TUM School of Life Sciences
Technische Universität München

Towards a System Biology Approach For Alternative Splicing

Zakaria Louadi

Technische Universität München
TUM School of Life Sciences

Towards a System Biology Approach For Alternative Splicing

Zakaria Louadi

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Mathias Wilhelm

Prüfer der Dissertation:

Prof. Dr. Dmitrij Frischmann

Prof. Dr. Jan Baumbach

Die Dissertation wurde am 04.04.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 14.08.2023 angenommen.

Acknowledgements

I would like to start by expressing my sincere gratitude to my family for their unconditional support and encouragement. Their love and understanding have been essential in enabling me to pursue my research in faraway countries and complete this thesis.

I am grateful to my PhD supervisors, Dr. Markus List and Dr. Olga Tsoy, for their invaluable guidance, mentorship, and support throughout my research. Their expertise and insightful comments have been instrumental in shaping this thesis. I would also like to extend my sincere thanks to Prof. Tim Kacprowski for his guidance and to the chair leaders, Prof. Jan Baumbach, and Prof. Dmitrij Frishman, for hosting me in their labs and for providing me with the resources and opportunities to conduct my research. I would also like to express my gratitude to Dr. Hilal Tayara who helped me get into bioinformatics earlier in my career.

I further extend my heartfelt thanks to my friends, colleagues and fellow Ph.D. students for their stimulating discussions, support, and friendship.

Finally, I would like to acknowledge the support and funding for this research provided by e:Med under under the Sys_CARE project.

Zakaria Louadi

Abstract

Alternative splicing (AS) refers to differences in the processing of transcripts (e.g. exon skipping, intron retention, etc.) allowing cells to synthesise various protein variants (isoforms) from the same gene. Protein isoforms differ in their functionality and can even have opposite roles. Thus, AS is an essential mechanism in cell maturation and differentiation but also in diseases such as cancer, heart, and kidney diseases. On the other hand, Protein-protein interaction (PPI) networks and pathway databases are important resources in systems biology. PPIs are identified in tedious experiments but due to the high number of possible interactions, efforts are limited to testing only major protein isoforms, neglecting the considerable influence of AS on the interactome. Similarly, pathway databases only include a single isoform per gene, although most isoforms don't share all biological functions.

In this work, I first developed DIGGER (Domain Interaction Graph Guided ExploreR), a user-friendly database and web tool to explore the functional impact of AS in human-protein interactions. DIGGER integrates the PPIs with Domain-domain interactions (DDIs) to identify the binding domains for each PPI. Notably, none of the existing resources annotates the role of individual exons, which is a prerequisite to studying the consequence of AS on DDIs. To mitigate this, DIGGER provides a unique mapping of interface residues of interacting proteins to exons, based on experimentally resolved structures in the Protein Data Bank. In this way, genomic information on a splicing event can be directly mapped onto three-dimensional protein structures and the impact of the AS event on the PPI interface can be assessed. Through DIGGER's user-friendly web interface, researchers can interactively visualise the domain composition for any protein isoform, with detailed information on the interacting domains between the selected protein and its partners in the PPI network.

To leverage the joint PPI and DDI network in DIGGER for studying the consequences of AS across two or more conditions, I further developed the python tool NEASE (Network Enrichment method for Alternative Splicing Events). The classical approach for studying differential alternative splicing focuses on alternatively spliced genes, rather than the exact exons. In contrast, NEASE considers interactions impacted by AS and identifies enriched pathways based only on these edges. The analysis presented in this thesis shows that NEASE largely outperforms classic gene set enrichment in the context of AS and gener-

ates meaningful biological insights on the impact of AS. Together, DIGGER and NEASE provide essential resources for studying the mechanistic consequences of AS in systems and network medicine.

Zusammenfassung

Das alternative Spleißen (AS) bezieht sich auf Unterschiede in der Verarbeitung von Transkripten (z.B. Exon-Überspringen, Intron-Retention usw.), die es Zellen ermöglichen, verschiedene Proteinvarianten (Isoformen) aus demselben Gen zu synthetisieren. Protein-Isoformen unterscheiden sich in ihrer Funktionalität und können sogar entgegengesetzte Rollen haben. Daher ist das AS ein wichtiger Mechanismus bei der Zellreifung und Differenzierung, aber auch bei Erkrankungen wie Krebs, Herz- und Nierenerkrankungen. Auf der anderen Seite sind Protein-Protein-Interaktionsnetzwerke und Datenbanken von Wirkungspfaden wichtige Ressourcen in der Systemsbiologie. PP-Interaktionen werden in aufwendigen Experimenten identifiziert, aber aufgrund der hohen Anzahl möglicher Interaktionen beschränken sich die Bemühungen darauf nur die wichtigsten Protein-Isoformen zu testen, wodurch die erheblichen Auswirkungen der AS auf das Interaktom unberücksichtigt bleiben. Ähnlich enthalten Datenbanken von Wirkungspfaden nur eine Isoform pro Gen, obwohl die meisten Isoformen nicht alle biologischen Funktionen gemeinsam haben.

In dieser Arbeit habe ich zunächst DIGGER (Domain Interaction Graph Guided ExploreR) entwickelt, eine benutzerfreundliche Datenbank und Web-Tool, um den funktionellen Einfluss von AS in menschlichen Proteininteraktionen zu erforschen. DIGGER integriert die PPIs mit Domain-Domain-Interaktionen (DDIs), um die Bindungsdomänen für jede PPI zu identifizieren. Bemerkenswert ist, dass keine der vorhandenen Ressourcen die Rolle einzelner Exons annotieren, was eine Voraussetzung für die Untersuchung der Auswirkungen von AS auf DDIs ist. Um dieses Problem zu lösen, bietet DIGGER eine einzigartige Zuordnung von Oberflächen-Residuen interagierender Proteine zu Exons auf der Grundlage experimentell aufgelöster Strukturen in der Protein Data Bank. Auf diese Weise kann genomische Information über ein Splicing-Ereignis direkt auf dreidimensionale Proteinstrukturen abgebildet und der Einfluss des AS-Ereignisses auf die PPI-Schnittstelle bewertet werden. Durch die benutzerfreundliche Web-Schnittstelle von DIGGER können Forscher interaktiv die Domänenkomposition für jede Protein-Isoform visualisieren, einschließlich detaillierter Informationen über die interagierenden Domänen zwischen dem ausgewählten Protein und seinen Partnern im PPI-Netzwerk.

Um das gemeinsame PPI und DDI-Netzwerk in DIGGER zur Untersuchung der Auswirkungen von AS zwischen zwei oder mehr experimentellen Gruppen

zu nutzen, habe ich außerdem das Python-Tool NEASE (Network Enrichment Methode für alternative Splicing-Ereignisse) entwickelt. Der klassische Ansatz zur differenzieller Analyse alternativen Spleißens konzentriert sich auf Gene und nicht auf betroffene Exons. Im Gegensatz dazu berücksichtigt NEASE Interaktionen die von AS betroffen sind und identifiziert angereicherte Pfade auf der Grundlage dieser Kanten. Die Analyse in dieser Arbeit zeigt, dass NEASE im Kontext von AS deutlich besser abschneidet als die klassische Gene-Set-Enrichment-Analyse und sinnvolle biologische Erkenntnisse über den Einfluss von AS liefert. Zusammen bieten DIGGER und NEASE unverzichtbare Ressourcen für die Untersuchung der mechanistischen Konsequenzen von AS in der Systems- und Netzwerkmedizin.

Publications record

Main thesis publications

- **Louadi, Z**, Yuan K, Gress A, Tsoy O, Kalinina OV, Baumbach J, Kacprowski T, List M. DIGGER: exploring the functional role of alternative splicing in protein interactions. **Nucleic acids research** 49, no. D1 (2021): D309-D318.

- **Louadi, Z**, Elkjaer ML, Klug M, Lio CT, Fenn A, Illes Z, Bongiovanni D, Baumbach J, Kacprowski T, List M, Tsoy O. Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases. **Genome Biology** 22, no. 1 (2021): 1-22.

Other publications and pre-prints

- **Louadi Z***, Lazareva O*, Kersting J, Baumbach J, Blumenthal D, List M. DysRegNet: Patient-specific and confounder-aware dysregulated network inference. **bioRxiv**. 2022 Jan 1. * These authors contributed equally.

- **Louadi Z**, Oubounyt M, Tayara H, Chong KT. Deep splicing code: Classifying alternative splicing events using deep learning. **Genes**. 2019 Aug 1;10(8):587.

- **Louadi Z***, Oubounyt M*, Tayara H, Chong KT. Deep learning models based on distributed feature representations for alternative splicing prediction. **IEEE Access**. 2018 Oct 8;6:58826-34. * These authors contributed equally.

- Lio CT, **Louadi Z**, Fenn A, Baumbach J, Kacprowski T, List M, Tsoy O. Systematic analysis of alternative splicing in time course data using Spycone. **Bioinformatics**. Accepted 2023.

- Fenn AM, Tsoy O, Faro T, Rössler F, Dietrich A, Kersting J, **Louadi Z**, Lio CT, Völker U, Baumbach J, Kacprowski T, et al. Alternative splicing analysis benchmark with DICAST. **bioRxiv**. 2022 Jan 1.

- Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm DG, Heumos L, Kacprowski T, Lazareva O, List M, **Louadi Z**, Pauling JK, et al. The

AIMe registry for artificial intelligence in biomedical research. **Nature Methods**. 2021 Oct;18(10):1128-31.

- Oubounyt M, **Louadi Z**, Tayara H, Chong KT. DeePromoter: robust promoter predictor using deep learning. **Frontiers in genetics**. 2019 Apr 5;10:286.

- Bongiovanni D, Klug M, Scheibling E, Tsoy O, **Louadi Z**, Han J, Laugwitz KL, Condorelli G, List M, Bernlochner I. Beyond the prothrombotic transcript of RPs: alternative splicing and circular RNAs. **Cardiovascular Research**. 2022 Jun;118(Supplement):cvac066-166.

Book Chapters

- **Louadi, Z***, Tsoy, O*, Baumbach, J, Kacprowski, T, List, M. (2022). Not Quite the Same: How Alternative Splicing Affects Protein Interactions. **Protein Interactions: The Molecular Basis of Interactomics**, Wiley, 359-379. * These authors contributed equally.

Contents

Chapter 1

Introduction

1.1 Biological background

1.1.1 Gene regulation

Nucleic acids are macromolecules that are composed of nucleotides. The double strands of deoxyribonucleic acid (DNA) are one such molecule that stores genetic information and instruction for the functioning of all known forms of life that exist, and also some types of viruses. A typical eukaryotic's DNA is divided into several chromosomes in the cell nucleus and the mitochondria. In a human cell, for example, there are 22 pairs of autosome chromosomes and a pair of sex chromosomes. In addition to the DNA, ribonucleic acid (RNA) is another nucleic acid that also carries genetic material. It differs from DNA by being a single single-stranded molecule, less stable, and found in multiple types. The messenger RNA (mRNA) is one type of RNA molecule that corresponds to the genetic sequence used by ribosomes to direct the synthesis of specific proteins (translation).

Approximately 1-2% of the human genome consists of protein-coding DNA [80]. This is the portion of DNA that contains the instructions for synthesizing proteins, and it incorporates regions of around 20,000 protein-coding genes [15]. The rest of the genome consists of non-coding DNA that is not used for translation but contains crucial patterns that regulate the genetic information flow between DNA to RNA and then to proteins, providing structural support for the genome. This process, often referred to as "gene regulation", is characterized by being very dynamic as a means to respond to different environmental changes and cell types [28]. In essence, this mechanism decides how much of each RNA or protein product is made and how stable are they on the latter stages.

The life of the RNA molecule starts with the transcription process where the DNA is used as a template to make an mRNA. The first step is the identification of the regulatory binding sites in the DNA, such as promoters and

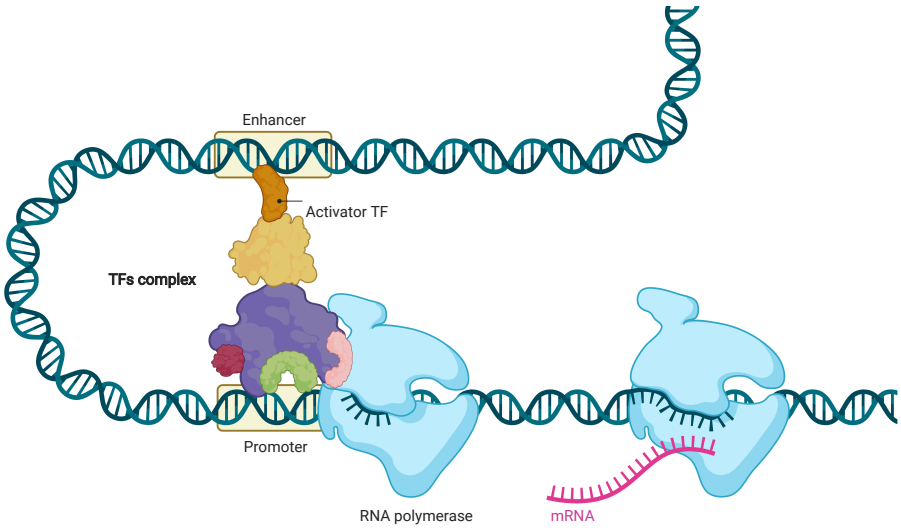


Figure 1.1: An illustration of the transcription process. The TF complex bends the DNA by binding in the enhancer and promoter regions and facilitates the binding of the RNA polymerase that produces the pre-mRNA. Created with BioRender.com.

enhancers, by specific proteins called transcription factors (TF). These proteins often work together and assemble a complex that can either promote or repress gene expression. Even though the binding sites can be thousands of base pairs away from each other as usually observed for enhancer regions, the TF complex can curve the DNA and bring it near a gene promoter region (Figure 1.1). This formulation makes it possible for the enzyme RNA polymerase to attach to the promoter and initiate the DNA unwinding to produce the RNA molecule.

A TF can also work as a repressor for a target gene by binding in the DNA regulatory region and blocking the RNA polymerase binding. Accordingly, a high expression of the activator results in an increase in the expression of the target gene. In the same way, an increase in a suppressor decreases the transcription. This is all regulated by each cell and it is crucial to make sure that the right amount of the target gene is expressed at the right time.

1.1.2 Alternative splicing

1.1.2.1 Overview

The primary RNA (pre-mRNA) is the initial molecule produced from the transcription of the DNA. It is made of exons, introns, and 5' and 3' untranslated regions (3'/5' UTR) [83]. Introns are the non-coding regions that are located between exons, which are the coding sequences of DNA and are often shorter

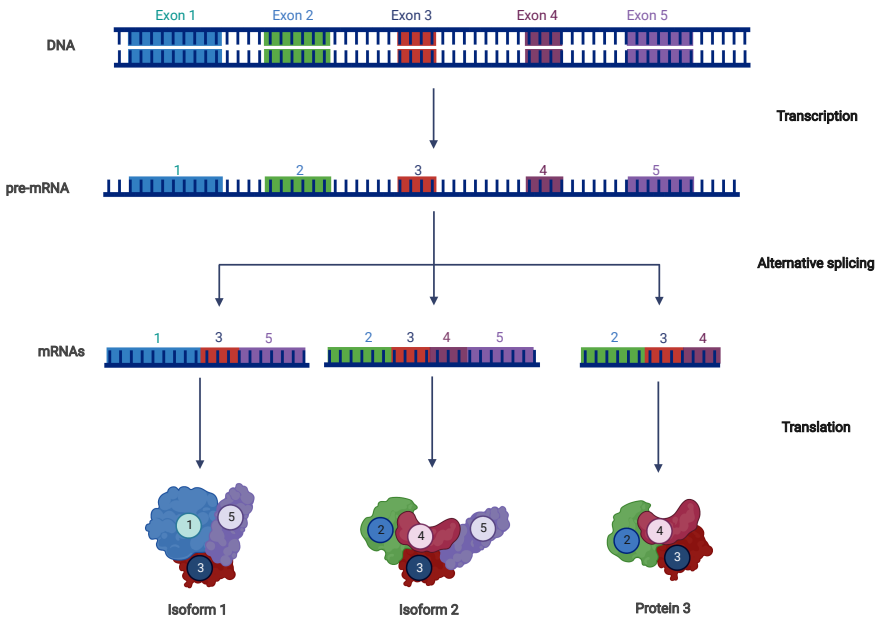


Figure 1.2: A pre-mRNA with 5 exons is produced from the gene's transcription. The alternative splicing process results in different combinations of exons that make 3 mature mRNA. The transcript variants, in the example, are protein-coding and thus transcribed to 3 proteins. These protein isoforms are similar in amino acid composition and share some of the structures. Created with BioRender.com.

and evolutionary more conserved [2]. Thus, before the pre-mRNA can be used to make a protein, it must undergo a process called splicing, in which the introns are removed and the exons are merged together. This is performed by a protein complex known as a spliceosome [83]. The resulting molecule of splicing is called mature mRNA and it mostly consists of exons and the 3'/5' UTRs. In some cases, an intron can also be retained with the rest of the exons [29].

Introns and exons can be removed and merged in different combinations, thus, one pre-mRNA frequently produces multiple transcripts and proteins. This process is called alternative splicing (AS) and it is the primary biological mechanism of interest in this thesis. AS is the regulatory mechanism that allows removing not just the introns but also some of the coding exons and alternatively including others together (Figure 1.2). For instance, in Figure 1.2 the pre-mRNA is composed of five exons, four being alternatively spliced and combined to produce 3 slightly different mature messenger RNA. Finally, the mature messenger RNA is translated to a protein if it has coding potential, otherwise, it is degraded often through the nonsense-mediated decay pathway. It is estimated that more than 95% of genes with multiple exons undergo AS [36]. From just around 20,000 coding genes in human genomes, AS can help produce over 120,000 transcripts and immensely boost human protein diversity [66].

AS is a highly regulated process. Even though the introns don't encode amino acids, they still play an important role in regulating the splicing process by enhancing or blocking the binding of splicing factors [83]. These are specialized proteins that compose the spliceosome and interact with the RNA molecule to decide the selection of exons. The most important regulatory element is the branch point, which is a conserved nucleotide, usually adenine, located near the 3' end of the intron. The branch point interacts with a component of the spliceosome called the U2 snRNP (small nuclear ribonucleoprotein particle), to covalently link the 5' end of the intron to the branch point. This process facilitates the formation of the lariat intermediate and allows the intron to be excised and the adjacent exons to be spliced together [18]. Additionally, both exons and introns contain certain sequences, known as splicing enhancers or silencers, that can affect the efficiency and accuracy of splicing [84].

Multiple types of AS exist including the skipping of a whole exon (or sometimes called cassette exon), retention of an intron, and the alternative usage of a 3' or 5' splice site (Figure 1.2). In addition, more complex events can also occur such as mutually exclusive exons where two exons never occur together. In mammalians, the most common and well-studied type is exon skipping [82].

The protein variants originating from the same gene are referred to as isoforms and often have different structures, functions, and locations within the cell. However, the true degree of the impact of AS is heavily debated; It is still not clear yet if all splicing variants have a coding potential or if they are functionally relevant in the cell [77]. Few argue that AS can also indirectly control gene expression even without producing coding variants since the expression of

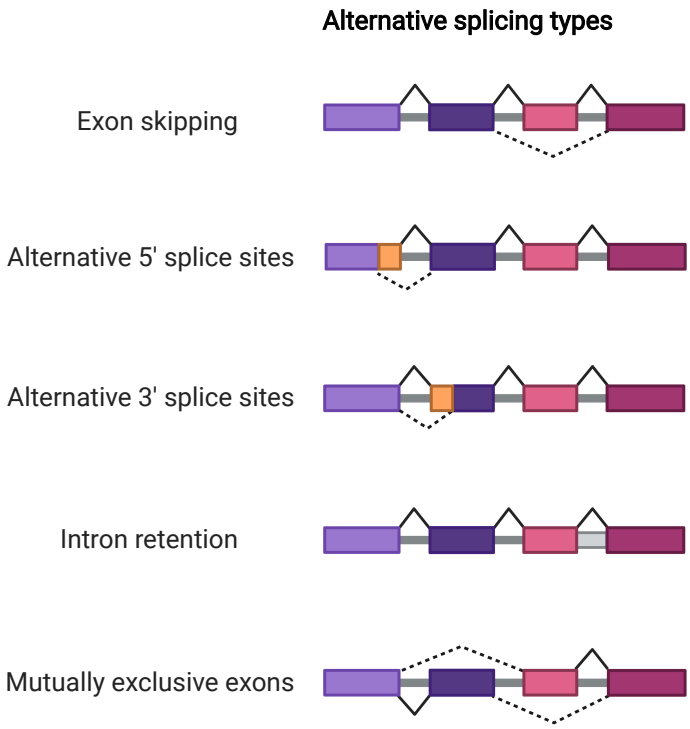


Figure 1.3: The most common alternative splicing types. Created with BioRender.com.

transcripts that are eliminated by nonsense-mediated decay negatively correlate with the expression of the protein-coding transcripts [91]. Therefore, these alternative non-coding transcripts produced by AS are also considered functionally significant.

1.1.2.2 Tissue-specific splicing

As discussed previously, the selection of the splice site depends on the regulatory sequence in the pre-mRNA (cis-regulatory elements). Other factors such as the amount of expression of a splicing factor in the cell are also important contributors to the exons selection decision.

Since the splicing factors can be expressed differently in some tissue, the same pre-mRNA can be processed in multiple ways depending on the cell type and state. Thus, AS is decisive for generating tissue-specific proteins by up- or down- regulating the inclusion of some exons. Numerous evidence shows that these exons are co-regulated and are likely to affect protein biochemical properties, structure, and interactions [83, 88]. A recent study validated these events at the proteomic level and uncovered a significant enrichment of tissue-specific exons in muscles and neurons[7]. These changes in isoform usage are vital for tissue identity, as well as cell differentiation, and maturation.

1.1.2.3 Mis-splicing in diseases

AS plays a crucial role in disease onset and development. It is estimated that mis-splicing is associated with at least 30% of all genetic diseases [83, 45]. The first known type of mis-splicing is cis-regulatory; such as mutations in a regulatory motif that can disturb the selection of the splice site, the branch point, or even cause the formation of a new exon (cryptic splicing). A prominent example of a cis-regulatory disruption of splicing is a single nucleotide variation in the gene SMN2 that causes mis-splicing at the junction of intron 6 to exon 8 in individuals affected by Spinal muscular atrophy [50]. The second type of mis-splicing is trans-acting, i.e. dysfunctional spliceosomal components. This latter is frequently characterized by disruption in multiple events or genes since one splicing factor can regulate the splicing of multiple genes. These defects are relatively rare, compared to cis-regulatory dysregulation, since they are often fatal [17]. Nevertheless, few diseases such as cancer and cardiomyopathy are linked to mutations and/or abnormal expression of RNA binding proteins [49]. One major example is the splicing factor “RNA Binding Motif 20” (RBM20). It has been shown that mutated or deficient RBM20 is present in up to 3% of individuals with familial dilated cardiomyopathy [13] and affects the splicing of at least 31 genes [30].

Thus, understanding the exact impact of mis-splicing is crucial to characterize a considerable amount of genetic diseases and to understand their progressions. Furthermore, in case of trans-acting mis-splicing such as a disruption of spliceosomes, the functional impact of AS should be examined systematically to

account for the combined effect of multiple disrupted and co-regulated events. This is one of the motivations behind developing the method NEASE, introduced in this thesis, by enabling the functional enrichment of mis-splicing (see subsection 2.3.2).

1.1.3 Proteins structure and interactions

1.1.3.1 Overview on proteins

The proteins are the result of translation, which is the process of decoding the mature mRNA in a ribosome into a chain of amino acids. Unlike the transcription and splicing processes, the translation occurs outside the nucleus, mainly in the cytoplasm. The ribosome operates by adding one amino acid at a time [19]. This procedure is decided based on the three-nucleotide subsequence (trinucleotide, codon) of the mRNA. Finally, the translation is terminated when a stop codon is found, which is also a trinucleotide. In this way, the final mRNA can be regarded as a template for the amino acid chain that comprises the protein [14].

Alongside the DNA and the RNA molecules, proteins are one of the most important molecules that perform a variety of functions within an organism, including providing cell structure, cell growth, and gene expression by enabling signal transduction and transporting other molecules [3]. They also act as enzymes and hormones. A few examples of such tasks were already covered in this introduction including DNA binding during the transcriptions and RNA binding during splicing which is maintained by specialized proteins called transcription factors and splicing factors respectively. As well as histone proteins that wrap the DNA and pack it around complexes [8]. All these mechanisms are dynamic and act following responses to stimuli. This complexity is often represented by biologists in terms of a pathway, where all the elements are included together, often in a chronological manner or a set of events, to describe a biological mechanism such as stimuli and responses [23].

1.1.3.2 Protein structure

Initially, the amino acid chain has a linear shape, known as the primary structure. The chain folds into a stable three-dimensional structure. The folding process goes through multiple steps starting with the secondary structure, where the arrangement of the amino acid chain occurs in two elements: Alpha helix and Beta strand [58]. The tertiary structure is the following level, where the entire amino acid chain (or polypeptide) folds into a three-dimensional structure [58, 52].

The fundamental unit of the tertiary structure is the protein domain, typically 40 to 350 amino acids long [52, 1]. The domain is self-stabilized and folds independently from the rest of the polypeptide [4]. Often, a single protein is composed of multiple such units with unique functions and properties

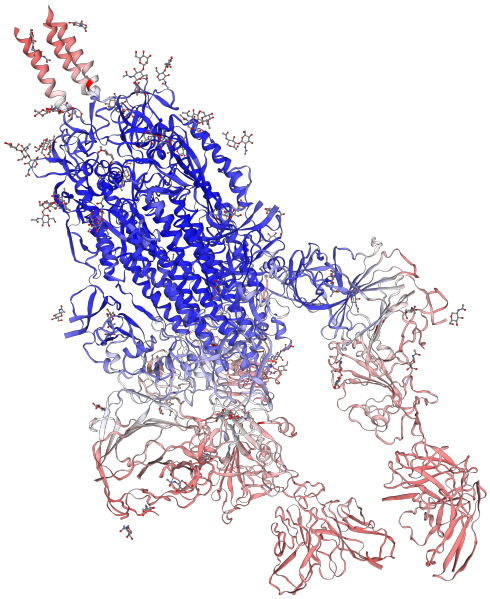


Figure 1.4: An example of a three-dimensional structure of the protein Spike glycoprotein (SPIKE_SARS2) from the virus SARS-CoV-2. Generated from (https://swissmodel.expasy.org/ [85]).

such as catalyzing chemical reactions, binding to other molecules, or providing structural support [52]. Domains with a similar structure can also appear in different proteins. Thus, the domains can be considered as building blocks of proteins used by evolution to make new proteins. In the case of multidomain proteins, the final arrangement of the domains is sometimes referred to as domain architecture [33] and it is at this level that the proteins become biologically functional.

Besides domains, intrinsically disordered regions are also an important class of polypeptide segments that do not have a unique three-dimensional structure but are characterized by dynamic state and multiple conformations [6]. The dynamic state of these regions helps with binding with other proteins as well as other molecules such as RNA, DNA, and ligands (More details in subsubsection 1.1.3.3).

Experimental protein structure identification is possible with techniques such as X-ray crystallography and nuclear magnetic resonance [51]. Furthermore, *in silico* methods for the prediction of a protein's three-dimensional structure are also attracting a lot of attention recently due to the remarkable progress in the accuracy of machine learning methods [89]. Notably, the machine learning algorithm AlphaFold 2 achieved an accuracy comparable to high-quality experiments, enabling a faster and more accurate annotation for the scientific

community [37]. Both experimental and high-quality predictions of the three-dimensional structure are deposited in large databases such as the Protein Data Bank (PDB) and the AlphaFold Protein Structure Database, where the users can either download or visualize the structure [10, 79].

Figure 1.4 shows an example of a three-dimensional structure of the protein Spike glycoprotein (SPIKE_SARS2) from the virus SARS-CoV-2 that causes *COVID-19* [85]. The protein consists of multiple domains that are coloured differently for visualization. The identification of the complete protein three-dimensional structure is a fundamental task for understanding the functional role of proteins and their evolution.

1.1.3.3 Protein-protein interactions

Proteins rarely act individually, instead, they tend to form complexes by binding to each other [20]. A single protein can have multiple partners or participate in different protein complexes. Nowadays, the majority of studies aiming to characterize the function of a protein within a living cell are done in the context of its interacting partners [92]. Thus, It is crucial to determine all protein-protein interactions within a cell. Furthermore, the goal of protein-protein interactions research is not just to identify the interacting proteins, but also to study the condition, the structure, and the stability of the interaction [55]. The answer to these fundamental questions can shed more light on the impact of genomics variants such as SNPs or transcript variants resulting from splicing or even help design new drugs and proteins.

Most often protein binding is a physical contact through a combination of hydrogen bonding and the hydrophobic effect. The strength of the binding depends on the size of the interacting region of the proteins [21]. The interaction is typically mediated either by a domain or a short linear motif. The latter are generally located in intrinsically disordered regions [61]. Since intrinsically disordered regions do not have a single unique tertiary structure, they are more adapted to binding to different interaction partners compared to domains [34].

To detect protein interactions, experimental biologists rely on a variety of techniques. The most common one is the Yeast Two-Hybrid which uses a yeast cell to test if two proteins interact using a reported gene. Affinity purification coupled with mass spectrometry is another popular technique that has the advantage of identifying a large number of interactions for a given protein at once [56]. Furthermore, structural-based methods such as X-ray crystallography provide valuable information about the structure of the complex at high resolution [12]. However, since they often require a high degree of protein purification and a specified environment to be successful, such methods can be less accessible and available.

We refer to the whole set of interactions as Interactome or Protein-Protein Interactions (PPI) and it is constructed by combining multiple experimental approaches for detecting interactions. Different variants of the interactome are

stored in publicly available databases such as STRING, Biogrid, etc [54, 72]. In system biology and most precisely the branch of network biology, the PPI is represented as a graph (or a network) with nodes as proteins and interactions as edges [22]. This representation is beneficial for a set of computational biology algorithms that aim to study the functionality of molecules within the cell systematically. More details about these methods, their applications, and their limitations are available in the section 2.2

1.2 Computational methods for transcriptomics

1.2.1 Overview of gene-level analysis

The most widely used method to quantify RNA in a sample is RNA sequencing (RNA-Seq). A typical RNA-Seq data analysis workflow starts with quality control of reads, then mapping them to the genome, and quantifying the number of reads per gene (gene expression) or per transcript (transcript expression). These steps result in a raw count matrix that represents the number of reads mapped to each gene or transcript in each sample. The matrix is then normalized to account for differences in library size and alternatively in gene/transcript length. The normalized counts are a good approximation of the amount of RNA from each gene in the sample. After correcting for technical variation, the gene counts still vary between samples, which is referred to as biological variation. Thus, using the count table and with a sufficient number of biological replicates, one can capture the association between TF and their target genes. This is possible because they are co-expressed; the amount of the RNA from the target gene is proportional to the amount of the regulating TFs at any given time. Such techniques are used to find modules of co-regulated genes and overlap them with known biological pathway pathways (Figure 1.5).

Clustering is another popular approach for identifying genes or/and patients with similar patterns. For instance, a cluster of patients could represent a disease subtype. Since clustering is unsupervised, it is not constrained to our traditional classification of phenotypes. This is especially useful for heterogeneous data or in the case of less characterized phenotypes. Clustering is also often used to compare the expression pattern of genes across tissues and cell types.

Often RNA-Seq experiments are designed in a control and case fashion. In such a case, one can also perform a supervised approach, such as differential expression analysis, to compare the mean expression of each gene between two or more groups. A well-known method is the use of generalized linear models such as Poisson or negative binomial that is fitted, for each gene individually, to estimate an expression mean and variance. Using packages such as DESeq2 and edgeR [48, 64], a statistical test is later applied to compare the expression between the two groups and results in a list of differentially expressed genes (Figure 1.5). This list is used in the downstream analysis to retrieve hypotheses about biological mechanisms that drive the difference between the two cases.

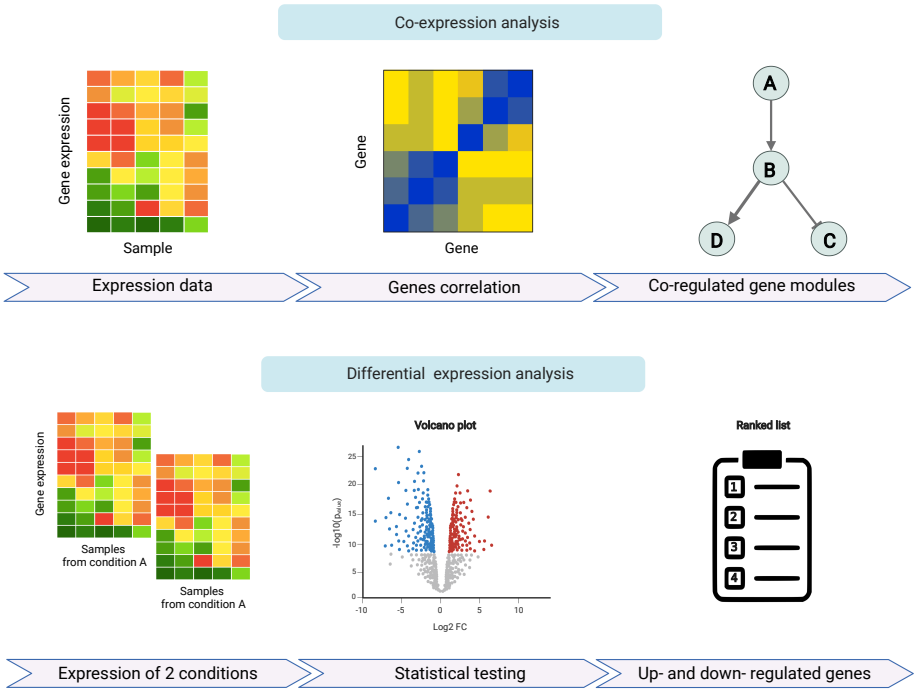


Figure 1.5: Overview of common transcriptomics data analysis methods. Created with BioRender.com.

1.2.2 Splicing-level quantification and differential analysis

In addition to quantifying gene expression, RNA sequencing technology can also be used to quantify individual transcripts or even exons to study AS patterns. The number of reads supporting exon inclusion is usually divided into the sum of reads supporting both exon inclusion and exclusion (Figure 1.6). The proportion is called the percentage spliced-in (PSI value) and represents the percentage of usage of each exon; a skipped exon can then be defined as an exon with a PSI value less than 1. The ratio can naturally be extended to quantify other types of AS events by comparing the number of reads supporting one splice site against all possible other splice sites [5]. This ratio is also normalized to account for some of the RNA-Seq biases such as exon length.

In some cases, the expression of a gene between two groups stays the same but the level of splicing differs and changes the proportions of produced isoforms. For this reason, similar to gene-level analysis, differential splicing methods are developed to compare the ratio of an exon inclusion across conditions. These tools output a metric often called Delta PSI (dPSI), which ranges from 0 to 1 representing the difference in the usage of the exon between the two groups. Multiple tools have been developed to identify and quantify splicing events, e.g, MAJIQ, Whippet, and rMATS [78, 71, 67].

Transcript-based approaches are another way of approaching the problem, where the focus is on quantifying the full transcript expression and comparing it, and then identifying "the switches" in the fraction of usage. These techniques have the advantage of being easier to interpret, unlike exon-based approaches. In particular, recent cancer studies have shown that multiple isoform switches are highly predictive of patient survival [81]. However, transcript-based approaches are restricted to annotated transcripts. Event-based and exon-based approaches on the other hand can identify new events and dysregulation and are less likely to be biased or limited by the availability of the reference genome annotation. One such straightforward application of event-based methods is the cryptic splice site search. This latter is a splice site that is generally regarded as dormant but could get activated because of a rare variant (often pathogenic) [39]. Since the resulting exons are usually not annotated, event-based methods are handier in such cases. For a detailed review and benchmark of methods used to quantify exon inclusion and differential splicing, I recommend the following studies [27, 86].

1.3 Overview of the thesis

1.3.1 Aim and motivation

At the time of writing, more than a decade has passed since the development of RNA sequencing (RNA-seq) [25]. During this period, a large amount of transcriptomics data was generated, allowing us to rapidly reduce the cost of experiments but at the same time increase the accuracy and the coverage of

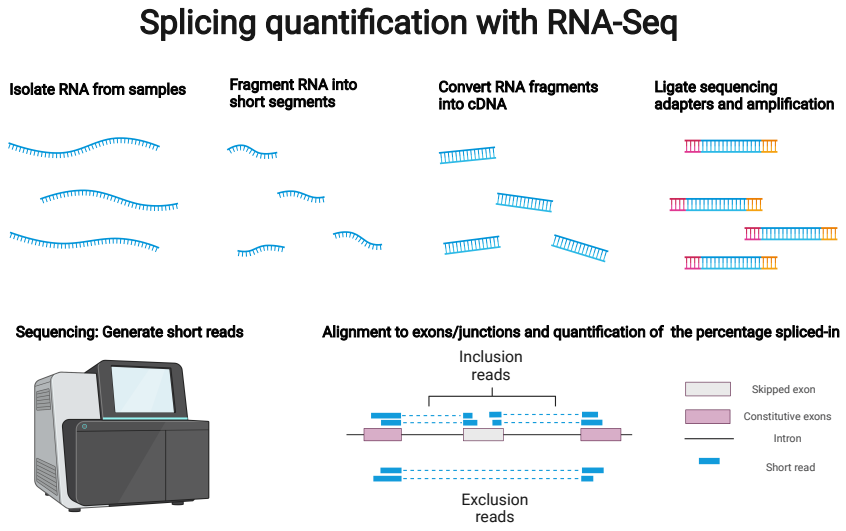


Figure 1.6: Splicing quantification with RNA-Seq. Created with BioRender.com.

sequencing. Unlike previous technologies such as microarray, RNA-Seq is not constrained to a reference [70]. This enabled a wider view of RNA biology, especially in the field of alternative splicing. The latter is, in simple words, the mechanism that allows a single gene to produce different protein variants. Thus, now we can discover many of these splicing variants easily and cheaply.

On the other hand, computational approaches to interpreting the functional role of transcriptome diversity still lag. A roadblock is the considerable knowledge gap concerning the function of most transcripts and proteins: the vast majority of splicing variants are still annotated with "unknown function". Protein-protein interaction databases are also often limited to the major protein isoform for every gene. Furthermore, since signalling and metabolic pathways only include the gene name instead of the protein variant, the widely used statistical methods such as gene set enrichment are as well gene-centred and neglect AS. Therefore, these resources are less suitable for interpreting the impact of splicing events, including tissue-specific exons/isoforms expression and differentially used exons. Both are crucial for understanding tissues and cell identity and disease pathomechanisms. For example, multiple splicing AS events have been recently linked to diseases including neurodegenerative disorders, cardiomyopathy, and cancer [40]. Hence, splice-aware methods are urgently needed to extract better biological insights from transcriptomics data.

In this thesis, I aim to develop computational methods to help address these limitations. In particular, I first introduce DIGGER, the first database and web

tool to explore the impact of AS in protein interactions. DIGGER integrates structural annotation and domain-domain interactions to construct an isoform-level protein-protein interaction network that can either be used for single gene analysis or transcriptome-wide analysis. Since DIGGER was not designed for multiple conditions analysis, I further extended it and developed NEASE, which is aimed to address the functional enrichment of AS events. The approach derives its power by focusing on exons and protein features they encode instead of the spliced genes as a whole. I evaluate NEASE in multiple datasets from both healthy and disease conditions and compared it with classical enrichment approaches. I show that NEASE provides unique and meaningful biological insights. The developed methods represent an important improvement over the state of the art and offer new opportunities to interpret the impact of AS at the system biology level.

1.3.2 Outline

This cumulative dissertation is based on two published manuscripts. In this chapter, I presented the essential biological and computational concepts. In Chapter 2, I introduce the state-of-the-art methods regarding AS analysis and their limitations to briefly motivate the need for a more robust approach for AS studies. In Chapter 3, I explain the methodology and resources used in the newly proposed methods. Chapter 4 summarizes the two publications of the tools DIGGER and NEASE and precisely describes my contribution. The detailed results are available in the full versions of the publications and embedded in Appendix A-B. Finally, Chapter 4 contains a general discussion of the thesis including the potential impact of the work, limitations, future work, and perspectives.

Chapter 2

State-Of-The-Art and Challenges

2.1 Overview

A wide range of approaches has been developed to interpret the results of RNA sequencing data. In this chapter, I briefly cover widely used methods and techniques. It is important to note that these methods were initially designed for gene-level studies (gene expression as a whole), but have later been extended to investigate alternative splicing (exons or junctions inclusion percentage).

The upstream strategies introduced in the last chapter such as differential expression, differential splicing, and co-expression usually result in a list of genes or exons that share a common pattern and are likely linked to the phenotype of interest. This is because biological molecules are likely to work in systematic and complex manners. Hence, system biology methods are designed to follow up and interpret the mechanistic link between the obtained entities. In this chapter, I will describe two categories of such methods: gene-set enrichment and network-based approaches (see section 2.2).

In the context of splicing, the tools for quantifying and comparing AS across conditions are available and relatively accurate. Yet, the recent availability of more AS data draws more attention to the challenges of interpretation of the functional impact of splicing. These challenges persist both at a single event and most critically for the systematic impact of AS. In the last section of the chapter, I explain the limitation of existing system biology approaches in the context of AS studies and the challenges that motivated the development of the methods DIGGER and NEASE.

2.2 Current approaches for interpreting expression profiles

Co-expression, clustering, differential expression as well as similar analyses often result in a list of thousands of genes. Unlike traditional approaches that focus on a single gene, system biology aims to extract meaningful biological insights from systematic data to understand the general state of the cells and tissues. The main argument in favour of relying on system biology approaches is the fact that genes do not work in isolation. For instance, the dysregulation of one TF can result in the down-expression of multiple genes that are either directly or indirectly regulated by the TF. Thus, It is essential to look at a biological system as a whole to understand the complete mechanism underlying the biological process.

Gene set enrichment analysis is the most popular of such approaches. It looks for over-representation of a gene set in already characterized pathways using hypergeometric or Fisher's exact tests. The prior knowledge about genes allows biologists to interpret an arbitrarily long list of genes with fewer human biases. Furthermore, these analyses break down the complexity of whole-genome or transcriptomics analysis to familiar concepts such as "cardiac development" or " cell cycle", to help interpret and generate new hypotheses. More details about the statistics side of the approach are presented in subsection 3.3.2. High-quality databases of non-redundant pathways are the key to success for the system biology approaches. An international community effort resulted in multiple such databases for different species such as KEGG and Reactome [38, 26]. These databases are curated and updated frequently to keep up with the most recent progress and discoveries. They also provide user-friendly visualization for a better understanding of the gene interactions [59].

Network approaches are another popular set of algorithms that can be performed to find enriched modules such as protein complexes or co-regulated genes (TFs and their targets). These methods are of particular interest since they are not limited to small and already known pathways but rather rely on the whole interactome or gene regulatory network. Therefore, they empower the possibility of finding new pathways or help discover novel gene candidates that were missing from the originally known pathway (biomarkers).

2.3 AS meets system biology: limitations and challenges

The AS mechanism, like the transcription, is co-regulated since the same splicing factors regulate multiple pre-RNAs. Thus, the system biology view for common AS patterns is important for deciphering the impact of this mechanism at the cell level. However, most of the RNA sequencing studies in recent literature were performed at the gene level only and without any splicing analysis at all.

The availability and the accuracy of tools to detect and quantify AS events are not the root cause of this practice but rather the difficulty of interpreting the results.

Even though some of the traditional gene-level approaches such as clustering and co-expression naturally extend to AS data, advanced interpretation techniques such as the PPI are not designed to address the immense variation in transcriptomics and proteomics produced by AS. For example, it is relatively easy to interpret the effect of the down-expression of a gene participating in a protein complex, but it is much more challenging to understand the impact of a skipped exon from the same gene in the interactions of the complex. The study of the systematic impact of AS such as interpreting AS-set events is also inherently more challenging than a gene set since the available pathway and gene set databases neglect the isoforms variants. Thus, interpreting the impact of AS events is a curtail roadblock to making AS study a routine part of transcriptome analysis.

2.3.1 AS impact on protein-protein interactions

The promising direction to interpret the functional consequences of AS is to evaluate the resulting changes in protein structure. Exon skipping, for instance, could cause the loss of a short linear motif, the shortness of a domain, or even its complete deletion. Thus AS can affect interaction mediated by these motifs or domains through the production of isoforms with different domains, or motif compositions. This rewiring of protein-protein interactions is immensely beneficial to cells since they can switch on or off different roles of the same gene depending on the physiological state of a cell, or disease phenotype. Recent experimental studies have shown that isoforms share less than 50% of interactions and less than 20% of isoforms are identical in terms of interactions [88].

However, the exact impact of rewiring protein interaction rewiring by AS is still far from understood because of the limitation of the current system biology approaches. To illustrate this limitation, let us assume an example where we are interested in gene A and its functional role in different tissues. Using the available large-scale transcriptomics datasets such as GTEx and TCGA ([16, 75], we could access its expression in different tissues and have an initial understanding of the gene activity. We could also search in the PPI databases for the interaction partners as well as their (co-)expressions, to further understand the activity of the gene in different environments (tissue or cell type). Now let us add the impact of AS to this complexity and assume that gene A produces two different isoforms called A1 and A2. The isoform A1 is expressed in a healthy heart and has three unique domains. While the isoform A2 is only expressed in the heart of patients with dilated cardiomyopathy and is slightly shorter with one less domain, because of the skipping of an exon (Figure 2.1 A-B). Using the current available PPI, we know already that gene A interacts with gene B, but since this information is only captured at the gene level it is not clear which of the isoforms A1 or A2 is interacting. Assuming the interaction is mediated by

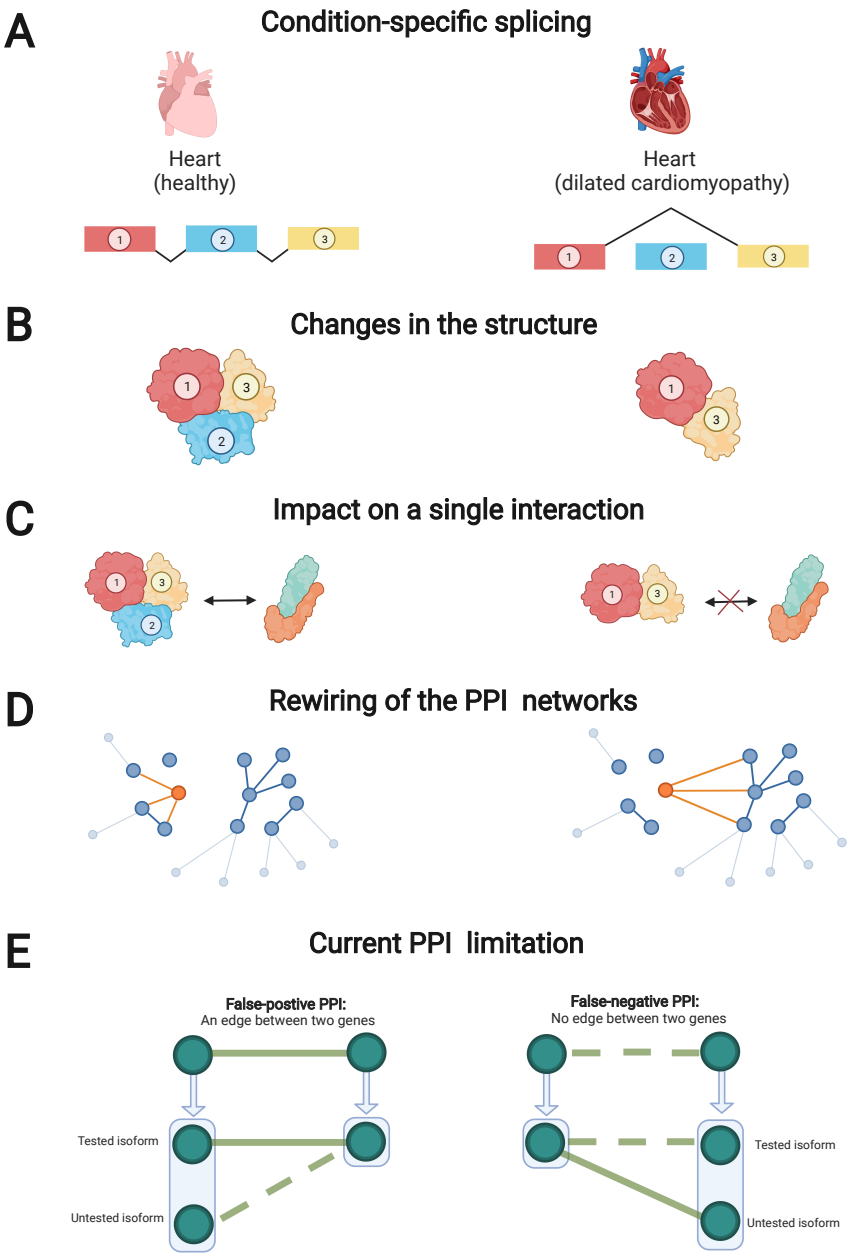


Figure 2.1: Impact of AS on protein-protein interactions and limitations of the current databases. (A-D) An illustration of two isoforms with different protein domains and interaction partners. (E) The current PPI representation neglects the effect of the alternative splicing and causes both false negative and false positive interactions. Created with BioRender.com

the additional domain missing in A1, this interaction will then be specific to the isoform A2 and thus specific to the heart tissue (Figure 2.1 C).

Considering that 95% of genes with multiple exons undergo AS in humans, both the interacting proteins in a PPI can have multiple isoforms with distinct expression patterns. Furthermore, the PPI rewiring from AS is often not limited to one edge but is observed globally in the network and includes multiple interactions and pathways (Figure 2.1 D). This complexity makes it very hard to experimentally test all possible combinations (Figure 2.1 E). As discussed by [73], the current method for testing gene interactions results in both false-negative and false-positive edges, and a more dynamic graph is needed to capture the complexity of the interactome. Hence a computational method to construct a condition-specific PPI is more appealing.

At the time of writing this thesis, only two methods address this challenge at the whole network level: PPIXpress [87] and DIGGER. A detailed comparison between DIGGER and PPIXpress is available in the first paper of this thesis (Appendix A). Briefly explained, DIGGER is built similarly to PPIXpress but extends it to a great extent, by including more protein structure information and offering other exclusive modes of analysis. One such addition is the exon analysis mode, which allows the users to examine the impact of new splicing events independently of transcript annotation. On the other hand, the isoform-level mode is another new feature that is designed to deliver an easy way to compare isoforms and their interactions with a particular focus on the "isoform-specific" interactions. Furthermore, the transcriptomics data analysis, provided in PPIXpress, is largely expanded in the DIGGER version to include protein complexes visualization and re-scoring of individual interactions. Instead of binary edges in the traditional PPI network, the scores suggested in DIGGER represent the confidence in each interaction given the isoforms in consideration. For example, by considering only one tissue-specific isoforms, DIGGER can generate a new re-scored PPI and focuses on rewired protein complexes. Finally, DIGGER offers, for the first time, a user-friendly and interactive web visualization to navigate between all of these modes.

2.3.2 Functional interpretation of AS events

As stated in the previous sections, differential expressions analysis is likely to output a large number of hits. This holds for both the analysis done at the gene or exon levels. To interpret a large list of hits, countless functional enrichment methods were introduced, but none of them was designed to address differential splicing analysis. The common approach used to overcome this limitation is by simply running enrichment on the genes originating from the differential exons or isoforms instead of the exons themselves. For instance, running a one-sided hypergeometric test on a list of differentially spliced genes to identify over-represented pathways (Figure 2.2). This simplification neglects the contribution of AS since genes can play multiple roles or participate in different pathways depending on the used isoform. Experimental studies even show that isoforms

could have an opposite role on a pathway. For example, the p38 isoforms are known to have an opposite role in the regulation of AP-1-dependent activities [57]. Similarly, it was revealed that TSC-22D1 isoforms have opposing functions in mammary epithelial cell survival [35]. While many such examples are being uncovered and validated experimentally, a computational method to truly scale our understanding of AS impact remains absent.

The main roadblock here is that pathway databases such as Reactome and KEGG are not isoform-specific, which requires more sophisticated methods than a simple statistical test. One promising strategy to address the challenge is the integration of different resources such as structure and PPI to predict the consequences of individual AS events on a pathway. Exon Ontology is the only tool, before NEASE, that was designed to perform a statistical test on a set of skipped exons [76]. It first determines the protein feature encoded by the skipping exons and then performs a permutation test to check if a well-characterized feature is hit by splicing more than expected by chance. The statistics used are indeed useful to identify common protein features that are regulated by splicing but it doesn't identify their functional impact.

For these reasons, NEASE (Network-based Enrichment method for AS Events) approach was introduced, as an extension of DIGGER, to tackle these limitations and to offer, for the first time, functional enrichment of an exon set as well as biological insights on the exact impact of AS. By focusing only on edges that are likely to be affected by splicing events, NEASE reduces the chances of false positive results (Figure 2.2). In the next chapter, I will explain the general method and statistics used in these tools.

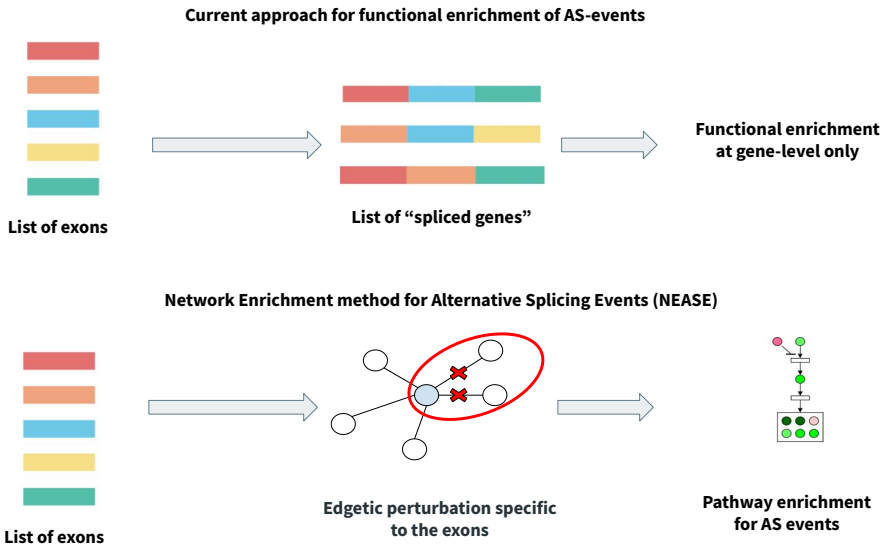


Figure 2.2: Current approach for functional enrichment of AS-events and a comparison with the method of NEASE. The proposed method relies only on the interactions of the domains and residues affected by splicing.

Chapter 3

Materials and Methods

3.1 Datasets and tools

3.1.1 Data sources

The PPI network used in this work is originally obtained from the BioGrid database BioGRID (v3.5) and the protein domain annotation from the Pfam database. The domain-domain interactions were downloaded from 3did (v2019) [53] and DOMINE (v2.0) [90] and the linear motifs instances and interactions from the ELM resource [41]. The co-resolved structure was obtained from the PDB [10] and pre-processed to extract residue-level and exon-level interactions. All ID mapping was performed using the Biomart mapping table [69]. The biological pathways collection was downloaded from the ConsensusPathDB database and includes 12 different pathway databases [32]. In the analysis presented here only the pathways from KEGG and Reactome were used, nevertheless, the Python package includes all 12 databases.

3.1.2 RNA-Seq datasets

All the RNA-Seq datasets used in this work are publicly available. Both TCGA and GTEx, which are the largest RNA-Seq datasets by scale available publicly, were used for validation. The Cancer Genome Atlas pan-cancer dataset [75] was downloaded via the Xena Browser (`https://xenabrowser.net/datapages`) and used in DIGGER's network-level analysis section to illustrate an example of condition-specific PPI. In this example, we construct a cancer-specific protein-interactions network to highlight lost interactions caused by cancer. GTEx was only used to validate the most critical tissue-specific exons from the enrichment results.

An extensive evaluation of NEASE performance, as well as a comparison against the other methods, was performed using three RNA-Seq experiments and the VastDB resource: which is a large-scale collection of AS events across

multiple datasets [74]. The evaluation scenarios, presented in the second publication of the thesis [47], include tissue-specific AS events, as well as a comparison between conditions both healthy and pathogenesis, as detailed below:

- A comparison between reticulated and mature platelets ([11], access numbers: GSE126448).
- A comparison between normal-appearing white matter and active lesions regions from postmortem white matter brains of multiple sclerosis patients ([24], access numbers: GSE138614).
- A comparison between heart tissue from healthy donors and dilated cardiomyopathy patients ([31], access numbers: EGAS00001002454).
- VastDB is a large collection of RNA-Seq experiments from different tissues with a focus on alternative splicing events ([74], https://vastdb.crg.eu/). We extracted neural-specific and muscle-specific exons from the available atlas of AS events.

Differentially splicing analysis was performed using MAJIQ [78] for the platelet and multiple sclerosis datasets. For the dilated cardiomyopathy dataset, the pre-processed data was used from the manuscript [31] and was originally performed using DEXSeq [62].

3.2 DIGGER: method description

3.2.1 Network biology notations

Network biology uses the mathematical notation from graph (or network) theory to represent complex biological relationships such as PPI and gene regulatory pathways. The most used notation to define a graph G is by defining an ordered pair $G = (V, E)$, where V is a set of vertices or nodes and E is a set of edges such as $E \subseteq V \times V$.

In the context of the PPI network, nodes represent either genes or proteins, and an edge is a physical interaction between them. By default, all PPI graphs are undirected, where all edges are bidirectional. As opposed to a directed graph where one vertice points to another. Mathematically a graph is say undirected if $(u, w) \in E$ necessarily implies $(w, u) \in E$.

Other definitions of interest include the order of the graph, which is the total number of vertices $|V|$, and the degree of a vertex v notated as $deg(v)$ and defined as the number of edges connected to it. Additionally, the degree sum formula states that the sum of the degree of all the vertices is twice the number of edges contained in it:

$$\sum_{v \in V} \deg(v) = 2|V|$$

.

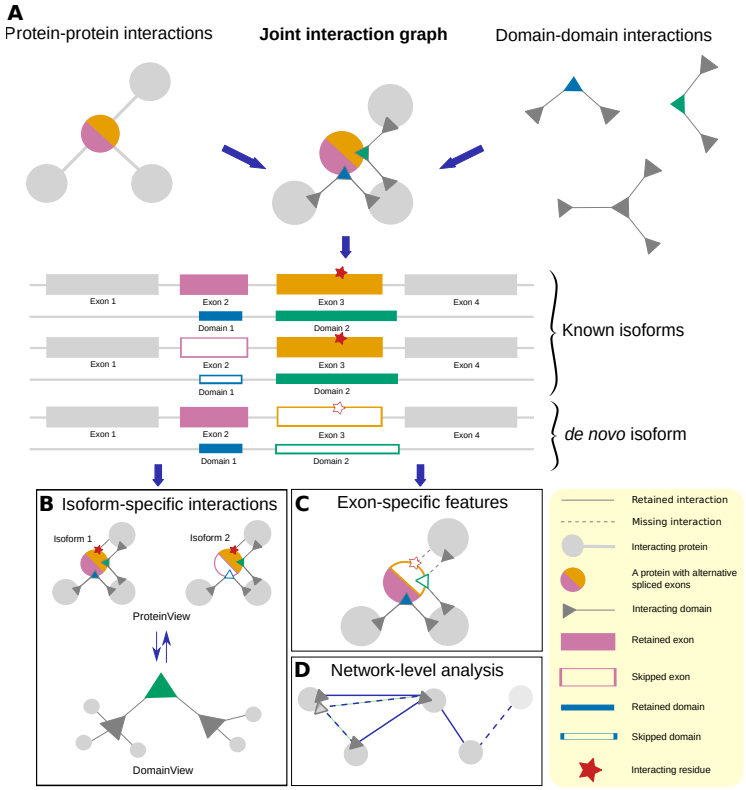


Figure 3.1: Overview of DIGGER method that incorporates protein-protein interaction with domain-domain interactions in a joint graph. DIGGERS offer three modes of analysis: exon-, isoform-, and network- levels (Reprinted from DIGGER's publication [46], an open-access article under the terms of the Creative Commons Attribution 4.0).

3.2.2 Joint graph construction

A protein domain can either interact with another domain or with a linear motif. Several studies have attempted to construct domain-centred interaction databases that store domain-domain interactions (DDI) and domain-motif interactions (DMI). These databases are based on the 3D structure of domains and proteins and/or known protein complexes [41, 53]. The same domain can be a part of multiple proteins, thus, unlike the PPI databases, DDI and DMI databases provide interacting interfaces independently of the originated genes. The integration of the PPI with structural information such as DDI and DMI can further enhance the details of protein interactions. Both DIGGER and NEASE rely on this concept and aim to construct a structurally annotated PPI. This network is, in principle, similar to the classical binary PPI network by the fact that it provides interacting genes, but additionally involves the exact position of the amino acids that mediate the binding. In addition to both DDIs and DMIs, we additionally enrich the PPIs with the available amino acid level interactions. These are derived from protein complexes with resolved structures in The protein Data Bank resource [10].

By annotating each interaction with the protein interfaces mediating it, DIGGER identifies interactions that are unique to only some of the isoforms Figure 3.1. Furthermore, the second mode of DIGGER allows the prediction of the effect of non-annotated splicing events such as exon skipping by mapping the skipped exon to its structure and interfaces. Finally, the last level of analysis provided by DIGGER is an extension of the same idea to handle the interaction of multiple isoforms. This mode takes as input a list of isoforms and transcripts from a couple of isoforms up to whole transcriptomics data. More details of this mode are available in the next section.

It is important to note that one interaction can have multiple interfaces too. In our scenario, the interaction can be mediated by either one or multiple domains, linear motifs, or known residues from the co-resolved structure in the PDB. In practice, to construct DIGGER's new PPI, we integrated both PPI and DDI into a single joint graph. The detailed pseudocode for constructing this graph is presented in the Algorithm 1. Briefly, we annotate every PPI with known DDIs of the interacting genes. Thus the nodes in the new graph represent unique protein domains and are defined by concatenating gene id and domain id. It is worth mentioning that in this representation, one single PPI can be annotated with multiple edges and more than two nodes. For both linear motif and residue level interactions, the information was saved as a database instead of a graph.

3.2.3 Network-level analysis of DIGGER

In addition to isoform and exon-specific interactions, DIGGER also provides a comprehensive view and analysis of multiple interactions between isoforms from RNA-Seq datasets from a specific tissue, condition, and developmental

Algorithm 1 Join PPI and DDI network construction.

```
for Every edge (X, Y) in the PPI do
    Get all the n domains of gene X: Di for i in i = [1, 2, ..., n].
    Get all the m domains of gene Y: Dj for j in i = [1, 2, ..., m].
    for Every domain Di of gene X do
        for Every domain Dj of gene Y do
            if The edge (Di, Dj) is in in the DDI graph then
                Append the joint graph with the edge (XDi, YDj).
            end if
        end for
    end for
end for
```

stage. We refer to this task simply as network-analysis mode but it is worth mentioning that previous studies used the term "condition-specific PPI" [87]. Typical examples of application include the analysis of isoform switches affecting one or multiple genes in a protein complex or a tissue-specific PPI. In our case, we do not just filter unexpressed genes in a given tissue but also involve the impact of splicing in rewiring interactions.

Constructing a condition-specific PPI using DIGGER requires either a user-defined list of isoforms or an expression table of transcripts. The DIGGER network-level algorithm is inspired by PPIXpress but extended by computing a confidence score for every interaction. The PPIXpress method on the other hand only uses the most expressed isoform for each gene and filters all interactions of the rest. Instead, DIGGER provides more flexibility by offering a filter option and re-scoring each interaction based on the ratio of missing interactions or isoforms. The output of this mode is a weighted PPI, where the score models the confidence of interaction in the condition. Accordingly, a score of 0 means that all interacting isoforms are absent and a score of 0.5 signifies that only half of the interaction interfaces are missing (Figure 3.2). Finally, DIGGER uniquely provides an interactive visualization together with a comprehensive analysis workflow that links the network-level analysis mode to other parts of the databases to further explore specific gene isoforms. A simplified workflow for the network-analysis mode is explained in the Figure 3.3.

3.3 NEASE: method description

3.3.1 Overview on the hypergeometric distribution

The hypergeometric distribution is frequently applied in enrichment analysis to estimate the significance of the results. It is a discrete probability distribution since its random variable is a count value. It is a very similar distribution to the binomial one, since it describes the probability of k success in n draw, from a population of size N that has K success on it. In both the binomial and

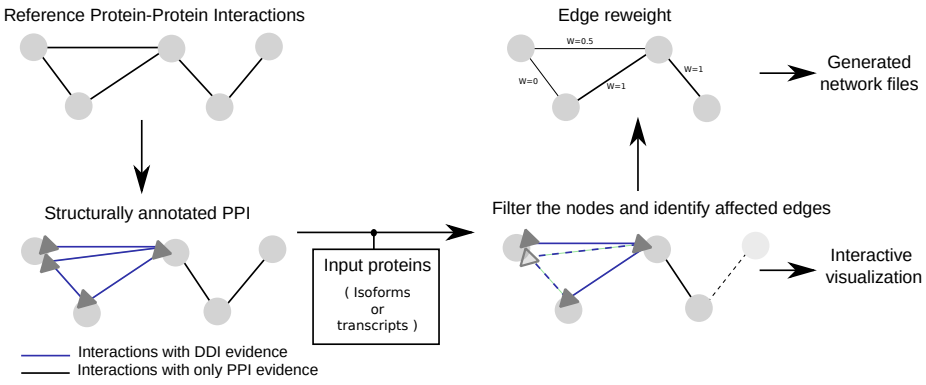


Figure 3.2: DIGGER constructs a condition-specific PPI and highlights domains absent in the user-submitted isoforms and their interactions (Reprinted from DIGGER's publication [46]).

hypergeometric distribution the result of each draw is binary (e.g; success or failure). But unlike the binomial, the hypergeometric distribution describes a drawing experiment without replacement, which means that a single object can only be drawn once. Consequently, the percentage of success and the population changes with every draw, unlike the binomial that describes the probability with replacement.

The probability mass function of a random variable X that follows that follow the hypergeometric distribution is:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Where k is the observed number of successes in n draws, out of K possible success in a population of size N. The symbol ! indicates the factorial operator and $\binom{A}{a}$ is the binomial coefficient and can be interpreted as the number of ways of choosing a elements out of A possibilities. It is defined as:

$$\binom{A}{a} = \frac{A!}{a!(A-a)!}$$

In the rest of the thesis, the probability mass function is simply denoted as:

$$\mathrm{X} \sim \mathrm{Hypergeometric}\,(\mathrm{n}, \mathrm{K}, \mathrm{N})$$

3.3.2 Fisher's exact test for enrichment analysis

In enrichment analysis, the hypergeometric distribution is used to calculate the probability that a list of genes of interest (e.g. differentially expressed genes)

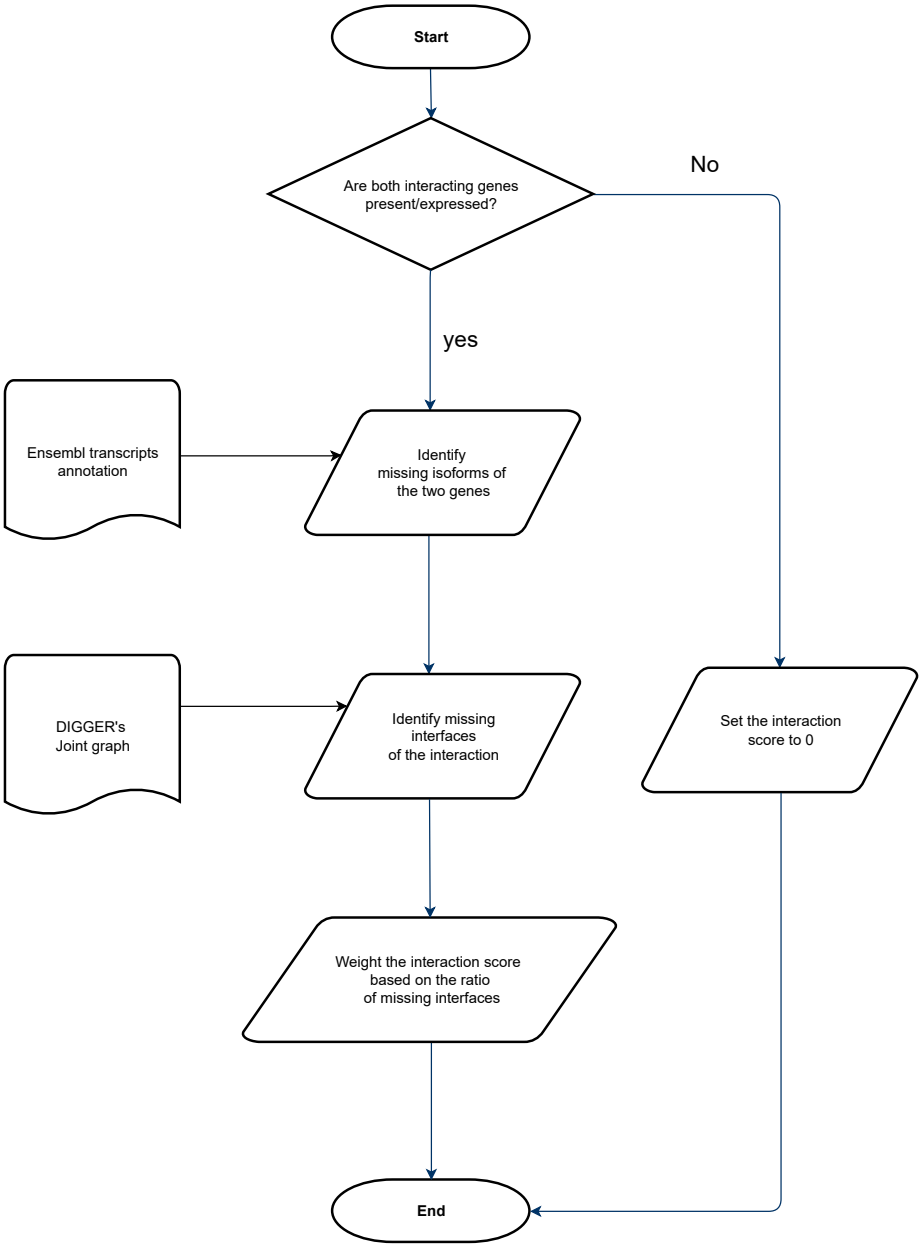


Figure 3.3: The workflow of the network-level analysis mode of DIGGER. DIGGER process the user input that contains a list of transcripts or protein IDs and constructs a condition-specific PPI by identifying the interactions specific to the isoforms in the list and removing the rest.

is over-represented in a gene set (pathways, co-expression in previous experiments, or location on a chromosome). Assuming we run differentially expressed analysis between healthy and tumour samples, we found 100 up-regulated or down-regulated genes. We would like to know if the list of differentially expressed genes is over-represented in a pathway of interest of 300 known genes. If that is the case we can hypothesize that this pathway is linked to the tumour.

In this case, we can use the analogy of randomly sampling 100 genes out of a population of all human genes (around 20,000). The drawing is without replacement because, in this scenario, a single gene can only be sampled once. Thus, the number k of genes overlapping between the differentially expressed genes and the list of genes of the pathway follows a hypergeometric distribution with the following parameters:

$$\mathrm{X} \sim \mathrm{Hypergeometric}\left(\mathrm{n} = 100, \mathrm{K} = 300, \mathrm{N} = 20,000\right)$$

Assuming the overlap is 5 genes (k=5) and using the probability mass function of the hypergeometric distribution, we can calculate the exact probability of randomly choosing 5 genes from the pathway:

$$P(X = 5) = \frac{\binom{300}{5}\binom{20{,}000-300}{100-5}}{\binom{20{,}000}{100}} = 0.01343$$

This is a low probability which could indicate a link between the pathway and the disease. However, in statistical testing, we are interested in getting a p-value that represents the probability of drawing k or more success. An event might have a low probability but a high p-value. In our case, we need to calculate the probability of getting 5 or more overlapping genes $P(X >= 5)$ by summing up multiple events. This probability (p-value) needs to be small enough to reject the null hypothesis that the genes are picked at random from the total gene population. Usually, a threshold of either 0.05 or 0.01 is used to reject this null hypothesis.

Since we are only interested in over-representation, the applied hypergeometric test is one-sided and is identical to the one-tailed version of Fisher's exact test. The test is used in the analysis of 2x2 contingency tables and has a wide range of applications including enrichment analysis [63].

Table 3.1: An example of a contingency table representation for enrichment analysis.

-	Diff. expressed	Not Diff. expressed	Total
Part of the pathway	a (=k)	b	a + b (=K)
Not a Part of the pathway	c	d	c+d
Total	a+c (=n)	b+c	a+b+c +d (=N)

Another formulation of the enrichment analysis is the 2x2 contingency table. As illustrated in Table 3.1, the list of differentially expressed genes is split the

two: the first is designated to genes that are also part of the pathway, their number is noted a and it is the same number of success k in the hypergeometric distribution that we want to inspect its significance. The second is the list of genes that are differentially expressed but not part of that pathway (c genes). Thus the number of draws n is equal to $a + c$. Similarly, the rest of the genes that are not differentially expressed are split depending if they are overlapping with the pathway or not. From the table, it is easy to observe that $a + b = K$ is the number of genes in the pathway (or possible successes). After forming a contingency table the p-value of Fisher's exact test can be calculated by summing the probability of observing the obtained table and the probability of more extreme tables (higher values). In the rest of the thesis and the second publication, we referee to this approach as either "classical enrichment" or "gene-level enrichment" and we afterwards introduce a new test at the edge level for AS study (see subsubsection 3.3.3.2)

In enrichment analysis, a common practice is to perform a test for every known pathway in one database. This practice yields simultaneous testing of more than one hypothesis and gave rise to multiple testing errors. Thus, multiple testing corrections have to be conducted to adjust the original p-values. The most popular method for adjustment is Bonferroni correction which divides the original p-value with the number of tests performed (m). Since this is a conservative approach, it might not be very appropriate for enrichment analysis. Alternative approaches, such as the Benjamini–Hochberg are more suitable [9] to control the false discovery rate. The Benjamini–Hochberg strategy works by first sorting the original p-values from the smallest to the largest and then calculating the critical value: $(r/m)\cdot\alpha$ where r is the rank of the original p-value, m is the number of tests and α is a selected false discovery rate. The procedure ends by locating the test with the largest p-value that is smaller than the critical value (the new significance condition). NEASE's enrichment approach uses the Benjamini–Hochberg method as the default for adjusting the p-values.

3.3.3 Statistics and hypothesis testing in NEASE

3.3.3.1 Overview of the method

The new method NEASE studies the impact of AS systematically. First, it detects all the edges affected by splicing. This is done by mapping the list of exons to their protein features (such as domain, residues, and linear motifs) and then identifying the exact interactions mediated by these regions (Figure 3.4-A and B). Thus, NEASE provides an exon-centric view of the PPI network where a list of exons is represented as edges rewired by AS.

To estimate the significance of these edgetic changes and calculate an enrichment p-value, NEASE projects the pathway of interest to the PPI (Figure 3.4-B) and counts the number of affected interactions that are directly linked to the genes of the pathway. Finally, the statistical significance of the connectivity between the AS edges and the pathway is calculated using a one-sided hyper-

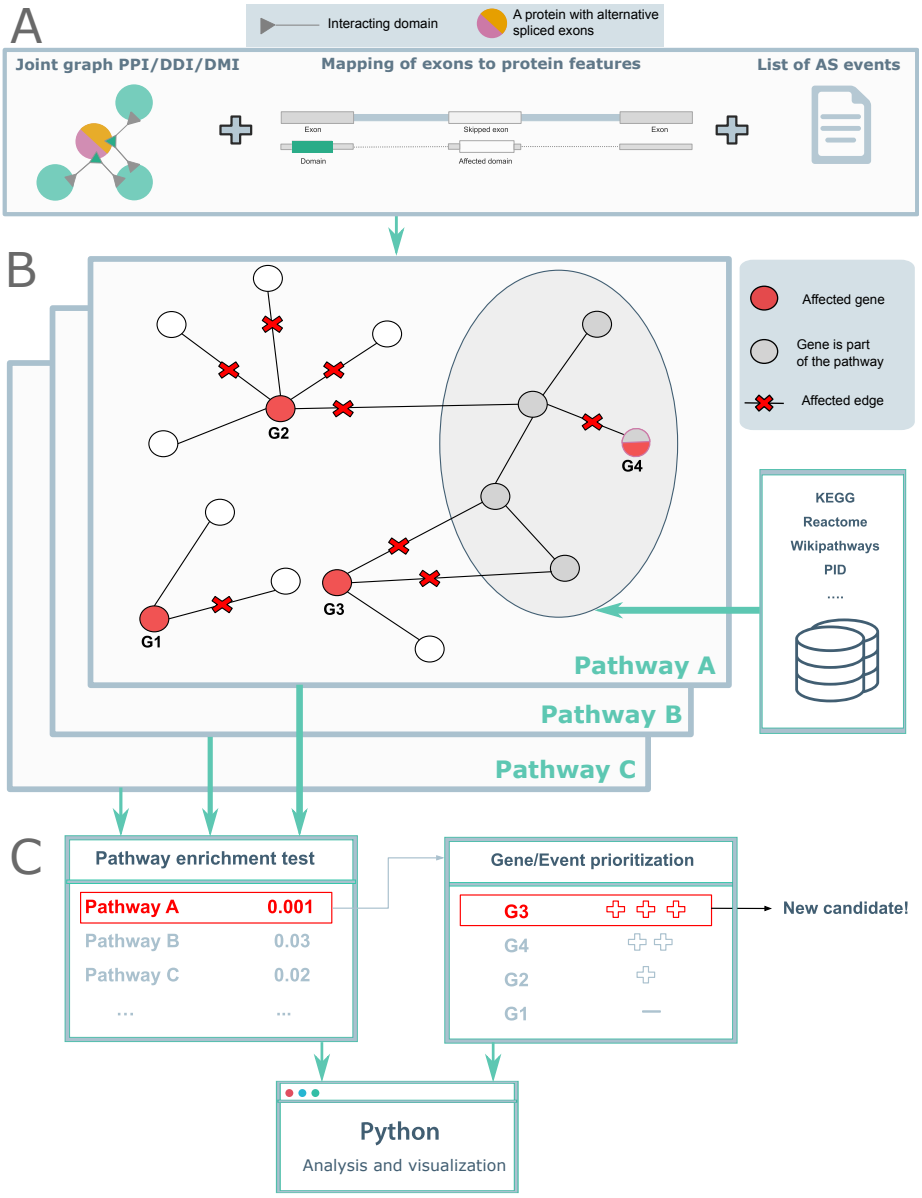


Figure 3.4: Overview of NEASE's procedure to run functional enrichment of alternative splicing events (Reprinted from NEASE's publication [47], an open-access article under the terms of the Creative Commons Attribution 4.0 License).

geometric test as explained in the next section. The procedure is repeated for every pathway to obtain a list of p-values that are afterwards adjusted for multiple testing.

The NEASE method can also run on a single event level by only considering the interactions affected by that event. This has the advantage of prioritizing the most relevant genes that are associated with a pathway. By running a test on single genes, we reduce the effect of hubs that could be connected to a pathway just by chance since they have a high node degree (many edges). The hypergeometric probability used calculates how likely the affected edges from the gene are just connected to the pathway by chance and the genes are sorted accordingly (Figure 3.4-C).

3.3.3.2 NEASE metrics for genes and pathways ranking

The one-sided hypergeometric test used in NEASE is similar to the classical test used for over-representation analysis (as previously explained in subsection 3.3.2) but revised to accommodate the edgetic effect of AS introduced by our approach. A similar method was introduced in [68] for general gene set enrichment at the gene level. The statistical test is conducted as follows:

- The total degree of the structurally annotated graph is calculated using the degree sum formula. This represents the population N of the distribution. It is important to note that the degree sum formula states that one edge is counted twice, which is in line with our representation here since one edge can be drawn in two ways (two interacting genes can be affected).
- The number of all affected edges from all exons is the number of draws n.
- For every pathway P, the total degree of all genes is summed up. This also includes edges coming from external genes of the pathway. This number represents the number of all possible successes K as defined previously in the hypergeometric distribution. In our case, it also denotes all the possible ways of selecting (drawing) an edge connected to the pathway. Hence, in our hypothesis, an edge can be internal or external since we assume that genes outside of the pathway can also have an impact on genes on the pathway.
- From all affected edges, NEASE detects the ones that are connected to the pathway and represent the observed number of successes k. This is also the number we want to check if it is significant since it describes the probability of association between the set of exons (or their interactions) and the pathway.

We then model the random variable that the outcome is k using a hypergeometric distribution as follows:

$$\mathrm{X} \sim \mathrm{Hypergeometric}\,(\mathrm{n}, \mathrm{K}, \mathrm{N})$$

Where:

- n: The number of affected edges.
- K: The degree of the pathway.
- N: The degree of the structurally annotated PPI using the degree sum formula.

The p-value is eventually calculated by adding up the probabilities of a k success or higher. If the test is performed for a single gene the hypergeometric distribution is alternatively modelled as the number of observed successes from the same gene or the number of edges from the gene that are connected to the pathways:

$$\mathrm{X} \sim \mathrm{Hypergeometric}\,(\mathrm{m}, \mathrm{K}, \mathrm{N})$$

Where m is the total number of affected edges from that gene. This gene-specific p-value is used to rank genes in a single pathway or to find new biomarkers in the case of disease studies.

NEASE also introduces a new weighted score as a complement to the enrichment p-value obtained from the hypergeometric test. The intuition behind the score is to further rank and prioritize pathways that have more significant genes and fewer hubs. Since the initial test considers all affected edges irrespective of what genes they come from, the probability can be affected by a single gene with a considerably higher number of interactions. Accordingly, the proposed score scales the original p-value by the number of significant genes (obtained from individual genes test), as follows:

$$\text{NEASE score} = -\sqrt{g} \times \log_{10}(p \text{ value})$$

The general pseudo code of the NEASE approach is presented in the Algorithm 2 and the gene-specific test in the Algorithm 3.

3.3.3.3 Permutation tests

Permutation tests were performed to further validate the robustness of the enrichment results obtained from NEASE. We initially ran NEASE on tissue-specific exons and found relevant significant pathways such as "Muscle Contraction" for the set of heart-specific exons and "Synaptic vesicle cycle" from the set of neural-specific exons. We then hypothesize that the enrichment was just due to chance and that tissue-specific exons are not any different than other skipped exons present in this tissue. Thus, the exact null hypotheses are:

Null Hypothesis 1 *Any random set of skipped exon events in expressed genes, of the same size, present in the heart can lead to an enrichment p-value of the "Muscle Contraction" pathway as low or lower than heart-specific exons using the NEASE approach.*

Algorithm 2 Pathway enrichment test.

Construct the structural annotated graph $\mathbf{G}'$.
Calculate $\mathbf{N}$ the total degree of $\mathbf{G}'$.
for Every exon in the submitted query list **do**
    Identify the list of affected domains, motifs, or residues.
    Update $\mathbf{e}_{G'}$ affected edges.
end for
Calculate $\mathbf{n}$ the total number of affected edges $\mathbf{e}_{G'}$ from the query.
for Every pathway $\mathbf{P}$ **do**
    Calculate $\mathbf{K}_P$ the pathway degree in the graph $\mathbf{G}'$.
    Calculate $\mathbf{k}$ number affected edges from $\mathbf{e}_{G'}$ that are part of $\mathbf{P}$.
    Calculate the one-sided $\mathbf{pvalue}_P = Hypergeometric(\mathbf{k}, \mathbf{n}, \mathbf{K}_P, \mathbf{N})$
    Calculate the gene-specific p values using **Algorithm 2**.
    Calculate $\mathbf{SCORE}_P$ the adjusted NEASE score using (Eq.: 1).
end for
Correct for multiple testing using Benjamini-Hochberg.
Rank pathways based on adjusted p values or NEASE scores.

Algorithm 3 Gene-specific enrichment test.

For a pathway of interest $\mathbf{P}$ with a degree $\mathbf{K}_P$, this function returns a p-value for every gene.
for Every spliced gene $\mathbf{i}$ **do**
    Calculate $\mathbf{n}'$ the number of affected edges from the gene $\mathbf{i}$.
    Calculate $\mathbf{k}'$ the number of edges that are part of $\mathbf{P}$.
    Calculate the gene-specific one-sided $\mathbf{pvalue}_{P_i} = Hypergeometric(\mathbf{k}', \mathbf{n}', \mathbf{K}_P, \mathbf{N})$
end for
Rank genes based on p values.

Null Hypothesis 2 *Any random set of skipped exon events, of the same size, present in the brain can lead to an enrichment p-value of the "Synaptic vesicle cycle" pathway as low or lower than brain-specific exons using the NEASE approach.*

From the set of highly confident skipped exons in VastDB (based on the number of samples supporting them), we randomly sample 10,000 random sets of exons of the same size as the original set. Let $P = \{p^{(1)}, \ldots, p^{(10,000)}\}$ be the set of 10,000 p-values obtained from running the permuted exon sets on NEASE enrichment and let $p^{(obs)}$ be the originally observed p-value from the up-regulated exon (tissue-specific). The empirical p-value is then obtained using the formula below. The +1 is added here to avoid an empirical p-value of exactly 0.

$$p^{(empirical)} = \frac{|\{p' \in P : p' \leq p^{(obs)}\}| + 1}{10,000}$$

3.4 Implementation of the tools and availability

The user-friendliness of the developed tools is an important aspect of this work. Both DIGGER and NEASE offer multiple and diverse options for the user to dig deeper into the acquired results.

The DIGGER web tool was designed using the Python web framework Django with graph visualization components using the Javascript library vis.js that was accommodated to supply the specialized features of DIGGER such as protein, domain, and missing domains and edges. The web tool offers three different modes that can be used interchangeably: Isoform-Level analysis, Exon-Level analysis, and Network-Level analysis with easy navigation between them (Figure 3.5): the user can, for instance, input a list of isoforms to construct a weighted subgraph of their interactions and visualize them. Likewise, this mode is connected to two other modes to allow the user to dig deeper into one specific gene of interest and compare its isoforms or even to a single exon and its encoded protein features.

NEASE is a Python package with object-oriented features that allow the user to run an enrichment job on multiple databases either separately or combined. After running the initial enrichment, the user can further focus on an individual pathway to prioritize the most relevant splicing events and biomarker candidates. The Python package was linked to the DIGGER database and individual events can be visualized there. NEASE also provides visualization for the whole pathways modules in the PPI together with the spliced genes and relevant edges.

The source codes for both DIGGER and NEASE are released as open-source under the GPLv3 license as well as the Python notebooks to reproduce most of the results of the two publications, as summarized in Table 3.2.

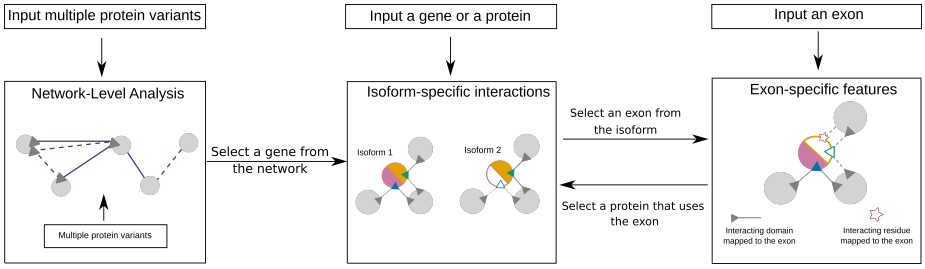


Figure 3.5: Navigation through DIGGER database (Reprinted from the Supplementary Information of DIGGER's publication [46]).

Table 3.2: Source code and data availability

Description	Link of the code
DIGGER webtool source code	https://github.com/louadi/DIGGER
DIGGER webtool live link	https://exbio.wzw.tum.de/digger/
DIGGER databases and graphs	https://exbio.wzw.tum.de/digger/download/
DIGGER documentation	https://zenodo.org/record/4010881
NEASE Python package (source code)	https://github.com/louadi/NEASE
NEASE Python package (PyPI)	https://pypi.org/project/nease/
Cancer Genome Atlas pan-cancer analysis with DIGGER	https://github.com/louadi/RNA-Seq-DIGGER
NEASE tutorials	https://github.com/louadi/NEASE-tutorials

Chapter 4

Publications

4.1 DIGGER: exploring the functional role of alternative splicing in protein interactions

The full citation of the paper

Louadi, Z, Yuan K, Gress A, Tsoy O, Kalinina OV, Baumbach J, Kacprowski T, List M. DIGGER: exploring the functional role of alternative splicing in protein interactions. **Nucleic acids research** 49, no. D1 (2021): D309-D318. https://doi.org/10.1093/nar/gkaa768

The full text and the license are available in Appendix A.

Summary of the paper

The paper describes the database DIGGER that integrates protein-protein interactions, domain-domain interactions, and residue-level interactions information to lift exon expression analysis to a network level. DIGGER includes a scoring system to account for limited evidence of multi-domain interactions, allowing for a more fine-grained consideration of the trade-off between false positive and false negative PPIs. As a user-friendly database, DIGGER allows users to seamlessly switch between isoform and exon-centric views of the interactome, making it an essential resource for studying mechanistic consequences of alternative splicing

The paper also provides several examples of how DIGGER can be used to create hypotheses or interpret experimental results concerning molecular consequences of alternative splicing. The application examples include the use of DIGGER to study the effects of exon skipping on the anaplastic lymphoma kinase (ALK) gene and the insulin receptor isoforms.

Contribution of the thesis author

Design of the method, implementation of the software, data analysis and visualization, literature review, results interpretation, and manuscript composition.

4.2 Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases

Full citation of the paper

Louadi, Z, Elkjaer ML, Klug M, Lio CT, Fenn A, Illes Z, Bongiovanni D, Baumbach J, Kacprowski T, List M, Tsoy O. Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases. **Genome Biology** 22, no. 1 (2021): 1-22. https://doi.org/10.1186/s13059-021-02538-1

The full text and the license are available in Appendix B.

Summary of the paper

This paper introduces NEASE, a tool for functional analysis of alternative splicing (AS) events. The goal of the tool is to identify pathways that are affected by AS events, which can be difficult to do using current methods that treat all interactions of the genes affected by AS equally. NEASE uses DIGGER joint graph (PPI and DDI) along with residue-level and domain-motif interactions, to identify interaction partners that are likely affected by AS. The paper presents a new gene set overrepresentation technique using an edge-level hypergeometric test, that only considers protein interactions that are likely affected by AS.

To benchmark the method, we have used multiple datasets from both healthy and disease cohorts. We show that it gives insights into the role of the muscle- and neural-specific exons, and reveals splicing-related differences between reticulated and mature platelets. We additionally demonstrate that it generates novel disease-relevant insights and provides valuable context to prior findings on altered RNA- and protein-expression levels consistent with recent literature. Examples include Dilated Cardiomyopathy and Multiple Sclerosis where NEASE highlights the impact of multiple biomarker genes.

Finally, the NEASE Python package is made available for community use with multiple functions and tutorials to help researchers deepen their analysis of AS.

Contribution of the thesis author

Design of the method, implementation of the software, data analysis and visualization, literature review, results interpretation, and manuscript composition.

Chapter 5

Discussion and Outlook

5.1 Impact and applications of the work

Despite the immense importance of AS in cellular differentiation and disease development, AS analyses are not routinely performed for RNA-Seq data. I argue that one main reason is the lack of tools to interpret the effect of splicing both at the individual event level but also systematically. In this thesis, I proposed two unique methods to address the challenge: DIGGER and NEASE. Using multiple validation steps and datasets, I demonstrated that DIGGER and NEASE confirm previous experimental results as well as provide unexplored insights on public datasets. In particular, DIGGER confirms the impact of a known event, on the anaplastic lymphoma kinase (ALK) gene, which is specific to non-small cell lung carcinoma cancer and causes a non-functional variant. NEASE, on the other hand, was compared with classical gene set enrichment in five different scenarios and has been shown to regularly outperform it. In the only scenario where NEASE results were not so different from the classic method, we have shown that NEASE shines in terms of finding novel disease candidates (biomarkers) that were not originally part of the pathway, which is a unique feature of NEASE's network-based method.

The developed tools are actively used by us and other researchers. In particular, Rodriguez-Polo *et al.* used DIGGER to characterize the exons 14 and 231 from the gene TTN that is linked to Non-Ischemic Dilated Cardiomyopathy [65]. Similarly, Liu *et al.* studied the consequence of different AS events of gene CD46 with its interaction with ITGA2 [44].

5.1.1 Application of NEASE in time series analysis

One notable application of NEASE is AS enrichment analysis for time course data, presented recently by Lio et al., where I am also a contributor [43]. Unlike the more straightforward methods that focus on a single condition or two conditions, time course data is characterized by a larger set of time points. For

example, the analysis of samples from different stages of progressing disease.

In Lio et al work, we present a framework named Spycone that first identifies isoform switches in temporal data and then determines gene clusters with similar switching patterns. The newly developed method incorporates a new metric to prioritize isoform switches, as well as multiple clustering algorithms. NEASE was later used for functional enrichment analysis of the obtained clusters from the SARS-CoV-2 infection dataset (accession ID GSE157490). The results show that the clusters were enriched in relevant pathways such as the MAPK pathway and TLR pathways [43].

5.2 Limitation and outlook

5.2.1 Proteomics approaches to further uncover AS impact

Numerous exciting technological improvements are greatly progressing the field of AS. Namely, the decrease in the cost of short-read deep sequencing and the advancements in long-read sequencing. But in my opinion, the most promising technology for AS study is proteomics [60].

DIGGER and NEASE allow a comprehensive study of AS impact. But a drawback of RNA-Seq is the fact that the AS events are not confirmed at the protein level: for instance, it is unknown if an event produces a functional protein variant or causes nonsense-mediated decay. Thus one possible future direction would be to confirm isoform switches events at the proteomic level before running NEASE enrichment and so omitting the events that yield non-coding transcripts. Taking into account that for these events, classical enrichment tests are more appropriate than the NEASE way. Accordingly, I believe that the combination of multi-omics data such as proteomics and transcriptomics will enormously boost the usefulness of DIGGER and NEASE.

5.2.2 More structural data is needed

The approach heavily relies on high-quality structural data that is currently very limited in terms of coverage such as domain and motif annotation as well as co-resolved structure. As a consequence, a large percentage of human exons are not covered in the present databases.

One possible extension is the utilization of machine learning approaches. For instance, the prediction of protein structure and interaction using deep learning is a newly emerging field with promising results. Carefully including high-confidence prediction of structure could help increase the coverage of DIGGER and NEASE and decrease the current biases in the databases toward certain genes.

5.2.3 Understanding the splicing regulation is crucial toward a system biology interpretation

NEASE and DIGGER offer a unique way to explore the impact of a given set of AS events. However, understanding how and why are these events happening in the first place is a crucial step toward complete system biology interpretation of splicing. For instance, it will help understand the directionality of these events and detect the causality and the drivers of a disease from the responses.

Future challenges include constructing a full-splicing regulatory network and extending the current gene regulatory networks to fully include regulation of splicing, polyadenylation, microRNA, etc.

5.2.4 Addressing patient specificity and heterogeneity

Transcription and splicing are often uniquely altered in diseases. Most of the current methods, including the ones in this thesis, are still focused on population-level comparison and thus ignore patient-specific dysregulation.

A new emerging topic in transcriptomics is the development of a statistical approach for patient-specific. In this direction, during my Ph.D., I also co-developed the method DysRegNet that aims to infer patient-specific regulatory alteration [42]. Briefly, DysRegNet uses a linear model to predict gene expression of the target gene from its transcription factor. The model is fitted on all control data and tested on each patient individually to detect outliers.

While DysRegNet was designed for the analysis of transcriptions. One possible and straightforward direction for future work is the extension of this method to model splicing. This can be done by correlating splicing factors with their target genes or exons and then identifying outliers splicing events. The obtained outlier events can be interpreted using NEASE and DIGGER in a system biology manner to help understand how individual patient events contribute to the phenotype or a rare disease.

Bibliography

[1] Bruce Alberts. "Molecular biology of the cell 5E". In: *Garland science* (2008), pp. 906–911.

[2] Bruce Alberts et al. "How genomes evolve". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[3] Bruce Alberts et al. "Protein function". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[4] Bruce Alberts et al. "The shape and structure of proteins". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[5] Simon Anders, Alejandro Reyes, and Wolfgang Huber. "Detecting differential usage of exons from RNA-seq data". In: *Nature Precedings* (2012), pp. 1–1.

[6] M Madan Babu. "The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease". In: *Biochemical Society Transactions* 44.5 (2016), pp. 1185–1200.

[7] Eman Badr, Mahmoud ElHefnawi, and Lenwood S Heath. "Computational identification of tissue-specific splicing regulatory elements in human genes from RNA-Seq data". In: *PLoS One* 11.11 (2016), e0166978.

[8] Jan Bednar et al. "Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin". In: *Proceedings of the National Academy of Sciences* 95.24 (1998), pp. 14173–14178.

[9] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[10] Helen M Berman et al. "The protein data bank". In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[11] Dario Bongiovanni et al. "Transcriptome analysis of reticulated platelets reveals a prothrombotic profile". In: *Thrombosis and haemostasis* 119.11 (2019), pp. 1795–1806.

[12] Shanshan YC Bradford et al. “Temperature artifacts in protein structures bias ligand-binding predictions”. In: *Chemical Science* 12.34 (2021), pp. 11275–11293.

[13] Katharine M Brauch et al. “Mutations in ribonucleic acid binding protein gene cause familial dilated cardiomyopathy”. In: *Journal of the American College of Cardiology* 54.10 (2009), pp. 930–941.

[14] Francois Chapeville et al. “On the role of soluble ribonucleic acid in coding for amino acids”. In: *Proceedings of the National Academy of Sciences* 48.6 (1962), pp. 1086–1092.

[15] Michele Clamp et al. “Distinguishing protein-coding and noncoding genes in the human genome”. In: *Proceedings of the National Academy of Sciences* 104.49 (2007), pp. 19428–19433.

[16] GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans”. In: *Science* 348.6235 (2015), pp. 648–660.

[17] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. “RNA and disease”. In: *Cell* 136.4 (2009), pp. 777–793.

[18] André Corvelo et al. “Genome-wide association between branch point properties and alternative splicing”. In: *PLoS computational biology* 6.11 (2010), e1001016.

[19] Francis H Crick. “On protein synthesis”. In: *Symp Soc Exp Biol.* Vol. 12. 138-63. 1958, p. 8.

[20] Javier De Las Rivas and Celia Fontanillo. “Protein–protein interactions essentials: key concepts to building and analyzing interactome networks”. In: *PLoS computational biology* 6.6 (2010), e1000807.

[21] Fausta Desantis et al. “Spatial organization of hydrophobic and charged residues affects protein thermal stability and binding affinity”. In: *Scientific Reports* 12.1 (2022), pp. 1–13.

[22] Marcus T Dittrich et al. “Identifying functional modules in protein–protein interaction networks: an integrated exact approach”. In: *Bioinformatics* 24.13 (2008), pp. i223–i231.

[23] Sorin Draghici et al. “A systems biology approach for pathway level analysis”. In: *Genome research* 17.10 (2007), pp. 1537–1545.

[24] Maria L Elkjaer et al. “Molecular signature of different lesion types in the brain white matter of patients with progressive multiple sclerosis”. In: *Acta neuropathologica communications* 7.1 (2019), pp. 1–17.

[25] Scott J Emrich et al. “Gene discovery and annotation using LCM-454 transcriptome sequencing”. In: *Genome research* 17.1 (2007), pp. 69–73.

[26] Antonio Fabregat et al. “The reactome pathway knowledgebase”. In: *Nucleic acids research* 46.D1 (2018), pp. D649–D655.

[27] Amit M Fenn et al. “Alternative splicing analysis benchmark with DICAST”. In: *bioRxiv* (2022).

[28] ER Gibney and CM Nolan. “Epigenetics and gene expression”. In: *Heredity* 105.1 (2010), pp. 4–13.

[29] David F Grabski et al. “Intron retention and its impact on gene expression and protein diversity: A review and a practical guide”. In: *Wiley Interdisciplinary Reviews: RNA* 12.1 (2021), e1631.

[30] Wei Guo et al. “RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing”. In: *Nature medicine* 18.5 (2012), pp. 766–773.

[31] Matthias Heinig et al. “Natural genetic variation of the cardiac transcriptome in non-diseased donors and patients with dilated cardiomyopathy”. In: *Genome biology* 18.1 (2017), pp. 1–21.

[32] Ralf Herwig et al. “Analyzing and interpreting genome data at the network level with ConsensusPathDB”. In: *Nature protocols* 11.10 (2016), pp. 1889–1907.

[33] Chia-Hsin Hsu, Chien-Kuo Chen, and Ming-Jing Hwang. “The architectural design of networks of protein domain architectures”. In: *Biology letters* 9.4 (2013), p. 20130268.

[34] Wei-Lun Hsu et al. “Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding”. In: *Protein Science* 22.3 (2013), pp. 258–273.

[35] CA Huser et al. “TSC-22D1 isoforms have opposing roles in mammary epithelial cell survival”. In: *Cell Death & Differentiation* 17.2 (2010), pp. 304–315.

[36] Wei Jiang and Liang Chen. “Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 183–195.

[37] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589.

[38] Minoru Kanehisa et al. “KEGG for linking genomes to life and the environment”. In: *Nucleic acids research* 36.suppl_1 (2007), pp. D480–D484.

[39] Yuri Kapustin et al. “Cryptic splice sites and split genes”. In: *Nucleic acids research* 39.14 (2011), pp. 5837–5844.

[40] Hyoung Kyu Kim et al. “Alternative splicing isoforms in health and disease”. In: *Pflügers Archiv-European Journal of Physiology* 470.7 (2018), pp. 995–1016.

[41] Manjeet Kumar et al. “ELM—the eukaryotic linear motif resource in 2020”. In: *Nucleic acids research* 48.D1 (2020), pp. D296–D306.

[42] Olga Lazareva et al. “DysRegNet: Patient-specific and confounder-aware dysregulated network inference”. In: *bioRxiv* (2022).

[43] Chit Tong Lio et al. “Systematic analysis of alternative splicing in time course data using Spycone”. In: *bioRxiv* (2022).

[44] Yunze Liu et al. “Identification and Validation of Novel Immune-Related Alternative Splicing Signatures as a Prognostic Model for Colon Cancer”. In: *Frontiers in oncology* 12 (2022).

[45] Núria López-Bigas et al. “Are splicing mutations the most frequent cause of hereditary disease?” In: *FEBS letters* 579.9 (2005), pp. 1900–1903.

[46] Zakaria Louadi et al. “DIGGER: exploring the functional role of alternative splicing in protein interactions”. In: *Nucleic acids research* 49.D1 (2021), pp. D309–D318.

[47] Zakaria Louadi et al. “Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases”. In: *Genome Biology* 22.1 (2021), pp. 1–22.

[48] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 1–21.

[49] Kiven E Lukong et al. “RNA-binding proteins in human genetic disease”. In: *Trends in Genetics* 24.8 (2008), pp. 416–425.

[50] Mitchell R Lunn and Ching H Wang. “Spinal muscular atrophy”. In: *The Lancet* 371.9630 (2008), pp. 2120–2133.

[51] Laurent Maveyraud and Lionel Mourey. “Protein X-ray crystallography and drug discovery”. In: *Molecules* 25.5 (2020), p. 1030.

[52] Jaina Mistry et al. “Pfam: The protein families database in 2021”. In: *Nucleic acids research* 49.D1 (2021), pp. D412–D419.

[53] Roberto Mosca et al. “3did: a catalog of domain-based interactions of known three-dimensional structure”. In: *Nucleic acids research* 42.D1 (2014), pp. D374–D379.

[54] Rose Oughtred et al. “The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions”. In: *Protein Science* 30.1 (2021), pp. 187–200.

[55] James R Perkins et al. “Transient protein-protein interactions: structural, functional, and network properties”. In: *Structure* 18.10 (2010), pp. 1233–1243.

[56] Lolita Piersimoni et al. “Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein Interactions A Method for All Seasons”. In: *Chemical Reviews* 122.8 (2021), pp. 7500–7531.

[57] Rocky Pramanik et al. “p38 isoforms have opposite effects on AP-1-dependent transcription through regulation of c-Jun: the determinant role of the isoforms in the p38 MAPK signal specificity”. In: *Journal of Biological Chemistry* 278.7 (2003), pp. 4831–4839.

[58] Ibraheem Rehman, Mustafa Farooq, and Salome Botelho. "Biochemistry, secondary protein structure". In: *StatPearls [Internet]*. StatPearls Publishing, 2021.

[59] Jüri Reimand et al. "Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap". In: *Nature protocols* 14.2 (2019), pp. 482–517.

[60] Marina Reixachs-Solé and Eduardo Eyras. "Uncovering the impacts of alternative splicing on the proteome with current omics techniques". In: *Wiley Interdisciplinary Reviews: RNA* (2022), e1707.

[61] Siyuan Ren et al. "Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions". In: *BMC genomics* 9.2 (2008), pp. 1–11.

[62] Alejandro Reyes et al. "Drift and conservation of differential exon usage across tissues in primate species". In: *Proceedings of the National Academy of Sciences* 110.38 (2013), pp. 15377–15382.

[63] Isabelle Rivals et al. "Enrichment or depletion of a GO category within a class of genes: which test?" In: *Bioinformatics* 23.4 (2007), pp. 401–407.

[64] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *bioinformatics* 26.1 (2010), pp. 139–140.

[65] Ignacio Rodriguez-Polo and Rüdiger Behr. "Exploring the Potential of Symmetric Exon Deletion to Treat Non-Ischemic Dilated Cardiomyopathy by Removing Frameshift Mutations in TTN". In: *Genes* 13.6 (2022), p. 1093.

[66] Caroline Seydel. "Diving deeper into the proteome". In: *Nature Methods* 19.9 (2022), pp. 1036–1040.

[67] Shihao Shen et al. "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data". In: *Proceedings of the National Academy of Sciences* 111.51 (2014), E5593–E5601.

[68] Mirko Signorelli, Veronica Vinciotti, and Ernst C Wit. "NEAT: an efficient network enrichment analysis test". In: *BMC bioinformatics* 17.1 (2016), pp. 1–17.

[69] Damian Smedley et al. "BioMart–biological queries made easy". In: *BMC genomics* 10.1 (2009), pp. 1–12.

[70] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". In: *Nature Reviews Genetics* 20.11 (2019), pp. 631–656.

[71] Timothy Sterne-Weiler et al. "Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop". In: *Molecular cell* 72.1 (2018), pp. 187–200.

[72] Damian Szklarczyk et al. “The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets”. In: *Nucleic acids research* 49.D1 (2021), pp. D605–D612.

[73] David Talavera, David L Robertson, and Simon C Lovell. “Alternative splicing and protein interaction data sets”. In: *Nature biotechnology* 31.4 (2013), pp. 292–293.

[74] 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 research* 27.10 (2017), pp. 1759–1768.

[75] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. “Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. In: *Contemporary Oncology/Współczesna Onkologia* 2015.1 (2015), pp. 68–77.

[76] Léon-Charles Tranchevent et al. “Identification of protein features encoded by alternative exons using Exon Ontology”. In: *Genome research* 27.6 (2017), pp. 1087–1097.

[77] Michael L Tress, Federico Abascal, and Alfonso Valencia. “Alternative splicing may not be the key to proteome complexity”. In: *Trends in biochemical sciences* 42.2 (2017), pp. 98–110.

[78] Jorge Vaquero-Garcia et al. “A new view of transcriptome complexity and regulation through the lens of local splicing variations”. In: *elife* 5 (2016), e11752.

[79] Mihaly Varadi et al. “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models”. In: *Nucleic acids research* 50.D1 (2022), pp. D439–D444.

[80] J Craig Venter. In: *The sequence of the human genome. Science* 291 (2001), pp. 1304–1351.

[81] Kristoffer Vitting-Seerup and Albin Sandelin. “The Landscape of Isoform Switches in Human CancersIsoform Switches in Cancer”. In: *Molecular Cancer Research* 15.9 (2017), pp. 1206–1220.

[82] Jianbo Wang et al. “Computational methods and correlation of exon-skipping events with splicing, transcription, and epigenetic factors”. In: *Cancer Gene Networks*. Springer, 2017, pp. 163–170.

[83] Yan Wang et al. “Mechanism of alternative splicing and its regulation”. In: *Biomedical reports* 3.2 (2015), pp. 152–158.

[84] Zefeng Wang et al. “Systematic identification and analysis of exonic splicing silencers”. In: *Cell* 119.6 (2004), pp. 831–845.

[85] Andrew Waterhouse et al. “SWISS-MODEL: homology modelling of protein structures and complexes”. In: *Nucleic acids research* 46.W1 (2018), W296–W303.

[86] Jennifer Westoby et al. “Simulation-based benchmarking of isoform quantification in single-cell RNA-seq”. In: *Genome biology* 19.1 (2018), pp. 1–14.

[87] Thorsten Will and Volkhard Helms. “PPIXpress: construction of condition-specific protein interaction networks based on transcript expression”. In: *Bioinformatics* 32.4 (2016), pp. 571–578.

[88] Xinping Yang et al. “Widespread expansion of protein interaction capabilities by alternative splicing”. In: *Cell* 164.4 (2016), pp. 805–817.

[89] Zhenyu Yang et al. “AlphaFold2 and its applications in the fields of biology and medicine”. In: *Signal Transduction and Targeted Therapy* 8.1 (2023), p. 115.

[90] Sailu Yellaboina et al. “DOMINE: a comprehensive collection of known and predicted domain-domain interactions”. In: *Nucleic acids research* 39.suppl_1 (2011), pp. D730–D735.

[91] Sika Zheng. “Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression”. In: *International Journal of Developmental Neuroscience* 55 (2016), pp. 102–108.

[92] Mi Zhou, Qing Li, and Renxiao Wang. “Current experimental methods for characterizing protein–protein interactions”. In: *ChemMedChem* 11.8 (2016), pp. 738–756.

Appendix A

Appendix: First publication

This is the copyedited PDF of the article originally published in the **Nucleic Acids Research** journal.

Citation: Louadi, Zakaria, Kevin Yuan, Alexander Gress, Olga Tsoy, Olga V. Kalinina, Jan Baumbach, Tim Kacprowski, and Markus List. "DIGGER: exploring the functional role of alternative splicing in protein interactions." Nucleic acids research 49, no. D1 (2021): D309-D318. https://doi.org/10.1093/nar/gkaa768.

Published online 25 September 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D309–D318***
doi: 10.1093/nar/gkaa768

DIGGER: exploring the functional role of alternative splicing in protein interactions

Zakaria Louadi[1], Kevin Yuan[1], Alexander Gress[2], Olga Tsoy[1], Olga V. Kalinina[2,3], Jan Baumbach[1,4], Tim Kacprowski[1,*,†] and Markus List[1,*,†]

[1]Chair of Experimental Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany, [2]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), 66123 Saarbrücken, Germany, [3]Faculty of Medicine, Saarland University, 66421 Homburg, Germany and [4]Department of Mathematics and Computer Science, University of Southern Denmark, 5230 Odense M, Denmark

Received July 17, 2020; Revised September 01, 2020; Editorial Decision September 02, 2020; Accepted September 04, 2020

ABSTRACT

Alternative splicing plays a major role in regulating the functional repertoire of the proteome. However, isoform-specific effects to protein-protein interactions (PPIs) are usually overlooked, making it impossible to judge the functional role of individual exons on a systems biology level. We overcome this barrier by integrating protein-protein interactions, domain-domain interactions and residue-level interactions information to lift exon expression analysis to a network level. Our user-friendly database DIGGER is available at https://exbio.wzw.tum.de/digger and allows users to seamlessly switch between isoform and exon-centric views of the interactome and to extract sub-networks of relevant isoforms, making it an essential resource for studying mechanistic consequences of alternative splicing.

INTRODUCTION

Alternative splicing (AS) refers to differences in the processing of transcripts (e.g. exon skipping, intron retention etc.) allowing to synthesize different protein variants from the same gene. These protein variants, called isoforms, can vary in their functionality or even have opposite roles (1). This mechanism is important in cell development and differentiation (2) but also in diseases such as cancer (3), heart and kidney diseases (4,5).

Protein-protein interaction (PPI) networks such as BioGrid (6) or STRING (7) are an important resource in systems biology. PPI interactions are identified in tedious experiments, mostly via affinity purification mass spectrometry or yeast two hybrid screens (8). Due to the high number of possible interactions (quadratic in the number of considered proteins), efforts are limited to testing only major protein isoforms, hence neglecting the considerable influence of AS on the interactome. For instance, it was shown that AS remodels the network of PPIs in a tissue-specific manner (9) and that protein variants from the same gene differ in their interactions due to changes in the structural domain composition (1,10). Yang *et al.* found that most isoforms share <50% of interactions and only 21% of isoforms pairs have identical interaction profiles (1). Furthermore, a high proportion of these isoforms are known to be expressed in a tissue-specific manner (11). Recently, Climenté-Gonzalez *et al.* showed that around 30% of all isoform switches in tumor cells affect domains that mediate protein interaction (12). This suggests a widespread impact of AS in the human interactome that is currently neglected (13).

Domain-domain interaction (DDI) databases provide an annotation of PPIs in a structural context. This structurally resolved interactome is frequently used to analyze the location of disease mutations in proteins (14). 3did visualizes DDIs as a graph but does not integrate this information with experimentally validated PPIs (15). In contrast, Interactome3D and INstruct add structural details such as DDIs and residues to the PPI networks but do not project this information to the level of isoforms or exons (16–17). Given the resolved structural composition of different isoforms, this annotation can be extended to predict isoform-specific interactions consistent with experimental results (1,18). It is further possible to identify residues located at the interface of a PPI to study PPI perturbation (19). However, existing efforts are mostly focused on studying mutations that perturb these interactions (20–21) but do not consider consequences of AS. Few existing tools address this gap to systematically study AS. The Cytoscape app DomainGraph (22) visualizes domain interactions simultaneously with protein interactions and analyzes the effect of differ-

*To whom correspondence should be addressed. Tel: +49 8161 71 2761; Email: markus.list@wzw.tum.de
Correspondence may also be addressed to Tim Kacprowski. Tel: +49 8161 71 2710; Email: tim.kacprowski@wzw.tum.de
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

ential exon usage. However, DomainGraph is limited to the output of the tool AltAnalyze (22). Ghadie *et al.* developed DIIP using a similar method to predict an isoform interactome (18). While their results were verified based on the experimentally validated isoform interactome reported by Yang *et al.* (1), their database covers only a fraction of the proteome with 2944 reference proteins and 4363 interactions. Exon Ontology (EXONT) characterizes protein domains and features that are affected by AS (23) but does not consider AS on the network level.

PPIXpress extends this idea to construct a condition-specific PPI network based on transcript expression (24). While covering the entire proteome, it was not intended for studying individual genes or protein variants. Neither DIIP nor PPIXpress provide a graphical visualization or support the analysis of a single splicing event such as the gain or loss of a domain. Furthermore, existing tools do not allow a side-by-side comparison of interactions of different protein isoforms, which, however, is crucial to understand the functional effect of an isoform switch between two conditions. To close this gap, we developed DIGGER (Domain Interaction Graph Guided ExploreR), a user-friendly database and web tool to explore the functional impact of AS on human PPIs. In contrast to existing tools (Supplementary Table S1), DIGGER includes residue-specific information, highlights consequences of exon skipping events, visualizes interactions between multiple isoforms and offers a user-friendly web interface.

MATERIALS AND METHOD

Joint PPI and DDI network

The human PPI network with 24 969 reference proteins and 410 961 interactions was obtained from BioGRID version Homo_sapiens-3.5 (6) and DDIs were downloaded from 3did (v2019_01) and DOMINE (v2.0) (15,25). 16,094 low-confidence interactions from DOMINE were removed. The remaining 2989 high- and 2537 mid-confidence interactions were integrated with all 13 499 reported interactions in 3did to obtain 17 349 interactions between 8190 Pfam domains (26). We implemented a joint network graph (Figure 1) that integrates PPIs and DDIs and in which nodes represent protein domains defined by concatenating Entrez and Pfam id. The edges between the nodes represent DDIs which are defined if the domains are known to interact and if the respective proteins are also PPI partners. The joint graph greatly speeds up the real-time processing of the requested data and can also be useful for studying the interacting regions of the proteins in other studies. Hence, we make the joint graph available as download in multiple formats on the DIGGER website.

Position-specific PPI network construction

We constructed a PPI network of the human proteome based on experimentally resolved structures in the Protein Data Bank (PDB) (27). First, we mapped individual amino acid positions to individual residues in experimentally resolved protein structures. To this end, we aligned the sequences of all protein isoforms in the human proteome to all protein chains with >95% sequence identity in the PDB. The second step was the identification of all interaction partners of a particular amino acid residue, which we defined as all amino acid residues from other protein chains co-resolved in the same three-dimensional structure and <5Å from the residue of interest (Figure 2). In total, we could identify 8991 DDIs, 3230 of which are also covered by BioGRID. Since a protein can be mapped to multiple structures, a single amino acid can be involved in multiple interactions with residues belonging to different interaction partner proteins. For proteins that have been experimentally resolved in complex with other human proteins, we can thus map every residue on the PPI interface to a particular position in the genome, and hence to a particular exon. Additionally, we obtain the same information for the interacting protein(s), creating a position-specific picture of the PPI interface.

Mapping of protein domains to exons

The exons and the domain composition of annotated proteins were obtained from Ensembl 99 using the Biomart webtool (28). We generated database tables for both genomic data, e.g. genes with their corresponding transcript and exon coordinates, and for proteins, e.g. isoforms and their domains. We converted the protein coordinates to genomic coordinates in the coding sequence and merged both tables to be able to map transcripts with their corresponding exons to the corresponding protein isoforms and Pfam domains. We further constructed a database table that maps position-specific residue annotations to exons to obtain an exon-level PPI network. The Biomart mapping table was also used to convert between Entrez, Ensembl and Uniprot ids. All data tables are available as downloads on our website.

RNA-Seq dataset analysis

The transcript expressions using RNAseq were obtained from the Cancer Genome Atlas pan-cancer dataset downloaded via the Xena Browser (29) (https://xenabrowser.net/datapages/) for the sample identifier TCGA-S9-A7J2-01. The isoform expression levels are originally estimated based on RSEM (30). In this analysis, all transcripts with an expression value above 1.0 are considered as abundant. The source code for the analysis is available at (https://github.com/louadi/RNA-Seq-DIGGER).

Web interface

DIGGER was developed using the Python web framework Django and is released as open source under the GPLv3 license (https://github.com/louadi/DIGGER). For visualization, we used the Javascript library vis.js (https://visjs.org/) with different graph layout parameters depending on the size of the generated network.

DATABASE CONTENT AND APPLICATIONS

DIGGER integrates the interactome from BioGRID (6) with DDIs of Pfam domains reported by DOMINE and 3did (15,25), comprising 9370 reference proteins and 52 083

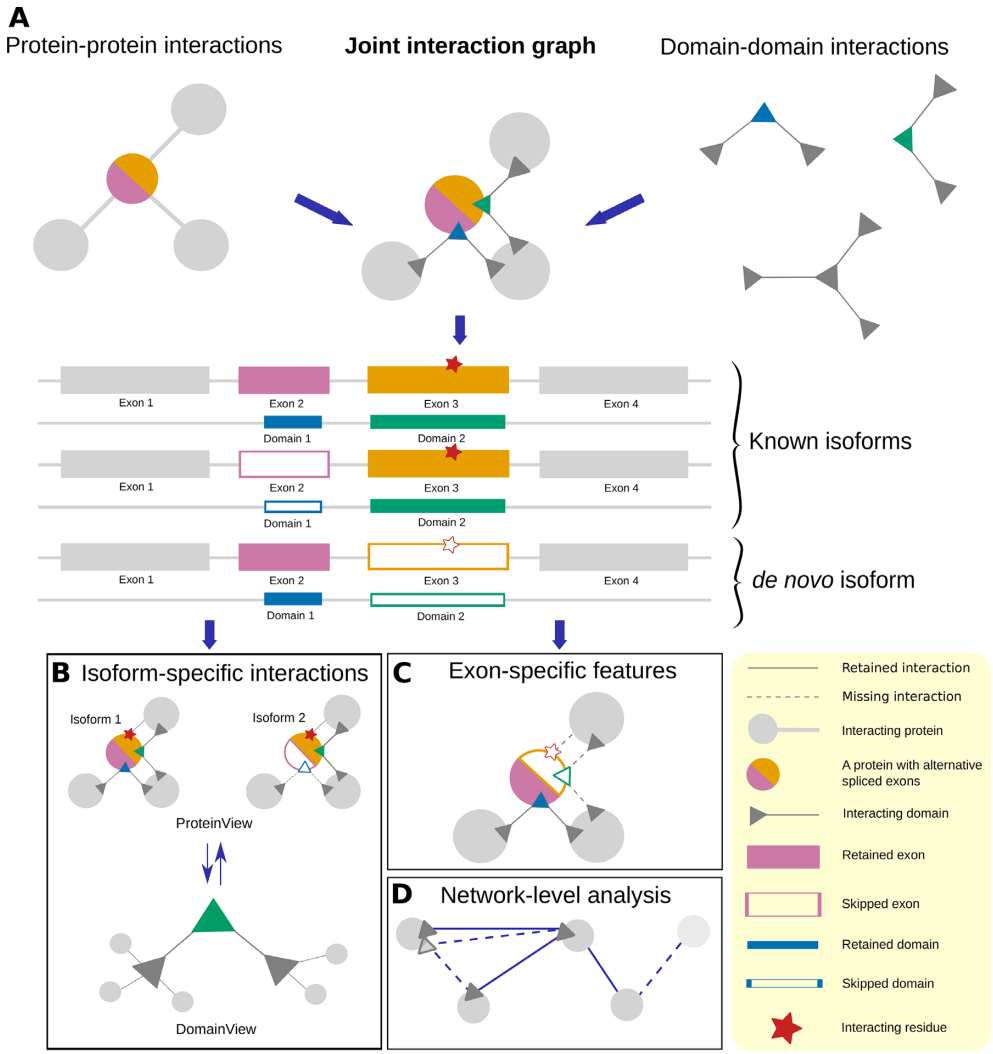


Figure 1. (**A**) Protein–protein interaction data is integrated with a domain–domain interaction data to construct structurally annotated interactions for every gene. This annotation is then used to compare between different protein variants in isoform-level analysis mode (**B**) and to identify the functional effect of a skipped exon in exon-level analysis mode (**C**). In the latter, residues located at the corresponding interaction interfaces are highlighted. Network-level analysis (**D**): DIGGER generates a subnetwork from a list of protein isoforms (see network-level analysis for details).

PPIs that are confirmed by at least one DDI and 17 390 PPIs mediated by multiple DDIs. Notably, none of the existing resources annotate individual exons, which we consider a prerequisite to study the consequence of AS on DDIs. To mitigate this, DIGGER provides a unique mapping of interface residues of interacting proteins to exons based on experimentally resolved structures in the Protein Data Bank (PDB) (27). We generated a PPI network resolving interactions on a residue-specific level, i.e. for each protein residue on an interaction interface, we derived information on all residues from the interacting protein that is in contact with it (see Materials and Methods for details). In this way, genomic information on a splicing event can be directly mapped onto protein three-dimensional structure and the impact of the AS event on the PPI interface can be assessed. Through DIGGER's user-friendly web interface, researchers can interactively visualize the domain composition for any protein isoform, with detailed information of the interacting domains between the selected protein and of its partners in the PPI network.

DIGGER offers three different modes (Figure 1) that can be used interchangeably. Here, we explain these modes individually and provide several use cases.

Isoform-level analysis

In this mode, users can query a protein isoform and visualize its composition including the exons and their corresponding domains as well as residues predicted to be part of the interface to interacting proteins. Interactions specific to the selected isoform are displayed as an interactive graph where users can toggle between the ProteinView and the DomainView to visualize interacting proteins or domains, respectively. Importantly, the ProteinView will highlight missing domains, i.e. domains that are not annotated in Pfam for a given isoform.

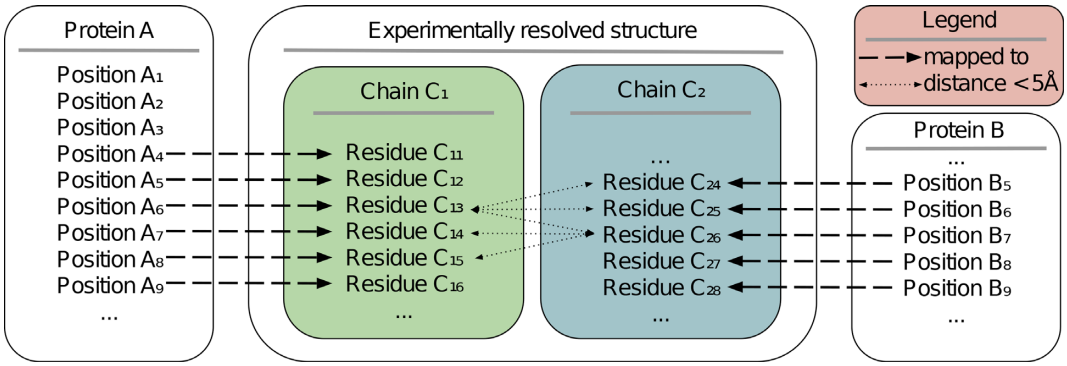


Figure 2. Schematic representation of the construction of the position-specific protein–protein interaction network. Identification of an interaction between proteins A and B based on their mappings to two different chains C1 and C2 in the same experimentally resolved structure. For example, the amino acid at position A_6 of protein A is defined to interact with the amino acids at positions B_5, B_6, and B_7 of protein B and vice versa.

Protein domains are often shared between different isoforms. The DomainView (Figure 1B) highlights domain-domain interactions together with potential protein interaction partners that utilize this domain and can be considered as a domain-specific interactome independent of the associated protein. This view is not only useful to study spliced domains but can also be extended for other applications such as studying coding disease variants affecting a protein domain or analysing specific drugs targeting a domain unit.

DIGGER scores multi-domain interactions to account for limited evidence. In contrast to existing methods that only consider a PPI missing if all its supporting DDIs are missing, DIGGER provides a score representing the percentage of missing domains for every interaction in a PPI. This allows for more fine-grained considerations and hence better control of the tradeoff between false positive and false negative PPIs. As an example for the usefulness of this feature, we consider a data set of 19 genes with 46 experimentally verified isoform-specific interactions (1). DIGGER could confirm 36 out of 46 experimentally verified splicing events that disrupt interacting domains (Supplementary Table S2). In 10 non-identified cases, no high quality structural annotated interactions were reported for the spliced domains. In one case, we observe that an isoform of CDK5 with a duplicated kinase domain interacts with the protein CCND2, while in another variant with only a single kinase domain, this interaction is missing.

Figure 3 illustrates two examples from the subset where isoforms of the genes BAG1 and NCK2 are shown to lose PPIs with their partners due to alternative domain usage. The first example (Figure 3A and B) shows that the interaction between the proteins BAG1 and HSPA8 is mediated by only one of the two domains of BAG1 (the BAG domain PF02179). This interaction is also confirmed by residue-level information. In contrast, we observe that for the interaction between NCK2 and ABI1 (Figure 3C and D), two domains of NCK2 participate in the interaction (SH2 domain PF00017 and SH3 domain PF00018), but the loss of the SH2 domain interaction disrupts the PPI.

This observation highlights a limitation of the current practice where an interaction is only considered as missing if all domain-domain interactions are missing (18,24). The exact domain(s) that mediate a PPI can not be precisely identified when multiple domains interact between two proteins as in the above example where the loss of domain SH2 alone is sufficient to disrupt the interaction (1). In total, we found 25 isoform-specific interactions that are mediated by multiple domains reported in (1), which motivates DIGGER's approach of scoring interactions rather than filtering them following an all or nothing strategy. The scores are available in a downloadable table in InteractionView.

Exon-level analysis

We propose that an exon-level view on PPIs and DDIs is best suited to recapitulate the effect of AS on the interactome. Thus, DIGGER maps domains and interface residues for all protein variants of a single gene to genomic coordinates and corresponding exons (Figure 1A–C, see Materials and Methods for details). In contrast to isoform-level analysis, the exon-level analysis mode allows the user to identify any domains encoded by the exon of interest and to visualize the interaction mediated by them. This method allows investigating the consequences of a putative or observed exon loss. For a better comparison, we also linked this feature with the isoform-level analysis by listing protein variants that contain the exon of interest. The user can, in a similar way to the previous modes, visualize all interactions of every partner individually, where the percentage of missing putative interactions is shown as a percentage score. Here, a missing domain is defined as any domain with a sequence overlap with the selected exon. To compare different possible scenarios resulting from different isoforms, the user can also visualize every DDI individually using the DomainView (Figure 1B).

DIGGER is unique in that it goes beyond domain-level annotation of PPIs to exon-level structural evidence of an interaction. An exon is considered to have structural evidence for a PPI if it codes for residues that are found within a distance of <5 Å in a co-resolved structure of the two proteins (see Materials and Methods for details). To run this mode, the user can input an exon Ensembl ID or a gene ID followed by the coordinates of the exon in hg38, which is similar to the output produced by most AS event detection tools. Another option to access this mode is from isoform-level analysis mode, by selecting a protein and then choos-

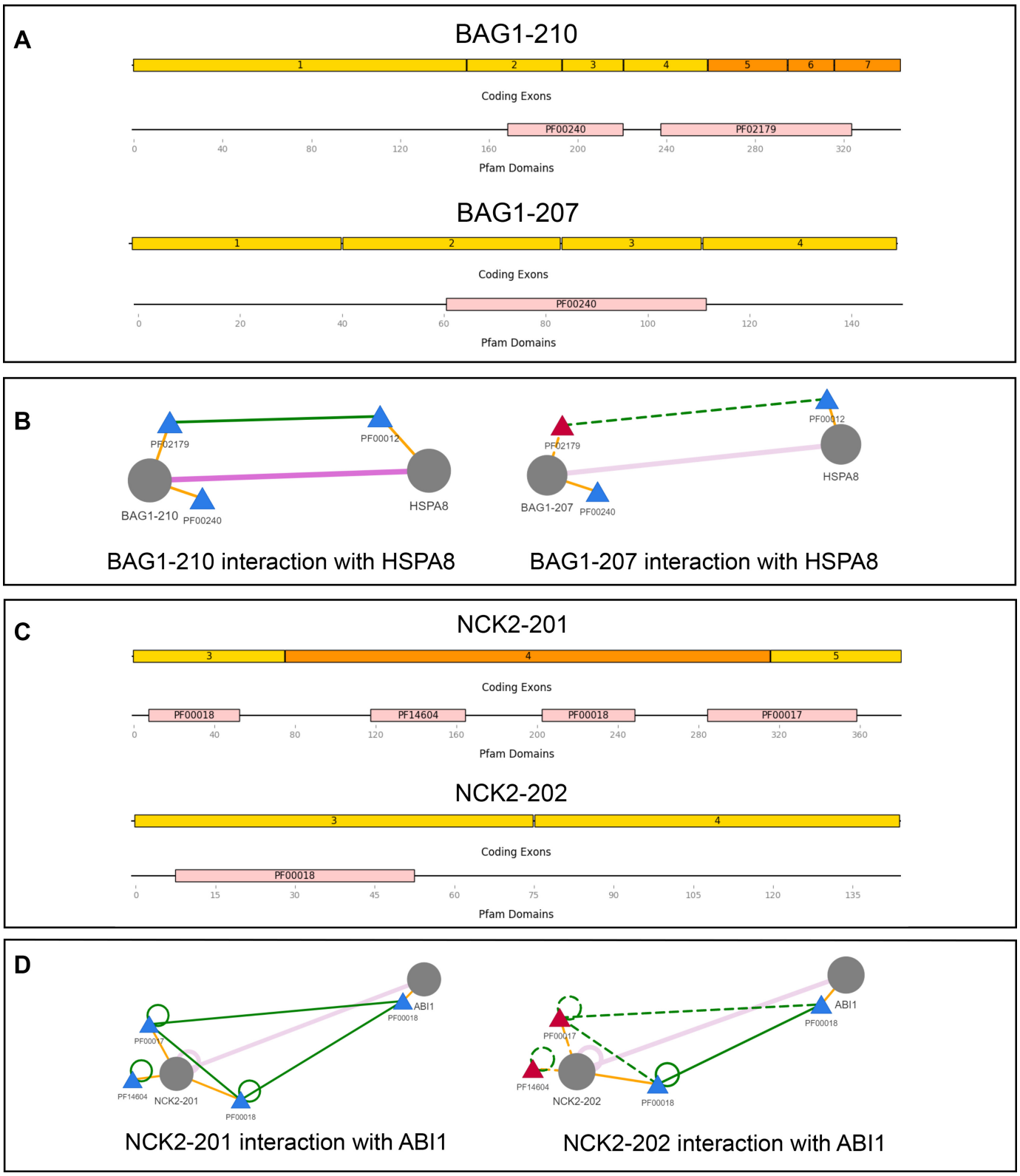


Figure 3. InteractionView example for comparing different isoforms. Circles represent proteins and triangles represent domains. A and B represent a comparison between the exon and domain structure of two isoforms of the gene BAG1 and their interaction with the protein HSPA8. (**A**) Isoform BAG1–207 lacks the domain PF002179 while for BAG1–210 both domains are preserved. (**B**) The effect of losing domain PF002179 is highlighted in the network by red triangle nodes and dashed edges. Notably, PF002179 is the only domain mediating the interaction with a domain of the protein HSPA8 suggesting that this interaction is missing for isoform BAG1–207. In the second example (**C**), two domains mediate the interaction between two proteins NCK2 and ABI2. (**D**) As one of them is spliced out for isoform NCK2–202, the interaction is scored 0.5 for this isoform and 1.0 for NCK2–201. Missing exons such as exon number 5–7 in BAG1–210 (**A**) are shown in orange if residues are predicted to be on the interface. In this case, the interface with HSPA8 is mapped to the exon 5 and also supported by residue-level evidence.

ing a specific exon from the exon or domain structure view (Figure 4).

Use Case 1: Truncated isoforms of anaplastic lymphoma kinase lose 97% of their PPIs due to AS. We demonstrate how explorative analysis in DIGGER can be used to create hypotheses or to interpret experimental results with respect to molecular consequences of alternative splicing. As a case study, we consider experimentally verified splicing variants of the tyrosine kinase receptor family.

In non-small cell lung carcinoma, Lobo de Figueiredo-Pontes *et al.* reported non-functional isoforms of anaplastic lymphoma kinase (ALK) that lack a functional kinase domain due to skipping of exons 23 and 27 (31). Figueiredo-Pontes *et al.* found that these isoforms are still able to fuse with EML4 but because of the lack of the kinase domain, the dimer EML4-ALK was unable to phosphorylate tyrosine sites. We can assess the consequence of skipping these exons using DIGGER's visualization, where we observe that no annotated isoform lacks exon 23 or 27 in our database or in the Ensembl transcript database. In ALK-201, the main isoform of the ALK gene, exons 21 to 28 encode for the domain tyrosine kinase (PF07714 in Figure 4A). By choosing the exon page (see exon-level analysis) of any of the two exons, we can contemplate the effect of losing one of these exons on ALK PPIs. Strikingly, the deletion of either exon 23 or 27 affects 31 of the 33 known structurally annotated interactions of the ALK gene (Figure 4B). This corroborates the experimental results showing that skipping of these exons leads to a translated but non-functional variant which likely lost 97% of its PPIs.

In another interesting example, Ellis *et al.* (9) found that the ability of gene GRB2 to self-interact was lost by deletion of a tissue-specific exon that overlaps with SH2 domain (PF00017) while the interaction with RAPGEF1 was retained. DIGGER could confirm that the self-interaction is mediated by the SH2 domain while the interaction with RAPGEF1 is mediated by SH3 domain (PF00018) and thus not affected in the isoform missing this exon.

Use case 2: AS leads to different insulin response. Denley *et al.* investigated two isoforms of the insulin receptor gene that respond differently to insulin (32,33), namely INSR-201 (ENST00000302850) and INSR-202 (ENST00000341500), which differ by the absence of exon 11 from the isoform INSR-202. Since the amino acids encoded in the skipped exon 11 (ENSE00001157509) are not a part of any annotated domain, we explored the existence of any known protein motifs. We found that the exon encodes the PKA phosphorylation site (MOD_PKA_1) according to the Eukaryotic Linear Motif resource (34), suggesting that a post-translational modification may be affected with possible consequences for protein signalling. Interestingly, DIGGER's exon-level analysis shows that this exon also contains residues that interact with four insulin isoforms (Figure 5). Consequently, the exon-level analysis results suggest that this interaction will be affected by skipping this exon. This observation shows the importance of residue-level interaction data as a complement to domain interactions and confirms the utility of this new feature that could be used to generate hypotheses of possible scenarios resulting from exon skipping.

Network-level analysis

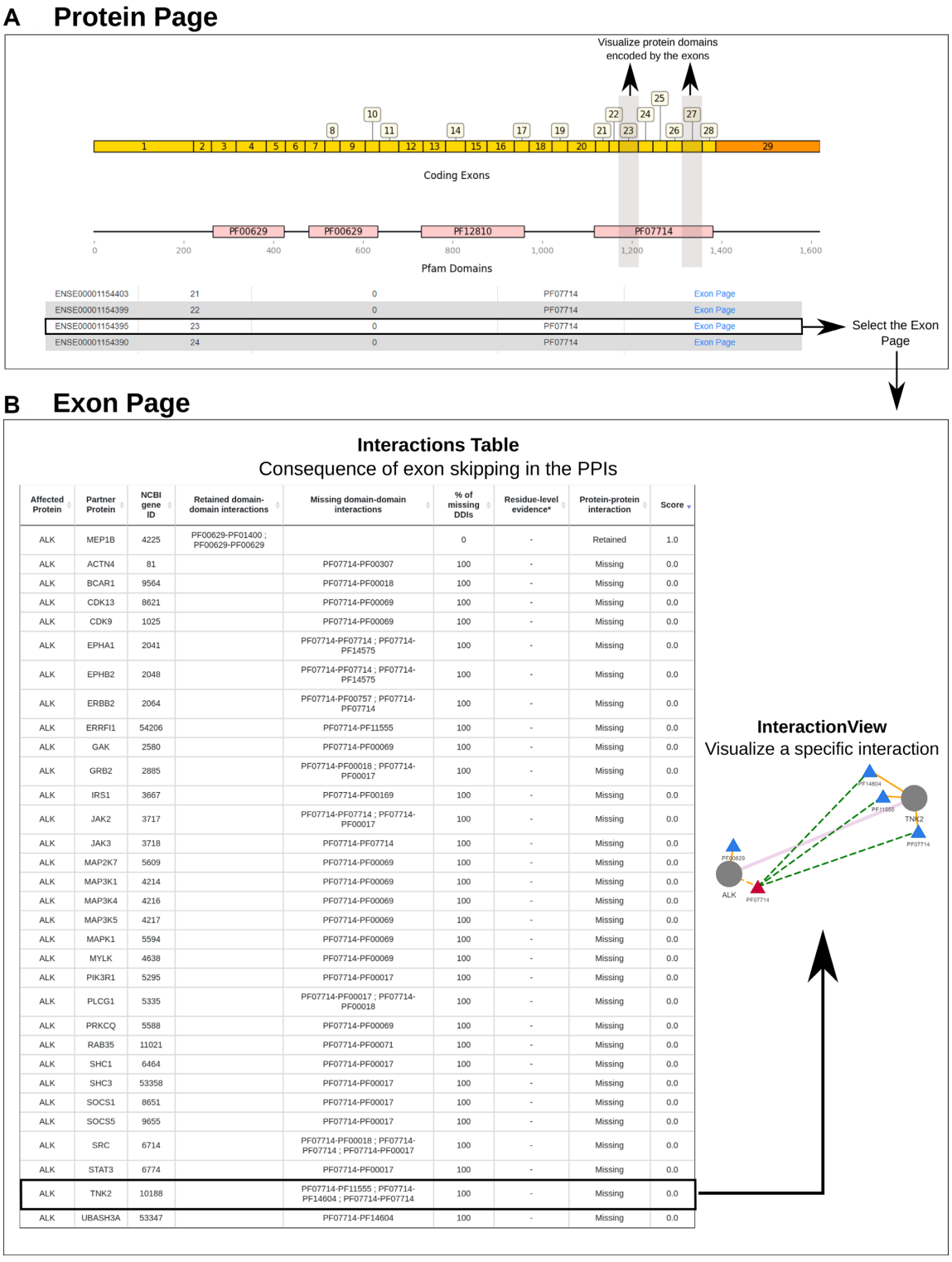
To study the effect of AS on PPIs and DDIs on a larger scale in systems and network biology, it is crucial to consider interactions between multiple protein isoforms or domains in a comprehensive view. Typical examples are the in-depth analysis of AS-driven interaction changes in a protein complex or a list of differentially expressed (or spliced) genes or proteins from transcriptomic or proteomic experiments. PPIXpress (24), the only other tool that constructs a subnetwork based on a list of transcripts, does not offer visualization of the network, affected edges or interacting domains. In contrast, DIGGER visualizes interactions between multiple proteins or isoforms. Users can input a list of gene, transcript or protein Ensembl identifiers to construct a subnetwork.

As illustrated in Figure 6, DIGGER generates a subnetwork of interactions showing domains putatively mediating these interactions. The interaction is labeled 'PPI', if there is no structural evidence for it. Otherwise, it is labeled 'PPI-DDI' and the specific DDIs are shown by one or multiple edges. Analogous to the isoform-level analysis, we provide a score for each interaction based on the fraction of annotated DDIs that are present. When the resulting network is exported, the score provides an edge weight for subsequent analysis.

Applying network-level analysis to the RNA-seq data from The Cancer Genome Atlas pan-cancer dataset (35), we could identify 41 449 edges with one or more DDIs, of which 3258 show at least one missing domain and 2,088 putative interactions that are likely completely missing. The details and code for this analysis can be found in the Materials and Methods section. These results corroborate the need for transcript and isoform-level network analysis to better reflect the proteome in disease-relevant conditions such as cancer.

DISCUSSION

With DIGGER, we present a versatile, user-friendly database and web tool to study the impact of AS on PPIs. DIGGER integrates PPI and DDI interactions into a joint graph and, as a key innovation, maps interacting residues to exons, allowing us to better assess the functional consequence of AS. Our analysis based on isoform-, domain- and exon-specific views of the human interactome shows a widespread effect of AS in concordance with experimental data. DIGGER is the first tool to score isoform-specific interactions based on the ratio of missing DDIs which facilitates the interpretation of interactions involving multiple domains. To facilitate systems and network biology analyses, DIGGER constructs a subnetwork of the joint PPI and DDI graph based on a list of isoforms or protein variants. Using this network-level analysis mode, it is possible to visualize affected DDIs. The resulting network can be exported for further analysis, e.g. for comparing different conditions. However, it is important to bear in mind that some interacting isoforms in the subnetwork are not



Affected Protein	Partner Protein	NCBI gene ID	Retained domain-domain interactions	Missing domain-domain interactions	% of missing DDIs	Residue-level evidence*	Protein-protein interaction	Score
ALK	MEP1B	4225	PF00629-PF01400 ; PF00629-PF00629		0	-	Retained	1.0
ALK	ACTN4	81		PF07714-PF00307	100	-	Missing	0.0
ALK	BCAR1	9564		PF07714-PF00018	100	-	Missing	0.0
ALK	CDK13	8621		PF07714-PF00069	100	-	Missing	0.0
ALK	CDK9	1025		PF07714-PF00069	100	-	Missing	0.0
ALK	EPHA1	2041		PF07714-PF07714 ; PF07714-PF14575	100	-	Missing	0.0
ALK	EPHB2	2048		PF07714-PF07714 ; PF07714-PF14575	100	-	Missing	0.0
ALK	ERBB2	2064		PF07714-PF00757 ; PF07714-PF07714	100	-	Missing	0.0
ALK	ERRFI1	54206		PF07714-PF11555	100	-	Missing	0.0
ALK	GAK	2580		PF07714-PF00069	100	-	Missing	0.0
ALK	GRB2	2885		PF07714-PF00018 ; PF07714-PF00017	100	-	Missing	0.0
ALK	IRS1	3667		PF07714-PF00169	100	-	Missing	0.0
ALK	JAK2	3717		PF07714-PF07714 ; PF07714-PF00017	100	-	Missing	0.0
ALK	JAK3	3718		PF07714-PF07714	100	-	Missing	0.0
ALK	MAP2K7	5609		PF07714-PF00069	100	-	Missing	0.0
ALK	MAP3K1	4214		PF07714-PF00069	100	-	Missing	0.0
ALK	MAP3K4	4216		PF07714-PF00069	100	-	Missing	0.0
ALK	MAP3K5	4217		PF07714-PF00069	100	-	Missing	0.0
ALK	MAPK1	5594		PF07714-PF00069	100	-	Missing	0.0
ALK	MYLK	4638		PF07714-PF00069	100	-	Missing	0.0
ALK	PIK3R1	5295		PF07714-PF00017	100	-	Missing	0.0
ALK	PLCG1	5335		PF07714-PF00017 ; PF07714-PF00018	100	-	Missing	0.0
ALK	PRKCQ	5588		PF07714-PF00069	100	-	Missing	0.0
ALK	RAB35	11021		PF07714-PF00071	100	-	Missing	0.0
ALK	SHC1	6464		PF07714-PF00017	100	-	Missing	0.0
ALK	SHC3	53358		PF07714-PF00017	100	-	Missing	0.0
ALK	SOCS1	8651		PF07714-PF00017	100	-	Missing	0.0
ALK	SOCS5	9655		PF07714-PF00017	100	-	Missing	0.0
ALK	SRC	6714		PF07714-PF00018 ; PF07714-PF07714 ; PF07714-PF00017	100	-	Missing	0.0
ALK	STAT3	6774		PF07714-PF00017	100	-	Missing	0.0
ALK	TNK2	10188		PF07714-PF11555 ; PF07714-PF14604 ; PF07714-PF07714	100	-	Missing	0.0
ALK	UBASH3A	53347		PF07714-PF14604	100	-	Missing	0.0

Figure 4. DIGGER can be used to study the putative effect of an exon skipping event. (**A**) in this use case, we consider an event resulting in a non-annotated protein. (**B**) We continue with exon-level analysis to show affected domains and interactions of the resulting protein. The dashed edges represent the interactions of a spliced domain that is encoded by the selected exon.

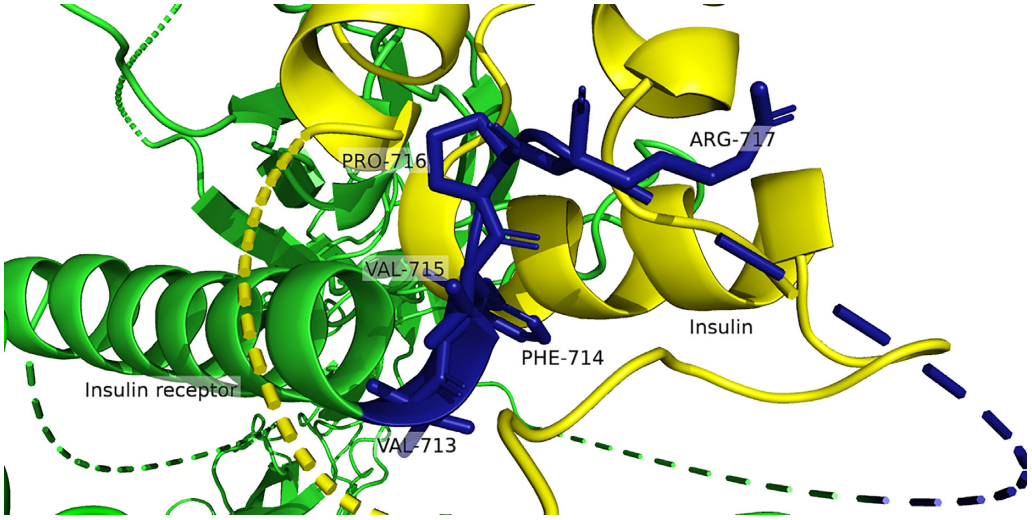


Figure 5. Three-dimensional structure of a complex with insulin (yellow) and the insulin receptor (green) (PDB id 6PXV, not resolved parts are drawn as dotted lines). The residues that are part of exon ENSE00001157509 (blue) are forming an interaction interface between the insulin receptor and insulin. Although only five residues (chain A 713–717) are resolved in the structure, they prove that the interface can be affected by the deletion of the exon ENSE00001157509.

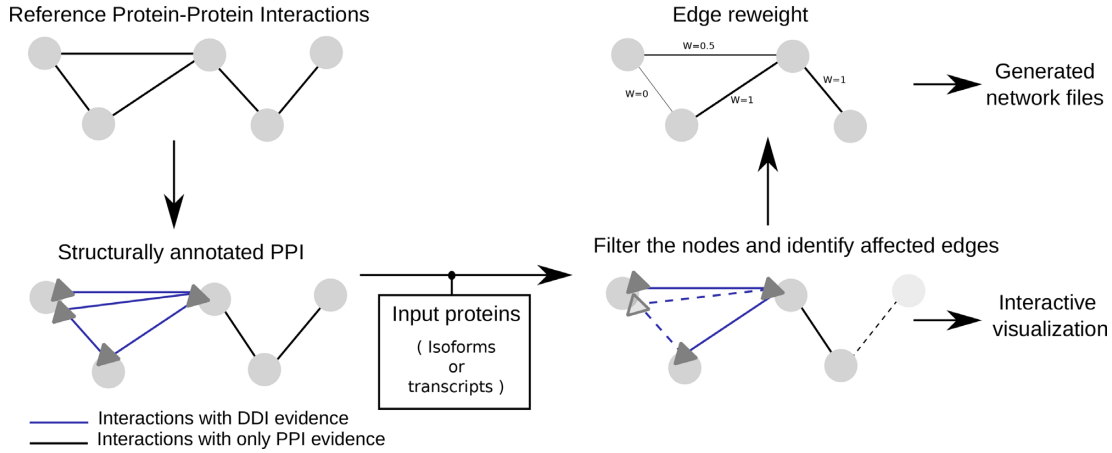


Figure 6. In network-analysis mode, DIGGER highlights domains absent in the user-submitted isoforms and identifies missing interactions mediated by these domains. Edges are scored according to the ratio of missing putative interactions. Results can be visualized or exported for further analysis in third party tools.

necessarily co-expressed in the same condition or tissue. The user can use this mode with RNA-Seq data (Option 2 in DIGGER Network-level analysis) to extract expressed transcripts and explore the specific interactions between them.

In addition to the visualization, DIGGER offers the following benefits over PPIXpress via its network analysis functionality. First, in PPIXpress a PPI is only considered missing if all associated DDIs are missing. In contrast, DIGGER allows more flexibility by offering a filter option based on the ratio of missing interactions. Note that filtering all interactions with weight equal to 0 will be equivalent to the PPIXpress algorithm. Second, PPIXpress only considers the most highly expressed transcript, which is arguably an oversimplification. In contrast, DIGGER combines all structural information from different isoforms of the same gene. As a result, missing domains are defined as those missing in all protein variants in the input list but known to be present in other variants that were not included. Again, users can choose to include one transcript or isoform from each gene, e.g. the most highly expressed one to obtain comparable results to PPIXpress. We believe that these improvements provide the user with considerably more flexibility and better interpretability of the results for in-depth analyses on the system or network level.

The general workflow of DIGGER provides the user with an easy and interchangeable navigation between these modes and the different views (Supplementary Figure S1). As a result, DIGGER is the only database that allows for a complete exploration of AS impact from the exon to the network level. In contrast, comparable tools and methods cover only individual aspects, such as the Ghadie *et al.*

method (18) and DomainGraph (22) that only focus on the isoform interactions or PPIXpress that analyzes transcript expression data (24). Furthermore, DIGGER is the only resource that combines DDIs with residue specific interactions to identify the consequence of skipping an exon.

We could show that DIGGER's ability to map interacting residues to exons enables us to study splicing events that result in hitherto unannotated protein isoforms with experimental evidence. While this is a powerful approach to assess the potential impact of alternative splicing events between two conditions, we caution that the structures used for this annotation are typically derived from the full-length transcript and mostly limited to the major isoform. They thus do not reflect the influence of the exon itself on protein folding. Nevertheless, identifying putatively interacting residues as well as domains encoded by the exonic region allows for exploring all possible scenarios that result from AS events such as exon skipping.

Naturally, the annotations found in DIGGER are limited by the quality of the integrated PPI and DDI data sets as well as the quality of the structural annotations of domains and residue interfaces. Currently, DIGGER covers 37% of the proteins and 13% of the interactions in BioGRID. Although the majority of proteins are annotated with at least one domain (26), the experimental coverage of DDIs is comparably poor. Furthermore, the DDI view of interactions neglects interactions mediated by disordered regions. Moreover, AS events occurring in these exons can possibly alter the translation or the folding of the protein. The function role of these exons is still not very well understood and even controversial (36–37). In the current database, around half of the annotated exons map to disordered regions (48% of the 307 219 annotated exons from protein coding transcripts) which limits the efforts towards a complete structurally annotated isoform interactome. By incorporating residue-level evidence, we increased the structural coverage by 4968 exons that were initially mapped to a disordered region. Another way to approach this problem is to explore linear binding motifs that could further expand our understanding of the rule of individual exons in the PPI. Thus, we plan to incorporate protein binding motifs in a future release of DIGGER and to integrate further data sources for DDIs and PPIs such as STRING (7).

Another challenge in the field is to determine the exact domains or exons responsible for a PPI when multiple domains are mapped to the interaction interface. Our analysis shows that 17 390 PPIs are annotated with multiple DDIs (33% of the structurally annotated PPI). Identifying the AS impact on these interactions is more difficult, since the role of individual domains or exons is not clear. To mitigate this, DIGGER scores the percentage of isoform-specific interactions missing associated domains. Here, users should be careful when choosing a threshold to avoid an excess in false positives or false negatives. Additional experimental results on isoform-specific interactions are needed to resolve this and to determine the best possible threshold. Another possibility to narrow down the regions corresponding to the interacting surfaces between the two proteins is the use of residue-level evidence provided in DIGGER at the exon-level. The existence of an interacting residue in a single specific interface provides strong support that the interaction is specific to that domain (or exon).

CONCLUSION

Recent studies emphasize the considerable influence of AS on human PPIs. As discussed previously by Talavera *et al.* (38), this may lead to a significant bias in network-driven systems biology analysis. For every PPI, there is a potentially large number of isoform combinations that would have to be experimentally validated (2). Given limited experimental data, it is essential to build computational approaches to distinguish between protein isoforms and to identify the function and interactions of putative new variants. DIGGER closes this gap in order to help biomedical researchers to address the complexity in visualizing and analyzing the functional impacts of AS in a user-friendly fashion and on multiple levels, ranging from protein isoforms, via domains, down to exons. DIGGER integrates state-of-the-art annotations of PPIs and DDIs and enriches them with a novel approach to gain residue-level information of PPI. We have shown that the results generated by DIGGER are consistent with experimental evidence in the context of isoform-specific interactions and exon skipping. DIGGER is ideally suited to investigate the differences between isoforms, to analyse the effect of an isoform-switch, or to explore how alternative splicing events such as exon skipping lead to altered interactions of protein isoforms. DIGGER provides a basis for network analysis, by re-weighting the reference PPI based on the structural evidence of the specific interacting proteins. In the future, we envision to extend DIGGER to provide network analysis features, such as de novo network enrichment (39) and to cover additional model organisms for which high-quality PPI networks are available.

DATA AVAILABILITY

DIGGER is accessible at https://exbio.wzw.tum.de/digger. Source code is available at https://github.com/louadi/DIGGER. 3did interactions were downloaded from https://3did.irbbarcelona.org/ (version 2019_01). DOMINE interactions were downloaded from https://manticore.niehs.nih.gov/cgi-bin/Domine (version2.0). BioGRID interactions were downloaded from https://thebiogrid.org/ (version 3.5.187). The latest joint graph as well as the datasets used in DIGGER are available at https://exbio.wzw.tum.de/digger/download. The source code for the RNA-Seq analysis is available at (https://github.com/louadi/RNA-Seq-DIGGER).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant

Sys_CARE [01ZX1908A]); J.B. is grateful for financial support he received through his VILLUM Young Investor Grant [13154] as well as through the European Union's Horizon 2020 project RepoTrial [777111]. Funding for open access charge: BMBF grant Sys_CARE; Technical University of Munich.
Conflict of interest statement. None declared.

REFERENCES

1. Yang,X., Coulombe-Huntington,J., Kang,S., Sheynkman,G.M., Hao,T., Richardson,A., Sun,S., Yang,F., Shen,Y.A., Murray,R.R. *et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, **164**, 805–817.
2. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.*, **5**, R74.
3. David,C.J. and Manley,J.L. (2010) Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.*, **24**, 2343–2364.
4. Beqqali,A. (2018) Alternative splicing in cardiomyopathy. *Biophys. Rev.*, **10**, 1061–1071.
5. Stevens,M. and Oltean,S. (2016) Alternative splicing in CKD. *J. Am. Soc. Nephrol.*, **27**, 1596–1603.
6. Oughtred,R., Stark,C., Breitkreutz,B.-J., Rust,J., Boucher,L., Chang,C., Kolas,N., O'Donnell,L., Leung,G., McAdam,R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
7. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
8. Rao,V.S., Srinivasa Rao,V., Srinivas,K., Sujini,G.N. and Sunand Kumar,G.N. (2014) Protein-protein interaction detection: methods and analysis. *Int. J. Proteomics*, **2014**, 147648.
9. Ellis,J.D., Barrios-Rodiles,M., Çolak,R., Irimia,M., Kim,T., Calarco,J.A., Wang,X., Pan,Q., O'Hanlon,D., Kim,P.M. *et al.* (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell*, **46**, 884–892.
10. Sulakhe,D., D'Souza,M., Wang,S., Balasubramanian,S., Athri,P., Xie,B., Canzar,S., Agam,G., Gilliam,T.C. and Maltsev,N. (2019) Exploring the functional impact of alternative splicing on human protein isoforms using available annotation sources. *Brief. Bioinform.*, **20**, 1754–1768.
11. Barbosa-Morais,N.L., Irimia,M., Pan,Q., Xiong,H.Y., Gueroussov,S., Lee,L.J., Slobodeniuc,V., Kutter,C., Watt,S., Colak,R. *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**, 1587–1593.
12. Climente-González,H., Porta-Pardo,E., Godzik,A. and Eyras,E. (2017) The functional impact of alternative splicing in cancer. *Cell Rep.*, **20**, 2215–2226.
13. Lee,L.Y.-H. and Loscalzo,J. (2019) Network medicine in pathobiology. *Am. J. Pathol.*, **189**, 1311–1326.
14. Wang,X., Wei,X., Thijssen,B., Das,J., Lipkin,S.M. and Yu,H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–164.
15. Mosca,R., Céol,A., Stein,A., Olivella,R. and Aloy,P. (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **42**, D374–D379.
16. Meyer,M.J., Das,J., Wang,X. and Yu,H. (2013) INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics*, **29**, 1577–1579.
17. Mosca,R., Céol,A. and Aloy,P. (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.
18. Ghadie,M.A., Lambourne,L., Vidal,M. and Xia,Y. (2017) Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS Comput. Biol.*, **13**, e1005717.
19. Ghadie,M. and Xia,Y. (2019) Estimating dispensable content in the human interactome. *Nat. Commun.*, **10**, 3205.
20. Mosca,R., Tenorio-Laranga,J., Olivella,R., Alcalde,V., Céol,A., Soler-López,M. and Aloy,P. (2015) dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods*, **12**, 167–168.
21. Sahni,N., Yi,S., Taipale,M., Fuxman Bass,J.I., Coulombe-Huntington,J., Yang,F., Peng,J., Weile,J., Karras,G.I., Wang,Y. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647–660.
22. Emig,D., Salomonis,N., Baumbach,J., Lengauer,T., Conklin,B.R. and Albrecht,M. (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.*, **38**, W755–W762.
23. Tranchevent,L.-C., Aubé,F., Dulaurier,L., Benoit-Pilven,C., Rey,A., Poret,A., Chautard,E., Mortada,H., Desmet,F.-O., Chakrama,F.Z. *et al.* (2017) Identification of protein features encoded by alternative exons using exon ontology. *Genome Res.*, **27**, 1087–1097.
24. Will,T. and Helms,V. (2016) PPIXpress: construction of condition-specific protein interaction networks based on transcript expression. *Bioinformatics*, **32**, 571–578.
25. Raghavachari,B., Tasneem,A., Przytycka,T.M. and Jothi,R. (2008) DOMINE: a database of protein domain interactions. *Nucleic Acids Res.*, **36**, D656–D661.
26. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
27. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
28. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.
29. Goldman,M.J., Craft,B., Hastie,M., Repečka,K., McDade,F., Kamath,A., Banerjee,A., Luo,Y., Rogers,D., Brooks,A.N. *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.*, **38**, 675–678.
30. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
31. de Figueiredo-Pontes,L.L., Wong,D.W.-S., Tin,V.P.-C., Chung,L.-P., Yasuda,H., Yamaguchi,N., Nakayama,S., Jänne,P.A., Wong,M.P., Kobayashi,S.S. *et al.* (2014) Identification and characterization of ALK kinase splicing isoforms in non-small-cell lung cancer. *J. Thorac. Oncol.*, **9**, 248–253.
32. Malakar,P., Chartarifsky,L., Hija,A., Leibowitz,G., Glaser,B., Dor,Y. and Karni,R. (2016) Insulin receptor alternative splicing is regulated by insulin signaling and modulates beta cell survival. *Sci. Rep.*, **6**, 31222.
33. Denley,A., Bonython,E.R., Booker,G.W., Cosgrove,L.J., Forbes,B.E., Ward,C.W. and Wallace,J.C. (2004) Structural determinants for high-affinity binding of insulin-like growth factor II to insulin receptor (IR)-A, the exon 11 minus isoform of the IR. *Mol. Endocrinol.*, **18**, 2502–2512.
34. Kumar,M., Gouw,M., Michael,S., Sámano-Sánchez,H., Pancsa,R., Glavina,J., Diakogianni,A., Valverde,J.A., Bukirova,D., Čalyševa,J. *et al.* (2019) ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.*, **48**, D296–D306.
35. Vivian,J., Rao,A.A., Nothaft,F.A., Ketchum,C., Armstrong,J., Novak,A., Pfeil,J., Narkizian,J., Deran,A.D., Musselman-Brown,A. *et al.* (2017) Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.*, **35**, 314–316.
36. Tress,M.L., Abascal,F. and Valencia,A. (2017) Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.*, **42**, 98–110.
37. Blencowe,B.J. (2017) The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.*, **42**, 407–408.
38. Talavera,D., Robertson,D.L. and Lovell,S.C. (2013) Alternative splicing and protein interaction data sets. *Nat. Biotechnol.*, **31**, 292–293.
39. Alcaraz,N., List,M., Dissing-Hansen,M., Rehmsmeier,M., Tan,Q., Mollenhauer,J., Ditzel,H.J. and Baumbach,J. (2016) Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Res.*, **5**, 1531.

Appendix B

Appendix: Second publication

This is the copyedited PDF of the article originally published in the **Genome Biology** journal.

Citation: Louadi, Zakaria, Maria L. Elkjaer, Melissa Klug, Chit Tong Lio, Amit Fenn, Zsolt Illes, Dario Bongiovanni, *et al.* "Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases." Genome Biology 22, no. 1 (2021): 1-22. https://doi.org/10.1186/s13059-021-02538-1.

Louadi et al. Genome Biology (2021) 22:327
https://doi.org/10.1186/s13059-021-02538-1

Genome Biology

METHOD Open Access

Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases

Zakaria Louadi[1,2], Maria L. Elkjaer[3,4,5], Melissa Klug[1,6,7], Chit Tong Lio[1,2], Amit Fenn[1,2], Zsolt Illes[3,4,5], Dario Bongiovanni[6,7,8], Jan Baumbach[2,9], Tim Kacprowski[10,11], Markus List[1*†] and Olga Tsoy[2*†]

* Correspondence: markus.list@wzw.tum.de; olga.tsoy@uni-hamburg.de
†Markus List and Olga Tsoy are joint last authors.
[1]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany
[2]Institute for Computational Systems Biology, University of Hamburg, Notkestrasse 9, 22607 Hamburg, Germany
Full list of author information is available at the end of the article

Abstract

Alternative splicing (AS) is an important aspect of gene regulation. Nevertheless, its role in molecular processes and pathobiology is far from understood. A roadblock is that tools for the functional analysis of AS-set events are lacking. To mitigate this, we developed NEASE, a tool integrating pathways with structural annotations of protein-protein interactions to functionally characterize AS events. We show in four application cases how NEASE can identify pathways contributing to tissue identity and cell type development, and how it highlights splicing-related biomarkers. With a unique view on AS, NEASE generates unique and meaningful biological insights complementary to classical pathways analysis.

Keywords: Alternative splicing, Differential splicing, Functional enrichment, Systems biology, Protein-protein interactions, Disease pathways, Platelet activation, Multiple sclerosis, Dilated cardiomyopathy

Background

Alternative splicing (AS) boosts transcript diversity in human cells [1] and thus contributes to tissue identity [2], cell development [3], and pathology in, e.g., cardiomyopathy [4], muscular dystrophy [5], or autoimmune diseases [6]. It is estimated that up to 30% of disease-associated genetic variants affect splicing [7]. RNA sequencing technologies (RNA-seq) allow the quantification of different types of AS events and detect splicing abnormalities in disorders. However, RNA-seq utility is currently limited by our incomplete understanding of the functional role of specific exons or the transcripts they contribute to.

A major challenge in AS analysis is the functional interpretation of a set of events, including isoform switching events and differentially spliced exons. The usual approach is to perform gene set enrichment or overrepresentation analysis [8–10]. This approach treats all genes affected by AS equally, neglecting that some AS events may not be functionally relevant at the protein level [11] or result from noise in the splicing

machinery [12]. Furthermore, functional differences between protein isoforms remain uncertain in many cases. A promising strategy to identify relevant AS events is to focus on those that lead to meaningful changes in the protein structure. Recent studies have shown that AS has the potential to rewire protein-protein interactions by affecting the inclusion of domain families [13] and linear motifs [14] or by activating nonsense-mediated decay [15].

This motivated the creation of databases and tools that predict the consequences of individual AS events or isoform switches. IsoformSwitchAnalyzeR [16], tappAS [17], DoChaP [18], and Spada [19] support transcript-level (as opposed to exon-level) analysis to identify isoform switches and their impact on the translation and the resulting isoforms features, such as domains, motifs, and non-coding sites. Exon Ontology [20] and DIGGER [21] support exon-level analysis to identify exon skipping events and their possible impact on the protein structure and function. Spada and DIGGER further consider the impact of AS on protein-protein interactions.

Most existing tools allow investigating AS-driven changes in an explorative fashion but tools for systematic analysis of functional effects of AS are lacking. Exon Ontology performs statistical tests to identify enriched features within a set of skipped exons. One example is domain families affected by AS across proteins more frequently than expected. However, none of the existing tools offer a systems biology view to specifically highlight functional consequences of AS events.

To tackle these limitations, we developed the first tool for functional enrichment of AS events. NEASE (Network-based Enrichment method for AS Events) first detects protein domains affected by AS and then uses protein-protein interactions (PPI) integrated with domain-domain interactions (DDI) [21], residue-level, and domain-motif interactions (DMI) [22] to identify interaction partners likely affected by AS. Next, it employs an edge-level hypergeometric test for gene set overrepresentation analysis. This approach is new in the way genes are selected for the enrichment test. Rather than considering only differentially spliced or expressed genes, which is currently the most common strategy, NEASE uses network information to select genes that are likely affected in the interactome. This is also superior to a simple network enrichment analysis, as we consider only those edges for which an AS contribution seems relevant and for which false positive results are less likely. We evaluated NEASE using multiple datasets from both healthy and disease cohorts. We show that the NEASE approach complements gene-level enrichment, and even outperforms it in scenarios where gene-level enrichment fails to find relevant pathways. Moreover, NEASE generates unique and meaningful biological insights on the exact impact of AS. Furthermore, since the statistical approach is network-based, NEASE can prioritize (differentially) spliced genes and find new disease biomarkers candidates in case of aberrant splicing. The NEASE Python package, freely available at https://github.com/louadi/NEASE, provides multiple functions for a deeper analysis and visualization of affected protein domains, edges, and pathways (individually or as a set).

Results

Overview of NEASE

NEASE uses a hybrid approach that combines biological pathways with PPIs and DDIs to perform functional enrichment of AS. First, we use the structural annotation of

known isoforms by mapping protein domains from the Pfam database [23] to the corresponding exons (Fig. 1A). Second, we construct a structural joint graph as previously reported [21] by enriching the BioGRID PPI [24] with DDIs (from DOMINE [25] and 3did [26]), DMIs from the Eukaryotic Linear Motif resource (ELM) [22], and interface residues from the Protein Data Bank (PDB) [27] (see Methods). In the joint graph, protein features such as domains, motifs, and residues are mapped to their mediated

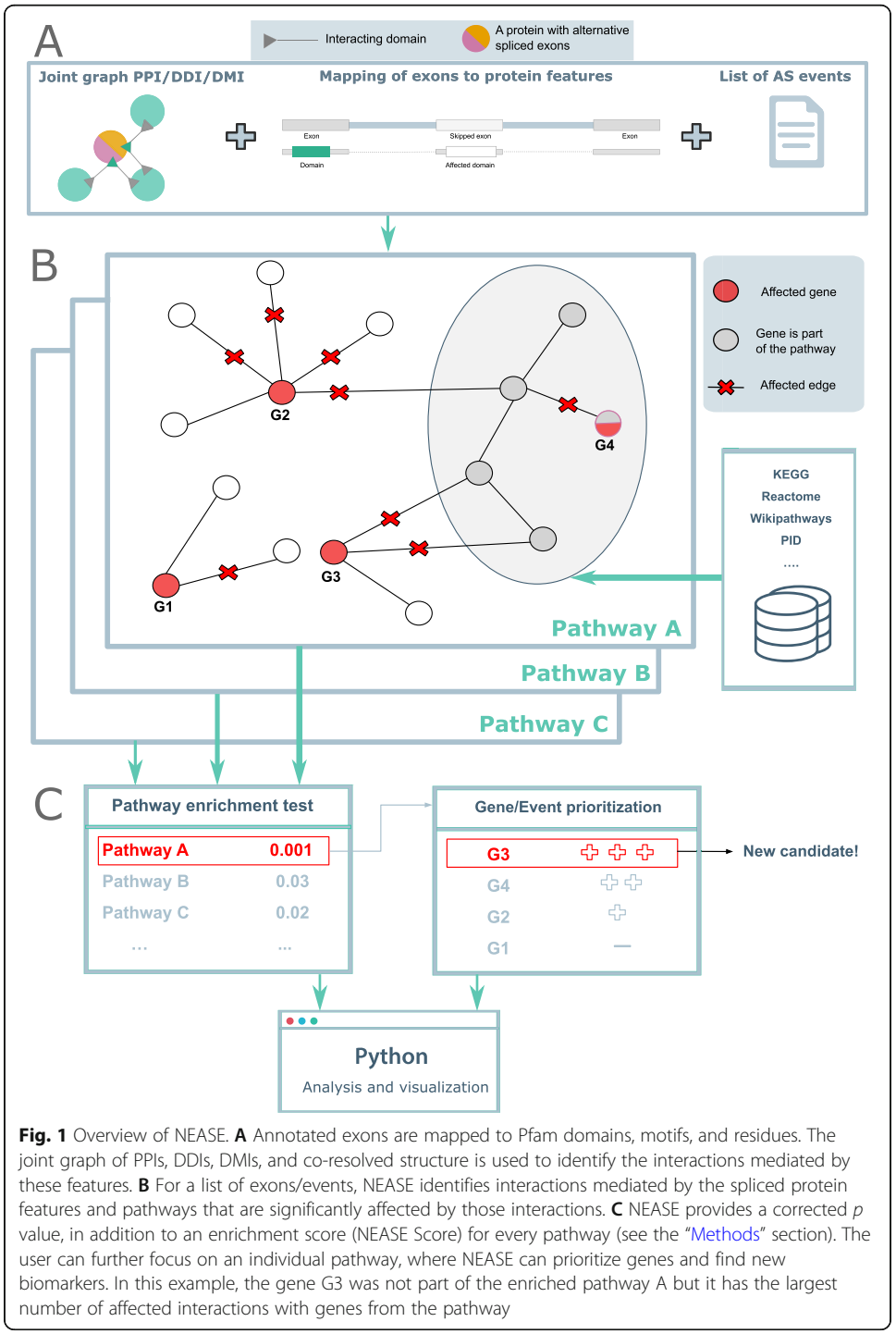


Fig. 1 Overview of NEASE. **A** Annotated exons are mapped to Pfam domains, motifs, and residues. The joint graph of PPIs, DDIs, DMIs, and co-resolved structure is used to identify the interactions mediated by these features. **B** For a list of exons/events, NEASE identifies interactions mediated by the spliced protein features and pathways that are significantly affected by those interactions. **C** NEASE provides a corrected *p* value, in addition to an enrichment score (NEASE Score) for every pathway (see the "Methods" section). The user can further focus on an individual pathway, where NEASE can prioritize genes and find new biomarkers. In this example, the gene G3 was not part of the enriched pathway A but it has the largest number of affected interactions with genes from the pathway

interactions. Thus, NEASE provides an exon-centric view of the interactome and addresses the limited exon-level annotation. Exons are represented by the features they encode, and interactions between features are represented by edges. In this way, the impact of AS can be seen as an edgetic change in the network. Analyzed AS events are viewed as a set of affected edges that represent gained or lost PPIs.

We then perform statistical tests to find enriched pathways and most likely responsible genes (Fig. 1B). Following, (differential) splicing analysis, a one-sided hypergeometric test is used to test for enrichment of a given pathway or gene set by considering all edges affected by AS in an experiment. A similar test is applied for each spliced gene to prioritize the most relevant events/genes that are affecting a pathway. We further introduce a weighted score (NEASE score) that penalizes hub nodes that are more likely to be connected to the pathway of interest by chance. Notably, this approach also considers genes that are not part of the existing pathway definition but show a significant number of interactions with the pathway, highlighting new putative biomarkers (see Methods and Additional file 1: Figure S6, for details).

The Python package provides an interactive analysis. Using a list of exons or events, users can run a general enrichment on 12 different pathway databases (collected from the ConsensusPathDB resource [28]), followed up by a specific analysis and visualization for a single affected pathway or module of interest (Fig. 1C). To provide analysis for individual isoforms and events, we linked NEASE to our previously developed database DIGGER, which provides an isoform- and exon-centric view of the interactome [21].

To check if the structurally annotated PPI is more biased to hubs than the standard PPI network, we computed the node degree distribution of the network before and after filtering for the structure evidence. As shown in Additional file 1: Figure S1, the two histograms show similar trends with an overall smaller number of edges in the structurally annotated PPI. The latter has a maximum node degree equal to 424, compared to 2887 in the full PPI. This observation shows that the structurally annotated PPI does not increase the bias towards hub genes of the interactome.

NEASE gives insights into the role of the muscle- and neural-specific exons

Recent studies suggest that the regulation of AS occurs in a tissue-specific manner and leads to remodeling of protein-protein interactions [29]. Understanding the functional impact of co-regulated exons is critical in understanding gene regulation. We applied NEASE to tissue-specific exons reported in VastDB, a resource that provides information on multiple types of AS events detected by RNA-seq from different tissue types and developmental stages [30]. We extracted 2831 exon skipping events and Percent Spliced In values (PSI) from 12 different human tissue types (Additional file 7, see Methods). We then performed hierarchical clustering on the z score standardized PSI values (Fig. 2A). The heatmap shows two distinct clusters, where neural-specific and muscle-specific (merged with heart-specific) exons are dominant.

Next, we extracted 56 skipped exons with a high PSI in the muscle tissues and 62 skipped exons with a high PSI in the neural tissues (z score $\geq +2$, see the "Methods" section). We checked how many of these events are overlapping with protein features. As shown in Fig. 2B, 27% of the upregulated exons in muscle tissues (13) and 28% of

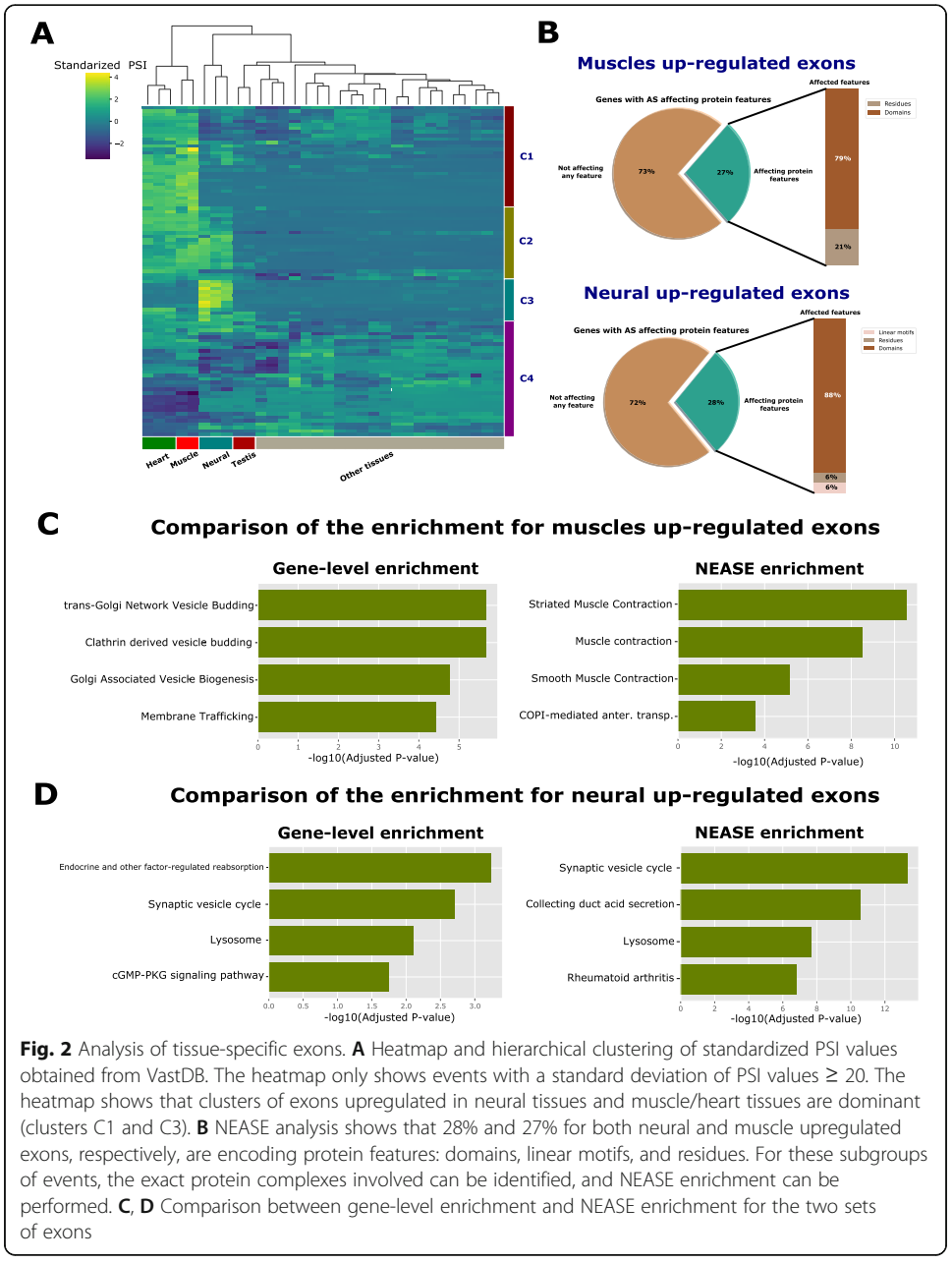


Fig. 2 Analysis of tissue-specific exons. **A** Heatmap and hierarchical clustering of standardized PSI values obtained from VastDB. The heatmap only shows events with a standard deviation of PSI values ≥ 20. The heatmap shows that clusters of exons upregulated in neural tissues and muscle/heart tissues are dominant (clusters C1 and C3). **B** NEASE analysis shows that 28% and 27% for both neural and muscle upregulated exons, respectively, are encoding protein features: domains, linear motifs, and residues. For these subgroups of events, the exact protein complexes involved can be identified, and NEASE enrichment can be performed. **C**, **D** Comparison between gene-level enrichment and NEASE enrichment for the two sets of exons

the upregulated exons in the neural tissues (17) overlap with protein features. NEASE also provides statistics of how many of these domains have known binding partners in the joint graph. In the two sets, around 60% of the affected domains have known interactions in our joint graph: 8 binding domains in the muscle tissues and 10 binding domains in the neural tissues (Additional file 2: Tables S4, S5 and Additional file 3: Tables S8, S9). We further identified one affected motif in the gene ATP2B1 in neural exons. For these groups of events, the exact protein complexes involved can be identified, and NEASE statistical analysis can be performed to determine affected pathways. However, it is important to keep in mind that not all affected domains are necessarily interacting domains but could also be regulating gene expression by binding to DNA or RNA [31].

First, we ran a gene set overrepresentation analysis (one-sided hypergeometric test), which we refer to as gene-level enrichment, to detect enriched pathways (see the "Methods" section). Next, we applied NEASE to the same genes to detect pathways affected by AS. Unlike the gene-level enrichment, the results obtained from NEASE in both sets better explain the functional role of the regulated exons (Fig. 2C, D). We also compared with the results from the Network Enrichment Analysis method NEA [32]. NEA is a PPI-based approach that considers all edges for statistical tests. In contrast, NEASE considers only AS-affected edges. For a fair comparison, we run NEA with the same PPI network (BioGRID) and same pathways databases (see the "Methods" section). The results of NEA did not improve over the classic gene-level enrichment (Additional file 1: Figure S2), which suggests that our exon-specific approach helps to narrow down the exact complexes/pathways affected by AS and reduces false positives.

To further validate the robustness of the enrichment obtained by NEASE, we further conducted permutation tests. Here, our null hypothesis is that the tissue-specific exons are not different from a random set of exons in terms of the quality of the functional enrichment (measured as the p values of the hypergeometric test). For a more realistic scenario, our background set of exons considers only exon skipping events that can actually be found in these tissues (see the "Methods" section for details). This approach will also help evaluate our methods against known and unknown biases. The empirical p values of the permutation test, which indicate the chance of finding an enrichment p value as low or lower than the one reported by NEASE, are 0.0008 and 0.0001 for neural and muscles upregulated exons, respectively. These results further demonstrate the robustness of our analysis.

The upregulated exons in heart and muscle tissues were enriched in "Muscle Contraction" pathways (Fig. 2C and Additional file 2: Table S7), while, in the gene-level enrichment, the pathways were related to very common subcellular functions such as the Golgi apparatus, which also is an organelle for collecting, modifying or destroying protein products (Fig. 2 C and Additional file 2: Table S6). NEASE provides detailed information about the affected domains and their interaction partners (Additional file 1: Table S1). The domain Tropomyosin (Pfam id: PF00261), which is part of the gene TPM1, e.g., is involved in the regulation of muscle contraction via actin and myosin. GAS2 (Pfam id: PF02187) is a domain of DST, a dystonin encoding gene, which plays a role in maintaining the integrity of the cytoskeleton. AS affects its binding with the gene CALM1 that encodes a calcium-binding protein involved in various calcium-dependent pathways like muscle contraction [33].

The exons upregulated in neural tissues showed enrichment in the synaptic vesicle cycle pathway responsible for the communication between neurons (Fig. 2D). Gene-level enrichment performed on par with NEASE, resulting in the same pathway but with a lower rank and significance (adjusted p values: 1.494631e−16 using NEASE and 0.0039 using gene-level, Additional file 3: Tables S10 and S11). Notably, NEASE also detected an enrichment in "oxidative phosphorylation", which is the initiator for powering all major mechanisms mediating brain information processing [34]. The neuron's energy demands are remarkable both in their intensity and in their dynamic range and quick changes [35–38]. Therefore, AS could modify oxidative phosphorylation to serve tissue-specific needs. Experimental studies have also found that several key enzymes in "oxidative phosphorylation" are

spliced, e.g., pyruvate kinase (PKM) that shifts from the PKM2 to the PKM1 isoform [39, 40]. NEASE also provides a detailed view on the affected mechanisms, such as an exon skipping event in the gene ATP6V0A1 overlapping with the V_ATPase_I domain (PFAM id: PF01496) and affecting the binding with seven other proteins from the complex vacuolar ATPase (V-ATPase) (p value: 5.853289e−17, Fig. 3, Additional file 1: Table S2). V-ATPase is required for synaptic vesicle exocytosis [41] The a1-subunit of the V0 domain in ATP6V0A1 was recently shown to be highly expressed in neurons and to be essential for human brain development [42, 43]. In another example, NEASE identified two co-regulated events of the genes CLTA and CLTB (Fig. 3). CLTA and CLTB genes are involved in Clathrin-dependent endocytosis which forms clathrin-coated vesicles. Both genes play a major role in forming the protein complex of the coated vesicle. Both events affect the same domain Clathrin light chain (Pfam id: PF01086). The Clathrin light chain domain binds to CLTC and CLTCL1 which are the Clathrin heavy chain genes (p value: 6.943483e−05). These results suggest that the formation of this complex is co-regulated by AS. A similar finding about the role of the Clathrin light chain in neurons was also described in [44]. NEASE highlights these co-regulated events at the network level (Fig. 3). As a sanity check, we manually checked the PSI values of these critical events identified by NEASE in the Genotype-Tissue Expression data set (GTEx), a comprehensive resource for tissue-specific gene expression and regulation [45]. VastDB includes the quantification of PSI values from 8378 samples (49 tissues and 543 individuals) from GTEx version 6 on their website (https://vastdb.crg.eu/). As shown in the examples in Additional file 1 Figures S4 and S5, the exons are confirmed to be highly upregulated in their respective tissues. The analysis generated from VastDB using NEASE agrees with the latest studies at transcriptomics and proteomics levels that emphasize the crucial role of AS in the function and development of brain and heart tissues [46–48].

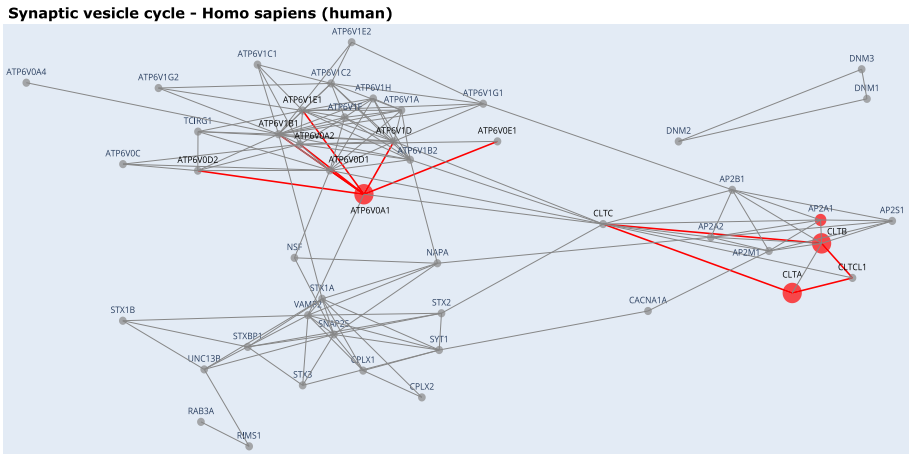


Fig. 3 NEASE visually highlights the impact of the AS regulation at the interactome level. The gray nodes represent proteins from the pathway and the red nodes represent genes with AS events. Red edges represent the affected interactions for the nodes with known DDIs, DMIs, or co-resolved structures. The visualization of the pathway "Synaptic vesicle cycle" from the KEGG database for the exons upregulated in the neural tissues shows that the splicing in the genes CLTA and CLTB is co-regulated and affects the interactions of the same complex. Similarly, NEASE highlights the importance of the domain ATP6V0A1 which is upregulated in neural tissues and binds seven proteins from the "Synaptic vesicle cycle" pathway

NEASE reveals splicing-related differences of reticulated and mature platelets

AS does not only drive tissue-specific regulation but also plays a major role in cell differentiation and maturation. To illustrate an example of the utility of NEASE in such studies, we used the RNA-seq data set from [49] which compares the transcriptome profiles of reticulated platelets and mature platelets from healthy donors. Reticulated platelets are younger [50], larger in size, and contain more RNA [51]. Moreover, they have a prothrombotic potential and are known to be more abundant in patients with diabetes, acute or chronic coronary syndrome, and in smokers [51–53]. Additionally, elevated levels of reticulated platelets in peripheral blood are predictors of insufficient response to antiplatelet therapies (e.g., aspirin and P2Y12 inhibitors) and are promising novel biomarkers for the prediction of adverse cardiovascular events in different pathological settings [52, 54]. A strong enrichment of pro-thrombotic signaling in reticulated platelets was observed in healthy donors [49]. Comparative transcriptomic analysis revealed a differential expression of several pathways in addition to an enrichment of prothrombotic pathways and transcripts of transmembrane proteins as the collagen receptor GPVI, the thromboxane receptor A2 and the thrombin receptors PAR1 and PAR4. Gene set enrichment analysis indicated an upregulation of entire prothrombotic activation pathways as the thrombin PAR1 and integrin GPIIb/IIIa signaling pathway in reticulated platelets.

Since AS has been described to occur in platelets [55], we wanted to investigate the splicing patterns between the previously defined reticulated and mature platelet subgroups. Using MAJIQ [56] (see the "Methods" section), we found 169 differentially spliced genes. From 25 affected protein domains, 17 have known interactions (68% of affected domains, Fig. 4A, Additional file 4: Tables S12 and S13). Other affected protein features include 6 residues involved in PPIs and one linear motif in the gene PAWR.

We observed that the enrichment at the gene-level using the Reactome [57] database ranks general cellular pathways higher, including "Membrane Trafficking" and "Vesicle-mediated transport," and "Golgi-to-ER retrograde transport." An exception is the "Circadian Clock" pathway, which is hypothesized to be related to platelet activation [58] (Fig. 4B). The pathway "Platelet activation, signaling and aggregation" was less significant in gene-level enrichment (adjusted p value: 0.061, Additional file 4: Table S14) compared to NEASE enrichment (adjusted p value: 0.004, Additional file 4: Table S15). Using NEASE, we obtained more meaningful results and unique pathways. As shown in Fig. 4C, the most significant pathways in reticulated platelets are G Protein-Coupled Receptor-related. G proteins are essential in the second phase of platelet-dependent thrombus formation [59]. Furthermore, GPCR isoforms are known to have distinct signaling properties [60]. Other relevant pathways associated with platelet activation are "Hemostasis," "Thromboxane signaling through tp receptor," and "Platelet homeostasis." The full tables for enrichment at the gene level and using NEASE are available in the Additional file 4: Tables S14 and S15. The upregulation of these pathways in reticulated platelets emphasizes their previously described prothrombotic phenotype and their involvement in several downstream signaling processes.

We also looked at the individual AS events driving this enrichment. For each affected feature, NEASE tests if it significantly interacts with the GPCR downstream signaling pathway (Additional file 4: Table S16, see the "Methods" section). Figure 4C illustrates affected genes and their p value ranking. The top gene is GNAQ (G-protein subunit

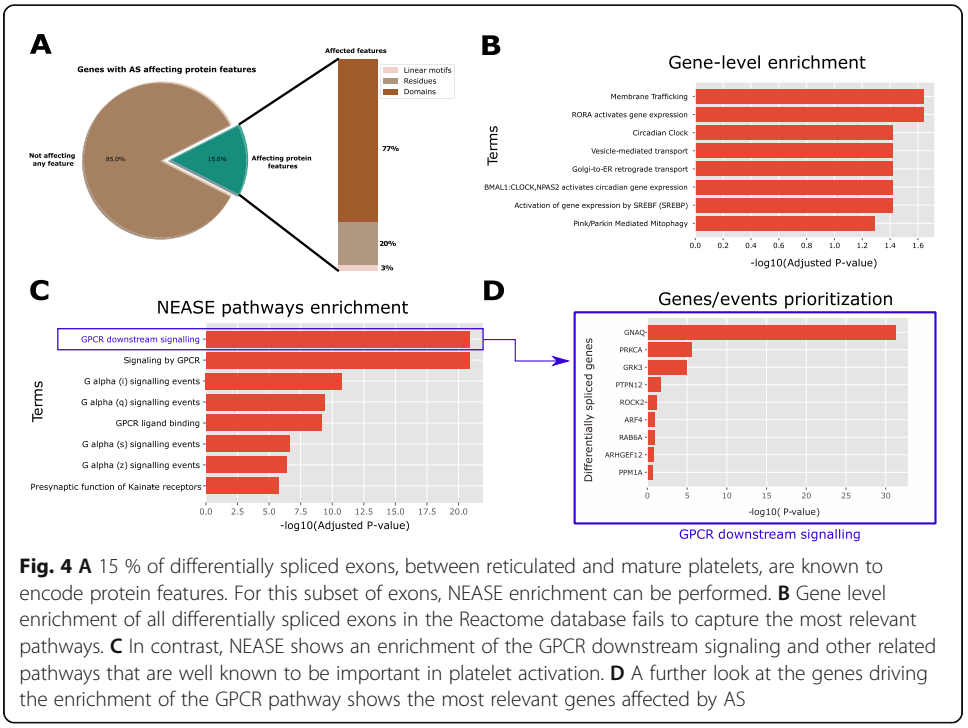


Fig. 4 **A** 15 % of differentially spliced exons, between reticulated and mature platelets, are known to encode protein features. For this subset of exons, NEASE enrichment can be performed. **B** Gene level enrichment of all differentially spliced exons in the Reactome database fails to capture the most relevant pathways. **C** In contrast, NEASE shows an enrichment of the GPCR downstream signaling and other related pathways that are well known to be important in platelet activation. **D** A further look at the genes driving the enrichment of the GPCR pathway shows the most relevant genes affected by AS

alpha q), which is known to be involved in signal transduction in platelets leading to platelet activation [61]. The regulation of the G-protein alpha subunit can be an indication that compared to mature platelets, reticulated platelets are more involved in various signal transduction pathways related to, e.g., pro-thrombotic processes [51]. PRKCA, which also showed different splicing patterns between the two platelet subgroups, plays a major role in the platelet formation process by modulating platelet function [62], megakaryocyte function, and development [63] and negatively regulates pro-platelet formation [64]. Moreover, the regulation of PRKCA binding in reticulated platelets might refer to the young nature of reticulated platelets, which have undergone the pro-platelet formation process more recently than mature platelets [50, 65].

NEASE characterizes complex disorders such as Multiple Sclerosis

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system. Early in the disease course, MS is characterized by focal lesions in the brain induced by an influx of systemic inflammatory cells. These active lesions infiltrated by immune cells and activated microglia are characterized by inflammatory demyelination and axonal loss [66]. The surrounding white matter tissue is termed normal-appearing white matter due to diffuse pathology without focal lesion activity and dense immune activity [67]. The etiology of MS remains unknown. Recently, a systematic literature review found 27 genes that were alternatively spliced in MS patients [68].

We used RNA-Seq of macrodissected areas from postmortem white matter tissue of patients with progressive MS [69]. We compared normal-appearing white matter and active lesions regions from postmortem white matter brains of MS patients. We found

109 differentially spliced genes with 19 affected domains and one linear motif with known interactions, in addition to 6 known interacting residues. In total, NEASE identified 156 affected interactions (Additional file 5: Tables S17 and S18).

Gene-level enrichment ranks high pathways likely irrelevant that are involved in muscle contraction, cardiac conduction, and membrane trafficking, with the exception of Ca2+ ion flow across membranes (Additional file 5: Table S19). Ca2+ is an essential signal molecule for all cell activity. Although deregulation of calcium signaling is related to the pathogenesis of multiple diseases [70], including neurological disorders [71], it is not specific to neuronal tissues. In line with the neurodegenerative and immune-mediated features of MS, NEASE found unique enriched pathways related to brain network signaling and neuronal pathways "Neurotransmitter receptors and postsynaptic signal transmission," "Transmission across Chemical Synapses," "Activation of NMDA receptor and postsynaptic events," "MAPK family signaling cascades," "Neuronal System"), as well as pathways related to immune responses ("interleukin-17 signaling," "Toll-Like Receptor 10 (TLF10) Cascade") (Table 1 and Additional file 5: Table S20). Two other pathways were related to the uptake of anthrax or bacterial toxins. This could be a result of clean-up from toxic inflammatory processes or increased presence of invaders due to the leaky brain-blood-barrier in MS [72–74]. Additionally, it also supports the theory of infections as the trigger of lesion damage in MS [75].

As shown in Table 1, the pathway "Uptake and function of anthrax toxins" has the best overall adjusted *p* value, calculated only based on the total number of edges affecting the pathway. When we also included the number of significant genes and calculated

Table 1 NEASE enrichment obtained from AS comparison between normal-appearing white matter and acute lesions, from multiple sclerosis patients. The highly enriched pathways belong to Neurotransmitter receptors, MAPK, and bacterial infection. Most of these pathways are hallmarks of MS. The NEASE score is obtained after combining the *p* value with the number of significant genes. The latter is obtained after individual tests for each gene in the column "Spliced genes" (see the "Methods" section)

Pathway name	Spliced genes (number of interactions affecting the pathway)	*p* value	adj *p* value	NEASE score
Neurotransmitter receptors and postsynaptic signal transmission	GRIA1 (7), ATP2B1 (2), BRAF (4), MAP2K4 (1), GRIN1 (4)	4.38e−09	0.000004	16.71
Uptake and function of anthrax toxins	ATP2B1 (1), BRAF (5), MAP2K4 (3)	2.98e−09	0.000004	14.76
Transmission across chemical synapses	GRIA1 (7), ATP2B1 (2), BRAF (4), MAP2K4 (1), GRIN1 (4)	5.65e−08	0.000010	14.49
Uptake and actions of bacterial toxins	ATP2B1 (1), BRAF (5), MAP2K4 (3)	3.46e−08	0.000009	12.92
Neuronal system	GRIA1 (7), ATP2B1 (2), BRAF (4), MAP2K4 (1), GRIN1 (4)	8.71e−07	0.000122	12.11
MAPK family signaling cascades	MYH10 (2), ATP2B1 (1), BRAF (17), MAP2K4 (5), GRIN1 (3)	1.52e−06	0.000184	10.07
Activation of NMDA receptor and postsynaptic events	GRIA1 (2), ATP2B1 (1), BRAF (4), MAP2K4 (1), GRIN1 (3)	2.12e−06	0.000241	9.82
FCERI mediated MAPK activation	MYH10 (1), BRAF (7), MAP2K4 (8)	2.52e−07	0.000038	9.33
RAF/MAP kinase cascade	MYH10 (1), ATP2B1 (1), BRAF (16), MAP2K4 (4), GRIN1 (3)	1.00e−06	0.000130	8.48
Signaling by moderate kinase activity BRAF mutants	MYH10 (1), BRAF (14), MAP2K4 (2)	8.30e−09	0.000004	8.08

NEASE scores (see the "Methods" section), NEASE ranks the pathway "Neurotransmitter receptors and postsynaptic signal transmission" first, and moves pathways such as "Transmission across Chemical Synapses" and "Neuronal System" higher in the rank. These observations illustrate the usefulness of the NEASE score as a complement to the global edge-based enrichment.

Two of the most significant genes in the "Neurotransmitter receptors" pathway were GRIN1 and GRIA1 (Additional file 5: Table S21). GRIN1 encodes GluN1, which is one of the two obligatory subunits for the NMDAR1 receptor, whereas GRIA1 encodes the AMPAR1 subunit. Their ligand is glutamate, and they are both ionotropic receptors and have been associated with MS disease severity [76–78]. Interestingly, AS of MAP2K4 appeared in both brain-related and immune-related pathways, significantly enriched in active lesions *vs* normal-appearing white matter (Table 1). MAP2K4 is a mitogen-activated protein kinase (MAPK) orchestrating multiple biological functions [79, 80]. AS of MAP2K4 has been found in rheumatoid arthritis [81], as well as in pathways of patients with other autoimmune diseases [82]. MS also precedes autoimmune attack, and therefore AS of MAP2K4 in active lesions detected with NEASE may represent dysregulated immune responses originating from the infiltrating immune cells or inflammatory-activated brain cells. This is supported by previous studies that found *(i)* overactivity of MAPK pathways in microglia (the resident immune cell of the brain) during neurodegeneration [83, 84], and *(ii)* increased phosphorylation of MAPK kinases in the systemic immune cells of MS patients [85, 86]. A recent study also characterized activated MS-specific pathways in immune cells from blood using phosphoproteomics. Here, MAP2K4 and its interaction partners (e.g., TAK1) were present in MS-specific signaling activity [87]. Future functional studies on the AS of MAP2K4 may help explain if AS could be the reason for increased phosphorylation and overactivity detected in MS. AS of MAP2K4 could result in switching protein conformation, increasing susceptibility to phosphorylation, or changing the downstream protein cascade.

With NEASE, we were able to specifically detect AS of genes and related pathways already known to be dysregulated within MS from excitotoxicity to inflammation. The detected AS genes in active lesions *vs* normal-appearing white matter demonstrate how major components in signaling activities may be fine-tuned/changed from regulation of a homeostatic state to an inflammatory state. Combining NEASE with functional experiments to understand the biological impact of AS could fuel new therapeutic opportunities for complex neurological diseases as MS. Novel developments in genome-editing tools and gene-specific strategies have made it possible to use antisense oligonucleotides or small modulators for splice modification. This is already used in the rare neuromuscular disease, spinal muscular atrophy, where an antisense oligonucleotide binds to a site near splicing to ensure the inclusion of an exon during the splicing event [78].

NEASE finds new biomarker candidates for dilated cardiomyopathy

AS might play a role in driving dilated cardiomyopathy (DCM) [88]. DCM is a common heart muscle disease that is often diagnosed with structural abnormalities resulting in impaired contraction. Previous studies have shown a large number of differentially used exons in DCM patients [4, 10]. In this analysis, we used a list of 1212 differentially used

exons between DCM patients and controls as reported by Heinig et al. [10]. 29% of these exons overlap with protein features, including 230 domains and 15 linear motifs. (Additional file 6: Tables S22 and S23). In this exon set, both the gene level enrichment and NEASE show very similar results (Additional file 6: Tables S24 and S25). In both methods, we found that the list of exons was enriched in the dilated cardiomyopathy (DCM) pathway from KEGG, as well as, "Adrenergic signaling in cardiomyocytes" and "Regulation of actin cytoskeleton".

In contrast to gene-level enrichment analysis, NEASE is able to score the contribution of alternatively spliced genes that are interacting with but are not part of the DCM pathway, allowing us to highlight putative biomarkers (Table 2, Additional file 6: Table S26, Additional file 1: Figure S3). The Myosin head domain from the gene MYO19 interacts with 6 other genes associated with DCM: (1) MYL2, which triggers contraction after Ca+ activation [89]; (2–5) TPM1/TPM2/TPM3/TPM4, which encode the TPM protein—the main regulator of muscle contraction [90]; and (6) ACTG, which encodes actin. Interestingly, MYO19 has not been investigated for its role in DCM, while its interacting genes are associated with DCM [91–94]. Additionally, the gene OBSCN has one affected interaction with the TTN gene [95]. The TTN gene itself is also differentially spliced and associated with DCM [95]. OBSCN was recently reported as a new DCM candidate [96, 97]. Another interesting example is CACNA1C (Calcium Voltage-Gated Channel Subunit Alpha1 C), an already known DCM candidate [98]. The differentially spliced exon overlaps with the domain Ion_trans (Pfam id: PF00520) which is essential for myocyte contraction [99]. The affected interaction identified is with the ryanodine receptor 2 (RYR2). In striated muscles, the excitation-contraction coupling is mediated by this complex [100]. Both CACNA1C and RYR2 are part of the KEGG DCM pathway [101]. Alterations in ryanodine receptors were repeatedly reported to be related to heart failure [102–104].

Discussion

In spite of its importance for biomarker and therapeutic target discovery, differential AS is still not a routine part of transcriptome analysis. A key reason for this could be the lack of suitable methods and software tools for AS-specific functional analysis. Our method NEASE closes this gap and provides a unique view on the impact of AS complementary to functional insights gained from traditional gene-level enrichment analysis. We applied NEASE to four diverse data sets and show that its results generate

Table 2 Enrichment of the pathway "Dilated cardiomyopathy (DCM)" from KEGG for the exons differentially used in DCM patients. The table shows the most significant genes (p value< 0.05) (see the "Methods" section)

Differentially spliced genes	DCM associated	Percentage of affected edges associated with DCM	*P* value	Affected binding (edges) associated with DCM
MYO19	No	6/51	0.000002	MYL2, TPM4, TPM3, TPM2, TPM1, ACTG
OBSCN	No	1/2	0.014	TTN
USP49	No	1/4	0.028	PRKACA
CACNA1C	Yes	1/4	0.028	RYR2

novel disease-relevant insights and provide valuable context to prior findings on altered RNA- and protein-expression levels consistent with recent literature.

In many cases, NEASE improves over gene-level enrichment analysis focusing on differentially spliced genes. One potential reason for this could be that not all AS events are necessarily functional [11, 12]. NEASE mitigates this by focusing on AS events that affect protein domains. However, it is important to keep in mind that this is not the only way to define functional AS events. AS also affects interacting disordered regions [14] or facilitates nonsense-mediated decay [105].

AS events could also lead to completely different functions or interactions [106], e.g., two isoforms can have different interaction partners depending on the inclusion or loss of a single domain [13]. Such changes in the interactome can not be captured with gene-level enrichment which has a strict focus on nodes rather than edges. With NEASE, we could show that integrating structural information at the exon level and PPI networks helps to identify the functional impact of differentially spliced and co-regulated exons. In practice, we consider both approaches as complementary and recommend running gene-level and edge-level enrichments together (both supported by the NEASE package). Note that while our analysis focuses on exon skipping events as the most studied event type, our method is generally agnostic to the event type.

NEASE relies on structurally annotated interactions and existing pathway annotations from databases such as KEGG [101] and Reactome [57]. Leveraging reliable structural information and established pathways likely removes many false positive PPI from considerations. While the structural annotations are generally of high quality, it should be noted that their coverage is still limited and, thus, the number of exons considered in our method is comparably low. For instance, the percentage of considered exons, in our example datasets, ranges between 15 and 30%, which is still far from being a global analysis of AS. Expanding the annotations at isoform-level and more widespread availability of structural information will greatly raise the usefulness of NEASE in the future. We also emphasize that while all events can potentially affect protein interactions on the domain level, not all AS events yield functional isoforms and other processes such as nonsense-mediated decay need to be considered as well. In the future, further progress is urgently needed to link transcriptomics and proteomics for better characterization and understanding of the exact impact of AS events. With our current approach, a large fraction of the PPI network remains unexplored, suggesting that adapting de novo network enrichment methods such as KeyPathwayMiner [107] towards AS could be a promising research direction to uncover previously unknown disease mechanisms. NEASE currently considers the immediate neighborhood of a pathway in the PPI network. When carefully considering the expected increase in false positives, one could also increase the size of the pathway neighborhood using, e.g., a fixed radius for shortest paths. While these are attractive approaches, the biases of the PPI towards hubs, as well as the high number of false (or missing) edges of PPI, in its current form, make such approaches hard to control and statistically challenging. Even though NEASE is relatively conservative, we demonstrated that it is simple, robust, and generates meaningful and interpretable results. Thus, it provides an unprecedented opportunity to understand the functional impact of tissue-, developmental- and disease-specific AS in a system biology manner.

Conclusions

While a plethora of gene set enrichment methods have been proposed in recent years, AS is typically not addressed specifically. Thus, NEASE closes an important gap in functional enrichment analysis of transcriptomics data. The analyses described here, confirm the widespread impact of AS in multiple biological processes and disorders. In the future, we plan to extend NEASE with further model organisms and to add structural annotations covering more types of AS events. Finally, we plan to integrate NEASE with the DIGGER web tool [21] for seamless downstream analysis of AS in the web browser with the vision of establishing functional AS event analysis as a routine step in the transcriptomic analysis.

Methods

NEASE data sources

We construct a human structurally annotated PPI as described previously [21]. Briefly, we integrate DDI and PPI information into a joint network where DDIs were obtained from 3did (v2019_01 [26]) and DOMINE (v2.0 [25] including high- and mid-confidence interactions) and PPIs were obtained from BioGRID 3.5 [24]. In summary, out of 410,961 interactions from the human interactome 52,467 have at least one domain interaction. The linear motif instances and their interactions were downloaded from the ELM website and mapped to the respective exons. We found 3926 PPIs that are confirmed by at least one source of DMI. Position-specific PPI based on experimentally resolved structure from the PDB was obtained from [21]. In total, 16,161 PPIs were enriched by at least one residue-level interaction. From the combination of all these resources the final structurally annotated graph contained in total 60,235 interactions. Each one of these interactions is annotated with one or multiple levels of evidence (DDIs, DMI, residues). The mapping of exons to their protein features was performed using the Biomart mapping table, Pfam, and ELM annotations [22, 23, 108]. We obtain the biological pathways with their gene list from KEGG [57] and Reactome [101] integrated into the ConsensusPathDB database [28].

Statistical tests and pathway scores

Gene-level enrichment is performed using a hypergeometric test from the package GSEAPY (a Python wrapper for Enrichr [109]) by considering all genes with (differential) AS events. Network enrichment analysis at the gene level was performed using the EviNet web server (www.evinet.org), which is an implementation of the randomization algorithm NEA [32]. To achieve a fair comparison with NEASE, we run NEA using BioGRID as a PPI database and Reactome and KEGG as pathways references to match the exact conditions of the NEASE analysis.

For NEASE enrichment, we filtered the PPI graph $G=(V, E)$, where V is the set of genes and E is the set of edges, to a subgraph $G'=(V', E')$ containing only structurally annotated interactions E' and their nodes V'. An interaction is considered structurally annotated if it is supported by at least one of these resources: domain-domain interactions, motif-domain interactions, or residue-level interactions. For a submitted query list of exons, NEASE first identifies affected domains, linear motifs, and residues that overlap with the exons and their interactions. Let N be two times the number of edges

in G' (the degree of the network) and n be the number of affected edges from the query. These edges are then considered using a test modified from [110]. For every pathway P with degree K, let k be the number of affected edges that are connected to P. We model X whose outcome is k as a random variable following a hypergeometric distribution:

$$\mathrm{X} \sim \mathrm{Hypergeometric}(\mathrm{n} = \text{number of affected edges}, \mathrm{K} = \text{degree of P}, \mathrm{N} = \text{degree of G}')$$

where k is considered as the number of observed successes out of n draws, from a population of size N containing K success. Subsequently, NEASE tests if the number k is significant using a one-sided hypergeometric test (over-representation). In contrast to the test proposed in [110], our test only includes structurally annotated edges and the ones likely to be impacted by AS in order to improve the signal-to-noise ratio. For illustration purposes, in the example of Fig. 1B, the overall number of affected edges by AS is n=7, and K=11 is the total degree of pathway A (11 possible success), the number of affected edges that are linked to the pathway is k=4. The enrichment p value of pathway A corresponds then to the significance of this last number. After testing for multiple pathways, the obtained p values for the edge-level enrichment are corrected, using the Benjamini-Hochberg method [111]. The detailed pseudocode of this algorithm is explained in Additional file 1: Figure S6, Algorithm 1.

For a pathway of interest, a similar test can be applied to determine if a splicing event significantly affects interactions of a specific gene with this pathway (Additional file 1: Figure S6, Algorithm 2 and Fig. 1B, C). Here, n is the number of all affected interactions (edges) of a spliced gene and k is the number of affected interactions (edges) across genes that are linked to the pathway of interest. In the example of Fig. 1B, C, the gene G2 can be connected to pathway A just by chance due to its high number of affected interactions. For this reason, it is ranked lower than the genes G3 and G4.

As a result, for every pathway, NEASE provides an overall p value, as well as the most significant genes. Since the p value only depends on the overall number of affected edges but not on the number of genes, the p value can be heavily influenced by hub genes. To reduce this influence, an optional score (NEASE Score, Eq.: 1) can be computed by NEASE to scale the natural logarithm of the p value with the total number of significant genes using a cutoff from the user (for instance p value ≤ 0.05):

$$\text{NEASE score} = -\sqrt{g} \times \log_{10}(p\,\text{value}) \quad (1)$$

where g is the total number of significantly connected genes obtained after testing individual spliced genes. Thus, the NEASE Score prioritizes pathways that are affected by a larger number of spliced genes rather than pathways that have a larger number of affected interactions (edges). The user can choose to rank enrichment based on the adjusted p value or by the NEASE score.

For permutation tests, the set of highly confident 2831 exon skipping events obtained after the initial quality filtering is considered as the background set of exons. We compared the enrichment obtained in the pathways “Muscle Contraction” and “Synaptic vesicle cycle” from the set of exons upregulated in muscles/heart and neural tissues respectively, with 10,000 random sets of exons of the same size and then derived distribution of p values. The empirical p values were then obtained by asking how likely it is to

obtain a p value as low or more extreme than the one reported by NEASE in the original set of neural and muscles upregulated exons.

VastDB events processing

PSI values of the exon skipping events from VastDB were quantified by the developers using vast tools [30, 112]. In our analysis, we extracted the PSI values for 32 experiments belonging to 12 main tissues: muscles/heart, neural (whole brain, cortex, and peripheral retina), placental, epithelial, digestive (colon and stomach), liver, kidney, adipose, testis, immune-hematopoietic, and ovary. We then filtered out the events with low read coverage (VLOW) and performed hierarchical clustering of standardized values (z scores). For every exon, we calculated the mean of PSI values from the samples of the same tissues. To extract muscles/heart and neural-specific exons and to ensure that we only consider functional events, we applied two filters: namely that the exon PSI value in the relevant tissue is higher than 20 and that the z score is higher than 2.

RNA-Seq analysis

Raw RNA-Seq reads for two types of platelets and multiple sclerosis patients were downloaded from the GEO repository (access numbers: GSE126448 and GSE138614). The number of samples and sequencing depth are reported in Additional file 1: Table S3. RNA-Seq reads were aligned to the reference human genome (hg38) using STAR 2.7 [113] in a 2-pass mode and filtered for uniquely mapped reads. Differential AS analysis was performed by MAJIQ [56] with default parameters, and with a threshold of P(dPSI > 20%) > 0.95.

NEASE: The Python package

NEASE's Python package relies on NumPy [114], pandas [115], NetworkX [116], SciPy [117], and Statsmodels [118]. The gene-level enrichment is also supported in the NEASE package using the Python implementation of Enrichr [109]. To speed up the edge hypergeometric test, the total degree of every pathway in the structural PPI, as well as the overall degree of the network were pre-computed. For visualization, we use the complete PPI (not the structural PPI) and extract connected subnetworks from each pathway as well as spliced genes and their interactions with the extracted modules. The position of nodes is computed using the Fruchterman-Reingold force-directed algorithm implemented in NetworkX [119]. The interactive visualization for individual genes and events is implemented with information from the DIGGER database and the Plotly package.

The package provides the option to automatically filter exons that are likely to disturb the open reading frame of the transcript based on the prediction in [30]. In the case of multiple AS events affecting the same genes, we consider every event individually and identify all protein features. The standard input of the package is a DataFrame object with the exon coordinates and Ensembl IDs of the genes. The package also supports the output of multiple AS differential detection tools such as rMATs [120], Whippet [121], and also tools that are event-based such as MAJIQ [56] where NEASE only considers annotated exons. NEASE is released as open-source under the GPLv3 license and is available at (https://github.com/louadi/NEASE). Step-by-step tutorials for running NEASE are available at (https://github.com/louadi/NEASE-tutorials).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02538-1.

Additional file 1: Tables S1-S3. Table S1. Enrichment of the pathway "Muscle contraction" from Reactome for the exons upregulated in the muscles generated by the NEASE package. **Table S2.** Enrichment of the pathway "Synaptic vesicle cycle" from KEGG for the exons upregulated in the neural tissues generated by the NEASE package. **Table S3.** RNA-Seq samples used in the study. **Figs. S1-S6. Fig. S1.** Node degree distribution of the classic PPI and structurally annotated PPI, the latter contains only interactions with evidence from DDIs and DMIs or residue-level evidence from the co-resolved structure. **Fig S2.** Network Enrichment Analysis using EviNet webtool for exons upregulated in muscles and neural tissues. **Fig. S3.** NEASE visualization highlights the interactions of differentially spliced genes with the DCM pathway. **Fig. S4.** The PSI values of two exon skipping events in the genes TPM1 and DST from the GTEx dataset confirm that both the exons are upregulated in muscles and heart tissues. **Fig. S5.** The PSI values of two exon skipping events in the genes CLTA and CLTB from the GTEx dataset confirm that both the exons are upregulated in neural tissues. **Fig. S6.** Pseudocode of NEASE algorithm.

Additional file 2: Tables S4-S7. The analysis of the upregulated exons in muscles. **Table S4:** List of spliced domains. **Table S5:** List of affected edges. **Table S6:** Gene-level enrichment. **Table S7:** NEASE enrichment.

Additional file 3: Tables S8-S11. The analysis of the upregulated exons in neural. **Table S8.** List of spliced domains. **Table S9.** List of affected edges. **Table S10.** Gene-level enrichment. **Table S11.** NEASE enrichment.

Additional file 4: Tables S12-S16. Differential splicing analysis between reticulated platelets and mature platelets. **Table S12.** List of spliced domains. **Table S13.** List of affected edges. **Table S14.** Gene-level enrichment. **Table S15.** NEASE enrichment. **Table S16.** Enrichment of the pathway "GPCR downstream signal".

Additional file 5: Tables S17-S21. Differential splicing analysis between normal-appearing white matter and acute lesion from multiple sclerosis patients. **Table S17.** List of spliced domains. **Table S18.** List of affected edges. **Table S19.** Gene-level enrichment. **Table S20.** NEASE enrichment. **Table S21.** Enrichment of the pathway "Neurotransmitter receptors and postsynaptic signal transmission".

Additional file 6: Tables S22-S24. Differential splicing analysis between Dilated Cardiomyopathy patients and controls. **Table S22.** List of spliced domains. **Table S23.** List of affected edges. **Table S24.** Gene-level enrichment. **Table S25.** NEASE enrichment. **Table S26.** Enrichment of the pathway "Dilated cardiomyopathy".

Additional file 7: Table S27. The PSI values for VastDB exons in different tissues.

Additional file 8. Review history.

Acknowledgements

Not applicable.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 8.

Authors' contributions

ZL, TK, OT, JB, and ML conceived the project. ZL performed the initial experiments, developed the method, and implemented the software. OT and ML supervised the project, provided critical feedback, and helped shape the research. OT prepared the RNA-Seq datasets for the differential splicing analysis. AF, CTL interpreted and discussed the biological insights of the results. MK and DB interpreted and discussed the platelet analysis. MLE and ZI interpreted and discussed the Multiple Sclerosis analysis. ZL wrote the initial draft of the manuscript. All authors contributed to writing the final manuscript and approved the final version.

Authors' information

Twitters: @ZakariaLouadi (Zakaria Louadi); @itisalist (Markus List); @janbaumbach (Jan Baumbach); @KacprowskiTim (Tim Kacprowski); @meeklug (Melissa Klug)

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept [grant 01ZX1908A (Sys_CARE)]. JB was partially funded by his VILLUM Young Investigator Grant (nr.13154). MLE is grateful for financial support from Lundbeckfonden (no. R347-2020-2454). ZI is grateful for financial support from Scleroseforeningen (no. A29926, A 31829, A33600). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

RNA sequencing data for reticulated platelets was provided by the authors [49] and it is freely available at GEO (access number: GSE126448). Multiple sclerosis Raw sequence was provided by the authors [69] and freely available at GEO (access number: GSE138614). Dilated Cardiomyopathy raw data is available in the European Genome-phenome Archive (Dataset ID: EGAS00001002454), in our analysis, we used pre-processed data from the manuscript [10]. VastDB dataset for humans (hg19) was downloaded from https://vastdb.crg.eu/wiki/Downloads. The linear motifs instances and interactions were downloaded from (http://elm.eu.org/). The generated joint graphs and the exon mapping databases are available on the DIGGER database website https://exbio.wzw.tum.de/digger/download. NEASE is released as open-

source under the GPLv3 license and is available at GitHub [122] and deposited to Zenodo [123]. All processed datasets, as well as step-by-step tutorials for using NEASE to reproduce the results presented in this paper, are available at https://github.com/louadi/NEASE-tutorials and deposited to Zenodo [124].

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany. [2]Institute for Computational Systems Biology, University of Hamburg, Notkestrasse 9, 22607 Hamburg, Germany. [3]Department of Neurology, Odense University Hospital, Odense, Denmark. [4]Institute of Clinical Research, University of Southern Denmark, Odense, Denmark. [5]Institute of Molecular Medicine, University of Southern Denmark, Odense, Denmark. [6]Department of Internal Medicine I, School of Medicine, University hospital rechts der Isar, Technical University of Munich, Munich, Germany. [7]German Center for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Munich, Germany. [8]Department of Cardiovascular Medicine, Humanitas Clinical and Research Center IRCCS and Humanitas University, Rozzano, Milan, Italy. [9]Institute of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5000 Odense, Denmark. [10]Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of Technische Universität Braunschweig and Hannover Medical School, Braunschweig, Germany. [11]Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany.

Received: 15 July 2021 Accepted: 10 November 2021
Published online: 02 December 2021

References
1. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, et al. Function of alternative splicing. Gene. 2005;344:1–20. https://doi.org/10.1016/j.gene.2004.10.022.
2. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. Genome Biol. 2004;5(10): R74. https://doi.org/10.1186/gb-2004-5-10-r74.
3. Baralle F, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017; 18(7) Available from:. https://doi.org/10.1038/nrm.2017.27.
4. Beqqali A. Alternative splicing in cardiomyopathy. Biophys Rev. 2018;10(4):1061–71.
5. Douglas AGL, Wood MJA. Splicing therapy for neuromuscular disease. Mol Cell Neurosci. 2013;56:169–85. https://doi.org/10.1016/j.mcn.2013.04.005.
6. Evsyukova I, Somarelli JA, Gregory SG, Garcia-Blanco MA. Alternative splicing in multiple sclerosis and other 673 autoimmune diseases. RNA Biology. 2010;7(4)462–73. https://doi.org/10.4161/rna.7.4.12301.
7. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 2005;579(9):1900–3. https://doi.org/10.1016/j.febslet.2005.02.047.
8. Karlebach G, Veiga DFT, Mays AD, Chatzipantsiou C, Barja PP, Chatzou M, et al. The impact of biological sex on alternative splicing. bioRxiv. 2020:490904. https://doi.org/10.1101/490904.
9. Tollervey JR, Wang Z, Hortobágyi T, Witten JT, Zarnack K, Kayikci M, et al. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. Genome Res. 2011;21(10):1572–82. https://doi.org/10.1101/gr.122226.111.
10. Heinig M, Adriaens ME, Schafer S, van Deutekom HWM, Lodder EM, Ware JS, et al. Natural genetic variation of the cardiac transcriptome in non-diseased donors and patients with dilated cardiomyopathy. Genome Biol. 2017;18(1):170. https://doi.org/10.1186/s13059-017-1286-z.
11. Tress ML, Abascal F, Valencia A. Alternative Splicing May Not Be the Key to Proteome Complexity. Trends Biochem Sci. Elsevier Ltd. 2017;42(2):98–110. Available from: https://pubmed.ncbi.nlm.nih.gov/27712956/. https://doi.org/10.1016/j.tibs.2016.08.008.
12. Melamud E, Moult J. Stochastic noise in splicing machinery. Nucleic Acids Res. 2009;37(14):4873–86. https://doi.org/10.1093/nar/gkp471.
13. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell. 2016;164(4):805–17. https://doi.org/10.1016/j.cell.2016.01.029.
14. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell. 2012;46(6):871–83.
15. da Costa PJ, Menezes J, Romão L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. Int J Biochem Cell Biol. 2017;91(Pt B):168–75.
16. Kristoffer V-S. Sandelin A. Genomics: The Landscape of Isoform Switches in Human Cancers; 2017; Available from:. https://doi.org/10.1158/1541-7786.MCR-16-0459.
17. delafuente L, Arzalluz-luque Á, Tardáguila M, Delrisco H, Martí C, Tarazona S, et al. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biol. 2020;21(1) Available from:. https://doi.org/10.1186/s13059-020-02028-w.
18. Gal-Oz ST, Haiat N, Eliyahu D, Shani G, Shay T. DoChaP: the domain change presenter. Nucleic Acids Res. 2021;49(W1): W162–8. https://doi.org/10.1093/nar/gkab357.

19. Ctor Climente-Gonzá Lez H, Porta-Pardo E, Godzik A, Correspondence EE, Eyras E. The Functional Impact of Alternative Splicing in Cancer. Cell Rep. 2017;20:2215–26.
20. Tranchevent L-C, Aubé F, Dulaurier L, Benoit-Pilven C, Rey A, Poret A, et al. Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res. 2017;27(6):1087–97.
21. Louadi Z, Yuan K, Gress A, Tsoy O, Kalinina OV, Baumbach J, et al. DIGGER: exploring the functional role of alternative 709 splicing in protein interactions. Nucleic Acids Res. 2020;49(D1):D309-D318. https://doi.org/10.1093/nar/gkaa768.
22. Kumar M, Gouw M, Michael S. Amano-S ´ Anchez HS´, Pancsa R, Glavina J, et al. ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020;48. Available from: https://academic.oup.com/nar/article/48/D1/D296/5611669. https://doi.org/10.1093/nar/gkz1030.
23. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32. https://doi.org/10.1093/nar/gky995.
24. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):D529–41. https://doi.org/10.1093/nar/gky1079.
25. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res. 2011;39(Database issue):D730–5. https://doi.org/10.1093/nar/gkq1229.
26. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014;42(Database issue):D374–9. https://doi.org/10.1093/nar/gkt887.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235.
28. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013;41(Database issue):D793–800.
29. Ellis JD, Barrios-Rodiles M, Çolak R, Irimia M, Kim T, Calarco JA, et al. Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. Mol Cell. 2012;46(6):884–92.
30. Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res. 2017;27(10):1759–68.
31. Seo PJ, Kim MJ, Ryu J-Y, Jeong E-Y, Park C-M. Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. Nat Commun. 2011;2(1):303. https://doi.org/10.1038/ncomms1303.
32. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics. 2012;13(1):226. https://doi.org/10.1186/1471-2105-13-226.
33. Tansey MG, Luby-Phelps K, Kamm KE, Stull JT. Ca(2+)-dependent phosphorylation of myosin light chain kinase decreases the Ca2+ sensitivity of light chain phosphorylation within smooth muscle cells. J Biol Chem. 1994;269(13):9912–20.
34. Hall CN, Klein-Flügge MC, Howarth C, Attwell D. Oxidative phosphorylation, not glycolysis, powers presynaptic and postsynaptic mechanisms underlying brain information processing. J Neurosci. 2012;32(26):8940–51. https://doi.org/10.1523/JNEUROSCI.0026-12.2012.
35. Sanganahalli BG, Herman P, Blumenfeld H, Hyder F. Oxidative neuroenergetics in event-related paradigms. J Neurosci. 2009;29(6):1707–18. https://doi.org/10.1523/JNEUROSCI.5549-08.2009.
36. Vergara RC, Jaramillo-Riveri S, Luarte A, Moënne-Loccoz C, Fuentes R, Couve A, et al. The Energy Homeostasis Principle: Neuronal Energy Regulation Drives Local Network Dynamics Generating Behavior. Front Comput Neurosci. 2019;13:49. https://doi.org/10.3389/fncom.2019.00049.
37. Du F, Zhu X-H, Zhang Y, Friedman M, Zhang N, Ugurbil K, et al. Tightly coupled brain activity and cerebral ATP metabolic rate. Proc Natl Acad Sci U S A. 2008;105(17):6409–14. https://doi.org/10.1073/pnas.0710766105.
38. Howarth C, Gleeson P, Attwell D. Updated energy budgets for neural computation in the neocortex and cerebellum. J Cereb Blood Flow Metab. 2012;32(7):1222–32. https://doi.org/10.1038/jcbfm.2012.35.
39. Magistretti PJ, Allaman I. A cellular perspective on brain energy metabolism and functional imaging. Neuron. 2015;86(4):883–901. https://doi.org/10.1016/j.neuron.2015.03.035.
40. Zheng X, Boyer L, Jin M, Mertens J, Kim Y, Ma L, et al. Metabolic reprogramming during neuronal differentiation from aerobic glycolysis to neuronal oxidative phosphorylation. Elife. 2016;10:5. Available from:. https://doi.org/10.7554/eLife.13374.
41. Hiesinger PR, Fayyazuddin A, Mehta SQ, Rosenmund T, Schulze KL, Zhai RG, et al. The v-ATPase V0 subunit a1 is required for a late step in synaptic vesicle exocytosis in Drosophila. Cell. 2005;121(4):607–20. https://doi.org/10.1016/j.cell.2005.03.012.
42. Aoto K, Kato M, Akita T, Nakashima M, Mutoh H, Akasaka N, et al. ATP6V0A1 encoding the a1-subunit of the V0 domain of vacuolar H+-ATPases is essential for brain development in humans and mice. Nat Commun. 2021;12(1):2107. https://doi.org/10.1038/s41467-021-22389-5.
43. Poëa-Guyon S, Amar M, Fossier P, Morel N. Alternative splicing controls neuronal expression of v-ATPase subunit a1 and sorting to nerve terminals. J Biol Chem. 2006;281(25):17164–72. https://doi.org/10.1074/jbc.M600927200.
44. Redlingshöfer L, McLeod F, Chen Y, Camus MD, Burden JJ, Palomer E, et al. Clathrin light chain diversity regulates membrane deformation in vitro and synaptic vesicle formation in vivo. Proc Natl Acad Sci U S A. 2020;117(38):23527–38.
45. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5. https://doi.org/10.1038/ng.2653.
46. Rodriguez JM, Pozo F, di Domenico T, Vazquez J, Tress ML. An analysis of tissue-specific alternative splicing at the protein level. Orengo CA, editor. PLoS Comput Biol. 2020;16(10):e1008287.
47. Raj B, Blencowe BJ. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron. 2015;87(1):14–27. https://doi.org/10.1016/j.neuron.2015.05.004.

48. Su C-H, Dhananjaya D, Tarn W-Y. Alternative Splicing in Neurogenesis and Brain Development. Front Mol Biosci. 2018;5:12.
49. Bongiovanni D, Santamaria G, Klug M, Santovito D, Felicetta A, Hristov M, et al. Transcriptome Analysis of Reticulated Platelets Reveals a Prothrombotic Profile. Thromb Haemost. 2019;119(11):1795–806. https://doi.org/10.1055/s-0039-1695009.
50. Ault KA, Knowles C. In vivo biotinylation demonstrates that reticulated platelets are the youngest platelets in circulation. Exp Hematol. 1995;23(9):996–1001.
51. Karpatkin S. Heterogeneity of human platelets. II. Functional evidence suggestive of young and old platelets. J Clin Invest. 1969;48(6):1083–7.
52. Cesari F, Marcucci R, Gori AM, Caporale R, Fanelli A, Casola G, et al. Reticulated platelets predict cardiovascular death in acute coronary syndrome patients. Thromb Haemost. 2013;109(05):846–53. https://doi.org/10.1160/TH12-09-0709.
53. Guthikonda S, Alviar CL, Vaduganathan M, Arikan M, Tellez A, DeLao T, et al. Role of reticulated platelets and platelet size heterogeneity on platelet activity after dual antiplatelet therapy with aspirin and clopidogrel in patients with stable coronary artery disease. J Am Coll Cardiol. 2008;52(9):743–9. https://doi.org/10.1016/j.jacc.2008.05.031.
54. Muronoi T, Koyama K, Nunomiya S, Lefor AK, Wada M, Koinuma T, et al. Immature platelet fraction predicts coagulopathy-related platelet consumption and mortality in patients with sepsis. Thromb Res. 2016;144:169–75. https://doi.org/10.1016/j.thromres.2016.06.002.
55. Nassa G, Giurato G, Cimmino G, Rizzo F, Ravo M, Salvati A, et al. Splicing of platelet resident pre-mRNAs upon activation by physiological stimuli results in functionally relevant proteome modifications. Sci Rep. 2018;8(1):498. https://doi.org/10.1038/s41598-017-18985-5.
56. Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife. 2016;5:e11752.
57. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55. https://doi.org/10.1093/nar/gkx1132.
58. Scheer FAJL, Michelson AD, Frelinger AL 3rd, Evoniuk H, Kelly EE, McCarthy M, et al. The human endogenous circadian system causes greatest platelet activation during the biological morning independent of behaviors. PLoS One. 2011;6(9):e24549.
59. Offermanns S. Activation of Platelet Function Through G Protein–Coupled Receptors. Circ Res. 2006;99(12):1293–304. https://doi.org/10.1161/01.RES.0000251742.71301.16.
60. Marti-Solano M, Crilly SE, Malinverni D, Munk C, Harris M, Pearce A, et al. Combinatorial expression of GPCR isoforms affects signalling and drug responses. Nature. 2020;587(7835):650–6. https://doi.org/10.1038/s41586-020-2888-2.
61. Jalagadugula G, Dhanasekaran DN, Kim S, Kunapuli SP, Rao AK. Early growth response transcription factor EGR-1 regulates Galphaq gene in megakaryocytic cells. J Thromb Haemost. 2006;4(12):2678–86.
62. Moore SF, van den Bosch MTJ, Hunter RW, Sakamoto K, Poole AW, Hers I. Dual regulation of glycogen synthase kinase 3 (GSK3)α/β by protein kinase C (PKC)α and Akt promotes thrombin-mediated integrin αIIbβ3 activation and granule secretion in platelets. J Biol Chem. 2013;288(6):3918–28. https://doi.org/10.1074/jbc.M112.429936.
63. Harper MT, Poole AW. Diverse functions of protein kinase C isoforms in platelet activation and thrombus formation. J Thromb Haemost. 2010;8(3):454–62. https://doi.org/10.1111/j.1538-7836.2009.03722.x.
64. Williams CM, Harper MT, Poole AW. PKCα negatively regulates in vitro proplatelet formation and in vivo platelet production in mice. Platelets. 2014;25(1):62–8.
65. Ault KA, Rinder HM, Mitchell J, Carmody MB, Vary CP, Hillman RS. The significance of platelets with increased RNA content (reticulated platelets). A measure of the rate of thrombopoiesis. Am J Clin Pathol. 1992;98(6):637–46. https://doi.org/10.1093/ajcp/98.6.637.
66. Bö L, Dawson TM, Wesselingh S, Mörk S, Choi S, Kong PA, et al. Induction of nitric oxide synthase in demyelinating regions of multiple sclerosis brains. Ann Neurol. 1994;36(5):778–86. https://doi.org/10.1002/ana.410360515.
67. Ludwin SK. The pathogenesis of multiple sclerosis: relating human pathology to experimental studies. J Neuropathol Exp Neurol. 2006;65(4):305–18. https://doi.org/10.1097/01.jnen.0000225024.12074.80.
68. Hecker M, Rüge A, Putscher E, Boxberger N, Rommer PS, Fitzner B, et al. Aberrant expression of alternative splicing variants in multiple sclerosis - A systematic review. Autoimmun Rev. 2019;18(7):721–32. https://doi.org/10.1016/j.autrev.2019.05.010.
69. Elkjaer ML, Frisch T, Reynolds R, Kacprowski T, Burton M, Kruse TA, et al. Molecular signature of different lesion types in the brain white matter of patients with progressive multiple sclerosis. Acta Neuropathol Commun. 2019;7(1):205. https://doi.org/10.1186/s40478-019-0855-7.
70. Gissel H. Ca2+ accumulation and cell damage in skeletal muscle during low frequency stimulation. Eur J Appl Physiol. 2000;83(2-3):175–80. https://doi.org/10.1007/s004210000276.
71. Maléth J, Hegyi P. Ca2+ toxicity and mitochondrial damage in acute pancreatitis: translational overview. Philos Trans R Soc Lond B Biol Sci. 2016;5(1700):371(1700). Available from:. https://doi.org/10.1098/rstb.2015.0425.
72. Minagar A, Alexander JS. Blood-brain barrier disruption in multiple sclerosis. Mult Scler. 2003;9(6):540–9. https://doi.org/10.1191/1352458503ms965oa.
73. Claudio L, Raine CS, Brosnan CF. Evidence of persistent blood-brain barrier abnormalities in chronic-progressive multiple sclerosis. Acta Neuropathol. 1995;90(3):228–38.
74. Ortiz GG, Pacheco-Moisés FP, Macías-Islas MÁ, Flores-Alvarado LJ, Mireles-Ramírez MA, González-Renovato ED, et al. Role of the blood-brain barrier in multiple sclerosis. Arch Med Res. 2014;45(8):687–97.
75. Ascherio A, Munger KL. Environmental risk factors for multiple sclerosis. Part I: the role of infection. Ann Neurol. 2007;61(4):288–99.
76. Baranzini SE, Srinivasan R, Khankhanian P, Okuda DT, Nelson SJ, Matthews PM, et al. Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis. Brain. 2010;133(9):2603–11. https://doi.org/10.1093/brain/awq192.
77. Strijbis EMM, Inkster B, Vounou M, Naegelin Y, Kappos L, Radue E-W, et al. Glutamate gene polymorphisms predict brain volumes in multiple sclerosis. Mult Scler. 2013;19(3):281–8. https://doi.org/10.1177/1352458512454345.

78. Wang JH, Pappas D, De Jager PL, Pelletier D, de Bakker PI, Kappos L, et al. Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data. Genome Med. 2011;3(1):3. https://doi.org/10.1186/gm217.
79. Keshet Y, Seger R. The MAP kinase signaling cascades: a system of hundreds of components regulates a diverse array of physiological functions. Methods Mol Biol. 2010;661:3–38. https://doi.org/10.1007/978-1-60761-795-2_1.
80. Chang L, Karin M. Mammalian MAP kinase signalling cascades. Nature. 2001;410(6824):37–40. https://doi.org/10.1038/35065000.
81. Shchetynsky K, Protsyuk D, Ronninger M, Diaz-Gallo L-M, Klareskog L, Padyukov L. Gene-gene interaction and RNA splicing profiles of MAP2K4 gene in rheumatoid arthritis. Clin Immunol. 2015;158(1):19–28.
82. Tuller T, Atar S, Ruppin E, Gurevich M, Achiron A. Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases. Genes Immun. 2013;14(2):67–82.
83. GJA t B, Bolk J, `t Hart BA, Laman JD. Multiple sclerosis is linked to MAPKERK overactivity in microglia. J Mol Med. 2021; Available from:. https://doi.org/10.1007/s00109-021-02080-4.
84. Mass E, Jacome-Galarza CE, Blank T, Lazarov T, Durham BH, Ozkaya N, et al. A somatic mutation in erythro-myeloid progenitors causes neurodegenerative disease. Nature. 2017;549(7672):389–93. https://doi.org/10.1038/nature23672.
85. Kotelnikova E, Kiani NA, Messinis D, Pertsovskaya I, Pliaka V, Bernardo-Faura M, et al. MAPK pathway and B cells overactivation in multiple sclerosis revealed by phosphoproteomics and genomic analysis. Proc Natl Acad Sci U S A. 2019;116(19):9671–6. https://doi.org/10.1073/pnas.1818347116.
86. Krementsov DN, Thornton TM, Teuscher C, Rincon M. The emerging role of p38 mitogen-activated protein kinase in multiple sclerosis and its models. Mol Cell Biol. 2013;33(19):3728–34.
87. Bernardo-Faura M, Rinas M, Wirbel J, Pertsovskaya I, Pliaka V, Messinis DE, Vila G, Sakellaropoulos T, Faigle W, Stridh P, Behrens JR. Prediction of combination therapies based on topological modeling of the immune signaling network in Multiple Sclerosis. Genome Medicine. 2021;13(1):1-6.
88. Maatz H, Jens M, Liss M, Schafer S, Heinig M, Kirchner M, et al. RNA-binding protein RBM20 represses splicing to orchestrate cardiac pre-mRNA processing. J Clin Invest. 2014;124(8):3419–30. https://doi.org/10.1172/JCI74523.
89. Sheikh F, Lyon RC, Chen J. Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease. Gene. 2015;569(1):14–20.
90. Matyushenko AM, Levitsky DI. Molecular Mechanisms of Pathologies of Skeletal and Cardiac Muscles Caused by Point Mutations in the Tropomyosin Genes. Biochemistry. 2020;85(Suppl 1):S20–33.
91. Caleshu C, Sakhuja R, Nussbaum RL, Schiller NB, Ursell PC, Eng C, et al. Furthering the link between the sarcomere and primary cardiomyopathies: restrictive cardiomyopathy associated with multiple mutations in genes previously associated with hypertrophic or dilated cardiomyopathy. Am J Med Genet A. 2011;155A(9):2229–35. https://doi.org/10.1002/ajmg.a.34097.
92. Brody MJ, Hacker TA, Patel JR, Feng L, Sadoshima J, Tevosian SG, et al. Ablation of the cardiac-specific gene leucine-rich repeat containing 10 (Lrrc10) results in dilated cardiomyopathy. PLoS One. 2012;7(12):e51621.
93. Gupte TM, Haque F, Gangadharan B, Sunitha MS, Mukherjee S, Anandhan S, et al. Mechanistic Heterogeneity in Contractile Properties of α-Tropomyosin (TPM1) Mutants Associated with Inherited Cardiomyopathies*. J Biol Chem. 2015;290(11):7003–15.
94. Huang W, Liang J, Yuan C-C, Kazmierczak K, Zhou Z, Morales A, et al. Novel familial dilated cardiomyopathy mutation in MYL2 affects the structure and function of myosin regulatory light chain. FEBS J. 2015;282(12):2379–93.
95. Herman DS, Lam L, Taylor MRG, Wang L, Teekakirikul P, Christodoulou D, et al. Truncations of titin causing dilated cardiomyopathy. N Engl J Med. 2012;366(7):619–28. https://doi.org/10.1056/NEJMoa1110186.
96. Marston S, Montgiraud C, Munster AB, Copeland O, Choi O, Dos Remedios C, et al. OBSCN Mutations Associated with Dilated Cardiomyopathy and Haploinsufficiency. PLoS One. 2015;10(9):e0138568.
97. Marston S. Obscurin variants and inherited cardiomyopathies. Biophys Rev. 2017;9(3):239–43. https://doi.org/10.1007/s12551-017-0264-8.
98. McNally EM, Mestroni L. Dilated Cardiomyopathy: Genetic Determinants and Mechanisms. Circ Res. 2017;121(7):731–48. https://doi.org/10.1161/CIRCRESAHA.116.309396.
99. Boczek NJ, Ye D, Jin F, Tester DJ, Huseby A, Bos JM, et al. Identification and Functional Characterization of a Novel CACN A1C-Mediated Cardiac Disorder Characterized by Prolonged QT Intervals With Hypertrophic Cardiomyopathy, Congenital Heart Defects, and Sudden Cardiac Death. Circ Arrhythm Electrophysiol. 2015;8(5):1122–32.
100. Mouton J, Ronjat M, Jona I, Villaz M, Feltz A, Maulet Y. Skeletal and cardiac ryanodine receptors bind to the Ca(2+)-sensor region of dihydropyridine receptor alpha(1C) subunit. FEBS Lett. 2001;505(3):441–4. https://doi.org/10.1016/S0014-5793(01)02866-6.
101. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
102. Ather S, Respress JL, Li N, Wehrens XHT. Alterations in ryanodine receptors and related proteins in heart failure. Biochim Biophys Acta. 2013;1832(12):2425–31. https://doi.org/10.1016/j.bbadis.2013.06.008.
103. Yano M, Yamamoto T, Kobayashi S, Matsuzaki M. Role of ryanodine receptor as a Ca2+ regulatory center in normal and failing hearts. J Cardiol. 2009;53(1):1–7. https://doi.org/10.1016/j.jjcc.2008.10.008.
104. Moccia F, Lodola F, Stadiotti I, Pilato CA, Bellin M, Carugo S, et al. Calcium as a Key Player in Arrhythmogenic Cardiomyopathy: Adhesion Disorder or Intracellular Alteration? Int J Mol Sci. 2019;16(16):20(16). Available from:. https://doi.org/10.3390/ijms20163986.
105. Jaffrey SR, Wilkinson MF. Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. Nat Rev Neurosci. 2018;19(12):715–28.
106. Schwerk C, Schulze-Osthoff K. Regulation of apoptosis by alternative pre-mRNA splicing. Mol Cell. 2005;19(1):1–13.
107. List M, Alcaraz N, Dissing-Hansen M, Ditzel HJ, Mollenhauer J, Baumbach J. KeyPathwayMinerWeb: online multi-omics network enrichment. Nucleic Acids Res. 2016;44(W1):W98–104.
108. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011;2011:bar030.
109. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7. https://doi.org/10.1093/nar/gkw377.

110. Signorelli M, Vinciotti V, Wit EC. NEAT: an efficient network enrichment analysis test. BMC Bioinformatics. 2016;17(1):352. https://doi.org/10.1186/s12859-016-1203-6.
111. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Stat Soc: Ser B (Methodological). 1995;57:289–300. Available from:. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
112. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159(7):1511–23. https://doi.org/10.1016/j.cell.2014.11.035.
113. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
114. Van Der Walt S, Chris Colbert S, Varoquaux G. The NumPy array: a structure for efficient numerical computation. arXiv [cs.MS]. 2011; Available from: http://arxiv.org/abs/1102.1523.
115. McKinney W, Others. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference. 2010;445:51–6. https://doi.org/10.25080/Majora-92bf1922-00a.
116. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using networkx. Proceedings of the 7th Python in Science Conference (SciPy2008). 2008;11-15.
117. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.
118. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference. https://doi.org/10.25080/Majora-92bf1922-011.
119. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Softw Pract Exp. 1991;21:1129–64. Available from:. https://doi.org/10.1002/spe.4380211102.
120. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111(51):E5593–601.
121. Sterne-Weiler T, Weatheritt RJ, Best A, Ha KCH. Whippet: an efficient method for the detection and quantification of 944 alternative splicing reveals extensive transcriptomic complexity. bioRxiv. 2017. https://doi.org/10.1101/158519
122. Louadi Z. NEASE: A network-based approach for the enrichment of alternative splicing events. Github. 2021. Available from: https://github.com/louadi/NEASE. Accessed 22 Nov 2021.
123. Louadi Z. NEASE: v.1.1.6. Zenodo; 2021. Available from:. https://doi.org/10.5281/zenodo.5653490.
124. Louadi Z. NEASE-tutorials: v1.2. Zenodo; 2021. Available from: https://doi.org/10.5281/ZENODO.5562626

Publisher's Note

Appendix C

Appendix: Teaching and Supervision Record

- Teaching Assistant for Systems BioMedicine class in TUM/LMU
- Teaching Assistant for Advanced Bioinformatics class in TUM/LMU
- Master's thesis supervision: "Methods forinferring patient-specific dysregulated networks using regression." Johannes Josef Kersting.
- Bachelor's thesis supervision: "Predicting tissue-specific splicing factor expression from whole blood expression.", Ningyue Zhou.
- Bachelor's thesis supervision: "Predicting alternative splicing in non-accessible tissues using machine learning.", Deniz Enes Hasler.
- Bachelor's thesis supervision: "Systematic identification of rare splicing events in a disease context." Anna Schuster.

List of Figures

LIST OF FIGURES

List of Tables