,2}}, f) = \sum_{u=1}^{U}\sum_{d=1}^{D} p(u,d)p(obs|d_{s,1,u}, d_{e,1,u}, E, d_{s,2,d}, d_{e,2,d}, f) \tag{3.35}$$

However, to directly apply this *a priori*, we would need to know the relative usage of each of the SJ. Ideally, we would perform inference of these basic probabilities using a more complex model taking into account all these biases, and derive the exon Ψ from them. In practice, we can approximate these SJ relative usage ratios assuming a multinomial distribution for reads sampled from each SJ without biases and take the MLE of the parameters of this distribution as $p(u), p(d)$.

$$p(\vec{obs_U}) = Multinomial(\vec{\theta}) \tag{3.36}$$

$$\hat{\theta}_u = \frac{N_u}{\sum_{k=1}^{U} N_k} \tag{3.37}$$

Incorporate known biases to models

We have seen along the last few sections that the probability of observing a sequencing read across a specific SJ depends on a number of parameters i.e. read length, minimum overhang, distance to transcript start and end and fragment size distribution, Thus probability can be calculated under some assumptions. However, our aim is not calculating these technical biases, but estimating the real Ψ taking them into account.

For that, we need a probability formula relating the observations (reads supporting inclusion I out of T total reads) and the underlying real Ψ. We can understand the process of sampling reads supporting exon inclusion as a coin tossing problem, such that the number of reads supporting inclusion out of the total follows a binomial distribution depending on Ψ^*.

$$p(I|T, \Psi^*) \sim Binomial(I|T, \Psi^*) \tag{3.38}$$

This Ψ^* is the probability of observing reads supporting exon inclusion when obtaining reads deriving from this ES event. It can be formulated as the probability of observing reads supporting exon inclusion, divided by the probability of observing any read across the ES event, which includes both inclusion and skipping reads. These probabilities can be decomposed into the probability of observing a read given that an inclusion or skipping event took place, $p(obs|inc)$ and $p(obs|skp)$, and the probability that the

inclusion or skipping event took place, which is given by Ψ, and $1-\Psi$, respectively

$$\Psi^* = \frac{p(obs|inc)p(inc)}{p(obs|inc)p(inc)+p(obs|skp)p(skp)} = \frac{p(obs|inc)\Psi}{p(obs|inc)\Psi+p(obs|skp)(1-\Psi)} \tag{3.39}$$

where the probabilities of observing reads of each type depend on the technical biases as previously calculated for the SJs supporting exon inclusion and skipping:

$$p(obs|inc) = p(obs|d_{e,1}, d_{s,1}, E, d_{e,2}, d_{s,2})$$

$$p(obs|skp) = p(obs|d_e, d_s)$$

In many cases, our aim is not the estimation of Ψs *per se*, but the characterization of other parameters, like the differences between experimental conditions e.g. tissues, developmental times or species, taking into account technical biases. In these cases, we often use the logit transformation to work in a regression framework. If we calculate the logit-transformation of the Ψ^*, we see that it can be decomposed in the sum of the real $logit(\Psi)$ and a bias term. Then, this term can be calculated for each sample and exon and incorporated to the full model to control for differences in both sequencing conditions across different samples and different transcript properties across species.

$$logit(\Psi^*) = log(\frac{\Psi^*}{1-\Psi^*}) = log\frac{p(obs|inc)\Psi}{p(obs|skp)(1-\Psi)} = log\frac{p(obs|inc)}{p(obs|skp)} + log\frac{\Psi}{(1-\Psi)} = logit(bias) + logit(\Psi) \tag{3.40}$$

3.3.3 Exon orthology identification

Markov clustering on best-reciprocal hits graph

As we aim to quantitatively compare the exon Ψ across different species to track how this character has changed over evolutionary time, a key step is the identification of exons that have a common ancestor and therefore are equivalent across species i.e. groups of orthologous exons. Wheareas there are several known tools for identifying gene orthologs e.g. OrthoFinder, InParanoid [68, 258] and systematic evaluations [8], none have been developed for inferring exon orthology relationships to our knowledge.

We assume that the gene orthology relationships are known and only want to identify exon orthologs within them. We propose to use two similar approximations to these methods, that are based on a common framework consisting on 2 main steps (Figure 3.7):

1. Build a best-reciprocal hit graph. For this, all possible pairwise comparisons are performed between exons of different species within groups of orthologous genes, using a certain metric for sequence comparison. One can use a wide variety of metrics for sequence comparison: from very simple and fast editing distance, which just counts the minimal number of changes that are required to change one sequence into the second one; to alignment based methods, which are more computationally expensive, especially if using algorithms for optimal pair-wise alignment like Smith-Waterman or Needleman–Wunsch with a given scoring matrix. Intermediate solutions may involve the use of more heuristic alignment methods like BLAST [9]. Whatever the distance used, we can then build a graph in which nodes represent different exons, which are connected whenever both exons are the best scoring exons in the other species.

2. Clustering on the best-reciprocal hit graph. Thus, we have obtained a graph in which highly similar exons across species will be joined together, forming modules or clusters isolated from each other. Thus, we can use methods for detecting modules or clustering on the graphs to identify putative

sets of orthologous exons. The most widely used clustering method for the identification of clusters of orthologous genes is Markov clustering (MCL) [159]. This method uses the principles of random walk on a graph to identify sets of nodes that tend to be recurrently traversed in a random walk. It uses an iterative algorithm with 2 steps called inflation and expansion, with their corresponding parameters. In the inflation phase, the transition probabilities are modified to enhance the more likely transitions and weaken the least likely ones by using a power of the transition probabilities followed by re-normalization. In the expansion step, the inflated matrix is then powered to calculate probabilities after certain number of steps in the random walk. Thus, after a number of iterations, the algorithm will converge and return sets of nodes that tend to remain associated [69]. Finally, if we want to obtain more confident sets of exon orthologs, we can select from each cluster only the sub-graph that is fully connected, this is, the set of exons that are all best-reciprocal hits. As exons from the same gene will never be connected in the graph, this already selects the best fitting exon in case of exon duplication. Within gene exon duplicates can otherwise be easily identified or removed if required for a particular analysis.

Modified multiple sequence alignment for order-aware exon distances

As exons are shorter than genes, it is more likely for find spurious similarities between them just by chance. On the other hand, the space search is much smaller (whole genome *vs* gene fragments), which may compensate the increased difficulty due to their shorter length. However, in contrast to genes, we have a very clear unit in which exon order is usually maintained: the gene. Thus, we propose to use a method that takes into account, not only sequence similarity, but also the relative order of exons across the different ortholgous genes for the definition of exon orthology relationships. With this purpose, we use MUSCLE [67] to perform Multiple sequence alignment (MSA) with a custom scoring matrix. We modified the pre-defined nucleotide scoring matrix in which we have added a new character to represent an intron, with a very high score for matching. This way, we force gene sequences to preferentially align exon boundaries and help aligning full exon sequences in case of poor sequence conservation. Thus, instead of calculating pairwise comparison between exon sequences without taking into account their relative position in the transcript, we define the score for building the best reciprocal hit graph as the number of aligned positions in the MSA between a pair of exons.

For our analysis, we have used the following scoring scheme for MSA, resulting in the scoring matrix at Table 3.1.

- +40 for matching introns
- -1 for gap extension
- -10 for gap opening
- 4 for nucleotide match
- -3 and -4 for transitions and transversion missmatches, respectively
- -10 for aligning intron with a nucleotide
- 0 for a match with "N" character

Moreover, we set the inflation and expansion parameters to 2 for MCL as provided by default in the MCL-Markov-Cluster python library, and applied the complete sub-graph method to filter high confidence orthologs with exons across at least 30% of the species included.

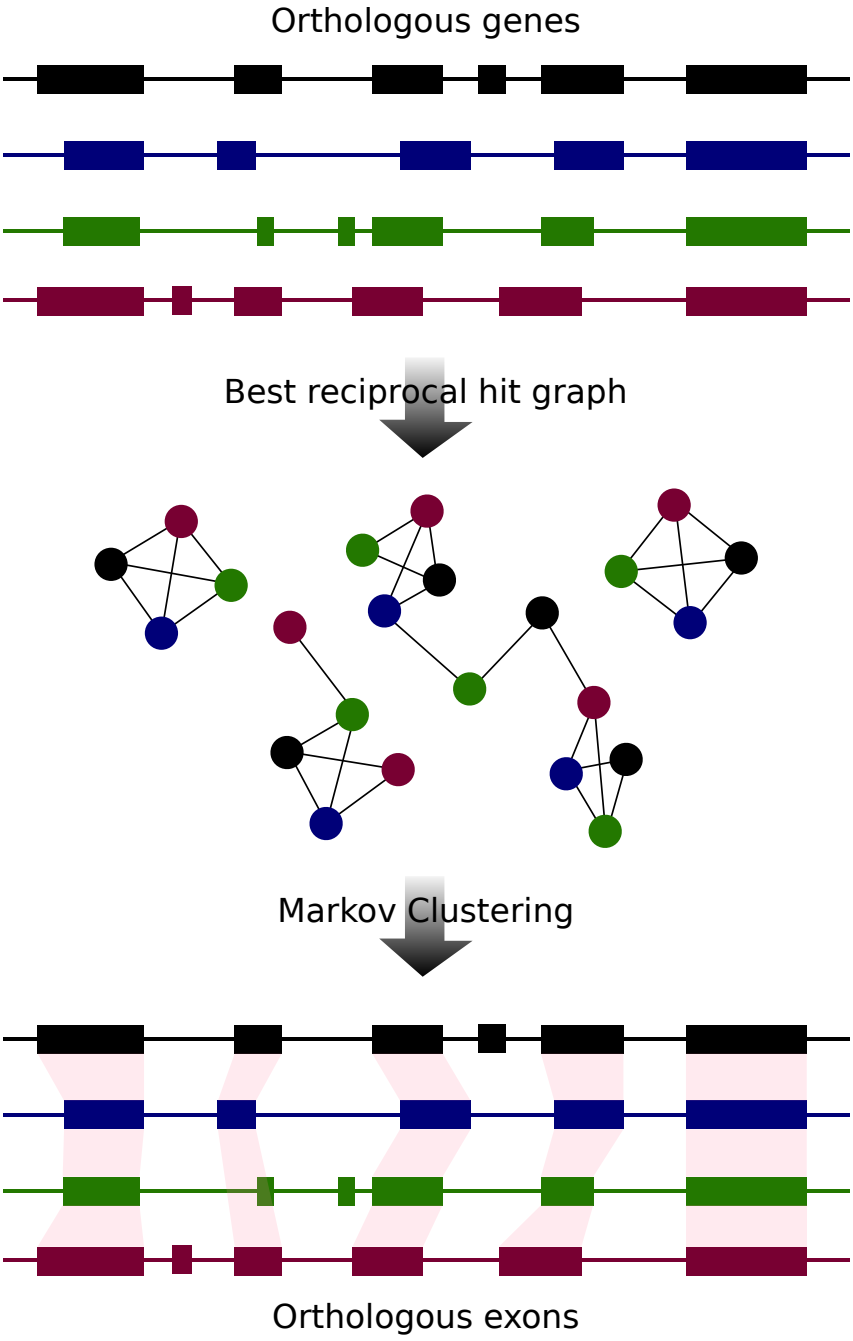

Figure 3.7: General framework for finding orthologous exons from orthologous genes using Markov clustering (MCL) on a best reciprocal hit graph

Table 3.1: Scoring matrix for Multiple sequence alignment (MSA)

Nucleotide	A	T	G	C	Intron	N
A	5	-4	-3	-4	-10	0
T	-4	5	-4	-3	-10	0
G	-3	-4	5	-4	-10	0
C	-4	-3	-4	5	-10	0
Intron	-10	-10	-10	-10	40	0
N	0	0	0	0	0	0

3.3.4 Models of evolution of quantitative traits

Brownian motion (BM) model

In a Brownian motion (BM), variations introduced in the mean of the trait in a population by random forces are accumulated over time without constraint at a constant rate τ^2, such that the trait is expected to diverge from the ancestral population linearly with time, on average (Figure 3.9).

$$d\bar{X} = \tau^2 dW_t \tag{3.41}$$

$$d\bar{W}_t \sim Normal(0, dt) \tag{3.42}$$

A population with a starting average trait value $\bar{X}_0$ will evolve by accumulating random variance over time, such that after some time t, the average $\bar{X}$ will follow a normal distribution centered at the starting value $\bar{X}_0$, with a variance proportional to t

$$\bar{X} \sim Normal(\bar{X}_0, \tau^2 t) \tag{3.43}$$

In a phylogenetic context, this process happens independently on each branch of the tree. Although we do not have data of the trait values at the internal nodes of the tree, we can model them as latent parameters and derive the joint probability of observing the data at the tips of the tree and the whole set of parameters $\Theta = \{\vec{\bar{X}}, \tau^2\}$.

In a tree with known topology and B branch lengths l_i, for each branch i:

$$p\left(\bar{X}_i | \bar{X}_i^{Parent}, l_i\right) = Normal\left(\bar{X}_i^{Parent}, \tau^2 l_i\right) \tag{3.44}$$

This probability can be calculated recursively for any combination of $\vec{\bar{X}}$, representing trait values for every node in the tree, and τ^2 values. The recursion ends up at the root of the tree $\bar{X}_0$, for which we need to specify a prior distribution $p(\bar{X}_0)$. In this case, as we have no prior expectation about the inclusion rates in the mammalian ancestor for a given exon, we set a uniform prior on the $\Psi_0 = InvLogit(\bar{X}_0)$.

$$p(\bar{\Psi}_0) = Uniform(0, 1) \tag{3.45}$$

However, we cannot directly measure population means $\bar{X}$, but only a few individuals in the population for each species. We assume that the within species variance is the same across species σ^2. For each individual n belonging to species j

$$p(X_n | \bar{X}_j, \sigma^2) = Normal(\bar{X}_j, \sigma^2) \tag{3.46}$$

This within species variance is directly related to τ^2 through *beta*, which represents the degree at which the variance present in the population accumulates with time.

$$\tau^2 = \beta\sigma^2 \tag{3.47}$$

Thus, we just now need to set priors on σ^2 and β, under which we allow flexible evolution of Ψs in a relatively long evolutionary period of 100 my, as shown in Figure 3.8.

$$\sigma^2 \sim Gamma(1, 1)$$

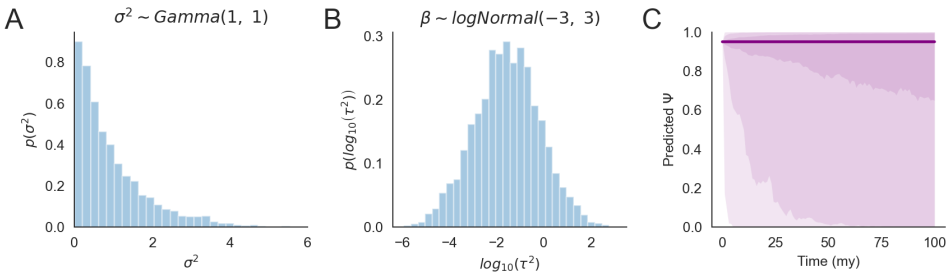

Figure 3.8: Prior distributions used for an exon-level BM model for exon inclusion rate. **A** Gamma prior for the within species variance. **B** Resulting prior on the evolutionary rate τ^2 by setting a logNormal prior on the relative rate β. **C** Predicted evolution of Ψ in 100 my from an exon starting at $\Psi_0 = 0.95$ under the specified priors. Shaded areas represent 2.5, 5, 10, 25 percentiles of the predicted Ψ at each time point

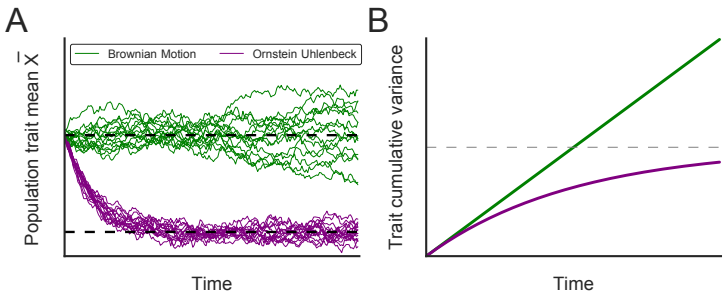

Figure 3.9: Traits evolving under Ornstein-Uhlenbeck (OU) and Brownian motion (BM) models over time (A) and how the variance of the expected distributions changes with time from the ancestral population (B)

$$\beta \sim logNormal(-3, 3)$$

Having specified the conditional and prior distributions across all model parameters, we can now fully specify $p(X, \Theta)$, which is directly proportional to the posterior probability $p(\Theta|X)$.

$$p(X, \Theta) = p(X|\Theta)p(\Theta) = p(\sigma^2)p(\beta)p(\Psi_0) \prod_{n=1}^{N} p(X_n|\bar{X}_n, \sigma^2) \prod_{i=1}^{B} p\left(\bar{X}_i|\bar{X}_i^{Parent}, l_i, \tau^2\right) \tag{3.48}$$

Ornstein-Uhlenbeck (OU) model for a single exon

The OU model is a generalization of the BM model, in which there is a pull towards an optimal value i.e. the value that maximizes the fitness of the population. The strength of the pull can be understood as the selective strength, which constrains the divergence from the optimal values: if random forces introduce variation too far from the optimal value, selection will eliminate this variation. Therefore, whereas in the BM model the variance increases linearly with time, the stabilizing selection introduced in an OU model drives the saturation of this variance (Figure 3.9).

We follow Butler's work for the main derivation of the OU model [42]. Let $\bar{X}$ be the mean of a continuous character evolving over an infinitesimal time dt with a unique optimal value μ and infinitesimal random fluctuations given by the infinitesimal variance τ^2. Then the expected infinitesimal change in

the mean of the character is

$$d\bar{X} = \alpha(\mu - \bar{X})dt + \tau^2 dW_t;\ dW_t \sim Normal(0, dt) \tag{3.49}$$

Starting from the mean state of the last common ancestor $\bar{X}_0$ at time t_0, the trait value is then normally distributed with mean

$$E[\bar{X}](t_0, \bar{X}_0, \alpha, \mu) = \bar{X}_0 e^{-\alpha t_0} + \mu\left(1 - e^{-\alpha t_0}\right) \tag{3.50}$$

and variance

$$Var[\bar{X}](t_0, \alpha, \sigma_\mu) = \frac{\tau^2}{2\alpha}\left(1 - e^{-2\alpha t_0}\right) . \tag{3.51}$$

Thus, the evolution across a l_i long branch i of a phylogenetic tree with a total B branches also follows a normal distribution

$$\bar{X} \sim Normal\left(X_0 e^{-\alpha l_i} + \mu(1 - e^{\alpha l_i}),\ \frac{\tau^2}{2\alpha}(1 - e^{-2\alpha l_i})\right) \tag{3.52}$$

As with the BM, we can add intra-species variance σ^2 and trait values at the internal branches as latent parameters of the model as shown by equation 3.46. We also need to specify a prior distribution for the trait value at the root of the tree $\bar{X}_0$. In contrast to the BM model, the OU model can provide a sensible prior: as time increases ($t \to \infty$), the expected distribution tends to reach a stationary state. Thus, if we assume that time evolving under the same regime before species divergence has been sufficiently long, we can expect the trait value at the root to be drawn from the equilibrium distribution.

$$X_0 \sim Normal(\mu,\ \frac{\tau^2}{2\alpha}) \tag{3.53}$$

The resulting joint probability can then be written as follows:

$$p(X, \Theta) = p(X|\Theta)p(\Theta) = p(\sigma^2)p(\alpha)p\left(\frac{\tau^2}{2\alpha}\right)p\left(X_0|\mu, \frac{\tau^2}{2\alpha}\right)p(\mu)\prod_{n=1}^{N} p(X_n|\bar{X}_n, \sigma^2)\prod_{i=1}^{B} p\left(\bar{X}_i|\bar{X}_i^{Parent}, l_i, \alpha, \frac{\tau^2}{2\alpha}, \mu\right) \tag{3.54}$$

Whereas the BM single parameter can be inferred with confidence with a relatively small comparative dataset, there is increasing evidence of strong bias in the inference of OU parameters, mostly α, using phylogenies fewer than 200 species [58, 232].

Global Ornstein-Uhlenbeck (OU) model

Despite our inability to infer every evolutionary parameter of the OU model for each exon independently, we can assume that exons will evolve under the same process i.e. selective constraint and neutral evolutionary rate. In other words, we can assume that the parameters that are hard to estimate from a single exon i.e. α, $\frac{\tau^2}{2\alpha}$, σ^2 are common across all of them. Even if this may not be fully accurate, we can at least use this approximation to obtain an average estimate across the whole set of exons under study. Similar approaches have been previously used to estimate biological variance in transcriptomic datasets with few samples [229, 171, 185].

To calculate the joint probability of the parameters and the data across the whole set of exons together, we assumed that exons evolve independently from each other, since this allows calculating the joint probability by multiplying the probabilities for individual exons.

$$p(\mathbf{X}, \Theta) = p(\Theta)\prod_{k=1}^{K} p(X_k|\Theta) \tag{3.55}$$

We assumed that all species share a common optimal value for the same exon, but each exon may have a different optimal value μ_k, which we assume to be drawn from a normal distribution; and root values are still drawn from the equilibrium distribution as previously described (equation 3.53).

$$\mu_k \sim Normal(\mu_0, \sigma_\mu^2) \tag{3.56}$$

Now, instead of using latent parameters representing the trait values across every internal node of the tree for each of the exons, which greatly expands the number of parameters to infer, we used an alternative multivariate parametrization [42] using covariance between samples separated by a given distance in the phylogenetic tree. For species i and j, the covariance between $\bar{X}_i$ and $\bar{X}_j$ depends on their distance in the tree $t_{i,j}$; and the total evolutionary time t_0 (sum of branch lengths involving the species i and j).

$$Cov(\bar{X}_i, \bar{X}_j | t_{i,j}, t_0, \alpha, \tau) = \frac{\tau^2}{2\alpha}\left(e^{-\alpha t_{i,j}} - e^{-2\alpha t_0}\right) \tag{3.57}$$

With the trait values evolving along a bifurcating tree with known divergence times $\mathbf{T}$, we can model the vector of trait values $\vec{\bar{X}}$ as a multivariate normal distribution with a common mean:

$$\vec{\bar{X}} \sim MvNormal\left(E[\bar{X}](t_0, \bar{X}_0, \alpha, \mu),\ Cov[\vec{\bar{X}}](\mathbf{T}, \alpha, \tau)\right) \tag{3.58}$$

For later use, we will denote $\Sigma_{OU} = Cov[\vec{\bar{X}}]\,(\mathbf{T}, \alpha, \tau)$ as the variance-covariance matrix for species means, which will be common to all exons. As such, we need to invert Σ or calculate its Cholesky decomposition only once for the whole dataset, becoming more computationally efficient than inferring internal nodes as in the single exon models.

To account for intraspecific variability, we model directly trait values across individuals rather than species. To do so, we can simply expand the covariance matrix to represent samples covariance rather that species covariance. Thus, we still use the OU covariance function (equation 3.57) based on the species to which each individual sample belong, but also add a common within species variance σ^2, common across species, to the diagonal elements of the covariance matrix

$$\Sigma = \Sigma_{OU} + \Sigma_{ind} = \Sigma_{OU} + \sigma^2\mathbf{I} \tag{3.59}$$

Using this, we can derive the joint probability of the data $\mathbf{X}$ and a certain combination of parameter values Θ as the product of conditional probabilities. Considering $\mathbf{X}$ as the matrix containing trait values for all K exons and N samples.

$$p(\mathbf{X}, \Theta) = p(\mu_0)p(\sigma_\mu^2)p(\alpha)p(\tau^2)p(\sigma^2)\left[\prod_{k=1}^{K} p\left(\vec{X_k}|\mu_k, \bar{X}_{0,k}, \alpha, \tau^2, \sigma^2, \mathbf{T}\right) p\left(\bar{X}_{0,k}|\mu_k, \alpha, \tau^2\right) p\left(\mu_k|\mu_0, \sigma_\mu^2\right)\right] \tag{3.60}$$

There remains to specify the prior distributions for the basic OU parameters. We aim to specify relatively informative prior distributions for the OU model, with the aim of setting soft bounds on the parameter space in which the OU model seems reasonable for evolving exon inclusion rates.

- Even if many genes can be alternatively spliced, as one major AS isoform is usually produced [70], we expect most exons to have high optimal inclusion rates. This allows us to specify prior distributions for the mean and variance of the optimal values in our model

$$\mu_0 \sim Normal(2, 3)$$

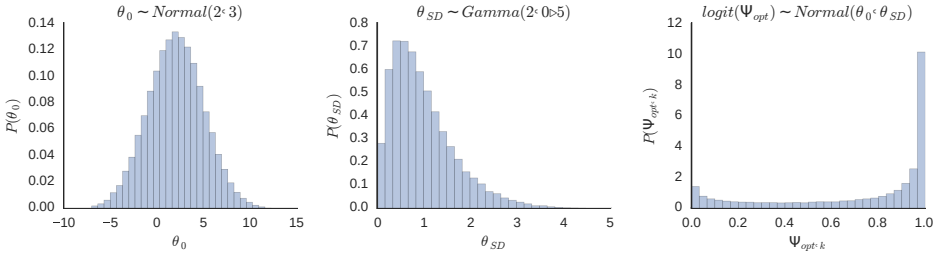

Figure 3.10: Prior distributions for parameters related with AS optimal values, i.e. mean μ_0 and standard deviation μ_{SD}, and the expected distribution of optimal inclusion rates across exon skipping events $\Psi_{opt,k}$.

$$\mu_{SD} \sim Gamma(2, 0.5)$$

These prior distributions are motivated by the resulting expectation on optimal inclusion rates $\Psi_{opt,k}$, resulting from the inverse logit transformation after sampling from a normal distribution with those means and variances. With these prior distributions, we obtain a relatively high amount of exons with Ψ_{opt} close to 1, but allowing certain probability for exon skipping at all possible rates (see Figure 3.10)

- To specify a prior distribution on selection strength α, we parametrized it as a function of the phylogenetic half-life $t_{\frac{1}{2}}$ i.e. the time required for selection to reduce the distance to the optimal value by half.

$$t_{\frac{1}{2}} = \frac{log(2)}{\alpha}$$

If selection is very weak, $t_{\frac{1}{2}}$ will be longer than the range of times included in the phylogeny. However, as this depends on the scale of the phylogeny, we set a prior on a scaled phylogenetic half-life that per time unit in the phylogenetic tree:

$$t^*_{\frac{1}{2}} \sim Gamma(2, 4) t^*_{\frac{1}{2}} = t_{\frac{1}{2}}/2.47$$

we put 95% of the prior probability mass between [2.53, 53.11] 100 m.y., consistent with the *a priori* expectation of relatively neutral evolution of exon inclusion rates (Figure 3.11).

- We set also a Gamma prior distribution on the equilibrium variance

$$\frac{\tau^2}{2\alpha} 10^{-2} \sim Gamma(2, 3)$$

which allows a very wide range of values for the equilibrium variance, enough to allow any inclusion rate in the equilibrium *a priori* but avoid very large and meaningless values.

- Finally, σ is parametrized as a function of τ, since rates of neutral evolution are expected to be proportional to the genetic variance. Therefore, we set a prior on the proportionality constant β

$$\beta \sim logNormal(3, 1) \sigma^2 = \frac{\tau^2}{\beta}$$

such that the resulting within species variance has 95% prior probability between 0.001 and 0.418 (Figure 3.11). This apparently strong informative prior on the logit scale actually allows relatively high variability in inclusion rates i.e. individuals from a species with an average $\bar{\Psi} = 0.5$ will show Ψ between 0.38 and 0.62 with a probability of 95% (Figure 3.11E).

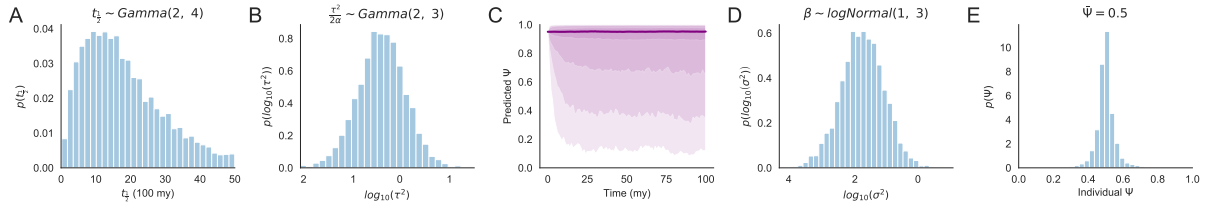

Figure 3.11: Prior distributions for parameters related with AS evolutionary rates and derive quantities. **A** phylogenetic half-life $t_{\frac{1}{2}}$. **B** Incremental variance τ. **C** Expected Ψ evolution under the specified prior distributions **D** Within species standard deviation σ **D** Expected within species Ψ distribution when the $\bar{\Psi} = 0.5$

.

Inference of optimal inclusion rates

While the global OU model allows inference of the specific optimal values for each exon included in the dataset, fitting it with the whole dataset is not feasible given the complexity of the model and the large computational burden that it implies. However, we know that, with about 2000 exons the estimates of the global parameters α, τ^2 and σ^2 become reliable. Thus, we can simplify the previous single exon model (equation 3.54) by plugging these estimates as known parameters.

$$p(X, \Theta) = p(X|\Theta)p(\Theta) = p\left(X_0|\mu, E\left[\frac{\tau^2}{2\alpha}\right]\right) p(\mu) \prod_{n=1}^{N} p(X_n|\bar{X}_n, E\left[\sigma^2\right]) \prod_{i=1}^{B} p\left(\bar{X}_i|\bar{X}_i^{Parent}, l_i, E[\alpha], E\left[\frac{\tau^2}{2\alpha}\right], \mu\right) \tag{3.61}$$

This leaves the trait values at each node $\bar{X}$ of the phylogenetic tree and the optimal value μ as only parameters, which can be easily inferred now for each exon independently. In this model, instead of setting a prior directly on μ, we specified a uniform prior for the optimal inclusion rate Ψ_{opt}, which can then be transformed to μ using the logit function.

$$\Psi_{opt} \sim Uniform(0, 1)$$

$$\mu = logit(\Psi_{opt}) = log\left(\frac{\Psi_{opt}}{1 - \Psi_{opt}}\right)$$

Adaptive evolution of exon inclusion rates

As before, we assumed that the inferred global OU parameters are common across every exon in the transcriptome. This provides an expectation of the change in Ψs that is allowed by random change and stabilizing selection in a period of time. Therefore, if there is strong evidence for an unusually large change at some branch of the phylogeny, we can reliably say that there has been a shift in the optimal value. We now have an optimal value μ_b for each branch b, which depend on the ancestral optimal value μ_0 and the change in this optimal value along each branch of the tree $\Delta\mu_b$.

$$\mu_b = \mu_{b,parent} + \Delta\mu_b$$

We can safely assume that most of the times there are no changes in the optimal values: they are zero. We can formalize this idea using a horse-shoe prior for regularization [46]. The horseshoe prior, a member of the family of hierarchical shrinkage priors, specifies a normal prior for $\Delta\mu_b$ with mean 0 and a standard deviation τ_b, where τ_b is not a fixed value, but drawn from a common half Cauchy distribution

with location parameter 0 and ρ scale. τ_b represents a local shrinkage parameter, as it only affects branch b, whereas ρ can be understood as a global shrinkage parameter, affecting the whole phylogenetic tree. We further set a half Cauchy prior in ρ centered at 0 with scale parameter of 0.1. While 1 was first recommended for general applications [46], it was later shown to add only a weak regularization [219], so we reduced it to 0.1.

$$p(\Delta\mu_b \mid \tau_b) = Normal(\Delta\mu_b \mid 0,\ \tau_b) \tag{3.62}$$

$$p(\tau_b \mid \rho) = Cauchy^{+}(\tau_b \mid 0,\ \rho) \tag{3.63}$$

$$p(\rho) = Cauchy^{+}(\rho \mid 0,\ 0.1) \tag{3.64}$$

With this model what we infer is the minimal combination of shifts, if any, that best explain the observed patterns across extant species given the known patterns of evolution in absence of shifts, as provided by the global fit.

Previous models allow the inference of the specific location of a number of shifts, which is a parameter itself [283]. This required the MCMC algorithm to jump between models of different complexity, as the number of parameters changes with the number of shifts, which is intractable in stan [45], which leverages the continuous nature of the parameter space to propose new samples with high acceptance probability. Instead, we specified potential fixed locations of those shifts and assume that most of them will be zero.

Global prevalence of adaptive evolution

Let p_a be the probability that an exon has experienced a shift along its evolutionary history or the proportion of exons that have adaptatively change during mammalian evolution. Whether an exon has a shift or not S is a binary outcome drawn from a Bernoulli distribution with parameter p_a

$$S \sim Bernoulli(p_a)$$

However, we do no have access to S directly, as our method for detecting shifts is not perfect and makes errors: there is a certain probability that our method detects a shift when there is a shift $p(shift|S = 1)$, which is true positive rate (TPR) sensitivity; and a probability that it detects a shift when there is not in reality $p(shift|S = 0)$, known as false positive rate (FPR). Thus, if we know how well the method performs through these parameters, we can derive the probability of observing a shift p_s with our method by summing the probabilities from both scenarios:

$$p_s = p(shift|S = 1)p(S = 1) + p(shift|S = 0)p(S = 0) = TPRp_a + FPR(1 - p_a) \tag{3.65}$$

At the same time, by fitting the shifts model to our data, we obtain the number of exons with observed shifts Q out of a total of K exons, which is drawn from a binomial distribution.

$$Q \sim Binomial(K, p_s)$$

Now, how do we infer the sensitivity and false positive rates of the test? We need to have some data in which we know the true states of the exons to be able to obtain information about those parameters. We generated such dataset by simulating data under the inferred OU parameter values with a probability of having a shift in each branch of 0.003. This probability results in a dataset in which about half of the trees have at least one shift, so that we can reliably evaluate the performance of the method for

identification of exons with a shift along their evolutionary history. To do so, we counted the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), which depend on the sensitivity and false positive rates:

$$TP \sim Binomial(TP + FN, TPR)$$

$$FP \sim Binomial(FP + TN, FPR)$$

We specified uniform priors for the 3 model parameters, as we do not have *a priori* information about the performance of the method or the p_a

$$p(p_a) \sim Uniform(0, 1)$$

$$p(TPR) \sim Uniform(0, 1)$$

$$p(FPR) \sim Uniform(0, 1)$$

which allow us to specify the joint probability of the model parameters $\Theta = \{p_a, TPR, FPR\}$ and the $data = \{Q, K, TP, TN, FP, FN\}$

$$p(data, \Theta) = p(Q|p_a, TPR, FPR)p(TP|TP + FN, TPR)p(FP|FP + TN, FPR)p(TPR)p(FPR)p(p_a)$$

In this model, we simultaneously infer the performance of the method on a finite dataset with known true values, and estimate the prevalence or proportion of exons with shifts using the observed data. This way, uncertainty in the estimation of the sensitivity and false positive rate is propagated to the estimation of p_a.

Parameter inference using Markov Chain Monte Carlo

Models were implemented in stan, a probabilistic programming language for bayesian inference [45]. Stan uses No-U Turn Sampler (NUTS) [106] to approximate the complex joint posterior distribution of the parameters. NUTS is a type of Markov Chain Monte Carlo (MCMC) algorithm based on Hamiltonian Monte Carlo (HMC) that aims to maximize the effective sample size (ESS) per computation time using proposal distributions with high acceptance probability and low autocorrelation [31]. The Markov chains visit points in the parameter space proportional to their posterior probability, so we can use the sample to estimate probabilities. For each dataset, 4 chains were run for 1000 iterations after 1000 warm-up iterations for sampler parameters optimization. We assessed convergence and mixing of the chains by measuring ESS and Gelman-Rubin R ($\hat{R}$) of the parameters of interest.

Code availability and reproducibility

Main methods to extract exon orthologs, extract exon biases, merge and handle counts matrices, and fit quantitative models for Ψ evolution are implemented in an in-house python library AS-quant. Specific code to reproduce the analyses performed in this section are found in a different repository AS-evolution

4. Results

4.1 Functional impact and regulation of alternative splicing in heart development and disease

4.1.1 Characterization alternative splicing patterns in the developing and diseased heart

To characterize the AS changes taking place during heart development and disease, we collected a large dataset of heart samples from mouse models at different developmental stages and disease conditions (Table S1). We grouped the collected samples according to 5 major phenotypes, including the three major developmental stages that have been previously characterized (embryo (E10.5-E17), post-natal (P0 to P7), and adult (from P10) stages) [126, 125, 92]; but also two common models of heart disease:

- myocardial infarction (MI), induced by permanent ligation of the left anterior descending (LAD) coronary artery. Thus, blood flow to a specific region of the heart is blocked as in humans, when a blood clot occludes a coronary artery. Thus, cells in this region die and we can study the transcriptional changes taking place in the infarcted, border or remote region.
- trans-aortic constriction (TAC), by tying a knot in the aorta that reduces its diameter as a model of increased arterial pressure. Thus, the heart is forced to grow to overcome the increased resistance to blood flow as in human hypertensive patients.

Using this categorization of samples, we first identified a set of over 20,000 AS events with at least one read supporting skipping or inclusion in 20% of the samples. We then focused on AS changes occurring in these specific transitions:

1. embryonic development (ED), by comparing neonatal with embryonic samples.
2. post-natal development (PD), by comparing uninjured adult samples with neonatal samples
3. TAC, by comparing samples from hypertrophic and uninjured adult hearts
4. MI , by comparing samples from infarcted and uninjured adult hearts

Differentially spliced events for each comparison were identified using a Generalized Linear Mixed Model (GLMM) with binomial likelihood and logit link function. To account for variability across experimental settings and sample types, we included the experiment, sample type and individual as random effects in the model. This model also allowed us to estimate the average Percent Spliced In (Ψ) in each of the phenotypes simultaneously removing unwanted sources of variability, like batch or sequencing conditions that may affect the observations in each of the collected samples.

AS changes were more abundant in the developmental transitions than in the disease models, suggesting more prominent roles of AS during embryonic and postnatal development (Figure 4.1A). Given

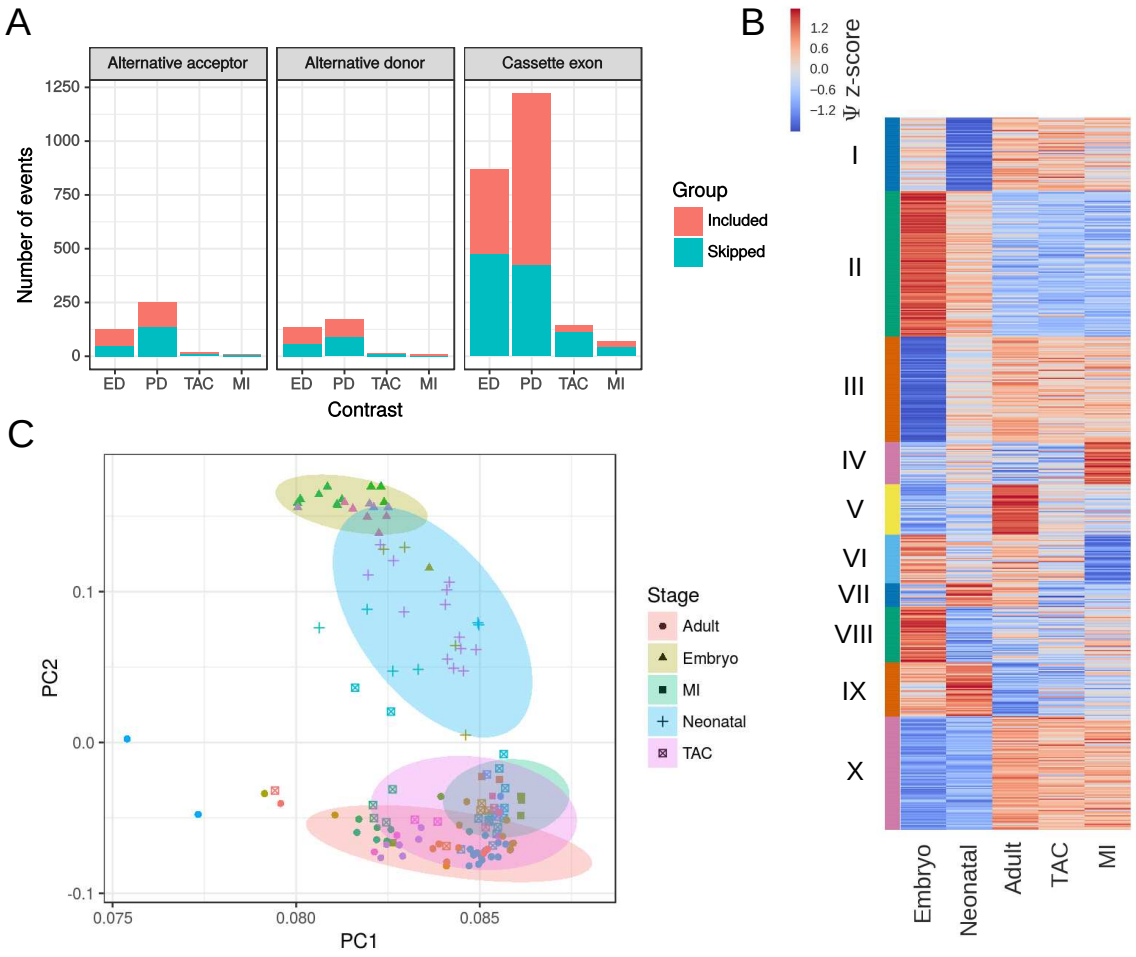

Figure 4.1: Alternative splicing landscape in heart development and disease. **A** Number of events showing significant differences in each comparison according to the event type: alternative acceptor, alternative donor, and exon skipping. **B** Heatmap representing z-scores calculated from estimated Ψ for each condition. Clusters were calculated using k-means on the normalized profiles. **C** Principal Component Analysis of all analyzed samples using exon cassette events without missing data in any of the samples. Different symbol colours represent different datasets or experiments. Ellipses were drawn according to each condition.

that the main AS changes in all comparisons were cassette exon events, we focused on this type of event in downstream analyses. Interestingly, whereas exon inclusion and skipping were observed in similar amounts during both embryonic and post-natal development, heart disease was mainly characterized by increased exon skipping (Figure 4.1).

In agreement with the observations in Figure 4.1A, K-means clustering of the standardized Ψ profile (Figure 4.1B) revealed that the largest clusters (II,III,IX,X) were those specific of developmentally regulated exons, with smaller clusters identified with specific changes in TAC and MI (clusters IV and VI). Interestingly, clusters V, VII and VIII show similar pattern in MI and in embryonic samples, suggesting partial re-expression or re-repression of the neonatal AS pattern after cardiac injury. Furthermore, Principal Component Analysis (PCA) showed a small displacement of TAC and MI samples toward neonatal samples (Figure 4.1C) reinforcing this idea. Embryonic and neonatal samples were clearly separated from the adult samples, supporting the notion that developmental stage is the main source of variability, over batch effects or sample source. This was nonetheless not specific to AS, as PCA of expression data showed a similar pattern (Figure 4.2).

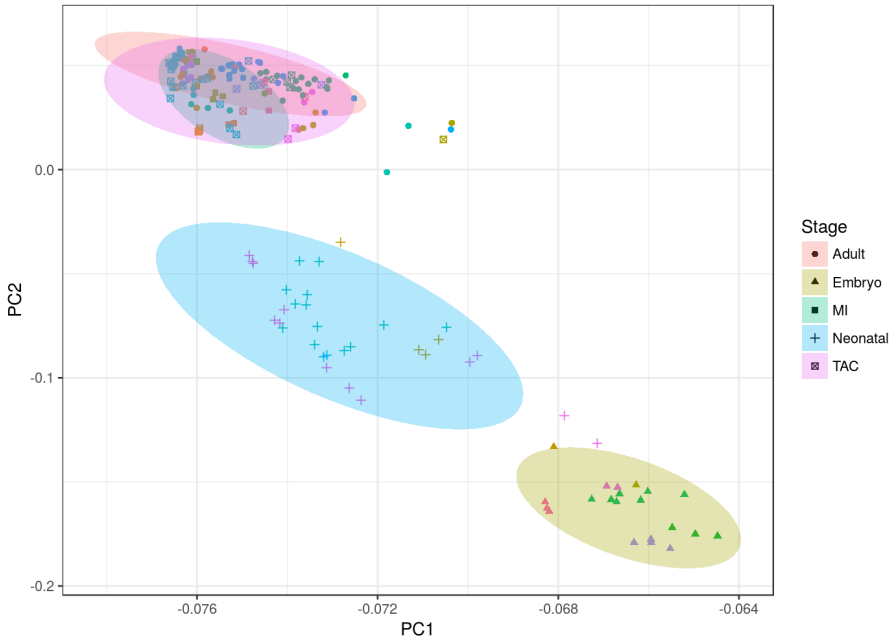

Figure 4.2: Principal Component Analysis using gene expression data. Normalized counts were transformed to z-scores, and genes with missing data were removed. Symbol shapes represent different conditions and symbol colors represent different experiments or datasets.

4.1.2 Functional impact of alternative splicing changes in the heart

Alternatively spliced exons potentially result in different protein isoforms both in developmental transitions and in disease

We next investigated the potential impact of AS changes on heart physiology. Two of the main potential roles of AS are the production of different protein isoforms from the same gene and the regulation of gene expression by RUST. Exons that generate alternative protein isoforms usually preserve the ORF to avoid early termination of protein translation. We found that exons undergoing AS changes in the heart in all studied conditions have a higher probability of having a length multiple of three, and thereby preserve the ORF, than those exons that were differentially spliced (Figure 4.3A; Fisher test p-value <0.01 for all comparisons except MI Included and TAC Included). To investigate the potential impact of AS on these potential proteins, we compared the exon length and found that exons preferentially included in developmental transitions and those preferentially skipped in TAC or MI tended to be shorter than those that showed no significant changes (Figure 4.3B, Mann-Whitney U test p-value < 0.05). These exons, which were the most abundant (Figure 4.1A), showed an under-representation of PFAM domains, suggesting that they do not tend to introduce strong changes in protein function. In contrast, a higher proportion of exons skipped in the ED or PD comparisons or included in MI encoded PFAM domains, and are therefore expected to have a greater impact on protein function (Figure 4.3C, Fisher test p-value < 0.05). These results, altogether, suggest that the prevalent role of AS in the heart is to generate slightly different protein isoforms rather than to modulate gene expression through NMD.

Modulation of protein-protein interaction networks by AS in heart disease

Alternative splicing protein isoforms have been previously shown to have different interaction partners [301]. To investigate whether AS changes in the heart regulate protein protein interactions we first compared the connectivity of proteins encoded by genes undergoing differential AS using the Intact

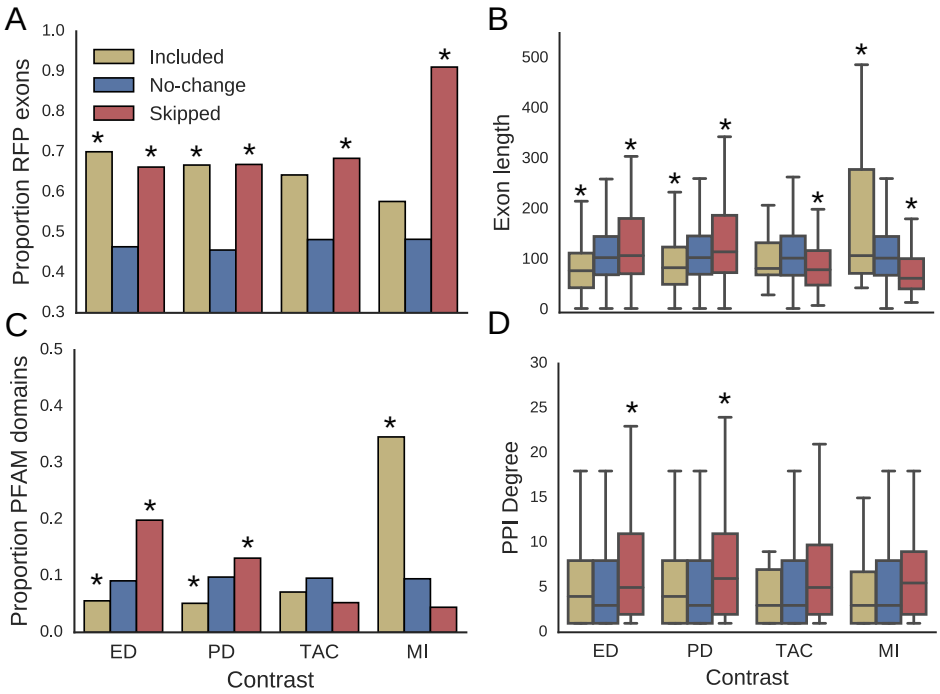

Figure 4.3: Properties of alternatively spliced exons in the heart. **A-D**, Values of different properties in each comparison according to whether inclusion levels where increased (Included), decreased (Skipped), or not significantly changed (No-change). **A** Proportion of exons with length that is multiple of 3 (reading frame preservation (RFP)) and therefore have no impact on the open reading frame. **B**, Exon length distribution. **C**, Proportion of exons overlapping with an annotated PFAM domain. **D**, Number of connexions or degree in the Intact protein-protein interaction (PPI) network.

Protein-Protein Interaction (PPI) network [206]. For all comparisons, genes with skipped exons showed more connections to other proteins than genes without significant AS changes (Figure 4.3D; Mann-Whitney U test p-value < 0.05 for PD and ED, p-value < 0.15 for MI and TAC). To check whether this was a general property of transcriptional changes, we compared the PPI degree distribution of proteins depending on their gene expression changes. In contrast to AS, DEGs did not show a higher number of connections in the PPI network (Fig. S2), suggesting that this feature is specific to AS.

However, this does not necessarily mean that these particular AS changes are modulating the interaction capabilities of these proteins. To determine whether PPIs are actually regulated by AS changes, we used information of domain-domain interaction (DDI) information [87], and assumed that exons located in a domain that mediates an interaction between two proteins are actually required for such interaction to take place. Thus, we can identify interactions that are increased or decreased depending on whether the inclusion of the exon spanning the domain increases or decreases, respectively. We found that exons included during TAC or MI affected more domain mediated PPIs than unchanged exons (OR=3.50 and OR=2.42, Fisher tests p=0.06 and p=0.21, respectively). Skipped exons in disease, if anything, avoided changing interactions (OR=0.51 an OR=0.65, p-values ¿ 0.2) (Figure 4.4A). Overall, these results suggest that AS changes can increase the number of interactions by increasing exon inclusion, but avoid reducing them through exon skipping.

The impact of increasing or decreasing the amount of interacting proteins may however depend on their position in the PPI network. In other words, reduction or even complete ablation of the binding affinity between two proteins by differential exon inclusion may only affect slightly the overall function of a protein complex, as cooperative binding of the remaining elements may compensate this lack of binding between some of their elements. On the other hand, if two protein complexes interact only through an interaction between a pair of proteins, the modulation of this interaction is expected to have greater functional consequences. Thus, we analysed how AS changes affect the structure of the PPI network beyond isolated interactions. To do so, we built an undirected graph using pairwise interactions for genes expressed in at least one condition and calculated the betweenness for each edge. DDI interactions potentially modulated by AS showed, in general, lower betweenness, compared to AS-insensitive interactions (Figure 4.4B, Mann-Whitney U test p-value$< 10^{-6}$). These results suggest that AS-modulated interactions tend to be located within closely interacting modules rather than connecting different protein complexes. When comparing across groups of modulated exons, we found that exons modulated during ED have significanly higher betweenness in the PPI network than the unchanged exons (Mann-Whitney U test p=0.01, Figure 4.4C), suggesting a stronger rewiring of the interaction networks during early heart development than in any other condition. No significant difference was found in the betweenness for exons modulated during heart disease (Mann-Whitney U test p¿0.1, Figure 4.4C).

Since proteins do not only interact through protein domains, we next studied AS-mediated PPI changes in experimentally built networks that are not limited to DDIs [305]. We found that 100% of exons changing in disease AS changes are located in genes with known AS-dependent interactions (Figure 4.4D), significantly greater than the approximately 60% observed for unchanged exons (Fisher test p¡0.0001 and p=0.53 for TAC and MI, respectively). These findings are specific to AS since GE changes showed the opposite trend: only developmentally regulated genes are associated to AS-modulated interactions (Figure 4.4E). To investigate the global impact of AS on the PPI network, we built an interaction network using only experimentally tested interactions in this dataset and calculated the edge betweenness, as before. Interactions affected by AS changes in TAC and MI showed higher betweenness than unchanged exons (Mann-Whitney U test p¡0.01 and p=0.05, respectively, 4.4F), suggesting a rewiring of the PPI network by AS in heart disease. Despite the low statistical power due to the small size of groups overlapping with available PPI data in each dataset, our results suggest that AS changes significantly alter PPIs networks in heart disease.

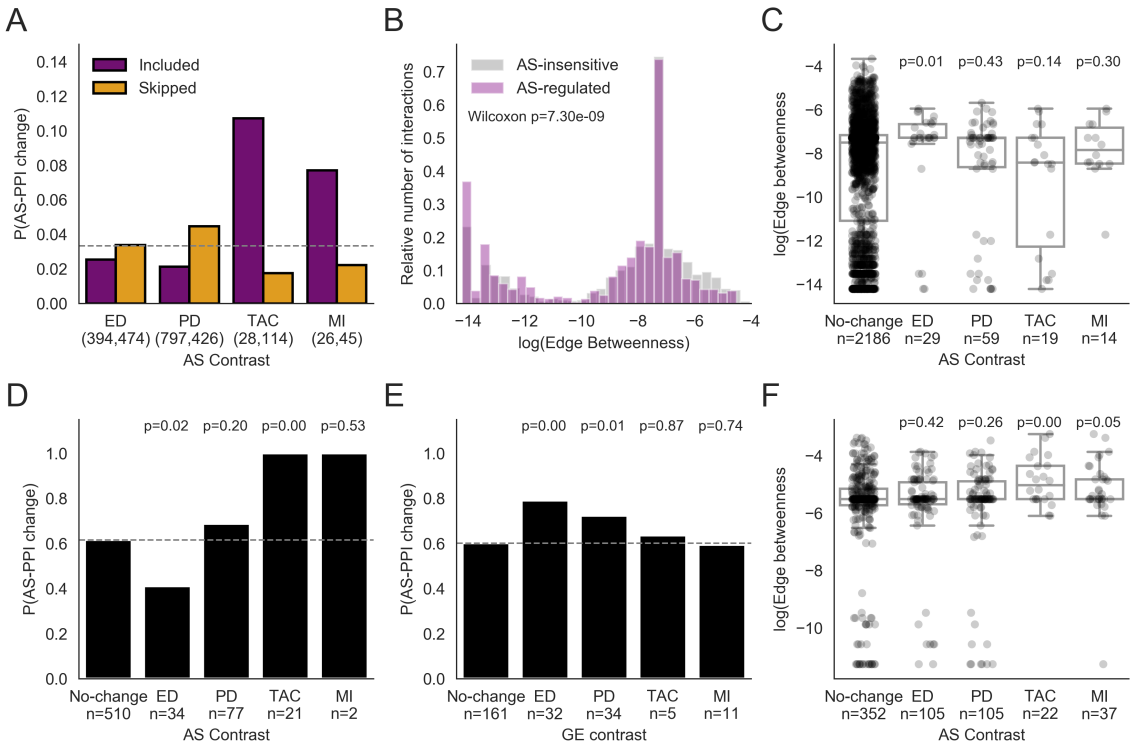

Figure 4.4: Impact of AS changes in the protein-protein interaction networks. **A**. Estimated proportion of exons that map to a domain mediating protein-protein interaction according to whether they are included, skipped or unchanged in the comparisons under study. Numbers in brackets represent the number of exons in the groups Included and Skipped for each contrast, out of 8893 total exons under study. **B**. Distribution of the log(Edge betweenness) for interactions that were found to be potentially regulated by AS in the heart compared to the remaining interactions. An interaction was considered to be potentially modulated by AS when an exon considered to be alternative in the heart encoded a domain mediating this interaction. **C**. Network edge betweenness for interactions depending on whether they were modified by AS in the conditions under study using domain mediated interactions [87]. **D,E**. Estimated proportion of interactions that differed between AS isoforms [305] in genes with significant differences in AS (D) or GE (E) in each contrast. **F**. Network edge betweenness for interactions depending on whether they were modified by AS in the conditions under study using data in [305]. Fisher tests were used to assess differences in proportions and Mann-Whitney U tests to analyze differences in edge betweenness

AS and GE control different biological processes in the heart

We next investigated the overlap between changes in AS and changes in GE. The proportion of differentially expressed genes was similar in genes undergoing differential AS and in those showing no AS changes, suggesting no association between AS and GE changes (Figure 4.5A). Even if GE and AS regulate different genes, these affected genes might regulate the same biological processes. To investigate this possibility, we performed Gene Ontology (GO) enrichment analysis in each comparison and then calculated the pairwise semantic similarity of the 10 most significantly enriched GO terms among all groups followed by hierarchical clustering based on the similarity profile (Figure 4.5B). This analysis clustered enriched processes for alternatively spliced and differentially expressed genes separately, regardless of the biological context (development or disease), indicating that AS and GE changes affect different biological processes in the heart. Interestingly, processes regulated by AS clustered separately for disease and development, whereas processes associated with GE changes clustered together (upregulated in development and downregulated in disease clustered separately from downregulated in development and upregulated in disease). This suggests a stronger functional reexpression of embryonic GE patterns than AS patterns in heart disease. Whereas changes in GE were mainly related to cell division, the respiratory chain, and extracellular matrix deposition, AS changes were more associated with cytoskeletal organization (Figure 4.5C).

4.1.3 Studying AS regulation in heart development and disease

MBNL1 drives major AS changes during embryonic and postnatal development in a position dependent manner

To identify the potential regulators of AS in the heart, we looked for over-represented binding sites of different RBPs across different potential regulatory regions. Binding sites were collected by integrating a series of databases of CLiP-seq experiments (see Methods). We first filtered those RBPs that were found to be significantly enriched ($p<0.01$, Fisher test) in at least one group of significantly changed exons. We then used the reduced set of enriched RBPs binding to different regulatory regions as substrate for regression analysis using a GLM with binomial likelihood to take into account co-linearities across binding profiles of different RBPs. This analysis was then applied to sets of exons that were found to change in any comparison (Figure 4.6A). Our results show that MBNL1 is strongly enriched in the upstream intron of exons that are skipped and in the downstream intron of exons that are included during both PD and ED. We also found that MBNL1 binding sites in exons showing changes tend to be more conserved across evolution at the sequence level, suggestive of functional importance (Figure 4.6B). Additionally, MBNL1 expression increases during development and remains unchanged in both TAC and MI (Figure 4.6C). To test whether different RBPs may regulate different biological functions, we looked for enrichment of GO terms in exons bound by each RBP compared to all those that changed in any of the comparisons under study (Figure 4.6D, one-tail Fisher test). We found that MBNL1 tends to bind to genes related with actin cytoskeleton dynamics and cell junctions, whereas other RBPs tend to bind more to exons of RNA binding proteins or proteins located in the nucleus. Whereas other RBPs may contribute to the regulation of AS changes during development, such as QK, RBFOX1 or PTBP1/2, our results suggest that MBNL1 is the main regulatory element.

PTBP1/2 drives a partial re-expression of AS neonatal patterns in heart disease

Although MBNL1 was found to be the main regulator of AS during ED and PD, we found only a mild enrichment of the binding sites in changing exons, and its expression remained unchanged in both models of heart disease: TAC and MI (Figure 4.6C). In contrast, PTBP1 and PTBP2 binding sites, among

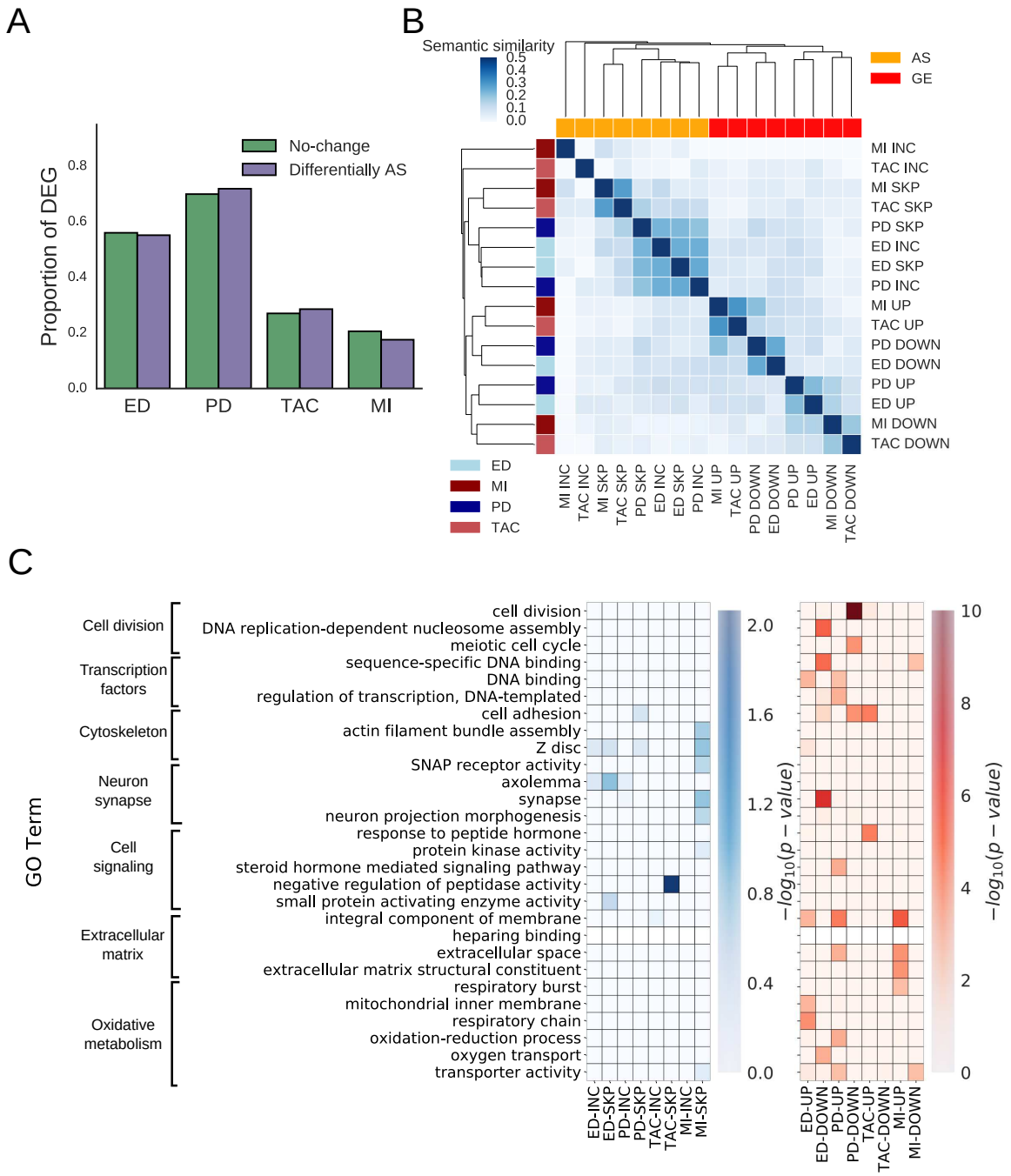

Figure 4.5: Functional impact of AS and GE changes and their overlap. **A**, Proportion of genes showing GE changes that also show changes in AS in each comparison. **B**. Pairwise semantic similarity among the most representative GO terms in each group of genes. Row colors represent the different comparisons studied and column colors represent AS or GE. Semantic similarity profiles were then clustered using hierarchical clustering. **C**, Heatmaps representing the –log10(p-value) of the functional enrichment analysis for top enriched categories across all groups for AS (left) and GE (right).

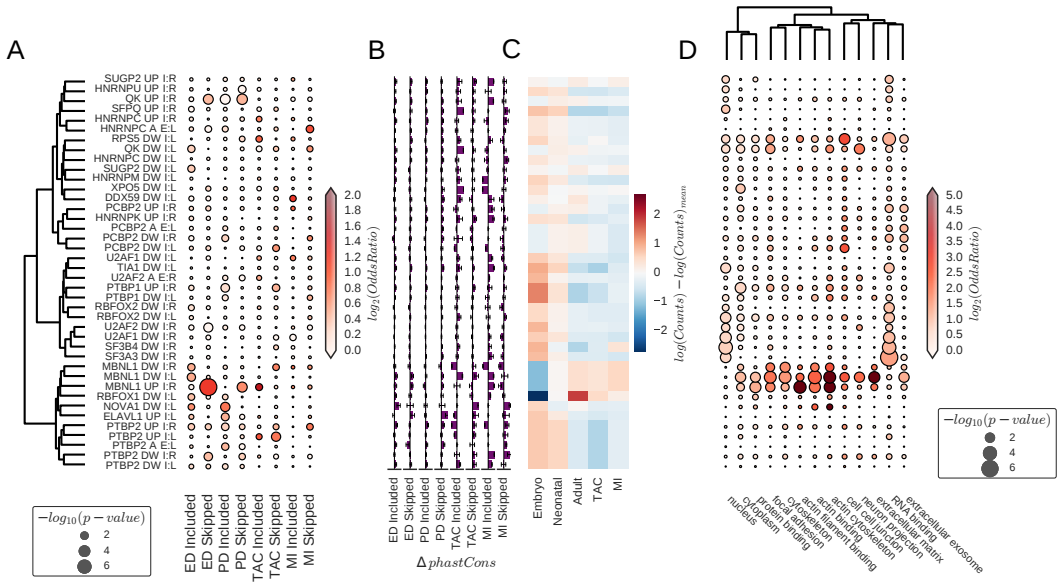

Figure 4.6: Direct regulation of alternative splicing changes. **A** Dotplot representing multivariate enrichment analysis for CLiP-seq binding sites in regulatory regions for RBPs that showed significant enrichment by univariate analysis for each group of exons. Regions are defined by combinations of the following terms: (I: Intron, A: Alternative exon, DW: downstream, UP: upstream, L: left, R: right). The dendrogram was calculated using the Ward method with distances between binding sites profiles to the exons included in the analysis. **B** Mean difference in phastCons scores in binding sites of exons showing changes compared to those that remained unchanged. Error bars represent the standard error of this difference. X-axis range from -0.5 to 0.5 in all cases. **C** Centered expression levels per condition under study for RBPs showing significant univariate enrichment from panel. **D** Dotplot representing the functional enrichment of genes with binding sites for each RBP and region and showing significant changes in at least one comparison. Genes with significant AS changes in at least one comparison were used as background for enrichment (one-sided Fisher test). The dendrogram represents distances between GO terms based on the proportion of shared genes using the Ward method.

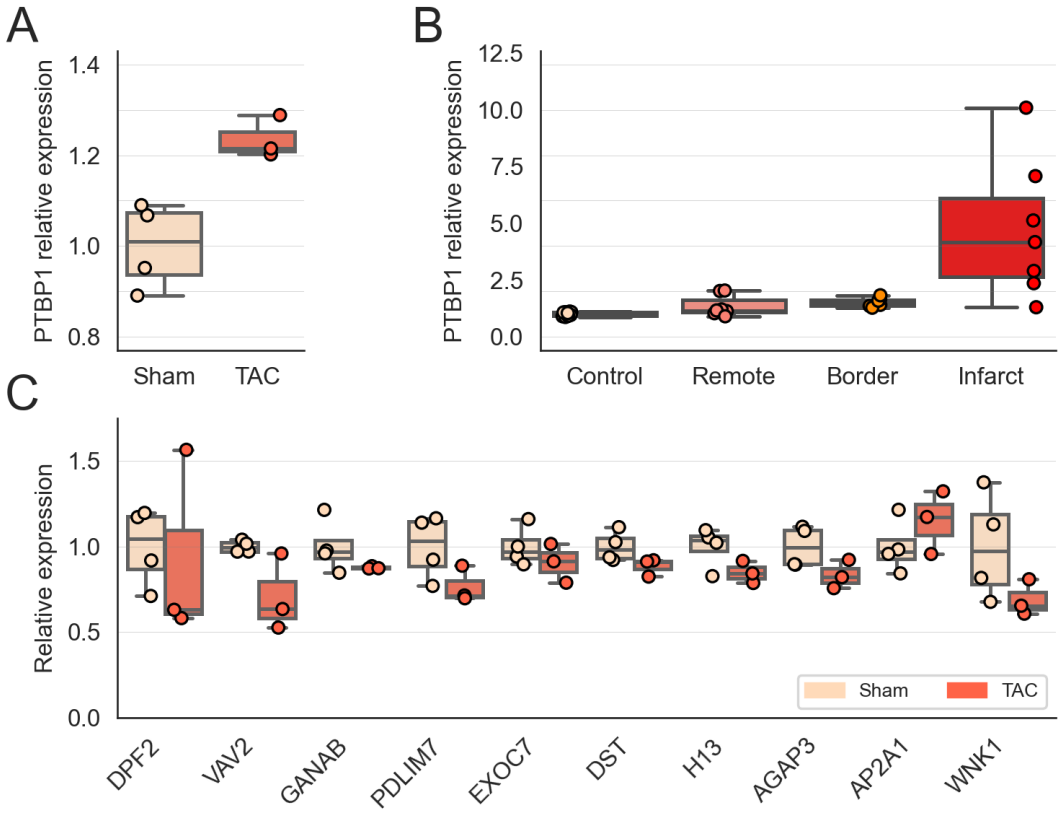

Figure 4.7: Validation of PTBP1 expression changes and predicted targets. **A** PTBP1 expression measured by qPCR in independent samples undergoing TAC and control treatment. **B** PTBP1 expression measured by qPCR in independent samples undergoing MI and control treatment, separating by infarcted, border and remote regions. **C** Relative expression of candidate alternative splicing changes measured by qPCR in the TAC experiment

others, were enriched in the upstream intron of skipped exons in TAC and MI, and included in ED and PD (Figure 4.6A). Additionally, PTBP1 expression decreased during ED and PD, and increased upon MI or TAC (Figure 4.6A), suggesting that its binding to the upstream intron of alternative exon inhibits exon inclusion. Although PTBP1/2 targets were not significantly enriched in any functional category (Figure 4.6D), their enrichment and expression patterns suggest that they mediate the partial re-expression of neonatal AS patters in heart disease.

To validate these findings, we performed independent experiments inducing hypetrophic growth by TAC and MI through LAD ligation. We confirmed that PTBP1 was significantly over-expressed after both treatments, much more so in the infarcted area and the border region separating it from the healthy tissue (Figure 4.7). Moreover, we selected a number of exons with the largest reduction in Ψ in TAC and MI, with binding sites for PTBP1 in the upstream intronic flank, as putative direct targets of PTBP1. Indeed, we found the TAC samples showed decreased inclusion rates compared with controls for all genes except AP2A1, significantly different in 7 out of 10 of them (p-value$\leq$0.1), providen support for PTBP1 upregulation to have an effect on AS patterns.

Regulatory roles for MBNL1 and PTBP1/2 are confirmed by experimental loss of function

To confirm the regulatory potential of the proteins identified above, we used published RNA-Seq data from loss-of-function experiments for MBNL1 and PTBP1/2 (either knock-out (KO) or knock-down (KD))

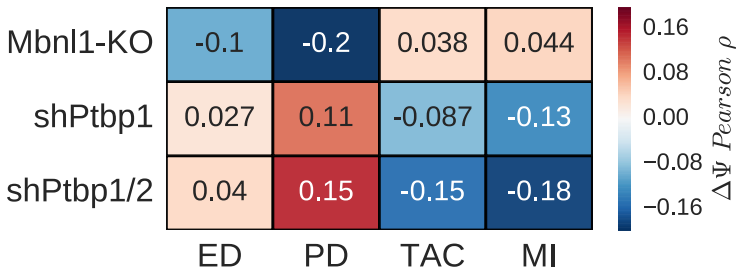

Figure 4.8: Comparison of AS changes in the heart with MBNL1 and PTBP1/2 LOF experiments. Heatmap representing Pearson correlation coefficients between changes in exon Ψ for MBNL1-KO, PTBP1-KD, PTBP1/2-KD and all comparison analyzed in the heart (ED, PD, TAC, MI).

[62, 93, 166, 153] (Figure 4.8). We analysed the correlation of AS changes between these loss of function (LOF) experiments and the conditions under study in the heart. AS changes in the PTBP1/2 double KD correlated with those in the cardiac PD, TAC, and MI contrasts, even though the KD experiment had been carried out in neural progenitor cells. The higher correlation coefficient with the double KD than with the individual PTBP1 KD suggests that PTBP1 and PTBP2 both actively regulate AS in the heart in all the conditions studied, in agreement with the binding-site–enrichment analyses (Figure 4.6). Interestingly, AS changes in the MBNL1 KD correlated with changes in ED, although the correlation was greater with changes in PD, suggesting an effect on AS throughout development that is stronger in the postnatal transition. Consistent with the enrichment results (Figure 4.6), no correlation was found with either TAC or MI, suggesting that MBNL1 does not mediate AS changes in these contexts (Figure 4.8).

Reduced coordination of RBP expression changes is associated with complex regulatory mechanisms of AS in heart disease

We have identified MBNL1 and PTBP1/2 as the main regulatory elements when analysed individually. However, more complex regulatory patterns, such as combinatorial binding of several RBPs, might contribute significantly to the observed AS changes. To tackle this question, we expanded our logistic regression model to include all pairwise combinations of RBP binding sites. As we expect most of these interactions to have no effect on the inclusion patterns, we added a Lasso penalization to promote sparsity of explanatory variables. We first optimized the regularizing constant using 10-fold cross-validation in terms of the AUROC, as a measure of predictive power (Figure 4.9A). As expected, interactions were less likely to contribute to predict inclusion or skipping of a particular exon (Figure 4.9B), even if the magnitude of the coefficients was comparable (Figure 4.9C). However, as there were many more pairwise combinations than single binding of RBPs, we found that the cumulative contribution of interactions to the prediction of inclusion or skipping groups is comparable in ED and PD to that of single RBPs, and this contribution was even higher in MI and TAC (Figure 4.9D). This suggests that the underlying regulatory patterns of AS in heart disease are more complex than those that control AS during development. Figure 4.9E,G show the specific interactions among RBPs that were found to be non-zero for included and skipped exons in ED and MI, respectively. We hypothesized that this increased complexity arises from the lack of coordination in RBPs expression changes. If the expression of two RBPs increased or decreased while maintaining their stoichiometry, they would act mostly in complexes and only their common targets would change. However, if there was an imbalance in their expression levels, both common and individual targets of these RBPs would change. Therefore, if lack of coordination underlied this increased regulatory complexity, we would expect a lower correlation among RBPs in TAC and MI than

in ED and PD. We found comparable normalized correlation coefficients among candidate RBPs in ED and PD to that of pairs of interacting proteins from Intact. The correlation was however lower in MI and TAC (Figure 4.9G). Furthermore, within each comparison, pairs of RBPs showing non-zero regulatory interactions at the binding site level showed lower correlation at the expression levels, consistent with our hypothesis (Figure 4.9H).

PTBP1 over-expression induces cardiac hypertrophy and diastolic disfunction

Our results suggests that PTBP1 upregulation mediates at least some of the AS changes in heart disease. To investigate whether PTBP1 upregulation alone, potentially through AS regulation, is sufficient to induce cardiac hypertrophy, we over-expressed PTBP1 specifically in the heart using an Adeno-Associated virus type 9 (AAV9) as vector carrying PTBP1 cDNA and injected in 10-12 weeks old wild-type (WT) mice in two independent experiments. These experiments were performed by Dr. Javier Larrasa, a former researcher in the group. Mice were sacrified after 28 days post-injection, when PTBP1 was over-expression was similar to the one achieved in TAC experiments (1.2 fold, Figure 4.10A). We also observed a significant increase in the expression of markers of cardiac dysfunction (MYH7 and BNP, p-value<0.01), with no significant increase of markers of fibrosis (LOC, COL1A1, COL3A1) (Figure 4.10B). Accordingly, no increase in fibrosis was observed in histological analyses (Figure 4.10C), suggesting that PTBP1 does not contribute to the characteristic fibrosis accompanying pathological cardiac fibrosis, at least up to 28 days after treatment. We also evaluated cardiac function *in vivo* using echocardiography before sacrifice. We found that mice over-expressiong PTBP1 showed an increased normalized cardiac mass (p-value=0.001, Figure 4.10D), particularly on the left ventricular posterior wall. Although no decrease in left ventricle ejection fraction was observed, suggesting normal systolic function, we found a reduction in the E/A ratio (p-value=0.016). Low E/A indicates a de-compensation on the relative contribution of passive and active left ventricle filling or, in more general terms, diastolic dysfunction (Figure 4.10E).

PTBP1 regulates a very small part of cardiac hypetrophy AS changes

To investigate whether PTBP1-driven cardiac hypertrophy was mediated by the previously characterized AS, we measured the relative isoform expression of candidate targets by qPCR, as in the previous section, and found significant changes in only 2 of them (PHDB1 and DST, p-value<0.05). We did not even observed an average reduction across all of them as in TAC, supporting that only few of the AS changes occurring in cardiac hypetrophy are mediated by PTBP1, and potentially mediate the development of cardiac hypetrophy and dysfunction (Figure 4.10F).

To investigate this issue more in depth, we performed RNA-seq of a reduced number of samples, to characterize PTBP1 mediated AS changes at a transcriptome-wide scale and confirm this trend. One of the mice treated with PTBP1-AAV9 showed no over-expression, and PCA showed a global transcriptomic pattern very similar to control samples, suggesting that it did not actually reach an over-expression of PTBP1, and was subsequently removed from the analysis (Figure 4.11). Even if we did not find any significantly changed gene, estimated $log_2(FC)$ were highly correlated (Pearson $\rho = 0.6$) with those observed in TAC, suggesting that PTBP1 over-expression recapitulates a great deal of expression changes induced in pathological cardiac hypetrophy models, even in absence of fibrosis (Figure 4.12A,B).

We found a large number of exon skipping events showing significant differences between groups, mostly increased upon PTBP1 over-expression. These changes, however, showed little relationship, even at the quantitative level, to those observed in TAC (Pearson $\rho = 0.1$). Even if this may be partly due to more noisy estimation of $\Delta\Psi$s than for GE measures, it indicates that only a very small part of AS changes in TAC is actually driven by PTBP1 (Figure 4.12C,D). To ensure that these AS changes are likely regulated by PTBP1, we selected exons with an estimated $\Delta\Psi$ larger than 0.1 and lower than

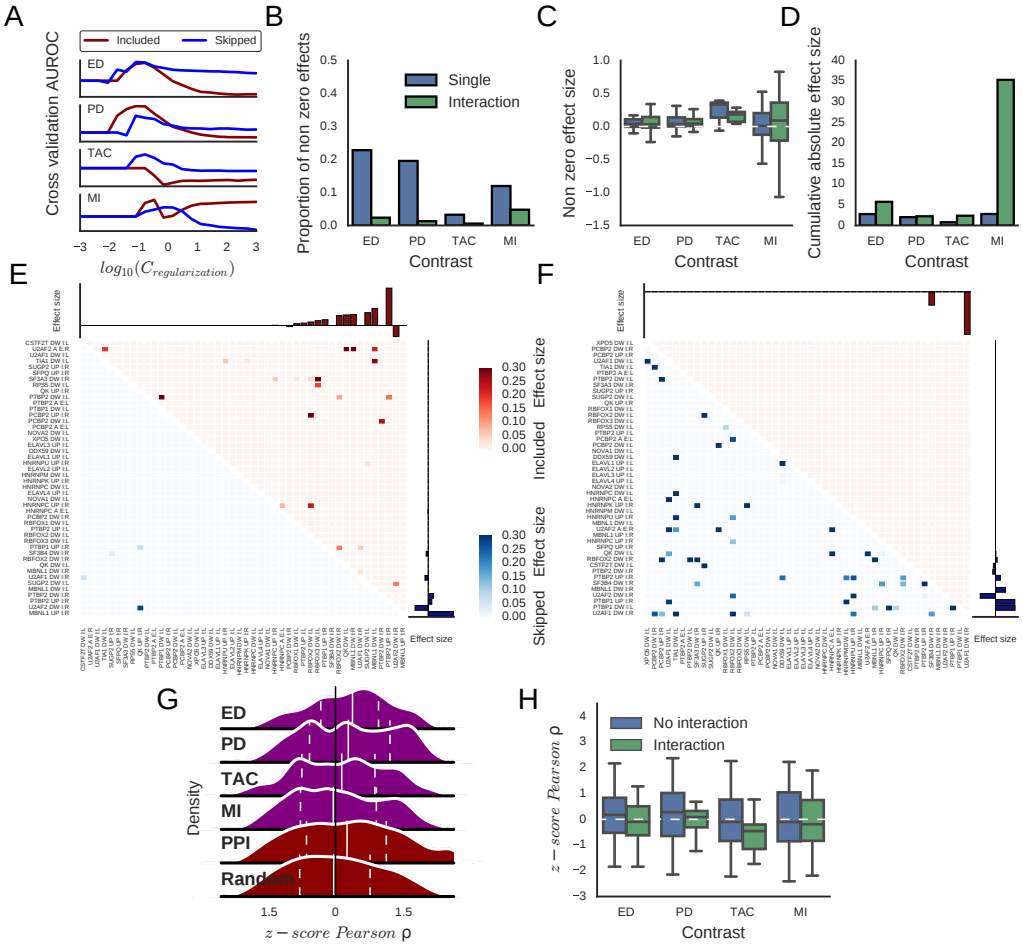

Figure 4.9: Analysis of AS regulatory complexity using regularized logistic regression with all pairwise combinations of RBPs binding sites. **A** area under the Receiver Operating Characteristic curve (AUROC) in 10-fold cross-validation analyses along a range of regularizing constants ($C_{regularization}$; the lower, the stronger the regularization), used to select the value with strongest predictive power of a particular group of exons. **B, C, D** Proportion of non-zero estimations (B), coefficient estimates (C), and cumulative absolute values of coefficient estimates (D), for coefficients corresponding to a single RBP and to an interaction between a pair of RBPs for each comparison under study. **E, F** Heatmap representing the estimate of the coefficients for each combination of RBPs for exons that are either included (Red) or skipped (Blue) for ED (E) and MI (F). Barplots represent the estimation of the coefficient for single RBPs. **G, H** Distribution of normalized correlation coefficients between expression levels of RBPs included in the regression model for each comparison (ED, PD, TAC, MI) and for pairs of interacting proteins and randomly selected pairs of genes as a whole (G) and separating pairs that showed non-zero coefficient (H).

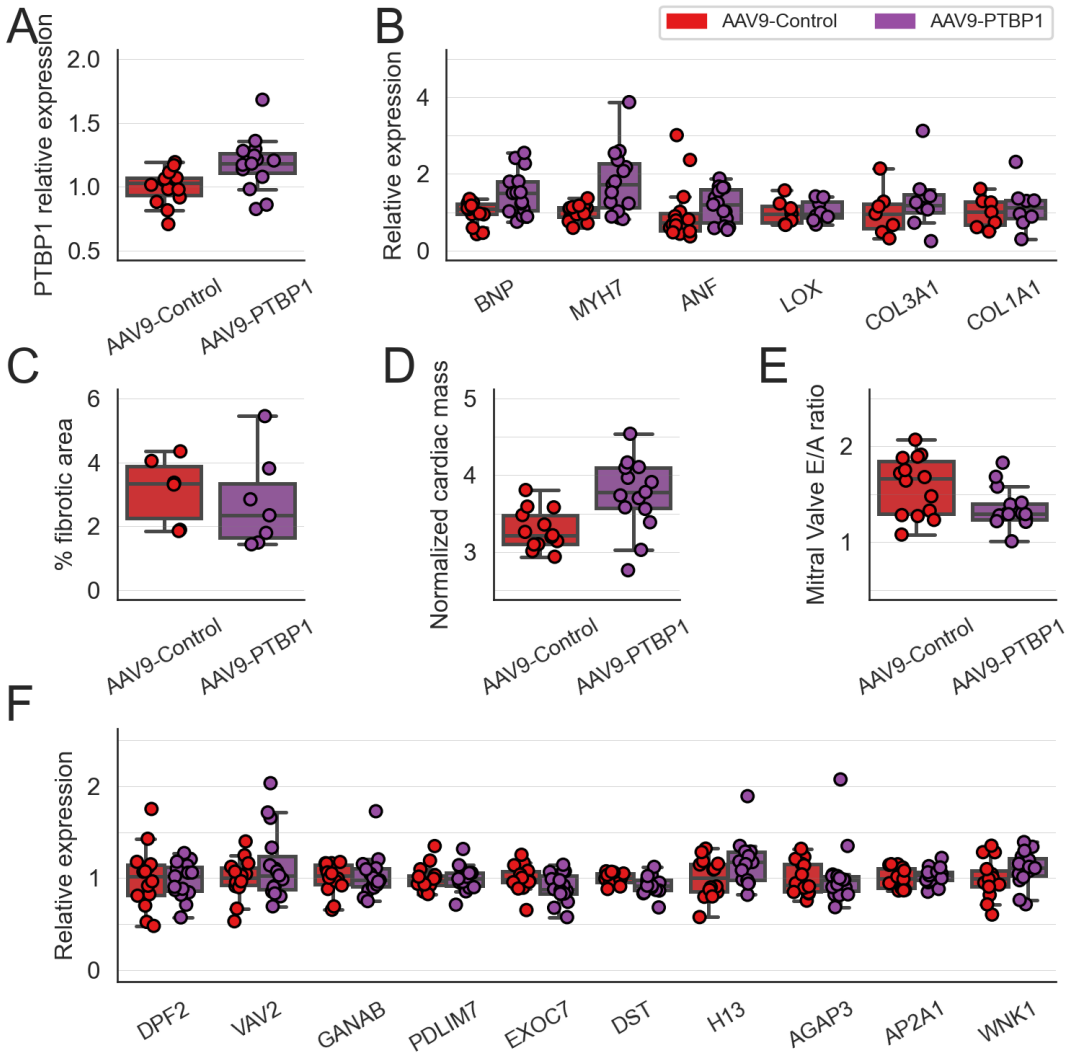

Figure 4.10: Phenotypic characterization of mice over-expressing PTBP1 using AAV9 vector. **A** PTBP1 expression measured by qPCR in mouse hearts injected with PTBP1-AAV9 and control virus. **B** Expression of markers cardiac dysfunction, hypetrophy and fibrosis measured by qPCR. **C** Percentage of fibrotic area in histological cuts of mouse hearts. **D** Normalized cardiac mass derived from echocardiography analysis. **E** Ratio of E to A flow velocities through the mitral valve assessed by echocardiography. **F** Relative expression of candidate alternative splicing changes measured by qPCR

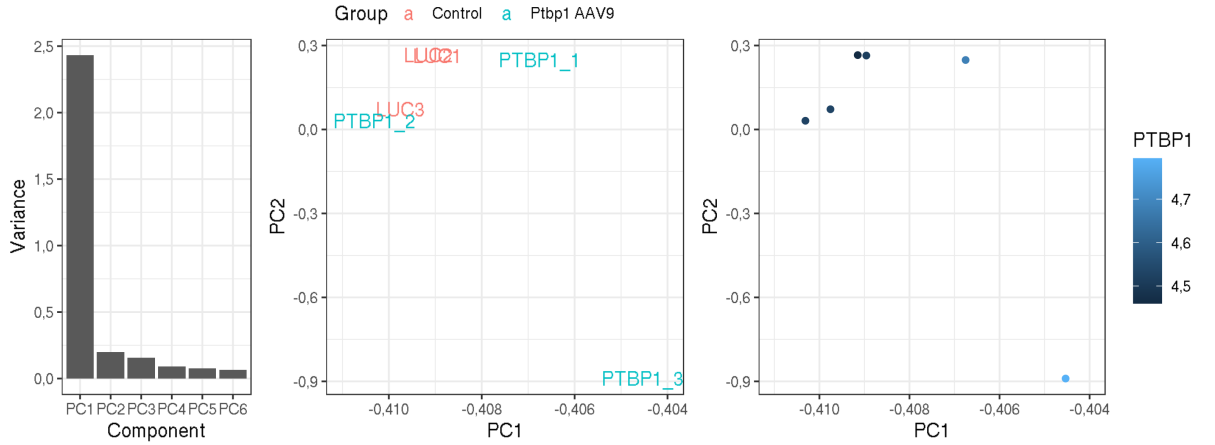

Figure 4.11: Transcriptomic characterization of mice over-expressing PTBP1 using AAV9 vector at the GE level by PCA. **A** Variance explained by each principal component. **B,C** PCA representation of samples according to the treatment group (B) and PTBP1 expression as measured in the RNA-seq (C)

-0.1, as included and skipped, respectively, and plotted the distribution of CLiP-seq binding sites along relevant regulatory regions, centered at the target exon. As in TAC and MI, skipped exons showed an enrichment of binding sites in the upstream intronic flank. However, we also observed a lower frequency of PTBP1 binding sites across the included exons, suggesting that binding of PTBP1 to the target exon actually enhances its inclusion (Figure 4.12E). Finally, we performed functional enrichment analysis on GO categories of genes undergoing GE and AS changes and found that upregulated genes are mostly associated to immune response, whereas downregulated genes are more associated to mitochondria and respiration. As before, we found different gene categories associated to AS changes: whereas genes with included exons were associated to microtubules and muscle cell devlelopment, genes with skipped exons were weakly related to regulation of cell contraction, more directly related to cardiac function (Figure 4.12F). Overall, our results suggests that over-expression of PTBP1 is sufficient to induce cardiac hypertrophy and may thus contribute to the underlying mechanisms of the disease. However, it seems unlikely to be mediated by global AS changes, which showed little similarity to those previosly observed in TAC. Whether this is actually mediate by only a few of the AS changes or other or other PTBP1-dependent effects remains to be elucidated.

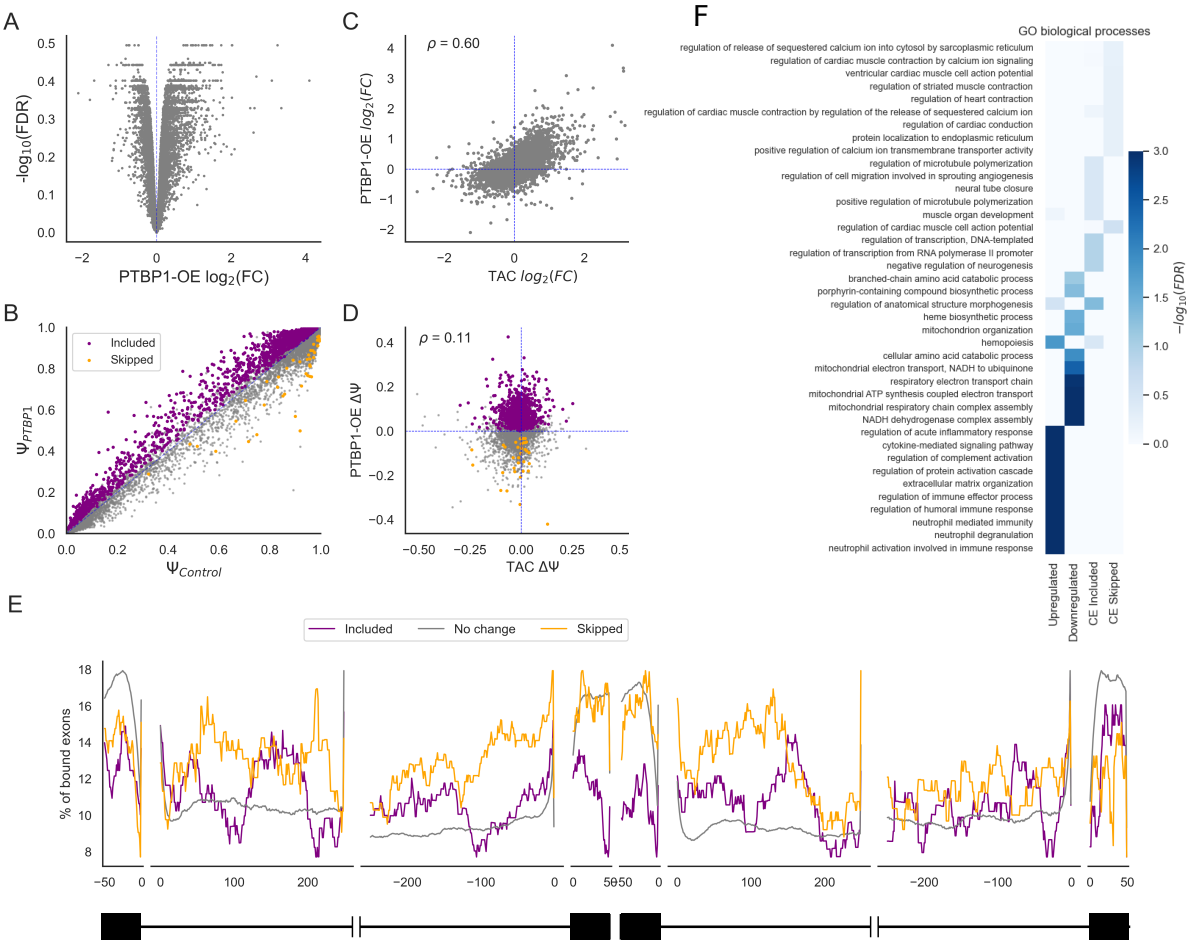

Figure 4.12: AS and GE characterization of mice over-expressing PTBP1 using AAV9 vector at the GE.

4.2 dSreg: A Bayesian model to integrate changes in AS and RBP activity

4.2.1 dSreg rationale for regulatory analyses

dSreg models simultaneously the changes in the Ψ of the whole set of AS events rather than inferring independently their changes fitting an independent model for each event. This allows modeling changes between conditions as latent variables that depend on the actual parameters of interest: the regulatory activity of the trans-regulatory proteins or RBPs. The key modeling assumption regards the relationship between AS changes and the regulatory activities. We assumed an additive model of regulatory effects θ_j on the $logit(\Psi)$ of RBPs bound to each AS event, given known binding preferences of each RBP to each exon k ($S_{k,j}$). We also add an additional ϵ_k term to model AS changes that are not explained by the regulators included in the dataset.

$$\beta_k = \Delta logit(\Psi)_k = \sum_{j=0}^{J} \theta_j S_{k,j} + \epsilon_k \tag{4.1}$$

An additional assumption made by dSreg is that the variance between individuals of the same conditions is common across all AS events in the dataset, since RNA-seq experiments are usually done with few samples, hindering the inference of an event-specific individual variance σ^2. This approach allows more robust estimation of teh variances for more reliable inference of the changes between conditions as previously shown for gene expression analysis [229, 171].

Under these assumptions, we can derive the joint posterior probability of the regulatory activities and changes in inclusion rates given the observed data across the whole transcriptome and explore this probability distribution through Markov Chain Monte Carlo (MCMC) sampling. This sample from the posterior allows calculating the expectation of the regulatory activities and provides an idea of the uncertainty of the inferences. We can easily calculate the probability of such activity to be higher or lower than certain threshold to prioritize regulatory candidates for experimental testing.

4.2.2 Evaluation using simulated data

To evaluate the performance of dSreg we simulated data under the same model and under a typical experimental design of 3 samples per condition. As we are interested in the evaluation of the statistical methods, rather than simulating heavy RNA-seq data with their technical artifacts, we directly simulated the number of reads supporting inclusion or skipping, assuming that read assignment and counting is comparably done across all methods. If our assumptions are realistic and hold in a real scenario, the comparison with other methods will provide a realistic idea about the advantages of dSreg against previous methods. We simulated data under different scenarios to evaluate the influence of different factors, like sequencing depths and regulatory sparsity on the performance of the methods. Since dSreg performs both inference of the regulatory activities and changes in the exon inclusion rates, we evaluated it from the two points of view.

Adding information about regulatory elements improves the detection of AS changes even at low sequencing depth

We first evaluated the performance standard GLM, as the one used by rMATS [249], for the detection of changes in AS at different sequencing depths (λ). From the quantitative point of view, we analyzed the correlation between the estimated $\hat{\beta_k}$ and the real β_k used for the simulations. This correlation was

generally low and did not increase with sequencing depth, suggesting that the limitations to correctly estimate AS changes may no lie on sequencing depth under this model, but on other factors like the limited sample size: with as few as 3 samples per condition, it is very likely that, even if we estimate perfectly the Ψ for a given sample with very high sequencing depth, it will hardly resemble the differences between the two populations.

Sometimes, we are not so much interested in the exact quantification of the change and only require selecting a reliable set of altered events. At this qualitative level, we can analyzed the performance of the GLM using classification metrics. At low sequencing depths ($log(\lambda) \leq 3$), the sensitivity at a 5% False discovery rate (FDR) was smaller than 10%, this is, we only detect about 10% of the true changes. As λ increased, so did the sensitivity (Fig. 4.13B). However, one may reach very high sensitivity by simply selecting every event as altered. Of course, this is not very reasonable, as it will come with a high degree of false positives. The F1 score is the harmonic mean of sensitivity and specificity and can provide a metric for taking into account these considerations. Interestingly, the F1 score saturated with depth (4.13C), suggesting that after some point, there was not much gain in performance by increasing sequencing depth. This may be because although sensitivity continuously increases (4.13B), differences between the groups of samples arisen by chance due to small sample sizes also become more reliable, decreasing the specificity. At this point, we may benefit more from including more samples in our experimental design, rather than increasing sequencing depth (Fig. 4.13C).

To avoid the need to select an arbitrary threshold to assess the performance of the different methods, we additionally calculated the Receiver Operating Characteristic (ROC) curves for each simulated dataset and the area under them (AUROC, Fig. 4.13D and E). These results showed that, at low sequencing depths ($log(\lambda) < 3$), the performance was rather poor, with AUROC values of 0.7 at most.

In order to check whether potential improvements of dSreg were due to the inclusion of binding sites and changes in RBPs activity in the model or just to variance pooling, we ran dSreg and a reduced model that only pools variance from all exons without taking into account of the binding sites and changes in regulatory activities (*Null model*). We defined as significantly changed events as those with a posterior probability higher than 95% of having a $\beta_k > 0$. The *Null model* already outperformed the GLM at the single exon level and improved quantitative estimation of β_j with depth (Fig 4.13A). However, dSreg showed a much greater improvement in correlation and sensitivity, even at very low sequencing depths ($log(\lambda) < 3$), when there was practically no information from individual events (Fig. 4.13). This increased sensitivity did not come with a decrease in specificity as could be expected, since it showed also very high F1 scores and AUROC, suggesting that differences in performance are intrinsic to the method and not threshold dependent (Fig. 4.13C,D and E). Results with the *Null model* suggest that pooling variance across events does only marginally improve the inference of splicing changes, at least with the low variance used in these simulations. dSreg, in contrast, additionally used the information about the underlying regulatory mechanisms to correct differences that may easily arise by chance in datasets with limited sample size, given that simulations were done with only 3 samples per condition.

dSreg improves the detection of the RBPs driving AS changes

Once AS changes have been identified, we focused on the detection of the regulatory elements potentially controlling these events. Using our simulated datasets, we compared dSreg with the traditional ORA and GSEA approaches. As FDR<0.2 filtering showed higher F1 score in the identification of splicing changes (Fig. 4.13C), we used this threshold to select significantly changed events to perform the downstream enrichment analyses. The dependency of Over-representation Analysis (ORA) on the detection of significant changes led to low F1 scores for GLM results at any tested FDR threshold, especially at low sequencing depths (Fig. 4.14A). We also used an in-house version of GSEA to take advantage of quan-

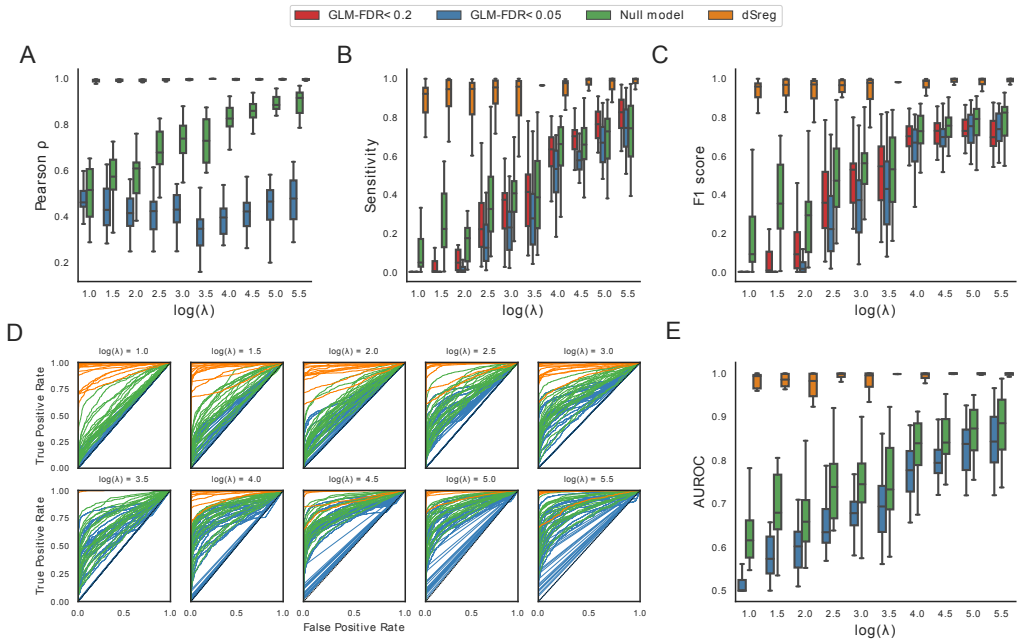

Figure 4.13: Comparison of the performance for the identification of different event inclusion rates of a standard method using a single GLM per exon considering two FDR thresholds (0.05 and 0.2),a bayesian model that pools variance across all exons (*Null model*) and dSreg. Performance was analyzed in simulations with increasing sequencing depths λ (the mean of the Poisson distribution used to simulate the total number of reads mapping to an exon skipping event). **A**. Pearson correlation between real and estimated β_j. **B** Sensitivity. **C** F1 score. **D, E** Receiver Operating Characteristic (ROC) curves (D) and the area under them (E)

.

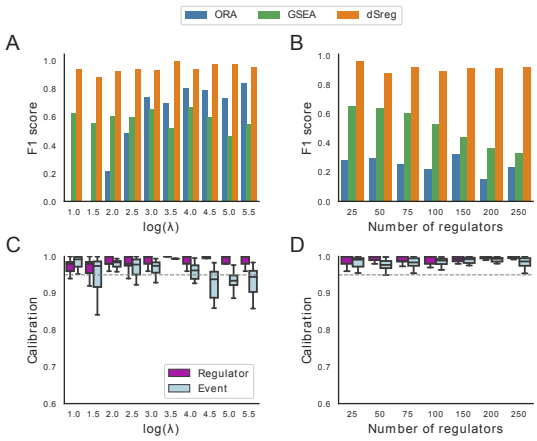

Figure 4.14: Performance of methods for the detection of regulatory elements: ORA with variable FDR thresholds (0.05 and 0.2), non-parametric GSEA and dSreg. Performance was analyzed in simulations with increasing sequencing depths λ, which is the mean of the Poisson distribution used to simulate the total number of reads mapping to an exon skipping event. **A, B**. Mean F1 scores obtained with different depths λ (A) and total number of regulators (B) for the different enrichment approaches. **C, D** Calibration, measured as the proportion of times the real value lies within the CI of differentially spliced exons and regulatory elements for increasing sequencing depth (C) or increasing number of total regulatory elements (D).
.

titative information in the identification of regulatory elements. Briefly, events were ranked according to their Maximum Likelihood Estimation (MLE) of the coefficient of the GLM, which represents the log of the odds ratio of inclusion between the two conditions. Then, we looked for non-random distributions of binding sites along the ranked list [264] (see Methods section for details). We found a substantial improvement over ORA, with higher F1 scores, especially at low sequencing depths, but did not seem to benefit from higher sequencing depths (Fig. 4.14A). dSreg outperformed both ORA and GSEA at every evaluation metric, and was barely affected by low sequencing depths (Fig. 4.14). Therefore, integration of the two sources of information improves results both in terms of inference of differential inclusion rates and the identification of the mechanisms driving those changes.

Increasing the number of potential regulatory elements does not decrease dSreg performance

We have so far used simulated data to explore the effect of sequencing depth on both the detection of splicing changes and on the identification of the key RBPs driving these changes. We next assessed the impact of the number of regulators, which may increase the number of false positives, particularly in presence of co-linearities among binding profiles of different RBPs. To study this potential limitation, we simulated datasets with only 5 active RBPs as in the previous simulations, but increasing the number of total RBPs included in the analysis up to 250. We found that the F1 score tended to decrease as the number of potential regulators increased with either ORA or GSEA, despite multiple test correction to control false discovery rate. Once more, dSreg outperformed both methods and remained unaffected by the inclusion of other inactive regulatory elements (Fig. 4.14B).

Model calibration remains robust while decreasing the proportions of active RBP

We further analyzed the performance of dSreg in terms of calibration. A model is well calibrated when inferred probabilities actually represent the real frequency of a given phenomena i.e. a model is calibrated

when the uncertainty of the parameter estimate matches the evidence contained in the data. Calibration was calculated as the proportion of events and regulators whose real change in logit-transformed inclusion rates (β_k) or activity (θ_j) is within the estimated CI. Whereas changes in inclusion rates were well calibrated, the uncertainty of the changes in the activity of RBPs seemed to be slightly overestimated, given that CI included the real values more often than 95% of the times, independently on the sequencing depth λ (Fig. 4.14C). We then tested how different numbers of total regulatory elements affected model calibration with the previous simulations using only 5 active out of an increasing number of candidate RBPs. We found that the total number of candidate regulators had no effect on calibration (Fig. 4.14D). These results suggest that dSreg is conservative when estimating the uncertainty of the regulatory activities θ_j based on the data, since the real value is within the CI more often than expected across all tested conditions (Fig. 4.14C,D).

4.2.3 Evaluation using real data

dSreg outperforms other methods using real data

To assess whether the better performance of dSreg could be confirmed with independent real data, we used an RNA-seq dataset (around 120M reads per sample) for which a subset of AS events were quantified using RASL-seq and can be used as gold standard [315]. We used CLiP-seq data of a number of RBPs binding to upstream and downstream flanks of exon skipping events as regulatory features for dSreg [64]. Since dSreg performed particularly better than other methods at low sequencing depths, we subsampled the sequencing reads by a factor of 2 up to 512 to analyze the extent of this advantage. We analyzed the data also with MISO, BRIE and DARTS. Both BRIE and DARTS use prior information to improve detection of splicing changes [128, 315, 107]. dSreg and the *Null model* showed the best performance, compared to all other methods, except in extremely low coverages (dilution factor ¿ 100), in which DARTS overcame dSreg (Figure 4.16A,B). In contrast to the results obtained from the simulated data, dSreg and *Null model* performed similarly, which suggests that the regulatory features that were added do not contribute much to the estimation of AS changes. However, it also shows that it remains robust to the inclusion of non-relevant regulatory features. Neither BRIE nor DARTS outperformed the *Null model*. We observed the same patterns when comparing the results to the full coverage RNA-seq dataset (Figure 4.15).

The main advantage and motivation of dSreg is the inference of the regulators driving AS changes, a feature that is not provided by any of the existing tools for AS analysis. To assess whether dSreg outperforms ORA and GSEA also with real data, we used the collection of RBP knock-down experiments from ENCODE [203]. Although it is difficult to know the actual regulatory mechanisms in each case, one may reasonably assume that at least some of the AS changes would be mediated by the down-regulation of the target RBP. dSreg detected the highest percentage of knock-down RBPs as regulatory elements compared to the random expectation in each case (Figure 4.16C). If the expression of other regulatory element is affected by the perturbation, we would expect them also to contribute to explain AS changes. Regulators detected by dSreg tended to be more often differentially expressed in the same experiment than expected by chance compared to other methods (Figure 4.16D). Finally, we observed that, when sorting the regulators by their evidence, the RBP that was knocked-down tended to appear higher in the ranking produced by dSreg than in those yielded by ORA and GSEA (Figure 4.16E and F, respectively). Altogether, these results suggest that dSreg also outperforms previous methods in the identification of regulatory elements using real data.

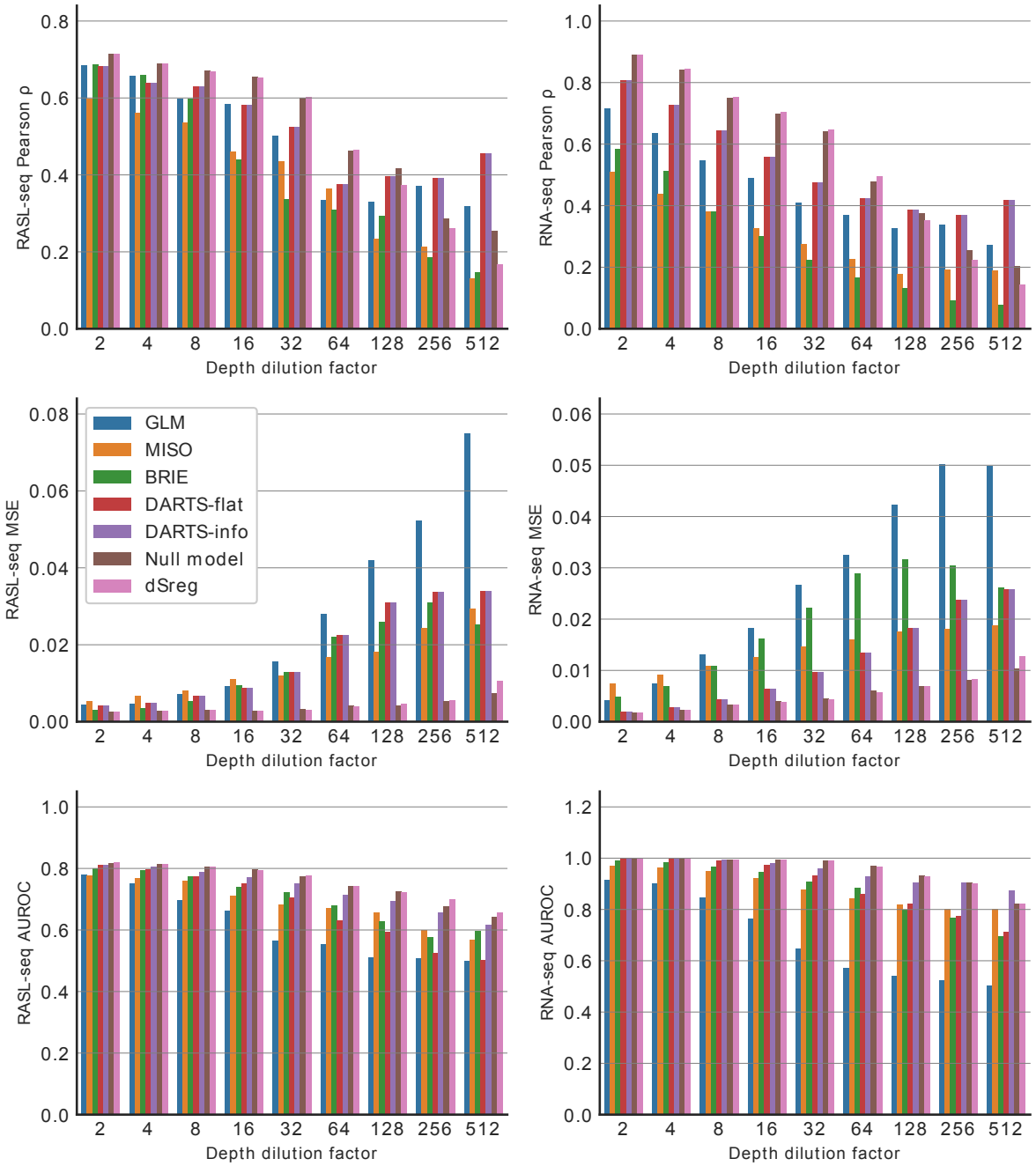

Figure 4.15: Evaluation of the identification of AS changes of dSreg with other methods on real data. Performance of differential splicing methods using RASL-seq quantifications (left column) and full coverage RNA-seq (right column) as true values, measured by Pearson correlation of $\Delta\Psi$, mean squared error (MSE) and AUROC. The different measures are represented in different rows

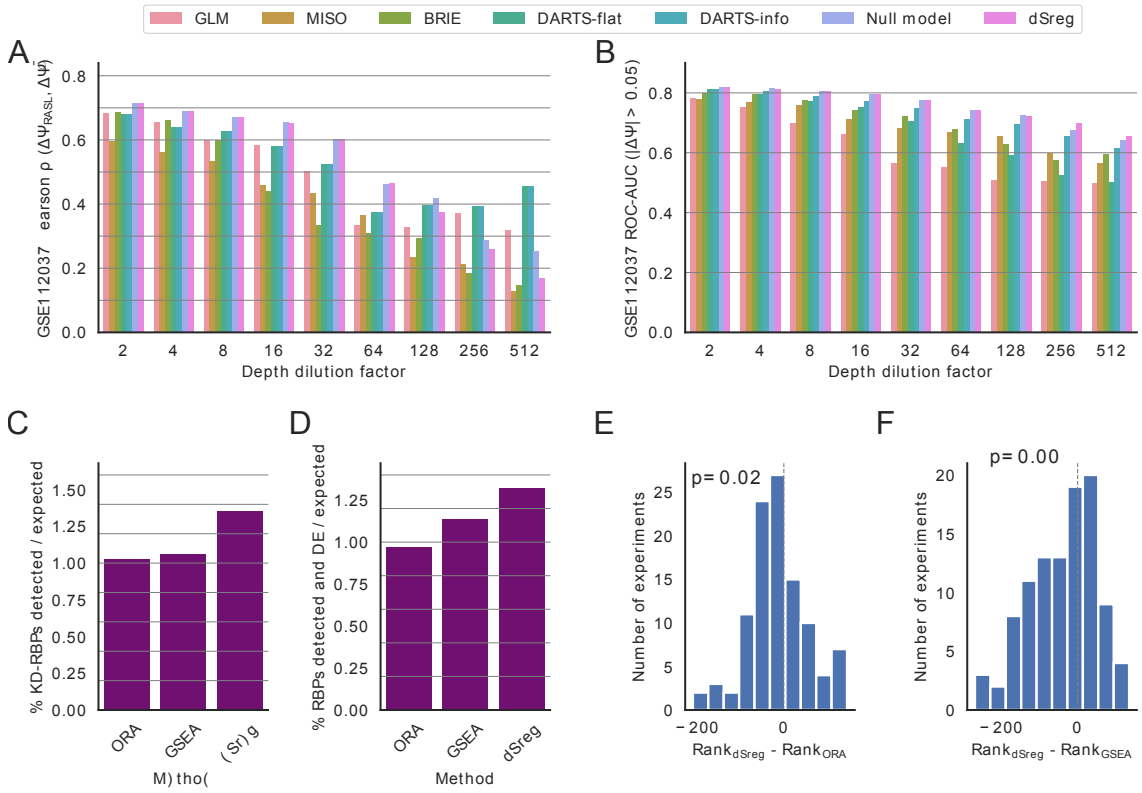

Figure 4.16: Evaluation of the performance of dSreg with other methods on real data. **A,B** Performance of differential splicing methods using RASL-seq quantification as true values, measured by Pearson correlation of $\Delta\Psi$ (A) and area under the Receiver Operating Characteristic curve (AUROC) for exons significantly changed, defined as those with a $|\Delta\Psi| > 0.05$. Methods include a GLM. MISO, BRIE, DARTS with and without using the predictions as prior (info and flat respectively), and dSreg and its *Null model*. **C** Percentage of experiments in which the knocked-down RBP was found among the regulatory elements compared to expectation. **D** Percentage of regulatory RBPs identified by each method that were detected to be differentially expressed (DE) compared to expectation. Expectations were calculated by 20000 random sampling of the same number of regulators. **E, F** Difference in rank occupied by the knocked-down RBP in the output of dSreg with that of ORA (E) and GSEA (F).

4.2.4 Analyzing alternative splicing regulation in cardiomyocyte differentiation

We then tested our model on a dataset of mouse cardiomyocyte differentiation from cardiac precursors (GSE59383) with 3 samples per condition as in our simulated scenario. Binding sites for a number of RNA binding proteins were obtained from CLiP-seq experiments and only those located in the upstream and downstream intronic flanking 250bp were used (see Extended Methods section for details). We run the 3 approaches explored in this work and found that ORA resulted in a high number of significantly enriched candidates, most of which are likely to represent false positives as in our simulation analysis (Fig. 4.17A). GSEA, on the other hand, showed no significant enrichment at $FDR < 0.05$, and only a few at nominal $p < 0.05$, which suggest that these p-values can easily arise by chance. Indeed, there is little concordance with results from ORA (Fig. 4.17A and B). dSreg did show an overall agreement with ORA results, but, as expected, dSreg provided a reduced number of RBPs whose combined action best explain the observed AS changes (Fig. 4.17, Table S1). Interestingly, a great deal of the identified regulatory RBPs are considered to be members of the core spliceosome (BUD13, EFTUD2, PRPF8, SF3A3, SF3B4), suggesting that changes in the activity of these particular components might be key for the AS changes underlying cardiomyocyte differentiation. In this regard, the core spliceosomal machinery has been shown to have extensive regulatory potential [211] and mutations in one of these genes (EFTUD2) have been associated with congenital heart defects, among other phenotypes [167].

4.3 Comparative study of exon inclusion rates across mammals

In this last section, we aimed to study how AS quantitatively changed during mammalian diversification. To do so, we will deal with Ψs as quantitative characters and use models of phenotypic evolution along a phylogenetic tree to characterize the underlying evolutionary process and the optimal inclusion rates across mammalian species. Using these models, we investigate not only the genetic forces driving AS evolution, but also the extent to which exon skipping is functional or contributes to lineage-specific adaptations.

4.3.1 Adapting models of phenotypic evolution for alternative splicing data

Quantitative traits may be assumed to derive from small contributions of a very large number of loci across the genome. Thus, as mutations affecting the trait accumulate at a constant rate, one also expects that mean value of the trait in the population diverges from the ancestral population. This can be approximated for long evolutionary times by a continuous time stochastic process known as Brownian motion (BM) model, in which the infinitesimal change in the trait value is proportional to a normal distribution with an infinitesimal variance. The proportionality constant τ^2 describes the evolutionary rate, this is the rate at which variance accumulates in an evolving population over time.

$$dX = \tau^2 dW$$

Based on this differential stochastic equation, we can derive the trait distribution after a finite period of time t, which may be a branch of the phylogenetic tree, with starting trait value X_0.

$$X \sim Normal(X_0, \tau^2 t)$$

There are numerous scenarios in which trait variance accumulates at a constant rate over time, including not only evolution by pure random forces i.e. mutation and drift [146], but also constant directional

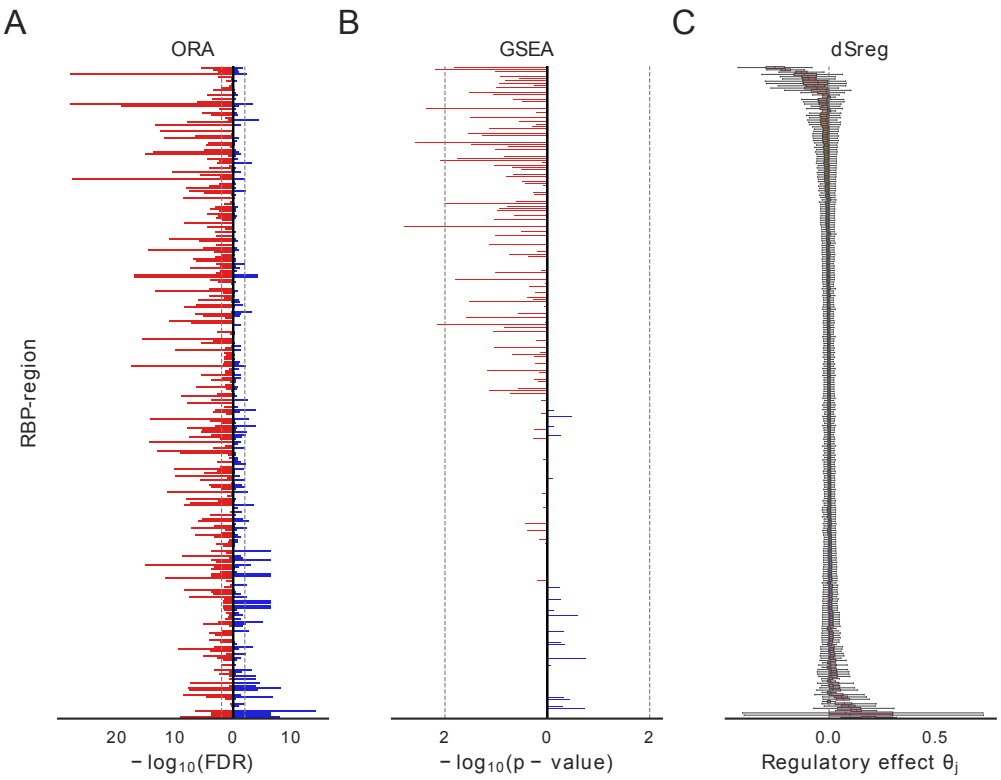

Figure 4.17: Comparison of ORA, GSEA and dSreg using a real RNA-seq dataset from a cardiomyocyte differentiation experiment. RBPs on the y-axis are sorted for the three panels according to the posterior mean of the regulatory effect θ_j inferred by dSreg. **A**. Candidate regulatory proteins derived from the ORA on the significantly included (blue) or skipped (red) exons represented by their significance expressed as the log transformation of the FDR. **B**. GSEA results represented by the nominal empirical p-value resulting from permuting the exon labels. RBPs with positive enrichment scores are represented on the right, and those with negative scores on the left. **C** Posterior distributions of the regulatory effects θ_j inferred by our model.

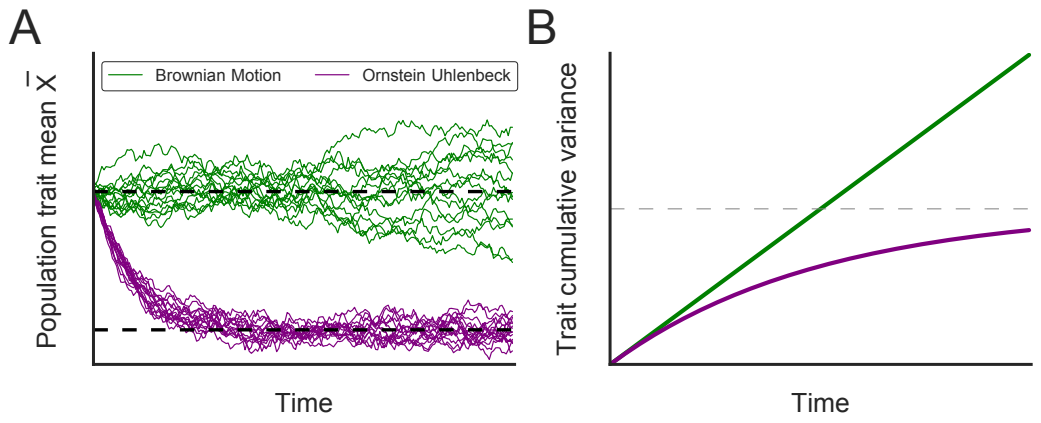

Figure 4.18: Traits evolving under Ornstein-Uhlenbeck (OU) and Brownian motion (BM) models over time (A) and how the variance of the expected distributions changes with time from the ancestral population (B)

selection [104]. The BM model can be extended with a pull towards a certain value μ with strength α in an Ornstein-Uhlenbeck (OU) model (Figure 4.18):

$$dX = \tau^2 dW + \alpha(X - \mu)$$

This pull, which can be explained by stabilizing selection towards a single optimal value μ [146], prevents the trait from accumulating variance indefinitely over time and reaches an equilibrium distribution in which the pull compensates all the variance that tends to accumulate over a short period of time (Figure 4.18).

$$X \sim Normal\left(X_0 e^{-\alpha t} + \mu(1 - e^{-\alpha t}), \frac{\tau^2}{2\alpha}(1 - e^{-2\alpha t})\right)$$

Trait X here is assumed to be continuous and unbounded. Inclusion rates (Ψs) are continuous values, but are bounded between 0 and 1. Thus, to be able to use these models to describe Ψ evolution, we conveniently transformed the Ψ by taking the commonly used logit transformation, and used the transformed variable as the evolving trait X. Hence, from now on, we are assuming that the underlying evolving trait is not the Ψ directly, but its logit transformation, which represents the log transformation of the odds ratio, or ratio of inclusion to skipping probabilities.

$$X = logit(\Psi) = log\frac{\Psi}{1 - \Psi}$$

Nonetheless, we do not quantify X or Ψ directly, but through counting the number of reads supporting exon inclusion I from a total number of reads T mapping to the ES event, naturally following a Binomial distribution depending on Ψ. Thus, if we want to estimate model parameters taking into account the uncertainty or error when estimating Ψ, we can simply expand our model with an additional binomial layer, taking into account a bias term that models the systematic biases introduced by the transcript structure, sequencing conditions and fragment size distribution specific of each sample (see Methods section for details)

$$I \sim Binomial(T, InvLogit(X + bias)) \tag{4.2}$$

Despite our models working on the unbounded logit scale, we will use the models and the inferred parameters to predict how evolution would take place at the Ψ scale to facilitate the interpretation of the

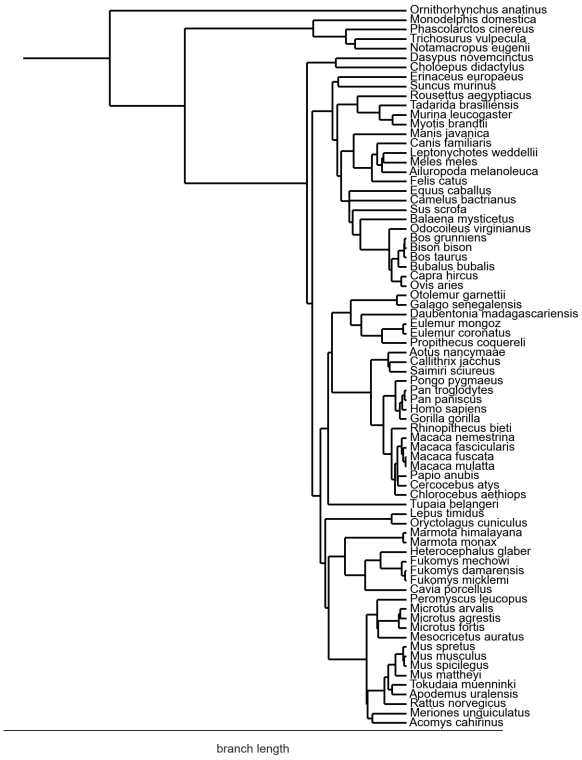

Figure 4.19: Phylogenetic tree of the species with liver RNA-seq data included in the study. Divergence times were taken from previous work [282]

results. In this sense, we can define the optimal inclusion rate Ψ_{opt} as the inverse logit transformation of the optimal value in the logit scale μ

$$\Psi_{opt} = InvLogit(\mu) = \frac{e^{\mu}}{1 + e^{\mu}} \tag{4.3}$$

4.3.2 Characterization of exon inclusion evolutionary rates with Brownian motion models

We collected a large number of RNA-seq datasets from livers of a total of 76 mammalian species with known phylogenetic relationships (Figure 4.19 [282]). To compare Ψs across different species, we first needed to identify sets of orthologous exons. For this, we used previously characterized sets of orthologous genes in mammals from OrthoMaM database [244], and derived 1 to 1 orthology relationships among exons composing those genes. Out of over 14000 sets of orthologous genes, we were able to identify 170400 sets of orthologous exons, characterized in at least 4 species. We the filtered a total of 27065 exons with sufficient coverage and some evidence of skipping (see Methods section for details) as a high quality set of exons for studying the evolution of their inclusion rates.

To obtain a first characterization of the evolutionary process underlying the quantitative changes in exon inclusion rates Ψ during mammalian diversification, we fitted a Brownian motion (BM) to each exon independently and calculated the posterior expectation of the evolutionary rate $\hat{\tau}^2$, i.e. the amount of variance accumulated per time unit. We found a unimodal distribution of evolutionary rates centered at around $\hat{\tau}^2 = 0.39$, with widespread variation across different exons, of about 2 orders of magnitude (Figure 4.20A). Naturally, exons showing higher intra-species variability have also a trend to evolve faster, even if the correlation between the two variables is only $\rho = 0.25$ (Figure 4.20B). We observed that exons from the same gene tended to have more similar $\hat{\tau}^2$ values, as clearly shown when comparing

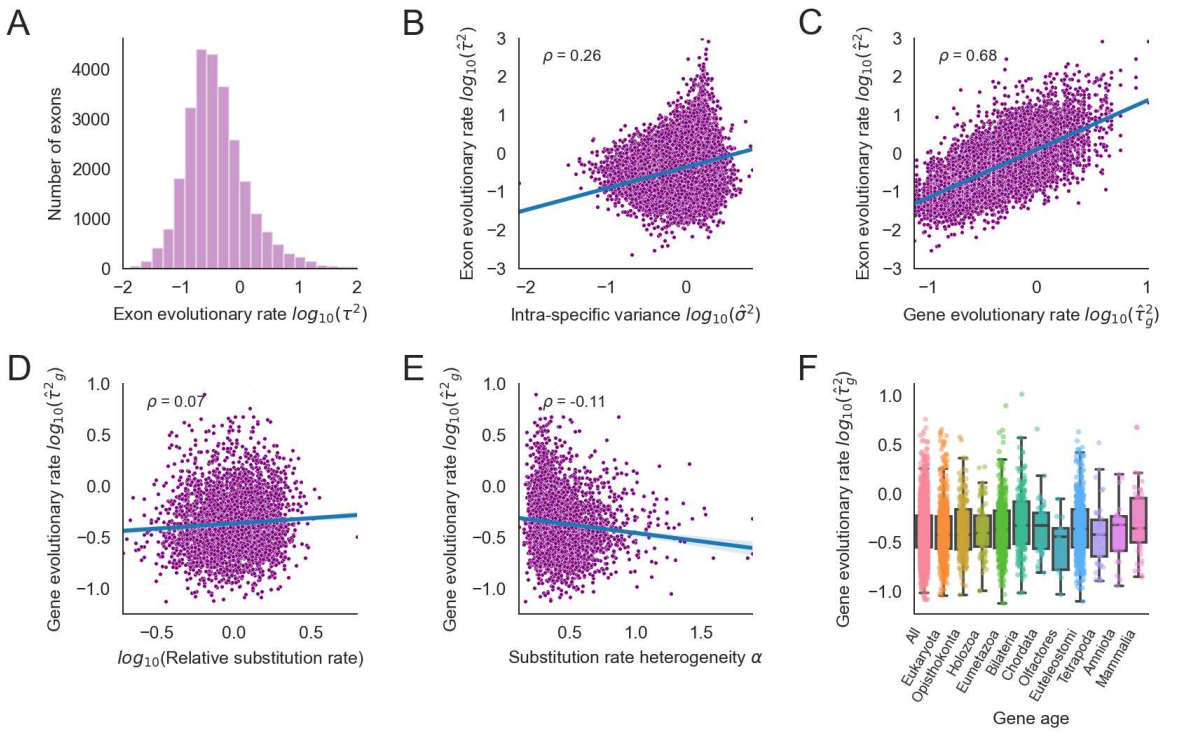

Figure 4.20: Evolution of exon Ψ under a BM model. **A** Distribution of exon inclusion evolutionary rates under a BM model in units $logit(\Psi)$ variance accumulated per million years (my). **B,C** Scatter plots showing the association of the inferred exon evolutionary rates $\hat{\tau}^2$ with within species variance σ^2 (B) and gene-wise evolutionary rate $\hat{\tau}^2_g$ (C). **D,E** Scatter plots showing the association of the inferred gene evolutionary rates $\hat{\tau}^2_g$ with gene relative substitution rates (D) and gene-wise substitution rate heterogeneity α (E). **F** Boxplots showing the distributions of evolutionary rates across genes with different phylogenetic ages.

them against the average across exons of the same gene (Figure 4.20C). Thus, an important part of the variability in the evolutionary rates of exon Ψ lies at the gene level. We next wondered whether the evolution of splicing rates is related with sequence evolution. We found a correlation of 0.07 between rates of evolution of exon inclusion rates and nucleotide substitution (Figure 4.20E) and of -0.11 between $log_{10}(\tau^2)$ and α_{Gene}, a measure of heterogeneity in the evolutionary rates across positions in the gene (Figure 4.20E). Therefore, even if the correlation coefficients are very small, indicating that only a small amount of variability is driven by these variables, genes with faster and more homogeneous evolutionary rates along their sequence tend to evolve slightly faster at the exon Ψ level.

Intron expansion and increased exon skipping events have been associate to particular evolutionary events during metazoan evolution [98, 97], particularly linked with the origin of bilaterians and changes in the genome architecture. To investigate whether genes originated at different points during evolution have evolve differently, we compared the rates of evolution for genes described to have originated at different time points in the past [168]. We found that, of all gene groups, the Ψs of bilaterian-old genes have evolved the fastest (Figure 4.20). These differences remained significant after simultaneously accounting for different gene properties using a multiple linear mixed model framework (See Table 4.1). Moreover, we additionally found a significant increase in evolutionary rates of boroeutherian and mammalian genes that may have been hidden by confounding factors.

Table 4.1: Gene and exon variables associated with the rate of evolution of Ψ in mammals in a multiple regression framework

Variable	Coefficient	p-value
log(Downstream Intron length)	-0.011	2.47e-05
$\log(\sigma^2)$	0.202	4.00e-186
Gene relative substitution rate	0.111	4.50e-10
Gene substitution rate heterogeneity	-0.453	7.74e-22
Amniota	0.09	3.46e-01
Bilateria	0.10	2.35e-03
Boroeutheria	0.30	7.10e-03
Chordata	0.07	3.26e-01
Eukaryota	0.01	7.92e-01
Eumetazoa	0.05	4.03e-02
Euteleostomi	0.11	4.77e-06
Holozoa	-0.02	6.35e-01
Mammalia	0.27	1.80e-04
Olfactores	-0.23	1.02e-01
Opisthokonta	0.04	2.43e-01
Tetrapoda	0.00	9.81e-01

4.3.3 Studying the contribution of stabilizing selection using Ornstein-Uhlenbeck models

Estimation of the contribution of selection to inclusion rate evolution

Although the evolutionary rate in a BM model provides an estimate of the average rate at which quantitative characters evolve, there are a number of reasons that suggest that a BM model may not completely describe the evolution of exons Ψ. First, an exon with an starting $\Psi = 0.99$ evolving at the average rate of $\hat{\tau}^2 = 0.39$, as estimated, will very quickly evolve low inclusion rates (Figure 4.22A). Thus, even if every ancestral mammalian exon was originally included at $\Psi_0 = 0.99$, about half of the exons will be included at very low rates after 250 my, with about 18% of them with $\Psi < 0.01$. This hardly resembles the observed Ψ distributions in any species (Figure 4.22C). Moreover, since splice sites are highly constrained due to their relevance to the splicing reaction [14], one would expect the resulting quantitative trait i.e. exon inclusion rate, to evolve under stabilizing selection.

The Ornstein-Uhlenbeck (OU) model is a generalization of the BM model, in which the average of the quantitative trait in the population does not only accumulate random variation, but also experiences a pull towards an optimal value. The increased number of parameters in the OU model hinders parameter inference for each exon independently with only 76 different species [232, 58]. However, one can assume that every exon evolves under the same regime and can be used as independent samples of the same underlying OU process to perform accurate parameter inference. As performing Markov Chain Monte Carlo (MCMC) on the complete dataset was too computationally demanding, we performed inference on increasing numbers of randomly selected exons and found that parameter estimates were already stable using up to 2000 exons (Figure 4.21A-E, Table 4.2). To investigate whether estimates were not only stable, but unbiased, we simulated data under an OU model with the inferred parameter values, and found that inferences converged to the true values similarly to random samples of exons (Figure 4.21F-J). The strength of selection in our model is represented by the phylogenetic half-life ($t_{\frac{1}{2}}$), the time required to reduce the distance to the optimal value by half. We inferred $\hat{t}_{\frac{1}{2}} = 25.72$ million years, supporting a weak but non-negligible selective force constraining exon Ψ around a common optimal value, in average. Thus, as with the BM model, we can use the inferred parameters to predict the evolution of exon Ψs across the genome from an ancestral $\Psi_0 = 0.999$, derived from the average optimal value. Now

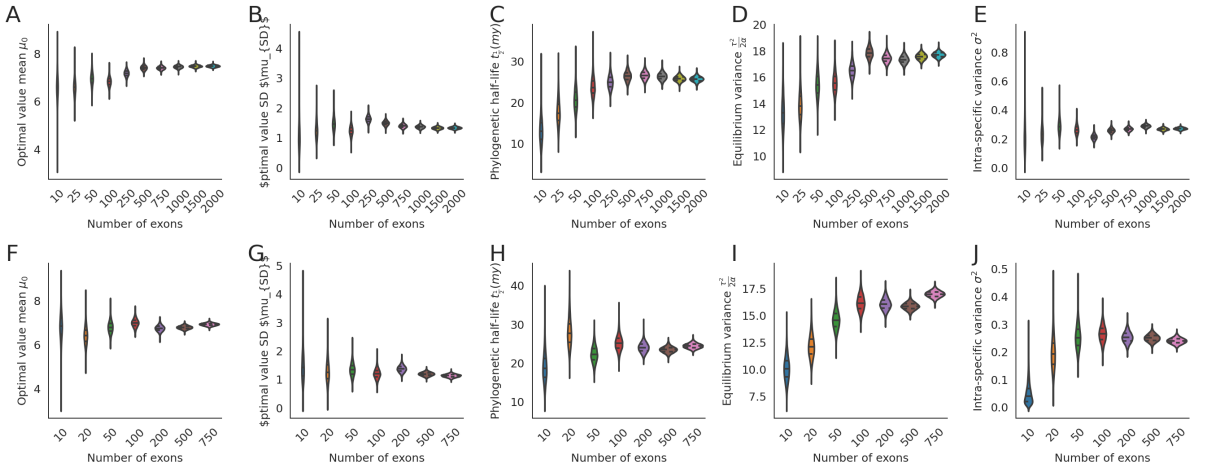

Figure 4.21: Inferred parameters of the OU for an increasing number of randomly selected exons. **A-E** panels show the posterior distribution for parameters inferred using the real dataset, with sets of up to 2000 exons. **F-J** show the posterior distribution of the same parameters inferred from simulated data under the previously inferred parameters with the biggest dataset. Simulations comprised sets of exons with increasing size up to 1000.

Table 4.2: Parameter estimation of an Ornstein-Uhlenbeck model for exon Ψ on a sample of 2000 exons

Parameter	Expectation	CI
$t_{\frac{1}{2}}$	25.72	[24.38, 27.01]
σ^2	0.27	[0.26, 0.29]
$\frac{\tau^2}{2\alpha}$	17.71	[17.03, 18.14]
μ_0	7.48	[7.40, 7.56]
μ_{SD}	1.22	[1.26, 1.41]

we predict that even with an $\Psi_{opt} \sim 1$, we expect a certain amount of sub-optimal inclusion rates in any species, with about 11% exons with $\Psi < 0.9$ once steady state is reached in about 100 my (Figure 4.22B). We then compared the predicted Ψ distributions by the BM and OU models from a common ancestral Ψ_0 distribution derived from the distribution of optimal values after 250 my of evolution with the estimations across the 76 different species under study (Figure 4.22C)). The OU model predicts much better the average observed patterns across the different species than the BM model. In summary, our results suggest that natural selection is limiting divergence of exon inclusion rates, but is not sufficiently efficient to enforce optimal inclusion rates for every exon in the genome.

Evolutionary forces underlying variability in Ψ evolutionary rates

The OU model allows inference of the relative contribution of selective and random forces to the evolution of quantitative traits. Thus, it allows, not only inferring the average behaviour across exons, but also to study what is their relative contribution to the highly heterogenous evolutionary rates that were inferred under the BM model (Figure 4.20A). To do so, we stratified exons according to the exon-level inferences of τ^2, and used samples of 200 exons from each group to infer evolutionary parameters under an OU model. We found comparable or even shorter phylogenetic half-lifes for fast evolving exons (Figure 4.23A), but very different equilibrium variances (Figure 4.23B). Thus, variability in evolutionary rates across exons is mainly driven by variation in the neutral evolutionary rates, rather than by differences in selective strength. To better understand the nature of these differences, we used the quantitative genetics parametrization of the OU model [146] based on the width on the fitness landscape w^2 and effective

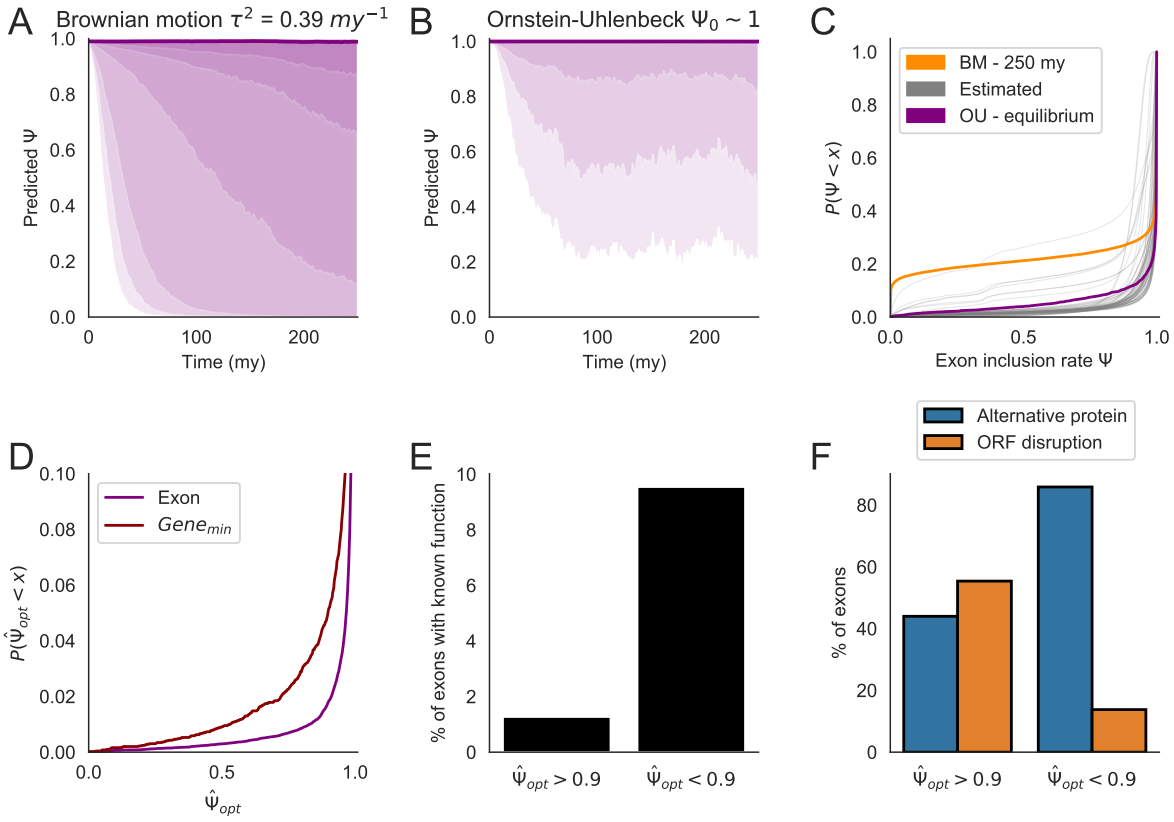

Figure 4.22: Evolution of exon Ψ under a OU model. **A** Predicted exon Ψ evolving under a BM model with an average $\tau^2 = 0.40 my^{-1}$ as inferred from the data. **B** Predicted Ψ evolution under the inferred OU model parameters with an optimal inclusion rate $\Psi_{opt} \sim 1$. Different degrees of shade represent, in order the 2.5, 5, 10, 25, 35, 40 percentiles in both A and B. **C** Ψ distributions estimated for the different species under study, as well as the predicted by the inferred BM after 250 my of divergence and the inferred OU model in the equilibrium. **D** Cumulative distribution of the inferred Ψ_{opt} across all gene and exon sets under study. For each gene we used the smallest estimated Ψ_{opt} across their exons. **E** Proportion of exons whose skipping is known to provide a different function from VASTDB depending on their inferred Ψ_{opt}. **F** Predicted protein impact in VASTDB human exons depending on their Ψ_{opt}.

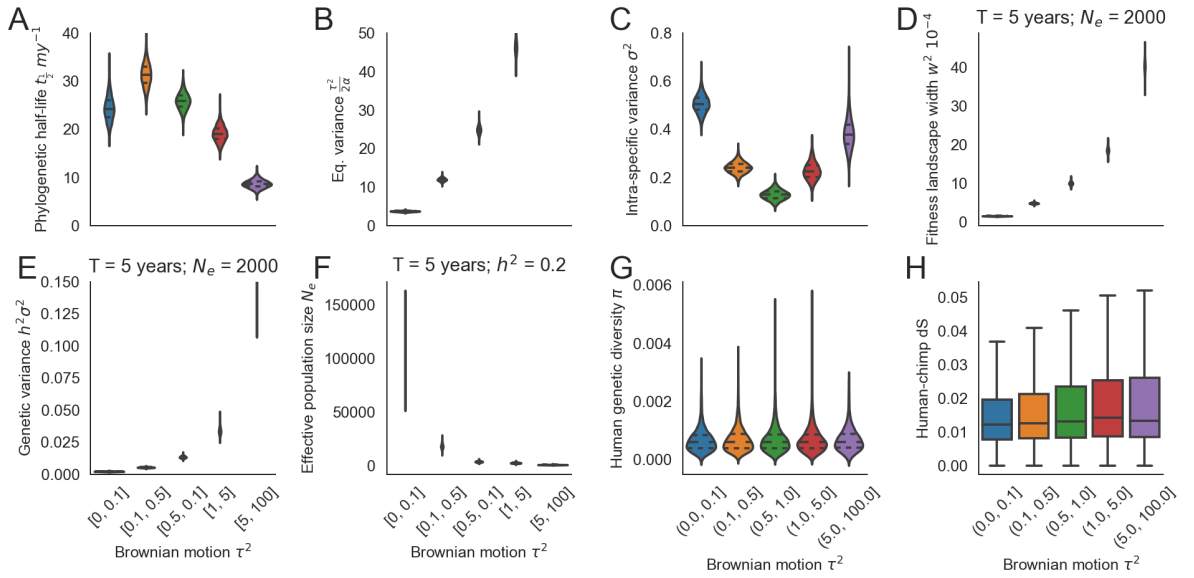

Figure 4.23: Random forces drive variability in average evolutionary rates of exon inclusion rates. Exons were stratified by their previously estimated average $\hat{\tau}^2$ under a BM model and subsets of 200 exons were used to fit OU models **A-F** OU inferred and derived parameters for sets of exons with increasing average evolutionary rates: phylogenetic half-life $t_{\frac{1}{2}}$ (A), equilibrium variance $\frac{\tau^2}{2\alpha}$ (B), within species phenotypic variance σ^2 (C), derived width of the fitness landscape w^2 (D) under a constant effective population size $N_e = 2000$, genetic variance $h^2\sigma^2$ (E) under a constant effective population size $N_e = 2000$, and effective population sizeN_e assuming a constant trait heritability of h^2 and average generation time of 5 years applying quantitative genetic models [146]. **G** Average genetic diversity in 10kb regions around exons evolving under varying rates estimated from allele frequencies annotated in dbSNP. **H** Synonymous substitution rates from human and chimp genomes extracted from Ensembl database for genes with exons with Ψs evolving at variable rates

population size N_e

$$V_{eq}^{OU}(X) = \frac{\tau^2}{2\alpha} = \frac{w^2 + \sigma^2}{2N_e} \tag{4.4}$$

If we assume that the effective population size is the same across all exons or at least with low variation, and arbitrarily set it to 2000, considering the inferred within species variance (Figure 4.23C), slowly evolving exons are characterized by a much narrower fitness landscape around the optimal values and smaller genetic variance (Figure 4.23D,E). These factors compensate each other to yield a similar pull towards the optimal value, as indicated by the phylogenetic half-life (Figure 4.23A), such that the time required to reach optimal inclusion rates remains constant across different average evolutionary rates. However, differences in the equilibrium variance may also be explained by variation in the effective population size N_e. Nonetheless, very large differences in the effective population size within the genome would be required to fully account for observed patterns (Figure 4.23F). To investigate this alternative explanation, we calculated genetic diversity in 10 kb regions around each exon using dbSNP153 common variants in humans, and found no differences across the different groups (Figure 4.23G). As genetic diversity may be influenced by selection, we also used synonymous substitution rates dS between human and chimp as a proxy for N_e. Whereas genes with rapidly evolving exons at the Ψ level were located in genes with higher dS (Figure 4.23H), these differences alone can hardly account for the large variation in observed in the equilibrium variance.

Table 4.3: Gene Ontology function enrichment analysis on those genes with an exon with $\hat{\Psi}_{opt} < 0.9$

GO:id	Description	OR	P-value	FDR
GO:0030054	cell junction	4.31	0.000	0.009
GO:0005096	GTPase activator activity	5.23	0.000	0.030
GO:0043547	positive regulation of GTPase activity	4.50	0.001	0.030
GO:0007155	cell adhesion	3.99	0.002	0.088
GO:0042995	cell projection	2.60	0.009	0.323
GO:0008284	positive regulation of cell proliferation	3.00	0.014	0.342
GO:0000165	MAPK cascade	3.22	0.017	0.342
GO:0016324	apical plasma membrane	3.13	0.019	0.342
GO:0015629	actin cytoskeleton	3.60	0.019	0.342
GO:0007165	signal transduction	1.97	0.020	0.342
GO:0005856	cytoskeleton	1.84	0.038	0.506
GO:0005737	cytoplasm	1.35	0.039	0.506
GO:0006468	protein phosphorylation	2.09	0.040	0.506
GO:0007186	G-protein coupled receptor signaling pathway	2.84	0.043	0.506

Estimating the proportion of mammalian wide functional exon skipping events

To identify optimally alternative exons across mammals, we used the posterior expectation of the OU parameters $t_{\frac{1}{2}}$, $\frac{\tau^2}{2\alpha}$ and σ^2 previously estimated (Table 4.2) to fit an OU model for each exon in our dataset independently, and inferred their optimal inclusion rates Ψ_{opt} as well as the Ψ in each node of the tree. We found 1.98 % of exons in 5.61% of the genes to have $\hat{\Psi}_{opt} < 0.9$, suggesting that, at least in liver, a small amount of genes has at least one exon with optimal intermediate inclusion rates. To validate our approach, only based on comparative Ψ data, we used human exons with experimentally validated functions as annotated in VASTDB v1.8 [268], and found that exons with $\hat{\Psi}_{opt} < 0.9$ have almost 10 fold probabilities of having an annotated function than those with higher optimal inclusion rates (11% compared with 1.2%) (Figure 4.22E). Moreover, mammalian alternative exons were twice as likely to encode an alternative protein compared with the remaining exons 55% of which disrupted the reading frame upon exon skipping (Figure 4.22F). To investigate whether optimal alternative splicing is associated to particular gene functions, we performed functional enrichment analysis using Gene Ontology (GO) categories as annotated in OrthoMaM, and found an over-representation of genes related to cell junction and signaling (Table 4.3).

Our findings suggest that even if a gene is alternatively spliced in a given species, it does not necessarily imply that it will produce functionally different isoforms. Thus, it is interesting to know how much evidence the hypothetical quantitative skipping of an exon provides about the functionality of such AS event. We can approach this question using our model by calculating probability of the Ψ_{opt} for an exon given its Ψ in a single species using Bayes theorem.

$$P(\Psi_{opt}|\Psi) = \frac{P(\Psi|\Psi_{opt})P(\Psi_{opt})}{P(\Psi)} \tag{4.5}$$

Since we have an empirical distribution for Ψ_{opt} and $P(\Psi)$ is constant, we can calculate this probability for any observed Ψ (Figure 4.24A), showing that, even observing a relatively low inclusion rate for an exon, the most probable value of Ψ_{opt} remains very close to 1 (Figure 4.24A). We also calculated the probability of Ψ_{opt} being below certain threshold given an observed Ψ (Figure 4.24B), showing that not only the highest probability value is close to 1, but that it is very unlikely that the Ψ_{opt} falls below a certain threshold. For instance, having observed an exon with a $\Psi = 0.5$, the probability that the actual Ψ_{opt} is higher than 90% is still 44%. Thus, we can never be sure that an exon is optimally alternatively spliced solely based on its inclusion rate, as intermediate inclusion rates can easily derive from optimally

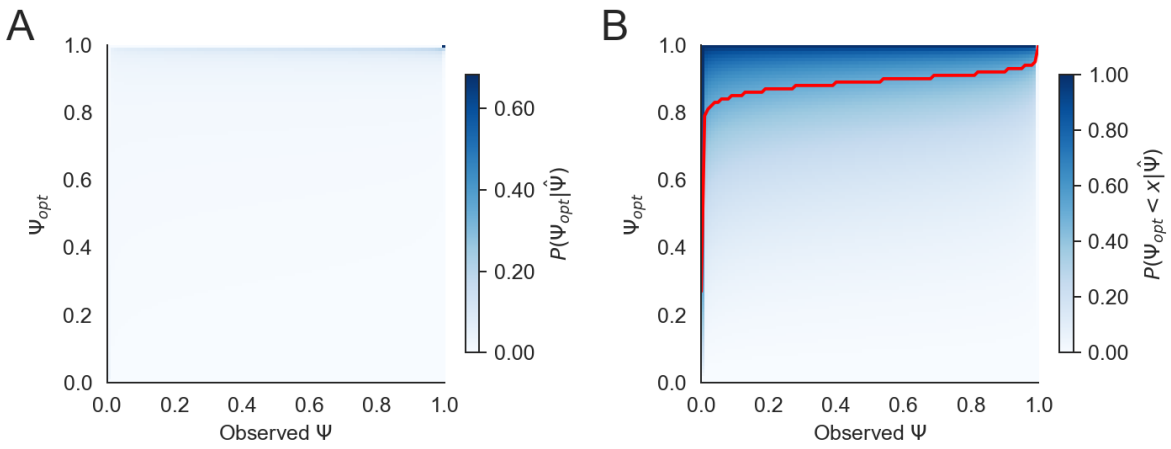

Figure 4.24: Inference of optimal inclusion rate from species inclusion rate in a single exon. **A,B** Heatmaps showing the probability density (A) and cumulative probability (B) of the optimal inclusion rate for a given exon Ψ_{opt} depending on the observed Ψ in a given mammalian species sharing the characterized Ψ_{opt} distribution

constitutive exons that are skipped to some degree, and optimally alternative exons are rare.

Characterization of the evolution of functionally skipped exons

To investigate whether optimally alternative exons evolve under a different regime, e.g. they evolve under different selective constraints or diverge at faster rates; we stratified exons into 5 different groups according to their Ψ_{opt}, and randomly selected a subset of 100 exons for each Ψ_{opt} range. We fitted independent OU models to each group and inferred the parameters that characterize their evolution. We found that the phylogenetic half-life remained comparable among the 5 different groups and always lower that the approximate 25 my inferred with the bulk exons (Figure 4.26A), suggesting that rate variation across exons may influence the inference of the selective strength. In contrast, there was a large difference in the equilibrium variance, which was markedly higher for exons with optimal inclusion rates below 80% (Figure 4.26B). Using these parameters, we can predict how exon Ψ may evolve depending on their Ψ_{opt}. The equilibrium is reached relatively fast across all groups of exons. However, whereas exons with high Ψ_{opt} reach stationary Ψ distributions near this optimal inclusion rate, the equilibrium variance in optimally alternative exons is so large that most exons are expected to have either high or low inclusion rates and only few exons would actually remain close to 50% inclusion (Figure 4.25).

To better understand the nature of these differences, we used the quantitative genetics parametrization of the OU model [146]. If we assume that the effective population size is the same across all exons, and arbitrarily set it to 2000, considering the inferred within species variance (Figure 4.26C), alternative exons would be characterized by a much wider fitness landscape at the logit scale (Figure 4.26D) and larger genetic variance (Figure 4.26E). These differences could also be explained by alternative exons being located in regions with lower effective population size, as shown by assuming a constant heritability and generation times (Figure 4.26F). We found no differences across the different groups in genetic diversity or synonymous substitution rates (Figure 4.26G,H). Again, this suggests that differences in the equilibrium variance are mostly explained by differences in the fitness landscape and genetic variance rather than by differences in the effective population size.

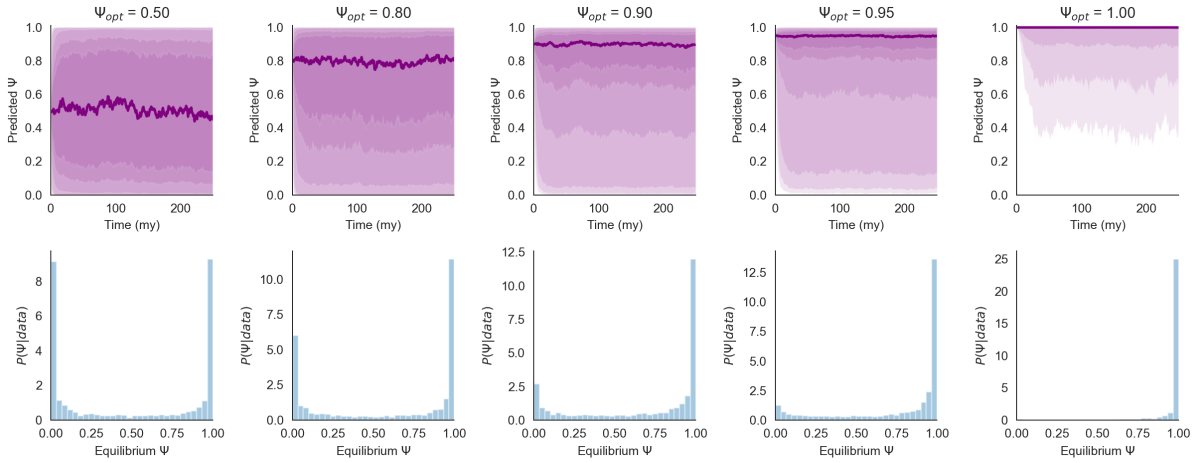

Figure 4.25: Prediction of exon inclusion rate evolution along time under independent OU models across exons with variable optimal inclusion rates Ψ_{opt} (shown in columns) Bottom panels show the expected Ψ distribution in the equilibrium for each optimal inclusion rate. These predictions were made based on the inferred parameters using subsets of exons in different ranges of optimal inclusion rates. Different degrees of shade are shown according to percentiles [2.5, 5, 10, 25, 35, 40, 50, 60, 65, 75, 90, 95, 97.5]

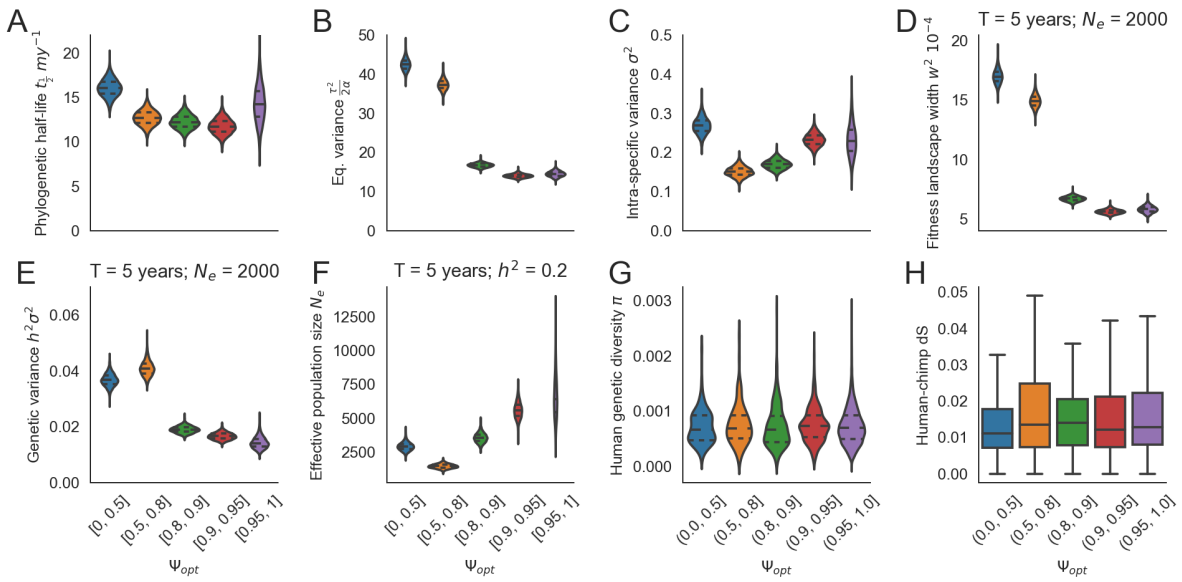

Figure 4.26: Alternative exons evolve under a different evolutionary regime. Exons were stratified by their previously estimated optimal inclusion rate $\hat{\Psi}_{opt}$ under a OU model and subsets of 100 exons were used to fit independent OU models **A-F** OU inferred and derived parameters for sets of exons with different optimal inclusion rates: phylogenetic half-life $t_{\frac{1}{2}}$ (A), equilibrium variance $\frac{\tau^2}{2\alpha}$ (B), within species phenotypic variance σ^2 (C), derived width of the fitness landscape w^2 (D) under a constant effective population size $N_e = 2000$, genetic variance $h^2\sigma^2$ (E) under a constant effective population size $N_e = 2000$, and effective population sizeN_e assuming a constant trait heritability of h^2 and average generation time of 5 years applying quantitative genetic models [146]. **G** Average genetic diversity in 10kb regions around exons with different optimal inclusion rates estimated from allele frequencies annotated in dbSNP. **H** Synonymous substitution rates from human and chimp genomes extracted from Ensembl database for genes with exons with variable optimal inclusion rates Ψ_opt

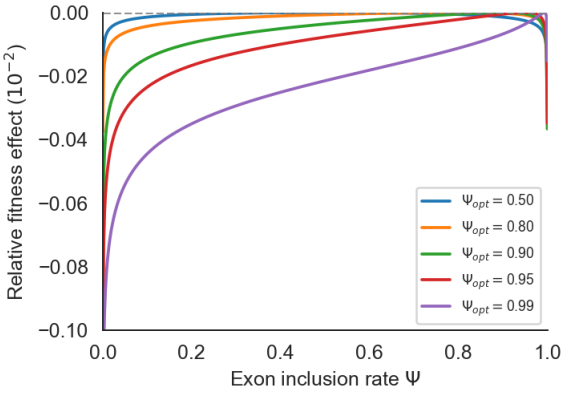

Figure 4.27: Inferred Ψ fitness landscapes under the OU model. Relative fitness associated to Ψ across range of different optimal inclusion rates Ψ_{opt}. This model assumes a quadratic fitness function on the $logit(\Psi)$ with parameter w^2

4.3.4 Alternative splicing fitness landscape

Following quantitative genetics derivation of the OU model, we assume a quadratic fitness function on the $logit(\Psi)$, characterized by a single parameter w^2 that provides an idea about the width of the fitness peak around the optimal value [146].

$$W(\Psi) = 1 - \frac{(logit(\Psi) - logit(\Psi_{opt}))^2}{w^2}$$

Assuming a certain effective population size, constant across different optimal inclusion rates, we can derive w^2 from the inferred equilibrium variance and study how the shape of the Ψ fitness landscape changes depending on the optimal inclusion rate Ψ_{opt} (Figure 4.27).

$$\frac{\tau^2}{2\alpha} = \frac{w^2 + \sigma^2}{2N_e}$$

Interestingly, we found that for optimally alternative exons, e.g. $\Psi_{opt} = 0.5$, the fitness barely changes between 0.2 and 0.8, suggesting that the exact proportions are not very important for their function, as long as there is certain amount of alternative product. In contrast, as the Ψ_{opt} increases, the fitness landscape starts to peak more steeply around the Ψ_{opt}, such that smaller deviations have greater impact on fitness, especially once we get near 100% optimal inclusion. Although the derived magnitude of the fitness effects change depending on the assumed effective population size, the relative shapes of the fitness landscapes remained constant with population sizes between 500 and 20000 (Figure 4.28). These landscapes imply that skipping of a optimally constitutive exon is expected to have greater fitness consequences than fully skipping or including an optimally alternative exon. Despite alternative exons having wider fitness peaks and therefore evolving under lower fitness gradients, the strength of the pull towards the optimal value remains relatively constant with Ψ_{opt} (Figure 4.26A). This can be explained by alternative exons having simultaneously higher genetic variance available for selection, which compensates the differences in the fitness gradient (Figure 4.26E).

4.3.5 Inference of lineage specific shifts in optimal inclusion rates

So far, our model assumes that there is a single optimal value across all mammalian species. If AS patterns are associated to liver function, which has remained similar across species, we expect the same for most exons Ψs. However, optimal inclusion rates in some exons may change in some lineages, for

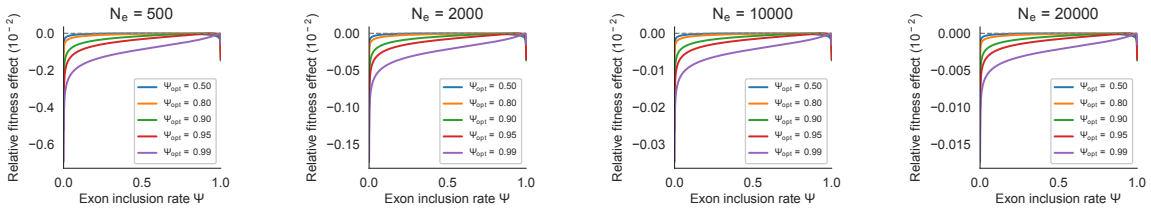

Figure 4.28: Fitness landscape shapes remain invariant to effective population size. Prediction of Ψ fitness landscapes for variable optimal inclusion rates Ψ_{opt} under different population sizes as indicated in each panel

instance, by exon skipping acquiring a new beneficial function in some environments. if we assume that such changes are rare, the inferred OU parameter are not expected to be affected by them. Therefore, using previously inferred parameters (Table 4.2), we can predict how the Ψ will evolve after a change in the Ψ_{opt}, in this case the acquisition of a new function by exon skipping of an original constitutive exon ($\Psi_{opt} \sim 1 \rightarrow \Psi_{opt} = 0.5$) has occurred (Figure 4.29A). These simulations show that the Ψs derived from the newly acquired optimal value are difficult to distinguish from those evolving under the original optimal constitutive inclusion rate.

To investigate this issue in more depth, we simulated data from exons evolving under the inferred OU parameters, but allowing the optimal value to change in each branch with certain probability. In order the infer those changes, we used a single exon-level OU model with known OU parameters allowing branch-specific variations in the optimal values. However, when we compared the simulated changes in Ψ_{opt} with the inferred ones, we find a very low correlation (Pearson $\rho = 0.1$), with high rates of false positives and negatives (Figure 4.29B). To investigate whether the problem was locating the shift in a specific branch of the tree or the identification of a shift across the evolution of a given exon, we assessed the performance of the method at the whole tree level. We calculated the Receiver Operating Characteristic (ROC) curve and calculated the area underneath it (AUROC=0.52), which was just slightly over the 0.5 representing random performance (Figure 4.29C). Although we cannot identify which exons have experienced a shift in the Ψ_{opt} over its evolution, we may still use the known sensitivity and specificity at a certain threshold to try to infer the frequency of those changes across the genome. However, the poor performance of the test for detecting shifts provided virtually no information about the % of exons that have experienced a shift in the simulated data, since the posterior highly resembles the prior distribution (Figure 4.29D).

To investigate whether this limitation is inherent to the method or depends on the parameter space and phylogenetic tree under study i.e. large neutral variance hindering identification of shifts in the optimal values, we simulated data with decreasing equilibrium variance $\frac{\tau^2}{2\alpha}$ and repeated the same procedure. Simulations showed a very different pattern when $\frac{\tau^2}{2\alpha} = 2.5$, as trajectories derived from exons acquiring a new $\Psi_{opt} = 0.5$ are clearly distinguishable from those maintaining their ancestral optimal inclusion rate (Figure 4.29E). The correlation between real and observed $\Delta\Psi_{opt}$ and $AUROC_{\Delta\Psi_{opt}}$ also increased as $\frac{\tau^2}{2\alpha}$ decreased (Figure 4.29F,G) suggesting that inference of shifts becomes possible when the rate of neutral evolution decreases. Even if the performance of the method remains poor, with AUROC values around 0.6, still close to random expectation, this is sufficient to at least infer the proportion of exons with shifts in Ψ_{opt} during mammalian evolution (Figure 4.29H). These results, altogether, suggest that the fast rates of neutral evolution driving Ψ divergence prevent us from estimating the prevalence of lineage-specific adaptive changes in inclusion rates during mammalian evolution.

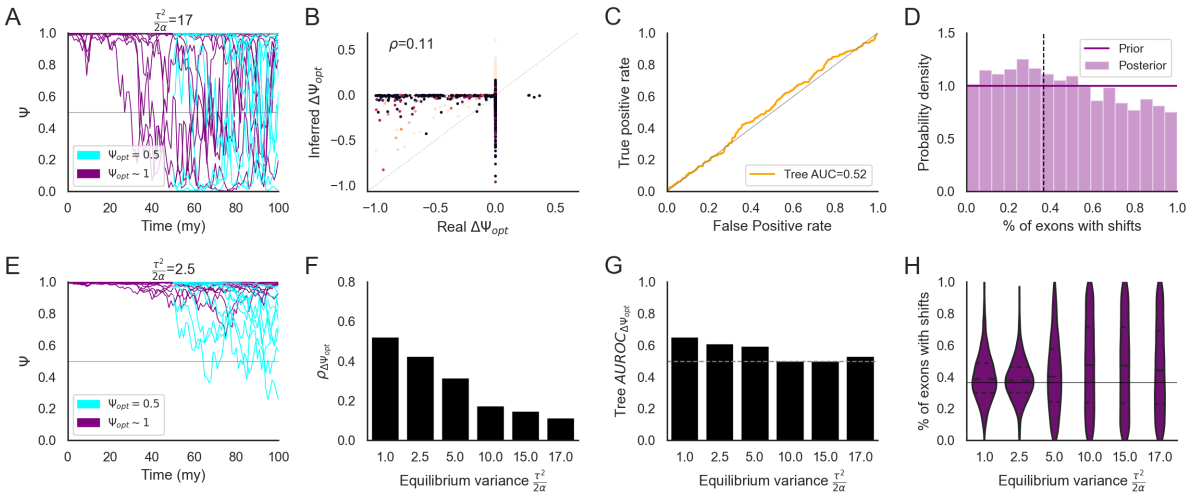

Figure 4.29: Fast neutral divergence prevents the inference of lineage specific-shifts in optimal inclusion rates. **A** Simulation of the evolution of 20 exons under the globally inferred OU model with a $\Psi_{opt} \sim 1$. After 50 my, new trajectories, shown in cyan, branch from the evolving exons to evolve under a $\Psi_{opt} = 0.5$. **B** Comparison of the simulated and inferred $\Delta\Psi_{opt}$ along the branches of the phylogenetic tree for 1000 simulated exons. **C** Receiver Operating Characteristic curve for the identification of exons with shifts in the optimal value based on the simulated data. **D** Prior and posterior distributions for the percentage of exons with shifts in the optimal value with known sensitivity and specificity. The vertical line shows the real percentage in the simulations. **E** Simulated Ψ trajectories as in A, but with using an equilibrium variance $\frac{\tau^2}{2\alpha} = 2.5$. **F-H** Correlation between simulated and inferred $\Delta\Psi_{opt}$ (F), AUROC for the identification of exons with shifts (G) and estimated percentage of exons with shifts (H) for varying sets of exons evolving under different equilibrium variances

5. Discussion

5.1 The importance of definitions: function and alternative splicing

In any scientific field, communication of ideas, results or controversies take place through language: we need to have a common framework, a set of words and concepts that we all understand in a similar way in order to build and reason about more complex ideas. There are some disciplines, like maths or physics, that are very aware of this issue and emphasize the importance of explicit definitions and axioms in their work. In contrast, biological sciences often come up with different intuitions about widely used concepts that are not always explicit and lead to apparent disagreements and lack of consensus. This happens most of the times because some concepts are rather difficult to define and have historically had different uses in different sub-fields of biology. This is the case of 'function'. While everyone has an intuition about the meaning of 'function', it is hard to actually put that intuition into a clear definition that allows distinguishing what is functional from what is not, at least conceptually. And this leads to apparent scientific disagreement over the interpretation of findings, as it happened with the first ENCODE publications claiming that 80% of the human genome was functional. This was viewed by evolutionary biologists as nonsense, not because they did not believe the results, but because they had a different understanding of 'function' [184]. There are two main different views of the concept of function: while some understand it as having some type of activity or effect, like binding to a particular molecule or catalyzing a reaction, evolutionary biologists usually understand it as having some effect on the fitness of the individuals carrying this element or variant [184]. The conflict arises as elements with some biological activity, e.g. transcriptional activity, can have no impact on fitness. The debate about the global functionality of alternative splicing [277, 34, 276] has a very similar nature, stemming from the lack of explicit definitions of both function and alternative splicing.

We can define alternative splicing as any variation in the splicing process. Under this view, given that biochemical reactions will never take place at 100% efficiency, every gene, even intron-less genes, are subjected to alternative splicing, even if it is detectable only in some of them. Indeed, as technology to detect AS variants and the amount of sequencing data improved, the estimates of the human genes that are alternatively spliced has increased from 74 with microarrays to 95 % with RNA-seq [122, 202]. Similarly, if we sub-sample sequencing reads from a RNA-seq experiment, we can see that our ability to find known major SJs saturates at lower sequencing depth compared with novel SJ, whose detection continuously increases with coverage. This is probably due to increasing power for detecting low frequency AS variants [292]. Similar results were found in a different study: as the number of human RNA-seq datasets increased, so did the number of new SJ, revealing more than 50000 unannotated SJ, which saturated only after over 7000 samples were included [201]. While some of these newly found SJ may derive from genetic variation within individuals, the statistical power with about to 7000 RNA-seq samples to detect rarely ocurring AS variants is also expected to increase and at least partially explain the results.

Therefore, the important scientific question remaining is: what is the proportion of detectable AS variants that are functional and how much of what we observe derives from noisy splicing?

In this thesis, we have approached this general question under the two different perspectives of 'function'. In the first section, one of the aims was the characterization of the functional impact of AS changes throughout heart development and disease from the point of view of the biological activity. In this sense, we have tried to characterize the biological functions associated with genes with varying AS patterns and how much they were expected to modify the protein sequence and molecular function e.g. binding to other proteins. The characterization of the global effects of AS changes provide no direct evidence of the associated fitness consequences of each alternative processing event. However, the existence of a regulatory network to induce AS changes during development or in response to injury in a reproducible manner, which are are additionally associated to specific biological activities, somehow suggests that these changes are required for proper heart functioning. Modification of the AS events, alone or in combination, would therefore have fitness consequences even if we do not know how that happens. Thus, an important part of the thesis has been devoted to the inference of the regulatory network underlying observed AS changes between different conditions. This not only provides a better understanding of the specific system under study i.e. heart function, but also provides some hints about the relevance of the AS changes as a whole. Indeed, we have seen how the perturbation of a single element of the regulatory network, PTBP1, leads to cardiac dysfunction. Given the importance of this task, we have also developed dSreg, a new statistical model for improved inference of the activities of AS regulators.

However, the relationship between the existence of a regulatory network and the fitness consequences of modifying the target nodes of that network is not straightforward: structured AS changes may also derive from propagation of noise in the trans-regulators through a regulatory network that may be functional in different conditions. In this scenario, whether these changes take place or not would not influence fitness and may not be deemed as functional from the evolutionary perspective even if they are associated to certain biological effects. Given this uncertainty, in the third section we have studied AS functionality from the evolutionary perspective, this is, aiming to characterize the fitness consequences of quantitative changes in exon inclusion rates.

5.2 Alternative splicing regulation

An important part in the development of this thesis has to do with the identification of the regulatory mechanisms driving AS changes between different biological conditions. In the following sections, we will discuss different aspects about the regulation of AS, ranging from the specific trans-regulators driving AS transitions in the heart to more general patterns of regulation and the methods to infer them from RNA-seq data.

5.2.1 Regulation of alternative splicing patterns in the mouse heart

We aimed to study the regulatory patterns driving AS changes specifically in mouse models of heart disease and how they resembled those modulating AS during developmental transitions in the heart. With this purpose, we analyzed a large dataset of new and previously published RNA-seq experiments, including several developmental stages and disease models i.e. myocardial infarction and cardiac hypertrophy through TAC. We identified sets of exons with reliably increased or decreased inclusion rates across 4 widely defined transitions: two developmental and two disease transitions. Using a set of CLiP-seq experiments, we identified RBPs whose binding sites were associated to AS changes more often than expected by chance. We analyzed their gene expression patterns, since a change in their expression is likely to be linked to a change in their regulatory activity. MBNL1 binding sites in the upstream

intronic flank were associated to decreased inclusion during heart development, while its binding to the downstream flank was associated to increased exon inclusion (Figure 4.6). These patterns can be explained by a position dependent effect of MBNL1 binding to different regulatory regions in the pre-mRNA relative to the target exon: it inhibits splicing when binding upstream and enhances it when binding downstream the target exon, as previously proposed [77]. This, together with a marked increase in its expression throughout development and the correlation with AS changes induced by MBNL1 knockout, suggests that MBNL1 regulates AS in the heart mainly during development, as previously characterized [126, 62]. Based on these findings, we hypothesize that the increase in MBNL1 expression is responsible for the highly dynamic AS regulation during development, as it promotes both exon inclusion and skipping in a position-dependent manner. Although its activity may still be regulated at the protein level during heart disease, both the lack of enrichment and expression changes suggest that its regulatory activity remains unchanged, at least for the disease models included in this study. This may explain, at least partially, why the re-expression of neonatal patterns in disease is not complete. While MBNL1 disruption leads to cardiac disease, as previously shown [62], it does not play a role in the development the heart diseases under study, limiting its therapeutic potential.

Then, what is modulating AS during heart disease? We hypothesize that PTBP1 is the main, but not only, driver. PTBP1 showed the greatest contribution to explaining AS changes independently from the other regulators in heart disease and, to a minor extent, in heart development. PTBP1 showed consistent opposite expression changes during development and disease, correlating with their targets being enriched in more included and skipped exons, respectively. These results are consistent with the expected inhbitory role of PTBP1 when binding to the upstream intronic sequence [**Gerogilis**, 284, 223]. The simultaneous regulation by PTBP1 in opposite directions during development and disease, in absence of MBNL1 changes in the latter, may underly the partial re-expression of the neonatal AS patterns in heart disease. While this strongly suggests that PTBP1 modulates AS changes during disease, it may very well be a downstream consequences of other molecular alterations rather than disease cause. However, previous work suggested that PTBP1 may indeed drive cardiac dysfunction, since it is required for the differentiation of iPSCs and fibroblasts to cardiomyocytes *in vitro* [169] and modulates the splicing of essential genes for cardiomyocyte function (e.g. Titin, Tropomyosin 1 and 2 and Mef2) [74]. Even if it was necesary for proper cardiomyocyte differentiation and function, its effect *in vivo* had not been evaluated to date. Thus, to investigate whether PTBP1 can actually cause heart disease, we over-expressed PTBP1 in the healthy myocardium of adult mice and found that these mice developed cardiac hypertrophy and diastolic disfunction compared with control mice. Importantly, this over-expression was very similar to that achieved in TAC, suggesting not only that our model was realistic, but also that small alternations (1.2 fold) of PTBP1 activity can have important phenotypic consequences. We hypothesize that this is effect takes place through modulation of AS patterns, but we can not rule out that PTBP1 has other functions in RNA metabolism besides AS regulation [284, 234], e.g. it modulates insulin mRNA stability in the cytoplasm [79], or that AS changes are secondary consequences of PTBP1-induced hypertrophy by other means. Despite the low overall correlation of AS changes between TAC and PTBP1 over-expression, the events affected in both situations were associated to PTBP1 binding. Moreover, there is a small subset of AS events that are coherently modulated in both scenarios that may drive the development of cardiac hypertrophy. The most affected event in both conditions is a muscle-regulated exon in LRP4 gene. LRP4 encodes a low-density lipoprotein (LDL) receptor, which is bound by Agrin and mediates MuSK activation, which is essential for the correct functioning of the neuromuscular junction [133]. In the heart, Agrin is required for cardiac regeneration in neonates and its over-expression was able induce regeneration in adults after injury [24] and to modulate cardiomyocytes contraction *in vitro* [105]. Therefore, defects in the function or amount of its receptor may affect the benefitial role of Agrin in the response to injury o hypertension. Moreover, mutations in LRP4 have been

associated to sindactyly and polysindactyly in several mammalian species, including mouse, cow and humans [253, 121, 164], which is usually accompanied by cardiac defects in Timothy Syndrome patients. Inclusion of this exon introduces a stop codon near the C-terminal region of the protein and thus is expected to produce a truncated version of LRP4 protein without the last exon. Even if the impact of this highly conserved and cardiac-specific splicing event in protein function remains unknown, it is an interesting candidate AS gene for better understanding PTBP1-mediated cardiac hypertrophy. PTBP1 is also known to regulate Titin splicing isoforms *in vitro* [74]. Although we did not find large changes across any exon in Titin upon PTBP1 over-expression, 27 of them showed estimated $\Delta\Psi < -0.04$, which, when considered together, may suggest a contribution to isoform shortening and increased passive stiffness of the cardiac muscle, as it happens during development [144]. Other interesting AS changes mediated by PTBP1 and potentially promoting cardiac disease affect LASS6 and CAD genes. The modulated exon in CAD shows high inclusion rates in skeletal and cardiac muscle compared to most other tissues, as its human ortholog does, suggesting that its cardiac-specific inclusion has functional consequences (Figure 5.1). CAD is a key enzyme for pyrimidine synthesis and aminoacid metabolism, and thus essential for nucleic acid synthesis and proliferation. The affected exon in LASS6 gene, on the other hand, is included specifically in neural tissues across both human and mouse genes (Figure 5.2). LASS6 participates in ceramide synthesis, compounds that can cause mitochondrial dysfunction and heart failure [214, 309, 118]. Genetic and pharmacological inhibition of ceramide synthesis improves cardiac function in mouse models of lipotoxic cardiomyopathy [214]. Thus, if LASS6 splicing event alters the ceramide synthesis pathway, it will probably play an important role in cardiac pathology.

Although PTBP1 appears to be the main driver of AS changes in heart disease, it is unlikely to be the only one. There are several sources of evidence pointing to a simultaneous contribution of its paralog PTBP2 to AS modulation in cardiac disease. Even if both proteins bind to relatively similar targets [166], we found that PTBP2 binding sites were independently associated to AS changes in heart disease thanks to our regression framework (Figure 4.6A), suggesting increased chances of modulation whenever having a PTBP2 binding site, regardless PTBP1 binding. This is further supported by the increased correlation of AS changes TAC with PTBP1/2 KD compared with that with PTBP1 KD alone (Figure 4.8). These results are consistent with the observed simultaneous up-regulation of both PTBP1 and PTBP2 in MI and TAC. This takes place even if PTBP1 actively inhibits PTBP2 expression through RUST [260, 36], suggesting the existence of additional and dominant mechanisms driving the simultaneous up-regulation of both genes. While PTBP2 keeps inhibited their common targets during later stages of neural differentiation, there are also some unique targets to each regulatory protein, like PBX1, which is specifically modulated by PTBP1 [166]. Altogether, we hypothesize that PTBP2 is also playing a role in AS regulation in cardiac disease and would further contribute to hypertrophic growth if over-expressed together with PTBP1.

Interestingly, neural and cardiac tissues also seem to share sequential regulation of AS by PTBP1 and MBNL1 [114, 148, 153, 57, 298]. These results sug Although there are additional regulators apparently specific to neural tissues, like SRRM4 [221, 113], differences in the AS patterns of muscle and brain tissues may be explained by quantitative differences in the expression of trans-regulators rather than on the structure of the regulatory network. This may be achieved if binding of RBPs to their targets depends on both the affinity of the binding site and the protein concentration, as recently shown for RBFOX1 [27]. If this is a general issue for every RBP, small changes in the expression of certain regulatory elements may drive the alternation of different AS targets and driver tissue-specific patterns.

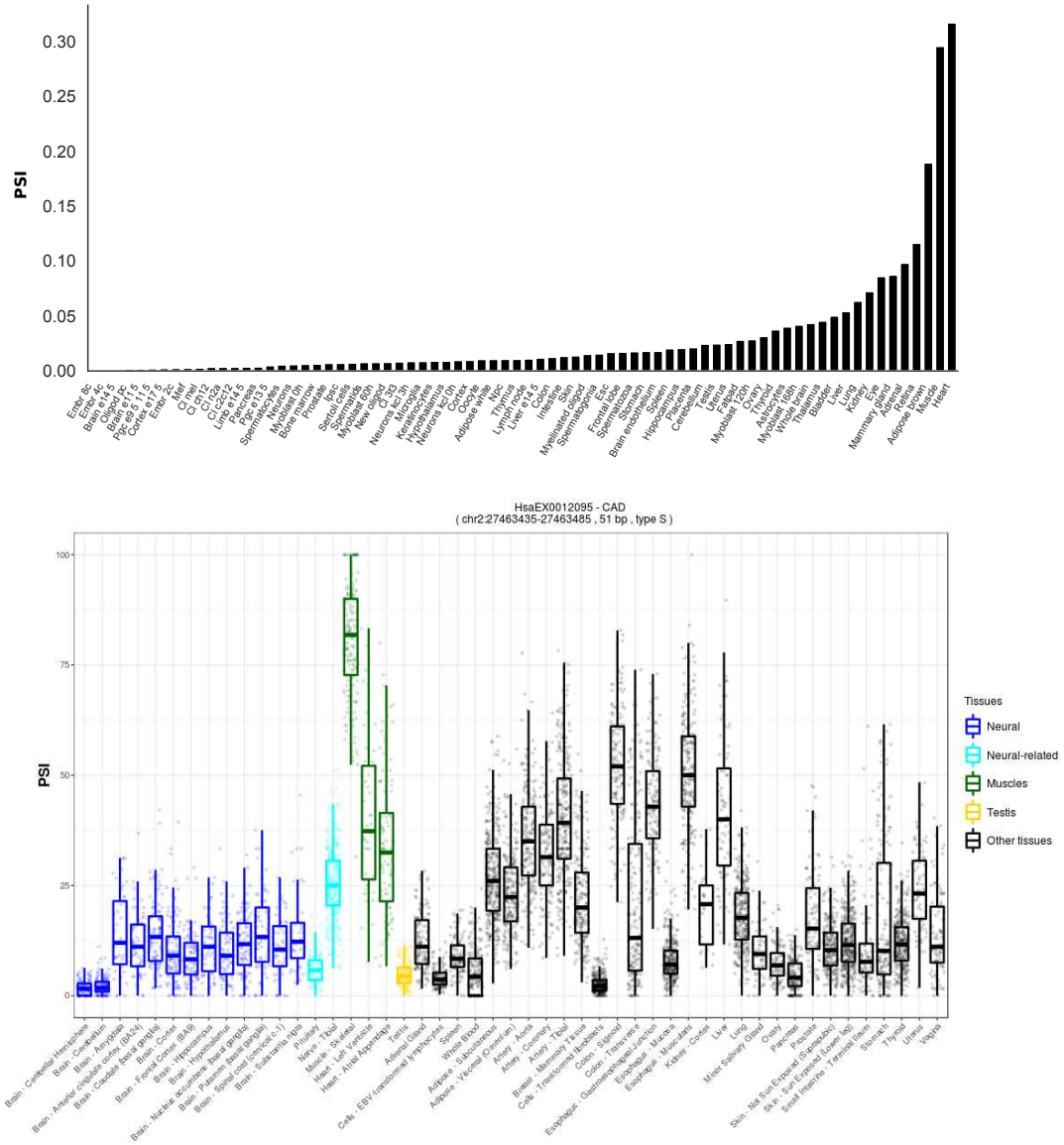

Figure 5.1: Exon inclusion rate of CAD alternative exon across mouse (top) and human tissues (bottom). Data was extracted from VASTDB [268]

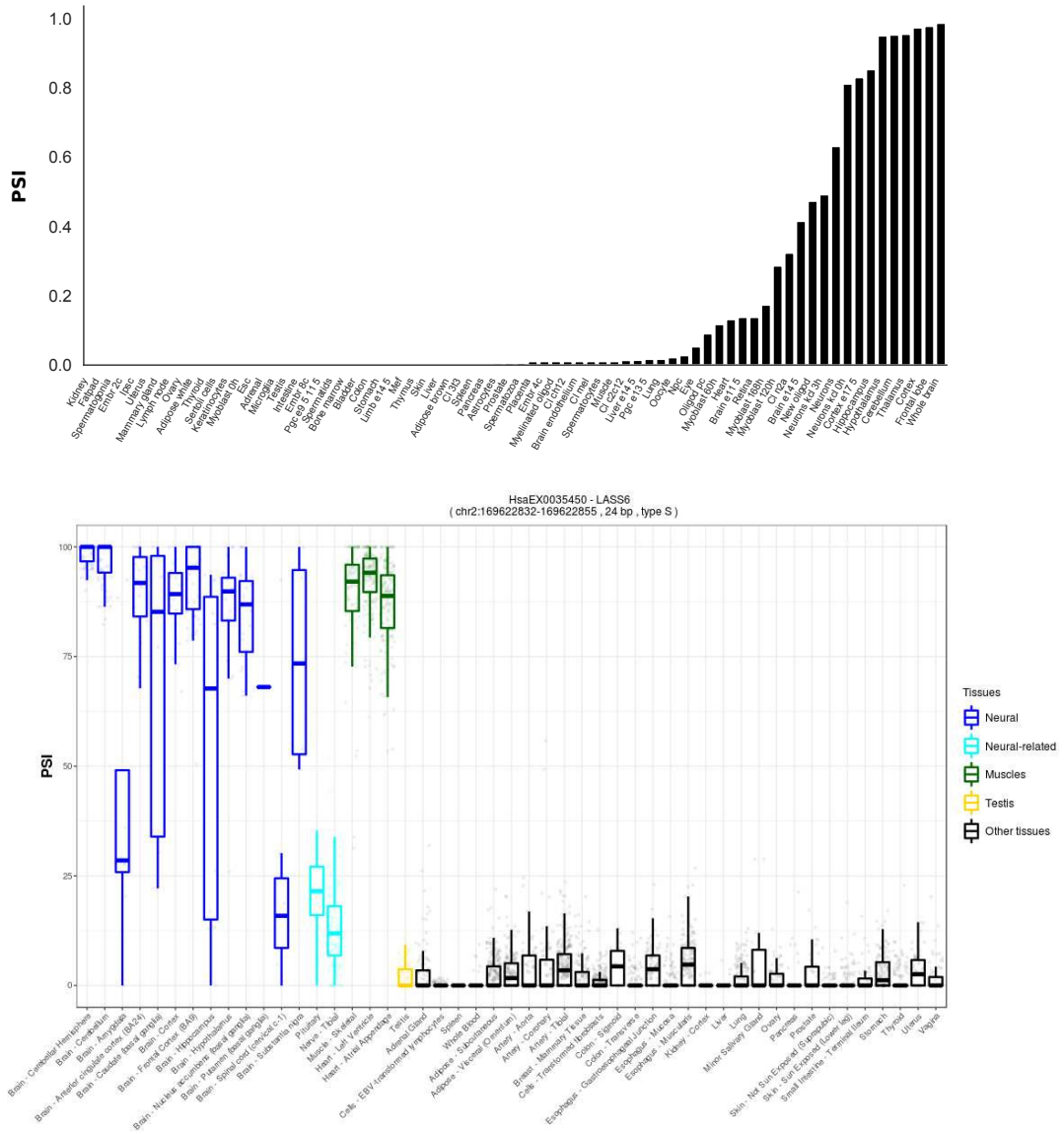

Figure 5.2: Exon inclusion rate of LASS6 alternative exon across mouse (top) and human tissues (bottom). Data was extracted from VASTDB [268]

5.2.2 Improving the inference of trans-regulatory elements activity with dSreg

During the development of the first section, we became aware of some important limitations of the enrichment approaches for identifying trans-regulators driving AS changes. First, as previously mentioned, different RBPs often bind to similar target sequences [225, 64, 203], which hinders the identification of the actual regulatory protein underlying the observed changes. To tackle this issue, we first performed the classical Fisher test to analyze the over-representation of binding sites of each RBP in altered exons individually. We used this as a first step to filter the potentially associated RBPs, rather than keeping these candidates alone. We then used them for a second analysis to identify the subset of RBPs that were independently associated to AS changes. In this second step, instead of a Fisher test, we worked in a logistic regression framework, using the binding sites of the candidate regulators simultaneously as predictors. While this approach was likely to reduce false positives in the identification of regulatory elements due to correlations in the binding sites of different regulatory elements, it still ignored the quantitative information in the data and remains highly dependent on the classification of exons into different discrete categories. Given this and other remaining methodological limitations, we developed dSreg, a new method that takes into account the quantitative changes in mRNA proportions and the uncertainty over such changes at a transcriptome wide level. dSreg is able to use quantitative information and gives the proper weight to each exon according to its coverage or degree of information, by directly conditioning on the number of reads supporting inclusion and skipping for each exon and sample. To deal with the co-linearity between binding profiles of different RBPs, dSreg uses a regression framework so that the changes in the $logit(\Psi)$ of a certain exon are derived as a function of the sum of effects of all the RBPs that bind to it.

To show the power of our hierarchical approach, we simulated data under this exact regulatory model with known binding sites and tried to recover the true parameter values, i.e. RBPs differential activities, giving rise to the observed data: the number of reads supporting exon inclusion and skipping across every sample in the experiment (Figure 3.1). Our results showed very good performance in the inference of trans-regulatory proteins, even at very low sequencing depths and few samples per condition (Figure 4.13). In these settings, there would certainly be no statistical power to test whether inclusion rate was different between the two conditions for any particular exon independently. As expected, ORA, reliant on such statistical power, is unable to identify any regulatory element at low depths and showed increasing F1 score with depth. In contrast, the performance of GSEA, which uses quantitative information, even if noisy, remains relatively unaffected by coverage like dSreg, but its performance is consistently lower. Also as expected, as we increase the number of non-active regulators, both ORA and GSEA increase their false positive rates, as they would report as significant both the driver regulators and those correlated binding sites. Of course, the good performance of dSreg in these simulations is conditioned on perfect knowledge about how splicing regulation takes place and the specific binding preferences of every potential trans regulator. Even if these are rather unrealistic assumptions, our results suggest that if we know the mechanisms by which AS is modulated and where each regulatory protein binds, we can very accurately infer which are the regulators acting in a given situation with very limited sequencing depth.

We also evaluated the performance of the different methods for the identification of trans-regulatory factors in real RNA-seq data. Unfortunately, there is no certain way to know which are the exact regulatory elements acting in a real situation. The closest we can get to that is by experimentally perturbing the expression of a certain factor or RBP, that may contribute to AS regulation through their ability to bind RNA. This does not ensure that this RBP is modulating AS patterns and that it is doing it alone. Still, over a large number of experiments, there should be at least an association between altered and driver RBPs. Under this assumption, we used dSreg, ORA and GSEA to infer the trans-regulators driving AS changes in a large collection of RBP knock-down experiments from the ENCODE

project. Surprisingly, we found that both ORA and GSEA identify the perturbed RBP as often as randomly selecting an number of RBPs equal to those deemed significant by each method. While dSreg shows a better performance, its improvement over random performance is still very low (Figure 4.16). Overall, these results suggest that our power to detect the perturbed RBP based on the AS patterns and the set of binding sites employed here is very small for any available method. There are several non-exclusive potential explanations for this. First, the set of derived binding sites from CLiP-seq experiments probably contains errors. The extent to which errors in the binding profiles affect the inference of the active regulatory elements remains unknown for every method. Unlike other issues, this can at least be tested using simulations and introducing errors in the binding profile. Second, certain amount of RBPs may not actually regulate splicing directly, and therefore there is no real association between AS changes and their binding sites. In this case, AS may be secondary, and driven by other RBPs whose expression is somehow dependent on the targeted protein.

Although the main motivation of dSreg is the inference of the RBPs driving AS changes, it simultaneously infers AS changes like widely used tools like MISO, rMATS, MAJIQ or Whippet [128, 249, 287, 261]. However, it does so taking into account the underlying regulatory network. These inferences were very much improved in the simulation experiments thanks to the propagation of information across the regulatory layers: when there is little evidence of change for the targets of a particular RBP, this strongly suggests that the activity of the RBP is altered. However, the information flows in both directions: as we know that an RBP activity is changing, then we have stronger evidence that their targets are changing, even if there is poor direct evidence in form of few reads mapping to each target exon. This may help improve estimates in case of errors due to small samples size. Similar attempts have been previously made to leverage the regulatory information to improve the inferences of the AS changes. First, BRIE [107] used k-mer composition to build an informative prior through a linear regression framework to infer Ψs in scRNA-seq data. In this type of data there are very few reads per cell but there are many cells that can compensate this lack of local information: having many cells can compensate the low coverage per cell if we are able to properly propagate information among different cells. Not surprisingly, BRIE outperformed classical tools designed for bulk RNA-seq experiments like MISO or rMATS [128, 249]. More recently, a similar approach was implemented in DARTS [315], which used a Deep Learning strategy to build an informative prior for each AS event for the probability of changing. In this case, DARTS includes much more information in its prior, as it uses data from the large collection of RBPs knock-down experiments from ENCODE [203], simultaneously integrating the expression of these RBPs with the downstream AS changes into the model. This prior is of course only available for human species and therefore limits its applicability in other model species like mouse, *Drosophila* or *C. elegans*. As it is already trained in previous data, it can directly provide a prediction of the AS changes in a dataset based only on the RBPs expression, or additionally incorporate the information from the RNA-seq experiment to generate a posterior probability of AS change for each event. The incorporation of this prior was shown to improve the detection of AS changes and outperformed classical tools like MISO and rMATS. In our work, we have compared these new tools with dSreg using deep sequencing data processed by DARTS [315] and sequentially reduced the sequencing depth. Assuming that the full sequencing depth or independent RASL-seq quantification of a small subset of exons provided the true value, we evaluated the dependency of these tools on coverage. This analysis showed that the performance of dSreg at quantifying and detecting differential splicing is, at least, similar to these previous approaches except at very low depths, at which DARTS outperforms the other methods. This happened regardless of using their model based informative prior (flat vs info), suggesting that the specific model specification of DARTS is more realistic when information is missing (Figure 4.16). The improvement of these approaches in comparison with our Null model are rather small, if any. This suggests that we are missing relevant mechanistic information in our regulatory models, given that information does not propagate through the regulatory

network to improve inferences as much as in our simulations. Whether this is driven by incorrect binding profiles or incomplete regulatory model remains unknown. Still, dSreg is able to identify whether there is information across the provided regulatory network and use it to derive trans-regulatory factors. It does it without perturbing the differential splicing inferences even under a possibly miss-specified regulatory model, only using information when it is actually there. Thus, there is still a lot of room for improvement, which can be monitored by comparing the relative improvements in performance achieved by methods in real and simulated data.

5.2.3 Definition of RBP-RNA interactions

A key and complex part of the regulatory analysis is the definition of the binding sites for the regulatory elements of interest. A whole variety of approaches have been taken in the literature, without a clear evaluation of which works best in practice for this specific aim.

First works on the determination of RBPs binding specificity showed that binding preferences for most RBPs could be confidently explained by 4-8 bases long stretches of RNA [173]. Based on these findings, a first simple and often used way to define regulatory features for AS modulation is to extract 6-mer profiles for exons and their adjacent sequences. This approach has been used for identifying which are the cis-regulatory signals affecting quantitative AS outcomes [235] or the specific factors modulating a set of exons e.g. microexons [161]. As previously discussed, 6-mer profiles have also served to build a global informative prior for inferring Ψ across exons in scRNA-seq experiments [107]. These cis-regulatory sequences can be later associated to specific trans-factors based on databases like cisbp-rna or ATtRACT [225, 91] using specific computational tools for motif comparison like TOMTOM [102].

In this work, we have taken advantage of the newly released ENCODE dataset with CLiP-seq data for a large amount of RBPs [64], together with previous smaller databases [301, 35, 131, 158], and used the obtained *in vivo* binding sites to define the binding profiles of the different RBPs across the whole dataset of exons. We have assumed that these binding sites are generally conserved and have lifted-over the coordinates from human experiments to mouse for our analyses in the mouse heart. These data provide direct evidence of *in vivo* binding of a RBP to a sequence in the RNA. This decision was motivated by the increasing evidence of differences between predictions from sequence motifs and *in vivo* binding from experimental data, potentially driven by the 3D RNA conformation [267, 64]. On the other hand, these data are highly dependent on the actual genes that are expressed in the experimental conditions: if a gene is not expressed, an RBP can hardly bind to its pre-mRNA to modulate AS. If this gene is expressed in the conditions we want to study, it may artificially be bound by no RBP at all. Thus, our inferences about the trans-regulators mediating AS transitions may be biased or at least guided mostly by genes that are expressed in the experimental settings in which the CLiP-seq experiments were performed. For most RBPs, CLiP-seq experiments are performed in cell lines like K562 or HepG2 [64], which may be very different from *in vivo* cardiac myocytes, for instance.

Ideally, instead of relying on directly determined binding sites, one could derive more complex models for the binding of each RBP to predict their binding sites across the transcriptome, overcoming the limitation of a gene being expressed in very specific experimental conditions. These models may improve simpler and purely sequence based Position Weight Matrix (PWM) as in Cisbp-rna [225] into complex Deep Learning approaches that take into account the RNA structure [311, 78, 187, 210, 6]. While PWMs accurately describe the binding of some RBPs, there are some, like PTBP1, for which the structural information i.e. both base pairing and 3D-structure, is a key determinant for *in vivo* binding to RNA [311]. Thus, although such models exist for a small number of RBPs, a comprehensive and systematic building of complex models for the *in vivo* binding is still missing. This development would allow the extrapolation of experimental results in very specific conditions to broader biological contexts and thus correct for gene

expression variability when defining the binding profiles for downstream regulatory analysis.

5.2.4 Interaction between trans-regulatory proteins

So far, most efforts to study AS regulation have focused on the identification of specific trans-regulatory elements that, alone, can explain at least part of the AS changes and potentially drive phenotypic differences [221, 126, 92]. However, there is increasing evidence that AS regulatory networks are more complex than that: different RBPs have very similar binding affinities, bind to common targets in the RNA [225, 64, 203] and modulate them [40]. How these RBPs organize in complexes and how they affect AS is not very well understood. Even considering a single RBP alone, several molecules may be required to bind stably to the target RNA e.g. PTBP1 requires several binding sites along the RNA for protein molecules to cooperatively bind simultaneously with higher affinity and specificity than individual proteins to isolated binding sites [54]. Assuming that AS modulation by RBPs depends on their binding affinities to their target sequences in the pre-mRNA, cooperative or competitive binding is expected to have an important impact on the resulting AS changes.

Nowadays, there is only indirect evidence about the role of interactions between RBPs in the regulation of AS. Previous work studying how AS rates change in large collections of mutant sequences in minigene constructs suggests that regulatory interactions contribute little to the determination of the Ψ: Accounting for pairwise interactions in the occurrence of 6-mers improved rather little the prediction of Ψs in the mutant libraries [**Rosenberg2015**]. However, this is contingent on the suitability of 6-mers to represent RBP binding: if they represent poorly the binding specificities, one can expect that interactions among these 6-mers will provide little to no information to explain observed Ψ patterns. A different study showed that epistatic interactions between pairs of nucleotides occurred mainly within windows of 6 bases, suggesting that they arise due to the complex nature of RBP-RNA binding rather than by the interaction between different regulatory elements [16].

Although the prevalence of interactions between RBPs in the AS regulatory network remains unclear, there are several known examples: PTBP1 molecules bind cooperatively among themselves and can interact with MBNL1 proteins and cooperatively bind and modulate the inclusion of Tropomyosin exon 3 [94] or with RBM20 to regulate Titin AS [74]. If interactions are widespread, a simple additive model with independent effects for all RBPs may fail to identify the actual regulatory elements, since their effects will be highly dependent on the factors that simultaneously bind each target. We have approached this question by using regularized regression to allow pairwise interactions between candidate RBPs when predicting whether an exon was changed in a particular transition in the mouse heart. These interactions were relatively abundant, and had a greater contribution to the regulation of AS in heart disease than to developmental changes. This does not necessarily imply a rewiring of the AS regulatory network in disease, but can be explained by non-coordinated changes in the expression of the regulators. In contrast, highly coordinated changes of RBPs during development resulted in a tightly regulated set of AS changes as if they were all regulated by a single RBP, in this case, MBNL1. Moreover, by taking into account these potential interactions, PTBP1 was unveiled as an important regulator whose contribution was validated in an experimental mouse model over-expressing PTBP1. Under widespread interaction between trans-regulators, it may not be surprising that PTBP1 up or downregulation (Figures 4.8 and 4.12) alone did not reproduce a large number of AS changes compared with disease. This would require a very specific modulation of a number of RBPs, according to our previous model, rather than the perturbation of a single RBP alone.

An important consideration in our approach is that we did not model quantitative changes in Ψs, but the qualitative definition of AS change in a certain transition. Although one can expect that bigger quantitative changes are more easily detectable, whether quantitative additivity on $logit(\Psi)$ would imply

additive effects on the detection of significant changes is a complex question. We have only studied rather superficially the role of regulatory interactions for AS regulation, but it can easily be improved by using dSreg accounting for co-occurrence of binding sites of combinations of RBPs. Including interactions would certainly increase the complexity of the model, since there are many more combinations of RBPs that single RBP when considered individually. However, as dSreg uses a Horse-shoe prior to promote sparsity on the regulatory activities, with little decrease in performance as the number of potential regulatory elements increase, it provides an interesting computational tool to study this question in more depth in the future. Still, additional studies would be required to validate this hypothesis experimentally, e.g. by systematic perturbation of pairs of RBPs.

5.2.5 Limitations of dSreg regulatory model and potential improvements

The good performance of dSreg on simulated data suggests that it is very powerful when the regulatory model is realistic and the binding sites for all potentially participating trans-regulators are perfectly known. In these conditions, we found that dSreg is able to very accurately identify the quantitative AS changes taking place for each exon even with few number of reads and samples supporting them in comparison with a model ignoring the regulatory information. Therefore, the difference in performance between the two models at identifying the AS changes upon decreasing sequencing depth may provide indirect evidence about the amount of variation that can be explained by our regulatory model and binding sites on real data. Indeed, in contrast to results in simulated data, the performance in the identification and quantification of AS changes of dSreg and its Null model were very similar when assessed on real data. While dSreg still provides better inferences about the trans-factors contributing to those changes, it suggests that there is room for improvement in probably both the regulatory model and the definition of binding sites.

We have previously discussed the difficulties we found at defining the binding sites for the potential trans-regulatory elements. But the regulatory effect of an RBP may not only depend on the binding affinity, but also its relative distance to a particular splice signal. For instance, we found that PTBP1 binding sites are mostly enriched in a window of about 100bp upstream of the splice acceptor, while differences between modulated and unchanged events at larger distances was rather small (Figure 4.9). The association of RBFOX2 binding sites with the modulated exons also decreases with the distance to exon splice donor, but may extend up to about 200 bases [255]. SF3B1, on the other hand, exerts its regulatory action by modulating the recognition of the branch point, and therefore the relative position at which the regulatory effect is maximized may be different [101]. As the position dependency appears to be unique to each regulatory protein or at least to regulatory modes, throughout this thesis, we have simply assumed that the regulatory effect is the same whenever the binding site was located within 250 intronic and 50 exonic bases. In the future, dSreg may be extended to have an optimal distance to any regulatory element, from which the regulatory effect would decrease according to certain function, e.g. Gaussian function, to simultaneously model the position dependency and regulatory activities.

Another important issue that is ignored by dSreg regulatory model is the expression of the trans-regulatory factor. In this model, we have assumed that the trans-factor would always bind to every provided cis-element in the binding sites matrix. However, if the RBP concentration is not very high, different cis-elements across the whole transcriptome may need to compete for the binding of the existing RBP molecules. In this situation, only a small amount of binding sites will be actually bound, such that only a small proportion of targets would be affected by the change in its activity. This has been recently shown for RBFOX1: whose up-regulation unveils its ability to bind to lower affinity targets and exert its regulatory function over them, while these targets are not sensitive to changes in lower concentrations of RBFOX1 [27]. Evidence for competition of binding sites for RBP binding can be found in DM. As

previously mentioned, expansion of GTC repeats in Dmpk 3'UTR out-competes other important cis-regulatory elements of MBNL1, leading to AS alternations and cardiac dysfunction [149]. As expected by the competition model, over-expression of MBNL1 in the skeletal muscle reverses myotonia [127] and satellite cell proliferation [257].

Although we also have information about the expression levels of each gene in the RNA-seq data, this is more difficult to incorporate to dSreg model. It would require to quantitatively characterize the binding affinity across each binding site and RBP and introduce a model for simultaneous competition among all binding sites depending on their pre-mRNA concentration. In this framework, one may not only consider competition for molecules of the same RBP but also of different RBPs and use it to investigate how changes in the stoichometry of RBPs may actually impact the resulting AS patterns. As we suggested for heart disease, changes in the stoichometry may explain the apparent interactions between regulators: if one regulator is limiting, then modification of the other may not have an effect. However, if the expression of the limiting regulator increases, it might unveil dependencies on other trans-factors.

We have used a simple logistic regression framework to integrate the regulatory effects into the resulting Ψ for a particular exons. Despite its mathematical convenience and relatively easy interpretation as the log-transformation of the ratio between inclusion and skipping, there is no mechanistic justification for trans-effects to be additive in the logit scale. Recent work has shown a global dependency of the effect of both mutations and trans-regulators on the actual Ψ. This dependency may be explained by a delayed competition model between splice sites characterized by the rates of recognition by the spliceosome for each competing splice site [16]. This model can accommodate other known important factors like intron size or nucleosome occupancy [110, 193, 172] as factors influencing the waiting time before splice site competition. This work provides a more mechanistic model for how the regulatory activity of a trans-regulator may be quantitatively translated into AS changes for building better models to identify trans-regulatory elements from RNA-seq data.

Finally, there is an additional assumption that may limit the ability of dSreg to learn the underlying regulatory patterns. dSreg assumes that the changes in Ψ between conditions depend on the combination of regulatory activities, but that biological variability between samples of the sample condition are completely unrelated. This is, there is no variability in the regulatory activity within the same condition. While this simplifies the amount of parameters in the model, there is no reason to think that the RBPs activities do not vary within experimental groups, as they may be also subjected to noise and biological variation. We find a very clear example of that in our own data: when injecting AAV9-PTBP1 into mouse hearts, there is some variability in the over-expression that we achieved in the heart. In fact, we found that one of the samples that were sequenced was not able to over-express PTBP1 sufficiently as to induce changes in the AS patterns. Therefore, we are missing the additional information provided by the relationships between exons that share binding sites within the same biological condition. If the variability within each experimental group is due to regulatory factors, we should be able to leverage that information. This can be incorporated in the model by assuming that each sample has a set of regulatory activities, which themselves depend on the experimental condition with some noise. If the variance within groups is driven, at least partially, by variability in the regulatory elements that propagates through the regulatory network to the target exons, we should be able to separate it from random independent noise.

5.3 Alternative splicing functionality

5.3.1 Functional impact of alternative splicing changes in the heart

In this section, we used an unprecedented breadth of samples and conditions to investigate not only the mechanisms that regulate AS changes in the heart, as previously discussed, but also their functional

consequences, here understood as the changes in the biological activity of the resulting mRNAs. We found that GE and especially AS patterns are more dynamically regulated during embryonic and postnatal heart development than after the induction of MI or TAC. Despite a partial recapitulation of developmental AS changes in heart disease, TAC and MI samples remained more similar to their adult controls than to neonatal samples, a trend also seen with GE changes. In addition, the biological processes affected by AS changes were mainly different from those altered by GE changes, regardless of the developmental or disease context, suggesting that AS and GE changes play distinct roles in the heart. These functions are often related with cytoskeleton, vesicle trafficking and acting dynamics, not only in myocytes, but also across neurons [114, 148, 153, 85]. These observations suggest that the genes modulated through AS changes are related to the same functions regardless, not only of the developmental stage in the heart, but also of the tissue and possibly other experimental conditions. Interestingly, exons with increased inclusion levels during development tended to be shorter and not to affect protein domains, whereas skipped exons tended to be of similar length to unchanged exons and to encode functional domains, which would presumably be disrupted by AS, suggesting strong effects on protein function. The latter changes are not recapitulated in heart disease, indicating that cardiac injury triggers only those AS changes with a lower impact on protein function compared to those taking place only during development.

We explored the global effect of AS changes on the PPI networks during heart development and disease. Besides AS modulated genes being central in the PPI network, as previously described in different contexts [114, 305], we found that AS changes tend to modify PPIs more often than expected by chance in both interaction datasets under study, as previously shown in cancer [57]. We also analyzed the relevance of the AS-dependent interactions in the global PPI network through the edge betweenness. This property reflects how many shortest paths between nodes go through each edge, in other words, how important those interactions are for the interconnection of different interacting modules. Our results suggest that AS changes in disease not only tend to affect more interactions than expected, but that these interactions are key in the PPI network. We did not find such association in domain mediated interactions. However, previous work suggests that most AS-mediated interactions do not affect protein domains, but short linear motifs [305]. Therefore, differences between the results in the two datasets may lay on the different nature of interactions under study. These results may be limited by the small size of exons groups overlapping between the interaction datasets and the exons that were characterized in our study. Therefore, increasing the number of interactions or exons would help to further verify our findings, besides careful experimental validation. In addition to this general trend, we observed some specific AS-mediated PPI changes that may have functional impact. Among these, we found MEF2A to have isoform specific interactions with MEOX1 and MAPK7/ERK5 and to show AS changes in both developmental transitions and MI. MAPK7 activates MEF2A in response to MEK5, which is alternatively spliced itself [247]. Since MEOX1 and MAPK7 expression and MEF2A splicing changes are known to modulate cardiac hypertrophy [247, 135, 154], these AS-modulated interactions may play a role in the development of the disease. In addition, among developmentally regulated AS-mediated interactions, we found the EGFR-ERBB2-ANKS1 interaction triad to be reduced. ANKS1 regulates EGFR and ERBB2 transport to the membrane, which is necessary for its ERBB2-mediated tumorigenesis [153, 213, 272]. ERBB2 deficiency has been shown to cause dilated cardiomyopathy [59], whereas its transient induction reactivates the regenerative potential of the neonatal heart in adults [60]. Therefore, changes in the interaction between ANKS1 and ERBB2 or EGFR mediated by AS are expected to have major consequences for the heart.

5.3.2 Approaching alternative splicing function through comparative studies

In the third chapter, we have characterized AS mammalian evolutionary patterns as quantitative characters evolving under a OU process. A key parameter of the OU model is the phenotypic optima i.e.

the value of the quantitative trait that maximizes the fitness of the species. Thus, we have been able to infer, for the first time, mammalian-wide optimal inclusion rates Ψ_{opt} for a subset of conserved exons in the mammalian genome. Both global and exon-level inferences suggest that most Ψ_{opt}s are very close to 100% inclusion despite occasional exon skipping at certain rates across a single or few lineages. In an attempt to answer the main question of how much of AS is functional, we may assume that an exon skipping event is functional when selection tends to maintain its skipping at least at 10% of the processed transcripts, i.e. $\Psi_{opt} > 0.9$. Under this assumption, we found that only about 2% of the exons with some evidence of skipping in at least one species actually have an optimal inclusion rate below 90% (Figure 4.22D). These exons belong to about 5-6% of the genes under study, suggesting that AS, or at least exon skipping events, are functional only in a minority of genes.

However, there are a number of factors that may influence the accuracy of our estimates. First, on one hand, there are many more exons in which no evidence of exon skipping was found, either because it is rare or gene coverage was low. This may lead to overestimating the percentage of functional exon skipping. On the other hand, AS patterns can be developmentally and tissue regulated, or even cell type specific [99, 194, 193, 21, 34]. Therefore, by studying only liver, a single tissue in adult stage, we may be overlooking some exons that are meaningfully skipped only in a single tissue or specific biological condition. In particular, brain shows quite distinct AS patterns at both RNA and protein levels [230]. Brain patterns are also more conserved than those of other tissues [21, 114, 194], suggesting that a good amount of functional AS that takes places exclusively in brain may be missed in this study. Second, AS functionality may not be detectable at the exon level, this is, there are indirect ways in which AS may perform a different function, such as RUST [151]. If AS fitness effects depend on the ability of exon inclusion or skipping to modulate gene expression, selection will not act to reach a certain Ψ for a particular exon, but on the proportion of transcripts to be degraded under certain condition. There are many different ways for a multiexonic gene to produce an aberrant transcript targeted by NMD e.g. inclusion of a poison exon (frame-shifting or with an in-frame stop codon), skipping of a frame-shifting exon or intron retention. Under this scenario, the selective constraint on a single of these options will be smaller, as there are many other available mechanisms that can perform the same function that can be easily evolved. Some RBPs have independent and recurrently evolved different ways of self-modulated non-productive splicing. In these particular cases, it seems likely that selection is acting to maintain these feedback loops regardless of the specific underlying mechanism [150]. More importantly, these alternative splicing-related mechanisms modulating gene expression can easily arise during evolution to compensate each other. How general these mechanisms are for and their functionality remains mostly unknown and is not addressed in our comparative study.

Given these known and other possibly unknown issues, our estimate of the % of functional AS may not be fully accurate. However, it provides completely orthogonal evidence to previous studies about AS functionality not being as widespread as previously thought, relying either on protein quantification or sequence analyses [277]. Indeed, our estimates are quite similar to those obtained from a literature based curated database of about 700 human genes, out of which between 5-13% were likely to be functional [32]. While this may seem disappointing, there was no rationale behind the expectation of AS greatly expanding the encoding potential of a genome or that doing so for a small percentage of the genes is not sufficiently advantageous to keep this mechanism. Indeed, it may very well be that not every gene can provide easily evolvable new functions. In this regard, most gene duplicates, which may be comparable to AS isoforms in terms of functionality, are actually lost and only a few are able to prove sufficiently advantageous to be retained, not necessarily by selection. Only 8-14% of duplicates resulting from the whole genome duplication in *Saccharomyces* [190] and between 4.4 and 16.2 % of duplicates from different genome duplication events in *Arabidopsis* are maintained nowadays [182]. Although there are many stochastic processes driving initial retention of duplicates and large chromosomal fragments

with many genes, we may assume that these percentages represent the proportion of duplicates that are functional. Under this assumption, the percentage of functional gene duplicates is similar to our estimates of AS functionality, pointing towards a proportion of the genome that can easily evolve new functions from pre-existing ones, either through gene duplication or AS.

Quantitative genetics theory links the OU model with the evolution of a quantitative trait under a single peak quadratic fitness landscape characterized by a single parameter w^2, that gives an idea of its width. Under some assumptions about the effective population size, we derived w^2 from the inferred parameters and obtained a first approximation of the Ψ-fitness map for mammalian species (Figure 4.27). If effective population sizes are in the order of thousands ($N_e \sim 2000$), the expected fitness consequences of quantitative changes in the Ψ are rather small, in the order of 10^{-4}. We could also investigate how the fitness landscape varied depending on the position of the fitness peak (Ψ_{opt}) and found that fitness effects are predicted to be much smaller for functional AS, understood as $\Psi_{opt} \sim 0.5$, than changes for optimal constitutive splicing rates ($\Psi_{opt} \to 1$). In other words, alternative exons do not require very specific inclusion rates, as long as there is certain production of the alternative isoform, whereas partially skipping a constitutive exon has much greater fitness consequences. This can be easily explained because variation in the proportion of functionally related AS isoforms, in which one can replace the other to some extent, is of course expected to have weaker impact than replacing it with nothing or an aberrant product.

5.4 Alternative splicing evolution

The application of models of phenotypic evolution to describe AS patterns in extant species not only allowed the estimation of the optimal inclusion rates and their fitness landscapes, as previously discussed, but also the characterization of rates and modes of Ψ evolution. This has been possible thanks to the time resolution provided the collection of a very large RNA-seq dataset including an unprecedented number of species in comparative transcriptomics [21, 194, 88]. Using these data, we have been able to obtain exon-level estimates of the average evolutionary rate under a BM model, which were highly variable along the genome, with 95% of exons spanning 2 orders of magnitude around a single peak at about $\hat{\tau}^2 = 0.39$ for every 100 my. Exons from the same gene evolved at similar rates, suggesting that there are gene properties that are associated to the rate of evolution of Ψs. We have explored what these gene properties may be and found that slowly evolving genes at the sequence level, possibly constrained by selection, also tend to evolve slowly at the Ψ level (Figure 4.20D), as previously reported for GE evolutionary rates [52]. In addition, we found a stronger negative association with the within-gene heterogeneity in the nucleotide substitution rate. In other words, the Ψs in genes with high variation in the nucleotide substitution rate along their sequence evolve at slower rates than genes with more homogeneous substitution rates. A potential explanation for this observation may be that genes that are highly constrained at the sequence level may introduce variability in the encoded protein through AS, allowing the partial removal of some protein segments. Alternatively, conserved AS, this is with low evolutionary rates, may allow heterogeneous nucleotide substitution rates. Other factor associated to variation in Ψ evolutionary rate is gene age. Bilaterian genes showed, in average, faster evolutionary rates than genes originated at different time points in the evolution. Interestingly, exon skipping became more prevalent in bilaterian genomes, possibly associated to changes in the genome architecture [97], providing further evidence of the association of exon skipping to the origin of bilaterians.

While the BM model provided an estimate of the average evolutionary rate of an exon across the mammalian phylogeny, it assumes that this rate is independent on the trait value, the Ψ for every exon. Not only that, but it predicts unrealistic Ψ distributions after some time of evolution, very different to the observed ones in extant species. All this suggests that the BM model is too simple to accurately

describe the evolutionary process driving Ψ change. To overcome some of the limitations of the BM model, we also used a generalized OU model for AS data. The OU models not only predicted better the observed Ψ patterns in extant species in comparison with the BM model (Figure 4.22), but also allowed us to study the evolutionary forces behind these variable Ψ evolutionary rates. We found that stabilizing selection constrains Ψ evolution, and does so with a comparable strength to that acting on gene expression patterns in *Drosophila* ($t_{\frac{1}{2}} = 19my$) [26]. Selection strength does not appear to contribute significantly to the variability in the average evolutionary rates, which are mainly explained by diverse rates of neutral evolution (Figure 4.23). Aiming for a more mechanistic interpretation, we used the known relationships between OU parameters and quantitative genetics theory [146]. A potential explanation for these differences is the effective population size N_e, since small populations are less sensitive to selection than bigger populations. N_e has been shown to vary within a single genome [95], spanning up to 1 order of magnitude, which seems insufficient to reach the high variability in N_e required to fully explain the data (Figure 4.23F). Moreover, even if fast evolving exons are located in genes showing higher rates of synonymous substitutions, supporting the lower N_e hypothesis, such differences seem unlikely to completely account for the variation observed in the equilibrium variances across exons (Figure 4.23H). Other non-exclusive factors are the fitness effects and the genetic variance. Our results suggest that rapidly evolving exons are characterized by large genetic variances, but also flatter fitness landscapes i.e. weaker fitness effects. Interestingly, these two factors seem to compensate each other to produce a relatively constant selective strength across exons. Still, random accumulation of phenotypic variance remains faster in these exons due to the larger genetic variance (Figure 4.23A).

We have also studied how evolution depends on the optimal inclusion rate and found that optimally alternative exons tend to evolve faster. Previous work suggested different evolutionary regimes for alternative and nearly constitutive exons [228] but, thanks to our comparative approach, we have been able to further dissect the nature of these differences. As with BM τ^2 variability, exons with different Ψ_{opt} evolve under similar selective strengths but variable rates of neutral evolution. These differences can be simultaneously explained by variation in the N_e across the genome, but also in the fitness landscapes and genetic variance (Figure 4.26). Recently, alternative exons i.e. with intermediate inclusion rates, have been shown to be more sensitive to mutations and binding of trans-regulatory elements both in deep mutagenesis experiments and human population data [16, 17]. This can be explained by a delayed competition model between splice sites, and provides an explanation to why genetic variance is larger when Ψ_{opt} are far from 100%, without necessarily having more mutations affecting the trait [16]. Similarly, it provides an additional explanation for the correlation between Ψ evolutionary rate of exons from the same gene: the time before competition of splice sites depends on intron length, but also on transcription elongation rate. The slower the elongation, the less competition between splice sites and less sensitive the resulting Ψ should be to mutations. In this line, we found a small but significant negative association between downstream intron length and the exon evolutionary rate (Table 4.1). Moreover, there are some factors, like post-transcriptional splicing, that affect the time before competition of splice sites. Thus, post-transcriptionally spliced genes are expected to be more sensitive to genetic variation, as recently shown [83], and therefore diverge faster during evolution.

Previous work proposed that fast divergence of AS patterns can be explained by lineage-specific adaptations [21, 88]. We have expanded our basic OU framework to account for potential changes in optimal inclusion rates along mammalian evolutionary history. This method, however, was not able to reliably identify shifts in simulated data under the inferred parameters, nor even its overall prevalence (Figure 4.29). We hypothesize that this is not really a methodological limitation, since decreasing the equilibrium variance in the simulated data significantly improves the performance of the method and enables inference, at least, of the proportion of exons with lineage specific-shifts. It is therefore limited by the large rates of neutral evolution that characterize the evolution of exon Ψs, which make adaptive

changes virtually indistinguishable from neutral divergence, at least at these long evolutionary distances. Although neutral divergence is a certainly more parsimonious explanation, as previously discussed [88], the prevalence of lineage-specific adaptive changes in AS rates remains unknown and difficult to infer with current methods and data. Denser phylogenetic data or even population level data, with higher time scale resolution may help to distinguish rapid adaptive change from neutral variation in future studies.

5.4.1 Study limitations

Despite the improvements and knowledge generated with the application of models of phenotypic evolution to AS patterns, this approach still has some important limitations, very similar to those already discussed for gene expression evolution [71]. The first issue is the assumption of independent evolution of exons. We have assumed independence at two different levels: genetic and selective independence. While the former suggests that mutations rarely affect different AS events simultaneously, such that accumulated mutations have independent phenotypic effects on inclusion rates, the latter assumes that the fitness consequences of quantitative variation in the Ψ of a given exon is independent from the Ψ in other exons.

Genetic independence is unlikely to hold for a number of reasons: AS variability is generally modulated by cis and trans regulatory elements. The mere existence of trans-regulatory elements and a regulatory network suggests that different AS events will be genetically linked through the underlying regulatory network driving variation along development, tissues and environmental conditions. As we showed in the first section, plastic AS changes take place through regulatory networks. The same would happen if mutations in trans factors influence their activity. Large genetic association studies like GTex have found genetic variants associated in trans to AS diversity [3]. As both cis and trans acting mutations fix and differentially accumulate in different populations and species, one can reasonably expect a co-evolution of co-regulated exons through trans-mechanisms. This is indeed the case, since analysis of parental and hybrid lines of mouse and *Drosophila* species suggests a substantial contribution of trans-effects to AS divergence [191, 155]. This, together with known factors affecting every exon in a gene e.g. elongation rate or post-transcriptional splicing, suggest that our assumption might not be fully accurate and provide a point for future development. If a multivariate BM accounting for correlations across morphological character proved superior for estimation of divergence times [10], we can expect that evolutionary rates inferred under known divergence times would also improve using a multivariate model for evolution that takes into account correlations between exons.

Despite no evidence of widespread non-independence in the fitness landscapes of exons Ψ, there are some known scenarios in which the fitness consequences of AS variation across different exons may not be independent. This is the case of mutually exclusive exons. These exons, often originated by tandem duplication, encode functionally similar protein sequences, such that inclusion of either of them separately yields a functional protein. However, if both were simultaneously included, it would certainly disrupt the protein function with the derived consequences on fitness. These pairs of exons, which are generally highly conserved [1, 2], would have anti-correlated fitness impacts: including one or the other but no the two simultaneously can produce a functional protein. Indeed, it may very well be that their mutually exclusive properties may not be random, but actually driven by selection. Another example of potentially correlated fitness consequences of different AS events is found in gene expression modulation through RUST [150]. In this case, fitness may depend on a combination of the outcomes of the AS events producing aberrant transcripts recognized by the NMD machinery. Therefore, the fitness consequences of changes in the inclusion of one exon may depend on how much unproductive splicing is generated by the remaining AS events. Linear fitness relationships across phenotypic dimensions e.g. exons, may be taken into account and inferred by extending the OU model to a multivariate framework [22, 56, 147]. However,

doing it for a large number of exons require working with increasingly larger covariance matrices, which poses additional computational challenges.

An additional assumption made by models of phenotypic evolution is that genetic variance, and hence rates of evolution, are independent of the phenotypic trait values [146]. This necessarily requires a distribution of mutational effects centered at 0 and independent of the trait value [71]. The latter is clearly violated by exons Ψ being bounded between 0 and 1, such that as Ψ gets closer to 1 or 0, the potential mutational effects start to be biased. We have somehow dealt with this limitation by assuming that the evolving character is its logit transformation $log(\frac{\Psi}{1-\Psi})$ or log-isoform ratios. Although this new character is mathematically unbounded, one may still expect fewer genotypes encoding for very large or very small Ψs, as suggested by our results on varying τ^2 across Ψ_{opt}. In this line, recent work has indeed showed that mutational effects depend on the starting Ψ [16]. This mutational bias may actually provide an alternative explanation for a constrained phenotypic divergence along evolution other than stabilizing selection [71]. Given the complex nature of the interaction between trait values and their evolutionary rates, more specific models taking into account the specific genetic architecture of exon Ψ and fitness landscape [188] will be required to better understand the genetic forces shaping AS evolution.

5.5 Integrative hierarchical modeling of alternative splicing data

There are two types of biological questions that we aim to answer when using transcriptomic data and bioinformatics analyses. The first, probably coming from a tradition in molecular biology, focuses on the specific details of the units under study. For example, when studying gene expression, the objective under this perspective is to identify whether each specific gene is changing in a certain biological process or remains constant. Thus, we need to have statistical methods that allow answering these questions specifically. This is the approach that most tools to infer gene expression levels and test whether they are different across different biological conditions take [215, 229, 275, 39, 157]. Similarly, when studying alternative splicing, we would want to know the details and have good estimations about the proportions of full isoforms [128, 39] or of each local variation in the splicing process [128, 248, 249, 114, 261, 287]. Under this perspective, genome-wide data becomes some sort of database or library to query specific biological information about the genetic element that we want to study e.g. gene, exon, transcript, enhancer or genetic variant. This is the natural statistical approach when one aims to identify which mutations or biomarkers e.g. genes, proteins, metabolites, are associated to disease susceptibility.

The second type of questions, on the other hand, shifts the focus from the individual entities e.g. genes or exons, to study the general patterns and rules that govern the observed profiles and their changes. This approach may be used to characterize the biological functions that are modulated or influenced by certain experimental treatment or mutated gene, and to understand the functional consequences in more general terms. It may also serve to identify the factors the drive the transcriptomic differences between conditions e.g. gene properties, transcription factor binding sites, or miRNA target sequences. From this perspective, the type of answers we aim to obtain are different: we are not really interested in knowing whether a specific gene is affected by a particular process, but we want to know, for instance, which are the trans-regulatory elements driving the transcriptomic changes between conditions.

Despite these philosophical differences, the common approach to tackle general questions still goes through performing statistical tests across every genetic element in the dataset e.g. identifying a set of elements for which we can confidently say that are changing in our experimental settings, even if this is not the question we want to ask our data. Then, clusters of elements are defined based on rather arbitrary thresholds to investigate which annotations or properties are different or more frequent in some groups compared with others, for which we need to perform again a large number of statistical tests on these newly generated datasets from our observations. These are the basis for a large collection of computational tools

and analysis aiming to characterize the biological functions that are regulated at the transcriptomic level or identify candidate regulatory elements driving them [290, 136, 189, 119]. This approach, although far from ideal, is what has been applied in this thesis to identify AS regulatory elements throughout heart development and disease. Although additional sources of information, especially gene expression changes, were used to identify PTBP1 as the main regulatory protein during heart disease, this enrichment approach has certainly helped to identify which AS regulator was driving AS changes in these two mouse models of heart disease.

The main limitation of this enrichment approach lies in the need to create new qualitative data from the actual observations: the assignment of each element to a group. By doing this, we are somehow assuming a discrete nature of the data, this is, that there are true underlying groups from which genes are sampled when we collect the data. Regardless of the validity of this assumption, this becomes something more similar to a clustering analysis, in which statistical significance seems a poor criteria to assign a group to each element. In practice, this makes the group definition highly dependent on the statistical power of tests whose answers were are not even interested in. Other tools, led by the popular GSEA, [264, 134, 15, 174, 314] go away the assumption of discrete gene categories associated to different biological process or regulatory features and start to use quantitative information to identify sets of genes that are particularly affected without relying on previous statistical tests. While these methods may be more statistically appropriate to the questions at hand, they still rely on an intermediate inferential step for the quantification of gene expression for each independent sample and gene, rather than on the raw observations. While we do not really know the influence of errors when estimating gene expression levels in the performance of GSEA, one may think it is small due to a large number of reads usually supporting each gene [39, 215, 160]. For the same reason, we can predict that error propagation will be worse when analyzing AS data, as there are much fewer reads that differentiate each alternative processing option within a gene, yielding our estimates much more noisy than when analyzing gene expression. Therefore, there are also practical reasons for a more principled and integrated statistical analysis of global patterns of alternative splicing.

In this thesis, we have made two attempts at extracting the information of interest directly from the data. First, we developed dSreg to infer the changes in the activity of regulatory proteins driving AS changes between different biological conditions. Second, we assumed that exons evolved under a common OU process to characterize the process that drives Ψ evolution along a phylogenetic tree. In both cases, we conditioned directly on the observations i.e. number of reads supporting inclusion and skipping of an exon, rather than on Ψ point estimates, which may be noisy for most of the exons. This allowed us to weight the contribution of each exon to the global parameters i.e. changes in the activities of regulatory factors or evolutionary parameters, according to the degree of information they provided. However, one may argue that to do so one needs to specify *a priori* a parametric model for the relationships between different exons or genes. In some situations, there are already theoretical models to describe the underlying biological process under specific parametrizations that provide the answer to our questions. This was the case of the OU model, which we merely extended to account for the binomial nature of AS data. This model extension allowed us to infer the contribution of stabilizing selection and phenotypic drift to AS evolution. In other situations, we do not know theoretical models to plug into our inferential framework, so we have to make principled assumptions about how the parameters answering our question relate to the observed data. This is what we did in dSreg, which models changes in the $logit(\Psi)$ as the sum of effects of the regulatory elements binding to each exon with some error. While the differences in performance of dSreg between simulated and real data suggest that this model may not be very realistic, it still outperforms the two-step approaches in simulated data and in the identification of the knocked-down RBPs across the ENCODE dataset. Thus, even if we are not sure about the underlying model, a simple linear approximation may still improve and outperform other strategies like ORA or GSEA.

Overall, these results show that carefully modelling the data generating process in our experimental design can not only provide additional biological information directly through the model parameters and their posterior distributions e.g. regulatory activities or evolutionary forces, but also improve our estimates of the derived inclusion rates thanks to information and uncertainty propagation across the different layers of the model. The inferential framework used here for modelling AS evolution and regulation can be extended to model other complex processes for better understanding how AS patterns change across different conditions and species.

5.6 Summary and perspective

The development of this thesis has greatly benefited from the incorporation of very different perspectives and frameworks to the study of AS in general. We have studied a diverse range of topics of different degree of generality and detail: from the identification of potential AS changes that may have an impact in the physiology and function of the heart and the specific factors that modulate it under different experimental conditions, towards a more general understanding of the regulatory patterns i.e. contribution of interaction between RBPs and their binding to the pre-mRNA, and how to infer them from the data. We have also zoomed out from a single species and investigated the contribution of different evolutionary forces to shape current AS patterns through phylogenetic comparative methods. During this path, we have developed the required statistical methods for each task, trying to improve on the limitations that previous and more general methods had for our specific aims. While we are still far from complete understanding of the importance and complexity of AS, how it is dynamically regulated to respond to different stimuli and the rules that govern this regulation and its evolution, we hope that this work may at least differentially get us closer to this aim.

6. Conclusions

Based on the presented results throughout this thesis and despite a large remaining uncertainty about many aspects, we have reached a number of conclusions:

1. Global AS patterns can distinguish disease samples from healthy adult hearts, as a partial regression in the developmental axis.

2. AS and GE changes affect independent genes and functions throughout developmental and disease transitions. While GE changes are associated to a large variety of functions, specific to the conditions under study, AS are always associated to cytoskeleton and actin binding functions, regardless of the tissue or experimental conditions.

3. AS changes in the heart are associated to changes in protein-protein interactions mediated by short motifs that are important for the overall connectivity of the network.

4. PTBP1 is the main driver of AS changes in heart disease and its over-expression is sufficient to induce heart dysfunction in mouse.

5. We have developed dSreg, a computational tool for the inference of active regulatory elements driving AS changes between different conditions that outperforms previous methods in both simulated and real data.

6. We have developed computational tools to apply models of phenotypic evolution to study AS patterns across mammals and proved that we can infer the parameters describing the evolutionary process under certain assumptions

7. Exon Ψ evolves under weak but significant stabilizing selection, comparable to that previously reported to constrain gene expression.

8. Diversity in exon Ψ evolutionary rates is mostly driven by variation in the random component rather than by the strength of selection acting on each exon.

9. We have inferred optimal inclusion rates across mammals and found that only about 5% of the genes are predicted to show functional exon skipping despite the widespread prevalence of AS across mammalian genomes

10. Neutral evolution of exon Ψ is so high that we cannot distinguish it from rapid adaptive changes at this large phylogenetic scale.

7. References

Bibliography

[1] Federico Abascal, Michael L. Tress, and Alfonso Valencia. "The evolutionary fate of alternatively spliced homologous exons after gene duplication". In: *Genome Biol. Evol.* 7.6 (2015), pp. 1392–1403. ISSN: 17596653. DOI: 10.1093/gbe/evv076.

[2] Federico Abascal et al. "Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level". In: *PLoS Comput. Biol.* 11.6 (June 2015). Ed. by Lukas Käll, pp. 1–29. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004325. URL: https://dx.plos.org/10.1371/journal.pcbi.1004325.

[3] François Aguet et al. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *bioRxiv* (2019), p. 787903. DOI: 10.1101/787903. URL: https://www.biorxiv.org/content/10.1101/787903v1.

[4] Gael P Alamancos, Eneritz Agirre, and Eduardo Eyras. *Methods to study splicing from high-throughput RNA sequencing data*. Ed. by Klemens J Hertel. Totowa, NJ, 2014. DOI: 10.1007/978-1-62703-980-2_26. URL: http://arxiv.org/pdf/1304.5952v2.pdf%20https://doi.org/10.1007/978-1-62703-980-2%7B%5C_%7D26 (visited on 01/18/2015).

[5] Alexander V. Alekseyenko, Namshin Kim, and Christopher J. Lee. "Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes". In: *Rna* 13.5 (2007), pp. 661–670. ISSN: 13558382. DOI: 10.1261/rna.325107.

[6] Babak Alipanahi et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". In: *Nat. Biotechnol.* 33.8 (Aug. 2015), pp. 831–838. ISSN: 15461696. DOI: 10.1038/nbt.3300. arXiv: 9605103 [cs]. URL: http://www.nature.com/articles/nbt.3300.

[7] Frederick W. Alt et al. "Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3 ends". In: *Cell* 20.2 (1980), pp. 293–301. ISSN: 00928674. DOI: 10.1016/0092-8674(80)90615-7.

[8] Adrian M Altenhoff et al. "Standardized benchmarking in the quest for orthologs". In: *Nat. Methods* 13.5 (May 2016), pp. 425–430. ISSN: 1548-7091. DOI: 10.1038/nmeth.3830. URL: http://www.nature.com/articles/nmeth.3830.

[9] Stephen F. Altschul et al. "Basic local alignment search tool". In: *J. Mol. Biol.* 215.3 (1990), pp. 403–410. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2.

[10] Sandra Álvarez-Carretero et al. "Bayesian Estimation of Species Divergence Times Using Correlated Quantitative Characters". In: *Syst. Biol.* 68.6 (2019), pp. 967–986. ISSN: 1076836X. DOI: 10.1093/sysbio/syz015.

[11] E. G. Eg G Ames et al. “Sequencing of mRNA identifies re-expression of fetal splice variants in cardiac hypertrophy”. In: *J. Mol. Cell. Cardiol.* 62 (Sept. 2013), pp. 99–107. ISSN: 00222828. DOI: 10.1016/j.yjmcc.2013.05.004. arXiv: NIHMS150003. URL: http://www.sciencedirect.com/science/article/pii/S0022282813001727?via%7B%5C%%7D3Dihub%20https://cnic.gtbib.net/sod/usu/M-CNIC/documentos/%7B%5C%%7D21FMCMARTI%7B%5C_%7DM-CNIC%7B%5C_%7D423327445%7B%5C_%7Djmolcellcardiol%7B%5C_%7D2013%7B%5C_%7D6299107..pdf.

[12] Simon Anders et al. “Detecting differential usage of exons from RNA-seq data.” In: *Genome Res.* 22.10 (Oct. 2012), pp. 2008–17. ISSN: 1549-5469. DOI: 10.1101/gr.133744.111. URL: http://genome.cshlp.org/content/22/10/2008%20http://precedings.nature.com/doifinder/10.1038/npre.2012.6837.

[13] Minna-Liisa Änkö et al. “The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes”. In: *Genome Biol.* 13.3 (2012), R17. ISSN: 1465-6906. DOI: 10.1186/gb-2012-13-3-r17. URL: http://genomebiology.com/2012/13/3/R17.

[14] Gil Ast. “How did alternative splicing evolve?” In: *Nat. Rev. Genet.* 5.10 (2004), pp. 773–782. ISSN: 14710056. DOI: 10.1038/nrg1451.

[15] C. Backes et al. “GeneTrail–advanced gene set enrichment analysis”. In: *Nucleic Acids Res.* 35.Web Server (May 2007), W186–W192. ISSN: 0305-1048. DOI: 10.1093/nar/gkm323. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm323.

[16] Pablo Baeza-Centurion et al. “Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing”. In: *Cell* 176.3 (Jan. 2019), 549–563.e23. ISSN: 10974172. DOI: 10.1016/j.cell.2018.12.010. URL: https://www.sciencedirect.com/science/article/pii/S0092867418316246?via%7B%5C%%7D3Dihub.

[17] Pablo Baeza-Centurion et al. “Mutations primarily alter the inclusion of alternatively spliced exons”. In: *bioRxiv* (2020), pp. 1–63.

[18] Francisco E. Baralle and Jimena Giudice. “Alternative splicing as a regulator of development and tissue identity”. In: *Nat. Rev. Mol. Cell Biol.* 18.7 (2017), pp. 437–451. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.27. URL: http://www.nature.com/doifinder/10.1038/nrm.2017.27.

[19] Marco Baralle and Francisco Ernesto Baralle. *The splicing code*. Feb. 2018. DOI: 10.1016/j.biosystems.2017.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0303264717303210?via%7B%5C%%7D3Dihub.

[20] Yoseph Barash et al. “Deciphering the splicing code”. In: *Nature* 465.7294 (2010), pp. 53–59. ISSN: 00280836. DOI: 10.1038/nature09000.

[21] Nuno L. Barbosa-Morais et al. “The Evolutionary Landscape of Alternative Splicing Vertebrate Species”. In: *Science (80-.).* 338.December (2012), pp. 1587–1593. DOI: 10.1126/science.1230612.

[22] Krzysztof Bartoszek et al. “A phylogenetic comparative method for studying multivariate adaptation”. In: *J. Theor. Biol.* 314 (Dec. 2012), pp. 204–215. ISSN: 0022-5193. DOI: 10.1016/J.JTBI.2012.08.005. URL: https://www.sciencedirect.com/science/article/pii/S0022519312003918.

[23] Giacomo Baruzzo et al. “Simulation-based comprehensive benchmarking of RNA-seq aligners”. In: *Nat. Methods* 14.2 (Feb. 2017), pp. 135–139. ISSN: 15487105. DOI: 10.1038/nmeth.4106. URL: http://www.ncbi.nlm.nih.gov/pubmed/27941783%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5792058%20http://www.nature.com/articles/nmeth.4106.

[24] Elad Bassat et al. "The extracellular matrix protein agrin promotes heart regeneration in mice". In: *Nature* 547.7662 (2017), pp. 179–184. ISSN: 14764687. DOI: 10.1038/nature22978.

[25] Douglas Bates et al. "Fitting Linear Mixed-Effects Models using lme4". In: *J. Stat. Softw.* 67.1 (2014). ISSN: 0092-8615. DOI: 10.18637/jss.v067.i01. arXiv: 1406.5823. URL: http://arxiv.org/abs/1406.5823.

[26] Trevor Bedford and Daniel L Hartl. "Optimization of gene expression by natural selection". In: *Proc. Natl. Acad. Sci.* 106.4 (Jan. 2009), pp. 1133–1138. ISSN: 0027-8424. DOI: 10.1073/pnas.0812009106. URL: http://www.pnas.org/content/106/4/1133.full.pdf?with-ds=yes%20http://www.pnas.org/cgi/doi/10.1073/pnas.0812009106%20http://www.ncbi.nlm.nih.gov/pubmed/19139403%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2633540.

[27] Bridget E Begg et al. "Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity". In: (). DOI: 10.1038/s41594-020-0475-8. URL: https://doi.org/10.1038/s41594-020-0475-8.

[28] Leslie R. Bell et al. "Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins". In: *Cell* 55.6 (1988), pp. 1037–1046. ISSN: 00928674. DOI: 10.1016/0092-8674(88)90248-6.

[29] Susan M Berget, Claire Moore, and Phillip A Sharp. "Spliced segments at the 5' terminus of adenovirus 2 late mRNA". In: *Proc. Natl. Acad. Sci. USA* 74.8 (1977), pp. 3171–3175.

[30] M. J. Betancourt and Mark Girolami. "Hamiltonian Monte Carlo for Hierarchical Models". In: *Arxiv* (2013). DOI: 10.1201/b18502-5. arXiv: 1312.0906. URL: http://arxiv.org/abs/1312.0906.

[31] Michael Betancourt. "A Conceptual Introduction to Hamiltonian Monte Carlo". In: (2017). arXiv: 1701.02434. URL: http://arxiv.org/abs/1701.02434.

[32] Shamsuddin A. Bhuiyan et al. "Systematic evaluation of isoform function in literature reports of alternative splicing". In: *BMC Genomics* 19.1 (Aug. 2018), p. 637. ISSN: 14712164. DOI: 10.1186/s12864-018-5013-2. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5013-2.

[33] C. C F Blake. *Do genes-in-pieces imply proteins-in-pieces?* 1978. DOI: 10.1038/273267a0.

[34] Benjamin J. Blencowe. *The Relationship between Alternative Splicing and Proteomic Complexity.* June 2017. DOI: 10.1016/j.tibs.2017.04.001. URL: http://www.sciencedirect.com/science/article/pii/S0968000417300701?via%7B%5C%%7D3Dihub%20http://www.ncbi.nlm.nih.gov/pubmed/28483376.

[35] K. Blin et al. "DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation". In: *Nucleic Acids Res.* 43.D1 (2015), pp. D160–D167. ISSN: 0305-1048. DOI: 10.1093/nar/gku1180. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1180.

[36] Paul L. Boutz et al. "A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons". In: *Genes Dev.* 21.13 (2007), pp. 1636–1652. DOI: 10.1101/gad.1558107.

[37] Robert K. Bradley et al. "Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution". In: *PLoS Biol.* 10.1 (Jan. 2012). Ed. by Laurence D. Hurst, e1001229. ISSN: 15449173. DOI: 10.1371/journal.pbio.1001229. URL: http://dx.plos.org/10.1371/journal.pbio.1001229.

[38] David Brawand et al. “The evolution of gene expression levels in mammalian organs”. In: *Nature* 478.7369 (2011), pp. 343–348. ISSN: 0028-0836. DOI: 10.1038/nature10532. URL: http://dx.doi.org/10.1038/nature10532%7B%5C%%7D5Cnpapers2://publication/doi/10.1038/nature10532.

[39] Nicolas L Bray et al. “Near-optimal probabilistic RNA-seq quantification”. In: *Nat. Biotechnol.* 34.5 (May 2016), pp. 525–527. ISSN: 1087-0156. DOI: 10.1038/nbt.3519. arXiv: 1505.02710. URL: http://www.nature.com/doifinder/10.1038/nbt.3519%20http://www.nature.com/articles/nbt.3519%20http://www.ncbi.nlm.nih.gov/pubmed/27043002.

[40] Angela N Brooks et al. “Regulation of alternative splicing in Drosophila by 56 RNA binding proteins”. In: *Genome Res.* 25.11 (2015), pp. 1771–1780. ISSN: 15495469. DOI: 10.1101/gr.192518.115. URL: https://genome.cshlp.org/content/25/11/1771.full.pdf.

[41] Marija Buljan et al. “Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks”. In: *Mol. Cell* 46.6 (June 2012), pp. 871–883. ISSN: 1097-2765. DOI: 10.1016/J.MOLCEL.2012.05.039. URL: https://www.sciencedirect.com/science/article/pii/S1097276512004844?via%7B%5C%%7D3Dihub.

[42] Marguerite A. Butler and Aaron A. King. “Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution”. In: *Am. Nat.* 164.6 (2004), pp. 683–695. ISSN: 0003-0147. DOI: 10.1086/426002. URL: http://www.jstor.org/stable/10.1086/426002%7B%5C%%7D5Cnhttp://www.jstor.org/discover/10.1086/426002?uid=2%7B%5C&%7Duid=4%7B%5C&%7Dsid=21103411612157%7B%5C%%7D5Cnhttp://www.journals.uchicago.edu/doi/10.1086/426002.

[43] John A. Calarco et al. “Global analysis of alternative splicing differences between humans and chimpanzees”. In: *Genes Dev.* 21.22 (2007), pp. 2963–2975. ISSN: 08909369. DOI: 10.1101/gad.1606907.

[44] Zhe Xu Cao et al. “Comprehensive investigation of alternative splicing and development of a prognostic risk score for prostate cancer based on six-gene signatures”. In: *J. Cancer* 10.22 (2019), pp. 5585–5596. ISSN: 18379664. DOI: 10.7150/jca.31725.

[45] Bob Carpenter et al. “Stan: A Probabilistic Programming Language”. In: *J. Stat. Softw.* 76.1 (2017). ISSN: 1548-7660. DOI: 10.18637/jss.v076.i01. URL: http://www.jstatsoft.org/v76/i01/.

[46] Carlos M Carvalho, Nicholas G Polson, and James G Scott. “Handling sparsity via the horseshoe”. In: *J. Mach. Learn. Res.* 5 (2009), pp. 73–80. ISSN: 1533-7928. URL: http://proceedings.mlr.press/v5/carvalho09a/carvalho09a.pdf%20http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Handling+Sparsity+via+the+Horseshoe%7B%5C#%7D0.

[47] Ana Catalán et al. “Evolution of sex-biased gene expression and dosage compensation in the eye and brain of heliconius butterflies”. In: *Mol. Biol. Evol.* 35.9 (Sept. 2018). Ed. by Gregory Wray, pp. 2120–2134. ISSN: 15371719. DOI: 10.1093/molbev/msy111. URL: https://academic.oup.com/mbe/article/35/9/2120/5017354.

[48] Francesco Catania, Xiang Gao, and Douglas G Scofield. “Endogenous mechanisms for the origins of spliceosomal introns.” In: *J. Hered.* 100.5 (2009), pp. 591–6. ISSN: 1465-7333. DOI: 10.1093/jhered/esp062. URL: http://www.ncbi.nlm.nih.gov/pubmed/19635762%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2877546.

[49] Francesco Catania and Michael Lynch. "A simple model to explain evolutionary trends of eukaryotic gene architecture and expression: How competition between splicing and cleavage/polyadenylation factors may affect gene expression and splice-site recognition in eukaryotes". In: *BioEssays* 35.6 (June 2013), pp. 561–570. ISSN: 1521-1878. DOI: 10.1002/bies.201200127. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201200127%20http://www.ncbi.nlm.nih.gov/pubmed/23568225%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4968935.

[50] Saurabh Chaudhary et al. *Alternative splicing and protein diversity: Plants versus animals*. June 2019. DOI: 10.3389/fpls.2019.00708. URL: https://www.frontiersin.org/article/10.3389/fpls.2019.00708/full.

[51] Mo Chen and James L Manley. "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." In: *Nat. Rev. Mol. Cell Biol.* 10.11 (Nov. 2009), pp. 741–54. ISSN: 1471-0080. DOI: 10.1038/nrm2777. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2958924%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[52] Nancy Chen et al. "Allele frequency dynamics in a pedigreed natural population". In: *Proc. Natl. Acad. Sci.* 116.6 (2019), pp. 2158–2164. ISSN: 0027-8424. DOI: 10.1073/pnas.1813852116.

[53] Ting-Wen Chen et al. "Interrogation of alternative splicing events in duplicated genes during evolution". In: *BMC Genomics* 12.Suppl 3 (2011), S16. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-S3-S16. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3333175%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[54] Dmitry Cherny et al. "Stoichiometry of a regulatory splicing complex revealed by single-molecule analyses". In: *EMBO J.* 29.13 (July 2010), pp. 2161–2172. ISSN: 02614189. DOI: 10.1038/emboj.2010.103. URL: http://emboj.embopress.org/cgi/doi/10.1038/emboj.2010.103.

[55] Louise T. Chow et al. "A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids". In: *Cell* 11.4 (1977), pp. 819–836. ISSN: 00928674. DOI: 10.1016/0092-8674(77)90294-X.

[56] Julien Clavel, Gilles Escarguel, and Gildas Merceron. "mv morph : an r package for fitting multivariate evolutionary models to morphometric data". In: *Methods Ecol. Evol.* 6.11 (Nov. 2015). Ed. by Timothée Poisot, pp. 1311–1319. ISSN: 2041210X. DOI: 10.1111/2041-210X.12420. URL: http://doi.wiley.com/10.1111/2041-210X.12420.

[57] Héctor Climente-GonzáLez et al. "The Functional Impact of Alternative Splicing in Cancer". In: *CellReports* 20.9 (Aug. 2017), pp. 2215–2226. ISSN: 22111247. DOI: 10.1016/j.celrep.2017.08.012. URL: http://dx.doi.org/10.1016/j.celrep.2017.08.012%20https://www.sciencedirect.com/science/article/pii/S221112471731104X?via%7B%5C%%7D3Dihub.

[58] Natalie Cooper et al. "A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies". In: *Biol. J. Linn. Soc.* 118.1 (2016), pp. 64–77. ISSN: 10958312. DOI: 10.1111/bij.12701. arXiv: arXiv:1408.1149.

[59] Steven A. Crone et al. "ErbB2 is essential in the prevention of dilated cardiomyopathy". In: *Nat. Med.* 8.5 (May 2002), pp. 459–465. ISSN: 1078-8956. DOI: 10.1038/nm0502-459. URL: http://www.nature.com/articles/nm0502-459.

[60] Gabriele D'Uva et al. "ERBB2 triggers mammalian heart regeneration by promoting cardiomyocyte dedifferentiation and proliferation". In: *Nat. Cell Biol.* 17.5 (May 2015), pp. 627–638. ISSN: 14764679. DOI: 10.1038/ncb3149. URL: http://www.nature.com/articles/ncb3149.

[61] J. Darnell. "Implications of RNA-RNA splicing in evolution of eukaryotic cells". In: *Science (80-.).* 202.4374 (1978), pp. 1257–1260. ISSN: 0036-8075. DOI: 10.1126/science.364651.

[62] Donald M. Dixon et al. "Loss of muscleblind-like 1 results in cardiac pathology and persistence of embryonic splice isoforms". In: *Sci. Rep.* 5.1 (Aug. 2015), p. 9042. ISSN: 2045-2322. DOI: 10.1038/srep09042. URL: http://www.nature.com/articles/srep09042.

[63] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bts635. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635.

[64] Daniel Dominguez et al. "Sequence, Structure and Context Preferences of Human RNA Binding Proteins". In: *Molecul* 70 (June 2018), pp. 854–7. ISSN: 10974164. DOI: 10.1101/201996. URL: http://linkinghub.elsevier.com/retrieve/pii/S1097276518303514%20http://dx.doi.org/10.1101/201996%7B%5C%%7D0Ahttps://www.biorxiv.org/content/early/2017/10/12/201996.

[65] W Ford Doolittle. "The origin and function of intervening sequences in DNA: A review". In: *Am. Nat.* 130.6 (1987), pp. 915–928.

[66] P. Early et al. "Two mRNAs can be produced from a single immunoglobulin $\mu$$*genebyalternativeRNAprocessingpath*
In: *Cell* 20.2 (1980), pp. 313–319. ISSN: 00928674. DOI: 10.1016/0092-8674(80)90617-0.

[67] R. C. Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic Acids Res.* 32.5 (Mar. 2004), pp. 1792–1797. ISSN: 1362-4962. DOI: 10.1093/nar/gkh340. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh340.

[68] David M. Emms and Steven Kelly. "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy". In: *Genome Biol.* 16.1 (Dec. 2015), p. 157. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0721-2. URL: http://genomebiology.com/2015/16/1/157%20http://dx.doi.org/10.1186/s13059-015-0721-2.

[69] A. J. Enright, S Van Dongen, and C A Ouzounis. "An efficient algorithm for large-scale detection of protein families". In: *Nucleic Acids Res.* 30.7 (Apr. 2002), pp. 1575–1584. ISSN: 13624962. DOI: 10.1093/nar/30.7.1575. URL: http://www.ncbi.nlm.nih.gov/pubmed/11917018%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC101833%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/30.7.1575.

[70] Iakes Ezkurdia et al. "Most highly expressed protein-coding genes have a single dominant isoform". In: *J. Proteome Res.* 14.4 (2015), pp. 1880–1887. DOI: 10.1021/pr501286b. arXiv: 15334406.

[71] J. C. Fay and P. J. Wittkopp. *Evaluating the role of natural selection in the evolution of gene regulation.* Feb. 2008. DOI: 10.1038/sj.hdy.6801000. URL: www.nature.com/hdy.

[72] Leanne E Felkin et al. "Calcineurin splicing variant calcineurin Aβ1 improves cardiac function after myocardial infarction without inducing hypertrophy." In: *Circulation* 123.24 (June 2011), pp. 2838–2847. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.110.012211. URL: http://www.ncbi.nlm.nih.gov/pubmed/21632490.

[73] Joseph Felsenstein. *Phylogenies and the comparative method.* 1985.

[74] Stefania Fochi et al. "The Emerging Role of the RBM20 and PTBP1 Ribonucleoproteins in Heart Development and Cardiovascular Diseases". In: *Genes (Basel).* 11.402 (2020), pp. 1–21. ISSN: 00471852. DOI: 10.1007/978-94-009-4285-1_6.

[75] Brent L. Fogel et al. "RBFOX1 regulates both splicing and transcriptional networks in human neuronal development". In: *Hum. Mol. Genet.* 21.19 (2012), pp. 4171–4186. ISSN: 09646906. DOI: 10.1093/hmg/dds240.

[76] Xiang Dong Fu. "Towards a splicing code". In: *Cell* 119.6 (2004), pp. 736–738. ISSN: 00928674. DOI: 10.1016/j.cell.2004.11.039.

[77] Xiang-Dong Dong Fu and Manuel Ares. "Context-dependent control of alternative splicing by RNA-binding proteins". In: *Nat. Rev. Genet.* 15.10 (Aug. 2014), pp. 689–701. ISSN: 1471-0056. DOI: 10.1038/nrg3778. URL: http://www.nature.com/doifinder/10.1038/nrg3778.

[78] Tsukasa Fukunaga et al. "CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data." In: *Genome Biol.* 15.1 (2014), R16. ISSN: 1465-6914. DOI: 10.1186/gb-2014-15-1-r16. URL: http://genomebiology.com/2014/15/1/R16.

[79] Rickard G. Fred, Linda Tillmar, and Nils Welsh. "The Role of PTB in Insulin mRNA Stability Control". In: *Curr. Diabetes Rev.* 2.3 (Aug. 2006), pp. 363–366. ISSN: 15733998. DOI: 10.2174/157339906777950570. URL: http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C&%7Dissn=1573-3998%7B%5C&%7Dvolume=2%7B%5C&%7Dissue=3%7B%5C&%7Dspage=363.

[80] L. M. Gallego-Paez et al. "Alternative splicing: the pledge, the turn, and the prestige: The key role of alternative splicing in human biological systems". In: *Hum. Genet.* 136.9 (2017), pp. 1015–1042. ISSN: 14321203. DOI: 10.1007/s00439-017-1790-y.

[81] Chen Gao et al. "RBFox1-mediated RNA splicing regulates cardiac hypertrophy and heart failure". In: *j Clin Invest* 126.1 (2015), pp. 1–12. ISSN: 15588238. DOI: 10.1172/JCI84015DS1.

[82] Theodore Garland, Albert F Bennett, and Enrico L Rezende. "Phylogenetic approaches in comparative physiology." In: *J. Exp. Biol.* 208.Pt 16 (2005), pp. 3015–3035. ISSN: 0022-0949. DOI: 10.1242/jeb.01745.

[83] Diego Garrido-Martín et al. "Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome". In: *bioRxiv* (2020), p. 2020.05.29.123703. DOI: 10.1101/2020.05.29.123703. URL: https://doi.org/10.1101/2020.05.29.123703.

[84] Alberto Gatto et al. "FineSplice, enhanced splice junction detection and quantification: A novel pipeline based on the assessment of diverse RNA-Seq alignment solutions". In: *Nucleic Acids Res.* 42.Ext 3309 (2014), pp. 1–11. ISSN: 13624962. DOI: 10.1093/nar/gku166.

[85] Lauren T. Gehman et al. "The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain". In: *Nat. Genet.* 43.7 (2012), pp. 706–711. DOI: 10.1038/ng.841.The.

[86] Andrew Gelman et al. *Bayesian Data Analysis, Third Edition (Texts in Statistical Science)*. 2014, p. 675. ISBN: 9781439840955. DOI: 10.1007/s13398-014-0173-7.2. arXiv: arXiv:1011.1669v3.

[87] Mohamed Ali Ghadie et al. "Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing". In: *PLoS Comput. Biol.* 13.8 (Aug. 2017). Ed. by Andrey Rzhetsky, e1005717. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005717. URL: http://dx.plos.org/10.1371/journal.pcbi.1005717.

[88] Lauren Gibilisco et al. "Alternative Splicing within and between Drosophila Species, Sexes, Tissues, and Developmental Stages". In: *PLoS Genet.* 12.12 (2016), pp. 1–19. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006464.

[89] W. Gilbert. "The exon theory of genes". In: *Cold Spring Harb. Symp. Quant. Biol.* 52 (1987), pp. 901–905. ISSN: 00917451. DOI: 10.1101/SQB.1987.052.01.098.

[90] Walter Gilbert. *Why genes in pieces?* 1978. DOI: 10.1038/271501a0.

[91] Girolamo Giudice et al. "ATtRACT-a database of RNA-binding proteins and associated motifs". In: *Database* 2016.November (2016), pp. 1–9. ISSN: 17580463. DOI: 10.1093/database/baw035.

[92] Jimena Giudice et al. “Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development.” In: *Nat. Commun.* 5 (2014), p. 3603. ISSN: 2041-1723. DOI: 10.1038/ncomms4603. URL: http://www.ncbi.nlm.nih.gov/pubmed/24752171.

[93] Jimena Giudice et al. “Neonatal cardiac dysfunction and transcriptome changes caused by the absence of Celf1”. In: *Sci. Rep.* 6.1 (Dec. 2016), p. 35550. ISSN: 2045-2322. DOI: 10.1038/srep35550. URL: http://www.ncbi.nlm.nih.gov/pubmed/27759042%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5069560%20http://www.nature.com/articles/srep35550.

[94] Clare Gooding et al. “MBNL1 and PTB cooperate to repress splicing of Tpm1 exon 3”. In: *Nucleic Acids Res.* 41.9 (2013), pp. 4765–4782. ISSN: 03051048. DOI: 10.1093/nar/gkt168.

[95] Toni I Gossmann, Peter D Keightley, and Adam Eyre-Walker. “The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes”. In: *Genome Biol. Evol.* 4.5 (2012), pp. 658–667. ISSN: 17596653. DOI: 10.1093/gbe/evs027. URL: http://www.ncbi.nlm.nih.gov/pubmed/22436998%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3381672.

[96] Gregory R. Grant et al. “Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)”. In: *Bioinformatics* 27.18 (2011), pp. 2518–2528. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr427.

[97] Xavier Grau-Bové, Iñaki Ruiz-Trillo, and Manuel Irimia. “Origin of exon skipping-rich transcriptomes in animals driven by evolution of gene architecture”. In: *Genome Biol.* 19.1 (Dec. 2018), p. 135. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1499-9. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1499-9.

[98] Xavier Grau-Bové et al. “Dynamics of genomic innovation in the unicellular ancestry of animals”. In: *Elife* 6 (July 2017). ISSN: 2050084X. DOI: 10.7554/eLife.26036.

[99] Brenton R. Graveley. “Alternative splicing: Increasing diversity in the proteomic world”. In: *Trends Genet.* 17.2 (2001), pp. 100–107. ISSN: 01689525. DOI: 10.1016/S0168-9525(00)02176-4.

[100] Wei Guo et al. “Splicing factor RBM20 regulates transcriptional network of titin associated and calcium handling genes in the heart”. In: *Int. J. Biol. Sci.* 14.4 (2018), pp. 369–380. ISSN: 14492288. DOI: 10.7150/ijbs.24117.

[101] Abhishek K. Gupta et al. “Degenerate minigene library analysis enables identification of altered branch point utilization by mutant splicing factor 3B1 (SF3B1)”. In: *Nucleic Acids Res.* 47.2 (Jan. 2019), pp. 970–980. ISSN: 13624962. DOI: 10.1093/nar/gky1161. URL: https://academic.oup.com/nar/article/47/2/970/5193340.

[102] Shobhit Gupta et al. “Quantifying similarity between motifs”. In: *Genome Biol.* 8.2 (Feb. 2007), R24. ISSN: 14747596. DOI: 10.1186/gb-2007-8-2-r24. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-2-r24.

[103] Hong Han et al. “Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors”. In: *Mol. Cell* 65.3 (Feb. 2017), 539–553.e7. ISSN: 10974164. DOI: 10.1016/j.molcel.2017.01.011. URL: http://www.sciencedirect.com/science/article/pii/S1097276517300370?via%7B%5C%%7D3Dihub.

[104] Luke J. Harmon. *Phylogenetic comparative methods*. 2019. DOI: 10.1016/j.cub.2017.03.049.

[105] Lutz G.W. Hilgenberg et al. “Agrin regulation of α3 sodium-potassium ATPase activity modulates cardiac myocyte contraction”. In: *J. Biol. Chem.* 284.25 (2009), pp. 16956–16965. ISSN: 00219258. DOI: 10.1074/jbc.M806855200.

[106] Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: 15 (2011), pp. 1351–1381. ISSN: 15337928. arXiv: 1111.4246. URL: http://arxiv.org/abs/1111.4246.

[107] Yuanhua Huang and Guido Sanguinetti. "BRIE: Transcriptome-wide splicing quantification in single cells". In: *Genome Biol.* 18.1 (Dec. 2017), p. 123. ISSN: 1474760X. DOI: 10.1186/s13059-017-1248-5. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1248-5.

[108] Jason T. Huff, Daniel Zilberman, and Scott W. Roy. "Mechanism for DNA transposons to generate introns on genomic scales". In: *Nature* 538.7626 (Oct. 2016), pp. 533–536. ISSN: 14764687. DOI: 10.1038/nature20110. URL: http://dx.doi.org/10.1038/nature20110.

[109] Camilla Iannone and Juan Valcárcel. "Chromatin's thread to alternative splicing regulation." In: *Chromosoma* 122.6 (Dec. 2013), pp. 465–74. ISSN: 1432-0886. DOI: 10.1007/s00412-013-0425-x. URL: http://www.ncbi.nlm.nih.gov/pubmed/23912688.

[110] Camilla Iannone et al. "Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells". In: (2015), pp. 360–374. DOI: 10.1261/rna.048843.114.minant.

[111] Luis P. Iñiguez and Georgina Hernández. "The Evolutionary Relationship between Alternative Splicing and Gene Duplication". In: *Front. Genet.* 08.February (2017), pp. 1–7. ISSN: 1664-8021. DOI: 10.3389/fgene.2017.00014. URL: http://journal.frontiersin.org/article/10.3389/fgene.2017.00014/full.

[112] Manuel Irimia, David Penny, and Scott W. Roy. "Coevolution of genomic intron number and splice sites". In: *Trends Genet.* 23.7 (July 2007), pp. 318–321. ISSN: 01689525. DOI: 10.1016/j.tig.2007.04.001.

[113] Manuel Irimia and Scott William Roy. "Origin of spliceosomal introns and alternative splicing". In: *Cold Spring Harb. Perspect. Biol.* 6.6 (June 2014). ISSN: 19430264. DOI: 10.1101/cshperspect.a016071. URL: http://www.ncbi.nlm.nih.gov/pubmed/24890509%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4031966.

[114] Manuel Irimia et al. "A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains". In: *Cell* 159.7 (2014), pp. 1511–1523. ISSN: 00928674. DOI: 10.1016/j.cell.2014.11.035. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867414015128.

[115] Manuel Irimia et al. "Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing". In: *BMC Evol. Biol.* 7.1 (Oct. 2007), pp. 1–12. ISSN: 14712148. DOI: 10.1186/1471-2148-7-188. URL: http://dx.doi.org/10.1186/1471-2148-7-188.

[116] Manuel Irimia et al. *Origin of introns by 'intronization' of exonic sequences*. Aug. 2008. DOI: 10.1016/j.tig.2008.05.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0168952508001765.

[117] Manuel Irimia et al. "Quantitative regulation of alternative splicing in evolution and development". In: *BioEssays* 31.1 (2009), pp. 40–50. ISSN: 02659247. DOI: 10.1002/bies.080092.

[118] Ruiping Ji et al. "Increased de novo ceramide synthesis and accumulation in failing myocardium". In: *JCI insight* 2.9 (2017), pp. 1–19. ISSN: 23793708. DOI: 10.1172/jci.insight.82922.

[119] Xiaoli Jiao et al. "DAVID-WS: A stateful web service to facilitate gene/protein list analysis". In: *Bioinformatics* 28.13 (July 2012), pp. 1805–1806. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts251. URL: /pmc/articles/PMC3381967/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3381967/.

[120] Lihua Jin et al. "The evolutionary relationship between gene duplication and alternative splicing". In: *Gene* 427.1-2 (2008), pp. 19–31. ISSN: 03781119. DOI: 10.1016/j.gene.2008.09.002. URL: http://dx.doi.org/10.1016/j.gene.2008.09.002.

[121] Eric B. Johnson et al. "Defective splicing of Megf7/Lrp4, a regulator of distal limb development, in autosomal recessive mulefoot disease". In: *Genomics* 88.5 (2006), pp. 600–609. ISSN: 08887543. DOI: 10.1016/j.ygeno.2006.08.005.

[122] Jason M. Johnson et al. "Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays". In: *Science (80-.).* 302.5653 (2003), pp. 2141–2144. ISSN: 00368075. DOI: 10.1126/science.1090100.

[123] Philippe Julien et al. "The complete local genotype-phenotype landscape for the alternative splicing of a human exon". In: *Nat. Commun.* 7.1 (May 2016), pp. 1–8. ISSN: 20411723. DOI: 10.1038/ncomms11558.

[124] Alex T. Kalinka et al. "Gene expression divergence recapitulates the developmental hourglass model". In: *Nature* 468.7325 (Dec. 2010), pp. 811–814. ISSN: 0028-0836. DOI: 10.1038/nature09634. URL: http://www.nature.com/doifinder/10.1038/nature09634.

[125] Auinash Kalsotra and T a Cooper. "Functional consequences of developmentally regulated alternative splicing". In: *Nat. Rev. Genet.* 12.10 (2012), pp. 715–729. ISSN: 1471-0064. DOI: 10.1038/nrg3052.

[126] Auinash Kalsotra et al. "A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart." In: *Proc. Natl. Acad. Sci. U. S. A.* 105.51 (Dec. 2008), pp. 20333–8. ISSN: 1091-6490. DOI: 10.1073/pnas.0809045105. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2629332%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[127] Rahul N. Kanadia et al. "Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy". In: *Proc. Natl. Acad. Sci. U. S. A.* 103.31 (2006), pp. 11748–11753. ISSN: 00278424. DOI: 10.1073/pnas.0604970103.

[128] Yarden Katz et al. "Analysis and design of RNA sequencing experiments for identifying isoform regulation". In: *Nat. Methods* 7.12 (Dec. 2010), pp. 1009–1015. ISSN: 1548-7091. DOI: 10.1038/nmeth.1528. arXiv: 9605103 [cs]. URL: http://www.nature.com/articles/nmeth.1528%20http://www.nature.com/doifinder/10.1038/nmeth.1528.

[129] Hadas Keren-Shaul, Galit Lev-Maor, and Gil Ast. "Pre-mRNA splicing is a determinant of nucleosome organization". In: *Epigenetics Chromatin* 6.Suppl 1 (2013), P46. ISSN: 1756-8935. DOI: 10.1186/1756-8935-6-s1-p46.

[130] Hadas Keren, Galit Lev-Maor, and Gil Ast. "Alternative splicing and evolution: diversification, exon definition and function." In: *Nat. Rev. Genet.* 11.5 (May 2010), pp. 345–355. ISSN: 14710056. DOI: 10.1038/nrg2776. URL: http://dx.doi.org/10.1038/nrg2776.

[131] M. Khorshid, C. Rodak, and M. Zavolan. "CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins". In: *Nucleic Acids Res.* 39.Database (2011), pp. D245–D252. ISSN: 0305-1048. DOI: 10.1093/nar/gkq940. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq940.

[132] Daehwan Kim et al. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.” In: *Genome Biol.* 14.4 (2013), R36. ISSN: 1465-6914. DOI: 10.1186/gb-2013-14-4-r36. URL: http://genomebiology.com/2013/14/4/R36.

[133] Natalie Kim et al. “Lrp4 Is a Receptor for Agrin and Forms a Complex with MuSK”. In: *Cell* 135.2 (2008), pp. 334–342. ISSN: 00928674. DOI: 10.1016/j.cell.2008.10.002. URL: http://dx.doi.org/10.1016/j.cell.2008.10.002.

[134] Seon-Young Kim and David J Volsky. “PAGE: Parametric Analysis of Gene Set Enrichment”. In: *BMC Bioinformatics* 6.1 (June 2005), p. 144. ISSN: 14712105. DOI: 10.1186/1471-2105-6-144. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-144.

[135] Tomomi E Kimura et al. “Targeted deletion of the extracellular signal-regulated protein kinase 5 attenuates hypertrophic response and promotes pressure overload-induced apoptosis in the heart”. In: *Circ. Res.* 106.5 (Mar. 2010), pp. 961–970. ISSN: 00097330. DOI: 10.1161/CIRCRESAHA.109.209320. URL: http://www.ncbi.nlm.nih.gov/pubmed/20075332%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3003662.

[136] D. V. Klopfenstein et al. “GOATOOLS: A Python library for Gene Ontology analyses”. In: *Sci. Rep.* 8.1 (Dec. 2018), p. 10872. ISSN: 2045-2322. DOI: 10.1038/s41598-018-28948-z. URL: http://www.nature.com/articles/s41598-018-28948-z.

[137] Fyodor A. Kondrashov and Eugene V. Koonin. *Evolution of alternative splicing: Deletions, insertions and origin of functional parts of proteins from intron sequences*. Mar. 2003. DOI: 10.1016/S0168-9525(02)00029-X. URL: https://www.sciencedirect.com/science/article/pii/S016895250200029X?via%7B%5C%%7D3Dihub.

[138] Sek Won Kong et al. “Heart Failure Associated Changes in RNA Splicing of Sarcomere Genes”. In: *Circ. Cardiovasc. Genet.* 3.2 (2010), pp. 138–146. DOI: 10.1161/CIRCGENETICS.109.904698.

[139] Eugene V. Koonin. *Intron-dominated genomes of early ancestors of eukaryotes*. 2009. DOI: 10.1093/jhered/esp056. URL: https://academic.oup.com/jhered/article-lookup/doi/10.1093/jhered/esp056.

[140] Eugene V. Koonin. “The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate?” In: *Biol. Direct* 1 (2006), pp. 1–23. ISSN: 17456150. DOI: 10.1186/1745-6150-1-22.

[141] Naama M Kopelman, Doron Lancet, and Itai Yanai. “Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms.” In: *Nat. Genet.* 37.6 (2005), pp. 588–589. ISSN: 1061-4036. DOI: 10.1038/ng1575. URL: http://www.ncbi.nlm.nih.gov/pubmed/15895079.

[142] Andrea N Ladd, Nicolas Charlet-b, and Thomas a Cooper. “The CELF Family of RNA Binding Proteins Is Implicated in Cell-Specific and Developmentally Regulated Alternative Splicing The CELF Family of RNA Binding Proteins Is Implicated in Cell-Specific and Developmentally Regulated Alternative Splicing Downloaded ”. In: 21.4 (2001), pp. 1285–1296. DOI: 10.1128/MCB.21.4.1285.

[143] Andrea N Ladd et al. “Cardiac tissue-specific repression of CELF activity disrupts alternative splicing and causes cardiomyopathy.” In: *Mol. Cell. Biol.* 25.14 (2005), pp. 6267–6278. ISSN: 0270-7306. DOI: 10.1128/MCB.25.14.6267-6278.2005.

[144] Sunshine Lahmers et al. “Developmental Control of Titin Isoform Expression and Passive Stiffness in Fetal and Neonatal Myocardium”. In: *Circ. Res.* 94.4 (2004), pp. 505–513. ISSN: 00097330. DOI: 10.1161/01.RES.0000115522.52554.86.

[145] Matthew J. Lambert et al. "Evidence for widespread subfunctionalization of splice forms in vertebrate genomes". In: *Genome Res.* 125.5 (2015), pp. 624–632. ISSN: 15495469. DOI: 10.1101/gr.184473.114.

[146] Russell Lande. "Natural selection and random genetic drift in phenotype evolution". In: *Evolution (N. Y).* 30.2 (June 1976), pp. 314–334. ISSN: 00143820. DOI: 10.1111/j.1558-5646.1976.tb00911.x. URL: http://doi.wiley.com/10.1111/j.1558-5646.1976.tb00911.x.

[147] Russell Lande and Stevan J. Arnold. "THE MEASUREMENT OF SELECTION ON CORRELATED CHARACTERS". In: *Evolution (N. Y).* 37.6 (Nov. 1983), pp. 1210–1226. ISSN: 00143820. DOI: 10.1111/j.1558-5646.1983.tb00236.x. URL: http://doi.wiley.com/10.1111/j.1558-5646.1983.tb00236.x.

[148] Enrique Lara-Pezzi et al. "Neurogenesis: Regulation by Alternative Splicing and Related Posttranscriptional Processes". In: *Neurosci.* 23.5 (Oct. 2017), pp. 466–477. ISSN: 1073-8584. DOI: 10.1177/1073858416678604. URL: http://journals.sagepub.com/doi/10.1177/1073858416678604.

[149] Enrique Lara-Pezzi et al. "The alternative heart: Impact of alternative splicing in heart disease". In: *J. Cardiovasc. Transl. Res.* 6.6 (2013), pp. 945–955. ISSN: 19375387. DOI: 10.1007/s12265-013-9482-z.

[150] Liana F. Lareau and Steven E. Brenner. "Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible". In: *Mol. Biol. Evol.* 32.4 (Apr. 2015), pp. 1072–1079. ISSN: 1537-1719. DOI: 10.1093/molbev/msv002. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msv002.

[151] Liana F Lareau et al. *The evolving roles of alternative splicing.* June 2004. DOI: 10.1016/j.sbi.2004.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0959440X04000764?via%7B%5C%%7D3Dihub.

[152] Eva Latorre and Lorna W. Harries. *Splicing regulatory factors, ageing and age-related disease.* Vol. 36. Elsevier B.V., 2017, pp. 165–170. ISBN: 1392406773. DOI: 10.1016/j.arr.2017.04.004. URL: http://dx.doi.org/10.1016/j.arr.2017.04.004.

[153] Ji-Ann Ann Lee et al. "Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes". In: *Neuron* 89.1 (Jan. 2016), pp. 113–128. ISSN: 10974199. DOI: 10.1016/j.neuron.2015.11.025. URL: http://www.sciencedirect.com/science/article/pii/S0896627315010314?via%7B%5C%%7D3Dihub%20http://www.ncbi.nlm.nih.gov/pubmed/26687839%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4858412.

[154] Kuy Sook Lee et al. "HB-EGF induces cardiomyocyte hypertrophy via an ERK5-MEF2A-COX2 signaling pathway". In: *Cell. Signal.* 23.7 (July 2011), pp. 1100–1109. ISSN: 08986568. DOI: 10.1016/j.cellsig.2011.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0898656811000076?via%7B%5C%%7D3Dihub.

[155] Sarah Leigh-Brown et al. "Regulatory divergence of transcript isoforms in a mammalian model system". In: *PLoS One* 10.9 (Sept. 2015). Ed. by Barbara E. Stranger, e0137367. ISSN: 19326203. DOI: 10.1371/journal.pone.0137367. arXiv: . URL: http://dx.plos.org/10.1371/journal.pone.0137367%20https://dx.plos.org/10.1371/journal.pone.0137367.

[156] Michael K K Leung et al. "Deep learning of the tissue-regulated splicing code." In: *Bioinformatics* 30.12 (June 2014), pp. i121–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu277. URL: http://www.ncbi.nlm.nih.gov/pubmed/24931975.

[157] Bo Li and Colin N Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." In: *BMC Bioinformatics* 12.1 (2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. arXiv: NIHMS150003. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323.

[158] J.-H. Li et al. "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data". In: *Nucleic Acids Res.* 42.D1 (2014), pp. D92–D97. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1248. URL: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1248.

[159] Li Li, Christian J Jr Stoeckert, and David S Roos. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes – Li et al. 13 (9): 2178 – Genome Research". In: *Genome Res.* 13.9 (2003), pp. 2178–2189. ISSN: 1088-9051. DOI: 10.1101/gr.1224503.candidates. URL: http://genome.cshlp.org/cgi/content/full/13/9/2178.

[160] Qin Li et al. "The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation." In: *Elife* 3 (2014), e01201. ISSN: 2050-084X. DOI: 10.7554/eLife.01201. URL: http://elife.elifesciences.org/content/3/e01201.abstract.

[161] Yang I. Li et al. "RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts". In: *Genome Res.* 25.1 (Dec. 2015), pp. 1–13. ISSN: 15495469. DOI: 10.1101/gr.181990.114. URL: http://genome.cshlp.org/content/25/1/1.

[162] Yang I. Li et al. "RNA splicing is a primary link between genetic variation and disease". In: *Science* 352.6285 (Apr. 2016), pp. 600–4. ISSN: 0036-8075. DOI: 10.1126/science.aad9417. URL: http://www.ncbi.nlm.nih.gov/pubmed/27126046%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5182069%20/pmc/articles/PMC5182069/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5182069/.

[163] Yumei Li et al. "Human exonization through differential nucleosome occupancy". In: *Proc. Natl. Acad. Sci. U. S. A.* 115.35 (Aug. 2018), pp. 8817–8822. ISSN: 10916490. DOI: 10.1073/pnas.1802561115. URL: http://www.ncbi.nlm.nih.gov/pubmed/30104384%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6126743.

[164] Yun Li et al. "LRP4 Mutations Alter Wnt/β-Catenin Signaling and Cause Limb and Kidney Malformations in Cenani-Lenz Syndrome". In: *Am. J. Hum. Genet.* 86.5 (2010), pp. 696–706. ISSN: 00029297. DOI: 10.1016/j.ajhg.2010.03.004.

[165] Donny D. Licatalosi et al. "Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain". In: *Genes Dev.* 26.14 (2012), pp. 1626–1642. ISSN: 08909369. DOI: 10.1101/gad.191338.112.

[166] Anthony J Linares et al. "The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation." In: *Elife* 4 (2015), pp. 1–25. ISSN: 2050-084X. DOI: 10.7554/eLife.09268. URL: http://www.ncbi.nlm.nih.gov/pubmed/26705333.

[167] Matthew A Lines et al. "Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly". In: *Am. J. Hum. Genet.* 90.2 (Feb. 2012), pp. 369–377. ISSN: 00029297. DOI: 10.1016/j.ajhg.2011.12.023. URL: http://www.ncbi.nlm.nih.gov/pubmed/22305528%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3276671.

[168] Thomas Litman and Wilfred D. Stein. "Obtaining estimates for the ages of all the protein-coding genes and most of the ontology-identified noncoding genes of the human genome, assigned to 19 phylostrata". In: *Semin. Oncol.* 46.1 (2019), pp. 3–9. ISSN: 15328708. DOI: 10.1053/j.seminoncol.2018.11.002. URL: https://doi.org/10.1053/j.seminoncol.2018.11.002.

[169] Ziqing Liu et al. "Single cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte." In: *Nature* 551.7678 (2017), pp. 100–104. ISSN: 0028-0836. DOI: 10.1038/nature24454. URL: http://dx.doi.org/10.1038/nature24454.

[170] James P. B. Lloyd. "The evolution and diversity of the nonsense-mediated mRNA decay pathway". In: *F1000Research* 7 (2018), p. 1299. ISSN: 2046-1402. DOI: 10.12688/f1000research.15872.1. URL: https://f1000research.com/articles/7-1299/v1.

[171] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12 (Dec. 2014), p. 550. ISSN: 1474760X. DOI: 10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

[172] Reini F Luco et al. "Regulation of alternative splicing by histone modifications." In: *Science* 327.5968 (Feb. 2010), pp. 996–1000. ISSN: 1095-9203. DOI: 10.1126/science.1184208. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2913848%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[173] Bradley M. Lunde, Claire Moore, and Gabriele Varani. *RNA-binding proteins: Modular design for efficient function*. June 2007. DOI: 10.1038/nrm2178. URL: www.nature.com/reviews/molcellbio.

[174] Weijun Luo et al. "GAGE: generally applicable gene set enrichment for pathway analysis". In: *BMC Bioinformatics* 10.1 (May 2009), p. 161. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-161. URL: http://www.biomedcentral.com/1471-2105/10/161.

[175] Michael Lynch. "Evolution of the mutation rate". In: *Trends Genet.* 26.8 (Aug. 2010), pp. 345–352. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2010.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0168952510001034.

[176] Michael Lynch. "Intron evolution as a population-genetic process". In: *Proc. Natl. Acad. Sci.* 99.9 (Apr. 2002), pp. 6118–6123. DOI: 10.1073/pnas.092595699. URL: http://www.ncbi.nlm.nih.gov/pubmed/11983904%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC122912%20http://www.pnas.org/cgi/doi/10.1073/pnas.092595699%20http://sansan.phy.ncu.edu.tw/%7B~%7Dhclee/ref/Lynch%7B%5C_%7DPNAS99%7B%5C_%7D02.pdf%20http://www.ncbi.nlm.nih.gov/pubm.

[177] Michael Lynch and John S Conery. *The Evolutionary Fate and Consequences of Duplicate Genes*. Mar. 2000. DOI: 10.1126/science.287.5461.2204. URL: http://www.ncbi.nlm.nih.gov/pubmed/11073452%20http://www.ncbi.nlm.nih.gov/pubmed/10731134%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2754258%20http://www.sciencemag.org/cgi/doi/10.1126/science.287.5461.2204.

[178] Michael Lynch and John S Conery. "The origins of genome complexity." In: *Science* 302.5649 (2003), pp. 1401–4. ISSN: 1095-9203. DOI: 10.1126/science.1089370. URL: http://www.ncbi.nlm.nih.gov/pubmed/14631042.

[179] Michael Lynch and John S. Conery. "The evolutionary demography of duplicate genes". In: *J. Struct. Funct. Genomics* 3.1-4 (2003), pp. 35–44. ISSN: 1345711X. DOI: 10.1023/A:1022696612931.

[180] Michael Lynch et al. “The Repatterning of Eukaryotic Genomes by Random Genetic Drift”. In: *Annu. Rev. Genomics Hum. Genet.* 12.1 (Sept. 2011), pp. 347–366. ISSN: 1527-8204. DOI: 10.1146/annurev-genom-082410-101412. arXiv: 15334406. URL: http://www.annualreviews.org/doi/10.1146/annurev-genom-082410-101412.

[181] Fu Chao Ma et al. “Profiling of prognostic alternative splicing in melanoma”. In: *Oncol. Lett.* 18.2 (2019), pp. 1081–1088. ISSN: 17921082. DOI: 10.3892/ol.2019.10453.

[182] Steven Maere et al. *No Title*. Apr. 2005. DOI: 10.1073/pnas.0501102102. URL: www.pnas.orgcgidoi10.1073pnas.0501102102%20http://www.ncbi.nlm.nih.gov/pubmed/15800040%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC556253.

[183] Daniel Mapleson et al. “Efficient and accurate detection of splice junctions from RNA-seq with Portcullis”. In: *Gigascience* 7.12 (2018), pp. 1–14. ISSN: 2047217X. DOI: 10.1093/gigascience/giy131.

[184] Diane Marie et al. “The meanings of’function’ in biology and the problematic case of de novo gene emergence”. In: *Elife* 8 (2019), pp. 1–12. ISSN: 2050084X. DOI: 10.7554/eLife.47014.

[185] Carlos Martí-Gómez, Enrique Lara-Pezzi, and Fátima Sánchez-Cabo. “dSreg: a Bayesian model to integrate changes in splicing and RNA-binding protein activity”. In: *Bioinformatics* (Dec. 2019). Ed. by Jan Gorodkin. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz915. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz915/5675498.

[186] Akio Masuda et al. “CUGBP1 and MBNL1 preferentially bind to 3 UTRs and facilitate mRNA decay”. In: *Sci. Rep.* 2 (2012), pp. 1–10. ISSN: 2045-2322. DOI: 10.1038/srep00209.

[187] Daniel Maticzka et al. “GraphProt: modeling binding preferences of RNA-binding proteins.” In: *Genome Biol.* 15.1 (2014), R17. ISSN: 1465-6914. DOI: 10.1186/gb-2014-15-1-r17. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053806%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[188] David M. McCandlish. “Long-term evolution on complex fitness landscapes when mutation is weak”. In: *Heredity (Edinb).* 121.5 (Nov. 2018), pp. 449–465. ISSN: 13652540. DOI: 10.1038/s41437-018-0142-6. URL: http://dx.doi.org/10.1038/s41437-018-0142-6%20http://www.nature.com/articles/s41437-018-0142-6.

[189] Michael G. McDermott et al. “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update”. In: *Nucleic Acids Res.* 44.W1 (July 2016), W90–W97. ISSN: 0305-1048. DOI: 10.1093/nar/gkw377. URL: http://www.ncbi.nlm.nih.gov/pubmed/27141961%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4987924%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw377.

[190] Casey L. McGrath et al. “No Title”. In: 24.10 (Oct. 2014). DOI: 10.1101/gr.173740.114. URL: http://www.ncbi.nlm.nih.gov/pubmed/25085612%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4199370%20/pmc/articles/PMC4199370/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4199370/.

[191] C Joel McManus et al. “Evolution of splicing regulatory networks in Drosophila”. In: *Genome Res.* 24.5 (May 2014), pp. 786–796. ISSN: 15495469. DOI: 10.1101/gr.161521.113. URL: http://www.ncbi.nlm.nih.gov/pubmed/24515119%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4009608.

[192] Eugene Melamud and John Moult. "Stochastic noise in splicing machinery". In: *Nucleic Acids Res.* 37.14 (Aug. 2009), pp. 4873–4886. ISSN: 03051048. DOI: 10.1093/nar/gkp471. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp471.

[193] Jason J. Merkin et al. "Origins and Impacts of New Mammalian Exons". In: *Cell Rep.* 10.12 (2015), pp. 1992–2005. ISSN: 22111247. DOI: 10.1016/j.celrep.2015.02.058. URL: http://linkinghub.elsevier.com/retrieve/pii/S2211124715002351.

[194] Jason Merkin et al. "Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues". In: *Science (80-.).* 338.December (2012), pp. 1593–1600. DOI: 10.1126/science.1228186. URL: http://www.sciencemag.org/content/338/6114/1593.full.

[195] Robert Middleton et al. "IRFinder: Assessing the impact of intron retention on mammalian gene expression". In: *Genome Biol.* 18.1 (Dec. 2017), p. 51. ISSN: 1474760X. DOI: 10.1186/s13059-017-1184-4. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1184-4.

[196] Jean Monlong et al. "Identification of genetic variants associated with alternative splicing using sQTLseekeR". In: *Nat. Commun.* 5.May (2014). ISSN: 20411723. DOI: 10.1038/ncomms5698.

[197] Jeffrey T. Morgan, Gerald R. Fink, and David P. Bartel. "Excised linear introns regulate growth in yeast". In: *Nature* 565.7741 (2019), pp. 606–611. ISSN: 14764687. DOI: 10.1038/s41586-018-0828-1. URL: http://dx.doi.org/10.1038/s41586-018-0828-1.

[198] Michaela Müller-McNicoll et al. "SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export". In: *Genes Dev.* 30.5 (2016), pp. 553–566. ISSN: 15495477. DOI: 10.1101/gad.276477.115.

[199] Valentine Murigneux et al. "Transcriptome-wide identification of RNA binding sites by CLIP-seq". In: *Methods* 63.1 (2013), pp. 32–40. ISSN: 10462023. DOI: 10.1016/j.ymeth.2013.03.022. URL: http://dx.doi.org/10.1016/j.ymeth.2013.03.022.

[200] Shiran Naftelberg et al. "Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure". In: *Annu. Rev. Biochem.* 84.1 (2015), pp. 165–198. ISSN: 0066-4154. DOI: 10.1146/annurev-biochem-060614-034242.

[201] Abhinav Nellore et al. "Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive". In: *Genome Biol.* 17.1 (Dec. 2016), p. 266. ISSN: 1474760X. DOI: 10.1186/s13059-016-1118-6. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1118-6.

[202] Timothy W. Nilsen and Brenton R. Graveley. *Expansion of the eukaryotic proteome by alternative splicing.* Jan. 2010. DOI: 10.1038/nature08909. URL: https://www.nature.com/articles/nature08909.pdf%20http://www.nature.com/articles/nature08909.

[203] Eric L Van Nostrand et al. "A Large-Scale Binding and Functional Map of Human RNA Binding Proteins Correspondence and requests for materials should be addressed to Brenton Graveley (graveley@uchc.edu), Chris Burge (cburge@mit.edu), Xiang-dong Fu (xdfu@ucsd.edu)". In: *bioRxiv* (2017), pp. 1–74. DOI: 10.1101/179648.

[204] Julia K. Nussbacher and Gene W. Yeo. "Systematic Discovery of RNA Binding Proteins that Regulate MicroRNA Levels". In: *Mol. Cell* 69.6 (Mar. 2018), 1005–1016.e7. ISSN: 10974164. DOI: 10.1016/j.molcel.2018.02.012. URL: https://www.sciencedirect.com/science/article/pii/S1097276518301102?via%7B%5C%%7D3Dihub.

[205] Susumu Ohno, Ulrich Wolf, and Niels B. Atkin. "Evolution from fish to mammals by gene duplication". In: *Hereditas* 1 (1967), pp. 169–187.

[206] Sandra Orchard et al. “The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases”. In: *Nucleic Acids Res.* 42.D1 (2014), pp. 358–363. ISSN: 03051048. DOI: 10.1093/nar/gkt1115.

[207] Paula Ortiz-Sánchez et al. “Loss of SRSF3 in cardiomyocytes leads to decapping of contraction-related mRNAs and severe systolic dysfunction”. In: *Circ. Res.* 125.2 (2019), pp. 170–183. ISSN: 15244571. DOI: 10.1161/CIRCRESAHA.118.314515.

[208] Qun Pan et al. “Alternative splicing of conserved exons is frequently species-specific in human and mouse”. In: *Trends Genet.* 21.2 (2005), pp. 73–77. ISSN: 01689525. DOI: 10.1016/j.tig.2004.12.002. URL: www.sciencedirect.com.

[209] Qun Pan et al. “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing”. In: *Nat. Genet.* 40.12 (2008), pp. 1413–1415. ISSN: 10614036. DOI: 10.1038/ng.259.

[210] Xiaoyong Pan et al. “Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks”. In: *BMC Genomics* 19.1 (2018), pp. 1–11. ISSN: 14712164. DOI: 10.1186/s12864-018-4889-1.

[211] Panagiotis Papasaikas et al. “Functional Splicing Network Reveals Extensive Regulatory Potential of the Core Spliceosomal Machinery”. In: *Mol. Cell* 57 (2015), pp. 1–16. DOI: 10.1016/j.molcel.2014.10.030.

[212] Julie Parenteau et al. “Introns are mediators of cell response to starvation”. In: *Nature* 565.7741 (2019), pp. 612–617. ISSN: 14764687. DOI: 10.1038/s41586-018-0859-7.

[213] Soochul Park. “Defective Anks1a disrupts the export of receptor tyrosine kinases from the endoplasmic reticulum”. In: *BMB Rep.* 49.12 (Dec. 2016), pp. 651–652. ISSN: 1976-6696. DOI: 10.5483/BMBRep.2016.49.12.186. URL: http://koreascience.or.kr/journal/view.jsp?kj=E1MBB7%7B%5C&%7Dpy=2016%7B%5C&%7Dvnc=v49n12%7B%5C&%7Dsp=651.

[214] Tae Sik Park et al. “Ceramide is a cardiotoxin in lipotoxic cardiomyopathy”. In: *J. Lipid Res.* 49.10 (2008), pp. 2101–2112. ISSN: 00222275. DOI: 10.1194/jlr.M800147-JLR200.

[215] Rob Patro et al. “Salmon provides fast and bias-aware quantification of transcript expression”. In: *Nat. Methods* 14.4 (Apr. 2017), pp. 417–419. ISSN: 15487105. DOI: 10.1038/nmeth.4197. arXiv: 1505.02710. URL: http://www.nature.com/articles/nmeth.4197.

[216] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python Gaël Varoquaux”. In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830. URL: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[217] Mihaela Pertea et al. “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown”. In: *Nat. Protoc.* 11.9 (Sept. 2016), pp. 1650–1667. ISSN: 1754-2189. DOI: 10.1038/nprot.2016.095. URL: http://www.nature.com/articles/nprot.2016.095%20http://www.nature.com/doifinder/10.1038/nprot.2016.095.

[218] Joseph K. Pickrell et al. “Noisy splicing drives mRNA isoform diversity in human cells”. In: *PLoS Genet.* 6.12 (Dec. 2010). Ed. by Emmanouil T. Dermitzakis, pp. 1–11. DOI: 10.1371/journal.pgen.1001236. URL: http://dx.plos.org/10.1371/journal.pgen.1001236.

[219] Juho Piironen and Aki Vehtari. “Sparsity information and regularization in the horseshoe and other shrinkage priors”. In: *Electron. J. Stat.* 11.2 (2017), pp. 5018–5051. ISSN: 19357524. DOI: 10.1214/17-EJS1337SI. arXiv: 1707.01694. URL: https://arxiv.org/pdf/1707.01694.pdf.

[220] Anthony M. Poole. *Did group II intron proliferation in an endosymbiont-bearing archaeon create eukaryotes?* 2006. DOI: 10.1186/1745-6150-1-36.

[221] Mathieu Quesnel-vallières et al. “Essential roles for the splicing regulator nSR100 / SRRM4 during nervous system development”. In: *Genes Dev.* (2015), pp. 746–759. DOI: 10.1101/gad.256115.114.4.

[222] Aaron R. Quinlan and Ira M. Hall. “BEDTools: A flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq033. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033.

[223] Bushra Raj et al. “A global regulatory mechanism for activating an exon network required for neurogenesis”. In: *Mol. Cell* 56.1 (2014), pp. 90–103. ISSN: 10974164. DOI: 10.1016/j.molcel.2014.08.011. URL: http://dx.doi.org/10.1016/j.molcel.2014.08.011.

[224] Vasily E. Ramensky et al. “Positive Selection in Alternatively Spliced Exons of Human Genes”. In: *Am. J. Hum. Genet.* 83.1 (July 2008), pp. 94–98. ISSN: 00029297. DOI: 10.1016/j.ajhg.2008.05.017. URL: https://www.sciencedirect.com/science/article/pii/S0002929708003510?via%7B%5C%%7D3Dihub.

[225] Debashish Ray et al. “A compendium of RNA-binding motifs for decoding gene regulation.” In: *Nature* 499.7457 (July 2013), pp. 172–7. ISSN: 1476-4687. DOI: 10.1038/nature12311. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3929597%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%20http://dx.doi.org/10.1038/nature12311.

[226] Anireddy S N Reddy et al. “Complexity of the Alternative Splicing Landscape in Plants”. In: *Plant Cell* 25.October (2013), pp. 3657–3683. DOI: 10.1105/tpc.113.117523. URL: https://www.ncbi.nlm.nih.gov/pubmed/24179125.

[227] Alissa Resch et al. “Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome”. In: *J. Proteome Res.* 3.1 (2004), pp. 76–83. ISSN: 15353893. DOI: 10.1021/pr034064v.

[228] Alejandro Reyes et al. “Drift and conservation of differential exon usage across tissues in primate species.” In: *Proc. Natl. Acad. Sci. U. S. A.* 110.38 (Sept. 2013), pp. 15377–82. ISSN: 1091-6490. DOI: 10.1073/pnas.1307202110. URL: http://www.ncbi.nlm.nih.gov/pubmed/24003148%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3780897.

[229] Matthew E. Ritchie et al. “Limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Res.* 43.7 (Apr. 2015), e47. ISSN: 13624962. DOI: 10.1093/nar/gkv007. URL: http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for.

[230] Jose Manuel Rodriguez et al. “An analysis of tissue-specific alternative splicing at the protein level”. In: *PLOS Comput. Biol.* 16.10 (2020), e1008287. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008287. URL: https://dx.plos.org/10.1371/journal.pcbi.1008287.

[231] Igor B. Rogozin et al. “Origin and evolution of spliceosomal introns”. In: *Biol. Direct* 7 (Apr. 2012), pp. 1–28. ISSN: 17456150. DOI: 10.1186/1745-6150-7-11. URL: http://www.ncbi.nlm.nih.gov/pubmed/22507701%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3488318.

[232] Rori V. Rohlfs, Patrick Harrigan, and Rasmus Nielsen. “Modeling gene expression evolution with an extended ornstein-uhlenbeck process accounting for within-species variation”. In: *Mol. Biol. Evol.* 31.1 (2014), pp. 201–211. ISSN: 07374038. DOI: 10.1093/molbev/mst190.

[233] Rori V. Rohlfs and Rasmus Nielsen. "Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution". In: *Syst. Biol.* 64.5 (2015), pp. 695–708. ISSN: 1076836X. DOI: 10.1093/sysbio/syv042.

[234] Maria Grazia Romanelli, Erica Diani, and Patricia Marie Jeanne Lievens. "New insights into functional roles of the polypyrimidine tract-binding protein". In: *Int. J. Mol. Sci.* 14.11 (2013), pp. 22906–22932. ISSN: 16616596. DOI: 10.3390/ijms141122906.

[235] Alexander B. B. Rosenberg et al. "Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences". In: *Cell* 163.3 (Oct. 2015), pp. 698–711. ISSN: 10974172. DOI: 10.1016/j.cell.2015.09.054. URL: https://www.sciencedirect.com/science/article/pii/S0092867415012714.

[236] Oliver Rossbach et al. "Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L." In: *RNA Biol.* 11.2 (2014), pp. 146–55. ISSN: 1555-8584. DOI: 10.4161/rna.27991. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3973733%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24526010.

[237] Scott W Roy. "Intronization, de-intronization and intron sliding are rare in Cryptococcus". In: *BMC Evol. Biol.* 9.1 (Aug. 2009), p. 192. ISSN: 14712148. DOI: 10.1186/1471-2148-9-192. URL: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-192.

[238] Scott William Roy and Manuel Irimia. "Mystery of intron gain: new data and new models". In: *Trends Genet.* 25.2 (Feb. 2009), pp. 67–73. ISSN: 01689525. DOI: 10.1016/j.tig.2008.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0168952508003132.

[239] Scott Roy and Manuel Irimia. "Intron mis-splicing: no alternative?" In: *Genome Biol.* 9.2 (2008), p. 208. ISSN: 1465-6906. DOI: 10.1186/gb-2008-9-2-208. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-2-208.

[240] Terrie Sadusky, Andrew J Newman, and Nicholas J Dibb. "Exon junction sequences as cryptic splice sites: Implications for intron origin". In: *Curr. Biol.* 14.6 (Mar. 2004), pp. 505–509. ISSN: 09609822. DOI: 10.1016/j.cub.2004.02.063. URL: https://www.sciencedirect.com/science/article/pii/S0960982204001538?via%7B%5C%%7D3Dihub.

[241] Helen K. Salz. *Sex determination in insects: A binary decision based on alternative splicing*. 2011. DOI: 10.1016/j.gde.2011.03.001.

[242] Claudia Scheckel and Robert B Darnell. "Microexons — Tiny but mighty". In: 34.3 (2015), pp. 2014–2016.

[243] Schraga Schwartz, Eran Meshorer, and Gil Ast. "Chromatin organization marks exon-intron structure". In: *Nat. Struct. Mol. Biol.* 16.9 (2009), pp. 990–995. ISSN: 15459993. DOI: 10.1038/nsmb.1659.

[244] Celine Scornavacca et al. "OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes". In: *Mol. Biol. Evol.* 36.4 (Apr. 2019). Ed. by Koichiro Tamura, pp. 861–862. ISSN: 0737-4038. DOI: 10.1093/molbev/msz015. URL: https://academic.oup.com/mbe/article/36/4/861/5303840.

[245] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with Python". In: *Proc. 9th Python Sci. Conf.* 57.Scipy (2010), p. 61. URL: http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf.

[246] Endre Sebestyén et al. "Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks". In: *Genome Res.* 26.6 (June 2016), pp. 732–744. ISSN: 15495469. DOI: 10.1101/gr.199935.115. URL: http://www.ncbi.nlm.nih.gov/pubmed/27197215%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4889968.

[247] Jan Seyfried et al. "A novel mitogen-activated protein kinase docking site in the N terminus of MEK5alpha organizes the components of the extracellular signal-regulated kinase 5 signaling pathway." In: *Mol. Cell. Biol.* 25.22 (Nov. 2005), pp. 9820–8. ISSN: 0270-7306. DOI: 10.1128/MCB.25.22.9820-9828.2005. URL: http://www.ncbi.nlm.nih.gov/pubmed/16260599%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1280269%20http://mcb.asm.org/cgi/doi/10.1128/MCB.25.22.9820-9828.2005.

[248] Shihao Shen et al. "MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data." In: *Nucleic Acids Res.* 40.8 (Apr. 2012), e61. ISSN: 1362-4962. DOI: 10.1093/nar/gkr1291. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3333886%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[249] Shihao Shen et al. "rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data". In: *Proc. Natl. Acad. Sci.* 111 (2014), E5593–E5601. ISSN: 0027-8424. DOI: 10.1073/pnas.1419161111. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1419161111.

[250] Shihao Shen et al. "SURVIV for survival analysis of mRNA isoform variation". In: *Nat. Commun.* 7 (2016), pp. 1–11. ISSN: 20411723. DOI: 10.1038/ncomms11548. URL: http://dx.doi.org/10.1038/ncomms11548.

[251] Hossein Shenasa and Klemens J. Hertel. "Combinatorial regulation of alternative splicing". In: *Biochim. Biophys. Acta - Gene Regul. Mech.* 1862.11-12 (July 2019), p. 194392. ISSN: 18764320. DOI: 10.1016/j.bbagrm.2019.06.003. URL: https://doi.org/10.1016/j.bbagrm.2019.06.003%20https://www.sciencedirect.com/science/article/pii/S1874939919301026?via%7B%5C%%7D3Dihub.

[252] Cedric Simillion et al. "Avoiding the pitfalls of gene set enrichment analysis with SetRank". In: *BMC Bioinformatics* 18.1 (Dec. 2017), p. 151. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1571-6. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1571-6.

[253] Dominique Simon-Chazottes et al. "Mutations in the gene encoding the low-density lipoprotein receptor LRP4 cause abnormal limb development in the mouse". In: *Genomics* 87.5 (2006), pp. 673–677. ISSN: 08887543. DOI: 10.1016/j.ygeno.2006.01.007.

[254] Pooja Singh et al. "The Role of Alternative Splicing and Differential Gene Expression in Cichlid Adaptive Radiation". In: *Genome Biol. Evol.* 9.10 (Oct. 2017), pp. 2764–2781. ISSN: 1759-6653. DOI: 10.1093/gbe/evx204. URL: http://academic.oup.com/gbe/article/9/10/2764/4259059.

[255] Ravi K. Singh et al. "Rbfox2-coordinated alternative splicing of Mef2d and Rock2 controls myoblast fusion during myogenesis". In: *Mol. Cell* 55.4 (Aug. 2014), pp. 592–603. ISSN: 10974164. DOI: 10.1016/j.molcel.2014.06.035. arXiv: NIHMS150003. URL: https://www.sciencedirect.com/science/article/pii/S1097276514005693.

[256] Ben Smithers, Matt Oates, and Julian Gough. "'Why genes in pieces?'-revisited". In: *Nucleic Acids Res.* 47.10 (2019), pp. 4970–4973. ISSN: 13624962. DOI: 10.1093/nar/gkz284.

[257] Kai Yi Song et al. “MBNL1 reverses the proliferation defect of skeletal muscle satellite cells in myotonic dystrophy type 1 by inhibiting autophagy via the mTOR pathway”. In: *Cell Death Dis.* 11.7 (2020). ISSN: 20414889. DOI: 10.1038/s41419-020-02756-8. URL: http://dx.doi.org/10.1038/s41419-020-02756-8.

[258] Erik L.L. Sonnhammer and Gabriel Östlund. “InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic”. In: *Nucleic Acids Res.* 43.D1 (Jan. 2015), pp. D234–D239. ISSN: 1362-4962. DOI: 10.1093/nar/gku1203. URL: http://academic.oup.com/nar/article/43/D1/D234/2439077/InParanoid-8-orthology-analysis-between-273.

[259] Rotem Sorek. “The birth of new exons: Mechanisms and evolutionary consequences”. In: *Rna* 13.10 (2007), pp. 1603–1608. ISSN: 13558382. DOI: 10.1261/rna.682507.

[260] Rachel Spellman, Miriam Llorian, and C. W J Smith. “Crossregulation and Functional Redundancy between the Splicing Regulator PTB and Its Paralogs nPTB and ROD1”. In: *Mol. Cell* 27.3 (2007), pp. 420–434. ISSN: 10972765. DOI: 10.1016/j.molcel.2007.06.016. URL: http://dx.doi.org/10.1016/j.molcel.2007.06.016.

[261] Timothy Sterne-Weiler et al. “Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop”. In: *Mol. Cell* 72.1 (Oct. 2018), 187–200.e6. ISSN: 1097-2765. DOI: 10.1016/J.MOLCEL.2018.08.018. URL: https://www.cell.com/molecular-cell/fulltext/S1097-2765(18)30678-6%20http://www.ncbi.nlm.nih.gov/pubmed/30220560.

[262] Peter Stoilov et al. “Human tra2-beta1 autoregulates its protein concentration by infuencing alternative splicing of its pre-mRNA”. In: *Hum. Mol. Genet.* 13.5 (2004), pp. 509–524. ISSN: 09646906. DOI: 10.1093/hmg/ddh051.

[263] Arlin Stoltzfus et al. “Testing the exon theory of genes: The evidence from protein structure”. In: *Science (80-.).* 265.5169 (1994), pp. 202–207. ISSN: 00368075. DOI: 10.1126/science.8023140.

[264] Aravind Subramanian et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proc. Natl. Acad. Sci. USA* 102.43 (2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. URL: http://www.pnas.org/content/102/43/15545.abstract.

[265] Yoichiro Sugimoto et al. “Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.” In: *Genome Biol.* 13.8 (2012), R67. ISSN: 1465-6914. DOI: 10.1186/gb-2012-13-8-r67. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053741%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

[266] David Talavera et al. “The (In)dependence of alternative splicing and gene duplication”. In: *PLoS Comput. Biol.* 3.3 (2007), pp. 0375–0388. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.0030033.

[267] J Matthew Taliaferro et al. “RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation”. In: *Mol. Cell* (2016), pp. 1–13. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2016.08.035. URL: http://dx.doi.org/10.1016/j.molcel.2016.08.035.

[268] Javier Tapial et al. “An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms”. In: *Genome Res.* 27.10 (2017), pp. 1759–1768. ISSN: 15495469. DOI: 10.1101/gr.220962.117.

[269] Jamal Tazi, Nadia Bakkour, and Stefan Stamm. “Alternative splicing and disease”. In: *Biochim. Biophys. Acta - Mol. Basis Dis.* 1792.1 (2009), pp. 14–26. ISSN: 09254439. DOI: 10.1016/j.bbadis.2008.09.017. URL: http://dx.doit.org/10.1016/j.bbadis.2008.09.017.

[270] B. G. Tenchov et al. “A probability concept about size distributions of sonicated lipid vesicles”. In: *BBA - Biomembr.* 816.1 (1985), pp. 122–130. ISSN: 00052736. DOI: 10.1016/0005-2736(85)90400-6.

[271] Hagen Tilgner et al. “Nucleosome positioning as a determinant of exon recognition”. In: *Nat. Struct. Mol. Biol.* 16.9 (Sept. 2009), pp. 996–1001. ISSN: 15459993. DOI: 10.1038/nsmb.1658. URL: http://www.nature.com/articles/nsmb.1658.

[272] Jiefei Tong et al. “Odin (ANKS1A) Modulates EGF Receptor Recycling and Stability”. In: *PLoS One* 8.6 (June 2013). Ed. by Laszlo Buday, e64817. ISSN: 19326203. DOI: 10.1371/journal.pone.0064817. URL: http://dx.plos.org/10.1371/journal.pone.0064817.

[273] Cole Trapnell, Lior Pachter, and Steven L Salzberg. “TopHat: discovering splice junctions with RNA-Seq.” In: *Bioinformatics* 25.9 (May 2009), pp. 1105–11. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp120. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp120%20http://www.ncbi.nlm.nih.gov/pubmed/19289445%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2672628.

[274] Cole Trapnell and Steven L. Salzberg. “How to map billions of short reads onto genomes”. In: *Nat. Biotechnol.* 27.5 (2009), pp. 455–457. ISSN: 10870156. DOI: 10.1038/nbt0509-455.

[275] Cole Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nat. Protoc.* 7.3 (2012), pp. 562–578. ISSN: 17542189. DOI: 10.1038/nprot.2012.016. URL: http://dx.doi.org/10.1038/nprot.2012.016.

[276] Michael L. Tress, Federico Abascal, and Alfonso Valencia. “Alternative Splicing May Not Be the Key to Proteome Complexity”. In: *Trends Biochem. Sci.* 0.2 (2017), pp. 1760–1774. ISSN: 13624326. DOI: 10.1016/j.tibs.2016.08.008. URL: http://dx.doi.org/10.1016/j.tibs.2016.08.008.

[277] Michael L. Tress, Federico Abascal, and Alfonso Valencia. “Most Alternative Isoforms Are Not Functionally Important”. In: *Trends Biochem. Sci.* xx.6 (2017), pp. 1–2. ISSN: 09680004. DOI: 10.1016/j.tibs.2017.04.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S0968000417300713.

[278] Juan L. Trincado et al. “SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions”. In: *Genome Biol.* 19.1 (Dec. 2018), p. 40. ISSN: 1474760X. DOI: 10.1186/s13059-018-1417-1. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1417-1.

[279] Yihsuan S. Tsai et al. “Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors”. In: *Oncotarget* 6.9 (2015), pp. 6825–6839. ISSN: 19492553. DOI: 10.18632/oncotarget.3145.

[280] Braunschweig U et al. “Widespread intron retention in mammals functionally tunes transcriptomes”. In: *Genome Res.* 24 (2014), pp. 1774–1786. ISSN: 1549-5469. DOI: 10.1101/gr.177790.114.1774.

[281] Jernej Ule et al. “iCLIP predicts the dual splicing effects of TIA-RNA interactions”. In: *PLoS Biol.* 8.10 (2010). ISSN: 15449173. DOI: 10.1371/journal.pbio.1000530.

[282] Nathan S. Upham, Jacob A. Esselstyn, and Walter Jetz. “Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation”. In: *PLoS Biol.* 17.12 (Dec. 2019). Ed. by Andrew J. Tanentzap, e3000494. ISSN: 15457885. DOI: 10.1371/journal.pbio.3000494. URL: https://dx.plos.org/10.1371/journal.pbio.3000494.

[283] Josef C. Uyeda and Luke J. Harmon. "A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data". In: *Syst. Biol.* 63.6 (Nov. 2014), pp. 902–918. ISSN: 1076836X. DOI: 10.1093/sysbio/syu057. URL: https://academic.oup.com/sysbio/article/63/6/902/2847928.

[284] Juan Valcárcel and Fátima Gebauer. "Post-transcriptional regulation: The dawn of PTB". In: *Curr. Biol.* 7.11 (Nov. 1997), R705–R708. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(06)00361-7. URL: https://www.sciencedirect.com/science/article/pii/S0960982206003617.

[285] Maarten M.G. Van Den Hoogenhof, Yigal M. Pinto, and Esther E. Creemers. "RNA Splicing regulation and dysregulation in the heart". In: *Circ. Res.* 118.3 (2016), pp. 454–468. ISSN: 15244571. DOI: 10.1161/CIRCRESAHA.115.307872.

[286] Maarten M.G. Van Den Hoogenhof et al. "RBM20 mutations induce an arrhythmogenic dilated cardiomyopathy related to disturbed calcium handling". In: *Circulation* 138.13 (2018), pp. 1330–1342. ISSN: 15244539. DOI: 10.1161/CIRCULATIONAHA.117.031947.

[287] Jorge Vaquero-Garcia et al. "A new view of transcriptome complexity and regulation through the lens of local splicing variations". In: *Elife* 5.FEBRUARY2016 (2016), pp. 1–30. ISSN: 2050084X. DOI: 10.7554/eLife.11752. arXiv: arXiv:1011.1669v3.

[288] Julian P Venables et al. "Identification of alternative splicing markers for breast cancer." In: *Cancer Res.* 68.22 (Nov. 2008), pp. 9525–31. ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-08-1769. URL: http://cancerres.aacrjournals.org/content/68/22/9525.abstract.

[289] Jacy L. Wagnon et al. "CELF4 Regulates Translation and Local Abundance of a Vast Set of mRNAs, Including Genes Associated with Regulation of Synaptic Function". In: *PLoS Genet.* 8.11 (2012). ISSN: 15537390. DOI: 10.1371/journal.pgen.1003067.

[290] Jing Wang et al. "WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit". In: *Nucleic Acids Res.* 45.W1 (July 2017), W130–W137. ISSN: 13624962. DOI: 10.1093/nar/gkx356. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx356.

[291] Kai Wang et al. "MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery". In: *Nucleic Acids Res.* 38.18 (2010), pp. 1–14. ISSN: 03051048. DOI: 10.1093/nar/gkq622.

[292] Liguo Wang, Shengqin Wang, and Wei Li. "RSeQC: quality control of RNA-seq experiments". In: *Bioinformatics* 28.16 (Aug. 2012), pp. 2184–2185. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bts356. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts356.

[293] Qian Wang et al. "Prognostic Potential of Alternative Splicing Markers in Endometrial Cancer". In: *Mol. Ther. - Nucleic Acids* 18.December (2019), pp. 1039–1048. ISSN: 21622531. DOI: 10.1016/j.omtn.2019.10.027. URL: https://doi.org/10.1016/j.omtn.2019.10.027.

[294] Zefeng Wang and Christopher B Burge. "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code." In: *RNA* 14.5 (2008), pp. 802–813. ISSN: 1355-8382. DOI: 10.1261/rna.876308.

[295] Chad M. Warren et al. "Titin isoform changes in rat myocardium during development". In: *Mech. Dev.* 121.11 (2004), pp. 1301–1312. ISSN: 09254773. DOI: 10.1016/j.mod.2004.07.003.

[296] Robert J. Weatheritt, Norman E. Davey, and Toby J. Gibson. "Linear motifs confer functional diversity onto splice variants". In: *Nucleic Acids Res.* 40.15 (2012), pp. 7123–7131. ISSN: 03051048. DOI: 10.1093/nar/gks442.

[297] Robert J. Weatheritt, Timothy Sterne-Weiler, and Benjamin J. Blencowe. "The ribosome-engaged landscape of alternative splicing". In: *Nat. Struct. Mol. Biol.* 23.12 (Dec. 2016), pp. 1117–1123. ISSN: 15459985. DOI: 10.1038/nsmb.3317.

[298] Sebastien M. Weyn-Vanhentenryck et al. "Precise temporal regulation of alternative splicing during neural development". In: *Nat. Commun.* 9.1 (Dec. 2018), p. 2189. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04559-0. URL: http://www.nature.com/articles/s41467-018-04559-0.

[299] Mandy S. Wong, Justin B. Kinney, and Adrian R. Krainer. "Quantitative Activity Profile and Context Dependence of All Human 5 Splice Sites". In: *Mol. Cell* 71.6 (2018), 1012–1026.e3. ISSN: 10974164. DOI: 10.1016/j.molcel.2018.07.033. URL: https://doi.org/10.1016/j.molcel.2018.07.033.

[300] Hui Y. Xiong et al. "The human splicing code reveals new insights into the genetic determinants of disease". In: *Science (80-.).* 347.6218 (2015). ISSN: 10959203. DOI: 10.1126/science.1254806.

[301] Yu-Cheng T Yang et al. "CLIPdb: a CLIP-seq database for protein-RNA interactions". In: *BMC Genomics* 16.1 (2015), p. 51. ISSN: 1471-2164. DOI: 10.1186/s12864-015-1273-2. URL: http://www.biomedcentral.com/1471-2164/16/51.

[302] Chengkun Yang et al. "Genome-wide profiling reveals the landscape of prognostic alternative splicing signatures in pancreatic ductal adenocarcinoma". In: *Front. Oncol.* 9.JUN (June 2019), p. 511. ISSN: 2234943X. DOI: 10.3389/fonc.2019.00511. URL: https://www.frontiersin.org/article/10.3389/fonc.2019.00511/full.

[303] Li Yang and Ling-Ling Chen. "Microexons Go Big". In: *Cell* 159.7 (Dec. 2014), pp. 1488–1489. ISSN: 00928674. DOI: 10.1016/j.cell.2014.12.004. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867414015748.

[304] Weiwei Yang et al. "EGFR-Induced and PKCϵ Monoubiquitylation-Dependent NF-κB Activation Upregulates PKM2 Expression and Promotes Tumorigenesis". In: *Mol. Cell* 48.5 (Dec. 2012), pp. 771–784. ISSN: 1097-2765. DOI: 10.1016/J.MOLCEL.2012.09.028. URL: https://www.sciencedirect.com/science/article/pii/S1097276512008283.

[305] Xinping Yang et al. "Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing". In: *Cell* 164.4 (2016), pp. 805–817. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.01.029. URL: http://www.cell.com/article/S0092867416300435/fulltext.

[306] Paul Yenerall, Bradlee Krupa, and Leming Zhou. "Mechanisms of intron gain and loss in Drosophila". In: (2011). DOI: 10.1186/1471-2148-11-364. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3296678/pdf/1471-2148-11-364.pdf.

[307] Paul Yenerall and Leming Zhou. "Identifying the mechanisms of intron gain: progress and trends." In: *Biol. Direct* 7 (Sept. 2012), p. 29. ISSN: 1745-6150. DOI: 10.1186/1745-6150-7-29. URL: http://www.ncbi.nlm.nih.gov/pubmed/22963364%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3443670.

[308] G. Yu et al. "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products". In: *Bioinformatics* 26.7 (Apr. 2010), pp. 976–978. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq064. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq064.

[309] Jin Yu et al. "JNK3 signaling pathway activates ceramide synthase leading to mitochondrial dysfunction". In: *J. Biol. Chem.* 282.35 (2007), pp. 25940–25949. ISSN: 00219258. DOI: 10.1074/jbc.M701812200.

[310] Jianzhi Zhang. *Evolution by gene duplication: An update*. 2003. DOI: 10.1016/S0169-5347(03)00033-8.

[311] Sai Zhang et al. "A deep learning framework for modeling structural features of RNA-binding protein targets." In: *Nucleic Acid Reseach* 44.10 (2015), pp. 1–14. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1025. URL: http://www.ncbi.nlm.nih.gov/pubmed/26467480.

[312] Xiang H-F Zhang, Christina S Leslie, and Lawrence a Chasin. "Computational searches for splicing signals." In: *Methods* 37.4 (Dec. 2005), pp. 292–305. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2005.07.011. URL: http://www.ncbi.nlm.nih.gov/pubmed/16314258.

[313] Xiang H.F. Zhang and Lawrence A. Chasin. "Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons". In: *Proc. Natl. Acad. Sci. U. S. A.* 103.36 (2006), pp. 13427–13432. ISSN: 00278424. DOI: 10.1073/pnas.0603042103.

[314] Yun Zhang et al. "FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene set enrichment analysis". In: *Bioinformatics* 33.13 (July 2017). Ed. by Ziv Bar-Joseph, pp. 1944–1952. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx104. URL: https://academic.oup.com/bioinformatics/article/33/13/1944/3038397.

[315] Zijun Zhang et al. "Deep-learning augmented RNA-seq analysis of transcript splicing". In: *Nat. Methods* 16.4 (Apr. 2019), pp. 307–310. ISSN: 15487105. DOI: 10.1038/s41592-019-0351-9. URL: http://www.nature.com/articles/s41592-019-0351-9.

[316] Liucun Zhu et al. "Patterns of exon-intron architecture variation of genes in eukaryotic genomes". In: *BMC Genomics* 10.1 (Jan. 2009), p. 47. ISSN: 14712164. DOI: 10.1186/1471-2164-10-47. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-47.