UNIVERSITÄT
DES
SAARLANDES

TRANSCRIPTOMIC AND PROTEOMIC REWIRING IN TISSUE-SPECIFIC REGULATION

Dissertation
zur Erlangung des Grades
der Doktorin der Naturwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes
von

Hoang Thu Trang Do

Saarbrücken, 2025

Tag des Kolloquiums	*09.01.2026*
Dekan der Fakultät	*Prof. Dr. Roland Speicher*

Prüfungsausschuss	
Vorsitz	*Prof. Dr. Verena Wolf*
Gutachter	*Prof. Dr. Volkhard Helms*
Gutachter	*Prof. Dr. Olga Kalinina*
Akad. Beisitzer	*Dr. Alexander Gress*

Abstract

The identity of a cell is characterized by its distinct physiology and behaviors, which develop from a single embryonic cell during the course of development. The differences between cell types or tissues within an organism are reflected at multiple levels, from its genetic components in DNA and RNA, to protein interactions and characteristic signaling pathways. In this doctoral thesis, I systematically investigate the rewiring events at different omics levels that are linked by various cell-dependent regulatory factors, ultimately to deepen the understandings of cells fate and identity.

In the first part of this dissertation, we focused on detecting tissues-specific alternative variants in the transcriptomes as well as relevant components of alternative splicing (AS)-modulating mechanisms including epigenetics and RNA-binding proteins (RBPs). Project 1 identified histone modifications-associated exon selection in 19 human cells and tissues in the Human Atlas Epigenomes project, essentially to understand the cell- and tissue-specific transcriptomic changes in relation to epigenetic regulation along human development timeline. Here, we introduced the term *"Epispliced genes"* to refer to genes with strong association between changes in exon usage and in histone modification enrichment. In Project 2, we detected in various brain tissues the expression of "STIM2.3", an alternate transcript of STIM2 gene and clarified its central roles in neuronal differentiation, dendritic spine formation and synaptic activity that potentially contribute to the evolution of hominoids brain development. Project 3 investigated the binding pattern and targets of the RBP IGF2BP2 *in vivo* in mice hepatocytes using a novel protocol named HyperTRIBE, revealing its mRNA stabilizing effects that contribute substantially to apoptosis and autophagy regulation in mice liver. The finding that IGF2BP2 influences autophagy is a novel discovery of our project.

Splicing decisions are supported by distinct RBPs and lead to systematic changes across the entire proteome. Therefore, the projects in the second part of this dissertation aim to associate stage and tissue-specific RBP topology to those at transcriptomic level. To this end, we first conducted a comparative study of current approaches for analyzing protein-protein interaction networks (PPINs) and rewiring events (Project 4). There we addressed the demand for an analysis workflow that facilitate robust and reliable research on context-dependent PPINs. In this context, we developed two new webservers, PPIxpress and PPICompare, to streamline the two-stepped workflow for constructing and comparing contextualized PPINs (Project 5). On premise of *epispliced genes* analysis and these two new webservers

for differential PPIN analysis, in Project 6, we built the rewired RBP networks across 19 cells and tissues in Project 1, using the same transcriptomic data in the detection of *epispliced genes*. We managed to identify subnetworks of RBPs that are responsible for specific cellular functions from the binding dynamics and co-occurence of RBPs in rewired networks. As RBPs showed distinct rewiring patterns across various cells and tissues which largely resemble the patterns of epispliced genes, this study presents a potential approach to studying splicing decision, machinery, and regulations though connecting RBPs, differential exons and histone modifications.

In summary, this doctoral work explores the alterations at transcriptomic and proteomic levels that guide the cells to their fates, while establishing new approaches, workflows and webservers for differential analysis to compare multiple cells and tissues. It thus lays a foundation for further attempts to unravel splicing mechanism and its regulation, as well as to systematically understand splice decisions and alternative variants by characterizing connections between rewiring events at transcriptome and proteome levels.

Zusammenfassung

Die Identität einer Zelle wird durch ihre spezifische Physiologie und ihr spezifisches Verhalten charakterisiert, die sich im Laufe der Entwicklung aus einer einzigen embryonalen Zelle herausbilden. Die Unterschiede zwischen Zelltypen oder Geweben innerhalb eines Organismus spiegeln sich auf mehreren Ebenen wieder, von den genetischen Komponenten in DNA und RNA über Proteininteraktionen bis hin zu charakteristischen Signalwegen. In dieser Doktorarbeit untersuche ich systematisch die Verschaltungsereignisse auf verschiedenen Omics-Ebenen, die durch verschiedene regulierende Faktoren miteinander verbunden sind. Dadurch wird die Erklärung der Zellidentität vertieft.

Im ersten Teil dieser Dissertation konzentrierten wir uns auf die Erkennung gewebespezifischer alternativer Varianten in den Transkriptomen sowie auf relevante Komponenten zur Modulation des alternativen Spleißens (AS), darunter Epigenetik und RNA-bindende Proteine (RBPs). Das Projekt 1 identifizierte auf Basis von Daten des Human Atlas Epigenomes-Projekts die mit Histonmodifikationen verbundene Exonauswahl in 19 menschlichen Zellen und Geweben. Damit charakterisierten wir die zell- und gewebespezifischen transkriptomischen Veränderungen in Bezug auf die epigenetische Regulation entlang der menschlichen Entwicklungszeitachse. Hier führten wir den Begriff *"Epispliced genes"* ein, um Gene zu bezeichnen, bei denen ein starker Zusammenhang zwischen Veränderungen in der Exonverwendung und der Anreicherung von Histonmodifikationen besteht. Im Projekt 2 wiesen wir in verschiedenen Hirngeweben die Expression von "STIM2.3" nach, einem alternativen Transkript des STIM2-Gens und klärten seine zentrale Rolle bei der neuronalen Differenzierung, der Bildung dendritischer Dornen und der synaptischen Aktivität auf, die möglicherweise zur Evolution der Gehirnentwicklung von Hominoiden beiträgt. Das Projekt 3 untersuchte das Bindungsmuster des RBPs IGF2BP2 *in vivo* in Maushepatozyten unter Verwendung eines neuartigen Protokolls namens HyperTRIBE und deckte dabei seine mRNA-stabilisierenden Effekte auf, die wesentlich zur Apoptose- und Autophagie-Regulierung in der Leber von Mäusen beitragen. Dass IGFBP2 Autophagie beeinflusst, ist durchaus eine neuartige Entdeckung unseres Projekts.

Spleißentscheidungen werden durch unterschiedliche RBPs verstärkt und führen zu systematischen Veränderungen im gesamten Proteom. Daher zielten die Projekte im zweiten Teil dieser Dissertation darauf ab, die entwicklungsstadien- und gewebespezifische RBP-Topologie mit der Transkriptomebene in Verbindung zu bringen. Zu diesem Zweck führten wir zunächst eine vergleichende Studie der

aktuellen Ansätze zur Analyse von Protein-Protein-Interaktionsnetzwerken (PPINs) und Verschaltungsereignissen (Projekt 4) durch. Dabei entwickelten wir einen Analyse-Workflow, der eine robuste und zuverlässige Identifizierung von kontextabhängigen PPINs ermöglicht. In diesem Zusammenhang entwickelten wir zwei neue Webserver, PPIxpress und PPICompare, um den zweistufigen Workflow zum Aufbau und Vergleich kontextualisierter PPINs zu optimieren (Projekt 5). Auf der Grundlage der Analyse von *episplicedgenes* und diesen zwei neuen Webservern für die differentielle PPIN-Analyse konstruierten wir in Project 6 die neu verdrahteten RBP-Netzwerke über 19 Zellen und Gewebe aus Project 1 , wobei wir dieselben Transkriptomdaten wie bei der Erkennung von *episplicedgenes* verwendeten. Wir identifizierten Teilnetzwerke von RBPs, die für bestimmte zelluläre Funktionen verantwortlich sind, und zwar anhand der Bindungsdynamik und des gemeinsamen Auftretens von RBPs in neu verdrahteten Netzwerken. Da RBPs in verschiedenen Zellen und Geweben unterschiedliche Umstrukturierungsmuster aufweisen, die weitgehend den Mustern von *episplicedGenen* ähneln, stellt diese Studie einen möglichen Ansatz zur Untersuchung von Spleißentscheidungen, Mechanismen und Regulationen dar, indem sie RBPs, differentielle Exons und Histonmodifikationen miteinander verbindet.

Zusammenfassend untersuchte diese Doktorarbeit die Veränderungen auf Transkriptom- und Proteomebene, die das Schicksal von Zellen beeinflussen, und etabliert gleichzeitig neue Ansätze, Arbeitsabläufe und Webserver für die Differentialanalyse zum Vergleich mehrerer Zellen und Gewebe. Sie bildet damit den Grundstein für weitere Versuche, den Spleißmechanismus und dessen Regulation zu entschlüsseln sowie Spleißentscheidungen und alternative Varianten systematisch zu verstehen, indem Verbindungen zwischen Verschaltungsereignisse auf Transkriptom- und Proteomebene hergestellt werden.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Volkhard Helms. In all these years, he has been the most supportive and the kindest mentor who has given me solid guidance and unwavering encouragement during my PhD journey. His perspectives, insights and attitudes in research and in life have inspired me greatly and will always resonate with me in the years to come. I also thank Kerstin Gronow-Pudelek whole-heartedly for her helps in all administrative matters and for being a kind, wonderful colleague.

My appreciation extends to our research collaborators, Prof. Dr. Barbara Niemeyer, Prof. Dr. Alexandra Kiemer, Prof. Dr. Sonja Kessler, Prof. Dr. Martin Simon, and all other co-authors, as their valuable insights and contributions have enabled and greatly enriched this dissertation. I would also like to thank Prof. Dr. Andreas Keller and Prof. Dr. Sven Rahmann for their reviews and constructive feedbacks on my work during the qualifying exam.

No journey should ever be undertaken without companions. That is why I am sending most heartfelt cheers to my research group members, past and present, especially Sudharshini Thangamurugan, Hanah Robertson, Aram Papazian, Debarshee Sengupta, Kevin George and Markus Hollander. We have grown so much together, from all the silliest "yaps" and pranks, to serious discussions about work and about life. The joyous moments we share make distant memories seem ever so vivid. To all the friends in ZBI, especially Shusruto Rishik and Paula Kramer, thank you for enriching my life with your friendship and supports. Knowing that you were there is an anchor for me in most daunting times.

The most special thanks are reserved to my family. To my parents, I am forever grateful that you have raised me to be the person I am today. The unconditionally love, freedom and supports you give are the reason I am able to pursue my dreams and overcome all the hardships when I am so much away from home.

I am blessed to have you all in my life.

Contents

List of Figures

List of Tables

List of Tables

List of Acronyms

DNA Deoxyribonucleic Acid
mRNA messenger Ribonucleic Acid
AS Alternative Splicing
snRNA small nuclear RNAs
snRNP small nuclear ribonucleoproteins
RBP RNA-binding protein
SF Splicing Factor
TF Transcription Factor
SRE Splicing Regulatory Elements
PPI Protein-Protein Interaction
PPIN Protein-Protein Interaction Network
DDI Domain-Domain Interaction
DDIN Domain-Domain Interaction Network
UTR Untranslated Region
TSS Transcription Start Site
SOCE Store-Operated Ca^{2+} Entry
TPM Transcripts Per Kilobase Million
FPKM Fragments Per Kilobase Million
LFC Log Fold Change
GLM Generalized Linear Model
LSV Local Splicing Variation
PSI Percent Spliced In
FDR False Discovery Rate
FWER Family-wise Error Rate
BH Benjamini-Hochberg
KS Kolmogorov-Smirnov

List of Tables

Chapter I.

Introduction

1. Motivation

The cell is the most fundamental unit of an organism: From the simplest life forms to the most complex creatures, the cells constitute the living entity and carry out functions that allow it to live, thrive and procreate. As every living body and its components are shaped by and respond to its environment, it is natural to correlate the complexity of cell characteristics to that of its living space, a slice of the highly-dimensional reality full of latent factors. Such manifestation of environmental influences on the cell morphology and behavior is enabled through the genetic information encoded in the cell's genome. This information is passed down from the Deoxyribonucleic Acid (DNA) blueprint to messenger Ribonucleic Acid (mRNA) via transcription, then from RNA to proteins through translation as described by the Central Dogma, a paradigm for the unidirectional genetic information flow that is cornerstone to traditional genomics.

However, a quick glance at the human genome, for example from the RefSeq's genome assembly for Homo sapiens (GRCh38.p14) retrieved in April 2025, reveals large discrepancies between the number of genes (42257), mRNAs (185558), non-coding RNAs (49289), and proteins (136859) [178]. As the numbers speak, transcription and translation are not sufficient to enrich the transcripts pool that gives base to myriads of unique phenotypes and behaviors of the cells. Another evidence for the complexity of the genetic information transfer shows that transcriptomes and proteomes are not equivalent [250, 270]. For example, Schwanhaeusser et al. reported that mRNA levels and transcription rates could only explain around 30-40% of variance in protein levels in mammalian cells, while the effects of translational regulations and protein turnover were starkly dominant [217]. Another example by Yang et al. shows how isoform-specific partners within protein complexes resulted from alternative transcripts, thereby change the interactome profiles that lead to different physiological outcomes in the cells [270]. In the study of cells development and the functional abnormalities leading to their diseased states, the key factors that guide transcriptome diversification and enrichment with high impacts on the

downstream proteome must be identified and their connectivity should be studied systematically.

2. Background

2.1. Alternative Splicing and Transcriptome Diversity

2.1.1. Components and Mechanism

Alternative Splicing (AS) is a post-translational process where the cells generate multiple mRNA transcript variants from single gene templates, essentially by removing introns from the mRNA precursors (pre-mRNA) and joining exons together [258]. The action is executed by the *spliceosome*, a large ribonucleoprotein complex composed of small nuclear ribonucleoproteins (snRNP)s functional units consisted of small nuclear RNAs (snRNA)s and various associated proteins [258, 160]. During its course of actions, spliceosome is assembled dynamically: Starting with a pre-spliceosome early complex of U1 and U2 snRNPs assembled onto the pre-mRNA at the splice site via base pairing, a pre-catalytic spliceosome is formed as U4-U6 tri-snRNP is recruited to the site. After rearrangement which releases U1 and U4 snRNPs from the complex, the spliceosome is catalytically active and exerts its function in two main steps, namely "branching" and "exon ligation". In the "branching" step, the spliceosome cleaves the pre-mRNA at the 5' splice site, resulting in a free 5' exon and a lariat-intron-3' exon intermediate. Splicing is completed by "exon ligation" step, when the 5' and 3' exons are ligased to form the mRNA, the lariat intron is cleaved and the U2, U5 and U6 snRNPs are released and recycled. The splicing cycle is shown in Figure I.1A by Matera et al. [160] and the resulting AS types identified in mammals are illustrated by Tao et al. in Figure I.1B [237].

2.1.2. Regulation of Alternative Splicing

Splicing decision is determined as early as when transcription initiates and continues. This is a process referred to as co-transcriptional splicing that is regulated by a variety of factors including epigenetic modifications, transcription factors and other DNA-binding proteins [133, 144]. As pre-mRNA maturation takes place, splice site recognition and inclusion of exons or introns are modulated by *cis-regulatory elements* and their antagonistic proteins *trans-acting factors* on the mRNA [237, 258, 160]. While other facets like splice site recognition strength, RNA architecture, nucleosome positioning or order of intron removal also aid in understanding the splicing mechanism and its regulation [4, 2, 58], this dissertation focuses on histone modifications and RNA-binding protein (RBP)s as main drivers in co-transcriptional splicing and pre-mRNA maturation.

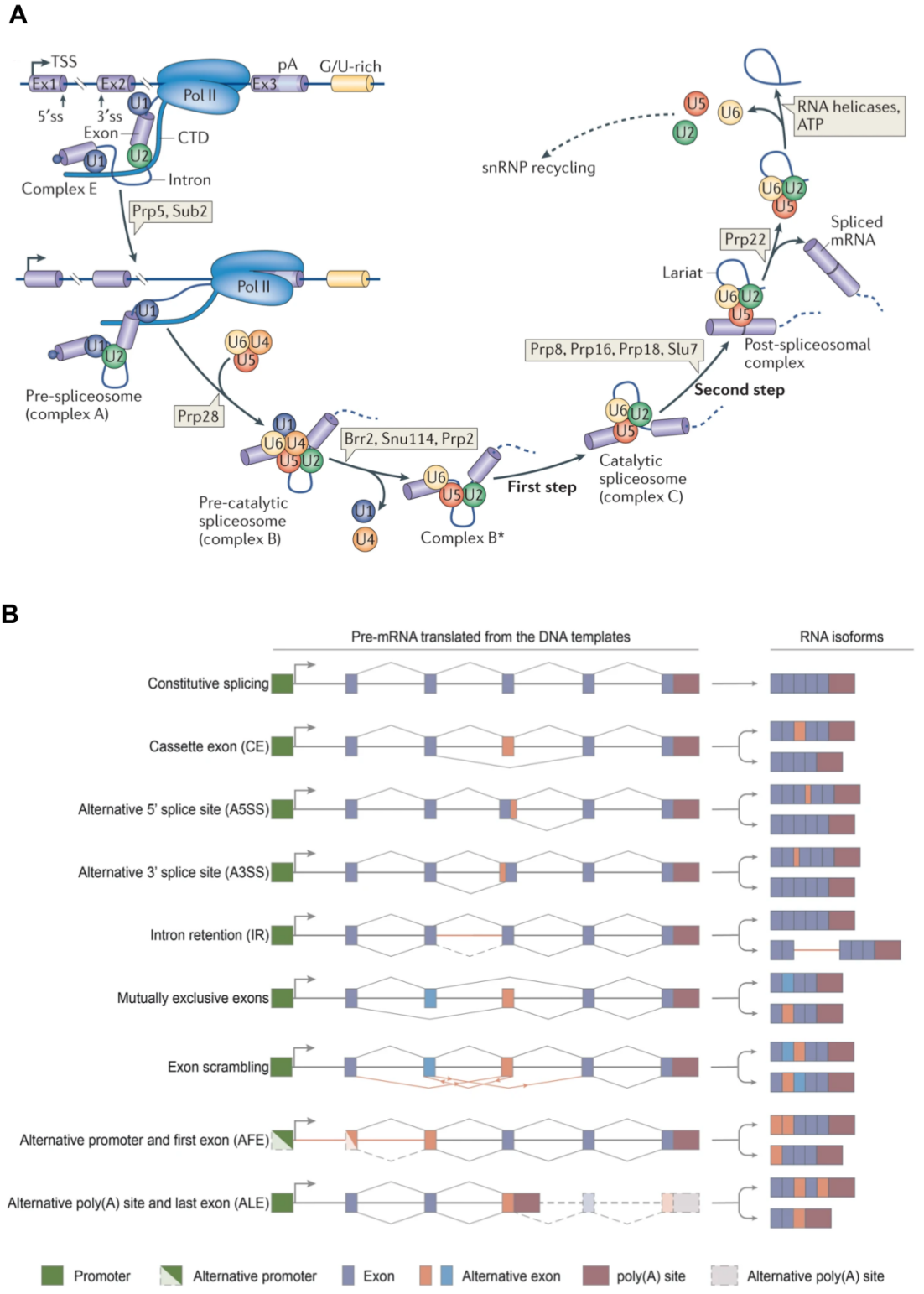

Figure I.1.: Splicing machinery and alternative splicing events. A. Spliceosome assembly takes place at sites of transcription (Schematic taken from Matera et al. [160]). Various splicing factors are recruited to the splice sites for the formation of snRNA-ribonucleoprotein complexes, as well as for catalyzing activities that cleave introns from pre-mRNA and joining exonic regions. B. Identified alternative splicing in mammals (Schematic taken from Tao et al. [237])

Histone Modifications Histone modifications are post-translational epigenetic modifications that control gene activity through altering chromatin state while preserving DNA composition [122, 127, 31]. Chromatin consists of complexes of DNA and proteins with subunits known as nucleosomes. Each nucleosome constitutes 146-147 DNA base pairs wrapped around a histone core of 8 subunits, as shown in Figure I.2A [134]. Stretches of disordered amino acid residues at the end of the nucleosome, referred to as "histone tail', are exposed outwards and prone to enzymatic modifications that could restructure the chromatin via nucleosome rearrangement. The type of modification and its target amino acid constitutes its nomenclature or "histone code". For examples, H3K4me3 refers to trimethylation of lysine 4 on histone 3, while H3K27ac refers to acetylation of lysine 27 on histone 3.

The histone codes control gene expression regulation from chromatin level, as chromatin can be in condensed (tightened) or relaxed (loosened) state that locally affects the activities of Transcription Factor (TF)s, chromatin remodelers and other DNA-binding factors [127, 31]. Histone acetylation like H3K27ac is associated to chromatin relaxation and is enriched in actively transcribed regions, whereas deacetylation leads to chromatin compression and gene silencing. Histone methylation also can either promote or inhibit gene activity. Several examples include H3K4me3 enriched at Transcription Start Site (TSS) facilitating the DNA-RNA polymerase II binding for transcription initiation [23, 97], H3K36me3 at 3' Untranslated Region (UTR) supporting transcription elongation [194, 97], or methylation on H3K9 inhibiting transcription via chromatin condensation [216, 194]. Naturally, histone modifications are not only involved in gene expression regulation, but as well in exon selection [194, 144, 97, 96].

Connecting histone modifications to rewired splicing events therefore requires a systematic approach that takes into account the interplay between different epigenetic patterns and DNA/RNA-binding proteins in both co-transcriptional splicing and post-transcriptional modification processes. Zhou et al. proposed two co-existing models for epigenetically mediated splicing regulation as shown in Figure I.2B-E [279]. The "kinetic coupling" model in subfigures I.2B and C shows that the elongation rate by RNA polymerase II affects exon selection as a result of the kinetic competition between transcriptional elongation and splicing. While the H3K9me2/3 marks are recognized by HP1γ protein and induce exon inclusion by reducing elongation rate, hyperacetylation decondenses the chromatin, leading to higher transcription rate and skipping of the exon. Alternatively, the "chromatin-splicing adaptor systems" model associates local changes in histone or DNA methylation patterns with alternative splicing events especially through chromatin remodeling proteins [279, 218, 153]. H3K36me3 enriched gene regions may be recognized by different proteins such as MRG15 or Psip1, which reinforce exon exclusion or inclusion by recruiting specific trans-acting proteins like PTB or SRSF1 (Figure I.2D and E). Furthermore, factors contributing to splice outcomes such as intron removal order [2, 58], involvement of different RNAs (siRNA/sRNA,

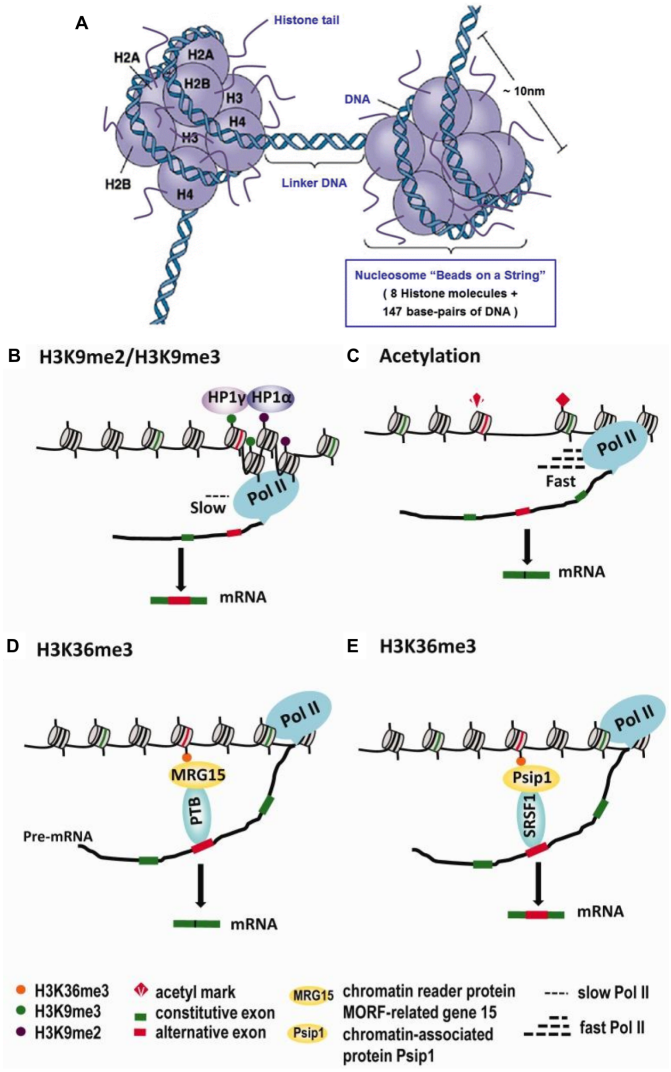

Figure I.2.: Nucleosome structure and histone modification-guided alternative splicing. A. Schematic of nucleosomes as chromatin subunits [134]. Each nucleosome consists of a stretch of DNA wrapping around a protein core of 8 histone proteins. Histone tails are amino acid residues at one end of the protein structure that are sites of enzymatic post-translational modifications. B-E. Proposed mechanisms for histone modification-guided alternative splicing by Zhou el al. [279]. In subfigures B and C, the "kinetic coupling" mechanism show how H3K9me2/3 and histone acetylation decides for or against exon inclusion via controlling transcriptonal elongation rate. The "chromatin-splicing adaptor systems" mechanism is illustrated for H3K36me3 mark in subfigures D and E, where recruitment of different adaptor proteins and splicing factors leads to different splice decisions. Subfigure A is taken from Lee et al's study[134] and subfigure B-E from Zhou et al.'s study [279] studies,

lncRNA and snoRNA) or tissue or context specificity [171, 191, 228, 96] are also studied intensively and add more complexity to the splicing mechanism.

Cis-Regulatory Elements and Trans-Acting Factors Cis-elements or Splicing Regulatory Elements (SRE) are short sequences located in the pre-mRNA intronic or exonic regions that can act as enhancers or silencers for splicing activities. SRE are largely categorized into four groups: exonic splicing enhancer (ESE), exonic splicing silencer (ESS), intronic splicing enhancer (ISE) and intronic splicing silencer (ISS). They control splicing decision by recruiting trans-acting factors to join or co-act with the bound splicesome. Trans-acting factors, or Splicing Factor (SF)s, are primarily consisted of RBPs with high specificity to SREs and can reinforce or inhibit splice activities. SRSFs and HNRNPs are two examples of SF classes that often act competitively during the selection for splice sites or retaining exons [72]. While SRSFs are often recruited by ESEs or ISEs to promote exon exclusion, HNRNPs are often bound to ESSs or ISSs and inhibit the splicing process, leading to exon retention. SFs can also be recruited by other splicing-modulators like histone modifications or other RBPs or chromatin-binding proteins. As shown previously, SFs like PTB and SRSF1 are parts of the chromatin-splicing adaptor systems recruited by H3K36me3 histone mark that either enable or disable exon inclusion (Figure I.2D and E).

2.2. Proteome and Interactome

Proteins are made up of long sequences of amino acids that are encoded by and translated from codons (sets of trinucleotides) on the mRNA. They achieve spatial conformation through four folding levels: Primary - simple amino acid sequences that make up polypeptide chains, secondary - arrangement of polypeptides into α-helices and β-sheets, tertiary - further organization of polypeptides into 3D space and quaternary - highest organizational level that may constitute multiple polypeptide chains [192]. To support specific folding at different levels, proteins require the support of various molecular interactions with varying strengths, from weak bindings like hydrogen bonds or Van der Waals bonds, to stronger ones like ionic bonds or disulfide bridges. Protein activities are determined by their structures and specifically by "domains". A domain is defined as a structural or functional subunit of protein with the ability to fold, evolve or function independently from the rest of the protein it resides in [272]. As the basis for Protein-Protein Interaction (PPI), the binding of a domain from one protein to that of another target protein (Domain-Domain Interaction (DDI)) is highly partner-specific and central to protein functionality. Since proteins are the last product in the downstream flow from genotypes to phenotypes, proteomic research provides the most straightforward explanation to cell behaviors and characteristics leading to great advances in disease diagnosis and therapeutics development [192, 41].

Proteins rarely act alone but rather in complexes with other proteins while serving fundamental functions and processes in living organisms [93, 51]. As a result, proteomic research often analyze not only protein abundance, but also co-existence, localization and interaction among proteins at a certain time. The protein components and their connectivity are represented by Protein-Protein Interaction Network (PPIN)s and expanded by Domain-Domain Interaction Network (DDIN) that map the whole interactome [51]. While a full interactome implies all possible interactions among proteins in the cells, context-dependent interactomes can be extracted as subnetworks that reveal the protein complexes responsible for certain cell responses and cell state progression [93]. This approach to study proteome demands modern methods to detect fine differences in large-scale PPINs with high confidence and explain rewired interactions to domain and molecular levels. Simultaneously, discovering the underlying causes at transcriptome level for disturbances in PPINs and DDINs is an important task in understanding cell behaviors and development of modern gene therapy. To this date, the mapping between transcriptome and proteome still contains unresolved gaps as a result of highly diverse transcripts library and post-translational controls of protein abundance [238, 92, 217] and the coordinated, context-dependent rewiring events at these levels have not been matched completely [93].

2.3. Study of Cell Development under Alternative Splicing

2.3.1. Linking Transcript Diversity to Cell Fates

In studies of cell development and physiology, changes in transcriptomic profile add complexity that goes beyond the gene expression layer and are often key to studying specific cell conditions. From a DNA template, different transcript variants may arise under the action of AS (Figure I.1B) and co-exist as a mixture of variants or isoforms. Naturally, the cell characteristics and behaviors in both normal and diseased states can be linked to canonical isoforms or the changes in isoform composition in the mixture [248, 48]. At the same time, many AS events are highly specific for investigated tissues, developmental stages, or growing conditions and can reveal important processes that shape the cell fates [237, 270]. As an example, the interplay between two transcript variants of FGFR2 gene containing either exon *IIIb* or *IIIc* is central to the epithelial-mesenchymal stem cell transition, along with exon-specific histone modifications like H3K36me3, H3K27me3, H3K4me1 and H3K4me3 [153, 218]. Another evidence for how AS greatly deepens understandings on cell physiology is the emergence of studies on BCRA1 isoforms that are associated with breast cancer [179]. As the reportedly low somatic mutation rate in BRCA1 gene in breast and ovarian tumors was not consistent with the high incidence of tumorous cells, many studies shifted their focus on the diversity of BRCA1 alternative transcripts and found various transcript variants responsible for tissue-specific tumorigenesis. TODO: Note, however, that the mutation frequencies are higher in patients from high-risk families.

These examples demonstrate how understanding the coordinated diversification of transcriptome under AS and its modulating factors is a powerful approach to study cell development and physiology. Meanwhile, as genetic information is passed down from DNA to mRNA and then to protein, the connection between transcriptome and proteome, specifically how different transcript variants determine protein isoforms and their interactions, is an other important facet that ultimately reveals how different cell types grow and behave.

2.3.2. Transcriptome to Proteome: Lost in Translation

AS creates a vast variety of transcript variants from a single gene template, which in turn leads to the production of different protein isoforms. The mRNA and isoform abundance, however, do not correlate directly with protein behaviors and functions, which are primarily revealed by protein complexes and interactions [217, 92]. Thus, many studies attempted to explore the transcriptome-proteome connection with focus on the correlation and discrepancy between AS and protein abundance or behaviors.

Alternative isoforms vary in structure and naturally in behavior, be it their binding preferences to other isoform-specific partners, localization or enzymatic activities [12, 113, 67, 92]. Functional differences between isoforms of the same gene can be so substantial that they behave like entirely distinct proteins, exhibiting different binding capabilities and leading to distinct phenotypes [270]. In other words, AS may remodel the interactome by altering protein expression and their binding property, leading to highly tissue-specific systematic changes that involve SF, RBP and epigenetic modifications. Certainly, transcriptome diversification under AS is the basis for the complexity of the proteome, as reflected by high correlations between global variations in transcript and protein expressions across various species including human [238, 19, 213, 66].

Although the coordinated changes in the proteome can effectively be associated to canonical transcript rewiring events, transcriptome and AS do not completely explain the interactome. While the global "across-gene" correlation between mRNA and protein copy numbers ranges from moderate to high as mentioned, this only applies to analysis of large genesets with little variations when the same samples, tissues or cell lines are studied [238]. The "within-gene" correlation that measures the gene-specific agreement between transcript and protein expression levels is often low and highly inconsistent across genes, tissues, individuals, species or experimental conditions [238, 19, 66, 252]. While the presence of multiple transcripts for a single gene and the resulting isoform diversity may explain this low consistency, the protein abundance and interactions are also affected by other factors such as translational or post-translational regulations [92], protein turnover and degradation rates [217] and interactions with protein partners [92].

3. Research objectives

The two key ideas in section 2.3 constitute the premises of my dissertation, where I aim to study the deciding factors during transcriptomic and proteomic divergence that lead different cell types to their fates under specific environmental and developmental contexts. The first part of this doctoral work including projects 1, 2 and 3 investigates the interplay between AS, histone modifications and RBPs in the context of cell development and physiology. The second part containing projects 4, 5 and 6 reviews and contributes to the advances in context-dependent proteomic research with two webservers, as well as associates the rewired interactions among RBPs to alternative transcripts during human development. Specifically, the projects in this dissertation contribute to the research aims as follows:

Project 1 : As introduced in the Background subsection 2.1.2, AS and epigenetic deregulation have been individually associated to cell identity, differentiation and reprogramming, and a few models for splicing regulation via histone modifications have been proposed. However, a systematic approach linking genome-wide exon rewiring events to local histone modifications at organism level was not established. In this project, we developed an analysis workflow to identify "Epispliced genes", which are genes with exon splicing events significantly correlated to histone modifications enriched at alternative exon flanks in 19 cells and tissues from the Human Epigenome Atlas project. Key processes in cell development were revealed from the analysis of exon-/histone patterns-rewiring events in tissue-specific epispliced genes. The pipeline identifying splicing events based on either exon usages or splice junctions established in this project is also used in projects 2 and 6.

Project 2 : This project is a collaboration with the group of Prof. Barbara Niemeyer where we elucidate the roles of STIM2.3/STIM2G, a unique short variant of STIM2 gene, in human brain development. STIM2 is a prominent gene in brain formation and functions in hypoxia induced neuronal death, dendritic spine formation and regulation of neurotransmission via Store-Operated Ca^{2+} Entry (SOCE) channels. The investigation on expression and regulation of STIM2.3/STIM2G variant via Calcium channels contribute to the understanding of neurogenesis processes and neuronal degenerative diseases like Huntington's disease, which aligns with the research aim to study how alternate transcripts affect the cell physiology and development.

Project 3 : In collaboration with Prof. Alexandra Kiemer, we study the RBP IGF2BP2 in terms of binding capability and biological impacts on its target genes in mouse hepatocytes. Using an experimental protocol named HyperTRIBE, we identify the binding pattern of IGF2BP2 in the transcriptome and elucidate the cellular processes modulated by such RBP-mRNA interaction. As the regulation of

AS is usually not feasible without compatible RBPs (Background subsection 2.1.2), the study on binding activity of IGF2BP2 and its genome-wide targets opens up possibilities to investigate splicing-specific binding events and thus, is a part of the research objectives.

Project 4 : This project is the literature and technical review where we discussed various state-of-the-art databases and softwares for the construction of PPINs from transcriptomic data. Ultimately, we emphasize strongly the importance of using context-dependent interactomes to unravel the complexity of proteomes, as well as the gaps between transcriptomes and proteomes, which are addressed in the Background subsections 2.2 and 2.3.2 respectively.

Project 5 : Following the review in Project 4, we developed two webservers, PPIXpress and PPICompare, to facilitate the construction of context-specific PPINs and their comparison. The webservers can be used independently or in combination as a seamless workflow that is user-friendly, scalable and accessible. Used in Project 6 to construct and compare context-specific RBP networks from Human Epigenome Atlas' transcriptomic data, the webservers are two important tools that support my study aim of mapping the rewiring events at transcriptomic and proteomic levels.

Project 6 : The study aims to investigate the connection between rewired transcripts and the functional interactomes of RBPs that may be involved in AS or histone modification-associated alternative exons. Using PPIXpress and PPIXCompare webservers from Project 5, I investigate the differences in transcriptome-inferred RBP networks of 19 cells and tissues used for finding Epispliced genes in Project 1. The binding dynamics of RBPs in differential RBP networks also reveal splicing-relevant interactions and complexes, demonstrated through a case study comparing CD4 and CD8 cells.

4. Outline of research topics

Project 1 Do, H.T.T., Shanak, S., Barghash, A. and Helms, V., 2023. **Differential exon usage of developmental genes is associated with deregulated epigenetic marks**. *Scientific reports*, 13(1), p.12256.

Abstract: Alternative exon usage is known to affect a large portion of genes in mammalian genomes. Importantly, different splice isoforms sometimes possess distinctly different protein functions. Here, we analyzed data from the Human Epigenome Atlas for 11 different human adult tissues and for 8 cultured cells that

mimic early developmental stages. We found a significant enrichment of cases where differential usage of exons in various developmental stages of human cells and tissues is associated with differential epigenetic modifications in the flanking regions of individual exons. Many of the genes that were differentially regulated at the exon level and showed deregulated histone marks at the respective exon flanks are functionally associated with development and metabolism.

Project 2 Poth, V., Do, H.T. T., Förderer, K., Tschernig, T., Alansary, D., Helms, V. and Niemeyer, B.A., 2023. **A better brain? Alternative spliced STIM2 in hominoids arises with synapse formation and creates a gain-of-function variant**. *bioRxiv*, pp.2023-01.

Abstract: Balanced Ca2+ homeostasis is essential for cellular functions. STIM2 mediated Store-Operated Ca2+ Entry (SOCE) regulates cytosolic and ER Ca2+ concentrations, stabilizes dendritic spine formation and drives presynaptic spontaneous transmission and ER stress in neurons. Recently identified alternative spliced variants expand the STIM protein repertoire, uncover unique functions and facilitate our understanding of tissue specific regulation of SOCE. Here, we describe an addition to this repertoire, a unique short STIM2 variant (STIM2.3/STIM2G) present only in old world monkeys and humans with expression in humans starting with the beginning of brainwave activity and upon synapse formation within the cerebral cortex. In contrast to the short STIM1B variant, STIM2.3/STIM2G increases SOCE upon stimulation independently of specific spliced in residues. Basal cluster formation is reduced and analyses of several additional deletion and point mutations delineate the role of functional motifs for Ca2+ entry, NFAT activation and changes in neuronal gene expression. In addition, STIM2.3/STIM2G shows reduced binding and activation of the energy sensor AMPK. In the context of reduced STIM2.3 splicing seen in postmortem brains of patients with Huntington's disease, our data suggests that STIM2.3/STIM2G is an important regulator of neuronal Ca2+ homeostasis, potentially involved in synapse formation/maintenance and evolutionary expansion of brain complexity.

Project 3 Do, H. T. T., Both, S., Kröhler, T., Pirritano, M., Van Wonterghem, E., Franzenburg, S., Simon, M., Biswas, J., Helms, V., Kessler, S., Kiemer, A. (2024). **Identification and analysis of IGF2BP2-binding pattern from HyperTRIBE protocol in mouse livers**. *Manuscript in revision at NAR Molecular Medicine.*

Abstract: Targets of RNA-binding proteins (RBPs) are often investigated using variants of the cross-linking and immunoprecipitation methodology, which show several disadvantages in target detection. The m6A reader RBP insulin-like growth factor 2 mRNA binding protein (IGF2BP2/IMP2) exerts important pathophysiological functions as a metabolic regulator and tumor promoter that has been

described to affect the stability, localization, and translation of its targets. The aim of this study was to determine the RNA targets of IGF2BP2 using HyperTRIBE in an in vivo mouse model.

IMP2-associated adenosine-to-inosine editing sites were identified by hydrodynamic transfection of mouse livers with an IMP2-ADAR (adenosine deaminase acting on RNA) construct. The results of functional enrichment and motif analysis suggest IMP2-facilitated target stabilization and confirmed the m6A binding motifs. In addition, the overlap with data of a TRIBE experiment in mouse embryonic fibroblasts and with those of differential gene expression was investigated. Comparative transcriptomics between IMP2, wild-type, and control samples (mCherry-ADAR) revealed an enrichment of IMP2-bound mRNAs associated with autophagy. Functional knockdown experiments in HepG2 cells showed increased autophagic flux, supporting the involvement of IMP2 in autophagy.

In conclusion, this study shows that HyperTRIBE can efficiently profile RBP targets in vivo. In addition, it indicates a possible role of IMP2 in autophagy.

Project 4 Hollander, M., Do, T., Will, T. and Helms, V., 2021. **Detecting rewiring events in protein-protein interaction networks based on transcriptomic data**. *Frontiers in Bioinformatics*, 1, p.724297.

Abstract: Proteins rarely carry out their cellular functions in isolation. Instead, eukaryotic proteins engage in about six interactions with other proteins on average. The aggregated protein interactome of an organism forms a "hairy ball"-type protein-protein interaction (PPI) network. Yet, in a typical human cell, only about half of all proteins are expressed at a particular time. Hence, it has become common practice to prune the full PPI network to the subset of expressed proteins. If RNAseq data is available, one can further resolve the specific protein isoforms present in a cell or tissue. Here, we review various approaches, software tools and webservices that enable users to construct context-specific or tissue-specific PPI networks and how these are rewired between two cellular conditions. We illustrate their different functionalities on the example of the interactions involving the human TNR6 protein. In an outlook, we describe how PPI networks may be integrated with epigenetic data or with data on the activity of splicing factors.

Project 5 Do, H.T.T., Thangamurugan, S. and Helms, V., 2025. **PPIXpress and PPICompare Webservers infer condition-specific and differential PPI networks**.*Bioinformatics Advances*, p.vbaf003.

Abstract: We present PPIXpress and PPICompare as two webservers that enable analysis of protein-protein interaction networks (PPINs). Given a reference PPIN and user-uploaded expression data from multiple samples, PPIXpress constructs context-dependent PPINs based on major transcripts and high-confidence domain interactions data. To derive a differential PPIN that distinguishes two groups

of contextualized PPINs, PPICompare identifies statistically significant altered interactions between multiple context-dependent PPINs from PPIXpress. We present a case study where PPIXpress and PPICompare webservers were used in combination to construct the PPINs specific for melanocytic nevi and primary melanoma cells, and to detect the rewired protein interactions between these two sample types.

PPIXpress and PPICompare webservers are available at https://service.bioinformatik.uni-saarland.de/ppi-webserver/index_PPIXpress.jsp and https://service.bioinformatik.uni-saarland.de/ppi-webserver/index_PPICompare.jsp, respectively. Alternatively, the webservers and application updates can be found at https://service.bioinformatik.uni-saarland.de/ppi-webserver/

Project 6 Do, H.T.T., Helms, V., 2025. **Tissue-specific RNA binding protein networks provides insights on splicing processes**.*Draft.* In a former study (Project 1), I investigated the connection between histone modifications and alternative exons in the transcriptomes of 19 human cells/tissues from the Human Epigenomes Atlas. To expand the knowledge on how these epigenetics-associated rewired transcripts control the cell activities and development through altering the functional interactomes, in this study, I inspect the differences in transcriptome-based RBP networks specific to each of the same 19 cells and tissues and link them to the results of our previous study. Using the implemented webservers PPIXpress and PPICompare (Project 5), I constructed the sample-specific RBP networks and compared them to identify the rewired interactions that distinguish the cells/tissues. The binding dynamics of RBPs in these networks were also analyzed to validate the plausibility and biological functions of these interactions.

Chapter II.

Materials and Methods

1. Differential analysis of omics data

Studying the canonical differences that define the cell behaviors and development under various circumstances is a recurring theme throughout my doctoral work. In the first project, differential exon levels and differential histone modifications were associated to detect AS events that might be influenced by post-transcriptional changes (Chapter 1). In the second project (Chapter 2), we detected the presence and expression of a STIM2's transcript variant that affect neurogenesis using differential splice analysis. In Chapter 3, genes with differential expression were compared to target genes of the RBP IGF2BP2 as a part of our approach to study IGF2BP2-binding patterns and its biological roles (Chapter 3). Furthermore, methods for detecting rewired protein interactions for proteomic studies were discussed extensively in our review article (Chapter 4), which supports the application of two newly implemented webservers that tackle the research gaps in this topic (Chapter 5). Finally, Chapter 6 presents my preliminary study on AS-coordinated changes in RBP networks. In this Materials and Methods section, the main approaches to detect differential features at multiple omics levels are summarized, together with data availability and general preprocessing methods for a thorough overview.

1.1. Transcriptomics

1.1.1. Data retrieval and processing

Source and availability Transcriptomic data used for differential expression analysis are mostly derived from common high-throughput sequencing methods, such as RNA-seq sequencing. Depends on the research questions, refined assays like poly-A plus RNA-seq or long-read sequencing can be used. Poly-A plus RNA-seq is used to capture only the mRNA transcripts without degradable RNAs fractions and is hence suitable for studies on transcript-centered processes such as RNAs splicing, post-translational modifications, etc. [44, 277]. Alternatively, long-read sequencing can be used to capture the full-length transcripts, which is useful for studies on

isoform-level differential analysis [207]. The FastQ-formatted raw RNA-seq reads can be retrieved from public databases such as NCBI Sequence Read Archive (SRA) [137] or ENCODE [56], where pre-processed data in SAM, BAM, BED or GTF formats might as well be available. For studies of transcriptome-epigenome or transcriptome-proteome associations in human cell developments, the Human Epigenomics Atlas from the Roadmap Epigenomics project is a large, reliable data source with standardized experiments and data processing methods across more than one hundred cell/tissue types [210].

Overview of processing methods As the basis for differential analysis is the comparison of gene expression levels between different conditions, raw reads are processed and quantified in a series of steps including trimming, alignment, quantification and normalization. Trimming step serves the removal of adapters, UMIs, barcodes, or other sequencing artifacts and improves the read quality for downstream analysis [282]. Trimmed reads are mapped and aligned against a reference genome in alignment step using tools like STAR alignment [140], HISAT2 [118], or BWA [63]. Read counts can be quantified based on indexed reads using methods like RSEM [138], Cufflinks [240], or estimated independently from alignment step using pseudo-aligned transcripts with Salmon [188] or Kallisto [33]. The estimated read counts are however prone to biases from inconsistencies in library sizes, sequencing depths, gene lengths, or from batch effects. Thus, normalization methods are applied to quantified read counts to minimize effects from systemic experimental variations and ensure comparability across genes and samples.

Typically, the frequency distribution of gene expression data is heavily left-skewed as dominated by low counts and has a long right tail associated to highly expressed genes [25]. While count data can generally be modeled by a binomial distribution or a Poisson distribution, the latter seems more appropriate for gene count data considering the large number of genes available from RNA-seq. Nonetheless, the assumptions of Poisson distribution, such as equal mean and variance and homoscedasticity, are often violated in RNA-seq data due to high biological and technical variability. Additionally, weakly expressed genes are often observed with higher variance when different mouse strains are compared [30] Thus, negative binomial distribution, a generalization of Poisson distribution that uses an additional "dispersion parameter" to estimate heterogeneity, is often preferred to effectively model gene expression profile.

Apart from straightforward approaches such as Transcripts Per Kilobase Million (TPM) (Equation II.1) and Fragments Per Kilobase Million (FPKM) (Equation II.2), methods based on expression data modeling such as DESeq2 [150], edgeR [211], or limma [209] are also commonly used for normalization and differential analysis [cite].

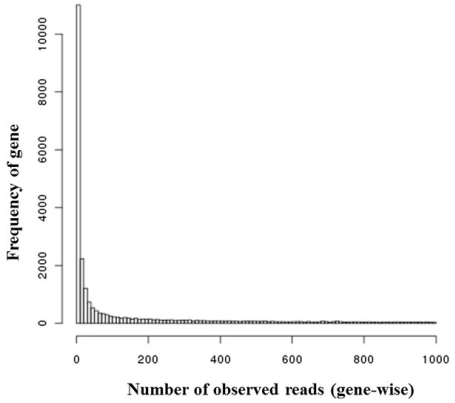

Figure II.1.: Frequency distribution of an example gene expression data. Distribution of number of observed reads per gene for genes with read count less than 1000

$$TPM = \frac{\text{Mapped reads to transcript/Transcript length}}{Sum(\text{Mapped reads to transcript/Transcript length})} \tag{II.1}$$

$$FPKM = \frac{1,000,000 \times \text{Mapped fragments}}{\text{Gene length} \times \text{Total mapped fragments}} \tag{II.2}$$

While detecting alternative transcripts has been extensively discussed and practiced, it is possible that one might be interested in transcript-defining regions like exons and exon-intron junctions, introns or untranslated regions. It is thus utterly important to tailor the processing and analysis pipeline according to biological contexts and mechanisms. For instance, the differential analysis of exon expression (or usage) would require additional normalization across exons from all considered genes to account for gene-wise bias. At the same time, deep understanding on existing differential analysis approaches allows proper data preparation and ensures selection of the most suitable method for the research question.

1.1.2. Tools for differential analysis

Gene-level

Gene expression ratio In classical differential expression analysis, the detection of significantly deregulated transcripts or genes relied on ratio-based decision that uses "expression ratio" or "Log Fold Change (LFC)" to assess the magnitude of differential gene expression [197]. LFC is simply the binary logarithm of the fraction between the normalized expression level of a gene in one sample over that in the control sample II.3, and thus, a LFC of 1 indicates a 2-fold higher gene expression in the sample.

$$\text{LFC} = \log_2 \left(\frac{\text{Gene expression level in sample}}{\text{Gene expression level in control}} \right) \tag{II.3}$$

As it requires low computing effort and has high interpretability, LFC is still one of the most commonly used metrics in differential analysis today and is central to more complicated models like limma [209] or DESeq2 [150]. The identification of differentially expressed genes by solely LFC is however severely challenged as it lacks the statistical basis to draw conclusive results. To elaborate, LFC-normalization fails to eliminate noises within sample and cross-replicates variability, and must be coupled with other hypothesis testing methods like t-test for significance assessment [197].

DESeq2 DESeq2 ranks as one of the most popular softwares for differential gene expression analysis [150]. As mentioned in the previous section, gene expression data typically follow a negative binomial distribution dominated by low gene counts with higher variance observed across samples, as compared to the much smaller portion of less varying, higher gene counts. DESeq2 handles this commonly encountered problem of heteroscedasticity in RNA-seq data by estimating the dispersion of gene count using an empirical Bayes approach. Given a gene i in sample j, its read count K_{ij} following the binomial distribution is described as:

$$K_{ij} \sim \mathrm{NB}(mean = s_{ij}q_{ij}, dispersion = \sigma_{ij}) \tag{II.4}$$

$$\log q_{ij} = \sum_r x_{jr}\beta_{ir} \tag{II.5}$$

where s_{ij} are the normalization factors that are considered constant within sample j ($s_{ij} = s_j$) and q_{ij} is the normalized expression level of gene i in sample j. The LFC $\log q_{ij}$ is estimated by the coefficients β_{ir} in the design matrix x_{jr}, where β_{ir} approaches a zero-centered normal distribution:

$$\beta_{ir} \sim \mathrm{N}(0, \sigma_\beta^2) \tag{II.6}$$

Dispersion shrinkage by empirical Bayes approach can be described in three main steps: (i) First, gene-wise dispersions estimates are derived using maximum likelihood (MLEs), then (ii) the MLEs-mean dependence is modeled by a smooth curve representing the prior means for the gene counts, and (iii) finally, the MLEs are fitted to MLEs-mean curve, or "shrunk" toward prior means, resulting in maximum *a posteriori* (MAPs) as final dispersion estimates. Likewise, LFC estimates are shrunk toward zero by empirical Bayes method. Shrinkage estimation of dispersions and LFCs are central to DESeq2 as the magnitude of shrinking depends on provided sample size and how good the fit is between dispersion values and estimates. This means in genes where LFC variance and mean count ratio blows up (ie. in small sample sizes), the shrinkage will be more aggressive to handle the heavy heteroscedasticity. To detect genes with differential expression, DESeq2 performs Wald tests to assess the significance of the shrunken LFCs and adjust the p-values using Benjamini-Hochberg procedure. The criteria for differential expression are typically set to a p-value threshold of 0.05 and an absolute LFC of 1.

Exon-level While differential analysis of transcriptomic data at gene-level can reveal up- or down-regulated genes or transcripts, it provides little basis to explain the diversification of isoforms through alternative splicing under influence of post-transcriptional modifications. For this purpose, differential analysis of exon or isoform composition is indeed more appropriate. Various approaches have been implemented with focus on AS detection and can be generally categorized into exon-based (DEXSeq [10], edgeR [211]), isoform-based (cuffdiff2 [239], DiffSplice[98]) and event-based methods (MAJIQ [246], rMATS [221]) [163]. The tools are vastly different in terms of requirements on the availability of data (reference genome availability, transcriptomic or genomic data), experiment designs (two groups or multiple groups design), or type of AS events (skipped exons, alternative 5'/3' splice sites, retained introns, mutually exclusive exon usage).

DEXSeq DEXSeq [10] is a software developed for differential analysis at exon-level. Instead of modeling gene/transcript expression like DESeq2, DEXSeq aims to quantify exon expression, referred to as "exon usage", and identify differential exon usage using a χ^2 likelihood-ratio test on Generalized Linear Model (GLM)s for exon counts with Cox-Reid dispersion estimate. This approach thus requires a "flattened" exon model to ensure non-overlapping count values that in principle should, similarly to gene/transcript counts, follow a negative binomial distribution. Figure I.7A in Chapter 1 describes the flattened model, where overlapping exonic regions across transcripts would define a new exon, or "exon bin". The exon count K_{ijl} in sample j and exon bin l of gene i is modeled as:

$$K_{ijl} \sim \text{NB}(mean = s_j \mu_{ijl}, dispersion = \alpha_{ij}) \tag{II.7}$$

$$\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ij}^S + \beta_{i\rho_j l}^{EC} \tag{II.8}$$

where s_j is the normalization factor constant within sample j and μ_{ijl} is the normalized exon expression level of gene i in sample j and exon bin l. The LFC for exon expression $\log \mu_{ijl}$ is predicted by a linear function including the baseline expression β_i^G for gene i, the expected fraction of reads that map to gene i that overlap with counting bin l β_{il}^E, the fold change of gene i in sample j β_{ij}^S, and the effect which condition ρ_j has on the expected count at exon bin l in gene i in sample j $\beta_{i\rho_j l}^{EC}$. This design allows DEXSeq to derive the effect a sample condition has on a specific exon bin in addition to the overall impact from gene/transcript counts and samples and increase power in the test for differential exon usage. To estimate the dispersion α_{ij}, Cox-Reid likelihood estimation is used. Finally, counting bins with significant exon usage are detected by a χ^2 likelihood-ratio test that compares the full model with the reduced model where the condition effect $\beta_{i\rho_j l}^{EC}$ is removed. The p-values from likelihood-ratio test for each gene are adjusted by Benjamini-Hochberg method for multiple testing correction. As changes in exon usage reflect the inclusion or exclusion of an exon in an isoform, DEXSeq is

a powerful tool to detect exon switching-typed AS events in connection to local transcript modifications like histone modification.

MAJIQ (Modeling Alternative Junction Inclusion Quantification) As another differential exon analysis tool used in this doctoral work besides DEXSeq, MAJIQ[246] offers another approach to identify AS events at exon-intron junctions with Local Splicing Variation (LSV)s. LSVs are splice junctions identified throughout the reference genome and RNA-seq reads which make up a splicegraph, where AS events across a gene region are demonstrated by source/target exon junctions. To correct for measurement error of true read rates at a specific junction across different replicates, MAJIQ performs nonparametric sampling and keeps the bootstrapped coverages for modelling of junction read counts in the later stage. Instead of quantifying exon expression like DEXSeq, MAJIQ models the number of reads aligned to each junction j in a LSV r_j, assuming read counts follows a binomial distribution:

$$r_j \sim \mathrm{B}(n = \sum_{j \in LSV} r_j, p = \Psi_j) \tag{II.9}$$

$$\Psi_j \sim \beta(\frac{1}{J}, 1 - \frac{1}{J}) \tag{II.10}$$

$$\Psi_j | r_j^{'} \in LSV \sim \beta(\frac{1}{J} + r_j, 1 - \frac{1}{J} + \sum_{j' \neq j} r_j^{'}) \tag{II.11}$$

where Ψ_j is the probability of the splice junction r_j. As Ψ prior distribution was observed to spike around zero or one, MAJIQ estimates the priors using a beta function with a Jeffrey's prior J which is invariant under reparameterizations (II.10). As mentioned before, MAJIQ stores bootstrapped read coverages from different replicates at each junction that can now be used to estimate Ψ posteriors, following Equation II.11. The final Ψ posterior is the ensemble of posteriors averaged across replicates and MAJIQ's quantifier for an AS event at a specific junction known as Percent Spliced In (PSI). By comparing the PSIs between two sample groups ($dPSI(\Delta\Psi) = \Psi_1 - \Psi_2$), users may draw conclusion on whether AS event occurs at a specific junction.

1.2. Epigenomics

1.2.1. Data retrieval and processing

Source and availability Profiling of epigenome and specifically histone modification has been enabled by chromatin immunoprecipitation-based methods like ChIP-seq, ChIP-on-chip or qChIP. Immunoprecipitation refers to the use of a specific antibody to tag the histone of interest in a DNA fragment, allowing the fragment to be retained in the later DNA purification step. As a high-throughput

approach that allows whole-genome profiling at base-pair resolution, ChIP-seq is the most popular assay with the highest data availability in databases such as GEO (NCBI) [178], BioStudies (EMBL-EBI) [95] or ENCODE [56]. For histone modification studies in human cells and tissues, Human Epigenomics Atlas [210] provides a large collection of high-quality tissue-specific ChIP-seq data in both raw (FastQ) and processed formats (SAM, BAM, BED, GTF, etc.) that can easily be linked to other assays like mRNA-seq, DNA methylation, etc.. This enables extensive and systematic research of multi-omics data that shed lights on the intricate machinery steering cell characteristics and behaviors.

Overview of processing methods The typical processing pipeline for ChIP-seq data includes read alignment, peak calling, normalization and differential analysis. Similar to RNA-seq data, raw ChIP-seq data containing raw reads are stored under FastQ format and must be aligned against a reference genome to identify the genomic locations of the reads. To achieve this, softwares like STAR [63] or BWA [140] are commonly used. In the next peak calling step, the enriched regions in the aligned reads known as "peaks" are identified to give information about the binding sites of a histone-mark specific antibody to the chromatin. Peak calling tools such as MACS2 [76] or HOMER [90] differ greatly in terms of algorithms and ability to serve specific purposes. For example, MACS2 is a powerful peak caller that enable the identification of both broad and narrow peaks, which is extremely important considering the patterns of the histone modification being studied, while HOMER on the other hand has been adopted more widely for de novo motif discovery associated with binding sites [275, 76, 90]. Although reliable peak calling methods already consider variations across replicates generally, the read counts across samples are still prone to batch effects and other systematic biases. Thus, normalization methods are necessary and are typically integrated into differential peaks analysis tools. Some widely used tools for identifying differential ChIP-seq including MAnorm, HOMER, MACS, or DiffBind [229] provide a wide range of functionalities that enable users to work with predefined peaks or any regions, replicated or non-replicated samples, or narrow or broad ChIP-seq binding regions.

1.2.2. Tools for differential analysis

MAnorm MAnorm[219] is one among established methods for identification and interference of differential ChIP-enriched regions. The method fundamentally applies rescaling and normalization according to the intensities of common peaks found across the two considered samples, assuming the degrees of enrichment are consistent since they result from the same binding mechanism of one common chromatin-associating protein. This assumption additionally enables the handling of the high signal-to-noise problem in ChIP-seq data, as well as the identification of differential binding patterns that strongly correlate to differential expression in

cell type-specific contexts. After identifying all peaks from two samples, MAnorm calculates the log2 ratio of ChIP-seq read counts M, or the LFC between the samples, and the average log2 read counts A for each peak.

$$M = \log_2 \frac{R_1}{R_2} \tag{II.12}$$

$$A = \log_2 \frac{R_1 \times R_2}{2} \tag{II.13}$$

where R_1 and R_2 represent the read density at a peak region in sample 1 and 2, respectively. In the next step, MAnorm identifies common peaks and models the M values as a function of A values for these peaks using a robust linear regression model with a bi-square weighting function. For common peaks, M values between two samples should be close to 0 as expected from the "invariant intensities across sample" assumption. Thus, MAnorm performs a coordinate transformation that overlaps the M-A fitted regression line to the y-axis, as new M and A values are now normalized values for the common peaks. The extrapolation to all non-common peaks is done by the same transformation. Finally, differential peaks are identified by a p-value from a Bayesian model that quantifies the difference between normalized read intensities from the two samples.

1.3. Proteomics

1.3.1. Data retrieval and processing

The data materials for proteomic and differential proteomes studies come from various sources and are available in public databases as summarized in Table IV.1 in Project 4. In general, data on protein/domain interactions can be obtained from Yeast Two-Hybrid assays, co-immunoprecipitation or Tandem Affinitiy Purification experiments. An important step toward comparing different interactomes is conceptualization of PPINs/DDINs, which entails identifying canonical interactions specific for a cell type or condition. The approaches and tools serving this purpose are discussed in detail in section 3 of Project 4.

1.3.2. Tools for differential analysis

In Project 4 section 4, we reviewed the existing tools for differential interactomes analysis in terms of resources, workflows and features with a summarization in Table IV.1. Additionally, we provided a use-case review for prominent softwares and databases to illustrate the differences in their functionalities and applicability in section 5.

1.4. Standards in differential analysis

Differential analysis of omics data, especially for large-scale systemic studies, involves multi-staged, highly biologically specific processing workflows that are

unavoidably prone to errors and biases. At any step, thorough understandings on technical and biological aspects of the experiments are required alongside with rigorous quality control measures. Data quality is agreeably the most deciding factor for reliable outcomes in any analysis. In differential studies of omics data, many analyses depend on statistical modeling and hypothesis testing for identifying deviations from expectations. Using low quality data, such as those with low read depth or complexity, high heterogeneity or high noise-to-signal ratio, would fail to add meaningful insights to the statistical models, if not to say invalidate the analysis methods with unaccountable noises and biases, ultimately leading to wrong conclusions. On this basis, outdated data or those without labels, sources and descriptions are huge risk factors not to be included as inputs. To minimize the impact of inadequate data which is most often unavoidable, technical errors from experiment stage as well as computational errors during analysis stage should be tightly monitored. For example, the use of FastQC and MultiQC [73] to inspect reads quality and generated outputs during trimming, alignment, peak calling, etc. is highly recommended.

2. Statistical hypothesis and testing

Statistical inference is an indismissible part of any study, from forming a research question to conducting a validating experiment, from collecting results to interpreting them and drawing the final conclusions. The process is methodically based on posing a research-oriented question that leads to a hypothesis which can be proven or disproven using a suitable statistical test. It is thus essential to understand the nature of insights that constitute the research assumptions, especially in biological research, to come up with correct hypothesis and testing methods.

2.1. Methodology

2.1.1. Formation of null and alternative hypothesis

Statistical inference begins with forming a "null hypothesis" H_0 that commonly assumes the non-existence of the effect being investigated, and an "alternative hypothesis" H_1 that negates H_0 [70]. As an example, H_0 might assume there is no difference in between two experimental groups and the failure to reject H_0 would mean there is not enough power to conclude there is a statistically significant difference between the groups:

$$H_0 : \mu_1 = \mu_2 \tag{II.14}$$
$$H_1 : \mu_1 \neq \mu_2 \tag{II.15}$$

These hypotheses can also be directional, assuming the effect to be positive or negative:

$$H_0 : \mu_1 \leq \mu_2 \tag{II.16}$$
$$H_1 : \mu_1 > \mu_2 \tag{II.17}$$

2.1.2. Rejecting the null hypothesis

It is common to set up an experiment where the effect that differentiates two samples or a sample from a population is assumed to be not significant in statistical sense [70]. Rejecting the null hypothesis, as a result, indirectly confirms the effect to be extreme or significant enough to make a difference in the comparison. To reject a null hypothesis, a test statistic is computed from the sample/-s, either in one-sample test where the sample is compared against a population or two-sample test between two samples, and compared to a critical value. This requires the assumption that the samples follow a specific distribution with known descriptive parameters like mean, standard deviation, variance, etc.. The critical value is a threshold that defines the "critical region" or "region of rejection", a range of values in the extremity that if the test statistic falls within, the null hypothesis is rejected. Naturally, one can select a critical value that is more or less extreme to tighten

Table II.1.: Confusion matrix for hypothesis testing

		Null hypothesis is	
		True	**False**
Null hypothesis was	Rejected ($p - value \leq \alpha$)	False positive (Type I error = α)	True positive (Test power = 1-β)
	Not rejected ($p - value \geq \alpha$)	True negative (Probability = 1-α)	False negative (Type II error = β)

or relax the condition for rejecting the null hypothesis. It is in general a more common and intuitive practice to determine the critical threshold through selecting a "significance level" denoted by α, which is the probability of falsely rejecting the null hypothesis when it is actually true. In another word, if a significance level is set to 0.05, there is a 5% chance of false positive or "type I error" (Table II.1).

Alternative to using test statistic and critical value, one may decide to reject or accept the null hypothesis based on the probability of obtaining a test result that is as least as extreme as the actually observed result, a concept commonly known under the term "p-value". As a smaller p-value indicates a lower probability for the test result to fall in the extreme zone of rejection, the p-value can be compared against the chosen significance level to conclude whether there is strong enough evidence to reject the null hypothesis. In an one-tailed test where the hypotheses are directional, the p-value needs to be smaller than the significance level, whereas in a two-tailed test where extreme values are distributed at both ends of the distribution, the p-value needs to be smaller than half of the significance level to reject the null hypothesis.

On the explained basis of hypothesis formation and testing, researchers can choose a suitable statistical test that is most appropriate for the data type and research question. Some of the most common statistical tests, their assumptions and applications are discussed in the following sections.

2.1.3. Multiple testing correction

As the decision for rejecting the null hypothesis is based on significance level that tolerates a certain threshold of false positive rate, the more statistical tests are conducted, the more likely it is to observe a wrongly rejected null hypothesis. For example, repeating a test 1000 times while setting false positive rate to 5% would allow 50 cases to appear as significant by chance, when they are in fact not. To address this issue, multiple testing correction methods like Bonferonni correction or Benjamini-Hochberg (BH) correction are used to adjust the p-values of the tests through controlling Family-wise Error Rate (FWER) or False Discovery

Rate (FDR). Given m tests and a significance level α, FWER is the probability of making at least one type I error in a family of tests and FDR is the likelihood of finding false positives among all rejected hypotheses:

$$FWER = 1 - (1 - \alpha)^m \tag{II.18}$$

$$FDR = \frac{False\ positive}{True\ positive + False\ positive} \tag{II.19}$$

Bonferroni correction Bonferonni correction controls the FWER during multiple testing by restricting the significance level of each test to α/m. Consider a family of null hypotheses $H_1, H_2, \ldots H_m$ and their corresponding p-values $p_1, p_2, \ldots, p_m$, Bonferroni correction rejects the null hypothesis H_i if $p_i \leq \frac{\alpha}{m}$. FWER of the test family is thereby controled at α and the tests in the family are not assumed to be independent from each other.

Bonferroni procedure is however a highly stringent procedure when it involves a large number of tests m, as it could blow up FWER exponentially as m increases and lower the statistical power (Equation II.18). This stringency introduces a higher false negative rate that is especially unfavorable in biology context where overlooking significant genes, exons or any other events might be gravely misleading while explaining a biological system or mechanism.

Benjamini-Hochberg correction The Benjamini-Hochberg procedure controls the FDR (Equation II.19) and is thus commonly known as FDR correction. Unlike Bonferonni procedure, BH correction assumes independency among the tests and is a less stringent correction as it considers false positive rate instead of the whole test family. Considering the largest number of tests k out of a total of m tests such that the corrected probability is still bound by the significance level α ($P_{(k)} \leq \frac{k}{m}\alpha$, where $\frac{k}{m} = FDR$ (Equation II.19)), BH correction rejects the null hypothesis of test i for $1 \leq i \leq k$.

As BH correction also takes the true positive rate into account, it is a more powerful method than Bonferroni correction and is widely used in omics data analysis. BH correction is thus chosen as the default method for multiple testing correction in this thesis, unless stated otherwise.

2.2. Statistical tests in biological problems

As discussed in detail in Methods section 1, omics data are greatly different in distributions, ranges, central tendency, frequency and dispersion. These characteristics dictate the choice of statistical tests, together with the research question and the design of experiment. Some of the commonly encountered questions during biodata analysis that can be solved by hypothesis testing are:

- Are the gene expression levels of two sample groups significantly different?

- Is the binding affinity of an RBP to a specific RNA sequence significantly different from the background?
- Are the biological functional annotations enriched for a set of deregulated genes?
- Is there in general a stronger correlation between the expression of an RBP and its target genes, compared to non-target genes?
- ...

Considering the variety and complexity of available statistical tests, only those that are most relevant to the thesis are discussed in the following sections in terms of purposes, approaches, applications and scenarios, where alternative methods are necessary.

2.2.1. t-test

t-test is a parametric test used for comparing the means of two samples, or the mean of one sample to that of a population. Three variations of t-test can be applied depending on the scenario and testing purposes, including one-sample t-test, two-sample t-test and paired t-test. Essentially, t-test assumes the samples to (i) follow a normal distribution, (ii) have equal variance, and (iii) be independent from each other.

Considering the normality assumption (i), it is a common practice to check for data normality by Shapiro-Wilk test or Q-Q plot before performing a t-test. Notably, when the underlying data distribution is not known or not perfectly normal, given the sample size is larger than 30 and the standard deviation is known, z-test should be used in place of t-test. This is enabled by the Central Limit Theorem, which states that for a population of sufficiently large size with a finite mean and standard deviation, the sampling distribution of sample mean will be approximately normal. This also means that the z-distribution is a standard normal distribution with one degree of freedom. In contrast, the t-distribution is sensitive to the degrees of freedom and has wider tails than normal distribution, making it more stringent to reject the null hypothesis in the t-test. As a result, t-test is more effective than z-test for small samples. Alternatively, one can use a non-parametric test like Wilcoxon signed-rank test for one-sample test or Mann-Whitney U test for two-sample test when the (i) assumption is violated. If the (ii) assumption for homoscedasticity is broken, a Welch's t-test can be used as an alternative to Student's t-test as it is modified to be more robust to unequal variances.

2.2.2. Fisher's exact test

The Fisher's exact test is a non-parametric test for non-random association or overlap between two categorical samples. It requires a contingency table to represent

the frequency of each category and calculates the probability of obtaining this table among all possible permutations under hypergeometric distribution. As this step may cause extremely high computational costs when the sample size increases, a Chi-square test is favored for larger sample sizes or when more than two categories are involved in the experiment. Chi-square test assumes the expected counts to be greater than 5 in each cell of the contingency table. It computes the Chi-square statistics based on the expected and observed frequencies, then rejects the null hypothesis based on the Chi-square distribution, and thus is advantageous over Fisher's exact test computation-wise.

A closely-related measure for association strength between two categorical variables that requires a contingency table is the odds ratio. While the odds ratio may reflect how weak or strong the association is, it does not reveal the statistical significance of such relationship. For this reason, a Fisher's exact test may be used to determine the significance of the odds ratio and show whether the association between two groups occurs by chance.

In genome research, Fisher's exact test is intensively used for gene set enrichment analysis to determine whether a set of genes is significantly enriched for specific biological processes, molecular functions or cellular pathways. Here, the test confirms whether the overlap in gene ontology terms between a small geneset and a background set is random.

2.2.3. Kolmogorov-Smirnov test

Kolmogorov-Smirnov (KS) test is a non-parametric test used for comparing the empirical distribution functions of two samples, or the empirical distribution of a sample to the cumulative distribution of a population. Thus, KS test can be used as a normality test to check if a sample follows a normal distribution, or as a two-sample test to compare the distributions of two samples. Since KS test is highly sensitive to differences in empirical distribution functions, it is not ideal for small sample sizes. In such cases, Mann-Whitney U test is a better alternative, as it compares distributions using rank sums and is more sensitive to median shifts. For testing normality in small samples, Shapiro-Wilk test can also be used instead of KS test.

2.2.4. Correlation test

Correlation test, or correlation t-test, is used to determine the strength and direction of the relationship between two continuous variables. For normally distributed data with assumed linear relationship between the variables, Pearson's correlation is commonly selected. On the other hand, Spearman's correlation, a method based on rank order of the data points, is preferred when the data is not normally distributed or the relationship is non-linear. As a result, Pearson's correlation is more sensitive to outliers, making it a conservative test that reduces the chance of false positives. For both Pearson's and Spearman's correlation, their coefficients lie in the [-1, 1]

range, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. The test statistic for correlation test follows a t-distribution and can be used to determine the significance of the correlation coefficient.

3. Management of Bioinformatics workflow

In bioinformatic studies, workflow management, which includes effective handling of tasks like data retrieval and pre-processing or results reporting and visualization, are as equally important as the data analysis part. The workflow may contain many sequential and parallel tasks that are often interdependent and require a systematic approach to manage. For example, finding the association between diffferential gene expression and differential histone modification (Project 1) requires integrating the RNA-seq and ChIP-seq data for 5 histone patterns from 19 different cells/tissues and involves multiple analysis steps such as filtering, pairwise diffferential analysis, corrections and correlation analysis. Another instance is the detection of alternative RBP-complexes from differential gene expression with DESeq2 and differential protein analysis with PPIXpress and PPICompare (Project 6). In both studies, processing an enormous amount of highly inhomogenous data, continuously executing and optimizing the analysis pipeline while handling systemic failure are technical challenges that require sophisticated solutions (Methods section 1). This shines light on the importance of using workflow management frameworks to deploy robust and flexible bioinformatic analysis pipelines.

3.1. Aims of workflow management

Considering the tremendous abundance and variability of biological data as facilitated by modern high-throughput sequencing methods, omics-data analysis pipelines or softwares must be able to accommodate high computing standards and users demands, in terms of automatization, distributability and scalability.

Automatization Workflows can be automated as cascade of steps where the next rule is dependent on the output from the previous. This way, automation can bypass many decision-making steps to save computing costs and time, while optimizing the usage of computing resources. In reality, automization of bioinformatic workflow is a highly complicated task. It deals not only with prototyping analysis processes for biodata that are unhomogenous in sizes, formats, quality, etc., or integrating myriads of interdependent processes, but also with the need to handle exceptions and errors that may arise at any stage [57, 102].

Distributability Distributability refers to the practice that allows users to share and execute softwares or workflows with highest independency from their operating systems and platforms. Workflows can be made portable by containerization, which encapsulates the software and its dependencies into a single package to be executed on prefered infrastructures. In bioinformatics, many common analysis pipelines, such as reads trimming, peaks calling, etc., are containerized and reused as modules to build even more complex, high-ordered workflows [60, 170]. This attribute is

thus important for bioinformatic research as it enables reproducible, collaborative studies and promotes shared usage of reusable bioinformatic tools [57, 60, 102].

Scalability As the scope of projects may vary from preliminary studies to extensive large-scale investigations, analysis workflows must be designed with scalability to adapt to research demands. A scalable workflow should maintain consistent performance across datasets of various sizes, sometimes of up to terabytes [57], which is reflected through its ability to parallelize tasks, distribute computing resources and minimize numerical instability across different computational platforms during batch processing [82, 60, 170]. Scalability ensures synchronization across various scales and platforms, facilitates testing and troubleshooting of the analysis pipeline, and most importantly, enables future expansion to accommodate more complex inputs and analyses.

3.2. Bioinformatics workflow management tools

Two of the most widely adopted workflow management frameworks in bioinformatics include nf-core [74] and Snakemake [126]. The two tools differ greatly in dependencies, workflow design principles and operations [57, 60, 74, 126] as summarized below.

Table II.2: Comparing Nextflow and Snakemake workflow management tools

Nextflow	Snakemake
Automation	
Domain-specific language	
• Groovy	• Python
Jobs dependency	
• State dependency: A Nextflow workflow constitutes a set of independent processes that are not linked by data dependency, but can communicate via a *channel* to be executed asynchronously. Processes reactively listen to the respective input channel and are executed if the inputs are available.	• Data dependency: A snakemake workflow constitutes of a set of *rules* connected by data dependency. The workflow endpoint is defined by the output of a target rule, and its inputs are outputs from other rules. By this design, snakemake initially builds a directed acyclic graph where nodes and edges represent jobs and data dependencies between them.
Design patterns	

Continued on next page

Table II.2: Comparing Nextflow and Snakemake workflow management tools (Continued)

• Supports various design patterns (generic, scatter/gather, output organization, conditional execution, batch, benchmarking)	• Supports various design patterns (generic, scatter/gather, output organization, conditional execution, batch, benchmarking)
Automated unit test generation	
• Yes	• Yes
Distributability	
Modularization	
• Allows modularization to create nested workflows with Nextflow DSLv2 modular language.	• Allows integrating modular rule files, subworkflows or modules to build new pipeline.
Container integration	
• Docker, Singularity, Conda	• Docker, Singularity, Conda
Conda integration	
• Native: For a workflow, a Conda environment containing all Conda-dependencies is automatically containerized.	• Per-rule: For each rule, a container image can be created.
Standardized code linting and formatting	
• Yes	• Yes
Workflow versioning	
• Yes	• No
Scalability	
Job scheduling	
• Jobs are natively handled with HPC schedulers (SLURM, Kubernetes, AWS/Google Cloud Batch)	• Jobs are submitted to local cluster engines (SLURM) or HPC schedulers (Kubernetes) for cloud execution (GCP, AWS) via explicitly configured cluster profiles
Parallelization	

Continued on next page

Table II.2: Comparing Nextflow and Snakemake workflow management tools (Continued)

• Dataflow model enables dynamic parallel execution as soon as input is available.	• Parallelism requires explicit declaration for multi-cores execution or cluster job submission.
Caching between workflows	
• Automatic caching of intermediate results enables workflow re-execution or cache-sharing across runs/clouds/HPCs.	• Caching based on file timestamps and checksums enabled workflow re-execution but less flexible for cache-sharing.
Streaming processing	
• Enables real-time data streaming via channels	• Requires all inputs files to exist before execution begins

Chapter I.

Project 1: Association between Differential Exon Usage and Deregulated Epigenetic Marks in Development

My contribution to this project is designing and implementing the workflow for associating differential exon usages to differential histone modifications, performing data analysis to unveil the biological relevance of epigenetics-associated exon rewiring events and drafting the manuscript. This work has been published as "Do, H. T. T., Shanak, S., Barghash, A., Helms, V. (2023).**Differential exon usage of developmental genes is associated with deregulated epigenetic marks**. *Scientific reports*, 13(1), 12256".

1. Background

Alternative splicing (AS) or differential exon usage (DEU) was reported to occur in 90.95% of all human multi-exon genes [168, 125] and leads to a substantial expansion of the eukaryotic proteome [177]. AS is an integral part of differentiation and developmental programs and contributes to cell lineage and tissue identity as reported by Wang et al. for nine different human tissues [253]. Based on the transcriptomes of 15 different human cell lines, the ENCODE project reported that up to 25 different transcripts can be produced from a single gene and up to 12 alternative transcripts may be expressed in a particular cell [61].

It is well established that AS is often tightly associated with respective epigenetic chromatin modifications [6, 152, 279, 120]. A contribution of chromatin to AS was first suggested by Adami and colleagues who found that two copies of the same adenovirus genome in the same nucleus gave rise to differentially spliced RNAs [1]. Another well-documented example where H3K36me3 influences AS of a mammalian transcript is the fibroblast growth factor receptor (FGFR2). *FGFR2* was reported by Misteli and co-workers to accumulate histone modifications H3K36me3 and

H3K4me1 along the alternatively spliced region in mesenchymal cells, where exon *IIIc* is included. In contrast, H3K27me3 and H3K4me3 were found to be enriched in epithelial cells, where exon *IIIb* is used [153]. *FGFR2* is one of the rare cases where an exclusive exon switching process has been unraveled even in mechanistic terms. Precisely, in mesenchymal cells, H3K36me3 is recognized by the MRG15 protein that recruits the splicing factor PTB to the intronic splicing silencer element surrounding exon IIIB to repress its inclusion in these cells [153]. Recently, Luco and co-workers manipulated the flanks of CTNND1 exon 20 and of *FGFR2* exon *IIIb* using Crispr-Cas and showed that a single change in H3K27ac or H3K27me3 levels next to the alternatively spliced exon is necessary and sufficient to alter splicing and thereby affect EMT-related processes such as cell motility and invasiveness [218].

Multiple studies also established a relationship between AS or DEU and differentiation or development. In 2011, Kalsotra and Cooper reviewed the roles of AS in cell division, cell fate decisions and in tissue maturation [108]. More recently, Baralle and Giudice reviewed the connection between AS and cell differentiation as well as with epigenetic landscapes, and the role of splicing processes in the brain, striated muscle and other tissues and organs [15]. More focused studies addressed, for example, how the splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program that is essential for mammalian development [20] and what role AS plays in neural development [257]. Although the pairwise connections between AS and epigenetic modifications and between AS and differentiation or development have each been characterized in detail, the intertwined connections between AS, epigenetic modifications and development have apparently received relatively little attention so far. As mentioned, Baralle and Guidice summarized some work describing such an interplay in brain and general neurological development [15]. Furthermore, an interesting study from the Heller lab related the enrichment of histone post-translational modifications (hPTMs) to AS regulation during tissue development in mice. They found, for example, that enrichment of histone modifications H3K36me3 and H3K4me1 in exon flanking regions was wired to skipped exon selection with strong evidence across all investigated embryonic tissues and developmental time points [96].

How can one understand the postulated relationship between AS and epigenetic modifications in mechanistic terms? The most important region for epigenetically regulated AS was shown to be the exon-intron boundary. For example, Guan et al. reported strong association between epigenetic signals and cassette exon inclusion levels in both exon and flanking regions [144]. Along the same lines, flanking areas annotated with exon skipping and alternative splice site selection events were found to be statistically enriched with DNA methylation, nucleosome occupancy and histone modifications [280]. The considered exon flank should be of certain dimension, enabling a mechanistic crosstalk between a DNA position where chromatin reader proteins may recognize specific histone marks, and the downstream position on the synthesized and post-processed mRNA where splicing factors may bind. In a recent study based on ENCODE human data, Gerstein

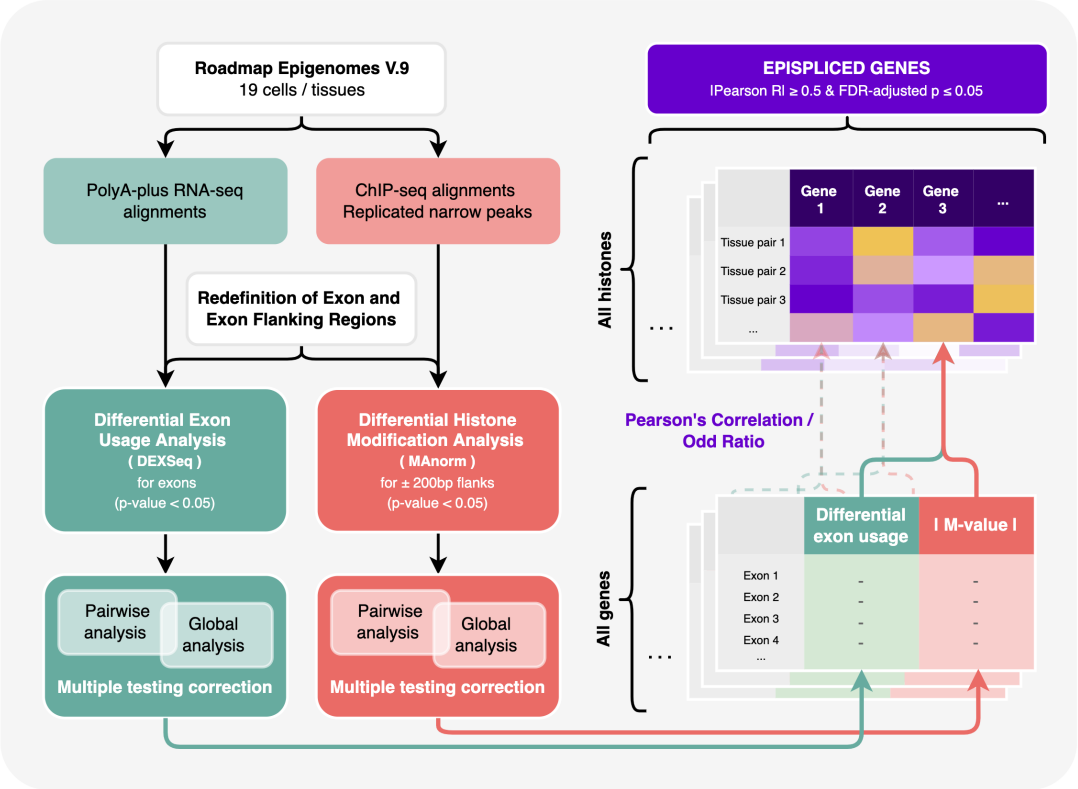

Figure I.1.: Schematic workflow to identify epispliced genes. Expression data and histone enrichment data were collected from Human Epigenomes Atlas and were subjected to differential exon usage (DEU) and differential histone modification (DHM) analysis, respectively. For each gene, we computed the Pearson correlation of the two features (DEU and DHM). Epispliced genes are required to have absolute R-value ≥ 0.5 and FDR-adjusted pvalue ≤ 0.05. Functional enrichment analysis based on Gene Ontology terms was performed against the background of all genes having either differentially used exons or differentially deregulated histone marks at the exon boundaries.

and co-workers showed that a combination of particular histone marks can be used to reliably predict using a trained machine learning classifier whether exons are included or not. Precisely, they used spatio-temporal epigenetic features extracted from exon flanks to model splicing regulation, and characterized H3K36me3, H3K27me3, H3K4me3, H3K9me3 and H3K27ac as highly influential features in the splicing regulatory model [133]. It was not the main point of our study to test again the hypothesis that one or only a few specific histone modifications mark the boundaries of either exons chosen for inclusion in a mRNA or excluded from an mRNA again. Instead, our manuscript focuses on which type of genes show differential splicing associated with deregulated epigenetic marks and in which context, particularly in cell fate transitions.

Based on data from the Human Epigenome Atlas [210] for adult human tissues and cultured stem cells, we aimed to correlating differential exon usage to epigenetic

modifications of different histone marks at the exon boundaries. The detailed workflow for the analysis performed in this study is illustrated in Figure I.1. Indeed, we found an overall enrichment of cases where differentially used exons overlap with differential histone marks. The involved genes were enriched in functional annotations related to the regulation of signaling and to developmental processes. When inspecting the overlap of such genes between different tissues and cell lines, we noticed a stronger overlap between cell lines corresponding to early developmental stages, whereas differentiated tissues had smaller overlaps. Besides a pooled analysis, we additionally present a detailed analysis of the two genes *FGFR2* and *LMNB1* (Lamin-B1).

2. Results and Discussion

2.1. H3K36me3 mark is most relevant to AS events

The first task was to prepare a suitable data set where differentially used exons (DEUs) can be clearly associated with individual genes. Hence, out of 19,240 clusters of protein coding genes generated for the DEXSeq analysis (see Materials and Methods section), we excluded 275 clusters of 679 genes partially overlapping with each other, 17 genes spanning more than one genomic region and 1,050 genes containing only a single exon. After processing this data with DEXSeq, we filtered the detected DEUs for significance, whereby only those DEUs having a $p_{FDR} \leq 0.0001$ in any of the pairwise comparisons are retained. The remaining superset of gene clusters with at least one annotated significant DEU consists of 13,837 genes. Next, we decided to focus on DEUs that were only detected in a limited number of pairwise comparisons. As revealed by the cumulative distribution in Supplementary Figure A.1A, approximately 95% of the entire DEUs library were detected in at most 25 out of 171 pairwise comparisons and are thus identified as "non-ubiquitous DEU". These DEUs belong to 9,321 genes, which are used for the global DEXSeq analysis for multiple testing correction. In total, 103,781 DEUs from 8,887 genes were identified by both DEXSeq pairwise and global analysis. Those genes make up the final dataset of interest that will be analyzed in detail in the remainder of this study. The steps to identify these "non-ubiquitous DEUs" are summarized in Supplementary Figure A.1B.

As just mentioned, all considered genes contain at least one non-ubiquitous DEU in the epigenomes that we investigated. When all differentially modified histones are pooled, the total number of coinciding DHMs and DEUs clearly outnumbered the other three categories (Table I.1). This is reflected by the total odds ratio of 3.68 (computed as $(8{,}198 \times 79{,}888) \div (5{,}149 \times 34{,}585)$ following Equation I.1). However, not all considered histone marks shared high overlap with the detected AS events. In fact, only the mark H3K36me3 $OR = 4.38$ gave a pooled OR above 1, all the other four marks had ORs under or around 1 suggesting that DEUs

		Not DEU	DEU	Baseline OR	95% CI
All histones	Not DHM	8,198	5,149	3.68^{***}	[3.54, 3.82]
	DHM	34,585	79,888		
H3K27ac	Not DHM	30,856	55,840	1.35^{***}	[1.32, 1.39]
	DHM	11,927	29,197		
H3K27me3	Not DHM	27,103	66,522	0.48^{***}	[0.47, 0.49]
	DHM	15,680	18,515		
H3K36me3	Not DHM	20,657	14,928	4.38^{***}	[4.27, 4.50]
	DHM	22,126	70,109		
H4K3me3	Not DHM	34,068	65,623	1.16^{***}	[1.12, 1.19]
	DHM	8,715	19,414		
H3K9me3	Not DHM	34,079	71,022	0.77^{***}	[0.75, 0.80]
	DHM	8,704	14,015		

Table I.1.: The number of detected DEU and DHM events in terms of overlap and non-overlap. DEU-DHM co-occurrence is measured by odds ratio (OR) with Fisher exact test (FET) significance and 95% confidence interval. OR was calculated as shown in Equation I.1. An OR greater than 1 implies a higher odd for DEU occurrence in the presence of DHM and vice versa, while OR of 1 indicates no association between the differential events. FET was used for statistical testing to determine whether the nonrandom overlap is significant (*** indicates FDR-adjusted p-value < 0.001) The 95% confidence intervals give the estimate of the precision of the ORs.

occurred rather independently from the presence of these DHMs. This matches previous reports that the H3K36me3 is most prominently associated with AS [194].

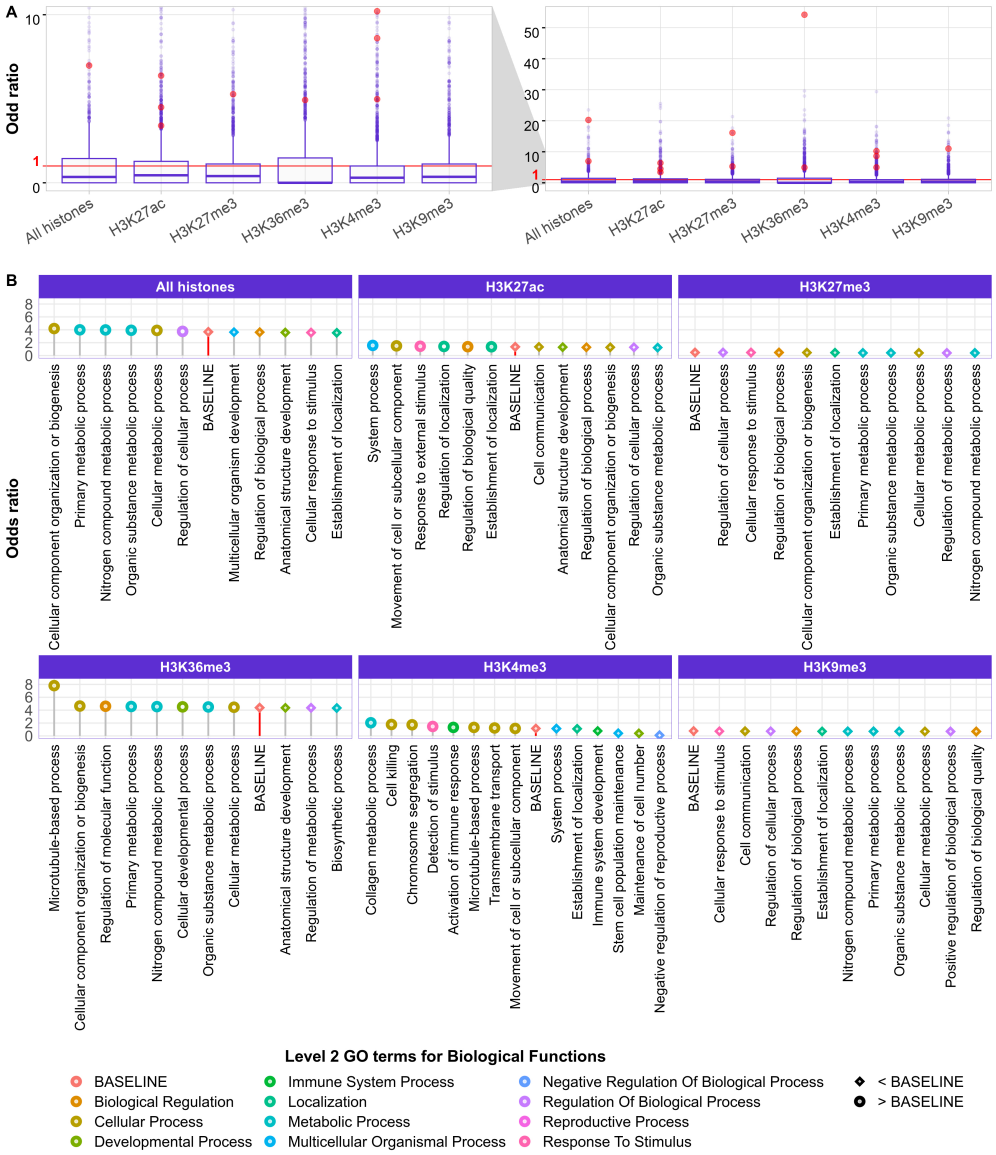

Figure I.2.: Genewise and pooled analysis of differential exon usage (DEU) and differential histone modification (DHM) co-occurrences using odds ratio (OR) and Fisher exact test (FET) significance. (A) Distribution of the OR for 8,887 genes with at least 1 DEU and 1 DHM event, whereby genes with significant nonrandom DEU-DHM overlap ($OR \geq 1$ and $p_{FET} \leq 0.05$) are highlighted in red. In (B), OR was calculated on all genes belonging to the same term at the third level of Gene Ontology (GO) terms hierarchy. These level 3 terms are colored by the GO level 2 term that they belong to. For every histone pattern, the baseline OR was computed based on all genes with at least 1 DHM of such histone type and 1 DEU event. The enriched level 3 terms with OR higher or lower than the baseline's OR are denoted by round and diamond shapes, respectively.

In total, 6,116 out of 8,887 genes had $OR \geq 1$ (Figure I.2A). Nonetheless, after applying the necessary multiple-testing correction only 11 genes among them had a p_{FET} significance below 0.05 (Supplementary Figure A.2 in Supplementary Materials). Interestingly, most of them are known to have prominent roles in cell signaling and extracellular matrix organization. Out of these 11 genes, the DEUs of two genes were associated with DHMs of all five histone marks, three genes were associated with H3K27ac, three other were associated with H3K4me3, two genes with H3K27me3, two other with H3K36me3 and one with H3K9me3. To check for a potential bias of the gene-length, Figure A.2 in Supplementary Materials plots gene-wise odds ratios as a function of exon number. Obviously, there exists a certain tendency that larger odd ratios are predominantly found for genes having fewer exons. However, the 11 genes remaining after the FET significance have quite variable numbers of exons.

We next performed the same types of analysis also for separate subgroups of genes annotated with a specific biological process term out of all level 2 or level 3 categories of the Gene Ontology. Whereas none of the subgroups annotated with level 2 terms had $OR > 1$, this was the case for several level 3 terms. Figure I.2B shows odds ratios of these collective level 3 GO terms in decreasing OR order for each histone context. In each panel, the entries labeled as BASELINE are the same values listed in Table I.1 for the superset of all considered genes. Figure I.2B illustrates that most terms with higher OR than the baselines refer to cellular processes, localization and communication, especially for the marks H3K27ac, H3K36me3 and H3K4me3. In the scenarios where all histone marks were considered altogether or for H3K36me3 mark, the most enriched terms are associated with growth and development.

2.2. Histone patterns and splicing decisions are tightly connected in *epispliced* genes

As previously shown, differential placement of chromatin marks has a substantial impact on post-transcriptional processes including alternative splicing [96, 144, 278, 71]. With the aim of delineating their role in human development, we now identified those genes where differential exon usage is linearly correlated to the degree of histone mark deregulation at the exon boundaries. These regions, alternatively referred to as "flank" or "flanking regions", were defined as a span 200-bp up- or downstream from the exon start or end points as suggested in related studies (Figure I.7B) [96, 278, 184]. Based on the analysis of odds ratios presented above, we conclude that there exists in fact a significant association between DEU and DHM at least for a fraction of genes. Only the top 5% of the investigated genes had a DEU-DHM correlation higher than the absolute Pearson correlation coefficient $|R| = 0.5$, see (Figure I.3A). Hence, we used this value as suitable threshold to identify "epispliced" genes.

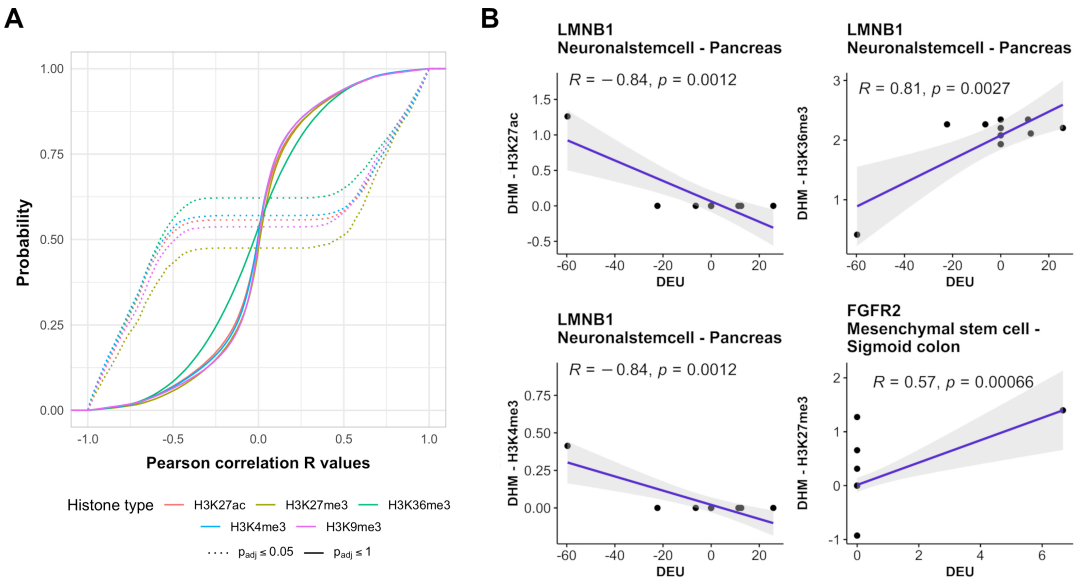

Figure I.3.: Linear association between differential exon usage (DEU) and differential histone modification (DHM). (A) Cumulative distribution of Pearson correlation between DEUs and DHMs for all genes. The dashed lines illustrate the cumulative distribution of all genes with the top 5% highest correlation level (FDR-adjusted p-value = 0.05). (B) Pearson correlation between differential exon usage (detected by DEXSeq) and deregulation of histone marks ($M - values$ detected by MAnorm) for the two genes *LMNB1* and *FGFR2*, respectively. For *LMNB1*, exon usage and histone modification were compared between neuronal stem cell and pancreas, and between mesenchymal stem cell and sigmoid colon for *FGFR2*.

As examples, Figure I.3B shows the Pearson correlations for the gene *FGFR2* in mesenchymal stem cells against sigmoid colon tissue and for *LMNB1* in neuronal stem cells against pancreas tissue. For *FGFR2*, AS events and splicing mechanisms have been frequently discussed [152, 279, 153, 108]. We found that DEU of *FGFR2* was positively correlated to H3K27me3 DHMs with a coefficient of 0.57. One may question whether the DEU-DHM correlation plot for *FGFR2* (Figure I.3B) represents a meaningful linear relationship. We note, however, that generally DEU only affects at most 3 exons of a gene (43.30% of all DEU cases detected from 171 pairwise comparisons). As a result, correlation plots such as the one shown for *FGFR2* are quite common. In this plot, the non-zero correlation is basically due to one point with high DHM and high DEU values. Note, however, that our DEU-DHM association analysis is based on data measured for multiple samples each and we only consider values that remained after statistical significance testing. Hence, this point does not represent an outlier that typically confuses Pearson correlation analysis, but it is a true data point. Those points having $DEU = 0$ but different DHM values are typical non-DEU exons where epigenetic deregulation may also affect other processes. As second example, we show the gene *LMNB1* which had relatively high correlations between DEUs and this was the case for 3 out

of 5 considered differential histone marks ($R = -0.84, 0.81$ and -0.84 for H3K27ac, H3K36me3 and H3K4me3, respectively), which only occurred for a few genes. For comparison, Podlaha et al. reported Spearman rank correlations of protein-coding genes between H3K36me3 enrichment level and splicing exon inclusion rate of at most 0.36 for six normal human cell lines [194]. From now on, we will use the term "epispliced genes" to refer to genes showing significant absolute correlations greater than 0.5 (threshold obtained from $p_{FDR} \leq 0.05$, Figure I.3A). For clarity, we accompany Figure I.3B with a more detailed representation of the transcript architecture of the same two genes *FGFR2* and *LMNB1* in Figure I.4.

Case study 1: *FGFR2* gene

For *FGFR2*, DEUs between mesenchymal stem cells and sigmoid colon were initially detected for exons 2, 5 and 22. These exon numbers refer to a flattened exon model used for the DEXSeq analysis. However, exons 2 and 5 were subsequently excluded from the analysis since they are the first exons of several transcript variants. The only detected AS event in this comparison was the exon skipping at exon 22 in mesenchymal stem cells as shown in the orange panel for exon usage of Figure I.4A. When mapped to the NCBI reference genome, exons 21-23 correspond to exons IIIa, *IIIb* and *IIIc* discussed in previous studies [64]. In fact, those three exons are known to determine the two most prominent, mutually exclusive transcripts of *FGFR2*, namely *FGFR2b* and *FGFR2c* (as shown by V2 and V1 transcripts in the "Transcripts" panel of Figure I.4A). The inclusion of exon *IIIb* (exon 22 in our annotation) and exclusion of *IIIc* (exon 23) give rise to the epithelial-specific *FGFR2b* variant, whereas the opposite case results in the mesenchymal-specific *FGFR2c* variant [20]. Using DEXSeq, we found strong evidence for the dominance of *FGFR2c* in mesenchymal stem cells and of *FGFR2b* in sigmoid colon tissue. Meanwhile, MAnorm detected a significantly higher H3K27me3 signal in mesenchymal stem cells (red) at the flank regions of exons *IIIb* and *IIIc* that in fact coincides with the recent experimental findings by Luco and coworkers [218]. These authors also reported anti-correlation between the inclusion level of exon *IIIc* and the localized enrichment level of H3K27me3 during epithelial-mesenchymal transition that is evident in our comparison between mesenchymal stem cells and sigmoid colon (Figure I.4A - dashed black box). Additionally, the enrichment of the H3K27me3 mark at the *FGFR2* promoter has also been linked to the down-regulation of exon *IIIb* [112]. In their previous study, Luco et al. reported an enrichment of H3K36me3 over the length of the *FGFR2* gene that is linked to exon *IIIb* skipping in mesenchymal stem cells [152, 153]. They speculated that the histone mark represses exon inclusion by recruiting two RNA-binding proteins MRG15 and PTB to the splice sites. Here, even though such enrichment can be observed in the last panel, we did not find a significant correlation between differential modification of H3K36me3 and *FGFR2* alternative exon usage. Nonetheless, it has recently been confirmed experimentally that the

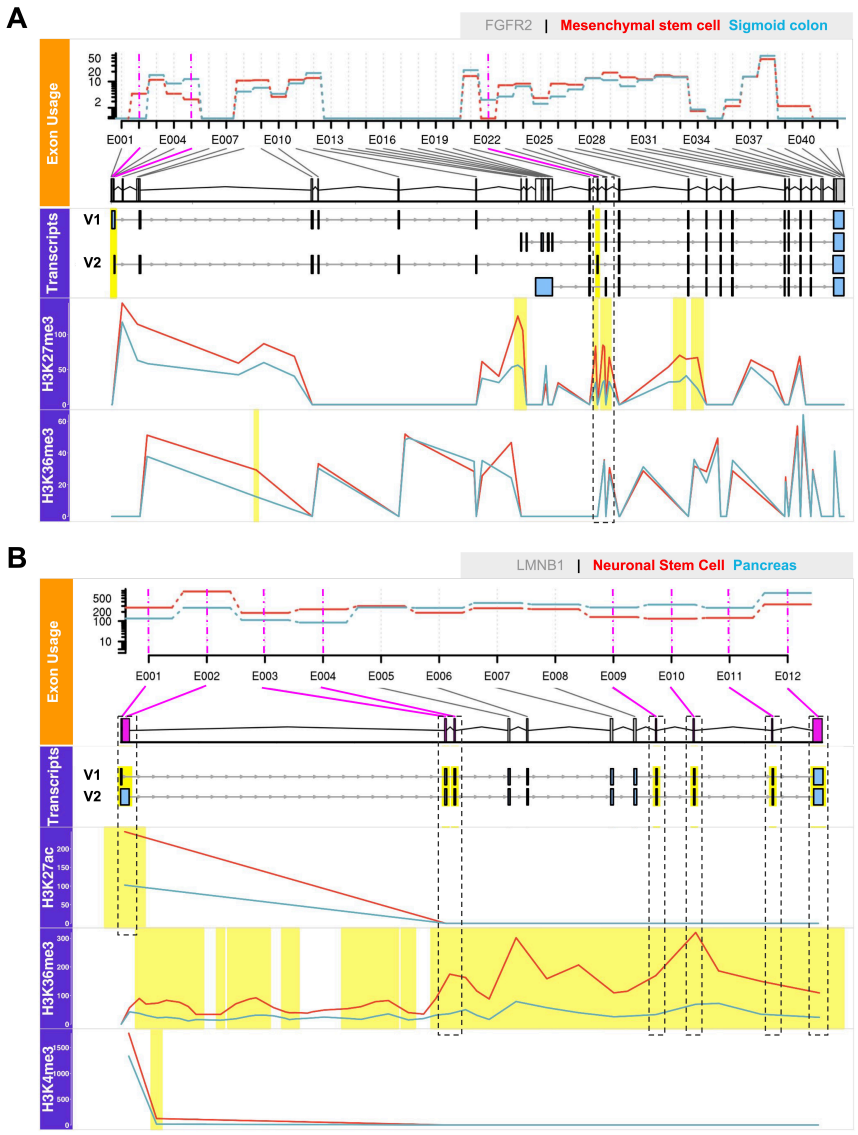

Figure I.4.: Two case studies (*FGFR2* and *LMNB1* genes) of deregulated epigenetic modifications associated with alternative splicing. The upper, orange-labeled panels that illustrate exon usage were produced by the DEXSeq package. In this case, they highlight the differential exon usage (DEU) of the *FGFR2* gene between mesenchymal stem cell and sigmoid colon (A), and of the *LMNB1* gene between neuronal stem cells and pancreas (B). Significantly differentially used exons (FDR-adjusted p-value $p_{FDR} \leq 0.05$) are marked in pink. The panels shown below that are colored in violet illustrate the association of DEUs and epigenomic modifications for the same two genes and tissues. Regions highlighted in yellow represent exons with DEUs identified from DEXSeq ($p_{FDR} \leq 0.05$) and significant differentially abundant peaks of histone modifications detected by MAnorm ($p_{FDR} \leq 0.05$ and $|M - value| \geq 1$). The boxes in the "Transcripts" panel show transcript variants found in the investigated cell types as retrieved from NCBI Refseq. The figure was generated using the Gviz package.

localized H3K36me3 mark rarely showed correlation to the changes in exon *IIIc* inclusion level [218]. This good match with experimental findings for individual well-studied genes emphasizes the importance of genome-wide examination of histone modification in AS contexts as is done here.

Case study 2: *LMNB1* gene

For *LMNB1* (Figure I.4B), two transcript variants including NM_005573 and NM_001198557 are presented as V1 and V2 in the "Transcripts" panel. While the first variant produces an isoform that includes all presented exons, the latter yields a shorter isoform consisting of only exons 4-12 due to a different 5' UTR [103]. Here, the exons 1-4 and 9-12 are clear examples of strong differential exon usage between neuronal stem cells and pancreas. All these exons also showed significantly modified histone patterns at their flank regions as highlighted by the black boxes. Since we decided to exclude all first exons of any annotated transcript to cast aside any transcription-related histone signals, the left-most box encloses only exon 2. As mentioned above, the H3K27ac, H3K27me3 and H3K36me3 marks are significantly correlated to DEU for *LMNB1*. Figure I.4B shows these marks in the three lowest rows. The two marks H3K27ac and H3K4me3 are more pronounced around the boundaries of exon 2 that have an elevated exon usage in neuronal stem cells. Furthermore, the H3K36me3 level in neuronal stem cells (red) is higher than in pancreas (blue) at exons 6-12, which intriguingly overlaps with the lower exon usage in neuronal stem cells. Considering that the elevated usage of these exons might signify a higher abundance of the shorter variant of *LMNB1* (NM_001198557) in pancreas, the histone mark H3K36me3 could serve a substantial role in the selection of alternative isoforms in these cell types.

2.3. Histone modification influences alternative splicing in developmental genes

The main biological aim of this paper was to investigate a possible relationship between episplicing and development. Thus, we were less interested in detecting genes that are alternatively spliced in a similar manner in any pairwise epigenome comparisons. Rather, we focused on those genes showing differential exon usage coupled to epigenetic rewiring in relatively few tissue comparisons. Such a subset of genes is captured by filtering for the least ubiquitously occurring DEUs (see Materials and Methods section). Table I.2 shows the subsets of *epispliced* genes which have DEU events that were detected in a limited number of tissue comparisons (1-25 out of 171). If we find such an event in the comparison of two tissues A and B, this event is counted both for A and B. Sometimes, a gene may show correlated DEU and histone mark levels for multiple histone marks. The last column lists the total number of non-ubiquitous *epispliced* genes where these overlaps are omitted. As mentioned before, these genes contain the non-ubiquitous DEUs that appeared

in a limited number of pairwise tissue comparisons. The three stem cells including neuronal stem cells, H1 stem cells and mesenchymal stem cells featured the largest number of non-ubiquitous *epispliced* genes, whereas aorta and esophagus are among the ones having the fewest of such genes (together with psoas muscle and sigmoid colon).

After identifying *epispliced* genes for individual epigenomes, we analyzed which cell types shared the most or fewest non-ubiquitous *epispliced* genes. For this, we computed pairwise similarities between epigenomes by taking their Jaccard index (Equation I.2) based on the non-ubiquitous *epispliced* genes listed in Table I.2. As an example, the largest overlap of 628 shared non-ubiquitous *epispliced* genes exists between H1 cells and mesendoderm, while their union of non-ubiquitous *epispliced* genes is 1292 genes. This then gives a Jaccard similarity of 0.486 (Figure I.5A). The similarity values are generally not remarkably high, reflecting clear differences in isoform expression between any two cell types. On the other hand, there are also clear similarities between certain epigenome pairs. Thus, for each mark, we applied hierarchical clustering to group cell types with higher similarities into clusters. As the same time, the epigenomes were described in four ways: by potency, sample type, origin and life stage. To quantify which labeling type was associated most strongly with the clustering obtained, we computed adjusted Rand indices that quantify how well the labeling scheme matches the clustering results (Table I.3).

Figure I.5 shows a clustered heatmap of the similarity of non-ubiquitous *epispliced* genes between pairs of epigenomes. For H3K27me3 (panel B) and H3K9me3 (E), only relatively small similarities were found between all cell types. For the H3K27ac mark (panel A), the largest similarities were found between neuronal stem cells, H1 cells and mesendoderm as well as between CD4 and CD8 immune cells. Differentiated tissues showed again rather low similarities among each other and with multipotent and pluripotent cells. For all histone marks, samples belonging to the same type shared most non-ubiquitous *epispliced* genes with relatively high Rand indices ranging from 0.703 to 0.911 (Table I.3). This is reflected by the fact that all differentiated tissues were clustered together. Among these, the tissue pair CD4 and CD8 cells always shared the highest similarity. We also observed a cluster of six pluripotent and multipotent cells (neuronal stem cells, H1 cells, trophoblast or ectodermal cell, mesendoderm, mesodermal cell, endodermal cell) sharing fairly high similarity in all histone contexts, especially for H3K27ac (A). This matched the Rand indices that show high clustering purity according to potency and life stage (0.518 and 0.602) for this mark. For those two categories, the clusters in H3K9me3 were dissimilar to those found from other histone modifications, as demonstrated by the low Rand indices for potency, origin and life stage (Table I.3).

Overall, stem cells and multipotent cells shared the largest number of non-ubiquitous *epispliced* genes especially for the two histone marks H3K27ac and H3K4me3, whereas differentiated cells tended to have rather low similarities for all five histone marks. The only exceptions to this were the immune cell types CD4 and CD8 that also had high similarities for H3K27ac, H3K4me3 and H3K9me3.

Tissue	H3K27ac	H3K27me3	H3K36me3	H3K4me3	H3K9me3	Total number of "epispliced" genes (with overlaps)	Total number of "epispliced" genes (without overlaps)
Adipose tissue	425	-	-	-	-	425	425
Aorta	414	166	383	324	136	1423	1125
CD4-positive alpha beta T cell	717	310	658	554	377	2616	1875
CD8 positive alpha beta T cell	750	241	648	548	302	2489	1843
Ectodermal cell	791	286	-	575	277	1929	1463
Endodermal cell	799	315	683	630	449	2876	2073
Esophagus	470	169	420	354	198	1611	1241
H1 cell	977	538	832	677	449	3473	2457
Mesenchymal stem cell	931	333	744	679	369	3056	2264
Mesendoderm	943	271	838	667	-	2719	2033
Mesodermal cell	727	-	651	490	312	2180	1688
Neuronal stem cell	963	435	816	822	621	3668	2611
Pancreas	593	221	629	405	295	2143	1646
Psoas muscle	544	202	485	403	165	1799	1363
Sigmoid colon	517	193	480	321	164	1675	1313
Small intestine	625	223	552	338	225	1963	1483
Spleen	491	201	530	367	302	1891	1484
Stomach	532	254	530	364	275	1955	1513
Trophoblast	942	305	790	-	369	2406	1963

Table I.2.: Number of "epispliced" genes with non-ubiquitous DEU events across all cell types in different epigenomics contexts. To account for non-ubiquitous exons, the genes with alternative splicing events occurring in a limited number of (1-25) tissue comparisons were selected from the differential exon usage analysis. "Epispliced" genes are genes where exon inclusion is correlated to differential modification of either H3K27ac, H3K27me3, H3K36me3, H3K4me3 or H3K9me3. The two rightmost columns list the count of "epispliced" genes with or without inclusion of repeating cases. (-) denotes cases where ChIP-seq histone peaks data was not available.

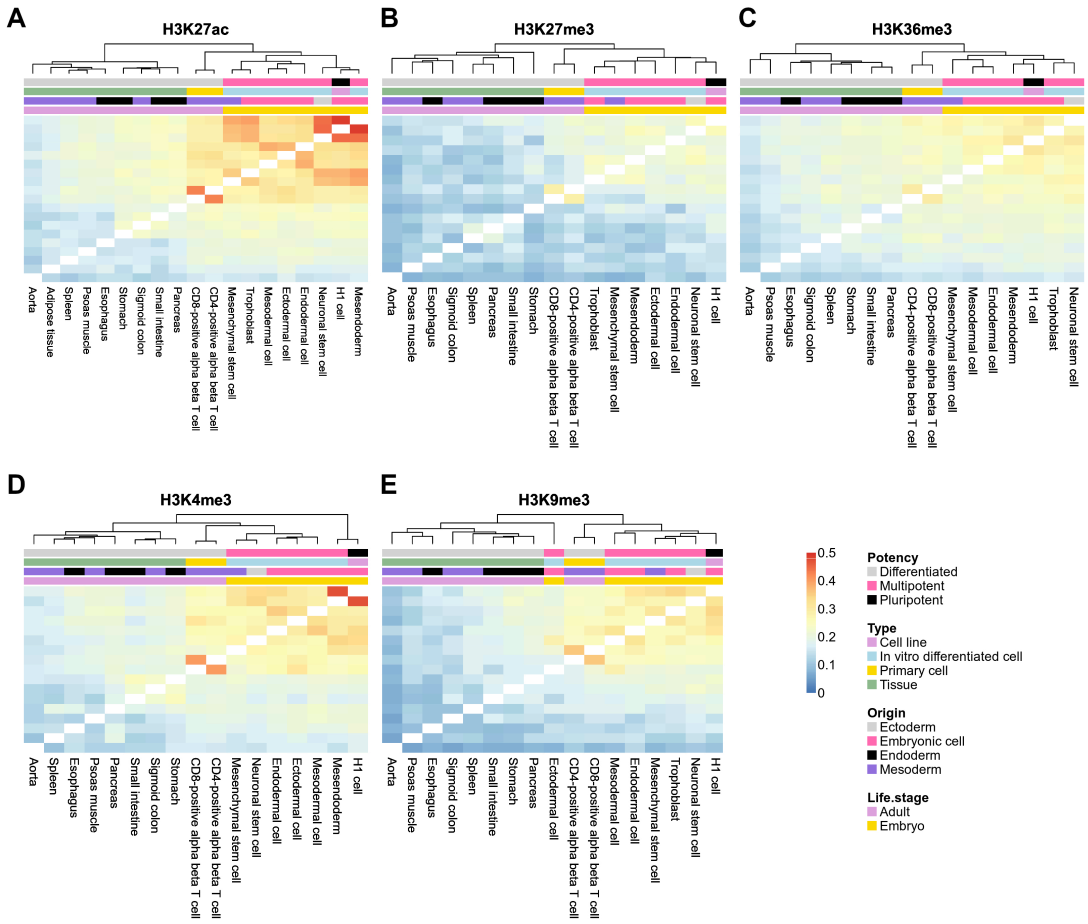

Figure I.5.: Heatmaps representing hierarchical clustering based on the similarity in non-ubiquitous "*epispliced*" genes in different epigenetic contexts. In total, 19 cell types were considered for H3K27ac, H3K27me3, H3K36me, H3K4me3 and H3K9me3 (A-E). The pairwise similarity between cell types was measured by the Jaccard index, which is the ratio between the number of mutual *epispliced* genes and the total number of *epispliced* genes in the union sets of two cell types (Equation I.2). All heatmaps use the same color scale ranging from 0 to the highest Jaccard index across all tissue pairs and for different histone marks. Investigated epigenomes were annotated on the top by their differentiation potency, type of sample, germ layer origin and the life stage when their samples were taken.

	H3K27ac	H3K27me3	H3K36me3	H3K4me3	H3K9me3
Number of available cell types	19	17	17	17	17
Potency	0.518	0.467	0.467	0.467	0.319
Type	0.911	0.903	0.703	0.740	0.711
Origin	0.349	0.314	0.314	0.181	0.160
Life stage	0.602	0.558	0.558	0.558	0.381

Table I.3.: Adjusted Rand indices measuring the similarity between heatmap hierarchical clustering and tissue label schemes. Investigated cell types were separated by potency, sample type, origin and life stage and compared to the cluster labels from hierarchical clustering, separately for differential exon usage correlated with the five histone modifications labeled in the table header. The second row lists the number of cell types analyzed for each histone mark.

One may wonder if analyzing shared DEU or DHM events alone would yield a similar clustering of tissues. This is analyzed in Supplementary Figure A.3 and Supplementary Table A.6 in Supplementary Materials. Obviously, the clustering based on either DEU or DHMs does not produce a meaningful clustering and gives only lower-valued Rand indices. In our view, this emphasizes the value of performing an integrative analysis of shared DEU and DHM events as is done in Figure I.5.

Finally, we performed functional enrichment analysis of the non-ubiquitous *epispliced* genes separately for each histone mark. Figure I.6 shows the results of gene-set enrichment analysis based on the Gene Ontology annotations of *epispliced* genes. The terms are arranged into three broad GO-SLIM categories, including cell signaling, developmental processes and cellular/metabolic processes. It turned out that the category of developmental processes played a dominant role with the highest number of terms shared between *epispliced* gene sets of different histone marks (Figure I.6C). The mark H3K27me3 seemed to have the largest contribution in this. Coincidentally, H3K27me3 also gave the second clearest separation according to sample types and origins of investigated tissues (*Rand indices* = 0.903 and 0.314, respectively). In a similar GO term enrichment analysis performed on the set of *epispliced* genes with correlation and anticorrelation separately, many of these biological annotations are found to associate with the direction of histone mark deregulation (Figure A.4 in Supplementary Materials).

The functional annotations related to the H3K27ac and H3K27me3 histone marks had the largest overlap of developmental GO terms at level 3 hierarchy (Figure I.2B). Besides, H3K27ac yielded the highest purity in clustering the tissues by potency, sample type, origin and life stage (*Rand indices* = $0.518, 0.911, 0.349, 0.602$, respectively). On the other hand, the GO terms in other categories of H3K27ac and H3K27me3 had little in common: "Epispliced" genes with deregulated H3K27me3 marks were mainly enriched with cell signaling functions (Figure I.6A), while those with deregulated H3K27ac marks were rather involved in cellular or metabolic

processes, specifically in post-translational modification (Figure I.6B). Another histone mark contributing prominently to the developmental category was H3K9me3 with many unique GO terms related to systemic development. Indeed, these results appear to have much clearer biological consequences than our initial analysis of DHM-DEU overlaps based on *OR*s, which did not show significantly enriched biological functions for many histone marks, especially for H3K27me3 and H3K9me3 (Figure I.2B). For the two marks H3K36me3 and H3K4me3 which shared less similarity in GO terms with others, epigenetic regulation of differential exon usage was also important for several rather general metabolic and signaling processes.

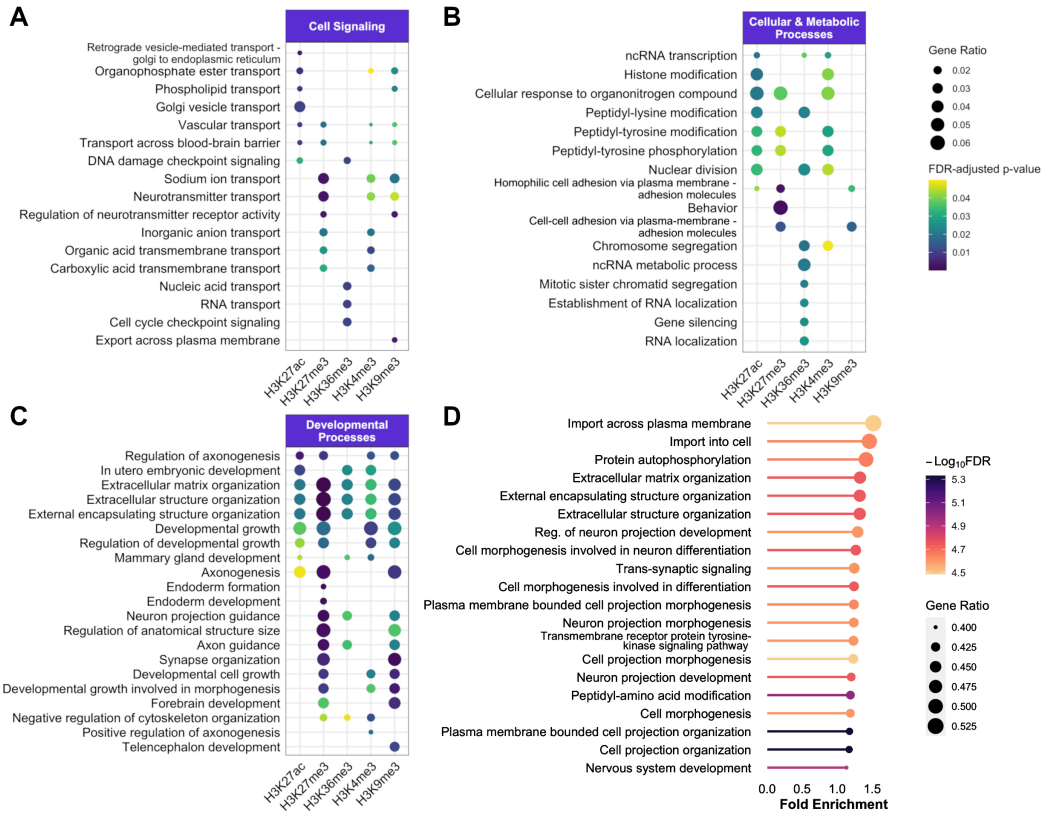

Figure I.6.: Gene ontology (GO) enrichment analysis for biological functions of non-ubiquitous *epispliced* genes for each histone type. The top enriched GO terms (FDR-adjusted p-value ≤ 0.05) annotated to *epispliced* genes that were correlated either with H3K27ac, H3K27me3, H3K36me, H3K4me3 or with H3K9me3 differential histone modifications were sorted in decreasing order of significance and of mutual functions between the histone marks. The GO terms are grouped into three main categories, namely cell signaling (A), cellular and metabolic processes (B) and developmental processes (C). (D) shows the terms enriched for the union set of *epispliced* genes detected from all histone contexts in decreasing order of fold enrichment. In the enrichment analysis, the respective *epispliced* gene sets were compared against the background set of all genes having either differentially used exons or differentially deregulated histone marks at the exon boundaries.

Interestingly, these *epispliced* genes with non-ubiquitous DEU events also had important roles specifically in post-translational modification of proteins (Figure I.6B). Upon considering the direction of DHM-DEU relationship, we also found that most of the development-related terms were enriched for genes where exon usage was anti-correlated to transcriptional silencing marks H3K27me3 and H3K9me3 or correlated to activation marks H3K27ac and H3K4me3 (Supplementary Figure A.4 in Supplementary Materials). An exception to this observation is the set of genes enriched in extracellular matrix organization which were associated with both suppressed and enhanced histone modification signals. We furthermore noticed the lack of enriched terms for transcriptional processes, despite the evident influence of histone modification on transcription [184]. This effect likely resulted from our decision to remove the first exons of any transcript variant from our analysis.

The same type of functional enrichment analysis was also carried out for the union set of *epispliced* genes detected across different histone modification contexts. The result of such an analysis revealed that cell morphogenesis and neurogenesis sub-processes have the highest fold enrichment after cell import and protein autophosphorylation (Figure I.6D). Again, the enriched terms for combined histone marks contained more significant and development-centric GO terms than those from the DEU-DHM co-occurrence analysis (Figure I.2B). One should note that other studies have already linked such histone pattern alterations to developmental processes. For instance, genes with H3K27ac-enhanced regions have been previously associated with GO functions that are characteristic for multipotent stem cells, such as anatomical structure development and nervous system development [50]. Broad H3K4me3 domains were also reported to have distinctive roles in neuronal development during stem cell and human brain tissue differentiation, which is in concordance with our findings [23]. Furthermore, the H3K4me3 and H3K27me3 promoter bivalency was established as a prominent epigenetic mechanism for lineage-specific activation or repression of developmental genes in embryonic and neural stem cell differentiation [36, 167]. For H3K36me3, we described that many GO terms contributed to cellular component organization or RNA processing and regulation besides morphogenesis, which opens up the possibility that the histone mark contributes to developmental processes via transcriptional regulation. In mouse embryonic stem cells, crosstalk between H3K36me3 and the RNA modification m6A mediates the maintenance of pluripotent state and initiates differentiation via recruitment of RNA methyltransferase complexes [100].

Finally, we add a word of caution about a possible limitation of our study where we mixed data from cell lines with data from tissues. Grouping data by "type" indeed gave rather high Rand indices in Table I.3. Interestingly, this was not the case when clustering was based on shared DEUs or DHMs alone (Table A.6 in Supplementary Materials), which speaks against a general bias of this mixing approach. We agree that, ideally, all data should either come from cell lines or from tissues. Unfortunately, to our knowledge such data is currently not publicly

available. In future, a similar type of analysis could possibly be done based on single-cell data.

3. Conclusions

Epigenetic histone marks at the exon-intron boundaries do not only play a role in defining the elements for the mRNA transcript to be expressed. Rather, as shown before, they can also contribute to regulating and controlling the relative abundance of different transcripts or protein isoforms that map to the same chromosomal region across tissues. Here, this relationship was captured by identifying genes where exon usage and histone marks at the exon flanks show concerted differential changes. We showed that there is a global enrichment of simultaneous differential exon usage and differential histone marks that is statistically significant for different subgroups of developmental genes. Taking *FGFR2* and *LMNB1* as examples, we highlighted exon-intron junctions as hot-spots for local epigenetic modifications which potentially have roles as splicing regulatory elements. Furthermore, we observed that the relationship between differentially used exons and differentially modified histone marks seems to be most prominent in early embryonic development, which suggests differential regulation across developmental stages. While this finding applied to the five studied histone marks, our assessment of *epispliced* genes also revealed further biological roles annotated to such genes for individual modification patterns. "Epispliced" genes related to H3K27me3 and H3K9me3 are mainly involved in cell signaling processes. On the other hand, the alternatively spliced genes associated to H3K27ac, H3K36me3 and H3K4me3 are potential key factors in chromatin remodeling and post-translational protein modifications, which in turn reinforce the epigenetic regulation of transcriptional and splicing activities.

4. Materials and Methods

4.1. Data Preparation

4.1.1. Transcriptomic and epigenetic data sets from the Human Epigenome Atlas

We examined the association between the differential usage of exons and epigenetic marks using RNA-seq and ChIP-seq data for histone modifications from the Human Epigenome Atlas (release 9) [210]. The data belongs to the Roadmap Epigenomics Project [210] and was downloaded from the ENCODE portal [56] at https://www.encodeproject.org/ for the histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me3 and H3K9me3. Cells or tissues that either lacked biological replicates, were flagged for poor quality controls, or had unclear developmental origin were excluded from the study. For the sake of homogeneity, only embryonic and adult samples were considered. In total, we analyzed 19 epigenomes including

one cell line, seven in vitro differentiated cells, two primary cells and nine tissues passing the described filters, each with minimum 2 and maximum 5 biological replicates. The samples were categorized by their potency, the life stage at their harvest time and the germ layer from which they arise. Table A.1 in Supplementary Materials lists the tissues and cell lines included in the current analysis, while metadata reporting all retrieved samples in details with regard to sources, biosample types and used parameters for bio-assays can be found in Tables A.2- A.4.

4.1.2. Annotation of gene body and flank regions

The gene components of interest were annotated based on the NCBI human reference genome GRCh38. The GTF-formatted reference files were retrieved and flattened following Anders et al [10]. In the first step, we excluded overlapping genes that share at least one exon to avoid misannotation when mapping differential events to the reference genome. Instances of duplicated genes, genes spanning more than one genomic region and single-exon genes were discarded as well. Next, we extracted the unique exons and defined new gene clusters based on these exons using the HTSeq package [9]. If any two exons from different transcripts of the same gene were mapped to the same genomic region, they were rearranged by HTSeq and assigned to a new non-overlapping classification of exons that mapped to that region (Figure I.7A). These redefined exons and gene clusters were subjected to differential usage analysis by DEXSeq in the subsequent step [10, 9].

As introduced before, we assume a mechanistic foundation for epigenetically regulated splicing events that implies the crosstalk between splicing factors at a specific splice-site and the chromatin readers that are recruited in the vicinity. The effective range where such crosstalk is highly probable are termed “exon flanks” and were defined as 200-bp up- and downstream from an exon’s start or end sites (Figure I.7B) as was done in previous studies [96, 280, 97]. Data annotation for differentially modified histones was performed using the *intersect* command from the package BEDtools [198]. Note that the differential signals were annotated using the flattened exon model that is explained previously in this section.

4.2. Differential analysis

4.2.1. Differential exon usage analysis

For the quantification of exon usage deregulation, the transcript and exon abundance in the polyA-plus RNA-seq alignment files were taken from the ENCODE database in BAM format. These BAM files were sorted lexicographically and converted to SAM format via SAMtools [139]. Using HTSeq, we obtained the read counts for flattened exons in each replicate of a sample from SAM files and used those as input for DEU analysis with the Bioconductor package DEXSeq [10] for all possible pairs of samples between the 19 epigenomes.

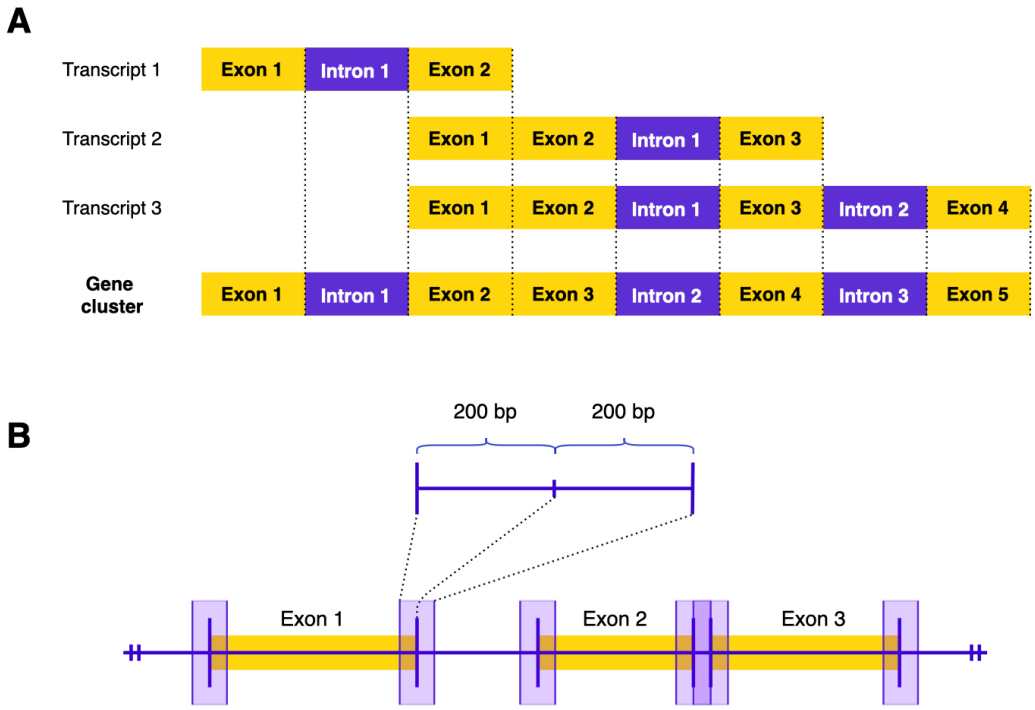

Figure I.7.: Redefinition of exons and exon flanks. (A) Overlapping transcript variants of a gene are collapsed and numbered in the flattened gene cluster following the strategy of Anders et al. [10]. Based on the read counts annotated to such redefined exons, DEXSeq compares the normalized exon usage between a tissue pair and determines differential exon usage (DEU) events. (B) Differential histone marks (DHMs) were detected by the tool MAnorm and annotated to exon borders (exon flanking regions), which were defined as the 200-bp regions around exon-intron junctions. These are flattened exons that are redefined following the scheme explained in (A).

In a pairwise comparison and for each exon, DEXSeq returns a statistic for differential usage and an FDR-adjusted p-value (p_{FDR}). The threshold of 0.05 was used to define significantly differentially expressed exons. Since we focused on the impact of DHMs on alternating splicing activity, we excluded the first exon of any transcript from the DEXSeq results, assuming that these are cases of alternate promoters where transcriptional regulatory effects of the investigated histone marks are more dominant [184].

4.2.2. Differential histone modification analysis

As materials for the analysis, we procured the GRCh38-assembled and BAM-formatted alignment files and the BED-formatted replicated or pseudo-replicated peak files from histone ChIP-seq analysis for the five mentioned histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me3 and H3K9me3. If multiple alignment files or peak files exist for specific histone type and epigenome, they were merged

using *merge* commands from SAMtools or BEDtools, respectively. To account for potential technical noise in the data and identify differentially modified histone regions, we modeled the epigenomic read counts using regression analysis in a pairwise manner across all epigenomes with MAnorm [219]. MAnorm returns the log2 ratio of read density between two samples ($M - value$) and a p_{FDR} which we subsequently mapped to the flanking regions of each exon in the reference genome. The criteria for a flank to be differentially modified $p_{FDR} \leq 0.05$ as well as $|M - value| \geq 1$

4.2.3. Multiple-comparison correction

The results from the pairwise comparisons across 19 cell types needed to be subjected to a multiple-testing correction to avoid an accumulation of false positives. This correction was implemented in the following manner: First, we computed the frequency of an exon having significant differential usage ($p_{FDR} \leq 0.0001$) in one or more of the 171 pairwise comparisons across 19 tissues. As revealed by the cumulative distribution (Supplementary Figure A.1A in Supplementary Materials), about 95% of the respective individual exons have DEUs in only 1-25 comparisons. Those exons were labeled as "more tissue specific" due to their non-ubiquitous occurrences. For all following analyses, we only considered the set of genes containing such exons. For this restricted set of genes, we performed a "pooled" DEXSeq analysis using the full collection of samples belonging to all 19 selected cell types. This analysis reports all exons that are differentially used in at least one sample with respect to all other samples, as opposed to the previous pairwise DEXseq analysis. Performing this pooled analysis with DEXSeq on all exons for 171 pairwise comparisons would have been computationally prohibitive as observed in a preliminary test for a small subset of the data. Based on this integrated analysis, we identified all individual exons showing "pooled" differential usage with $p_{FDR} \leq 0.05$ and filtered the results of the pairwise comparisons by keeping only these "overall significant" DEU exons. In the final dataset, we retained their DEU values from the pairwise comparisons, while setting the true values of non-significant exons to zero.

A multiple testing correction was likewise applied to the differentially abundant histone peaks that had been annotated to the exon flank regions of AS genes. For each region and each pairwise analysis, we retained the peak with the highest significance annotated to that region and performed an FDR correction on the results from all possible pairs. As significant DHM events, only those peaks with $p_{FDR} \leq 0.05$ were retained.

4.3. Identification and analysis of genes with strong DEU and DHM association

4.3.1. Overall and genewise co-occurrence of DEU and DHM

Previous work suggested that alteration of histone modifications contributes mechanistically to alternative splicing [6, 152, 279, 120, 1, 153]. Hence, we first identified those exons where both types of rewiring events coincide. The frequency of such DEU-DHM co-occurrences was quantified by odds ratio (OR) as defined in Equation I.1.

$$OR = \frac{DEU \cdot DHM \times \neg DEU \cdot \neg DHM}{DEU \cdot \neg DHM \times \neg DEU \cdot DHM} \tag{I.1}$$

here $DEU \cdot DHM$ refers to the number of exons where both types of differential events were detected and $\neg DEU \cdot \neg DHM$ where none of the event types occurred. Exons with $DEU \cdot \neg DHM$ or $\neg DEU \cdot DHM$ were identified with either DEU or DHM events, respectively. An OR greater than 1 indicates a higher odd of occurrence for DEU in the presence of DHM, while ORs of 1 and less than 1 reflect that DEUs are either unaffected by DHMs or even underrepresented, respectively [236]. To determine the significance of these ORs, the p-values from Fisher Exact Tests (FET) (p_{FET}) were also computed and adjusted across all accounted exons.

For each type of histone modification, we first used a contingency table to categorize all exons based on their $DEU \cdot DHM$ overlaps to compute the OR and p_{FET} significance for the set of genes where this hPTM type occurred (Table I.1) and consider this as a "global" OR analysis. Second, we performed the analysis separately for all individual genes by means of computing gene-wise ORs and their statistical significance. The genes with strong evidence for a nonrandom association between epigenetic marks and splicing activity were defined y $p_{FET} \leq 0.05$ and $OR \geq 1$ (Table A.5 in Supplementary Materials). Finally, we performed the same nalysis separately for all subgroups of genes annotated to separate biological process terms in the second or third ierarchy level of the Gene Ontology (GO). The point of this was to find out whether the co-occurrence of DEU and DHM was enriched or depleted in certain biological processes.

4.3.2. Combined differential expression analysis

Our next objective was to associate differential epigenetic profiles to exon rewiring of individual genes. For each individual gene and each pairwise comparison of epigenomes, we calculated the Pearson correlation between the DEXSeq-generated DEU values for all its exons and the respective $M - values$ computed by MAnorm mapped to their flanking regions (Figure I.7B). To enhance the contrast between differential and non-differential features, all DEU and DHM values with non-adjusted p-value > 0.05 were set to zero before computing the correlations. The

top 5% of genes having the highest FDR-adjusted correlation of all genes between DEU and DHM (Figure I.3A) are referred to as "epispliced genes" in our study. Figure I.1 provides an overview of the entire analysis.

We found many instances for *epispliced* genes where only one or a few exons show DEU-DHM overlaps and all other exons are annotated either to have only DEU or DHM events or even none of them. For our analysis, where we associate differential splicing with differential histone modifications, those rare DEU-DHM exons should be considered as true signals and should not be mistaken as outliers. Figure A.1C in Supplementary Materials compares results from both Spearman rank correlation and from Pearson correlation. In most cases, Spearman correlation gave slightly smaller coefficients than Pearson correlation and identified approximately half as many *epispliced* genes. However, 88% of the *epispliced* genes identified by Spearman were also identified by Pearson and all downstream analyses showed the same trends.

4.3.3. Association between *epispliced* genes and human development

For each histone modification type, we counted how many *epispliced* genes or gene clusters (identified in any pairwise comparison involving this sample) are shared between two cell types. As a similarity measure of shared episplicing between two cell types, the Jaccard index (Equation I.2) was used:

$$J(E_1, E_2) = \frac{E_1 \cap E_2}{E_1 \cup E_2} \tag{I.2}$$

here E_1 and E_2 are the sets of *epispliced* genes identified for a pair of cells or tissues.

Additionally, we quantified how well the cell type labels matched the similarity of episplicing on the basis of adjusted Rand indices. For this, the epigenomes were first annotated based either on their potency (potency), the sample type retrieved from ENCODE database (sample type), the germ layer they originate from (origin) or the life stage to which they belong (life stage). Then, we defined pairwise distances between epigenomes by subtracting their Jaccard similarity index of shared *epispliced* genes from 100%. These distances were then used for hierarchical clustering of the epigenomes. Using *adj.rand.index()* function from the CRAN package *fossil*, the matching between the true labels and *epispliced* genes-based clusters was quantified.

Finally, all non-ubiquitous *epispliced* genes (identified in 1-25 pairwise comparisons) collected for each histone mark were subjected to GO term enrichment analysis according to the biological process hierarchy of the PANTHER classification system [164]. The background gene set used for computing enrichment comprises all genes having either DEU or DHM events at their exon flank regions. GO term enrichment analysis was performed using the Bioconductor package *clusterProfiler* with a cutoff $p_{FDR} \leq 0.05$ for significant enrichment level [263]. Enriched GO terms were sorted in decreasing order of fold enrichment.

5. Data Availability

RNA-seq and ChIP-seq data used in this study are parts of the Roadmap Epigenomics Project [210] and are available on ENCODE database [56] at https://www.encodeproject.org/. The detailed descriptions on samples used for the analysis can be found in Supplementary Tables A.2- A.4. All analysis code and additional data supporting the study are accessible via https://github.com/dhtt/ENCODE_episplicing.git.

6. Acknowledgements

We thank Barbara Niemeyer, Fabian Müller, and Markus Hollander for helpful comments on the text.

7. Author Contributions

HTTD designed, implemented and performed the data analysis. SS and AB contributed to preparing data and to developing the analysis workflow. VH contributed to data analysis. All authors contributed to writing and editing the text.

8. Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9. Funding

This work was supported by a grant of Deutsche Forschungsgemeinschaft to VH via CRC 1027 (project C3).

Chapter II.

Project 2: A better brain? Alternative spliced STIM2 in hominoids arises with synapse formation and creates a gain-of-function variant

In this project, my contribution includes conducting the bioinformatic analysis to detect the presence of the new STIM2 variant and its abundance in public data. This computational analysis is presented in the following section, while the experimental part developed and written by our collaborators is not included here but can be found in the bioRxiv preprint "Poth, V., Trang Do, H. T., Förderer, K., Tschernig, T., Alansary, D., Helms, V., Niemeyer, B. A. (2023). **A better brain? Alternative spliced STIM2 in hominoids arises with synapse formation and creates a gain-of-function variant**. *bioRxiv*, 2023-01". I also developed a light version of RNA-seq analysis in Nextflow for our collaborators, with the goal of providing a simple, direct and reproducible way to process RNA-seq data and prepare mapped reads in suitable formats for IGV visualization [https://github.com/dhtt/SpliceView.git].

1. Introduction

Alternative splicing takes place in about 95% of human genes [185, 253]. With the rapid development of RNA microarrays, bulk RNA sequencing and sc-RNA sequencing technologies, the complexity of regulated and dynamic alternative splicing is just beginning to be understood [243, 161, 222] and it becomes increasingly evident that resulting alternate protein functions can shape both the physiology and pathophysiology of organisms [157]. Proteins derived from alternative splicing can alter infrared sensing in bats [87], sex determination in fruit flies [22] as well as change cell proliferation, cancer and neurological disorders [157]. So perhaps it is not surprising that a ubiquitous mechanism essential for regulation of cellular Ca2+ homeostasis and transcription factor activation is subject to cell-type specific alternative splicing. Store-operated Calcium Entry (SOCE) is triggered when activation

of cell surface receptors induce depletion of Ca2+ from the endoplasmic reticulum (ER). The decrease in luminal Ca2+ is differentially sensed by STIM proteins with STIM2 -having a reduced EF-hand affinity-responding to small changes in intraluminal Ca2+ whereas STIM1 requires more substantial depletion [32, 230]. Activated STIM molecules gather at ER-PM junctions where they bind and activate ORAI ion channels residing in the plasma membrane (reviewed in [196]). STIM genes (STIM1 and STIM2), have a related genomic structure with 12 conventional exons [261] and long introns harboring partly unknown small exons (reviewed in [176]), thus may utilize alternate splicing to adapt SOCE to cell type specific needs. The first described STIM1 splice variant shows an alternatively spliced extension of exon 11, leading to the longer protein variant STIM1L found in skeletal muscle [55]. A dramatic switch in STIM2 function can be seen with alternative exon inclusion into the region encoding the ORAI ion channel activating domain (SOAR/CAD), with splice inclusion reverting STIM2 from a channel activator to an inhibitor of channel function [166, 203]. Besides the inhibitory splice variant STIM2.1 (STIM2β), we have recently described how two alternate tissue specific splice variants of STIM1, which although modifying SOCE to a much lesser degree, can profoundly influence the efficacy of synaptic transmission in a frequency dependent manner, in the case of the neuronal-specific STIM1B [201], or differentially affect gating of ORAI1, protein interactions and NFAT translocation in the case of the more broadly expressed STIM1A, [121]. This same variant (STIM1A) has also been described as STIM1β, and shown to alter glioblastoma proliferation and wound healing [264], although both of these reports postulate different molecular mechanisms leading to altered cellular function.

With the discovery of a neuron-specific STIM1 splice variant [201] and the finding that STIM2 is prominently expressed in the brain, where it contributes to hypoxia induced neuronal death [24], but also stabilizes dendritic spine formation and protects spines from amyloid synaptotoxicity [195, 232] as well as positively affects spontaneous excitatory neurotransmission and drives Synaptotagmin7 dependent neurotransmitter release [42], the aim of this study was to investigate expression and function of the alternate STIM2 splice variant STIM2.3/STIM2G.

2. Materials and Methods

2.1. Bioinformatics

The following three datasets were downloaded from the NCBI Sequence Read Archives (SRA): SRP331938, GSE181813 representing 6 male and 6 female non-alcohol use disorder postmortem control donors, SRP346150: analyzing GSE207713 with iPSC derived astrocytes from 4 control donors, and GSE188847, which analyzes aged Covid-19 unaffected postmortem controls (including 6 males, and 4 females), ages of age from 45 to 64 yrs., data years old. Additional datasets were obtained for development stages Carnegie stage 22 and

9PCW9 post conception weeks, downloaded from the Biostudies database, a data infrastructure belonging to The European Bioinformatics Institute https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-4840 [143].

A bioinformatics workflow was constructed to first quantify gene expression and to analyze splicing events for selected genes. Files in FASTQ format for all datasets were downloaded and used as input for the nf-core rnaseq pipeline to map and quantify reads abundance using Salmon's fast mapping algorithm [188]. Mapped reads in BAM format were sequentially subjected to splicing variations detection by MAJIQ software. Among the current tools for splice events study, MAJIQ is a comparably fast and reliable tool [163]. MAJIQ yields a Percent Selected Index (PSI) for each recognized and de-novo junctions as quantification measurements for splice variations, which efficiently facilitates our current study to identify and compare STIM2 alternative events across various samples. For both mapping and splicing analysis stages, the Human Reference Genome build 38 (hg38) retrieved from NCBI in GTF format was used. For easier comparison to qRTPCR data, single exon 13 splice border probabilities were averaged to give the overall exon 13 inclusion probability.

3. Results

During initial screening for novel STIM2 splice variants [166], we detected an alternative spliced exon within STIM2's critical channel activating region, leading to slightly longer STIM2.1 (NM_001169118.2) and expressed in several cell types. However, we were unable to detect any significant amount of the predicted variant STIM2.3 (STIM2G) (NM_001169117.2) in lymphocytes or cell lines. Detailed analysis of novel splice variants of STIM1, meanwhile, revealed a new STIM1 variant with an alternative exon inserted between conventional exons 11 and 12, resulting in the short protein variant STIM1B, present in neuronal cells [201]. Although the alternate STIM1 exon is not conserved in STIM2, we find that the putative alternate exon 13 also resides within the intronic region between conventional exons 11 (now 12) and 12 (now 14), located on chromosome 4p15.2; with variant STIM2.3 described according to HGVS as: NC_000004.12 (NM_020860.4): c.1763_1764 ins1764-1518_1764-1451, resulting in an mRNA coding sequence of 2061nt, compared to 2502 nt for Stim2.2 (Figure II.1A). The inserted exon results in termination of the protein after 686 amino acids and contains 12 unique amino acids. The abbreviated protein lacks 159 of the original C-terminal residues including the serine/proline rich region (SP) as well as the C-terminal poly-basic domain (PBD) (Figure II.1B). In contrast to all previously analyzed variants STIM2.1/STIM2β, STIM1B, STIM1A [121], this splice event evolved more recently and the inserted unique protein domain can only be found in Catarrhini (old world monkeys): *Theropithecus gelada* (gelada baboons) and Hominoids (apes), implying a selective advantage that arose around 20 Million years ago (Figure II.1C). In contrast to

the alternate exon B for STIM1, exon 13 also contains a short poly-adenylation site. Using conventional and splice-specific primers (Figure II.2A) as well as counting reads derived from RNA seq data (exemplary reads shown in Figure II.2B), we detected STIM2.3 (STIM2G) in postmortem human brain probes extracted from different brain regions (Figure II.2C), with variability as expected from tissues derived from different postmortem times but potentially with a slightly higher expression of both STIM2 and STIM2.3 (STIM2G) in female compared to male donors by qRTPCR analysis (Figure II.2C). We also used commercially available qRTPCR primers for general STIM1 and STIM2 detection and observed slightly increased STIM1 versus STIM2 expression in cerebellum. In addition, we also find ORAI2 as the most abundant ORAI isoform in human cerebellum (Figure II.2D). We next quantified the fraction of exon 13 inclusion, showing a slightly increased abundance in female donors (Figure II.2E). However, exon read analysis using published RNA seq data of postmortem tissues from different sources, while confirming inclusion of exon 13 in around 20 of all reads, did not reveal gender differences (Figure II.2F). To estimate expression during early development, we analyzed RNA seq data obtained from the human developmental biology resource (HDBR) database to determine exon inclusion at around 7 (Carnegie stage 22) and 9 weeks post conception and find significantly reduced exon inclusion during fetal brain development, indicating a potential role of STIM2.3 in synapse formation/pruning at later developmental stages (Figure II.2G). We did not detect splice inclusion in samples (0/4) derived from astrocytes differentiated from iPSCs, indicating that STIM2.3 splicing may indeed be neuron-specific.

4. Discussion

The present study identifies STIM2.3 (STIM2G) as a unique splice variant arising with the evolution of Hominoids and *Theropithecus*, indicating a potential selective advantage for the development or function of more complex brains. Indeed, a human-specific variant ARHGAP11B which evolved by a single splice site mutation has been shown be causative for neocortical expansion by enhancing basal progenitor generation [80, 79, 265]. Highest expression of the alternative spliced-in exon 13 is found within the cerebellum, however, we detected exon 13 expression in all investigated brain regions. Cerebellar expansion rate significantly increased relative to neocortex in the phylogenetic branch of apes compared to related non-ape branches, indicating an important role of cerebellar specialization in cognitive evolution including technical complexities such as production and use of tools and learning of complex motor skills [17], pointing to a potential contribution of STIM2.3. Generation of a stable STIM2.3 expressing cell line with subsequent analysis of its differentially expressed genes compared to the cell line expressing the non-spliced variant indeed indicates a significant upregulation of genes involved in neurogenesis, neuronal differentiation and axon development. STIM2.3 likely is

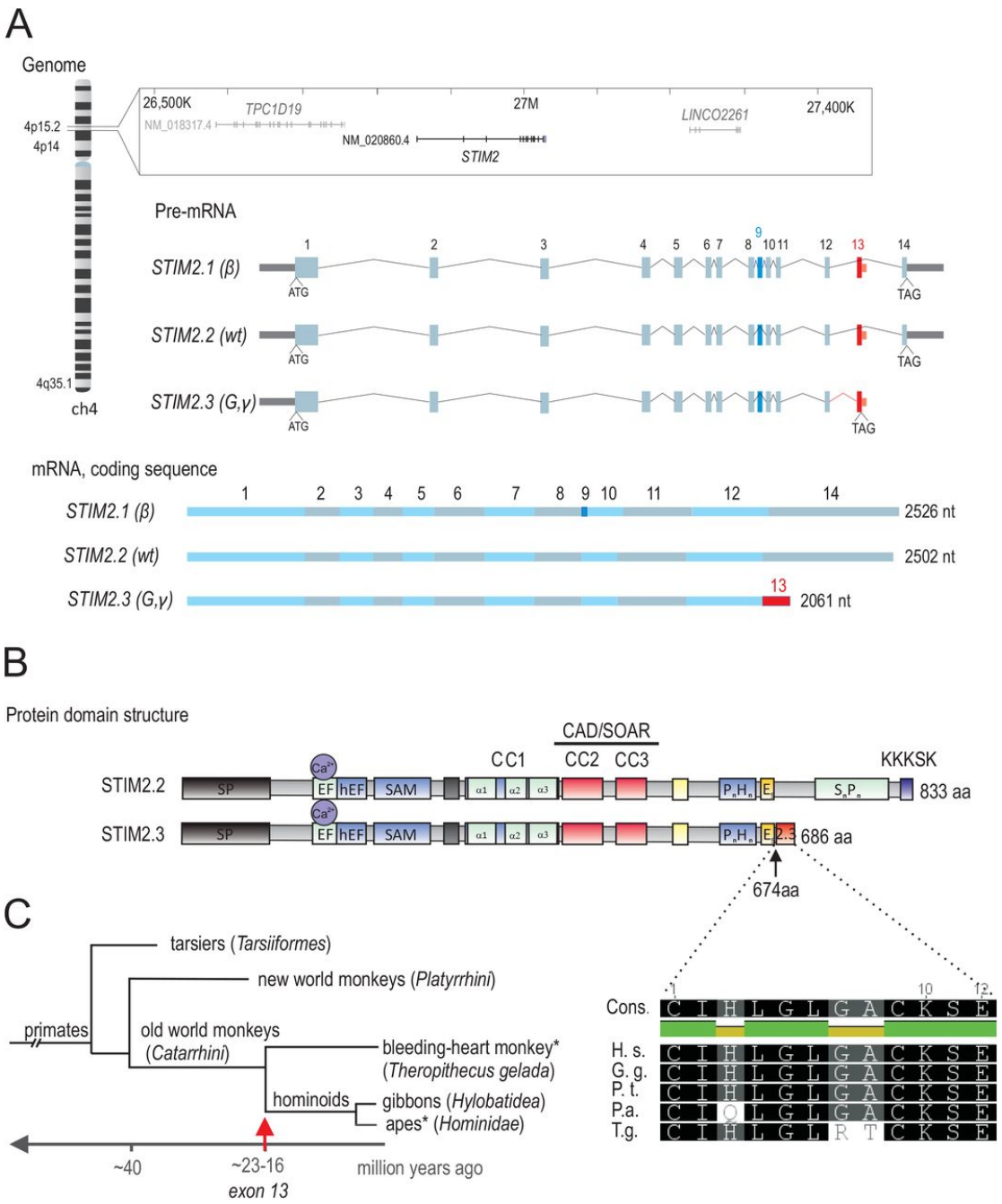

Figure II.1.: A. Genomic context and gene structure of human STIM2. Predicted pre-mRNA of the isoforms STIM2.1 (β), STIM2.2 (wt) and STIM2.3 (C). Translation start (ATG) and stop (TAG) codons are indicated. Conventional exons are depicted in light blue, STIM2.1 specific exon (9) is highlighted in dark blue and STIM2.3 specific exon (13) is highlighted in red. Lines indicate intronic regions. Coding sequence of STIM2.1, STIM2.2 and STIM2.3. Alternating light blue and grey rectangles indicate individual exons. STIM2.3-specific exon (13) is highlighted in red. B. Schematic protein structure of STIM2 displaying functional domains. STIM2.3-specific domain (red) is inserted after poly-E domain (E5, yellow) at aa 674. C. Phylogenetic tree. Red arrow indicates evolutionary splice event of exon STIM2.3. Evolutionary conservation of domain STIM2.3 in *Homo sapiens* (H. s.), *Gorilla gorilla gorilla* (G. g.). *Pan troglodytes* (P. t.). *Pongo abelii* (P. a.) and *Theropithecus gelada* (T. g.). Identical residues are within black boxes.

of pathophysiological relevance as analysis of gene splicing in postmortem probes of Huntington's disease (HD) patients, revealed that usage of the specific splice borders of alternative exon 13 showed a highly significant decrease in the PSI (percent splice inclusion: 0.175 ± 0.035 wt vs 0.004 ± 0.006 in HD) in striatum of HD patients compared to healthy controls [69] (supplemental table S7). The percent striatal exon splice inclusion correlates well with data shown in Figure II.2. A steady progression of motor dysfunction is a hallmark of Huntington's disease, however, the molecular origins of the motor dysfunction are not well understood [225] but correlate with a loss of cerebellar Purkinje neurons [224]. Utilizing patient-specific iPSC derived neurons, STIM2 has been found to mediate excessive SOCE in HD patient cells, potentially contributing to a juvenile form of HD, although splicing has not been investigated in this study [249].

5. Conclusions

In summary, our work identifies STIM2 splicing as a promising regulator of neuronal SOCE. Its evolutionary late and brain specific addition to the STIM repertoire, with increased SOCE, reduced AMPK activation and differential activation of neurogenesis genes, leads us to hypothesize that STIM2.3 may be an important regulator of neuronal differentiation, dendritic spine formation or pruning and homeostatic synaptic activity, and that mis-splicing of STIM2 as observed in HD patients may contribute to neuronal degeneration.

6. Acknowledgements

We thank Drs. Marcus Grimm and Tobias Hartmann, Saarland University for initial transfection of SH-SY5Y cells with the STIM2 targeting construct, Dr. Nicole Ludwig, Saarland University, for additional samples of RNA.

7. Author Contributions

VP conducted all imaging and microscopy experiments, biochemistry and cell line generation, analyzed data and contributed to writing. HTTD and VH provided bioinformatic analysis, KF performed RTPCR analysis, TT operated and provided postmortem tissues, DA helped with microscopy and initial experiments, BAN conceptualized the study, analyzed data and wrote the MS.

8. Competing Interests

None declared.

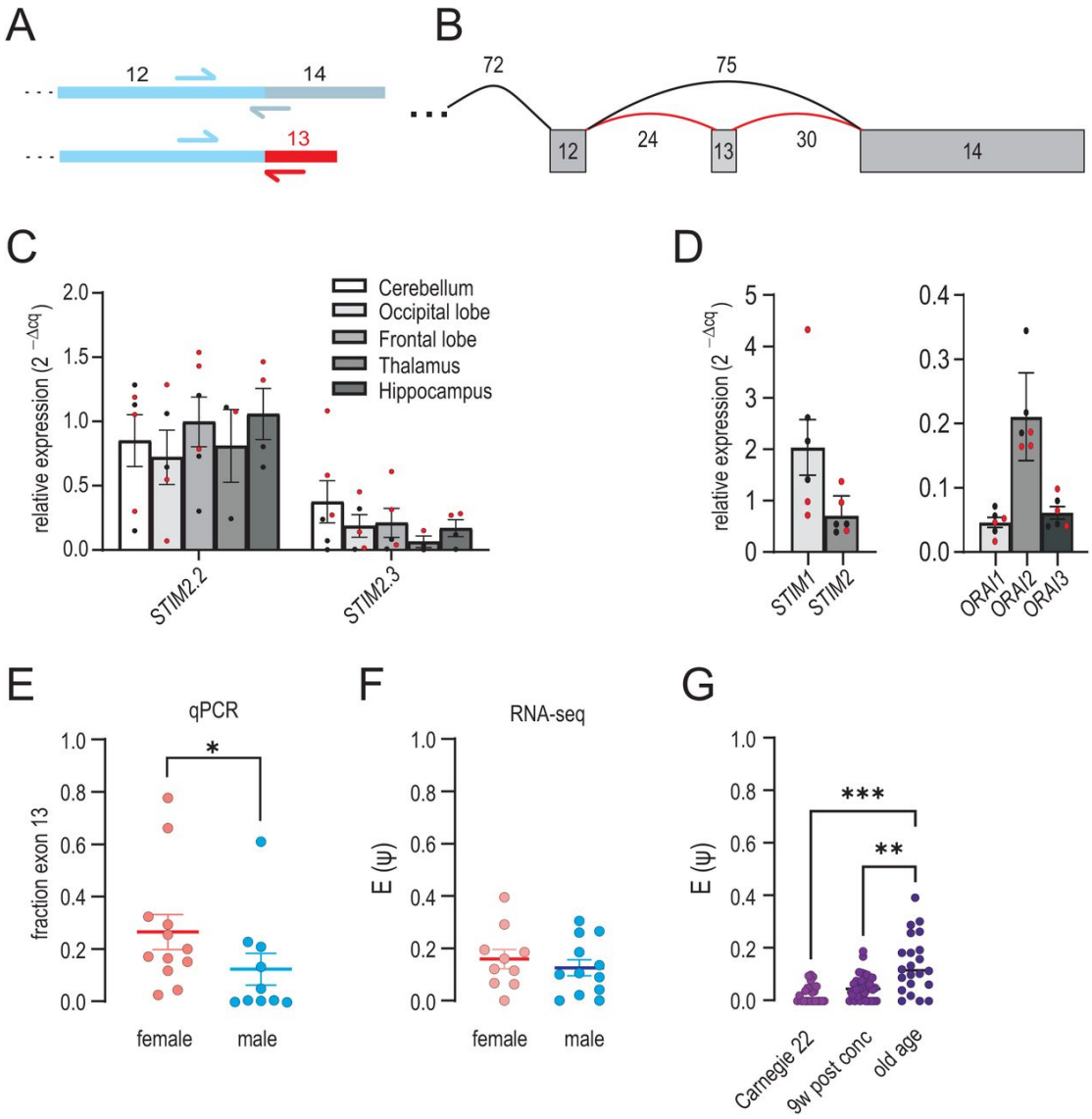

Figure II.2.: A. Schematic representation of primer annealing sites on coding mRNA. B. Representative RNA Seq read numbers of splicing of exon 13. C. Relative expression $(2-(\Delta Cq)\pm SEM)$ of STIM2.2 and STIM2.3 using splice specific primers in different postmortem human brain regions as indicated derived from six different donors. Male donors are shown in black, female donors are shown in red. D. Relative expression $(2-(\Delta Cq)\pm SEM)$ of both STIM1 and STIM2 and Orai1-3 in cDNA of postmortem human cerebellum derived from six different donors. Male donors are shown in black, female donors are shown in red. E. Quantification of fraction exon 13 normalized to sum of splice-specific and wt-specific STIM2 measured in D in female or male donors. Quantification of spliced-in exon 13 in female or male donors using RNA Seq data of cerebellum and frontal cortex. F. Quantification of spliced-in exon 13 in different developmental stages of female and male donors using RNA Seq data of cerebellum, cortex and telencephalon. G. Exon inclusion at 7 (Carnegie stage 22) and 9 weeks post conception showing significantly reduced exon inclusion during fetal brain development.

9. Funding

Funding was provided by the Deutsche Forschungsgemeinschaft (DFG) grants SFB894 (A2) and SFB1027 (C4) to BAN.

Chapter III.

Project 3: HyperTRIBE identifies IGF2BP2/IMP2 targets in vivo and links IMP2 to autophagy

This work is in preparation for submission as "Do, H.T.T, Both, S., Kröhler, T., Pirritano M., Van Wonterghem, E., Franzenburg, S., Simon, M., Biswas, J., Helms, V., Kessler S.M., Kiemer, A.K.. **HyperTRIBE identifies IGF2BP2/IMP2 targets in vivo and links IMP2 to autophagy**". My contributions to this project include conducting the bioinformatic analysis and writing the first draft for the manuscript. The *in vivo* HyperTRIBE experiments and validating assays were conducted by our collaborators and are not shown in the following sections where, only my bioinformatic analysis and relevant results are presented.

1. Introduction

The insulin-like growth factor 2 mRNA binding protein (IGF2BP2/IMP2) has originally been described as a selective binder to the 5' UTR of the insulin-like growth factor 2 (IGF2) mRNA, but has meanwhile been suggested to bind to a broad range of RNAs [201, 27, 88]. IMP2 affects the stability, localization, and translation of its target RNAs [251, 99]. IMP2 is typically regarded as exhibiting an oncofetal expression pattern [52], whereby it controls the expression and translation of multiple oncogenes in different ways, thus promoting several hallmarks of cancer, including proliferation, migration, chemoresistance, and dysregulation of cellular metabolism, in gastrointestinal and other cancer entities [116, 117, 53, 105, 114]. In addition to its crucial role in cancer, altered expression of IMP2 has been linked to the development of metabolic disorders in adult tissue [204, 131]. In vivo studies have demonstrated that transgenic IMP2 overexpression in mice can induce a steatotic phenotype [242]. Furthermore, a liver-specific overexpression model has shown that this process can also result in elevated hepatic iron deposition and increased production of hepatic free cholesterol [223].Thus, IMP2 overexpression can be considered a methodical and substantiated approach to investigate non-

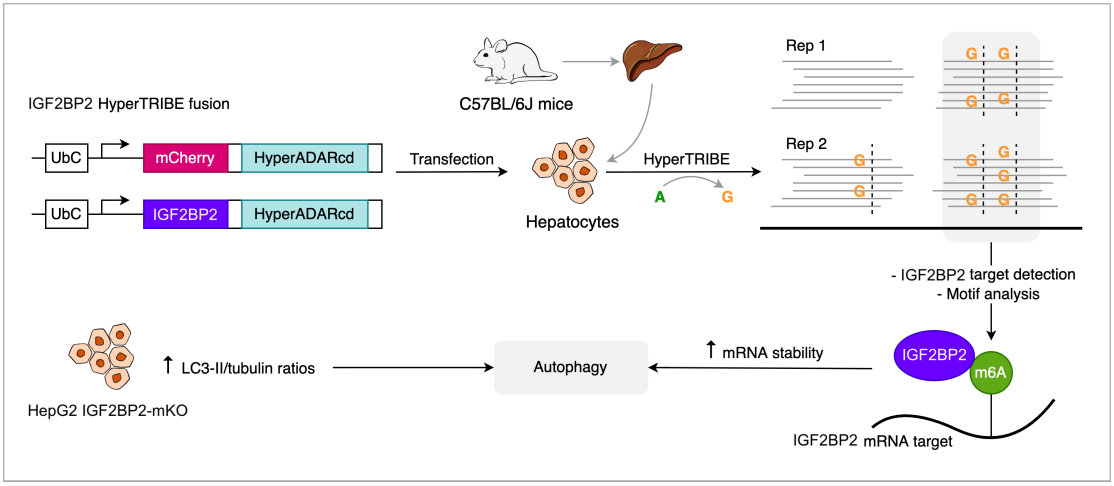

Figure III.1.: Graphical abstract

alcoholic fatty liver diseases. For a comprehensive review of these findings, see [254]. Recently, IMP2 was identified as a binder to mRNAs containing the most prevalent modification in eukaryotic mRNAs, i.e. N6-methyladenosine (m6A) [99], moving it even more into the focus of current research. Interestingly, IMP2 was also shown to stabilize USP13, which deubiquinates and thereby stabilizes autophagy-related protein 5 (ATG5) in a m6A-dependent manner [81].

All IMPs share two RNA recognition motifs (RRM1-2) in the N-terminal region and four KH homology domains (KH1-4) in their C-terminal region [175]. Several RNA recognition elements have been described to be recognized by IMP family members [27, 43, 186, 215]. The RNA recognition motif (RRM) or hnRNP K homology (KH) domains bind about 3-5 bases of RNA [83, 3], while only a few studies have focused on uncovering specific binding motifs of RNA targets. For all members of the IMP family the consensus motif CAUH was found by PAR-CLIP in more than 75% of the top 1000 targeted transcripts, followed in more than 30% by a second motif, predominantly within a distance of three to five nucleotides [88]. Concordantly, RNA Bind-N-seq applied to recombinant full-length IMP1 and IMP2 revealed CA-rich motifs that are enriched in a set of 3' UTR-enriched targets in eCLIP-defined binding sites [49]. Van Nostrand and colleagues compared 5-mers in RNA Bind-N-Seq (RBNS)-bound sequences and to corresponding enrichments in eCLIP peaks and, however, achieved different motifs using these two techniques [245].

Interactions between mRNAs and RNA-binding proteins (RBPs) are often studied by variants of the cross-linking and immunoprecipitation (CLIP) methodology, where the RBP-mRNA complexes are subjected to immunopurification utilizing an RBP-specific antibody. RBP-mRNA interactions can also be directly monitored via a protocol termed TRIBE (targets of RNA-binding proteins discovered by editing [162]) that is based on the expression of a fusion protein between an RBP and the catalytic domain of the RNA-editing enzyme ADAR (adenosine deaminases acting

on RNA). In the spatial vicinity of the expressed protein, the deaminase domain of ADAR catalyzes conversions of adenosine bases to inosine when bound to mRNAs. Subsequently, such conversions can be detected by sequencing. The HyperTRIBE protocol [200] is a refined version of TRIBE and utilizes a catalytically hyperactive E488Q mutant of the ADAR protein. HyperTRIBE was reported to work well with low amounts of starting material, but - in contrast to CLIP - cannot reveal the exact binding position of the RBP on the mRNA [200]. HyperTRIBE allows to determine cell-specific targets of RBPs [162] and has mostly been employed in cultured cells so far. However, it also proved to be suitable for the analysis of RBP target mRNAs in vivo in Drosophila [135] and *Plasmodium falciparum* [145].

Here, we aimed at determining IMP2 binding target mRNA molecules in hepatocytes as an important cell type for IMP2 action in vivo. We applied hydrodynamic gene delivery (HGD) in mice to specifically overexpress the ADAR-IGF2BP2 construct in hepatocytes. HGD "is considered the most efficient nonviral technique for gene delivery into rodents because of its high efficiency, simplicity, safety, and reproducibility" [220].

To this aim, we compared HyperTRIBE editing sites in the livers of a wild-type mouse strain to livers of mice transfected with a plasmid encoding human IGF2BP2-ADAR or a control plasmid, respectively, and also related these findings on IMP2 targets to their transcriptomic behavior. For comparison, we also determined HyperTRIBE editing sites in mouse embryonic fibroblasts (MEFs). To our knowledge this is the first study applying the HyperTRIBE approach in vivo in mice.

2. Materials and methods

2.1. Transcriptomic Data Processing

We first extracted UMIs from the first 16 bps of the raw FASTQ-formatted forward reads by umitools extract command [226]. Using cutadapt [159], we trimmed sequencing adapters from paired reads with marked UMIs while simultaneously removing poly-A tails according to the manufacturer's script (Diagenode). We also discarded reads shorter than 20bp from the final library. The trimmed reads were aligned against the *Mus musculus* reference genome version GRCm38 (mm10) from Refseq [178] using STAR alignment [62] and sorted them by samtools sort command afterwards [54]. Finally, the aligned reads were deduplicated using umitools dedup command and quantified with Salmon transcript quantifier [187].

We examined the homogeneity of WT, mCherry, and IMP2 samples by computing Pearson's or Spearman correlation coefficients and performed Principal Component Analysis (PCA) on the log-transformed gene expression data for the replicates. Pairwise differential gene expression analysis was performed with DESeq2 [151], whereby the gene expression in WT, mCherry and IMP2 samples was normalized, transformed and compared combinatorically or between each sample pair.

2.2. Identification of IMP2 binding targets by HyperTRIBE

We identified IMP2-targeted sites in the processed transcripts using the software tool HyperTRIBE [266] available at https://github.com/rosbashlab/HyperTRIBE/. At first, HyperTRIBE generates lists of editing sites identified by comparing a control against an experimental group. Then, the editing sites are filtered whether the coverage exceeds a certain threshold for the average editing percentage (0%, 1% or 5%) and are reported with genome coordinates in bedgraph files. As editing percentage indicates the fraction of reads with adenosines edited to inosines among total reads at a specific site [266] , the high-stringency threshold of 5% requires a site to be edited in at least 5% of all transcripts. Furthermore, all sites retained after the filter with the thresholds of 1% and 5% were required to have a coverage of at least 20 reads. The edited sites identified by HyperTRIBE were sequentially refined in three steps. First, we applied different replicate collapsing schemes to combine A2G sites that distinguish WT, mCherry and IMP2 samples that were identified in multiple pairwise replicate comparisons. Background correction of identified IMP2 targets was sequentially performed to remove results relevant to only control samples. Lastly, we selected one final replicate collapsing scheme with high sensitivity and robustness based on the concordance between HyperTRIBE analysis results for mouse liver and mouse embryonic fibroblast, as well as between IMP2-target genes and deregulated genes. The selection strategy for appropriate replicate collapsing scheme is shown and discussed in detail in Supplementary Information Section 5.1 “Replicate collapsing scheme selection”.

2.2.1. Replicate collapsing schemes

We compared each replicate of a sample group (WT, mCherry or IMP2) against all replicates from other groups. Based on the documentation of HyperTRIBE [266], we used several schemes to select common IMP2 targets that were identified consistently in comparisons of replicates belonging to two different sample groups. These replicate-collapsing schemes combine nearby editing sites across samples in different ways. This matches the experimental condition where ADARcd only marks the vicinity of IMP2-binding sites with A2G sites, instead of competing with IMP2 for the exact binding sites [200]. Along with the editing sites resulting from HyperTRIBE analysis, we also examined editing regions where multiple sites were detected in. Three editing regions were considered, (i) gene spans of size 10 bps, (ii) coding sequences (CDS) or untranslated regions (UTRs), and (iii) full transcripts. In case (i), The 5-bp stretch up/down-stream from a detected editing site was identified using the slop command from bedtools [199]. The editing sites or regions were either included or excluded in two different ways (see Figure III.2). In the INTERSECT scheme, we require an editing site to be present in all three replicates of an experiment group, whereas in the UNION scheme, we require this site to be found in at least 2 out of 3 replicates. The overlapping sites were extracted using the bedtools command intersect, while duplicated sites and regions

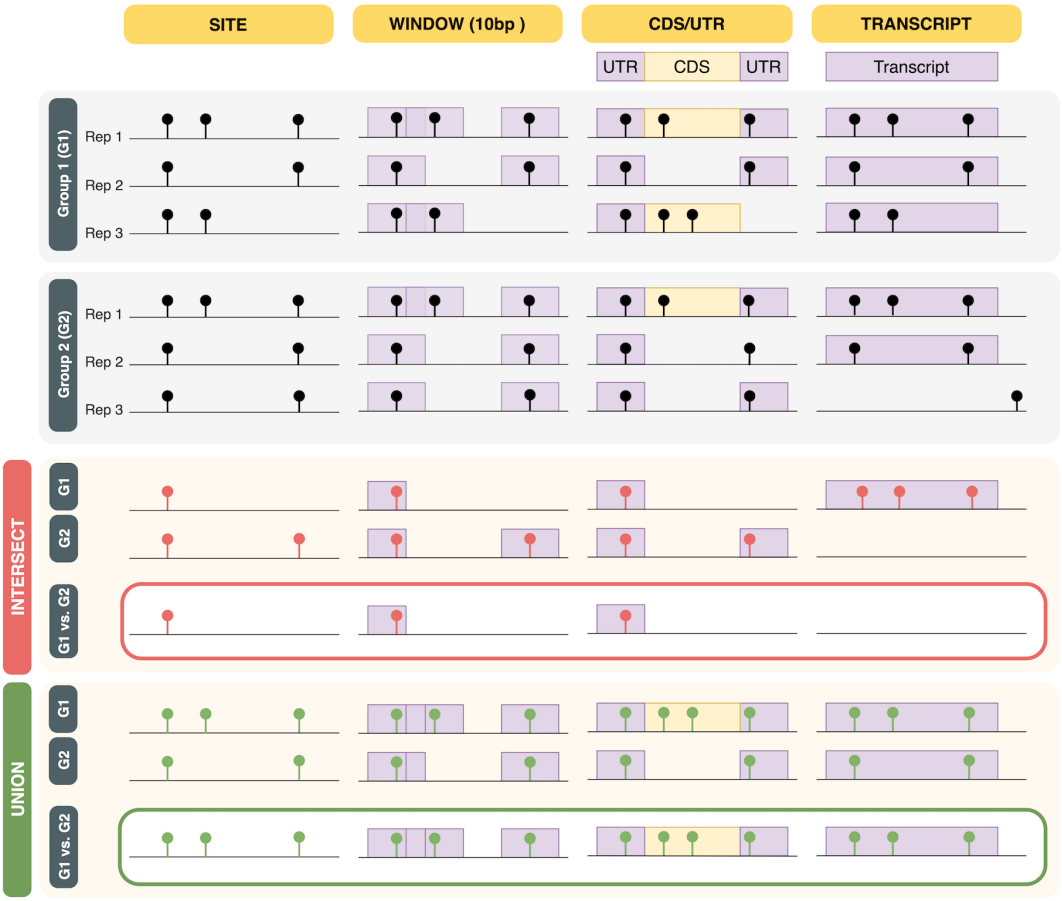

Figure III.2.: Replicate collapsing schemes. The editing sites or regions are combined across replicates by different strategies. (Middle) The red-colored INTERSECT scheme requires a site/region to be present in all replicates of a sample to be included, whereas (bottom) the green-colored UNION only requires the presence in at least two replicates. Three different genomic regions for combining the editing sites were investigated, including an overlapping window of 10 bps, CDS or UTRs, and whole transcripts. If a region contains at least one editing site in any replicate within a sample group, it will be retained or omitted according to the INTERSECT/UNION schemes.

were merged using the bedtools groupby command. As HyperTRIBE compares each replicate from the first sample type to each replicate from the second sample, we collapsed the results for the replicates of the second group, and subsequently for the first group to obtain the coordinates of editing sites/regions found between those two sample types. If a region is retained, all editing sites detected in this genome span will be included in the final set of outcomes, as HyperTRIBE can only summarize the results based on sites and not regions.

2.2.2. Background correction

After applying replicate collapsing schemes, the A2G sites or regions found in all replicate comparisons are filtered and merged for each pair of samples among WT, mCherry, and IMP2. As we are focusing on altered binding sites in IMP2 samples, we considered the set of IMP2 targets between IMP2 and WT or mCherry minus those that are only related to control groups. We refer to this group as "background corrected IMP2-specific genes" with the label "WT/mCherry vs. IMP2 - WT vs. mCherry" and focus on these genes in further analyses, see Supplementary Information section 5.2 "Background activity of ADAR". The set of differential expressed genes identified from the comparisons between IMP2 and either WT or mCherry in the previous section was corrected in a similar manner. Here, the background genes are deregulated genes detected by comparing two control samples WT and mCherry. We removed these genes from IMP2-related differentially expressed genes (DEGs) and labeled the result as "background corrected DEGs".

2.2.3. Concordance between IMP2 targets

Overlap between HyperTRIBE results in mouse liver and mouse embryonic fibroblasts. We used HyperTRIBE to identify editing sites between a control and an experimental group (IMP2 against WT or mCherry) or between the two controls (mCherry vs. WT). The perl script *summarize_results.pl* provided by HyperTRIBE generates a list of genes having edited adenosine sites, with the number of editing sites and the average editing percentage. To examine and compare the robustness of the replicate collapsing schemes, we compared the geneset derived from a specific replicate-collapsing scheme to the HyperTRIBE results for mouse embryonic fibroblast (MEF) samples (see Methods). We computed the Jaccard Index as a measurement for similarity between the sets of the same average editing percentage-threshold, as well as the number of overlapping genes and its percentage in the newly generated genesets.

Overlap between HyperTRIBE results and DESeq2 differentially expressed genes. Next, we measured the similarity between the set of HyperTRIBE genes containing at least one A2G site and the set of differentially expressed genes according to DESeq2 by the Jaccard index. We finally report and perform validating experiment using deregulated IMP2 target genes identified from overlapping HyperTRIBE and DESeq2 results in each pairwise comparisons across WT, mCherry and IMP2, and after background correction. Note that this experiment has been conducted by our collaborators, however its results have not yet been discussed and incorporated into the manuscript at this point.

2.3. Analysis of HyperTRIBE results

2.3.1. Gene ontology enrichment analysis

Gene ontology enrichment analysis for biological processes was performed on DEGs using the R package ClusterProfiler [274]. Similar to DEGs, HyperTRIBE hits were subjected to gene ontology enrichment analysis. The analysis was performed for the sets of IMP2-target genes identified when IMP2 is compared to WT or mCherry, background genes from WT and mCherry comparisons, and background-corrected IMP2-specific genes.

2.3.2. Profiling of identified editing sites

We identified preferential editing regions for detected IMP2-targets by annotating the found sites to the CDS, 3' and 5' UTR regions. As WT and mCherry were both used in our experiment as control samples for HyperTRIBE detection of IMP2-binding sites, we analyzed the concordance of the results when either one of the controls was compared against the IMP2 sample, or when the controls were compared to each other. To this aim, we plotted the resulting genes using the number of detected A2G sites detected in different pairwise comparisons. These plots reveal whether the control samples show similar differences with respect to IMP2 and which genes are affected, as well as whether the number of editing sites detected for a gene is over- or underestimated in a comparison, respectively. Using these plots, we detected genes marked with high editing frequency in the comparisons between both control groups against IMP2.

2.3.3. Effects of IMP2 binding on targeted transcripts

To check the stabilizing effect of IMP2 on targets which was reported in previous study [114], we investigated whether the expression levels of IMP2 target genes (IMP2+) would be more stable compared to of genes without A2G editing sites (IMP2-). Thus, we plotted the cumulative distribution of LFCs of these IMP2+ and IMP2- genes and performed Kolmogorov-Smirnow tests to determine whether their LFC-distributions are significantly different (Supplementary Figure B.11).

2.3.4. Motif enrichment analysis

IMP2 belongs to the set of m6A readers and exerts its stabilizing activity on transcripts through promoting m6A modification [21, 205]. This motivated us to investigate IMP2 editing sites and their proximity for enriched binding motifs. Based on a previously reported observation that 72% of HyperTRIBE editing sites overlapped with CLIP sites that were found in the 500 bp-range [266], we analyzed sequences in a [-500 bp, +500 bp] region around each A2G site. Motif enrichment analysis for each HyperTRIBE comparison (WT vs. IMP2, mCherry vs. IMP2 and WT vs. mCherry) was performed using HOMER software [90]. We also performed

differential motif discovery for the comparisons of WT vs. IMP2 and mCherry vs. IMP2 using HOMER. The identified motifs were considered significantly enriched if they had an adjusted p-value less than 1e-10 according to HOMER.

3. Results

3.1. Identification of IMP2 binding targets by HyperTRIBE

As mentioned in the methods section, we applied 16 different schemes to combine A2G sites identified in different pairs of replicates when comparing two samples (Supplementary Table B.1). Since the number of identified IMP2 targets varied greatly across the examined schemes, we selected the most robust and specific scheme based on the concordance between the resulting genes and genes with editing sites detected in mouse embryonic fibroblasts by HyperTRIBE analysis (Supplementary Table B.2) and additionally with differentially expressed genes identified by DESeq2 (Supplementary Table B.3). As we are interested in genes which distinguish IMP2 samples from WT or from mCherry but behave similarly for the two controls, we first gathered the genes from WT vs. IMP2 and mCherry vs. IMP2 comparisons and subsequently removed the genes found in WT vs. mCherry from this group Gene sets corrected in this manner are referred to as "background corrected IMP2-specific gene" or labeled "WT/mCherry vs. IMP2 - WT vs. mCherry". The same background correction procedure was applied to DEGs from DESeq2 to ensure that HyperTRIBE and DESeq2 results are comparable for concordance analysis (Supplementary Table B.3). Based on the selected replicate collapsing scheme, IMP2-target genes were defined as those with A2G editing sites in their 3'/5'-UTR/CDS that had transcripts coverage higher than 1% and were identified in at least two of out three replicates. The background corrected IMP2-target genes that resulted from this scheme showed largest overlaps with those from HyperTRIBE analysis of MEF samples and DEGs from DESeq2 in comparison to other schemes (Supplementary Tables B.1 to B.3, Supplementary Figures B.5 and B.6). The results that explain the selection of this collapsing scheme are summarized and discussed in Supplementary Information section 5.1 "Replicate collapsing scheme selection". Genes that were both significantly deregulated according to DESeq2 and contained HyperTRIBE-identified A2G sites before and after background correction are listed in Table III.1.

3.2. Profiling Of Identified Editing Sites

3.2.1. Functional enrichment analysis of IMP2 target genes and deregulated genes

Identifying the biological processes enriched among IMP2-target genes or control genes should help in better understanding how IMP2 targets potentially impact

Comparison	Deregulated IMP2 target genes	Gene count	Jaccard similarity (x1000) between IMP2 target genes and DEGs
WT *vs.* IMP2	Calu, Ccdc90b, Cdadc1, Clta, Cped1, Cplane1, Creb1, Crem, Cyp2c38, Dck, Ddx11, Ddx58, Dhx58, Dlg1, Ecm1, Elmod3, Esam, Fktn, Gm20604, Grk6, Hck, Igtp, Macf1, Mad2l2, Me2, Met, Mff, Mmrn2, Mtif3, Mug1, Numb, Parp14, Pecam1, Plec, Ppil3, Rabep1, Reln, Retreg1, Rnf38, Slc66a2, Stx5a, Syt12, Taf6, Tardbp, Tdrd7, Tgtp1, Tmem62, Tnpo1, Trim39, Vps39, Wnk1	51	27.55
mCherry *vs.* IMP2	Creb1, Crem, Fam219a, Fktn, Gm20604, Map7d1, Med16, Nfix, Prpf40b, Ptbp3, Rabep1, Smim14, Taf6, Tardbp, Tnpo1, Vps39	16	12.26
WT *vs.* mCherry	Add3, Cdadc1, Cyp27a1, Dhx58, Dlg1, Dusp12, Fktn, Gm11837, Hsd3b5, Lrp11, Mmrn2, Oasl1, Pigw, Plec, Pnpt1, Prpf40b, Retreg1, Rnf38, Slfn4, Tdrd7, Ttbk2, Ubr5, Wnk1, Zfp708	24	17.32
Background corrected IMP2 target genes	Agfg1, Ccdc90b, Creb1, Cyp2c38, Dck, Mad2l2, Med16, Rabep1, Taf6, Tardbp, Tmem62, Tnpo1, Trim39	13	12.62

Table III.1.: Genes with deregulated transcript levels and identified as IMP2 targets. Genes with A2G sites from HyperTRIBE analysis (using 1% average editing threshold, combined for mutual CDS/UTR across at least two replicates) and differential expression from DESeq2 analysis ($|LFC| < 1$, FDR-adjusted p-value < 0.05) are listed in the second column. The gene counts are reported for each pairwise comparison among WT, mCherry and IMP2 samples and for "background corrected IMP2 target genes", which are genes identified in the comparison between any control and IMP2, but not in the WT vs. mCherry comparison. The similarity between IMP2-specific genes and DEGs was measured using Jaccard Index x 1000.

cell activities through altering transcriptomic profiles. Thus, we performed GO enrichment analysis for the sets of genes which contrasted control and IMP2 samples in the HyperTRIBE experiment and differential transcripts analysis.

Figure III.3A lists the biological processes that were enriched in IMP2-bound control genes and in background-corrected genes (left side) and in deregulated genes (right side). As mentioned above, we removed control genes which belong to the WT vs. mCherry comparison to obtain background corrected genes for both the set of IMP2 target genes and of DEGs. The background IMP2-bound mRNAs were enriched in catabolic processes and cellular organization (Figure III.3A, top left). For IMP2 target genes after background removal, the majority of catabolism-related terms were pruned, whereas autophagy terms were retained (Figure III.3A, bottom left). Meanwhile, background DEGs were enriched in cellular defense mechanisms (Figure III.3A, top right). Interestingly, eliminating control DEGs resulted in omitting most terms related to immune responses, whereas GO terms for apoptosis and catabolic process now emerged with high frequency (Figure III.3A, bottom right). While IMP2 has been linked to both autophagy and apoptosis in previous studies on cancer progression [254, 267], the tight interconnection between these processes has been portrayed as sequential [158] whereby the Wnt/β-catenin signaling pathway plays an important role [155]. Additional results of gene ontology

enrichment analysis for IMP2 target genes and DEGs of each pairwise comparison across three samples are shown in Supplementary Figure B.8.

In order to experimentally validate the suggested effect of IMP2 on autophagy, we performed LC3 specific western blot analysis on the hepatocellular carcinoma cell line HepG2. Whereas IMP2 is highly abundant in HepG2 wild type cells, it is significantly downregulated in our previously published CRISPR/Cas9 generated monoallelic knock out mutant HepG2 mKO [53]. Indeed, after inhibition of late-stage fusion of autophagolysosomes with bafilomycin, HepG2 mKO showed significantly higher levels of autophagy marker LC3-II compared to the wild type cells (Figure III.4).

3.2.2. Distribution of A2G sites in genes

At the beginning of the section presenting HyperTRIBE results, we decided to select CDS/UTR as regions to combine A2G sites present in 2 over 3 three replicates of each sample group. Figure 3B shows the distribution of those sites in CDS, 3'UTR and 5'UTR regions in background and background-corrected IMP2-specific genes, while Supplementary Figure B.10 shows the same results for each pairwise comparison across WT, mCherry and IMP2 samples. Since we used the mm39 reference genome which contains overlapping transcripts, some regions may have multiple labels at the same time. However, there are only very few cases of regions annotated both as 3'UTR and 5'UTR so the overall statistics is not distorted. In each pairwise comparison between WT, mCherry and IMP2 samples, most IMP2 targets were edited in their CDS (42.82-48.49%) or in their 3'UTR (32.45-37.35%) (Supplementary Figure B.10). When filtering out A2G sites that were also detected in comparisons of the two controls (WT vs. mCherry), the proportion of sites increased remarkably to 56.63% for CDS and decreased to 23.44% for 3'UTR (Figure 3B). Meanwhile, we only found marginal changes in the percentages of other annotated regions.

3.2.3. Effects of IMP2 binding on targeted transcripts

Previous studies showed that IMP2-binding tends to stabilize targeted mRNAs [114]. We analyzed our dataset to see whether such a connection exists here as well. We divided the full geneset into two groups, genes with and without identified A2G editing sites (IMP2+ and IMP2-), and plotted the cumulative distribution of their LFC for all pairwise comparisons (Supplementary Figure B.11). The LFC distributions of background genes (WT vs. mCherry) and background-corrected IMP2-related genes (WT/mCherry vs. IMP2 - WT vs. mCherry) are shown in Figure 3C. Kolmogorov-Smirnov test was used to test whether there is a significant difference between the LFC distributions. In all cases, genes with A2G sites showed a significantly lower LFC than genes without A2G sites, suggesting mRNA stabilization in IMP2-targets, which matches previous findings (Figure 3C and Supplementary Figure B.11).

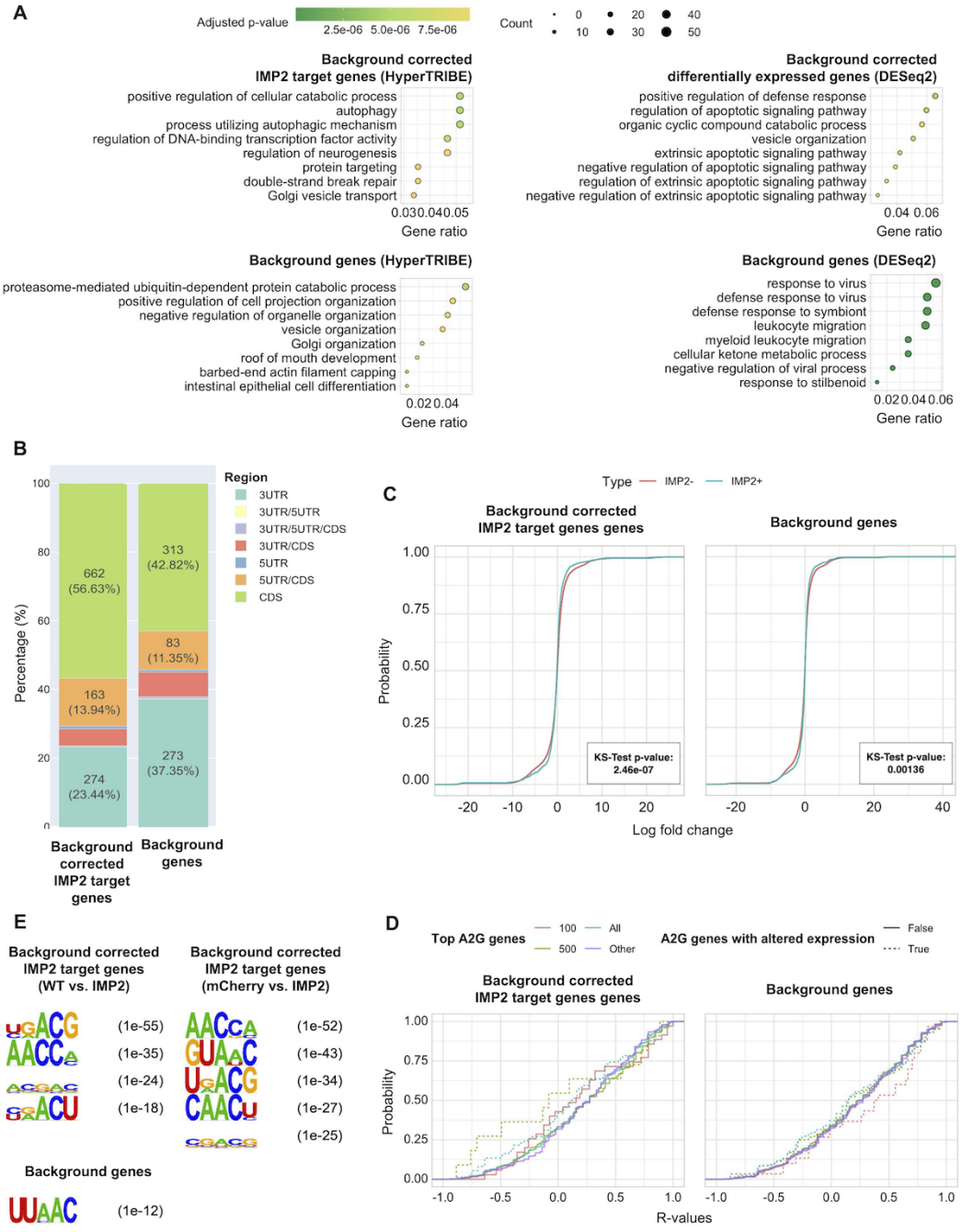

Figure III.3.: Profiling of background genes and background-corrected IMP2-specific target genes identified by HyperTRIBE. **A.** Most significantly enriched biological process terms (FDR ≤ 0.05) from gene ontology enrichment analysis of IMP2 target genes (left side) and differentially expressed genes (DEGs) (right side). Enrichments were computed with the R package clusterProfiler. **B.** Distribution of A2G sites remained for the selected replicate collapsing scheme (T=1%, UNION, CDS/UTR). The regions containing A2G sites which were identified with at least 1% transcripts coverage (T=1%) in CDS or 3'/5'-UTR in at least two out of three replicates (UNION) were annotated using *Mus musculus* reference genome mm39. *(to be continued)*

(continue) **C.** Cumulative distribution of Log Fold Change (LFC) for IMP2 target genes. The DESeq2-computed LFCs showing change in gene expression levels were plotted for genes with and without A2G sites separately (denoted by IMP2+ and IMP2- with blue and red colors, respectively). Kolmogorov-Smirnow tests were used to compare the cumulative distribution between any IMP2- and IMP2+ set of LFCs. **D.** Cumulative distribution of Spearman correlation coefficients in expression level between IMP2 and other genes. The genes were categorized into and colored according to four groups, namely all genes with A2G sites (All), top 100 and 500 A2G genes with the highest editing percentage (100 and 500), and genes with no A2G sites (Other). The distributions for A2G genes with deregulated expression identified by DESeq2 listed in Table 1 are plotted in dashed lines. **E.** De novo motif analysis and differential analysis of m6A-related motifs enriched in IMP2 binding regions using HOMER Motif Discovery tool. De novo motif analysis was performed for sequences in a [-500 bp, +500 bp] region around each A2G site for each comparison group. In differential motif analysis, all positions identified in the WT vs. mCherry comparison (bottom left) were removed from the WT vs. IMP2 (top left) and mCherry vs. IMP2 (top right) comparisons. Significantly enriched motifs (FDR $\leq$ 1e-10) are tabulated in Supplementary File 2. Only motifs similar to RAC/RACH/DRACG variants of the m6A consensus motifs and their adjusted p-values are shown here (where D = A/G/U, R = A/G, and H = U/A/C). Full reports of these analyses for all pairwise comparisons are shown in Supplementary Figures (Supplementary Figures B.8, B.10, B.11, B.12 and B.14 correspond to subfigures A-E, respectively), where background genesets are labeled "WT vs. mCherry" and background-corrected genesets are labeled "WT/mCherry vs. IMP2 - WT vs. mCherry".

In the light of this finding, we speculated that there might be a negative association in expression levels between IMP2 and IMP2 target genes. Hence, we computed the Spearman correlation coefficient between each IMP2 target gene and IMP2 itself. Similar to LFC distribution, we show the cumulative distribution of Spearman coefficients for the same 4 comparisons in Supplementary Figure B.12. Figure 3D contrasts the correlation of background genes (WT vs. mCherry) with the correlation of background-corrected IMP2-specific genes (WT/mCherry vs. IMP2 - WT vs. mCherry). To this aim, we split IMP2 target genes by whether they were deregulated ($|\text{LFC}| > 1$, FDR < 0.05) or not, and ranked them by their LFC into top 100 and 500 genes. In general, deregulated IMP2 targets differ noticeably from deregulated non-targets (purple solid lines) and from IMP2 targets that are not deregulated (panels A-D) (Supplementary Figure B.12). The distribution of deregulated top 100 IMP2 targets has a larger anticorrelation than non-targets (red dashed lines versus solid purple lines in panels A-C). For the deregulated IMP2 targets ranked in and under top 500 (green and blue dashed lines versus purple solid lines), this effect pertains to weaker degrees in all pairwise comparisons (panels A-C), but is particularly noticeable when we omitted background genes from identified IMP2 targets (Figure 3D).

We additionally performed the same analysis for 4 public datasets retrieved from GEO NIH with series numbers GSE14520 (cohorts 1 and 2), GSE57957 and GSE54236 (Supplementary Figure B.13). For GSE14520, data from both cohorts show higher anticorrelation in top 100 A2G genes compared to genes without A2G sites (red lines versus purple lines, Panel B and C for the first three groups). This is not the case for GSE57957 and GSE54236, where the majority of A2G-gene

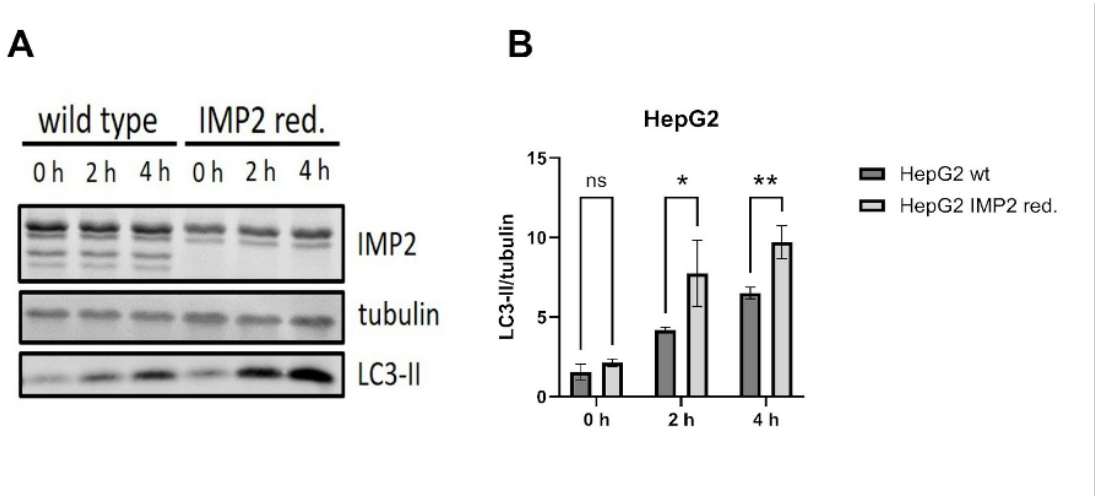

Figure III.4.: Western blot analysis of IMP2 and its splice variants as well as LC3-II in HepG2 wild type and monoallelic IMP2 knockout cells after treatment with 10 nM bafilomycin for 0 - 4 h. (A) Representative images of three (n = 3) biological replicates are shown. (B) Quantification of the LC3 II/tubulin ratios. Data are expressed as mean ± standard deviation. Statistical differences were calculated using multiple unpaired t-tests (Bonferroni-Dunn method). *p < 0.05; **p < 0.01

expressions correlate positively to IMP2 expression. For all datasets, correcting IMP2 targets revealed A2G genes with weaker correlation to IMP2. Thus, at least true for the largest data set GSE14520, human liver cancer tissues confirmed anticorrelation of IMP2 targets, again suggesting mRNA target stabilization by IMP2.

3.2.4. Motif analysis of IMP2 binding targets

To identify putative molecular details how IMP2 may stabilize mRNA transcripts, we performed motif enrichment analysis in regions of 500bp up-/downstream from an A2G site. The enriched motifs from this analysis with p-value less than 1e-10 are listed in Supplementary File 2. In total, the HOMER software identified 27 de novo motifs to be enriched around A2G sites detected by HyperTRIBE for WT vs. IMP2, mCherry vs. IMP2 and WT vs. mCherry. The number of significant hits was lowest for the group WT vs. mCherry. While more motifs were enriched in WT vs. IMP2 and mCherry vs. IMP2 groups, these two sets of results shared only one common motif (UUACC). Subsequently, sequences of genes having edited A2G sites from WT vs. mCherry were extracted from WT vs. IMP2 and mCherry vs. IMP2 groups as background for the differential motif analysis by HOMER (Figure 3E). After background removal, the similarities between these two groups became more pronounced with 13 common motifs among 34 distinct enriched motifs. Out of these 13 motifs, four (UGACG, AACCA, AACCA, UGACG) are similar to RAC/RACH/DRACG variants of m6A consensus motifs [11, 111, 174]

(Supplementary Figure B.11 A-C). In total, differential motif analysis identified 7 out of 34 motifs that resemble m6A motif variants (Figure 3E).

4. Discussion

RBP target discovery largely relies on CLIP-based methods that feature high specificity and robustness. However, immunoprecipitation experiments require large amounts of input materials and are limited by the efficiency of RBP-protein cross-linking, as well as by the sensitivity of transcripts to enzymatic activity in the subsequent RNase digestion step [91]. To overcome these shortcomings, the HyperTRIBE protocol utilizes a fusion construct between an RBP of interest and the enzyme ADAR. In subsequent sequencing, it detects readouts with modified ADAR-deaminated adenosine-to-inosine bases (A2G sites) that hence reveal RBP binding sites [200]. Targets identified by HyperTRIBE were previously found to be consistent with those detected by CLIP [200]. Yet HyperTRIBE features reduced experimental cost and complexity, lower sequencing depth bias, and its results are not affected by crosslinking efficiency. In this study, we adapted HyperTRIBE to identify IGF2BP2 binding targets in mouse hepatocytes in their tissue environment. This was facilitated by hepatocyte-specific transfection with the experimental plasmid IGF2BP2-ADAR and a control plasmid mCherry-ADAR in vivo.

In the first computational step, we mapped A2G sites detected in the sequenced reads to the mouse reference genome and contrasted the results for IMP2 samples against those from mCherry and wild-type (WT) samples. Whereas mCherry samples acted as positive controls for transfection in hepatocytes, WT samples were negative controls for both transfection and IMP2 binding assay. Furthermore, HyperTRIBE uses ADAR with a E488Q hyperactive mutation that could in principle be biased to certain sequences and structures, although likely to a much smaller extent than the TRIBE protocol. Such a bias might manifest to a certain degree in mCherry samples and could be detected when comparing mCherry to WT [200, 266]. Hence, we inspected A2G sites present in IMP2 and not in either of two controls (WT vs. IMP2 and mCherry vs. IMP2) alongside with sites present in none of the controls (WT/mCherry vs. IMP2 - WT vs. mCherry). Genes with A2G sites in the latter group are referred to as "background-corrected IMP2-specific genes" as they do not contain background genes with A2G sites which distinguish only control samples (WT vs. mCherry). We discussed the selection of background genesets to account for biases from ADAR background activities in Supplementary Information section 5.2 "Background activity of ADAR".

One should note that (1) edited A2G sites should not be considered as IMP2 binding sites as ADAR does not compete with RBP binding spots; (2) For each edited site, HyperTRIBE quantifies the editing percentage based on read counts in IMP2 samples and background samples. The first observation suggests that defining IMP2 targets solely based on individual A2G sites without considering

proximal regions may not yield adequate and accurate conclusions. Our decision of considering entire CDS and UTR regions is reinforced by previous findings that RBP binding modes are highly diverse, with preference to certain sequences (i.e GC-rich sequences), structures (i.e loop regions), or genomic regions (i.e UTR) that open a vast space of possibilities for unconventional RBP-binding domains [124]. The latter observation implies that selecting significant A2G sites based on editing percentage is greatly impacted by how one combines results across replicates in IMP2 or background samples. In fact, the authors of HyperTRIBE recommended selecting only A2G sites shared across all replicates to ascertain a high level of confidence, or to use a high threshold (10%) to define the significant editing percentage [200]. Here, we compared the outcomes of different ways of selecting A2G sites across replicates and of defining IMP2 targets based on ADAR-edited sites or regions. This comparison was termed "replicate-collapsing schemes" in the Methods section.

We assessed these schemes by their power and stability in detecting IMP2 targets, and additionally by their concordance to the results from HyperTRIBE experiments on mouse embryonic fibroblasts and their association with deregulated genes. Those outcomes collectively suggested that CDS, 3'UTR and 5'UTR are appropriate genomic spans for grouping A2G sites into "edited regions" if they contain any edited sites in two out of three replicates with 1% editing percentage across all transcripts (1%-UNION-CDS/UTR). The selected scheme preserves reproducibility by requiring the edited sites to be found across different replicates while relaxing the constraint for exact site overlap, which enables it to generate sufficient results for further analyses (Supplementary Figure B.5, Supplementary Tables B.1 to B.3). Additionally, selecting CDS/UTR as target-defining regions showed good robustness, as these were the only regions that detected the same IMP2 targets (Supplementary Figure B.6). Annotating A2G sites to CDS/UTR regions also put the main focus on regions previously identified as IGF2BP2 or ADAR binding hot-spots [202, 124]. As shown in Supplementary Figure B.10, in all pairwise comparisons among IMP2, WT and mCherry samples, IMP2 showed strong binding preference for regions within 3'UTR (32.38-37.35%) and CDS (42.82-48.49%). This finding agrees with previously conducted studies about the binding preferences of IMP2[59, 276]. Interestingly, in the set of background corrected IMP2 target genes, we found considerably fewer IMP2 targets in 3'UTR (23.44%) and more in CDS (56.63%) (Figure 3B).

Differential gene expression analysis with DEseq2 showed that more genes were up-regulated in mCherry or IMP2 samples than in WT (Supplementary Figures B.3 and B.4). Raw expression levels and differentially expressed genes were more similar between mCherry and IMP2 groups than with those of the WT group (Supplementary Figures B.1, B.2, B.3 and B.4). As a result, it is not surprising that the deregulated genes in the WT vs. IMP2 and WT vs. mCherry comparisons share more similar and more significant biological processes with each other than with those in the mCherry vs. IMP2 comparison (Supplementary

Figure B.8E-G). This is, interestingly, not the case in the analysis of IMP2 targets with HyperTRIBE. The enriched biological functions for IMP2 targets share more overlap between WT vs. IMP2 and mCherry vs. IMP2 than with WT vs. mCherry, specifically for autophagy and regulation of catabolic processes (Supplementary Figure B.8A-C). Additionally, we found only a small subset of IMP2 targets that were also deregulated (Table III.1 and Supplementary Tables B.3). This discrepancy between DEseq2 and HyperTRIBE analysis implies that IMP2 can affect the transcriptome without drastically altering the expression of target genes and emphasizes the complex nature of RBP activity in cells. However, in both DESeq2 and HyperTRIBE analysis, background correction on the set of identified deregulated genes or of IMP2-specific genes revealed genes enriched with cellular catabolic processes and either autophagy (HyperTRIBE) or apoptosis (DESeq2) (Figure III.3A, bottom). Autophagy and apoptosis are two deeply connected processes that control cell death and survival through mediating the cellular and organismic homeostasis in response to stresses [75]. In particular, genes involved in regulation of autophagy were reported to overlap with apoptosis genes [28]. This connection suggests that IMP2-bound genes might be responsible for immune responses and apoptotic activation through autophagy.

Increased LC3-II levels upon IMP2 knockdown in the HCC model cell line HepG2 (Figure III.4A and B) support our hypothesis that IMP2 affects autophagy in a yet to be clarified manner. A potential link between IMP2 and autophagy could be the ability of IMP2 to induce steatosis in mice [242, 223, 8] which is strongly associated with dysfunctional autophagy [154]. Although the complex role of autophagy in lipid metabolism in steatosis models remains to be fully elucidated, the literature suggests a negative correlation between autophagy and steatosis [154]. In addition, it has been reported that modulated endogenous cholesterol synthesis by transgenic overexpression of IMP2 in the liver of mice activates NF-κB transcription factors and several of their targets [223]. NF-κB is bidirectionally linked to the regulation of autophagy in a variety of physiological and pathological contexts [189, 241, 281] and may also serve as an indirect mediator between IMP2 and autophagy. Furthermore, [141] reported decreased p62/SQSTM1 levels and increased LC3-II/LC3-I ratio in glioma cells upon IMP2 knockdown, similar to our findings in hepatocellular carcinoma cells. IMP2 was also shown to promote the progression of breast cancer by degrading the RNA transcript encoding a subunit of v-ATPase [81, 132], which is a major regulator of LC3 [94]. In contrast, [89] showed that IMP2 increases the stability of the lncRNA MALAT1, which in turn upregulates the autophagy-related gene ATG12, thereby activating autophagy in non-small cell lung cancer. Although the exact mechanism of action remains unclear, we suggest that IMP2 deregulation may also affect autophagy in the liver.

As previously reported, IMP2 is linked to mRNA stabilization in colorectal cancer [114, 81]. In our study, we also found IMP2 targets to be more stabilized by contrasting the changes in expression between IMP2-target and non-target genes (Supplementary Figures B.10 and B.12, Figure III.3C-D). Although DEseq2

showed that gene expression levels in mCherry and IMP2 were more similar than in WT (Supplementary Figures B.3 and B.4), the cumulative distributions of the LFC derived from DESeq2 analysis, separated by whether a gene is a IMP2-target or not, show more significant differences when comparing mCherry or WT against IMP2 (Supplementary Figure B.11 and Figure III.3C). Additionally, IMP2 targets with strong deregulation show anticorrelation to IMP2 (Supplementary Figure B.12). While these findings suggest IMP2's potential to destabilize IMP2 target genes, a portion of deregulated targets are observed with higher correlation to IMP2 instead (green and blue dashed lines in panel B versus solid purple line, Supplementary Figure B.12B). This effect, although unobservable when IMP2 targets were background-corrected (Figure III.3D), suggests that studies on specific target genes for IMP2 are needed for understanding the destabilizing and stabilizing effects of IMP2.

IMP2 has been reported as an important m6A-reader and prefers binding to m6A's consensus motif DRACH in Mus muculus and its variants (RRACH, URACH, RACH, RRAC, DRAY, RAC and DRACG, where D = A/G/U, R = A/G, H = U/A/C, and Y=U/C) on target mRNAs [11, 111, 174]. In our current study, targeted transcripts are more resistant to changes in expression levels, suggesting a connection to IMP2-directed stabilization through initiating m6a modification. We thus performed de novo motif analysis to discover IMP2's binding preference within +/-500bp around an A2G site and its potential link to m6A motifs. Remarkably, UGACG, a motif similar to DRACG, was found recurrently around edited sites in the de novo analysis of WT vs. IMP2 and in both differential motif analyses with top enrichment (for background-corrected WT vs. IMP2 and mCherry vs. IMP2 genes). Out of 7 common motifs between corrected results for WT vs. IMP2 and mCherry vs. IMP2, four motifs were found to contain or resemble RAC/RACH/DRACG variants. We suggest from the results that IMP2-targeted mRNAs might have been stabilized through IMP2's preferential binding to m6A modification sites.

In summary, our study demonstrates the successful in vivo adaptation of the HyperTRIBE protocol to identify IMP2 binding targets in mouse hepatocytes. By focusing on CDS and UTR regions, we were able to ensure reproducibility and robustness of target identification and align our results with known IMP2 binding sequence preferences. Furthermore, our results showed a direct interaction of IMP2 with transcripts involved in autophagy. So far, mainly target-stabilizing effects have been described for IMP2. Thus, further investigation of its destabilizing effects is required to clarify their functional implications.

5. Supplementary Information on Methods

5.1. Replicate collapsing scheme selection

Summary: We established and applied 16 different schemes to combine A2G sites identified in different pairs of replicates when comparing two samples. As mentioned in Methods section (Figure 1, main text), the schemes are combinations of two editing thresholds used for coverage-based filtering of A2G sites (1% and 5%), four gene regions in which A2G sites should be combined (editing sites, 10bp windows, CDS/UTR, full transcript) and two ways of combining A2G sites (UNION: requires A2G sites to occur in at least two out of three replicates; INTERSECT: requires A2G sites to occur in all three replicates). The examined schemes resulted in greatly varying number of identified IMP2 targets (Supplementary Table B.1); Thus, we compared the resulting genes to with editing sites detected by HyperTRIBE in mouse embryonic fibroblasts, and to genes detected by DESeq2 to be differentially expressed (Supplementary Information section 5.1.1 and 5.1.2, respectively). The largest overlap and highest robustness were obtained when using the 1% average editing threshold and requiring detection of A2G sites in 3'UTR, 5'UTR or CDS in at least two out of three replicates (Supplementary Tables B.2 and B.3). Hence, we used this replicate collapsing scheme for the downstream gene set enrichment analysis and when discussing the biological relevance of our results.

5.1.1. Overlap between HyperTRIBE results in mouse liver and in mouse embryonic fibroblast

We measured the overlap of gene sets reported by HyperTRIBE in murine liver and in MEF samples either as count of mutually found genes, the percentage of mutual genes among the results, or by the Jaccard index between the two gene sets. These results are shown in Supplementary Table B.2. Similar to the number of genes containing A2G sites, the overlap between mouse liver and MEF samples was larger when we compared larger genome spans, lower average editing percentage or fewer required replicates. The overlap was quite small when using a 5% threshold for the average editing percentage (Supplementary Table B.2). Thus, we selected the results from the 1% threshold to ensure better interpretability and robustness. On the other hand, filtering genes with A2G sites belonging to the same 10 bp window or CDS/UTRs yielded the highest percentage of mutual edited genes between mouse liver and MEF samples for the comparisons between any control and IMP2 (61.27% and 60.59% for 10 bp window and CDS/UTRs, respectively, for IMP2 vs. WT comparison and 60.42% and 59.98% for IMP2 vs. mCherry comparison) (Supplementary Table B.2). Interestingly, although the number of mutual genes almost doubled when we expanded the editing regions from CDS/UTRs to full transcripts, the percentage of mutual genes in fact decreased, indicating a lower sensitivity. se

5.1.2. Overlap between HyperTRIBE results and DESeq2 differentially expressed genes

Next, we compared the editing sites in mouse liver detected by HyperTRIBE genes against the DEGs reported to be differentially expressed by DESeq2. Supplementary Table B.3 lists Jaccard indices measured for the four editing sites/regions, where A2G sites were identified based on at least 1% average editing coverage and in at least two replicates, as reasoned above. As expected, the Jaccard similarity increased when considering larger editing regions (Supplementary Table B.3, columns A-D), but not as strongly as the increase in the numbers of detected genes having A2G editing sites (Supplementary Table B.1). Since we are interested in genes which distinguish IMP2 samples from WT or from mCherry but behave similarly for the two controls, we opted for the genomic span that could filter "control genes" most effectively. To this aim, we first gathered the genes from WT vs. IMP2 and mCherry vs. IMP2 comparisons (Supplementary Table B.3, column D) and subsequently removed these genes from the comparison between control samples (Supplementary Table B.3, column E). The fraction of genes that do not occur in the merged WT vs. mCherry result is reported in Supplementary Table B.3, column F. Although the number of A2G genes with altered expression increased 2-fold when using transcripts instead of CDS/UTR regions (26 and 13, respectively), the fraction of IMP2-related genes did not further decrease (both 0.19, Supplementary Table B.3, column F). Additionally, all IMP2-related genes from the CDS/UTR-based scheme were also identified in all other schemes, whereas 13 out of 26 genes from the transcript-based scheme were not shared by any other scheme (Supplementary Figure B.6). In general, aggregating all A2G sites in one CDS/UTR seems preferable as this resulted in a relatively large overlap between the HyperTRIBE and DEG gene sets without losing sensitivity and robustness in comparison to replicate collapsing-schemes using 10bp window or transcript. The genes that are both significantly deregulated according to DESeq2 and contain HyperTRIBE-identified A2G sites are listed in Table III.1.

5.2. Background activity of ADAR

Summary: We used two control samples in the experiment, including mCherry samples as positive control for transfection in hepatocytes and WT samples as negative controls for both transfection and IMP2 binding assay. Comparing IMP2 samples to either of these controls reveals IMP2 binding sites and regions through detecting A2G editing sites. As HyperTRIBE uses ADAR with a E488Q hyperactive mutation that reduces editing biases in certain sequences and structures as compared to TRIBE method, a certain degree of biases may still exist and can be detected when comparing any sample to WT samples (26, 39). Thus, we investigated the extent by which the three samples WT, mCherry and IMP2 differ from each other as means to measure ADAR background activities that might lead to editing biases and relevant changes in the transcriptomes. Surprisingly, the

pairwise comparisons of HyperTRIBE results showed high consistency in detected A2G sites between WT and mCherry against IMP2 (Supplementary Information section 5.2.2), mCherry shared more similarity in transcriptomic profile with IMP2 than with WT (Supplementary Information section 5.2.1). Thus, to remove ADAR's background activity that might result in incorrectly identified editing sites, we performed a background correction for the set of IMP2 target genes identified in any comparison that IMP2 samples were involved in. We defined "background genes" as genes with A2G sites from WT vs. mCherry comparison and subtracted these from genes with A2G sites from WT vs. IMP2 and mCherry vs. IMP2 comparisons. For the sets of genes with differential expression, the same background-correction procedure was applied.

5.2.1. HyperTRIBE's IMP2 target genes - Supplementary Figures B.8 and B.9, Supplementary Table B.1, B.2

To visualize the count distribution of editing sites per gene and to analyze whether the choice of control would affect the identification of editing sites, Supplementary Figure B.9 plots counts of editing sites identified in different comparisons. Supplementary Figure B.9C shows that both WT and mCherry identified highly consistent A2G numbers against IMP2 and can hence both serve as controls for identifying A2G sites in IMP2 samples. The genes identified in WT vs. mCherry generally had fewer A2G sites than in the comparisons WT vs. mCherry or in WT vs. IMP2 (Supplementary Figure B.9A-B). Supplementary Figure B.8 lists the biological processes that were enriched in IMP2-bound genes and in deregulated genes, respectively. IMP2-bound mRNAs were enriched in catabolic processes, autophagy and cellular organization (Supplementary Figure B.8A-C). In the last comparison (Supplementary Figure B.8D), we removed control genes which belong to the WT vs. mCherry comparison. Then, the majority of catabolism-related terms were pruned, whereas autophagy terms were retained.

5.2.2. DESeq2's differentially expressed genes - Supplementary Figure B.1 to B.5, B.8, Supplementary Table B.4

First, the homogeneity of each sample group, i.e., WT, mCherry, and IMP2, was assessed by principal component analysis (PCA). This showed that 97.38% of the total variance in the replicates is captured by the three first principal components (Supplementary Figure B.1A). The three WT samples revealed the largest variation and were placed separate from the other samples. Along the fourth principal component, mCherry and IMP2 samples then show a split as well (Supplementary Figure B.1B). The results of PCA analysis were consistent with correlation analysis of the raw read counts with Salmon using either Pearson or Spearman correlation (Supplementary Figure B.2). Using DESeq2, we normalized and log-transformed the gene expression data in a multifactorial design, whereby one sample group was compared to one of the two other sample groups. Heatmaps of the differentially

expressed genes then showed high agreement between the replicates of a sample group (Supplementary Figure B.3). MA plots generated by DESeq2 revealed a symmetric distribution of transformed counts (Supplementary Figure B.4). The comparisons between WT vs. IMP2, mCherry vs. IMP2, and WT vs. mCherry, yielded 2252, 860, and 2068 transcripts belonging to 920, 264, and 832 differentially expressed genes identified by DESeq2, respectively ($|\text{LFC}| \leq 1$, FDR-adjusted p-value ≤ 0.05). The Jaccard index of the sets of DEGs between WT and IMP2 and between WT and mCherry shows a good overlap of 0.49, Supplementary Table B.4, which is also reflected in the heatmaps for those comparisons in Supplementary Figures B.3A and C. On the other hand, the DEGs from the pair mCherry vs. IMP2 (Supplementary Figure B.3B) share little similarity to those from the other two comparisons (Jaccard indices of 0.13 and 0.14 for mCherry vs. IMP2 and WT vs. IMP2 in Supplementary Table B.4, respectively). While there are large overlaps in enriched gene ontology terms for IMP2 target genes from WT/mCherry to IMP2 comparisons, enriched terms for DEGs from mCherry/IMP2 to WT comparisons are more similar. DEGs from comparing mCherry/IMP2 to WT were enriched in cellular defense mechanisms (Supplementary Figure B.8E and G). Notably, the enriched biological processes from mCherry vs. IMP2 DEGs, most of them being related to DNA and RNA processing (Supplementary Figure B.8F), differ considerably from the DEGs derived from any comparison against the WT group (Supplementary Figure B.8E and G).

6. Data Availability

The codes for data retrieval, data processing and all analysis are made available on Zenodo https://doi.org/10.5281/zenodo.14627021. Mouse embryonic fibroblast (MEF) sequencing data are available on GEO with the accession number GSE260682.

7. Acknowledgements

We thank Dr. Gilles Gasparoni for valuable discussions and support in sample preparation.

8. Author Contributions

A.K., and S.M.K. designed the research project. T.K. prepared RNA from mouse, M.P., S.F., and M.S. performed the TRIBE sequencing experiments of mouse in vivo data, S.B. conducted autophagy assays. H.T.T.D. conducted the bioinformatics analysis of mouse in vivo data, J.B. performed the experiments and conducted the bioinformatics analysis of MEF data, M.P., M.S., J.B., V.H., S.M.K. and A.K.

contributed to the data analysis. H.T.T.D. wrote the draft of the manuscript, S.B., V.H., S.M.K., M.S., J.B. and A.K.K. edited the manuscript. All authors reviewed and approved the final version of the manuscript.

9. Competing Interests

None declared.

10. Funding

This work was supported by the DFG Research Infrastructure NGS_CC (project 407495230) as part of the Next Generation Sequencing Competence Network (project 423957469). NGS analyses were carried out at the Competence Centre for Genomic Analysis (Kiel). This study was funded, in part, by the Deutsche Forschungsgemeinschaft (DFG, KI702 to A.K.K. and KE2519 to S.M.K.).

Chapter IV.

Project 4: Review article: Detecting Rewiring Events in Protein-Protein Interaction Networks Based on Transcriptomic Data

This review has been published as "Hollander, M., Do, T., Will, T., Helms, V. (2021). **Detecting rewiring events in protein-protein interaction networks based on transcriptomic data**. *Frontiers in Bioinformatics*, 1, 724297". In this work, Markus Hollander came up with the review strategy and general structure for the manuscript, while I performed the literature research jointly with him and conducted the case study on existing softwares and webservers for protein interaction network analysis. The parts which were reviewed mainly by Markus Hollander, such as certain domain or protein interaction databases like STRING, IID or DomainGraph, are not included in this thesis. To maintain clarity and cohesiveness of the case study where we compared different tools to study the interaction networks of human TNR6 protein, the parts discussing PPIXpress and PPICompare written mainly by Markus Hollander are also included here.

1. Introduction

Protein-protein interaction (PPI) networks are a popular cornerstone of integrative or computational cell biology and are frequently used to interpret the findings from high-throughput studies [123]. Typically, PPI networks provide a genome-scale picture of all physical interactions detected between pairs of proteins. In the past, such networks have been compiled by integrating the results from many small-scale experiments and from several high-throughput experimental methods such as Yeast Two-Hybrid or Tandem Affinity Purification coupled to mass spectrometry (TAP-MS) [14]. Full PPI networks provide a comprehensive picture of the interactome of the full proteome of an organism. However, in each cell at a particular moment in time, any physical protein-protein-contact can only be realized if both proteins are

expressed at the same time. To address this, it has become common practice to trim general PPI networks to the set of proteins encoded by the genes that are expressed in the same condition. In this manner, researchers have compared the protein interaction landscape across tissues [148] as well as the origin of tissue-specific diseases [16].

PPI networks have an interesting scale-free topology, whereby highly connected "hub" proteins occur at a higher frequency than expected in, for example, a random graph. Furthermore, there exist densely connected communities [85] of proteins participating in particular cellular functions or certain biological pathways. This ordering according to cellular function gives rise to a modular architecture of PPI networks. On a smaller scale, densely connected clusters are candidates for protein complexes and several algorithms exist to identify such complexes in PPI networks [173]. Interacting partners and members of the same protein complex tend to be co-expressed [106]. The stable association of two proteins often involves one or more distinct structural contacts between specific domains of the proteins [7]. Knowledge about protein domain annotations and domain-domain interactions (DDIs) thus provides a good basis to describe protein associations [7]. DDIs were used, for example, to predict protein complexes and to analyze protein-protein interaction networks [182, 156].

About 95% of all human multi-exon genes are subject to alternative splicing (AS) [185] which clearly affects the ability of the encoded isoforms of the proteins to interact with other proteins [34, 68]. Hence, it appears worthwhile to exploit the base-resolution of modern RNA sequencing technology to resolve context-specific PPI networks at isoform-resolution. In 2015, the Vidal group published the first large-scale experimental study on isoform-specific protein interactions [269]. They profiled the interactomes of 366 protein isoforms encoded by 161 genes and assayed them against a library of 13,000 genes. They found that accounting for isoforms gave a remarkable 3.2-fold increase of the number of PPIs. Strikingly, different isoforms of the same protein can interact with completely different proteins.

In the next section, we first give an overview of the numerous protein-level PPI databases that underpin the research effort in this field. These databases were recently reviewed in a comprehensive manner [14] and we will thus focus on a few popular meta webservices that offer integrated analyses and the ability to tailor full PPI networks to a particular cellular context. Afterwards, we present those tools and webservices in detail that support isoform-level analysis of protein interactions. Jalili and co-workers previously reviewed studies that integrated gene expression data with protein interaction networks [104]. Yet, their review focused on discovery of biomarkers and did not discuss the existing software tools, nor underlying domain models or protein isoform effects. Next, we review webservices and software tools that conduct differential comparisons of interactions between cellular contexts and thus facilitate the detection and study of PPI rewiring events. Finally, we will illustrate the usage and capabilities of some of them on the example of the interactions formed by the human TNR6 protein encoded by the FAS gene.

2. Web services providing protein-level data on protein-protein interaction networks

Manual and automated analyses of PPI networks require reliable and preferably large collections of PPIs. Consequently, many databases have been established over the years that collect, curate, and annotate protein interactions and make them available to the research community. These resources differ in their sources, curation and annotation policies, as well as their focus on, for example, particular species or interaction types, and the features of their interfaces.

Many of the available resources represent PPI networks at the level of integral genes, so that alternative splicing and protein isoforms are not considered. Already, the protein level enables powerful analyses of PPI rewiring between different conditions. Specifically, the human genome contains around 20,000 protein-coding genes [190, 193], while the human body consists of more than 200 different cell types [26]. Each one of them will only express a cell-type specific subset of the full proteome, e.g. about 8,000-12,000 proteins [65]. Hence, Bossi and Lehner argued that if two genes are co-expressed in a cell in a particular condition, their products may physically interact in that cell [29]. However, if the two proteins are not simultaneously expressed in a tissue, then the interaction obviously cannot occur in that tissue. An examination of the relationship between tissue-specificity and connectivity found that proteins with more pronounced tissue-specificity are involved in fewer protein interactions than more universally expressed proteins [29]. Furthermore, tissue-specific proteins are more likely to be recent evolutionary innovations than universally expressed proteins [136]. Turned around, the more conserved a protein is, and the larger the number of tissues where it is expressed, the more protein interactions it is likely to have [271]. Filtering global PPI networks to the subset of expressed genes or proteins thus became the workhorse for generating tissue- and other context-specific PPI networks.

Table 1 presents an overview of major primary and meta PPI databases and their features. Most of the resources discussed here provide a web-interface that enables users to query and download the underlying interaction data. In many cases this includes integrated visualization of the queried interactions. The subsequent sections discuss a few meta databases in more detail whose webservices additionally offer PPI (sub-)network analyses or processing of user provided data in a context-specific manner.

2.1. Human Integrated Protein-Protein Interaction Reference (HIPPIE)

The Human Integrated Protein-Protein Interaction Reference (HIPPIE) [5] is a meta database and webservice that facilitates access to and context-specific analysis of experimentally detected human PPIs. These interactions are consolidated from BIND [13], BioGRID [181], DIP [214], HPRD [115], IntAct [180], MINT [142], and

Table IV.1.: Overview of PPI databases and PPI network (PPIN) features of their webservice.

	Data Collection	**Source Type**	**Species**	**Webservice**	**Context Filter**	**Visual-ization**	**PPIN Analyses**
APID	meta	evidence	multiple	apid.dep.usal.es	no	yes	no
BIND	primary	evidence	multiple	-	-	-	-
BioGRID	primary	evidence	multiple	thebiogrid.org	no	yes	no
DIP	primary	evidence	multiple	dip.doe-mbi.ucla.edu	no	yes	no
HIPPIE	meta	evidence	human	cbdm.uni-mainz.de/hippie	tissues, diseases, functional	yes	enrichment
HPIDB	both	evidence, predicted	multiple	hpidb.igbb.msstate.edu	no	yes	no
HPRD	primary	evidence	human	hprd.org	no	no	no
HuRI	primary	evidence	human	interactome-atlas.org	tissued	yes	no
I2D	meta	evidence, predicted	multiple	ophid.utoronto.ca	no	no	no
IID	meta	evidence, predicted	multiple	iid.ophid.utoronto.ca	tissues, druggability, localization, diseases	yes	enrichment, topology
InnateDB	both	evidence, predicted	multiple	innatedb.com	tissues, cell types, diseases	yes	topology
IntAct	both	evidence	multiple	ebi.ac.uk/intact	no	yes	no
iRefWeb	meta	evidence, predicted	multiple	wodaklab.org/iRefWeb	no	no	no
MatrixDB	both	evidence, predicted	multiple	matrixdb.univ-lyon1.fr	tissues	yes	no
mentha	meta	evidence	multiple	mentha.uniroma2.it	no	no	paths
MINT	primary	evidence	multiple	mint.bio.uniroma2.it	no	yes	no
MPact	primary	evidence	yeast	-	-	-	-
MPIDB	both	evidence	microbes	-	-	-	-
MPPI	primary	evidence	mammals	mips.gsf.de/proj/ppi	no	no	no
MyProteinNet	meta	evidence	multiple	netbio.bgu.ac.il/myproteinnet2	tissue, expression, single cell	yes	no
PIP	meta	evidence, predicted	human	compbio.dundee.ac.uk/www-pips	no	no	no
PrePPI	meta	evidence, predicted		bhapp.c2b2.columbia.edu/PrePPI	no	yes	no
SPECTRA	meta	evidence	human	alpha.dmi.unict.it/spectra	tissue, tumors, expression	yes	network alignments
STRING	meta	evidence, predicted	multiple	string-db.org	no	yes	enrichment, clustering
TissueNet	meta	evidence	human	netbio.bgu.ac.il/tissuenet	tissue, expression	yes	no

MPPI [183]. HIPPIE assigns confidence scores to the PPIs based on the quality and reliability of the experiments supporting them. The tissue specificity of the PPIs is derived from tissue RNAseq data from the GTEx Consortium [147], and they are further annotated with functional information from the Gene Ontology [39].

The web-interface enables users to query the database with individual proteins or PPI networks in a tissue- and function- specific manner, with the option to define a custom context and to specify the desired interaction types and level of confidence. HIPPIE subsequently generates and visualizes a query-specific PPI network (Figure IV.1A) and can optionally perform edge direction inference, prediction of inhibitory or activating effects, and enrichment analysis of disease, process, function and cellular compartment annotations. In addition to the web-interface, HIPPIE offers an application programming interface (API) that facilitates automated queries and thus integration into analysis pipelines.

The integration of experimentally confirmed PPIs from expert-curated sources makes HIPPIE a reliable and resourceful reference database to be employed in various scenarios. Many studies availed the tool to collect high confidence PPIs to support their study hypothesis, verify experimental results or control the quality of analytic methods. Some of such use-cases include the work by Sundell et al., who assessed the performance of phosphomimetic proteomic peptide-phage display in detecting ligands of short linear motifs by comparing the identified ligands to those reported by HIPPIE [233], and the work by Kruse et al., who screened for candidate protein constituents of N-cadherin complexes based on HIPPIE PPI confidence scores [128]. Furthermore, the tool is widely employed to analyze pathogenesis or developmental processes where tissue- specificity has a substantial weight on defining the PPI networks. This is well illustrated by a comparative study by Verma et al. where the subtle differences in LRRK2 interactomes across different brain subregions, kidney and lung could be spotted using HIPPIE's tissue filter [247]. On the other hand, this webservice is often used as a meta-database and integrated in other more context-specific analysis methods such as GSA-SNP2 [273], PSSMsearch [129], or LncDEEP [268].

3. Resources and software tools on domain-level and isoform-level protein-protein interaction data

A few approaches that model context-specific PPI networks utilize an underlying domain model in which each protein is represented by one or more structural domains. Using such a domain model incorporates elements of three-dimensional protein structures. As domain-domain interactions (DDIs) tend to be evolutionary conserved [101], experimental evidence on domain interactions can be transferred between related organisms. Isoform-specific expression data enables detecting effects of alternative splicing at the level of full protein domains. Hence, DDI-based tools allow predicting the interactomes of specific protein isoforms in a particular

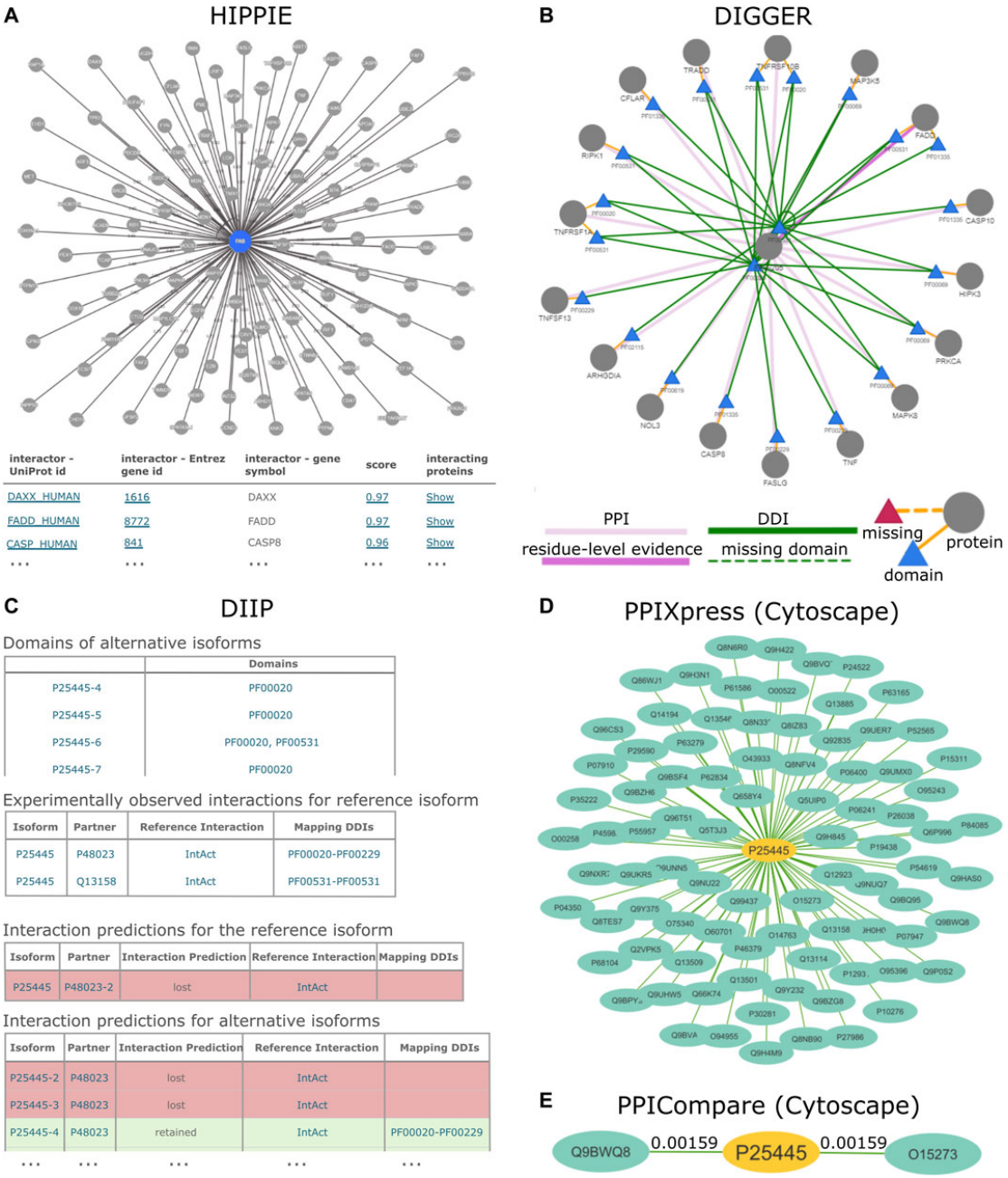

Figure IV.1.: PPI results for the human TNR6 protein encoded by the FAS gene (UniProt accession: P25445) from HIPPIE, DIIP, DIGGER, PPIXpress and PPICompare. (A) HIPPIE presents general interacting partners as a PPI network view and a table with evidence scores. (B) DIGGER presents an interactive PPI-DDI view that combines different aspects of a PPI for a single isoform. The domain and exon architecture view and more detailed results for individual interactions, domains and isoforms are not shown here. (C) DIIP compiles tables with the domains and predicted interaction retention for alternative isoforms. (D) PPIXpress was applied to RPKM expression data of neuronal stem cells and H1 stem cells from the Roadmap Epigenomics project [210]. Shown here are the interaction partners of TNR6 in the resulting neuronal stem cell specific PPI network. (E) PPICompare was subsequently applied to cell-specific PPIs generated by PPIXpress to produce the differential PPI network between the two cell types. Illustrated here are two interaction gains in the transition from H1 stem cells to neuronal stem cells with the respective adjusted p-value. The graphics for the subnetworks from PPIXpress and PPICompare were generated with Cytoscape.

condition or tissue. Table IV.2 presents a feature overview of such webservices and software tools.

Table IV.2: Features of software tools and webservices enabling the generation and analysis of domain- or isoform-level protein-protein interaction data (DIIP, DIGGER, PPIXpress), as well as the comparison of interaction rewiring (SPECTRA, DifferentialNet, PPICompare).

	DIIP	**DIGGER**	**PPXpress**	**SPECTRA**	**DifferentialNet**	**PPICompare**
Type	Webservice	Webservice	Stand-alone tool	Webservice	Webservice	Stand-alone tool
Species	Human	Human	Multiple	Human	Human	Multiple
Resources						
PPI	HI-II-14, IntAct	BioGRID	IntAct, mentha	BioGRID, HPRD, MIPS, IntAct	BioGRID, DIP, IntAct, MINT	PPXpress
DDI	3did, DOMINE	3did, DOMINE	DOMINE, IDDI, 3did, iPfam	Not available	Not available	PPXpress
Other	Pfam (domain annotations)	PDB (exon-specific residues)	UniProt + Ensembl (protein details), Pfam-A	Protein Atlas, ArrayExpress, GEO, TCGA (tissues, tumors)	GTEx, Human Protein Atlas (tissues)	Ensembl (protein details)
Query type						

Continued on next page

Table IV.2: Features of software tools and webservices enabling the generation and analysis of domain- or isoform-level protein-protein interaction data (DIIP, DIGGER, PPIXpress), as well as the comparison of interaction rewiring (SPECTRA, DifferentialNet, PPICompare). (Continued)

Protein query	Single gene or protein (UniProt, HGNC)	Single gene, transcript, protein (Ensembl, HGNC); Single exon (Ensembl, HGNC, coordinates)	Not available	All genes in SPECTRA	Up to 5 genes or proteins (Entrez, Ensembl)	Not available
Network query	Single protein and list of interacting partners	List of isoforms, transcripts, genes; Transcript expression counts	Gene or transcript expression data (UniProt, HGNC, Ensembl IDs); Reference PPI network (optional)	List of genes (Entrez, Ensembl, UniProt); Gene expression SPECTRA	Not available	Two sets of condition-specific PPI and DDI networks (PPXpress)
Integrating output from other tools or databases	Not available	Expression data: Cufflinks, Kallisto, count matrix	PPI networks: mentha, IntAct, BioGRID, HPRD, custom PPI list; Expression data: Cufflinks, Kallisto, (TCGA) RSEM, GENCODE GTF, count matrix	Expression matrix	Not available	PPI and DDI networks: PPXpress

Continued on next page

Table IV.2: Features of software tools and webservices enabling the generation and analysis of domain- or isoform-level protein-protein interaction data (DIIP, DIGGER, PPIXpress), as well as the comparison of interaction rewiring (SPECTRA, DifferentialNet, PPICompare). (Continued)

Output	Isoform domains; PPI prediction for isoforms	Isoform and exon domains; PPIs and DDIs for isoforms and exons	Context-specific PPI and DDI network(s)	Context-specific PPI network; Differential PPI subnetworks	Tissue-specific differential PPIs	Context-specific differential PPIN network; Minimal set of rewiring causes
Features						
Visualization	Not available	Integrated (protein, domain, and interaction view)	Network output available for Cytoscape	Integrated	Integrated	Network output available for Cytoscape
Context-specific	Not available	User-defined	User-defined	Tissues, tumors, user-defined	Tissues	User-defined

Continued on next page

Table IV.2: Features of software tools and webservices enabling the generation and analysis of domain- or isoform-level protein-protein interaction data (DIIP, DIGGER, PPIXpress), as well as the comparison of interaction rewiring (SPECTRA, DifferentialNet, PPICompare). (Continued)

Interaction scoring	DDI-based loss/retention prediction for isoforms	DDI-based missing interactions for isoforms	DDI-based presence for main isoform	PPIs: coverage, average weight; Differential subnetworks: expression log fold change	Difference between tissue-specific and median expression score	Significance of PPI rewiring event
Batch analysis	Not available	Not available	Yes	Not available	Not available	Yes

3.1. Domain-Domain Interaction Databases

Domain- and isoform-level methods tend to use Pfam domains [169], which are typically identified in experimentally determined three-dimensional structures of protein complexes by assessing whether neighboring domains are in close contact. This information is compiled by web services such as iPfam [77] and 3did [172]. The database 3did is regularly updated and currently contains 14,972 DDIs for 9,580 domain families in its current release (Pfam version 32.0, PDB version 2020_05). Furthermore, the structurally derived domain-level interactome can be enriched by computational predictions of DDIs between domain families [208]. The meta database DOMINE [272] integrated two databases of PDB-derived DDIs and 7 predicted data sources. Moreover, IDDI [119] combined data from three structure-based DDI sources and 20 computational datasets.

3.2. Domain-Based Isoform Interactome Prediction (DIIP)

Domain-based Isoform Interactome Prediction (DIIP) [84] is a method that uses reference PPIs and DDIs topredict isoform interactions, which can be queried with theaccompanying webservice (https://predict-isoform-interactome.herokuapp.com). Given a human query protein, the DIIPwebservice lists the domains in alternative isoforms,experimentally observed PPIs involving the reference isoform,as well as interaction predictions for alternative isoforms(Figure IV.1B).

To construct the underlying human isoform interactome,DIIP first builds a PPI network with experimentallydetermined PPIs from the HI-II-14 dataset [212] and the IntAct database [180]. The network proteins are annotated with Pfam domains [169], which then guide the mapping of DDIs from3did [172] and DOMINE [272] onto the PPIs. Finally, DIIP predicts isoforminteractions based on the presence of interacting domains inalternative isoforms retrieved from UniProt [244]. A PPI between two isoforms is considered lost if none of the supporting DDIs can be realized due to the requireddomains missing from the isoforms. Otherwise, the PPI ispredicted to be retained.

In the accompanying study, the authors showed thatalternative splicing is responsible for extensive network remodeling of protein interactions [84]. For about 22% of the genes with two or more isoforms in the predicted isoform interactome, at least one isoform lost an interaction. Furthermore, different interaction profiles were found for roughly 18% of the isoform pairs encoded by the same gene in the isoform interactome.

3.3. PPIXpress

PPIXpress [259] is a stand-alone tool developed in the Helms group that constructs condition- or sample-specific protein-protein interaction networks from transcriptomic data. It considers underlying domain-domain interactions and can be applied on the gene-level as well as the transcript-level, thereby capturing the effects of

alternative splicing in addition to those of differential expression. As outlined in Figure 2, the approach consists of a mapping step that relates protein-protein interactions to domain-domain interactions and a contextualization step that removes domain-domain interactions not supported by the given expression data, yielding a condition-specific PPI network (Figure 1D).

3.3.1. Condition-Specific DDI and PPI Network Construction

Besides one or more gene- or transcript-level expression data samples provided by the user, PPIXpress requires a reference PPI network with condition-unspecific interactions from the corresponding species. This reference network can be a custom one supplied by the user or can be automatically retrieved from the current versions of the databases mentha [38] or IntAct [180]. The STRING database [235] can be queried to add functional association scores. The most recent Ensembl [95] and UniProt [244] databases are queried for gene and transcript annotations, which are subsequently associated with Pfam-A domains [169] by InterProScan [107]. For physical domain-domain interaction data, a pre-compiled database of high-confidence data from DOMINE [272] and IDDI [119] is used, which is supplemented by automatically retrieved data from the iPfam [78] and 3did [172] databases.

The initial mapping stage (Figure 2, step 1) starts from the complete, condition-unspecific PPI network. Each interaction therein is then annotated with known interactions between domains in the longest isoform of the participating proteins. To ensure complete correspondence of the resulting domain-domain interaction network and the PPI network, artificial domains are added to interacting proteins if their interaction cannot be assigned to at least one domain-interaction. These artificial domains counteract the sparsity of domain-level data, where protein-protein interaction may not be accounted for by a known domain-domain interaction. In a similar previous approach, this improved the performance of protein complex prediction without negatively affecting precision [156].

In the subsequent contextualization stage (Figure 2, step 2), node pruning (Figure 2, step 3) is conducted by removing proteins from the network that are not supported by sufficient gene or transcript expression in the sampled condition, similar to established gene-based approaches. When PPIXpress is used on the gene-level, each protein is represented by its longest isoform just like in the initial mapping. Due to the direct correspondence between the two layers, pruning then stops here. When applied on the transcript-level, however, there is an additional edge pruning step (Figure 2, step 4), in which the protein-protein interactions that are not backed by at least one domain-domain interaction in the most abundant transcript are trimmed as well. In both cases, the result is a pruned, condition-specific PPI network annotated with the participating domain-domain interactions.

3.3.2. Applications: PPI Rewiring in Cancer

The accompanying case study [259] applied PPIXpress to 112 matched breast cancer and healthy tissue samples obtained from The Cancer Genome Atlas [256] and constructed a single differential PPI network of significantly rewired interactions between the two conditions from the combined individual condition-specific networks. The rewired interactions were associated with hallmarks of cancer if at least one of the proteins was annotated with a corresponding Gene Ontology (GO) [39] term or KEGG [109] pathway, and in addition GO term enrichment analysis was performed. To compare the performance, this approach was conducted with PPIXpress set to the gene-level and then to transcript-level filtering.

Although the gene-based approach generated larger networks, it detected fewer significant changes in interactions between the tumor and healthy samples. In contrast, the transcript-based network construction found significantly more rewiring events associated with the hallmarks of cancer in the differential network. Additionally, enriched KEGG and GO terms were related to carcinogenic processes, and the transcript-level differential network contained more enriched KEGG pathways and GO biological processes. Overall, the inclusion of domain-level and transcript-level information improved the performance and statistical significance of the results.

Frishman and colleagues used PPIXpress to generate 642 patient-specific pairs of interactomes corresponding to both the tumor and healthy tissues across 13 cancer types based on RNA-Seq datasets from The Cancer Genome Atlas [110]. The underlying hypothesis for this study was that isoform switching has been noted as a hallmark of cancer. Isoform switching often results in the loss or gain of domains mediating protein interactions and thus, the re-wiring of the interactome. Comparison of these interactomes between tumor and normal samples gave a list of patient-specific "edgetic" perturbations of the interactomes associated with the cancerous state. Interestingly, the majority of the rewiring events did not directly affect significantly mutated genes but were nonetheless strongly correlated with patient survival. The findings of this study are made available as EdgeExplorer: http://webclu.bio.wzw.tum.de/EdgeExplorer. The involved proteins are suggested both as a new source of potential biomarkers for classifying cancer types and as putative anti-cancer therapy targets.

3.4. Domain Interaction Graph Guided Explorer (DIGGER)

The Domain Interaction Graph Guided Explorer (DIGGER) [149] is a webservice (https://exbio.wzw.tum.de/digger) that leverages domain- and residue-level information to expedite studying the mechanistic effects of alternative splicing in humans. For domain-level analysis, DIGGER constructs a joint interaction network of human PPIs retrieved from BioGRID [181] and accompanying DDIs selected from 3did [172] and DOMINE [272]. With this data structure, the mapping of PPIs and DDIs to exon- and transcript-defined regions is facilitated for higher-leveled analysis of the interactomes. First, residues from experimentally resolved protein

structures from the Protein Data Bank (PDB) [35] are aligned with the amino acid residues from a protein in the joint network. Then, the interactions between the mapped resolved residues from different isoforms are used as evidence for the interactions between the domains in the isoforms. This allows DIGGER to locate the exons or transcripts associated with these domains using the mapping positions. Since a single amino acid residue or domain might be involved in the interaction between a protein and multiple partners, the authors defined an interaction scoring scheme based on the fraction of annotated DDIs present between two proteins. Finally, based on the score gradient one can assess to which degree a PPI is affected and draw inference on underlying mechanisms that impact those interactions, such as exon skipping. Notably, DIGGER combines all structural information from different isoforms of the same gene, whereas other tools such as PPIXpress [259] consider only a particular transcript, typically the most strongly expressed one.

The webservice offers three different use modes that can be used interchangeably. In the isoform-level analysis, the user can comprehensively visualize the interacting domains of proteins and compare the interactions of different isoforms (Figure IV.1C). In addition to displaying the associated PPIs, this mode visualizes the interactions annotated with a particular domain using the underlying protein and domain interaction data. Furthermore, each domain in the isoform can be selected to show a domain- centered interaction view. Lastly, this mode offers an overview of the domain and exon architecture of the selected isoform. The exon-level mode is similar and focuses on a particular exon. In the network-level analysis, the user can explore interactions between multiple isoforms and generate a specific subnetwork from a list of protein variants or transcripts. Here, the user can optionally upload a particular expression data set with transcript counts, thus accounting for user-defined contexts.

In their paper, Louadi et al. demonstrated how DIGGER can be used to confirm the effects of alternative splicing on PPIs and DDIs [149]. A case study where the PPI networks for two ALK transcripts were visualized using DIGGER revealed that 97% of the PPIs were lost in the truncated transcript. In another example, the self-interacting property of the GRB2 protein was removed by the loss of domain SH2 during the exclusion of a tissue-specific exon, while its interaction to gene RAPGEF1 was retained thanks to the unscathed domain SH3. This result could be confirmed by DIGGER using the exon-view of DDI networks for GRB2 and its interacting partners. Additionally, it is possible that the rewiring events for PPIs and DDIs resulting from exon skipping will emerge from the inspection of DIGGER-generated interaction networks for different transcripts. While DIGGER currently does not offer a downstream analysis after network construction, the authors expect to expand this webservice with pathway annotation of PPIs and DDIs and a focus on investigating the biological impacts of exon skipping events.

4. Detecting protein-protein interaction rewiring events

The tools presented in the previous section enable retrieving PPIs for specific organisms, tissues, and other conditions. Yet, one is often interested in detecting changes between two conditions. For example, Basha and co-workers recently analyzed differential protein interactomes for 51 tissues from the GTEx consortium [18]. In their study, they used single expression data sets and focused on establishing relationships between genes and hereditary disorders and Gene Ontology [39] terms. The main idea behind this was to identify gene candidates to explain tissue-selectivity of these hereditary disorders.

As described in the previous sections, the aforementioned protein- and domain-level tools can be used as basis to study PPI rewiring events. However, doing so often requires custom scripts to integrate, analyze, and compare the data sets. To facilitate such differential analyses, various tools have been developed in recent years that enable scientists to detect PPI rewiring events between samples belonging to two conditions. The features of the webservices and tools presented here are summarized in Table IV.2.

4.1. PPICompare

As a package to be used down-stream of PPIXpress, PPICompare [260] is another standalone tool developed in the Helms group that identifies significantly rewired interactions between two sets of condition-specific PPI networks. Based on information contained in the input networks, it can determine the reason for each rewiring event and assembles a small set of causes that can explain all such events. Figure 3 presents an overview of the method that is further explained below.

4.1.1. Differential PPI Network Construction

First, the PPICompare tool performs independent pairwise comparisons across the two sets of condition-specific PPI networks provided as input by the user (Figure 3, step 1). In each such inter-group comparison, it is noted which interaction is added or removed in the second PPI network, and the rewiring probability is calculated as the Jaccard distance between the interaction sets of the two compared PPI networks. Subsequently, the general rewiring probability is computed as the mean of the comparison-specific rewiring probabilities, and the number of additions (positive) and removals (negative) is summed for each interaction (Figure 3, step 2). Interactions without any changes or with a null-sum, indicating a balance of additions and removals, are not included in the differential network.

Given the overall rewiring probability, the statistical significance of each potential rewiring event is determined with a one-tailed binomial test followed by false discovery rate (FDR) correction for multiple hypothesis testing with the Benjamini-Hochberg method at a user-defined threshold (Figure 3, step 3). A differential PPI network consisting of significantly rewired interactions is provided as output. If the

input PPI networks contain information about the respective dominant isoform of each protein, PPICompare can additionally report for each rewiring event if it is caused by differential expression, dominant isoform switching, or possibly a combination of both (Figure 3, step 4).

On that basis, a bipartite graph of significantly rewired interactions and individual causal reasons is constructed (Figure 3, step 5). For each such reason, a score is computed from the number of significant rewiring events affected by it and the number of pairwise comparisons in which it took place. Determining a small set of causes that can explain all rewired interactions is then a weighted set-cover problem that is solved with the application of a greedy algorithm and the resulting collection of reasons is reported by the tool.

4.1.2. Application: Interactome Rewiring in Hematopoiesis

To evaluate PPICompare, a case study was performed on hematopoietic interactome rewiring [260] . First, PPIXpress was applied to 59 samples of 11 hematopoietic cell types from the BLUEPRINT epigenome project [45] to generate the cell type-specific PPI networks. The differential interactomes of adjacent cell types in the classical blood development progression model were then constructed with PPICompare and further analyzed. When comparing results on undersampled data sets, it turned out that a minimum number of three samples per group is required to yield robust results. Then, the statistical model employed by PPICompare is able extract most differentially altered interactions of possible relevance.

In most rewired interactions, differential gene expression of a single interaction partner was identified by PPICompare as the cause. A comparison with rewiring instances in which both interaction partners were deregulated showed that concurrent deregulation occurred more often in similar processes and known protein complexes, and thus possible functional modules more generally. A closer examination further indicated that different causes can be responsible for the same rewiring events. Underlining the importance of considering AS events in differential PPI analyses, alternative splicing was the identified reason for many differentially altered interactions relevant to the hematopoietic development transitions. Alternatively spliced proteins that were part of the set of most explanatory causes were associated with transcriptional control. Furthermore, proteins associated with hematopoiesis and targets of hematopoietic transcription factors were significantly overrepresented amongst the set of proteins participating in rewired interactions.

5. Use-case comparison

To compare the tools from a user's standpoint, we inspected the human tumor necrosis factor receptor protein 6 (TNR6) encoded by the FAS gene (UniProt accession: P25445) with the software tools and webservices reviewed here. Due to the large number of such tools, not all results can be discussed in detail, and we will

mainly focus on the protein-level webservice HIPPIE, the domain-based webservices and tools DIIP, DIGGER and PPIXpress, and the differential network analysis tool PPICompare. An overview of all protein-level resources can be found in Table IV.1, while Table IV.2 summarizes features of the tools DIIP, DIGGER, PPIXpress, SPECTRA, DifferentialNet, and PPICompare. Figure IV.1 illustrates the graphical output generated by selected tools when using the human protein TNR6 as input. We compare the tools with respect to user experience and significance of results.

Differences already emerge in the input scenarios. While most protein-level resources accept query proteins, or like HIPPIE, DIIP and DIGGER additionally accept a PPI network as input, PPIXpress strictly requires a PPI network with expression data to expose condition-specific subnetworks with confidence. The integration of expression data for interaction network construction is also possible with MyProteinNet, DIGGER, and SPECTRA, although this is optional. PPICompare requires two sets of context-specific PPI and DDI networks generated by PPIXpress and is thus optimally used downstream of PPIXpress for reliable discovery of protein rewiring events caused by differential expression. For that reason, we illustrate the use of PPIXpress for the TNR6 interaction network specific for neuronal stem cell and compare the output network to a H1 stem cell- specific network using PPICompare.

The chosen tools produced results that vary in terms of information and presentation. For HIPPIE, DIGGER and PPIXpress, the visualization of network topology is available at different levels of analysis. HIPPIE, as a protein-based tool, shows a network of genes interacting with FAS where edge weights indicate interaction strength (Figure IV.1A). The domain-based tool DIGGER additionally shows associated domains and DDIs in the network of interacting proteins, where the edges indicate whether an interaction is missing or if residue-level evidence is available (Figure IV.1C). With PPIXpress, one can use Cytoscape to visualize the condition-specific weighted PPI or DDI network (Figure IV.1D). Results produced by PPICompare, including information for gained and lost protein interactions or the statistical significance of the rewiring event, can also be illustrated with Cytoscape (Figure IV.1E). As a tool that does not generate graphical visualizations for the output networks, DIIP results are summarized as look-up tables where the interactions between each isoform of the queried protein and their interacting partners are listed in long format (Figure IV.1B). Similarly, IID generates a table with interactions involving the query proteins and optionally a table with significantly enriched annotations, a feature that HIPPIE provides as well. While DIIP classifies an interaction only as lost or retained, all other tools provide users with a metric to assess interaction confidence (HIPPIE, SPECTRA, DifferentialNet, and PPIXpress) or domain-based reliability of the found PPIs (DIGGER and PPICompare).

As a consequence of differences in the underlying data sources, methodologies and focuses, the results for TNR6 differ in terms of the number of found interactions and the type of information. For example, HIPPIE found 118 PPIs involving TNR6,

while IID found 367 including predicted interactions and 122 when only considering experimentally validated ones. Both computed significant enrichment of proteins associated with cancer among the interaction partners. In contrast, DIIP found two experimentally observed interactions with underlying DDIs for TNR6, namely with the tumor necrosis factor ligand 6 (TNFL6, UniProt accession: P48023) and the FAS-associated death domain protein (FADD, UniProt accession: Q13158). It predicted that for all but one TNR6 isoform the interaction with FADD would be lost, whereas most isoforms can enter domain-based interactions with TNFL6. DIGGER additionally offers isoform- and exon-level visualization of the interacting domains. It is thus possible to visually inspect how the domain presence leads to the loss or retention of the protein interaction in TNR6 isoforms predicted by DIIP. A similar analysis can be performed based on the results produced by PPIXpress but requires users to manually visualize and overlay the generated PPI and DDI networks. While comparing sample-specific PPI and DDI networks generated by PPIXpress from H1 stem cells to those from neuronal stem cells, PPICompare detected two statistically significant rewiring events ($p < 0.01$) involving TNR6. In neuronal stem cells, TNR6 gains interactions with FAIM2 (UniProt accession: Q9BWQ8) and TCAP (UniProt accession: O15273). FAIM2 is an important regulatory molecule for apoptotic control during neurological development and pathogenesis [206], while TCAP was associated to dendrite and axon formation during neurogenesis [262].

In summary, the results of the reviewed tools are diverse and suggest that users should weigh the advantages and disadvantages of each method carefully for the purpose of their specific needs. Depending on the use-case, it may be beneficial to cross-check and compare the results from multiple tools to obtain a richer picture.

6. Conclusion

In summary, there are now several sophisticated tools available to construct context-specific protein-protein interaction networks. These include both webservices as well as stand- alone software packages, some of which support domain-level and isoform-level analyses. Differential analysis of context- specific PPI networks is currently only possible with the tools SPECTRA, DifferentialNet and PPICompare. SPECTRA and DifferentialNet are available as easily accessible web-based resources for human PPIs. On the other hand, the stand- alone tool PPICompare more generally enables identifying statistically significant rewiring events between two groups of samples. Lastly, integration with other data types appears worthwhile.

7. Author Contributions

TW developed the tools PPIXpress and PPICompare, VH supervised the project and acquired funding. VH, MH, and TD wrote the manuscript.

8. Funding

Work in the Helms group that led to the development of the tools PPIXpress and PPICompare was funded by Deutsche Forschungsgemeinschaft (SFB 1027, project C3).

9. Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter V.

Project 5: PPIXpress and PPICompare Webservers infer condition-specific and differential PPI networks

The manuscript presented in this chapter has been published as "Do, H. T. T., Thangamurugan, S., Helms, V. (2025). **PPIXpress and PPICompare Webservers infer condition-specific and differential PPI networks**. *Bioinformatics Advances*, vbaf003". Meanwhile, PPIXpress and PPICompare webservers can be accessed at https://service.bioinformatik.uni-saarland.de/ppi-webserver/index_PPIXpress.jsp and https://service.bioinformatik.uni-saarland.de/ppi-webserver/index_PPICompare.jsp, respectively. My contributions to this project include implementing the PPIXpress and PPICompare webservers, as well as writing the initial draft of the manuscript.

1. Introduction

Protein-protein interaction networks (PPIN) play a pivotal role in organizing the proteomics landscape that controls most cellular activities. A wide range of softwares exist to setup such networks based on knowledge pooled across various databases. Some tools also allow to infer context-dependent interactions that contribute to the complexity of different biological systems [93] and provide insights into protein rewiring events. One workflow is based on PPIXpress [259] and PPICompare [260], two softwares developed in our research group. For example, PPIXpress has been employed to construct tissue-specific PPINs between normal and carcinoma tissues across 13 cancer types [110]. So far, PPIXpress-PPICompare remains one of only two workflows that enable users to prune a protein interaction network based on transcriptomic data of their choice [93, 149].
Here, we introduce two novel webservers for PPIXpress and PPICompare with core functionality similar to the standalone versions. Our services aim at streamlining

the PPIN analysis workflow and at providing a comprehensive view of the dynamics and alterations in network topology. The PPIXpress webserver enables users to infer tissue-specific networks from a self-provided PPIN or database-available PPIN using transcript/expression data uploaded as input. In a second step, significant rewiring events are identified by PPICompare by comparing two sets of networks constructed by PPIXpress for two different conditions. Both webservers provide downloadable visualization of condition-impacted protein interactions or domain interactions.

2. Methods

2.1. PPIXpress

PPIXpress infers condition-specific PPINs from a reference PPI network based on transcript abundances in user-defined samples by performing two main tasks: protein-domain mapping and network contextualization [259]. PPIXpress requires as inputs one reference protein interactions network file and at least one expression data file. In the first step, PPIXpress translates the reference PPIN into a reference domain-domain interaction network (DDIN) by mapping each pair of proteins involved in an interaction to two protein domains that have been documented to interact. This task requires domain annotations from the Pfam-A database [169] and high-confidence domain interaction data from DOMINE [272] and IDDI [119], supplemented by the most recent 3did/iPfam databases [172, 77]. Note that, if no Pfam domain annotation exists for a protein, PPIXpress defines a dummy domain for it. Using the expression data of a specific sample, PPIXpress then prunes the reference DDIN to a network that contains only interactions between dominantly expressed transcripts. The set of PPIs explained by at least one DDI remaining after trimming is the condition-specific PPIN as the main output of PPIXpress. Users may select optional auxiliary results including a list of major transcripts and corresponding DDINs.

2.2. PPICompare

PPICompare identifies a set of protein interactions that are statistically significantly rewired between two groups of samples represented by condition-specific PPINs [260]. First, for each inter-group pair of samples i, PPICompare defines a pairwise rewiring probability $P_{rewired_i}$as the fraction of rewired edges (both vanishing and appearing) over all edges present in all samples. The total rewiring probability for the initial differential network Δ between the two sample groups is computed by averaging all inter-group rewiring probabilities:

$$P_{rewired} = \frac{1}{N} \sum_{i}^{N} P_{rewired\_i} \tag{V.1}$$

For an edge alteration (u, v) between protein nodes u and v ($(u, v) \in \Delta$)to be a true randomly rewired event, its occurrence must be significantly greater than the occurrence of random edges in the differential network Δ. The likeliness for (u, v) to be detected due to random variations in expression is computed by a one-tailed binomial test:

$$p(u, v) = 1 - \sum_{j=0}^{|\Delta(u,v)|-1} \binom{N}{j} (P_{rewired})^j (1 - P_{rewired})^{N-j} \tag{V.2}$$

whereby $|\Delta(u, v)|$is the annotated number of rewiring events of this edge over all N pairwise comparisons. After applying a Benjamini-Hochberg correction to the p-values, only significant rewired events are retained in the final differential network. PPICompare also identifies a minimal portion of transcripts that could generate the systemic rewiring. By viewing transcripts and their edge alterations as two classes of a bipartite graph, finding the smallest set of transcripts that explain the largest number of rewiring events becomes a weighted set cover problem. Each transcript, or rewiring reason i, is assigned a score $s_i$and a weight $w_i$subsequently:

$$s_i = pw_i \times rw_i \tag{V.3}$$

$$w_i = s_{\max} - s_i \tag{V.4}$$

where $rw_i$is the number of alterations caused by reason i, $pw_i$is the number of pairwise comparisons where those alterations are found and $s_{\max}$is the maximum value of $s_i$over all considered transcripts. PPICompare solves the problem of minimizing the sum of weights using a greedy heuristic approach [47]. This final step results in a small set of transcriptomic changes that explain the systematic differences in the differential network between two cell conditions.

3. Workflow for differential protein interaction networks analysis

Our differential PPIN analysis workflow comprises of PPIXpress and PPICompare webservers, which can be run independently. Fig. V.1 shows the workflow of a typical project.

3.1. PPIXpress Webserver

PPIXpress analysis entails (i) Data upload and setting of parameters, (ii) Construction of condition-specific network(s), and (iii) Query and visualization of a protein network. In part (i), PPIXpress requires two types of input, a reference PPIN, which can be defined by the user or retrieved from IntAct [180] or mentha [37] databases given a taxon number, and at least one user-defined transcript/expression

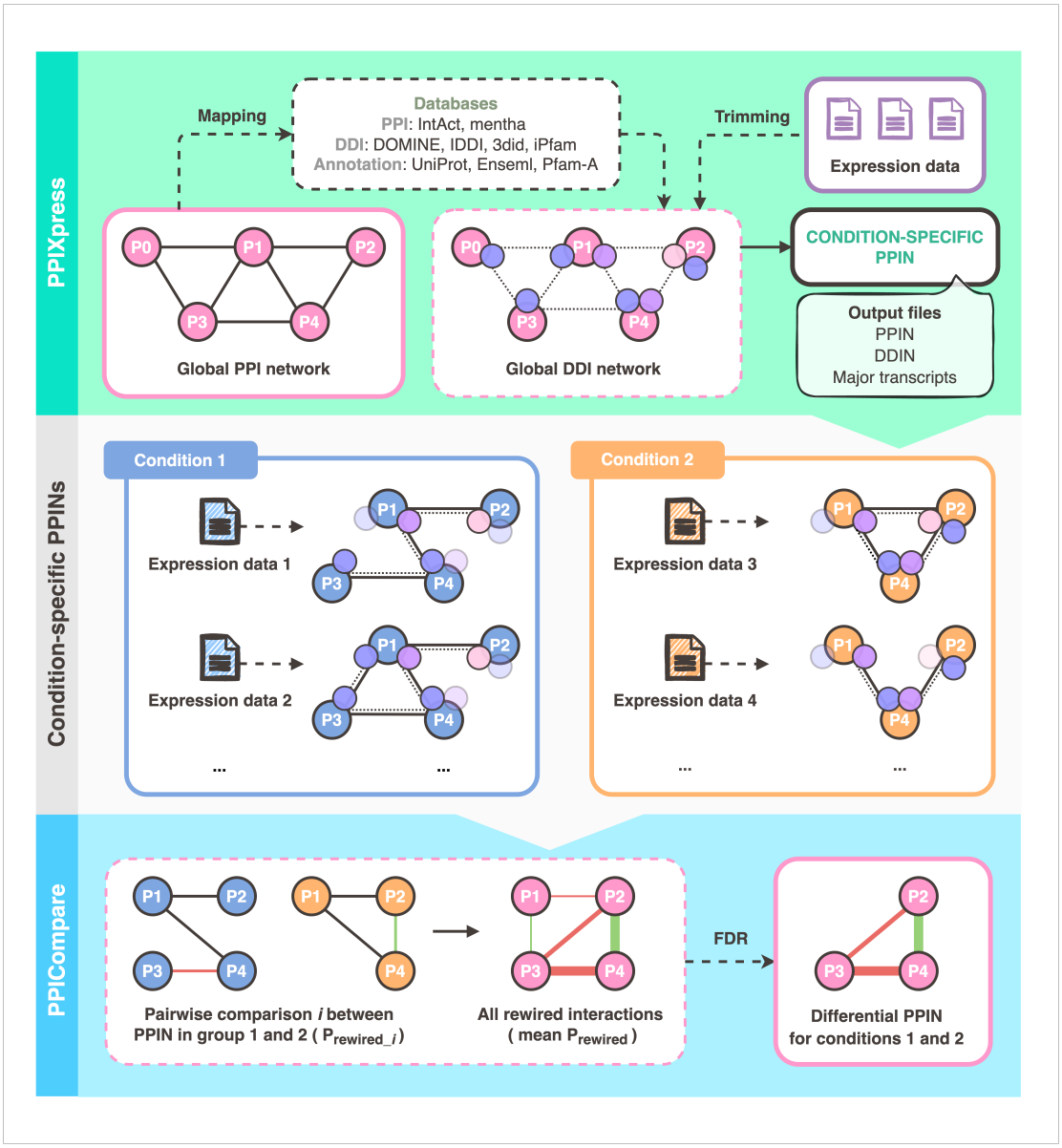

Figure V.1.: Workflow of PPIXpress and PPICompare webservers for protein network analysis. The boxes lined in solid represent input/output in the front-end that allow users to use and interact with the tools. Boxes lined in dashed include information about databases used by the tools or analysis processes in the back-end. Within a network, protein-protein interactions and domain-domain interactions are denoted by solid lines and dashed lines, respectively. In the middle panel, the faded nodes show strongly down-regulated or not expressed domains that are pruned from the global DDI network by PPIXpress when constructing the condition-specific protein-protein interaction networks.

data set. Expression data required for network contextualization can be provided in various file formats, including outputs of popular transcript alignment and

quantification tools (such as Cufflinks, GENCODE, Kallisto, RSEM, and TCGA RSEM), or plain text formats with either FPKM (Fragments Per Kilobase Million) or TPM (Transcripts Per Kilobase Million) values. The construction of condition-specific network(s) (ii) is carried out for each uploaded expression data set, presumably representing different tissues or experimental conditions. Using the set of most abundant genes or transcripts of a protein from a sample, PPXpress derives condition-specific DDIs which dictate if a DDI in the original PPIN should be retained. The remaining DDIs are mapped back to the reference PPIN, resulting in condition-specific PPIs.
In the visualization panel of the PPIXpress webserver, the user may query a protein to inspect its direct neighbors in the PPIN that was pruned using the expression data from a particular tissue or condition. The display of the DDIN underlying this PPIN can be toggled by expanding or collapsing a protein node. Currently, the PPIXpress webserver only allows single-protein query instead of full protein networks, as Cytoscape.js takes quite some time for rendering a large number of protein nodes, especially when domain nodes are added. Alternatively, the trimmed networks for each tissue or condition can be downloaded and used as input to a local installation of Cytoscape on the user's end. Apart from the condition-specific networks, users can download DDINs, and the most abundant genes or transcripts used for their construction, which are essential inputs for the differential PPIN analysis of PPICompare.
The PPIXpress webserver allows enriching the reference network with functional association scores from the STRING database [234], ELM motifs, or updated UniProt accessions numbers for accurate domain annotation and mapping [244]. Data for DDI mapping can be retrieved from 3did [172] as default, or from locally precompiled DOMINE [272], IDDI [119], and iPfam [77] data sets. While users can use and prioritize gene abundance over transcript abundance for DDIN construction, PPIXpress also allows the normalization of transcript-level expression data by transcript lengths. The threshold for a gene or transcript to be considered expressed can be adjusted in the input of PPIXpress webserver. Its optimal value may depend on the particular application scenario. The PPIXpress analysis can be further tuned by removal of nonsense-mediated decay transcripts. A Help page and FAQs section provide a detailed list of input requirements and execution options for PPIXpress, help with the interpretation of its output, and explain how PPIXpress results may be transferred to the PPICompare webserver.

3.2. PPICompare Webserver

The PPICompare webserver performs differential analysis on multiple condition-specific PPINs generated e.g. by PPIXpress, including (i) Data upload, (ii) Detection of significantly rewired interactions and transcriptomic changes responsible for such changes, and (iii) Visualization of altered protein interactions. Two datasets are required for PPICompare, each must contain three types of out-

puts from PPIXpress, namely the condition-specific DDINs (*_ddin.txt(.gz)), PPINs (*_ppin.txt(.gz)), and a list of major genes or transcripts (*_major_transcripts.txt(.gz)). In the core part (ii), PPICompare compares the interactomes between the uploaded conditions in a pairwise manner and derives the set of inter-group rewired interactions with a probability for each alteration. The false discovery rate (FDR) is set by default to 0.05 as a threshold for significant rewiring interactions. Fig. V.2 showcases the network of proteins involved in the transcriptomic alterations between the compared conditions, together with the lost and gained interactions between them.

4. Case study

The PPIXpress webserver was used to construct condition-specific protein-protein interaction networks (PPINs) for RNA-seq datasets of primary melanoma samples and melanocytic nevi (Gene Expression Omnibus series ID GSE112509) [130]. To retrieve the reference human interactome, the taxon ID 9606 was provided. The gene expression data for each sample was loaded with the options set to "Gene-level only" and the expression level threshold set to 1. The run options for "Output DDINs" and "Output major transcripts" were selected as they are required for further processing using PPICompare. The condition-specific networks for each sample were constructed and visualised using their corresponding protein IDs. As two examples, Fig. V.2A and Fig. V.2B show the interactors of proteins Q13263 and Q9UBF1 identified in a melanoma sample. A complete list of the interactors of Q13263 and Q9UBF1 in each sample of the melanoma and nevi groups is provided in Supplementary_Data.xlsx.

The primary melanoma- and melanocytic nevi-specific networks were loaded as two input groups into PPICompare. Then, we selected the option to return the attribute table and set the FDR for significantly altered PPIs to 0.05. The significant rewiring events between the two groups were constructed as a differential network and visualised in the "Network Visualisation" panel (shown in Fig. V.2C). The analysis found that the interaction between the proteins Q13263 (gene name: TRIM28) and Q9UBF1 (gene name: MAGEC2) is significantly rewired between those two groups. This is consistent with previous studies [227, 146] showing that MAGEC2 interacts with TRIM28 in melanoma cells. In tumour cells, the expression levels of MAGEC2 depend on the expression levels of TRIM28, as a significant decrease in MAGE28 levels was observed in TRIM28 depleted cells.

To our knowledge [93], there exists no tool comparable to PPICompare that detects interactome rewiring events and the according transcriptomic alteration for user-defined datasets. SPECTRA [165] and DifferentialNet [18] are two webservers that also generate context-specific differential PPINs, but both allow only the analysis of human data. There are several similar tools to PPIXpress; nonetheless,

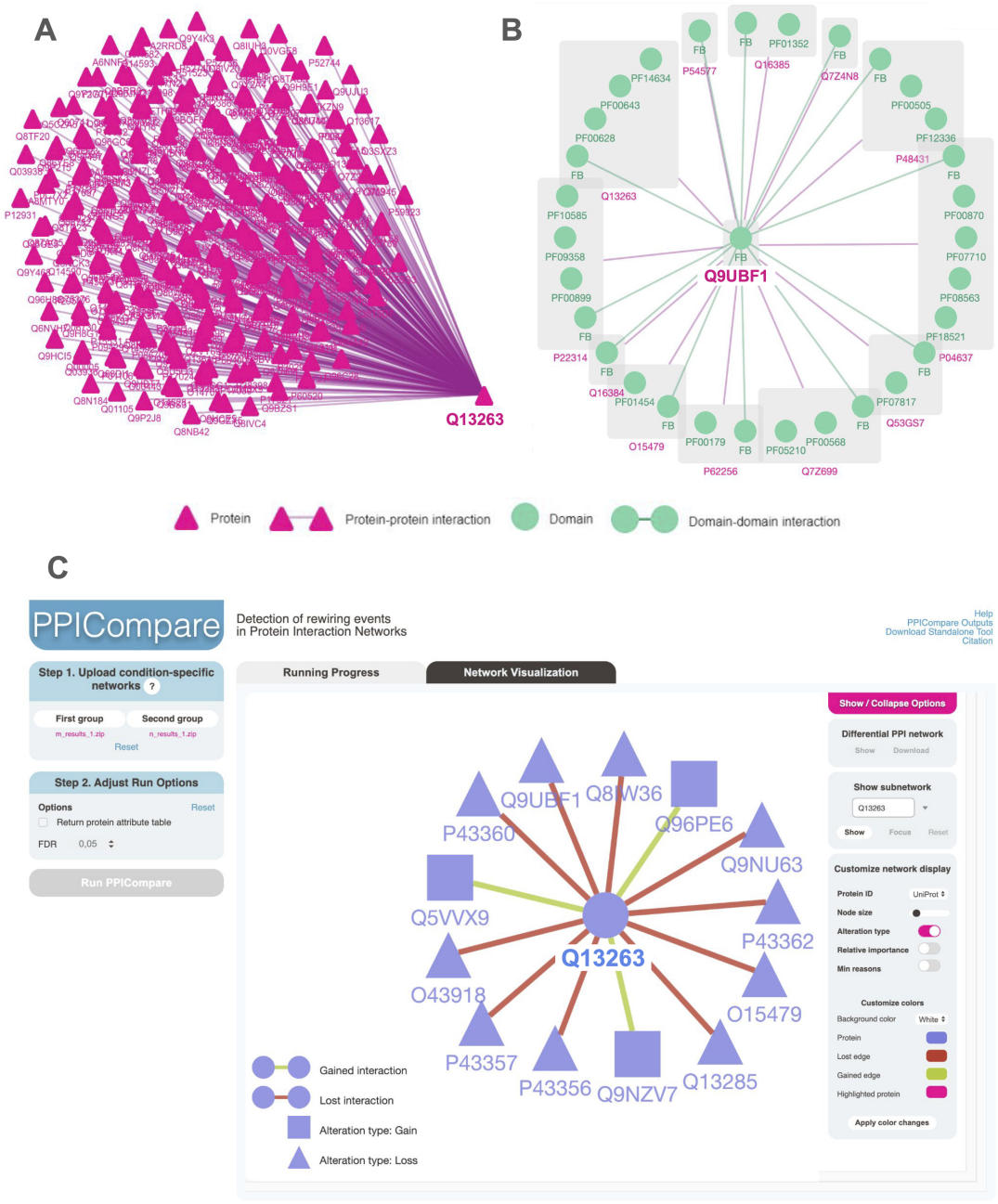

Figure V.2.: Case study of analyzing the differential protein-protein interaction network (PPIN) between melanocytic nevi and primary melanoma samples by PPIXpress and PPICompare webservers. (A) and (B) provide two examples for the network visualization by the PPIXpress webserver, where the PPINs specific for BO-027-SM_M2 sample from the dataset GSE112509 containing the proteins Q13263 (TRIM28) and Q9UBF1 (MAGEC2) are shown respectively. (C) shows a snapshot of a differential PPIN constructed by PPICompare that encapsulates the differences between nevi- and melanoma-specific PPINs from PPIXpress. Here, gained and lost interactions in nevi samples are respectively denoted by green and red edges, while the corresponding emerging proteins are represented by triangle nodes and vanishing proteins by square nodes.

they have different focuses, such as predicting alternative isoforms (DIIP [84]) or DDI-based missing interactions at isoform/exon-level (DIGGER [149]).

5. Conclusion

Two new webservers PPIXpress-web and PPICompare-web are introduced that enable biomedical scientists to identify statistically significant rewirings between PPINs of two conditions. The established workflow supports rapid inference of protein networks directly from gene or transcript expression data with network visualization and protein query in constructed condition-specific or differential networks.

6. Acknowledgements

We thank Thorsten Will for his contributions to the standalone versions of PPIXpress and PPICompare, as well as Markus Hollander and Aram Papazian for testing the new webservers and for helpful comments.

7. Author Contributions

Hoang Thu Trang Do implemented the webservers and wrote the article. Sudharshini Thangamurugan developed the case study and wrote the article. Volkhard Helms wrote and reviewed the article.

8. Conflict of Interest

No competing interest is declared.

9. Funding

This work was supported by a grant of Deutsche Forschungsgemeinschaft to V.H. via CRC 1027 (project C3).

Chapter VI.

Project 6: Tissue-specific RNA binding protein networks provide insights on splicing processes

This project is conceptualized at the end of my doctoral study, where I aim to study how rewiring events at transcriptomic and proteomic levels are linked.

1. Introduction

The gaps between transcriptomes and proteomes have been long addressed: While genetic information is passed down from RNA to proteins via translation with consistently high correlations between protein expression and mRNA levels, protein levels may variate strongly across different organisms and contexts without strong dependence on corresponding genes [238]. This phenomenon has been captured in various studies by low correlations (0.14-0.59) in the variabilities of protein expressions and transcripts across normal and tumor cells, or across pluripotent stem cells [238, 19, 213, 66]. While many post-translational regulations may affect the protein pool [92, 217], it is still unclear to what extent transcript alternative events may influence proteomic profiles and activities, as well as how much of the changes in transcript-level manifest in cell physiology and behaviors.

Alternative splicing (AS) is a major contributor to transcriptomic diversity, which is regulated by RNA-binding proteins (RBPs) and their interactions with splicing regulatory elements in pre-mRNA [237, 258, 160]. RBPs are also involved in the splicing machinery as components of the spliceosome, a large ribonucleoprotein complex that catalyzes the removal of introns from pre-mRNA after translation [258]. The great variety of alternate transcripts is as well a major factor contributing to high complexity of the protein pool, as different transcripts of the same gene might even interact differently [270]. Thus, the changes in the transcriptome due to AS may also be reflected by the perturbances in RBP networks (RBPN) to a certain degree.

In a former study (Project 1), I have investigated the connection between histone modifications and alternative exons in the transcriptomes of 19 human cells/tissues from the Human Epigenomes Atlas [210]. To expand the knowledge on how these epigenetics-associated rewired transcripts control the cell activities and development through altering the functional interactomes, in this study, I investigate the differences in transcriptome-based RBP networks specific to each of the same 19 cells and tissues and link them to the results of our previous study. Using the implemented webservers PPIXpress and PPICompare (Chapter 5), I constructed the sample-specific RBP networks and compared them to identify the rewired interactions that distinguish the cells/tissues. The binding dynamics of RBPs in these networks were also analyzed to validate the plausibility and biological functions of these interactions.

2. Methods and Methods

2.1. Materials and preprocessing

Human Epigenomes Atlas transcriptomic data The transcriptomic data used in this study were intended to match with those from Project 1 for the investigation on the connection between alternative transcripts and alternative RBP interactions. Thus, the methods and standards for RNA-seq data retrieval from the Human Epigenomes Atlas [210] and preprocessing were adopted directly from the Materials section 4.1.1. The included tissues and cells are listed in Table A.1 in Supplementary Materials and their metadata which contain sample sources, biosample types and used parameters for bio-assays are tabulated in Tables A.2- A.4.

RBP2GO reference RBP network RBP2GO [40] is a comprehensive database for RBPs and RBP-interactions that are curated from all currently available proteome-wide studies across 13 species. Among 22552 available RBP candidates, 6100 proteins with high confidence scores are collected from human studies. The database also provides STRING-derived PPIs for each RBP, which were used to construct the reference RBP network in this study.

mRNP Chronology time-resolved interactomes The database mRNP Chronology [46] contains the human interactome data with sole focus on RBPs that aligns well with our aim of investigating the temporal dynamics of RBPs in tissue-specific PPINs and rewired PPINs. Using longitudinal pulse-chase metabolic labeling and selective crosslinking experiments, mRNP Chronology records the binding dynamics across 10 time points for 734 RBPs. The RBPs are classified based on their normalized log2 protein intensity and peak binding time into 7 temporal clusters (I-VII) using k-means clustering. The clusters are further grouped into two broad categories, namely early binders and late binders, with evidently distinguished characteristics regarding localization, binding preferences and biological functions.

2.2. Constructing cell/tissue-specific RBP networks with PPIXpress

The construction of cell/tissue-specific RBP interactions networks was performed using PPIXpress webserver (Project 5) using default parameters. PPIXpress requires a reference PPIN and a set of transcriptomic data to construct sample-specific PPINs. Since the study is investigating rewired interactions in RBP networks, the PPIN containing only RBPs (or RBPN) was retrieved from RBP2GO [40] to be used as the reference PPIN. Using the transcriptomic data from samples of 19 human cells/tissues obtained from the Human Epigenomes Atlas [210], I constructed the sample-specific RBPNs. Since only cells/tissues with replicates are included in the study as a requirement for DESeq2 [151] and PPICompare [260], a cell/tissue-specific RBP network would consist of at least two sample-specific networks. For the sake of simplicity, tissue/cell-specific RBP networks are now preferred to as "tissue-specific RBPNs".

2.3. Comparing tissue-specific RBP interaction networks with PPICompare

To identify the rewiring events within the RBPNs of two different cells/tissues, PPICompare webserver (Project 5) was used with default settings. The software takes the PPIXpress-constructed sample-specific RBPNs for each comparison group and derives the "differential RBPN" of rewired interactions that distinguish the two tissues.

2.4. Binding time difference in RBPNs

The proteins in the tissue-specific and differential RBNPs were annotated with their peak binding time in minutes using the data from mRNP Chronology [46]. Binding time refers here to when the RBPs bind to the mRNA molecule after it was newly synthesized by RNA polymerase. To validate the possibility of PPI interactions in these RBPNs, the difference in peak binding time between interacting proteins was computed. With the hypothesis that the binding time differences in PPIs in RBPNs are lower than in a random set of PPIs, I generated a set of Mock PPIs (mPPIs) for comparison. The mPPIs were hypothetical, presumably not existing interactions that were created by subtracting the existing PPIs from all possible pairwise interactions between all proteins in the RBPN. The significance of the difference in binding time between PPIs and mPPIs was determined by Mann-Whitney U test with FDR-adjusted p-values.

Two reference RBPNs were also included in the analysis to emphasize the differences in behaviors of detected RBPs in tissue-specific networks against random RBPs. The first reference RBPN was the one used for the construction of tissue-specific RBPNs, which contains only interactions among RBPs retrieved from

RBP2GO. The second reference RBPN was constructed from mRNP Chronology data [46] by linking the available proteins to their BioGRID partners. Furthermore, only RBPs and PPIs which can be annotated and are experimentally validated in mRNP Chronology were considered, and PPIs and mPPIs containing any protein that does not exist in the database were removed.

2.5. Comparing differential RBPNs and RBP interactions

To compare the differential RBPNs generated for each pair of cells/tissues, Jaccard similarity index was used and is defined as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{VI.1}$$

where A and B are the PPIs set in two differential RBPNs, $|A \cap B|$ is the number of overlapping PPIs between the two sets, and $|A \cup B|$ is the number of all PPIs in both sets. For each cell/tissue, the set of tissue-specific PPIs from all differential RBPNs they are involved in was collected and the similarity between the tissue-specific PPIs was also computed using the Jaccard index above. Figure C.2 shows the clustering of differential RBPNs and tissue-specific PPIs based on the Jaccard similarity index. Since the PPINs from mRNA Chronology and RBP2GO were not constructed by PPIXpress and were not suitable for PPICompare, they were excluded from this analysis.

3. Results and Discussion

Using PPIXpress to construct sample-specific RBPNs for 19 cells and tissues of the Human Epigenomes Atlas project, I was able to identified the RBP interactions (or simply PPIs) that likely result from the distinct transcriptomic profiles of each cell and tissue type. In Project 1 , we clustered those 19 cells/tissues based on their shared "Epispliced genes" - genes with exon rewiring highly associated to histone modifications, and found clusters of cells/tissues in distinct developmental stages (Figure I.5). Remarkably, the clustering of samples based on sample-specific PPIs from PPIXpress in this study gives similar result to epispliced genes (Figure C.1). As the clustering based on only differential exon usage (Project 1, Figure A.3) was not successful like those based on histone-related alternative exons and rewired RBP interactions, future experiments with clustering based on only transcriptomic data would clarify whether the changes in interactome are linked to alternative splicing and histone modifications.

To validate the possibility of these PPIs assuming the interacting RBPs would bind to mRNA at similar timings, I annotated each RBP with its peak binding time in minutes using the data from mRNP Chronology and computed the difference in binding time for each interaction. For all 19 cells/tissues, the interacting proteins

in actual PPIs bind to their targets more simultaneously than proteins in random mPPIs in general (Figure VI.1). This co-occurrence is also observed in the reference RBPN constructed from RBP2GO data and mRNP Chronology data, although a much less visible discrepancy between actual PPIs and mPPIs is spotted in the latter. The lower discrepancy in binding times in RBP complexes as compared to in random RBP interactions was also found by Choi el al. [46]. Here, the authors of mRNP Chronology showed more synchronized binding among the RBPs belonging to known protein complexes retrieved from CORUM [86], or to RNase-insensitive, RNA-independent "apo-stable" complexes, but not to RNase-sensitive, RNA-dependent "structural" complexes. It would be interesting to investigate if RBPs in sample-specific RBPNs belong to certain functional complexes and their roles in distinct cells/tissues.

While the binding time differences in PPIs are generally smaller than in mPPIs in all investigated samples, their distributions show a striking difference between tissue-specific/RBP2GO RBPNs and mRNP Chronology RBPNs (Figure VI.2). The distributions of binding time differences in PPIs and mPPIs groups in mRNP Chronology's RBPN overlap completely and show almost no difference in co-binding activities with a single peak at 0 minute. While the distributions of mPPIs-binding time difference in tissue-specific/RBP2GO RBPNs are highly identical to that in mRNP Chronology's RBPN, the PPIs group show distinct distributions with two peaks at around 0-5 minutes (peak region I) and 50-60 minutes (peak region II). The first peak region I is likely to be the result of the co-occurrence of RBPs in the same RNP complex, which agree with Figure VI.1 and with the results for RNA-independent complexes in Choi et al.'s study [46]. The peak region II shows interactions between RBPs that do not share the same mRNA binding time. How are those interactions possible considering the temporal factor? This may be explained by the presence of RBPs with broader binding peaks, which are potentially recruited at multiple stages by a wider range of partners, or RBPs with more stable interactions with mRNA that are captured at later time points [46]. Indeed, the histograms of binding time clusters in Figure VI.3 show a strikingly higher proportion of RBPs belong to cluster IV in tissue-specific/ RBP2GO RBPNs (left) as compared to mRNP Chronology RBPN (right) (approximately 18-22.5% compared to 12%, respectively). While more than 50% of RBPs are still in "early binders" clusters II and III, the percentages of RBPs in "late binders" clusters IV and V are higher in RBP2GO and tissue-specific RBPNs, which explain the larger binding time differences shown in Figure VI.2. Interestingly, the authors of mRNP Chronology also found a slightly delayed binding temporal patterns for splicing-related and cytoplasmic mRNA remodeling factors, as compared to in early binding RBPs which are mainly involved in transcription and pre-mRNA 3' end processing.

PPICompare webserver (Project 5) was used to compare tissue-specific RBPNs in a pairwise manner to determine the RBP interactions that distinguish the cells/tissues. Due to the samples availability limitation and high PPI selection

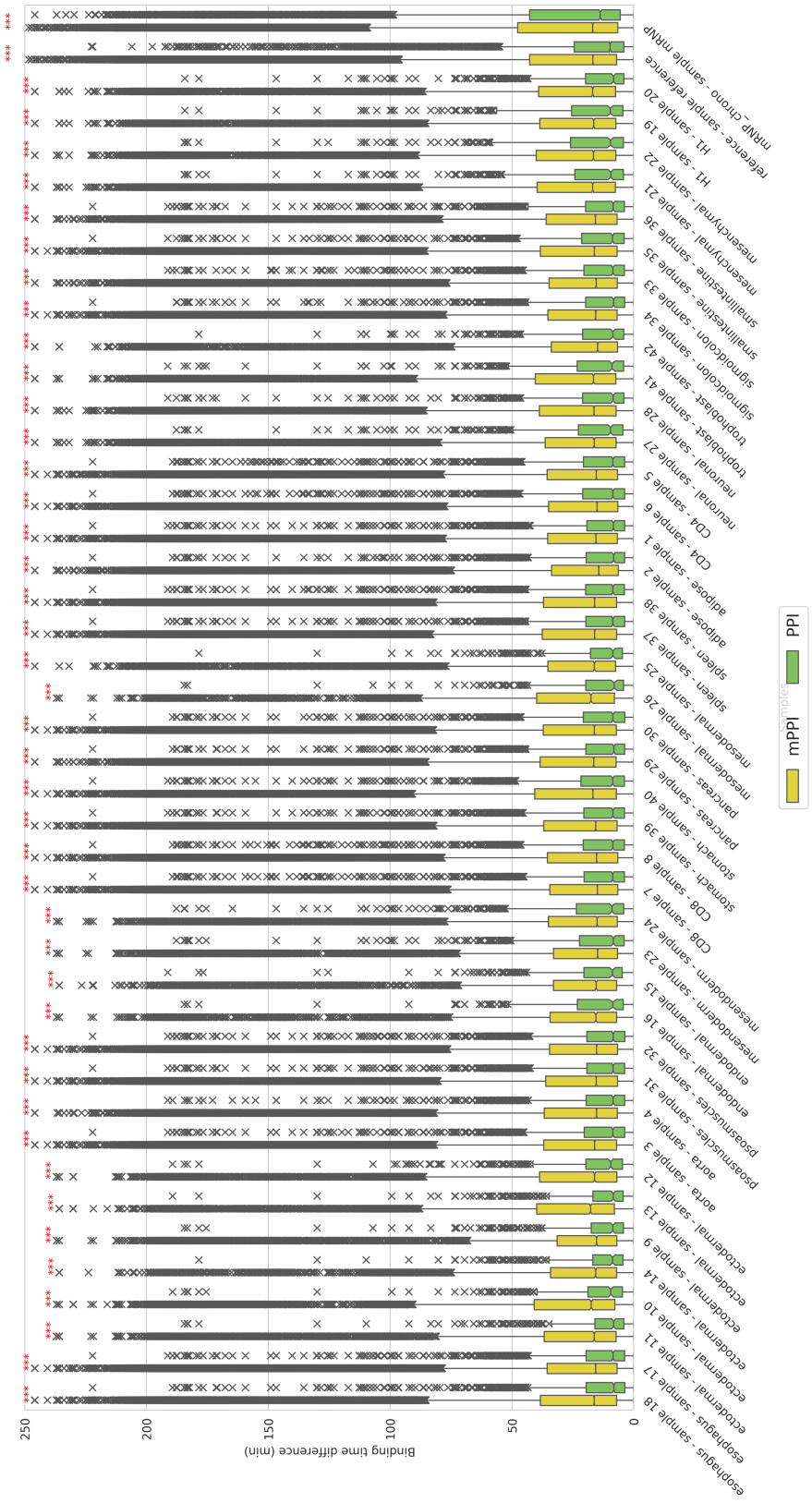

Figure VI.1.: Binding time difference (min) for PPIs and mock PPIs (mPPIs) in cells/tissues-specific RBPNs and reference RBPNs. For each sample of 19 cells/tissue in the Human Epigenomes Atlas, a PPIN specific for its transcriptomic data was constructed using PPIXpress. (*to be continued*)

(*continue*) For each sample of 19 cells/tissue in the Human Epigenomes Atlas, a PPIN specific for its transcriptomic data was constructed using PPIXpress. An absolute binding time difference was computed for each PPI using the peak binding time information in minutes of each partner protein provided by mRNP Chronology [46]. The PPIN containing interactions between RBPs exclusively from RBP2GO [40] was used as the reference PPIN for PPIXpress. The PPIN for mRNP Chronology data was built using the set of BioGRID interactors provided for each protein available on the platform. For each constructed PPIN, mPPIs were generated by subtracting the existing PPIs from all possible pairwise interactions among all proteins in the PPIN. Significant difference in binding time difference between PPIs (green) and mPPIs (yellow) was determined by Mann-Whitney U test (*** indicates FDR-adjusted p-value < 0.001).

stringency, only for 12 cells/tissues shown in Figure C.2, the differential RBPNs were successfully constructed. The similarity between the differential networks, or how much rewired interactions they share, was computed using Jaccard similarity index (Equation VI.1). Consistent with the clustering results from Project 1 and

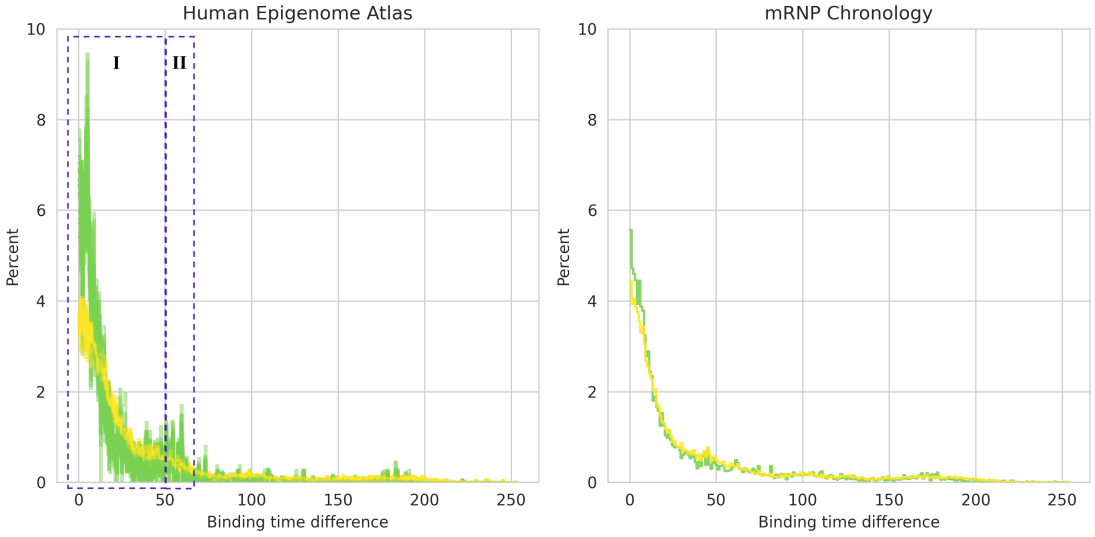

Figure VI.2.: Distribution of binding time difference (min) for PPIs and mock PPIs (mPPIs) interactions in cell/tissue-dependent RBPNs and reference RBPNs. The left figure shows the distributions of binding time difference within the PPINs specific for each sample of 19 cells/tissues of Human Epigenomes Atlas, as well as the reference PPIN that contains only interactions among RBPs retrieved from RBP2GO [40]. The right subfigure shows the distribution of binding time difference for mRNP Chronology PPIN [46]. The PPIN for mRNP Chronology data was built using the set of BioGRID interactors provided for each protein available on the platform. In both subfigures, the distributions for binding time difference in mPPIs are also included, while mPPIs were obtained by subtracting the existing PPIs from all possible pairwise interactions among all proteins in the PPIN of mRNP. Subfigure A is annotated with two peak regions I (0-5 minutes) and II (50-60 minutes). The distribution of binding time difference in PPIs and mPPIs is shown in green and yellow, respectively.

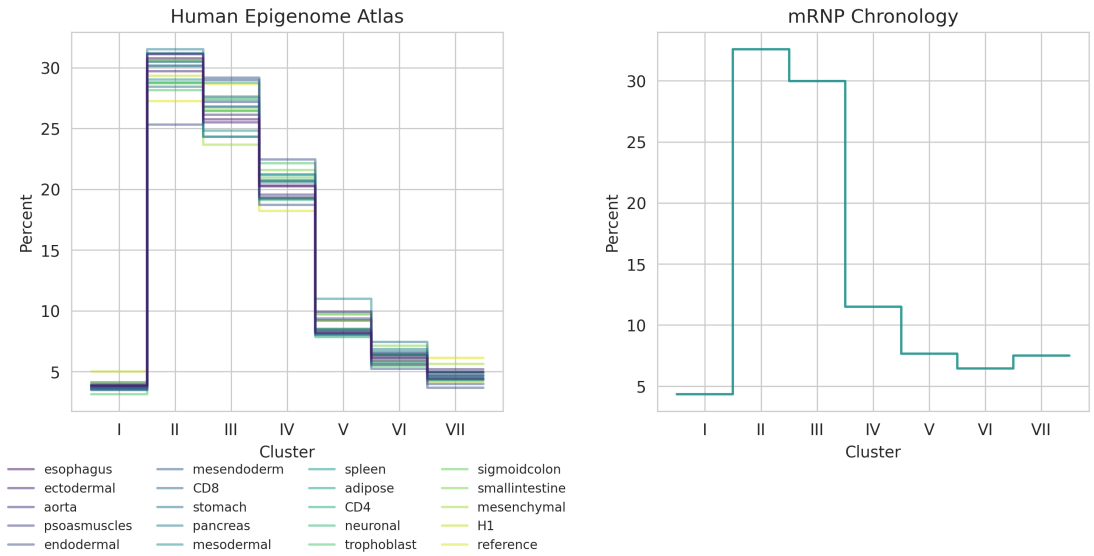

Figure VI.3.: Distribution of mRNP Chronology's clusters which proteins in cell/tissue-dependent RBPNs and references RBPNs belong to. Based on their binding dynamics, the RBPs in mRNP Chronology are categorized into 7 clusters numbered from I to VII. The left figure shows the distributions of clusters annotated to RBPs in PPIXpress-constructed PPINs for each sample of 19 Human Epigenomes Atlas cells/tissues, as well as the reference PPIN containing only RBPs retrieved from RBP2GO [40]. The right subfigure shows the distribution of RBP clusters for mRNP Chronology PPIN [46].

from PPIXpress (Figure C.1), many cells/tissues at the same developmental stages share more similarity in their differential RBPNs than with other cells/tissues.

To preliminarily inspect the functional relevance of the rewired interactions, I focused on the differential RBPN of CD4 and CD8 T-cells, two cell types that shared the largest fraction of shared epispliced genes in Project 1. Interestingly, CD4 and CD8 T-cells were also found here to share the most differential RBP interactions among all pairs of cells/tissues (Figure C.2B). Figure VI.4 shows the complete differential RBPN of CD4 and CD8 T-cells and the subnetworks of three splicing-related RBPs, namely HNRNPF, HSPA8 and PQBP1. These subnetworks show high connectivity both within and outside the RBP complexes. The HNRNPF subnetwork contains SFs of the hnRNA (HNRNPH3, HNRNPD, HNRNPM, $\cdots$) and SRSF families (SRSF1, SRSF2, SRSF6, $\cdots$) as well as key components of the spliceosome such as SNRNPs, U2AF2 and PQBP1 [72, 160, 255] (Figure VI.4B). PQBP1 is the core protein of another subnetwork of several other members of hnRNA SF family or spliceosomal SFs like SF3B1 and PRPF19 [160, 255] (Figure VI.4C). The heatshock protein HSPA8 with known scaffolding role in spliceosome assembly connects to both HNRNPF and PQBP1, as well as its DNAJ chaperone partners like DNAJA1, DNAJA2, DNAJC5 $\cdots$ [231]. The "losses" of these subnetworks in CD8 cells imply a potential shift to splicing down-

regulation. However, not only spliceosome units or trans-acting SFs are present in the differential RBPN and have the potential to reveal transcript rewiring events through unique sets of RBP interactions. Many other RBPs which are not directly involved in splicing, such as transcription initiation and elongation factors, chromatin remodelers, mRNA processing factors, must be considered to understand how transcriptomic rewiring events affect the interactomes and proteomes, resulting in distinct phenotypes [127, 279, 31]. Studies on binding targets, patterns and dynamics of specific RBPs in the differential RBPNs would provide more insights on systemic rewiring at multiple omics levels.

4. Conclusions

This study aims at connecting the histone-associated rewiring events in transcriptomes and the transcript-inferred RBP-interactomes for 19 cells and tissues of the Human Epigenomes Atlas project. Due to time limitations, the analysis focused on (i) showing the concordance in the analysis results of histone-linked AS events and rewired RBP interactions using the same transcriptomic data, and (ii) inspecting the differential binding dynamics of RBPs to suggest potential regulatory mechanisms underlying the observed transcriptomic changes. Future work should close the gaps between histone modifications, alternative exons/transcripts and rewired RBP interactomes by explaining how different RBP complexes and their binding behaviors are related to key epigenetic events.

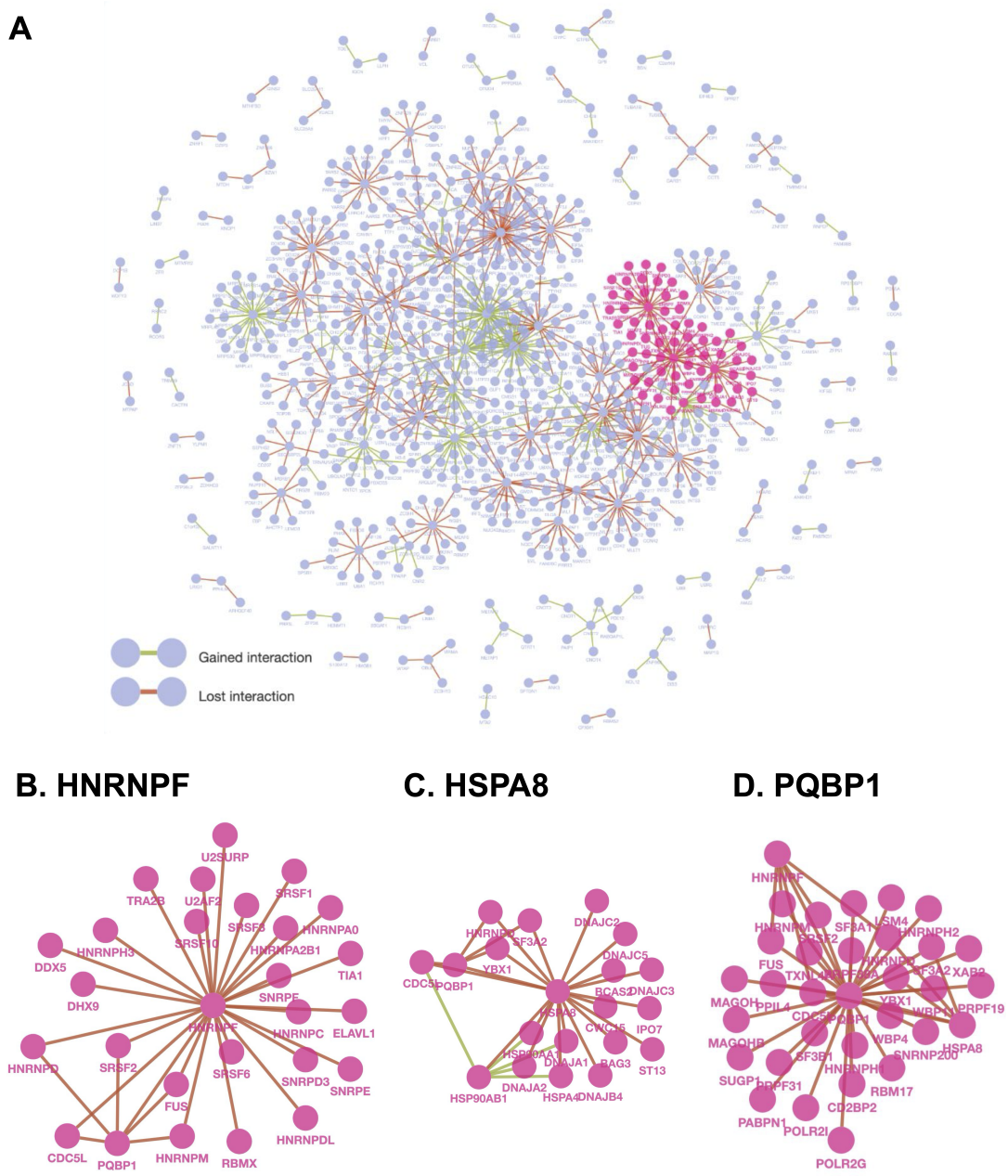

Figure VI.4.: Differential RNP-binding protein interaction network (RBPN) distinguishes key processes in CD4 and CD8 T-cells. The differential RBPN that highlights the unique interactions in CD4 and CD8 cells was constructed and visualized in PPICompare webserver using CD4/CD8-specific RBPNs. A. The complete differential RBPN with three subnetworks of major splicing-related proteins and their partners highlighted in pink. The subnetworks were selected based on literature and previous studies, with HNRNPF, HSPA8 and PQBP1 as central proteins to show highly connected RBP complexes which potentially contribute to alternative splicing between CD4 and CD8 cells. B-D. Closer snapshots of the splicing-related subnetworks of HNRNPF, HSPA8 and PQBP1, respectively. Gained interactions in the networks showing CD4-exclusive interactions are colored green, while lost interactions are in red.

Chapter VII.

Conclusion

This dissertation presents the approaches and findings of my doctoral research, which consists of six projects that collectively investigate rewiring events and key regulatory factors at transcriptomic and proteomic levels. Project 1 focuses on the interplay between AS and epigenetic deregulation, while Projects 2 and 3 explore the functional roles of specific transcript variant and RBPs in specific cell types and experiment contexts. The remaining projects, 4, 5, and 6, assessed and developed a new workflow consisting of two webservices for analyzing protein-protein interaction networks with the ultimate aim to study the differential interactions of RBPs across various cell types and tissues that may associate with AS.

Project 1 readdresses the relationship between AS and epigenomic alterations, however with an emphasis on systematic changes across developmental stages in 19 human cell and tissue types. By associating the changes in exon usages with changes in histone marks enrichment for each pair of cells/tissues, the study provides evidence for cell-specific splicing regulations tightly connected to cell fates and differentiation that might involve chromatin remodeling through histone modifications. This association, exemplified by histone-associated exon selection in the FGFR2 and LMNB1 genes, necessitates and reinforces further studies on the roles of splicing factors in connecting key players of the chromatin-splicing adaptor system. We coined the term "epispliced genes" to refer to genes with exon selection that is associated with histone modifications in this research work and used it to highlight the importance of epigenetic regulation in AS.

As illustrated by epispliced genes in Project 1, AS is condition-specific and may result in distinct transcriptomes which guide the cells into starkly different developmental trajectories. In Project 2, we managed to detect a new STIM2 variant, namely STIM2.3, in post-mortem brain tissues that plays a key role in synapse formation, neuronal SOCE regulation and homeostatic synaptic transmission. As this novel variant arose more recently along the evolution of Hominoids and *Theropithecus* compared to other STIM1/STIM2 variants, our findings hint toward the potential selective advantage for the development of more complex brains with STIM2.3 as a gain-of-function variant. Furthermore, we found the connection between mis-splicing of STIM2.3 and deregulated calcium signaling

in neurons of Huntington's disease patients, which urges more research on the pathological relevance of STIM2.3 variant, especially in neuronal degenerative diseases.

RBPs are crucial components and regulators in splicing processes and thus, are heavily involved in shaping the transcriptomes. In Project 3, we focused on the RBP IGF2BP2 and provided insights into its binding patterns, mRNA targets and functional roles in mice hepatocytes. The detection of IMPBP2's binding sites was enabled by HyperTRIBE, a novel protocol applied for the first time *in vivo* in mice. Data analysis of HyperTRIBE experiment shows IGF2BP2's stabilizing effects on its target-mRNAs, as reflected by overall reduced expression changes of target genes and the anticorrelating expressions between IGF2BP2 and its targets. Moreover, the enrichment of the m6A consensus motif among binding sites may imply IGF2BP2's potential role in post-transcriptional regulation. Finally, we unveiled the regulatory roles of IGF2BP2 in autophagy and apoptosis by functional analysis of its targets and deregulated genes, as well as by an experiment showing increased LC3-II levels upon IGF2BP2 knockdown. Collectively, these findings not only expand our understandings of IGF2BP2's regulatory mechanisms, but also establish HyperTRIBE as a reliable protocol for investigating mRNA-protein interactions *in vivo*.

AS causes systemic changes in the transcriptomes, which is enabled by regulatory factors like RBPs or SFs, and in turn leads to rewiring of the proteomes. Thus in Projects 4, 5, and 6, we focused on studying the contextualized proteomic changes in response to transcriptomic alterations. To this end, we reviewed in Project 4 the state-of-the-art databases, tools and webservices for protein network analysis; Thereby put emphasis on the lack of tools or workflows for effortless contextualized PPI analysis with user-define data, statistical evidence and domain-leveled explanation. This observation motivated us to develop two new webservers, PPIXpress and PPICompare, to streamline the construction and comparison of condition-specific interactomes from transcriptomic data in Project 5.

In the final Project 6, we attempted to study the gaps between alternatively spliced transcripts and rewired RBPs interactions. Using the implemented webserver-dual PPIXpress and PPICompare (Project 5), I constructed and compared the sample-specific RBP networks based on transcriptomic data of 19 human cell and tissue types in the epispliced genes analysis (Project 1). The resulting differential RBPNs reveal distinct rewiring patterns, which largely mirror the similarities between cells and tissues that were observed for epispliced genes in Project 1. Meanwhile, the binding dynamics and co-occurrence of rewired RBPs suggest the presence of separate RBP hubs which mainly involve in transcription and pre-mRNA processing, or are more related to splicing and cytoplasmic mRNA remodeling as illustrated by differential RBPNs of CD4 and CD8 cells. In short, by inspecting the rewired RBP-interactions across various samples, we found tissue-specific changes in the RBPNs that may affect or result from AS and might even be associate with histone modifications to a certain extent. Due to a time limit and

the intricate nature of splicing regulation, this study leaves open questions about the mechanism of action, specific links to epigenetic features and other functional relevance of these rewired RBPs for future studies.

Chapter A.

Supplementary Data for Chapter I

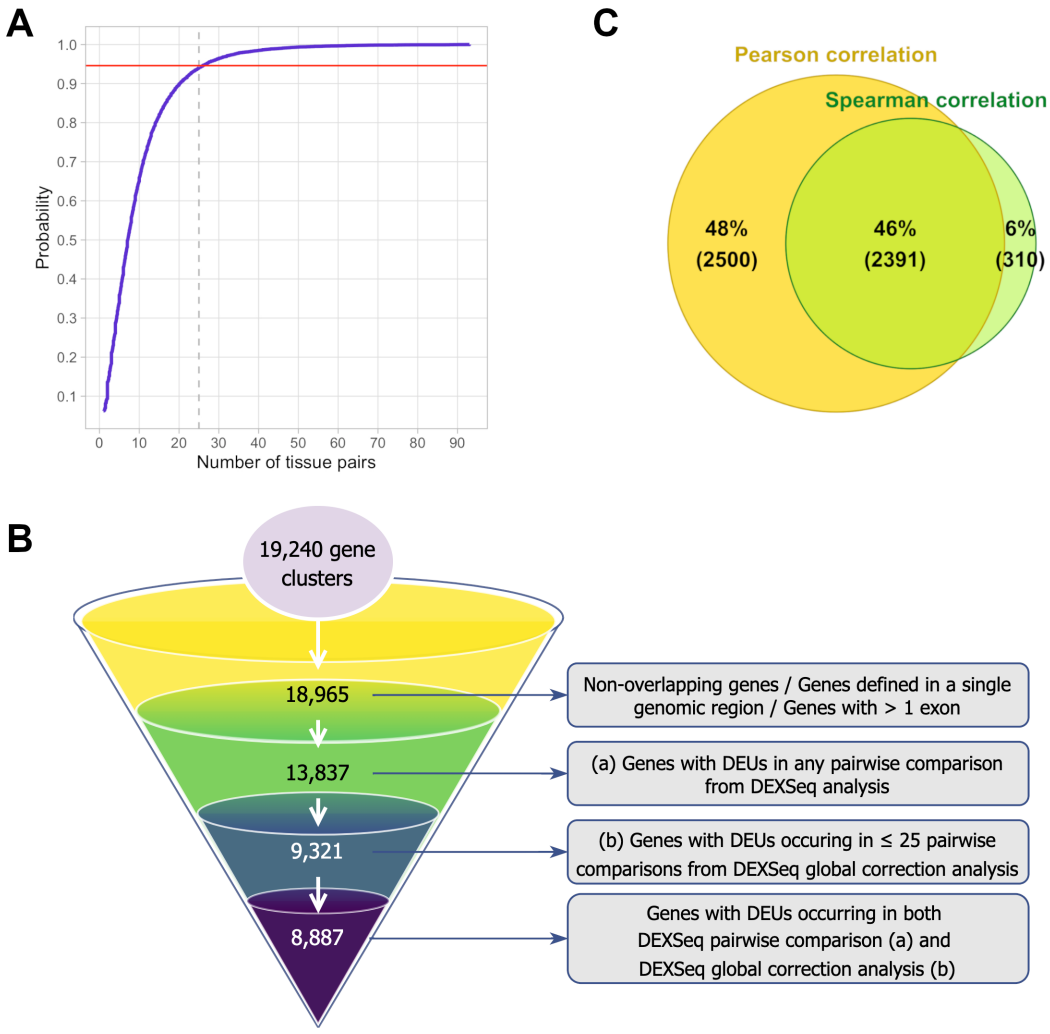

Figure A.1.: Preliminary results for defining the analysis method. (A) Cumulative distribution for the number of tissue pairs in which a differentially used exon is identified. Approximately 95% of the differentially used exons were detected in fewer than 25 pairwise comparisons across 19 tissues (171 pairs in total). These non-ubiquitous, differentially used exons belong to 9,321 genes that were considered for the identification of epispliced genes. (B) Diagram demonstrating the genes selection process for the correlation analysis between differential exon usage (DEU) and differential histone modification (DHM). From the initial 19,240 flattened gene clusters, 275 clusters of genes partially overlapping with each other, spanning more than one genomic region or containing only a single exon were first excluded. The remaining genes were used for pairwise DEU analysis with DEXSeq, resulting in the set of 13,837 genes containing at least one DEU event in any pairwise comparison (a). Using only 9,321 genes with non-ubiquitous DEUs detected in (A), a global DEXSeq analysis was performed to compare all samples against each other simultaneously for multiple testings correction purpose (b). The set of genes with DEUs occuring in both the separate pairwise comparison (a) and the global correction (b) consists of 8,887 genes and was subjected to the main correlation analysis of the study. (C) Venn diagram showing the overlap between the sets of epispliced genes from the analysis with Pearson correlation and Spearman rank correlation.

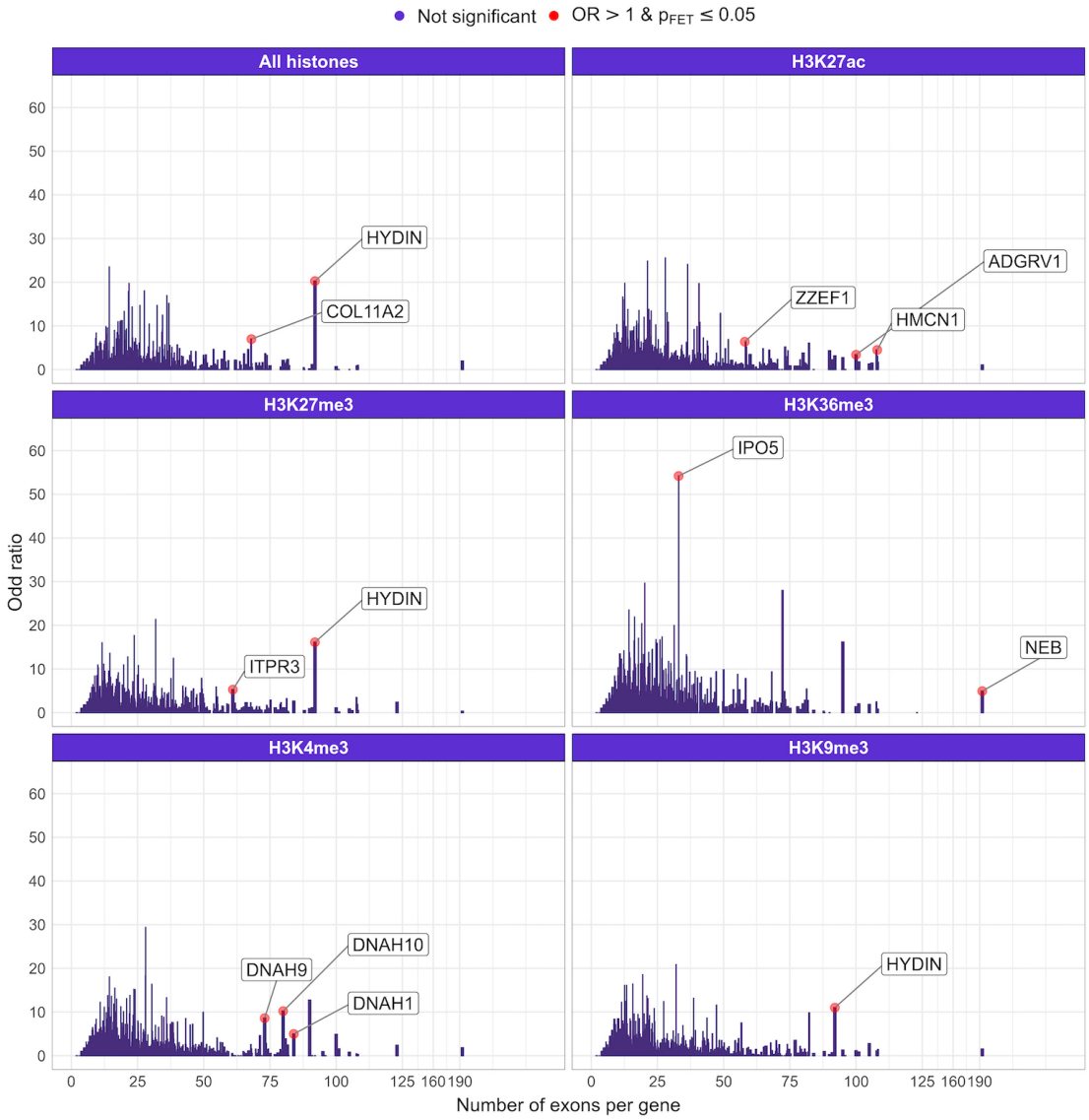

Figure A.2.: Genewise odds ratio with respect to number of exons. The genes with $OR > 1$ and FDR-adjusted $p - value \leq 0.05$ from Fisher Exact Test are highlighted in red. The odds ratios distribution with respect to the number of exons are shown for 5 histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me3 and H3K9me3, as well as for any detected differential histone peaks regardless of histone modification type (All histones).

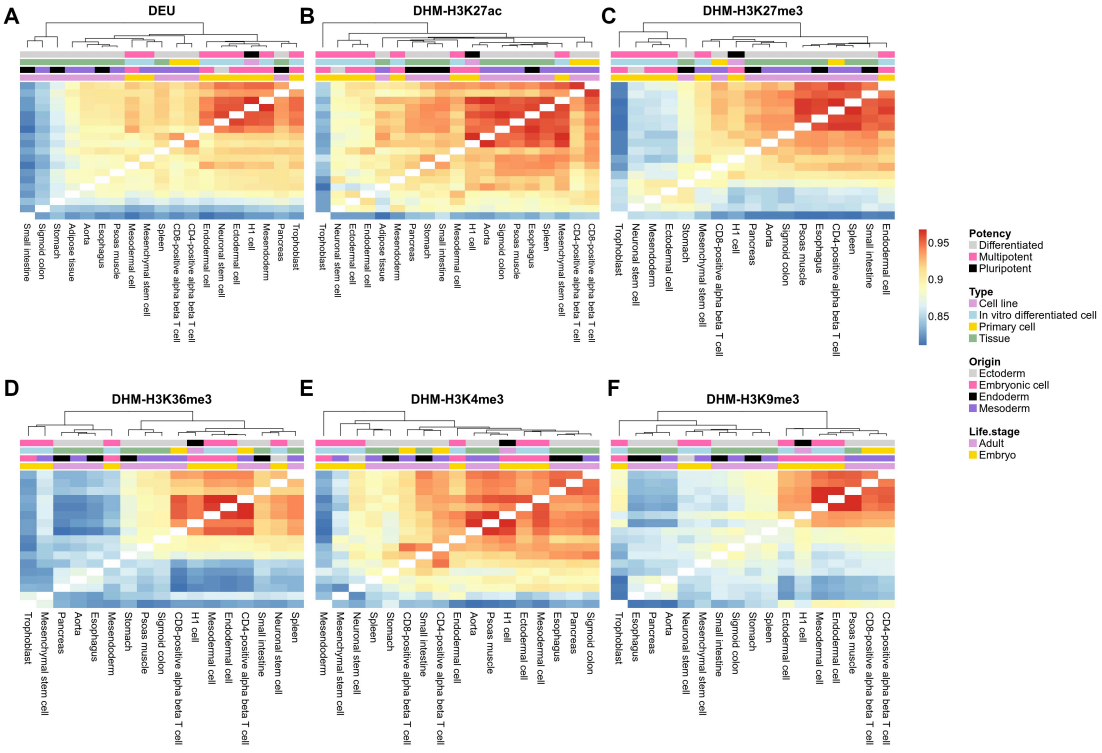

Figure A.3.: Heatmaps representing hierarchical clustering of studied tissues based on genes with differential exon usage (DEU) or differential histone modification (DHM). The pairwise distance across a total of 19 tissues were computed using the set of genes with differential features identified from DEXSeq DEU analysis (A) and MAnorm DHM analysis for H3K27ac, H3K27me3, H3K36me, H3K4me3 and H3K9me3 (B-F). For each differential feature, the tissue distance was measured by Jaccard index, defined as the ratio between the number of mutually affected genes and the total number of affected genes in the union sets of two cell types. All heatmaps use the same color scale ranging from 0.86 to the highest Jaccard index across all tissue pairs and for different differential features. Investigated epigenomes were annotated on the top by their differentiation potency, type of sample, germ layer origin and the life stage when their samples were taken.

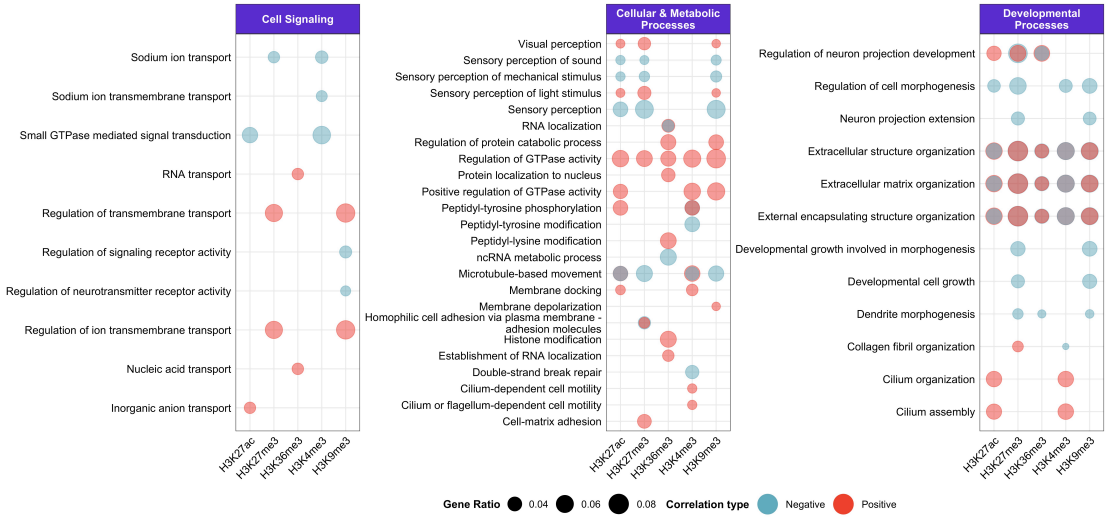

Figure A.4.: Gene ontology (GO) enrichment analysis for biological functions of non-ubiquitous epispliced genes for each histone type. The top enriched GO terms (FDR-adjusted $p - value \leq 0.05$) annotated to epispliced genes that were correlated either with H3K27ac, H3K27me3, H3K36me, H3K4me3 or with H3K9me3 differential histone modifications (DHM) were sorted in decreasing order of significance and of mutual functions between the histone marks. The terms annotated to the epispliced genes with positive correlation between differential exon usage (DEU) are represented by red and with negative correlation by blue. All GO terms are grouped into three main categories, namely cell signaling (A), cellular and metabolic processes (B) and developmental processes (C). In the enrichment analysis, the respective epispliced gene sets were compared against the background set of all human genes having either DEU or DHM events at the exon boundaries.

Epigenome	Potency	Type	Origin	Life stage
Adipose tissue	Differentiated	Tissue	Mesoderm	Adult
Aorta	Differentiated	Tissue	Mesoderm	Adult
CD4-positive alpha beta T cell	Differentiated	Primary cell	Mesoderm	Adult
CD8-positive alpha beta T cell	Differentiated	Primary cell	Mesoderm	Adult
Ectodermal cell	Multipotent	In vitro differentiated cell	Embryonic cell	Embryo
Endodermal cell	Multipotent	In vitro differentiated cell	Embryonic cell	Embryo
Esophagus	Differentiated	Tissue	Endoderm	Adult
H1 cell	Pluripotent	Cell line	Embryonic cell	Embryo
Mesenchymal stem cell	Multipotent	In vitro differentiated cell	Mesoderm	Embryo
Mesendoderm	Multipotent	In vitro differentiated cell	Embryonic cell	Embryo
Mesodermal cell	Multipotent	In vitro differentiated cell	Embryonic cell	Embryo
Neuronalstem cell	Multipotent	In vitro differentiated cell	Ectoderm	Embryo
Pancreas	Differentiated	Tissue	Endoderm	Adult
Psoas muscle	Differentiated	Tissue	Mesoderm	Adult
Sigmoid colon	Differentiated	Tissue	Mesoderm	Adult
Small intestine	Differentiated	Tissue	Endoderm	Adult
Spleen	Differentiated	Tissue	Mesoderm	Adult
Stomach	Differentiated	Tissue	Endoderm	Adult
Trophoblast cell	Multipotent	In vitro differentiated cell	Embryonic cell	Embryo

Table A.1.: List of tissues and cell types retrieved from the Human Epigenome Atlas with annotated potency, sample type, origin and life stage.

Biosample term name	File accession	Experiment accession	File format	Output type	Biosample type	Biological replicate(s)	Technical replicate(s)
adipose tissue	ENCFF717MIN	ENCSR686JJB	bam	alignments	tissue	1	1_1
adipose tissue	ENCFF491UPQ	ENCSR741QEH	bam	alignments	tissue	1	1_2
aorta	ENCFF081DYZ	ENCSR763NOO	bam	alignments	tissue	1	1_2
aorta	ENCFF864QLZ	ENCSR995BHD	bam	alignments	tissue	1	1_1
CD4-positive, alpha-beta T cell	ENCFF597BAH	ENCSR545MEZ	bam	alignments	primary cell	1	1_1
CD4-positive, alpha-beta T cell	ENCFF208TSZ	ENCSR463JBR	bam	alignments	primary cell	1	1_1
CD8-positive, alpha-beta T cell	ENCFF746WLK	ENCSR944FLL	bam	alignments	primary cell	1	1_1
CD8-positive, alpha-beta T cell	ENCFF716ABE	ENCSR861QKF	bam	alignments	primary cell	1	1_1
ectodermal cell	ENCFF985MTW	ENCSR872LTT	bam	alignments	in vitro differentiated cells	1	1_1
ectodermal cell	ENCFF148VCG	ENCSR872LTT	bam	alignments	in vitro differentiated cells	1	1_2
ectodermal cell	ENCFF549FQD	ENCSR382XJF	bam	alignments	in vitro differentiated cells	1	1_3
ectodermal cell	ENCFF370OWM	ENCSR851BRK	bam	alignments	in vitro differentiated cells	1	1_3
ectodermal cell	ENCFF686CNC	ENCSR593AMV	bam	alignments	in vitro differentiated cells	1	1_1
ectodermal cell	ENCFF154POQ	ENCSR593AMV	bam	alignments	in vitro differentiated cells	1	1_2
endodermal cell	ENCFF489LAR	ENCSR002CTR	bam	alignments	in vitro differentiated cells	1	1_3
endodermal cell	ENCFF093GMX	ENCSR472PBS	bam	alignments	in vitro differentiated cells	1	1_3
esophagus	ENCFF441FVJ	ENCSR993QGR	bam	alignments	tissue	1	1_1
esophagus	ENCFF251YUZ	ENCSR102TQN	bam	alignments	tissue	1	1_2
H1	ENCFF740OFO	ENCSR043RSE	bam	alignments	cell line	1	1_1
H1	ENCFF365DFR	ENCSR670WQY	bam	alignments	cell line	1	1_1
mesenchymal stem cell	ENCFF011JJG	ENCSR663WGC	bam	alignments	in vitro differentiated cells	1	1_1
mesenchymal stem cell	ENCFF317PIS	ENCSR275ZLF	bam	alignments	in vitro differentiated cells	1	1_1
mesendoderm	ENCFF738KXI	ENCSR700BEW	bam	alignments	in vitro differentiated cells	1	1_1
mesendoderm	ENCFF028PIJ	ENCSR976JGI	bam	alignments	in vitro differentiated cells	1	1_1
mesodermal cell	ENCFF107QNN	ENCSR500UOD	bam	alignments	in vitro differentiated cells	1	1_3
mesodermal cell	ENCFF091OZL	ENCSR433GXV	bam	alignments	in vitro differentiated cells	1	1_3
neuronal stem cell	ENCFF359YOQ	ENCSR977XUX	bam	alignments	in vitro differentiated cells	1	1_1
neuronal stem cell	ENCFF905ZXT	ENCSR572EET	bam	alignments	in vitro differentiated cells	1	1_1
pancreas	ENCFF199EFU	ENCSR629VMZ	bam	alignments	tissue	1	1_1
pancreas	ENCFF085TWC	ENCSR571BML	bam	alignments	tissue	1	1_1
psoas muscle	ENCFF677DKS	ENCSR502OTI	bam	alignments	tissue	1	1_1
psoas muscle	ENCFF738TTD	ENCSR843HXR	bam	alignments	tissue	1	1_1
sigmoid colon	ENCFF588TQX	ENCSR825GWD	bam	alignments	tissue	1	1_12
sigmoid colon	ENCFF050JVY	ENCSR999ZCI	bam	alignments	tissue	1	1_1
small intestine	ENCFF584GAJ	ENCSR039ICU	bam	alignments	tissue	1	1_1
small intestine	ENCFF894PUN	ENCSR719HRO	bam	alignments	tissue	1	1_1
spleen	ENCFF717MVQ	ENCSR510PSL	bam	alignments	tissue	1	1_2
spleen	ENCFF918SPI	ENCSR910QOX	bam	alignments	tissue	1	1_2
stomach	ENCFF268XDH	ENCSR721HDG	bam	alignments	tissue	1	1_1
stomach	ENCFF056CIU	ENCSR980UEY	bam	alignments	tissue	1	1_1
trophoblast cell	ENCFF137MXA	ENCSR762CJN	bam	alignments	in vitro differentiated cells	1	1_1

Biosample term name	File accession	Experiment accession	File format	Output type	Biosample type	Biological replicate(s)	Technical replicate(s)
trophoblast cell	ENCFF677DAD	ENCSR762CJN	bam	alignments	in vitro differentiated cells	2	2_1

Table A.2.: Metadata for the retrieved Human Epigenome Atlas poly-A plus RNA-seq. For each biosample, the associated information including the file accession number, experiment accession number, file format, sample types and replicates are listed. All files used the Human Reference Genome GRCh38 for assembly.

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
adipose tissue	ENCFF861HMY	ENCSR082SHT	bam	alignments	H3K27ac-human	1	1_1
aorta	ENCFF434DCE	ENCSR322TJD	bam	alignments	H3K27ac-human	1	1_2
aorta	ENCFF265HDL	ENCSR519CFV	bam	alignments	H3K27ac-human	2	2_1
CD4-positive, alpha-beta T cell	ENCFF082MIE	ENCSR561KOM	bam	alignments	H3K27ac-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF449YKU	ENCSR007HLH	bam	alignments	H3K27ac-human	1	1_2
ectodermal cell	ENCFF728FKI	ENCSR747HAM	bam	alignments	H3K27ac-human	1	1_2
ectodermal cell	ENCFF537TER	ENCSR747HAM	bam	alignments	H3K27ac-human	2	2_2
endodermal cell	ENCFF898KBV	ENCSR200ETW	bam	alignments	H3K27ac-human	1	1_2
endodermal cell	ENCFF668YUL	ENCSR200ETW	bam	alignments	H3K27ac-human	2	2_2
esophagus	ENCFF303TFY	ENCSR679OVD	bam	alignments	H3K27ac-human	1	1_1
esophagus	ENCFF166HVM	ENCSR645SYH	bam	alignments	H3K27ac-human	1	1_1
H1	ENCFF948UXT	ENCSR880SUY	bam	alignments	H3K27ac-human	2	2_1
H1	ENCFF663SAM	ENCSR880SUY	bam	alignments	H3K27ac-human	1	1_2
mesenchymal stem cell	ENCFF262VKZ	ENCSR013KEC	bam	alignments	H3K27ac-human	1	1_1
mesenchymal stem cell	ENCFF615PUT	ENCSR013KEC	bam	alignments	H3K27ac-human	2	2_1
mesendoderm	ENCFF710LJW	ENCSR473PNT	bam	alignments	H3K27ac-human	1	1_1
mesendoderm	ENCFF535JGD	ENCSR473PNT	bam	alignments	H3K27ac-human	2	2_1
mesodermal cell	ENCFF114JZY	ENCSR931WLE	bam	alignments	H3K27ac-human	1	1_2
mesodermal cell	ENCFF175GQI	ENCSR931WLE	bam	alignments	H3K27ac-human	2	2_2
neuronal stem cell	ENCFF063LDJ	ENCSR799SRL	bam	alignments	H3K27ac-human	1	1_1
neuronal stem cell	ENCFF402MPW	ENCSR799SRL	bam	alignments	H3K27ac-human	2	2_1
neuronal stem cell	ENCFF263UOB	ENCSR799SRL	bam	alignments	H3K27ac-human	3	3_1
pancreas	ENCFF015LEU	ENCSR402HFW	bam	alignments	H3K27ac-human	1	1_1
pancreas	ENCFF516INT	ENCSR612BWE	bam	alignments	H3K27ac-human	1	1_1
psoas muscle	ENCFF295IFC	ENCSR250NHD	bam	alignments	H3K27ac-human	1	1_1
psoas muscle	ENCFF755AFZ	ENCSR367WYJ	bam	alignments	H3K27ac-human	1	1_1, 1_2
psoas muscle	ENCFF305JAV	ENCSR791ISZ	bam	alignments	H3K27ac-human	2	2_1
sigmoid colon	ENCFF118ACN	ENCSR213SMK	bam	alignments	H3K27ac-human	1	1_1
sigmoid colon	ENCFF342MZW	ENCSR561YSH	bam	alignments	H3K27ac-human	1	1_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
small intestine	ENCFF429QHU	ENCSR543CPW	bam	alignments	H3K27ac-human	1	1_3
small intestine	ENCFF410WCE	ENCSR655XLM	bam	alignments	H3K27ac-human	1	1_1
small intestine	ENCFF915JKN	ENCSR454VRA	bam	alignments	H3K27ac-human	1	1_1
small intestine	ENCFF466VBP	ENCSR892XFG	bam	alignments	H3K27ac-human	1	1_2
spleen	ENCFF825ZUF	ENCSR170MAJ	bam	alignments	H3K27ac-human	1	1_1
spleen	ENCFF628GTH	ENCSR170MAJ	bam	alignments	H3K27ac-human	1	1_1
spleen	ENCFF850VCJ	ENCSR235ZBF	bam	alignments	H3K27ac-human	1	1_2
spleen	ENCFF500MVV	ENCSR235ZBF	bam	alignments	H3K27ac-human	1	1_2
spleen	ENCFF062ZGG	ENCSR086XCT	bam	alignments	H3K27ac-human	1	1_1
spleen	ENCFF626HSF	ENCSR086XCT	bam	alignments	H3K27ac-human	1	1_1
stomach	ENCFF735KUK	ENCSR001SHB	bam	alignments	H3K27ac-human	1	1_2
stomach	ENCFF972FYM	ENCSR437QMD	bam	alignments	H3K27ac-human	1	1_1, 1_2
stomach	ENCFF081CDJ	ENCSR743DDX	bam	alignments	H3K27ac-human	1	1_2
stomach	ENCFF805CCB	ENCSR582UTE	bam	alignments	H3K27ac-human	1	1_1
trophoblast cell	ENCFF233ZTK	ENCSR425PQI	bam	alignments	H3K27ac-human	1	1_1
trophoblast cell	ENCFF570NQD	ENCSR425PQI	bam	alignments	H3K27ac-human	2	2_1
trophoblast cell	ENCFF625RQZ	ENCSR425PQI	bam	alignments	H3K27ac-human	2	2_1
aorta	ENCFF127BNG	ENCSR196PGM	bam	alignments	H3K27me3-human	1	1_1
aorta	ENCFF654BWT	ENCSR128VHV	bam	alignments	H3K27me3-human	1	1_2
CD4-positive, alpha-beta T cell	ENCFF971YVS	ENCSR043SBG	bam	alignments	H3K27me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF624YMM	ENCSR733QOZ	bam	alignments	H3K27me3-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF783NPA	ENCSR103GGR	bam	alignments	H3K27me3-human	1	1_4
ectodermal cell	ENCFF761AOT	ENCSR690GLT	bam	alignments	H3K27me3-human	1	1_2
endodermal cell	ENCFF465YTB	ENCSR273IYV	bam	alignments	H3K27me3-human	2	2_5
endodermal cell	ENCFF249NJY	ENCSR273IYV	bam	alignments	H3K27me3-human	1	1_5
esophagus	ENCFF096RVZ	ENCSR641RQV	bam	alignments	H3K27me3-human	1	1_2
esophagus	ENCFF703GII	ENCSR088GXB	bam	alignments	H3K27me3-human	1	1_1
H1	ENCFF748KOZ	ENCSR186OBR	bam	alignments	H3K27me3-human	1	1_1
H1	ENCFF382GPJ	ENCSR186OBR	bam	alignments	H3K27me3-human	2	2_1
H1	ENCFF596SHE	ENCSR928HYM	bam	alignments	H3K27me3-human	1	1_2
H1	ENCFF830PVE	ENCSR928HYM	bam	alignments	H3K27me3-human	2	2_1
H1	ENCFF310SBN	ENCSR216OGD	bam	alignments	H3K27me3-human	2	2_1
H1	ENCFF083QQZ	ENCSR216OGD	bam	alignments	H3K27me3-human	1	1_1
mesenchymal stem cell	ENCFF356HLU	ENCSR832JVP	bam	alignments	H3K27me3-human	1	1_1
mesenchymal stem cell	ENCFF455IUR	ENCSR832JVP	bam	alignments	H3K27me3-human	2	2_1
mesendoderm	ENCFF803MMH	ENCSR405AXO	bam	alignments	H3K27me3-human	1	1_1
mesendoderm	ENCFF244AUB	ENCSR405AXO	bam	alignments	H3K27me3-human	2	2_1
neuronal stem cell	ENCFF230ZHL	ENCSR550XZG	bam	alignments	H3K27me3-human	2	2_1
neuronal stem cell	ENCFF101MMK	ENCSR550XZG	bam	alignments	H3K27me3-human	1	1_1
neuronal stem cell	ENCFF194JFG	ENCSR692CTK	bam	alignments	H3K27me3-human	1	1_1
neuronal stem cell	ENCFF055OMW	ENCSR692CTK	bam	alignments	H3K27me3-human	2	2_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
neuronal stem cell	ENCFF781FVD	ENCSR694LBI	bam	alignments	H3K27me3-human	1	1_1
neuronal stem cell	ENCFF591ZEN	ENCSR694LBI	bam	alignments	H3K27me3-human	2	2_1
pancreas	ENCFF530LYL	ENCSR186QKH	bam	alignments	H3K27me3-human	1	1_1
pancreas	ENCFF987RZA	ENCSR486NDF	bam	alignments	H3K27me3-human	1	1_1, 1_2
psoas muscle	ENCFF096XQY	ENCSR720SAS	bam	alignments	H3K27me3-human	1	1_1
psoas muscle	ENCFF539ZGZ	ENCSR843KHS	bam	alignments	H3K27me3-human	1	1_2
sigmoid colon	ENCFF124DET	ENCSR042RIW	bam	alignments	H3K27me3-human	1	1_1
sigmoid colon	ENCFF607WIJ	ENCSR897TGR	bam	alignments	H3K27me3-human	1	1_1
small intestine	ENCFF557ZVM	ENCSR877PAS	bam	alignments	H3K27me3-human	1	1_2
small intestine	ENCFF816GHG	ENCSR340OPI	bam	alignments	H3K27me3-human	1	1_1
small intestine	ENCFF651ZOQ	ENCSR859EIX	bam	alignments	H3K27me3-human	1	1_2
spleen	ENCFF558UBX	ENCSR608FDQ	bam	alignments	H3K27me3-human	1	1_2
spleen	ENCFF061EVZ	ENCSR608FDQ	bam	alignments	H3K27me3-human	1	1_2
spleen	ENCFF708NMK	ENCSR408ONP	bam	alignments	H3K27me3-human	1	1_1
spleen	ENCFF191LZK	ENCSR408ONP	bam	alignments	H3K27me3-human	1	1_1
stomach	ENCFF192EMA	ENCSR527BFF	bam	alignments	H3K27me3-human	1	1_1
stomach	ENCFF582NZD	ENCSR354IST	bam	alignments	H3K27me3-human	1	1_2
trophoblast cell	ENCFF904JLV	ENCSR960CWQ	bam	alignments	H3K27me3-human	1	1_1
trophoblast cell	ENCFF303YNU	ENCSR960CWQ	bam	alignments	H3K27me3-human	2	2_1
trophoblast cell	ENCFF482DKF	ENCSR960CWQ	bam	alignments	H3K27me3-human	2	2_1
aorta	ENCFF168IPJ	ENCSR673JYT	bam	alignments	H3K36me3-human	1	1_1
aorta	ENCFF230YGV	ENCSR989AMI	bam	alignments	H3K36me3-human	1	1_1
CD4-positive, alpha-beta T cell	ENCFF300XBB	ENCSR828WZG	bam	alignments	H3K36me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF151UEE	ENCSR774OKQ	bam	alignments	H3K36me3-human	1	1_3
CD8-positive, alpha-beta T cell	ENCFF814JIZ	ENCSR782NOO	bam	alignments	H3K36me3-human	1	1_3
endodermal cell	ENCFF778DVB	ENCSR677EZB	bam	alignments	H3K36me3-human	1	1_4
esophagus	ENCFF168XVF	ENCSR279MCN	bam	alignments	H3K36me3-human	1	1_1
esophagus	ENCFF192PWM	ENCSR034ZHF	bam	alignments	H3K36me3-human	1	1_1
H1	ENCFF805WZT	ENCSR496DCY	bam	alignments	H3K36me3-human	2	2_2
H1	ENCFF697AQU	ENCSR496DCY	bam	alignments	H3K36me3-human	1	1_1
H1	ENCFF295LHK	ENCSR476KTK	bam	alignments	H3K36me3-human	1	1_1
H1	ENCFF603HTP	ENCSR476KTK	bam	alignments	H3K36me3-human	2	2_1
H1	ENCFF619JFN	ENCSR925LJZ	bam	alignments	H3K36me3-human	1	1_1
H1	ENCFF044YAN	ENCSR925LJZ	bam	alignments	H3K36me3-human	2	2_1
mesenchymal stem cell	ENCFF412PPW	ENCSR824UNY	bam	alignments	H3K36me3-human	1	1_1
mesenchymal stem cell	ENCFF651ARP	ENCSR824UNY	bam	alignments	H3K36me3-human	2	2_1
mesendoderm	ENCFF077SHE	ENCSR144RXL	bam	alignments	H3K36me3-human	1	1_1
mesendoderm	ENCFF682KFP	ENCSR144RXL	bam	alignments	H3K36me3-human	2	2_1
mesodermal cell	ENCFF592HWS	ENCSR100LWU	bam	alignments	H3K36me3-human	1	1_4
mesodermal cell	ENCFF335KOW	ENCSR100LWU	bam	alignments	H3K36me3-human	2	2_2
neuronal stem cell	ENCFF095MCZ	ENCSR256ESY	bam	alignments	H3K36me3-human	1	1_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
neuronal stem cell	ENCFF964FXG	ENCSR256ESY	bam	alignments	H3K36me3-human	2	2_1
neuronal stem cell	ENCFF726UFJ	ENCSR238WMO	bam	alignments	H3K36me3-human	2	2_1
neuronal stem cell	ENCFF226NJT	ENCSR238WMO	bam	alignments	H3K36me3-human	1	1_1
pancreas	ENCFF693EDE	ENCSR943JOF	bam	alignments	H3K36me3-human	1	1_1
pancreas	ENCFF075MGN	ENCSR393HBQ	bam	alignments	H3K36me3-human	1	1_1
psoas muscle	ENCFF664YVG	ENCSR277PDE	bam	alignments	H3K36me3-human	1	1_2
sigmoid colon	ENCFF296GNA	ENCSR445RFF	bam	alignments	H3K36me3-human	1	1_1
sigmoid colon	ENCFF501JBW	ENCSR751JOQ	bam	alignments	H3K36me3-human	1	1_1
small intestine	ENCFF633TEH	ENCSR073YZL	bam	alignments	H3K36me3-human	1	1_1
small intestine	ENCFF674FLQ	ENCSR958DEW	bam	alignments	H3K36me3-human	1	1_1
small intestine	ENCFF739XVS	ENCSR205NEW	bam	alignments	H3K36me3-human	1	1_4
spleen	ENCFF011YZH	ENCSR466DUB	bam	alignments	H3K36me3-human	1	1_1
spleen	ENCFF232GTM	ENCSR466DUB	bam	alignments	H3K36me3-human	1	1_1
spleen	ENCFF361ULW	ENCSR078BHK	bam	alignments	H3K36me3-human	1	1_1
spleen	ENCFF259KHT	ENCSR078BHK	bam	alignments	H3K36me3-human	1	1_1
stomach	ENCFF814EYF	ENCSR697YSL	bam	alignments	H3K36me3-human	1	1_2
stomach	ENCFF749IJO	ENCSR269GMC	bam	alignments	H3K36me3-human	1	1_1
stomach	ENCFF731WJY	ENCSR552MZH	bam	alignments	H3K36me3-human	1	1_1
trophoblast cell	ENCFF079SNV	ENCSR038OIN	bam	alignments	H3K36me3-human	1	1_1
trophoblast cell	ENCFF763USC	ENCSR038OIN	bam	alignments	H3K36me3-human	2	2_1
trophoblast cell	ENCFF294APS	ENCSR005YZH	bam	alignments	H3K36me3-human	1	1_1
aorta	ENCFF176CXM	ENCSR957BPJ	bam	alignments	H3K4me3-human	1	1_1
aorta	ENCFF817YZO	ENCSR960EVO	bam	alignments	H3K4me3-human	1	1_2
CD4-positive, alpha-beta T cell	ENCFF681JNH	ENCSR263WLD	bam	alignments	H3K4me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF962FVV	ENCSR852FRR	bam	alignments	H3K4me3-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF470ZIH	ENCSR796CSH	bam	alignments	H3K4me3-human	1	1_1
endodermal cell	ENCFF347EDU	ENCSR446ZCY	bam	alignments	H3K4me3-human	1	1_5
endodermal cell	ENCFF937ZPT	ENCSR446ZCY	bam	alignments	H3K4me3-human	2	2_5
esophagus	ENCFF894DUH	ENCSR697GPO	bam	alignments	H3K4me3-human	1	1_2
esophagus	ENCFF585DLK	ENCSR577ILY	bam	alignments	H3K4me3-human	1	1_1
H1	ENCFF467XCU	ENCSR019SQX	bam	alignments	H3K4me3-human	1	1_1
H1	ENCFF494FNC	ENCSR019SQX	bam	alignments	H3K4me3-human	2	2_1
H1	ENCFF640RPS	ENCSR019SQX	bam	alignments	H3K4me3-human	3	3_1
mesenchymal stem cell	ENCFF071ETA	ENCSR501JET	bam	alignments	H3K4me3-human	1	1_1
mesenchymal stem cell	ENCFF707CSG	ENCSR501JET	bam	alignments	H3K4me3-human	2	2_1
mesendoderm	ENCFF948LKM	ENCSR441SAT	bam	alignments	H3K4me3-human	1	1_2
mesendoderm	ENCFF072FDU	ENCSR441SAT	bam	alignments	H3K4me3-human	2	2_2
neuronal stem cell	ENCFF396QPN	ENCSR354XWM	bam	alignments	H3K4me3-human	1	1_2
neuronal stem cell	ENCFF490QCW	ENCSR354XWM	bam	alignments	H3K4me3-human	2	2_1
neuronal stem cell	ENCFF943BCG	ENCSR956CTX	bam	alignments	H3K4me3-human	1	1_1
neuronal stem cell	ENCFF102CQQ	ENCSR956CTX	bam	alignments	H3K4me3-human	2	2_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
pancreas	ENCFF056MDM	ENCSR747VED	bam	alignments	H3K4me3-human	1	1_1
pancreas	ENCFF907SLS	ENCSR315LPR	bam	alignments	H3K4me3-human	1	1_3
psoas muscle	ENCFF215MSA	ENCSR245BEV	bam	alignments	H3K4me3-human	1	1_1
psoas muscle	ENCFF959CCK	ENCSR949OYZ	bam	alignments	H3K4me3-human	1	1_2
sigmoid colon	ENCFF219BET	ENCSR321SZE	bam	alignments	H3K4me3-human	1	1_1
sigmoid colon	ENCFF917ZBE	ENCSR421HUB	bam	alignments	H3K4me3-human	1	1_2
small intestine	ENCFF883KQO	ENCSR237QFJ	bam	alignments	H3K4me3-human	1	1_2
small intestine	ENCFF035OFJ	ENCSR944QSH	bam	alignments	H3K4me3-human	1	1_1
small intestine	ENCFF070DOR	ENCSR792IJA	bam	alignments	H3K4me3-human	1	1_1
spleen	ENCFF346MPS	ENCSR432KIH	bam	alignments	H3K4me3-human	1	1_1
spleen	ENCFF266SHE	ENCSR432KIH	bam	alignments	H3K4me3-human	1	1_1
spleen	ENCFF205USG	ENCSR448FZC	bam	alignments	H3K4me3-human	1	1_1
spleen	ENCFF566BTS	ENCSR448FZC	bam	alignments	H3K4me3-human	1	1_1
stomach	ENCFF630XTQ	ENCSR202RXT	bam	alignments	H3K4me3-human	1	1_2
stomach	ENCFF937AGY	ENCSR129NCV	bam	alignments	H3K4me3-human	1	1_1
trophoblast cell	ENCFF638UNA	ENCSR874WOB	bam	alignments	H3K4me3-human	1	1_1
trophoblast cell	ENCFF961MFR	ENCSR874WOB	bam	alignments	H3K4me3-human	2	2_1
aorta	ENCFF371ZWQ	ENCSR065ZNA	bam	alignments	H3K9me3-human	1	1_1
CD4-positive, alpha-beta T cell	ENCFF616YFF	ENCSR453GNY	bam	alignments	H3K9me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF877MMM	ENCSR787WLV	bam	alignments	H3K9me3-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF953NDJ	ENCSR824PXG	bam	alignments	H3K9me3-human	1	1_3
ectodermal cell	ENCFF405EWU	ENCSR235CEI	bam	alignments	H3K9me3-human	1	1_2
endodermal cell	ENCFF520XOV	ENCSR823BHO	bam	alignments	H3K9me3-human	2	2_5
endodermal cell	ENCFF184CLR	ENCSR823BHO	bam	alignments	H3K9me3-human	1	1_5
esophagus	ENCFF189GCC	ENCSR150GLE	bam	alignments	H3K9me3-human	1	1_1
esophagus	ENCFF072NBC	ENCSR200WDD	bam	alignments	H3K9me3-human	1	1_1
H1	ENCFF421AGL	ENCSR395USV	bam	alignments	H3K9me3-human	2	2_2
H1	ENCFF008ALX	ENCSR395USV	bam	alignments	H3K9me3-human	1	1_1
H1	ENCFF597CRW	ENCSR883AQJ	bam	alignments	H3K9me3-human	1	1_2
H1	ENCFF354TVY	ENCSR883AQJ	bam	alignments	H3K9me3-human	2	2_1
H1	ENCFF913CWL	ENCSR883AQJ	bam	alignments	H3K9me3-human	3	3_1, 3_2
mesenchymal stem cell	ENCFF009ADE	ENCSR746CUY	bam	alignments	H3K9me3-human	1	1_1
mesenchymal stem cell	ENCFF533ZII	ENCSR746CUY	bam	alignments	H3K9me3-human	2	2_1
mesodermal cell	ENCFF619CYH	ENCSR887ZPC	bam	alignments	H3K9me3-human	2	2_2
mesodermal cell	ENCFF334UGS	ENCSR887ZPC	bam	alignments	H3K9me3-human	3	3_2
neuronal stem cell	ENCFF483TNX	ENCSR391WDE	bam	alignments	H3K9me3-human	1	1_1
neuronal stem cell	ENCFF200TLD	ENCSR391WDE	bam	alignments	H3K9me3-human	2	2_1
neuronal stem cell	ENCFF369SMM	ENCSR800IIW	bam	alignments	H3K9me3-human	1	1_1
neuronal stem cell	ENCFF347GTB	ENCSR800IIW	bam	alignments	H3K9me3-human	2	2_1
pancreas	ENCFF705BUN	ENCSR533HDU	bam	alignments	H3K9me3-human	1	1_1
pancreas	ENCFF506OJR	ENCSR035QNZ	bam	alignments	H3K9me3-human	1	1_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
psoas muscle	ENCFF062DIE	ENCSR394DRL	bam	alignments	H3K9me3-human	1	1_1
sigmoid colon	ENCFF900CUX	ENCSR737NLJ	bam	alignments	H3K9me3-human	1	1_1
sigmoid colon	ENCFF046FUT	ENCSR636IDR	bam	alignments	H3K9me3-human	1	1_1
small intestine	ENCFF959CNM	ENCSR417RFS	bam	alignments	H3K9me3-human	1	1_1
small intestine	ENCFF036JLB	ENCSR773TWR	bam	alignments	H3K9me3-human	1	1_2
small intestine	ENCFF181EGL	ENCSR270VNK	bam	alignments	H3K9me3-human	1	1_1
spleen	ENCFF928TDB	ENCSR249XEB	bam	alignments	H3K9me3-human	1	1_1
spleen	ENCFF858WJB	ENCSR249XEB	bam	alignments	H3K9me3-human	1	1_1
spleen	ENCFF177AJJ	ENCSR421FPV	bam	alignments	H3K9me3-human	1	1_1
spleen	ENCFF694MHC	ENCSR421FPV	bam	alignments	H3K9me3-human	1	1_1
stomach	ENCFF606BDG	ENCSR885CMN	bam	alignments	H3K9me3-human	1	1_1
stomach	ENCFF642FEE	ENCSR639RKZ	bam	alignments	H3K9me3-human	1	1_1
stomach	ENCFF081KPW	ENCSR447GWQ	bam	alignments	H3K9me3-human	1	1_3
trophoblast cell	ENCFF220LTL	ENCSR044WCY	bam	alignments	H3K9me3-human	1	1_1
trophoblast cell	ENCFF932LWO	ENCSR044WCY	bam	alignments	H3K9me3-human	2	2_1
trophoblast cell	ENCFF915XKH	ENCSR044WCY	bam	alignments	H3K9me3-human	2	2_1

Table A.3.: Metadata for the retrieved Human Epigenome Atlas ChIP-seq alignment files. For each biosample, the associated information include the file accession number, experiment accession number, file format and replicates are listed 5 ChIP-seq targets, including H3K27ac, H3K27me3, H3K36me3, H3K4me3 and H3K9me3.

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
adipose tissue	ENCFF154CXY	ENCSR082SHT	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
aorta	ENCFF686UBD	ENCSR322TJD	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
aorta	ENCFF057LJS	ENCSR519CFV	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	2	2_1
CD4-positive, alpha-beta T cell	ENCFF901AMN	ENCSR546SDM	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	2	2_1
CD4-positive, alpha-beta T cell	ENCFF449ITG	ENCSR561KOM	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF321NYQ	ENCSR007HLH	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF243EDS	ENCSR835OJV	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_1, 2_1
ectodermal cell	ENCFF909CTU	ENCSR747HAM	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_2, 2_2
endodermal cell	ENCFF435HDB	ENCSR200ETW	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_2, 2_2
esophagus	ENCFF352GQN	ENCSR645SYH	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
esophagus	ENCFF535PRN	ENCSR679OVD	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
H1	ENCFF045CUG	ENCSR880SUY	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_2, 2_1
mesenchymal stem cell	ENCFF268FCW	ENCSR013KEC	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_1, 2_1
mesendoderm	ENCFF459UTL	ENCSR473PNT	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_1, 2_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
mesodermal cell	ENCFF997UPL	ENCSR931WLE	bed narrowPeak	replicated peaks	H3K27ac-human	1, 2	1_2, 2_2
neuronal stem cell	ENCFF722OBP	ENCSR799SRL	bed narrowPeak	replicated peaks	H3K27ac-human	1, 3	1_1, 3_1
pancreas	ENCFF064KHS	ENCSR402HFW	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
pancreas	ENCFF769XNX	ENCSR612BWE	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
psoas muscle	ENCFF779CYV	ENCSR250NHD	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
psoas muscle	ENCFF110BLF	ENCSR367WYJ	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1, 1_2
psoas muscle	ENCFF265IZO	ENCSR791ISZ	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	2	2_1
sigmoid colon	ENCFF077GUP	ENCSR213SMK	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
sigmoid colon	ENCFF096HVY	ENCSR561YSH	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
small intestine	ENCFF581EXO	ENCSR454VRA	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
small intestine	ENCFF434KGU	ENCSR543CPW	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_3
small intestine	ENCFF484AHL	ENCSR655XLM	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
small intestine	ENCFF458TBT	ENCSR892XFG	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
spleen	ENCFF233IPB	ENCSR086XCT	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
spleen	ENCFF015DHA	ENCSR086XCT	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
spleen	ENCFF158FZE	ENCSR170MAJ	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
spleen	ENCFF273NFQ	ENCSR170MAJ	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
spleen	ENCFF661WLX	ENCSR235ZBF	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
spleen	ENCFF145NKN	ENCSR235ZBF	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
stomach	ENCFF598JDA	ENCSR001SHB	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
stomach	ENCFF055FKA	ENCSR437QMD	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1, 1_2
stomach	ENCFF268OSV	ENCSR582UTE	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
stomach	ENCFF878SVN	ENCSR743DDX	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_2
trophoblast	ENCFF972AZZ	ENCSR507SRD	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	• 1	1_1
trophoblast	ENCFF832MMQ	ENCSR955XFL	bed narrowPeak	pseudo-replicated peaks	H3K27ac-human	1	1_1
aorta	ENCFF887VHI	ENCSR128VHV	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
aorta	ENCFF634IHR	ENCSR196PGM	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
CD4-positive, alpha-beta T cell	ENCFF614QJJ	ENCSR043SBG	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF310UPS	ENCSR043SBG	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_4, 2_2
CD4-positive, alpha-beta T cell	ENCFF211ERK	ENCSR733QOZ	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF565COF	ENCSR103GGR	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_4
CD8-positive, alpha-beta T cell	ENCFF200PGJ	ENCSR639HVJ	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
CD8-positive, alpha-beta T cell	ENCFF515XPE	ENCSR797GOJ	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
ectodermal cell	ENCFF940CGE	ENCSR690GLT	bed narrowPeak	replicated peaks	H3K27me3-human	1, 3	1_2, 3_2
endodermal cell	ENCFF901CSA	ENCSR273IYV	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_5, 2_5
esophagus	ENCFF472RYD	ENCSR088GXB	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
esophagus	ENCFF033JBF	ENCSR641RQV	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
H1	ENCFF156RHD	ENCSR186OBR	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
H1	ENCFF084QDP	ENCSR216OGD	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
H1	ENCFF098JFF	ENCSR687FDK	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
H1	ENCFF411ESN	ENCSR928HYM	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_2, 2_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
mesenchymal stem cell	ENCFF541LUL	ENCSR262VXI	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
mesenchymal stem cell	ENCFF759ILD	ENCSR832JVP	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
mesendoderm	ENCFF273OXZ	ENCSR405AXO	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
neuronal stem cell	ENCFF491GSQ	ENCSR550XZG	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
neuronal stem cell	ENCFF504BOD	ENCSR692CTK	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
neuronal stem cell	ENCFF044KMT	ENCSR694LBI	bed narrowPeak	replicated peaks	H3K27me3-human	1, 2	1_1, 2_1
pancreas	ENCFF581PLO	ENCSR186QKH	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
pancreas	ENCFF398GCM	ENCSR486NDF	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1, 1_2
psoas muscle	ENCFF593RXH	ENCSR720SAS	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
psoas muscle	ENCFF285KGM	ENCSR843KHS	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
sigmoid colon	ENCFF199HGP	ENCSR042RIW	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
sigmoid colon	ENCFF145YBF	ENCSR897TGR	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
small intestine	ENCFF128PPS	ENCSR340OPI	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
small intestine	ENCFF321PQU	ENCSR859EIX	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
small intestine	ENCFF954AUE	ENCSR877PAS	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
spleen	ENCFF948PIU	ENCSR408ONP	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
spleen	ENCFF492VAT	ENCSR408ONP	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
spleen	ENCFF218HXK	ENCSR608FDQ	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
spleen	ENCFF065SNS	ENCSR608FDQ	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
stomach	ENCFF787QYF	ENCSR354IST	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_2
stomach	ENCFF464IGY	ENCSR527BFF	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
trophoblast	ENCFF027NPI	ENCSR221ZRM	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
trophoblast	ENCFF740AIH	ENCSR495OEG	bed narrowPeak	pseudo-replicated peaks	H3K27me3-human	1	1_1
aorta	ENCFF220HID	ENCSR673JYT	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
aorta	ENCFF695OJS	ENCSR989AMI	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
CD4-positive, alpha-beta T cell	ENCFF062JZY	ENCSR774OKQ	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_3
CD4-positive, alpha-beta T cell	ENCFF670MKJ	ENCSR828WZG	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF687BTK	ENCSR828WZG	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_4, 2_2
CD8-positive, alpha-beta T cell	ENCFF190SHZ	ENCSR681OSD	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
CD8-positive, alpha-beta T cell	ENCFF430GVA	ENCSR694CDP	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
CD8-positive, alpha-beta T cell	ENCFF865NBY	ENCSR782NOO	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_3
endodermal cell	ENCFF280ASK	ENCSR677EZB	bed narrowPeak	replicated peaks	H3K36me3-human	1, 3	1_4, 3_2
esophagus	ENCFF693ZIC	ENCSR034ZHF	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
esophagus	ENCFF878XQI	ENCSR279MCN	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
H1	ENCFF093YFN	ENCSR476KTK	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
H1	ENCFF422MTS	ENCSR496DCY	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_2
H1	ENCFF736WAN	ENCSR925LJZ	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
mesenchymal stem cell	ENCFF115LHE	ENCSR555QHZ	bed narrowPeak	replicated peaks	H3K36me3-human	2, 3	2_1, 3_1
mesenchymal stem cell	ENCFF051RNB	ENCSR824UNY	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
mesendoderm	ENCFF771IYF	ENCSR144RXL	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
mesodermal cell	ENCFF049SYA	ENCSR100LWU	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_4, 2_2

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
neuronal stem cell	ENCFF286UMF	ENCSR238WMO	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
neuronal stem cell	ENCFF999NTP	ENCSR256ESY	bed narrowPeak	replicated peaks	H3K36me3-human	1, 2	1_1, 2_1
pancreas	ENCFF752OPF	ENCSR393HBQ	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
pancreas	ENCFF987DRQ	ENCSR943JOF	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
psoas muscle	ENCFF209YRO	ENCSR277PDE	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_2
sigmoid colon	ENCFF567DCF	ENCSR445RFF	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
sigmoid colon	ENCFF332VLM	ENCSR751JOQ	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
small intestine	ENCFF599NYJ	ENCSR073YZL	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
small intestine	ENCFF842DWY	ENCSR205NEW	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_4
small intestine	ENCFF384JUV	ENCSR958DEW	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
spleen	ENCFF294HWV	ENCSR078BHK	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
spleen	ENCFF206EGT	ENCSR078BHK	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
spleen	ENCFF224MXB	ENCSR466DUB	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
spleen	ENCFF462KMA	ENCSR466DUB	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
stomach	ENCFF435RTF	ENCSR269GMC	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
stomach	ENCFF845PNV	ENCSR697YSL	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_2
trophoblast	ENCFF584XPR	ENCSR482KJD	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
trophoblast	ENCFF493PZU	ENCSR746MTU	bed narrowPeak	pseudo-replicated peaks	H3K36me3-human	1	1_1
aorta	ENCFF223WWD	ENCSR957BPJ	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
aorta	ENCFF455KIC	ENCSR960EVO	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
CD4-positive, alpha-beta T cell	ENCFF736VPN	ENCSR263WLD	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF093RDC	ENCSR263WLD	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_4, 2_2
CD4-positive, alpha-beta T cell	ENCFF674ZVF	ENCSR852FRR	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF269CAD	ENCSR166ZZZ	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
CD8-positive, alpha-beta T cell	ENCFF820IRX	ENCSR231FDF	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
CD8-positive, alpha-beta T cell	ENCFF295CWC	ENCSR660KHZ	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
CD8-positive, alpha-beta T cell	ENCFF958QRD	ENCSR796CSH	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
ectodermal cell	ENCFF450NCT	ENCSR807LJO	bed narrowPeak	replicated peaks	H3K4me3-human	1, 3	1_2, 3_2
endodermal cell	ENCFF599VIR	ENCSR446ZCY	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_5, 2_5
esophagus	ENCFF812UXQ	ENCSR577ILY	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
esophagus	ENCFF232JVH	ENCSR697GPO	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
H1	ENCFF744ORJ	ENCSR003SSR	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
H1	ENCFF408FCY	ENCSR019SQX	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
H1	ENCFF277AOQ	ENCSR443YAS	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
mesenchymal stem cell	ENCFF493ZYQ	ENCSR004AKD	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
mesenchymal stem cell	ENCFF011GWS	ENCSR501JET	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
mesendoderm	ENCFF052VWO	ENCSR441SAT	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_2, 2_2
mesodermal cell	ENCFF870OTW	ENCSR959RHF	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_2, 2_2
neuronal stem cell	ENCFF043FGL	ENCSR354XWM	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_2, 2_1
neuronal stem cell	ENCFF480IBO	ENCSR956CTX	bed narrowPeak	replicated peaks	H3K4me3-human	1, 2	1_1, 2_1
pancreas	ENCFF918CGC	ENCSR315LPR	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_3

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
pancreas	ENCFF147VOD	ENCSR747VED	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
psoas muscle	ENCFF925HIR	ENCSR245BEV	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
psoas muscle	ENCFF783HFC	ENCSR949OYZ	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
sigmoid colon	ENCFF890JZD	ENCSR321SZE	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
sigmoid colon	ENCFF197CZF	ENCSR421HUB	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
small intestine	ENCFF012TFR	ENCSR237QFJ	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
small intestine	ENCFF563IRG	ENCSR792IJA	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
small intestine	ENCFF213LSU	ENCSR944QSH	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
spleen	ENCFF590BSY	ENCSR432KIH	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
spleen	ENCFF814FPR	ENCSR432KIH	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
spleen	ENCFF772AFU	ENCSR448FZC	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
spleen	ENCFF441DWZ	ENCSR448FZC	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
stomach	ENCFF805ZSK	ENCSR129NCV	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_1
stomach	ENCFF291FKK	ENCSR202RXT	bed narrowPeak	pseudo-replicated peaks	H3K4me3-human	1	1_2
aorta	ENCFF978QLH	ENCSR065ZNA	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
CD4-positive, alpha-beta T cell	ENCFF326XIB	ENCSR453GNY	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_4
CD4-positive, alpha-beta T cell	ENCFF255BCT	ENCSR453GNY	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_4, 2_2
CD4-positive, alpha-beta T cell	ENCFF323MFM	ENCSR787WLV	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_2
CD8-positive, alpha-beta T cell	ENCFF668CHY	ENCSR815PSO	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
CD8-positive, alpha-beta T cell	ENCFF552VSA	ENCSR824PXG	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_3
CD8-positive, alpha-beta T cell	ENCFF320HLH	ENCSR905SHH	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_1, 2_1
ectodermal cell	ENCFF482VRI	ENCSR235CEI	bed narrowPeak	replicated peaks	H3K9me3-human	1, 3	1_2, 3_2
endodermal cell	ENCFF446HIG	ENCSR823BHO	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_5, 2_5
esophagus	ENCFF040EBA	ENCSR150GLE	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
esophagus	ENCFF277ETB	ENCSR200WDD	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
H1	ENCFF250GSY	ENCSR395USV	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_1, 2_2
H1	ENCFF348GGB	ENCSR883AQJ	bed narrowPeak	replicated peaks	H3K9me3-human	1, 3	1_2, 3_1, 3_2
mesenchymal stem cell	ENCFF188ZUZ	ENCSR439EHQ	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_1, 2_1
mesenchymal stem cell	ENCFF022VRW	ENCSR746CUY	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_1, 2_1
mesodermal cell	ENCFF708SUO	ENCSR887ZPC	bed narrowPeak	replicated peaks	H3K9me3-human	2, 3	2_2, 3_2
neuronal stem cell	ENCFF718CGL	ENCSR391WDE	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_1, 2_1
neuronal stem cell	ENCFF304FHZ	ENCSR800IIW	bed narrowPeak	replicated peaks	H3K9me3-human	1, 2	1_1, 2_1
pancreas	ENCFF501UUF	ENCSR035QNZ	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
pancreas	ENCFF153MCW	ENCSR533HDU	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
psoas muscle	ENCFF368EHI	ENCSR394DRL	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
sigmoid colon	ENCFF824HBO	ENCSR636IDR	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
sigmoid colon	ENCFF417EMO	ENCSR737NLJ	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
small intestine	ENCFF482VIL	ENCSR270VNK	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
small intestine	ENCFF303WRT	ENCSR417RFS	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
small intestine	ENCFF274KHB	ENCSR773TWR	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_2
spleen	ENCFF561FFX	ENCSR249XEB	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1

Biosample term name	File accession	Experiment accession	File format	Output type	Experiment target	Biological replicate(s)	Technical replicate(s)
spleen	ENCFF688HWH	ENCSR249XEB	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
spleen	ENCFF932LXW	ENCSR421FPV	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
spleen	ENCFF424ZSF	ENCSR421FPV	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
stomach	ENCFF957ZGX	ENCSR447GWQ	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_3
stomach	ENCFF822CXT	ENCSR639RKZ	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
stomach	ENCFF496LZT	ENCSR885CMN	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
trophoblast	ENCFF309GOP	ENCSR356JTB	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1
trophoblast	ENCFF335ZGF	ENCSR612MZB	bed narrowPeak	pseudo-replicated peaks	H3K9me3-human	1	1_1

Table A.4.: Metadata for the retrieved Human Epigenome Atlas ChIP-seq peak files. For each biosample, the associated information include the file accession number, experiment accession number, file format and replicates are listed 5 ChIP-seq targets, including H3K27ac, H3K27me3, H3K36me3, H3K4me3 and H3K9me3.

Symbol	Gene description
A2ML1	alpha-2-macroglobulin like 1
COL11A2	collagen type XI alpha 2 chain
DNAH1	dynein axonemal heavy chain 1
DNAH10	dynein axonemal heavy chain 10
DNAH9	dynein axonemal heavy chain 9
HYDIN	HYDIN axonemal central pair apparatus protein
IPO5	importin 5
JAKMIP1	janus kinase and microtubule interacting protein 1
MYO1H	myosin IH
NEB	nebulin
RANBP17	RAN binding protein 17
TRPM5	transient receptor potential cation channel subfamily M member 5
USHBP1	USH1 protein network component harmonin binding protein 1
VWA5B1	von Willebrand factor A domain containing 5B1

Table A.5.: List of genes with odds ratio > 1 and Fisher Exact Test adjusted $p - value \leq 0.05$

	DEU	H3K27ac	H3K27me3	H3K36me3	H3K4me3	H3K9me3
Potency	0.091	0.481	0.423	0.056	0.052	0.105
Type	0.177	0.379	0.130	0.048	0.162	0.118
Origin	0.459	0.420	0.176	0.022	0.200	0.152
Life stage	0.016	0.304	0.238	0.031	0.033	0.029

Table A.6.: Adjusted Rand indices measuring the similarity between differential features-based hierarchical clustering based on and tissue label schemes. Investigated cell types were separated by potency, sample type, origin and life stage and compared to the cluster labels from hierarchical clustering.

Chapter B.

Supplementary Data for Chapter III

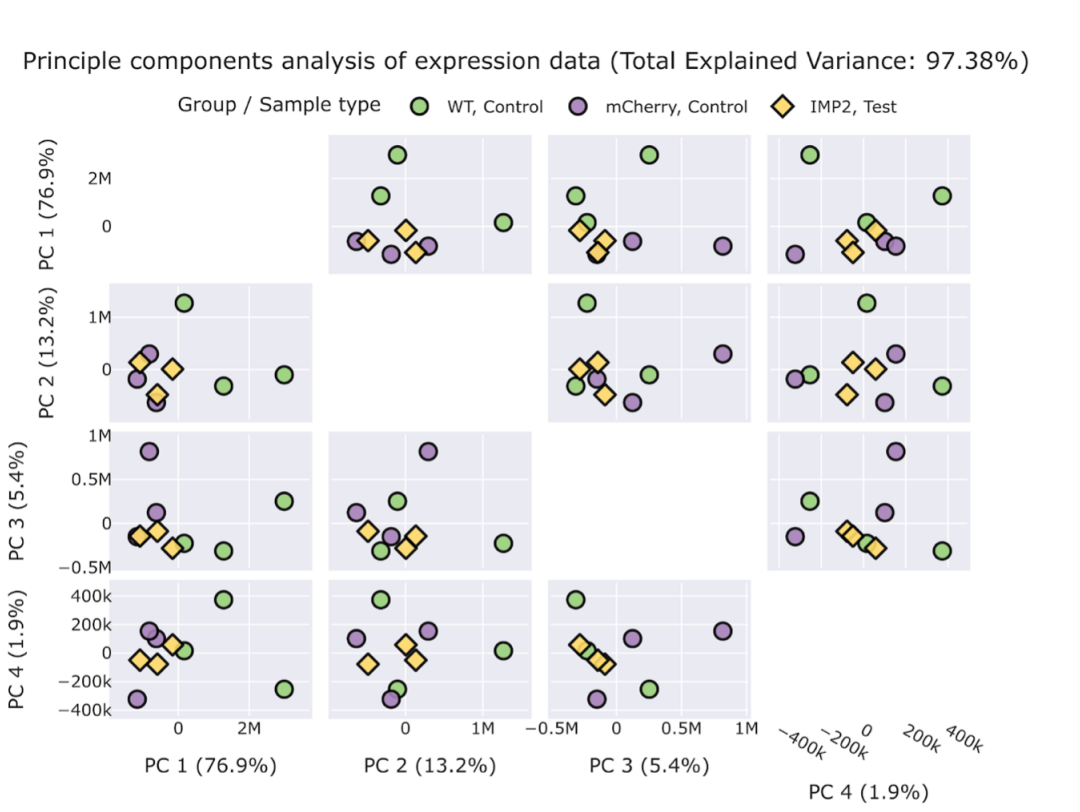

Figure B.1.: Principal Component Analysis (PCA) of gene expression across samples. PCA was performed on the matrix containing raw gene expression levels from individual replicates of three sample groups, i.e., wild type (WT), mCherry, and IMP2. These expression data were transformed and clustered using the three (A) or four (B) largest principal components. The replicates are grouped by colors, while control samples, including WT and mCherry, and experiment samples (IMP2) are distinguished by round and diamond shapes, respectively.

	INTERSECT (1%)			**INTERSECT (5%)**			**UNION (1%)**			**UNION (5%)**		
	WT vs. IMP2	mCherry vs. IMP2	WT vs. mCherry	WT vs. IMP2	mCherry vs. IMP2	WT vs. mCherry	WT vs. IMP2	mCherry vs. IMP2	WT vs. mCherry	WT vs. IMP2	mCherry vs. IMP2	WT vs. mCherry
Editing Site	276	321	96	59	55	8	366	362	119	83	77	14
Overlapping gene span (10bp)	500	635	315	68	66	10	692	720	363	95	93	14
CDS/UTR	715	976	532	111	167	18	982	1057	579	174	187	24
Transcript	1347	1837	1148	170	259	29	1888	2032	1251	362	119	44

Table B.1.: Number of genes with A2G sites for different replicate collapsing schemes. Using HyperTRIBE, we identified and compared the editing sites between IMP2 and either one of the controls (WT or mCherry) or between the controls. Listed here are gene counts with either exactly overlapping editing sites or showing overlap in a certain region (10 bp window, CDS or UTR regions, and transcript) that were identified either in all replicates of a sample group (INTERSECT) or in at least two out of three replicates (UNION). Two different thresholds were applied for the average editing percentage (1% and 5%). Overall, we identified more genes containing at least one A2G site when we considered less stringent thresholds for average editing percentage (1%). As TRIBE is not able to reveal the exact binding position of the RBP on the mRNA (26), retaining only the editing sites identified in all three replicates of one condition (INTERSECT) can therefore be considered as a fairly strict filter, which may cause loss of information. Finally, we selected the genomic span for the final replicate collapsing scheme by comparing the sets of IMP2 targets identified in this analysis to those in mouse embryonic fibroblasts, or the set of differentially expressed genes detected by DESeq2.

	INTERSECT						UNION					
	1%			5%			1%			5%		
	WT-IMP2	mCherry-IMP2	WT-mCherry	WT-IMP2	mCherry-IMP2	WT-mCherry	WT-IMP2	mCherry-IMP2	WT-mCherry	WT-IMP2	mCherry-IMP2	WT-mCherry
Editing Site	173 (62.68%)	195 (60.75%)	67 (69.79%)	33 (55.93%)	26 (47.27%)	1 (12.50%)	219 (59.84%)	215 (59.39%)	80 (67.23%)	36 (43.37%)	32 (41.56%)	5 (35.71%)
Overlapping gene span (10bp)	320 (64.0%)	388 (61.10%)	205 (65.08%)	36 (54.55%)	34 (50.00%)	2 (20.00%)	424 (61.27%)	435 (60.42%)	233 (64.19%)	41 (43.16%)	41 (44.09%)	5 (35.71%)
CDS/UTR	443 (61.96%)	594 (60.86%)	363 (68.23%)	59 (53.15%)	87 (52.10%)	7 (38.89%)	595 (60.59%)	634 (59.98%)	386 (66.67%)	84 (48.28%)	94 (50.27%)	12 (50.0%)
Transcript	756 (56.12%)	1019 (55.47%)	692 (60.28%)	7 (38.89%)	116 (44.79%)	10 (34.48%)	1048 (55.51%)	1118 (55.02%)	739 (59.07%)	121 (42.76%)	131 (43.23%)	15 (34.09%)

Table B.2.: The number and percentage of genes with A2G sites overlapping with HyperTRIBE results from mouse embryonic fibroblast (MEF) samples. For each replicate collapsing scheme, the set of genes with at least one editing site was compared to the HyperTRIBE results of MEF samples, which used the same average editing percentage thresholds of 1% and 5%. The number of overlapping genes between the two sets, as well as their percentage in all genes found with A2G sites, are reported for the comparisons between IMP2 and either controls (WT or mCherry) or between the controls. The overlap was quite small when using a 5% threshold for the average editing percentage. Thus, we selected the results from the 1% threshold to ensure better interpretability and robustness.

	WT *vs.* IMP2	**mCherry *vs.* IMP2**	**WT *vs.* mCherry**	**WT/mCherry *vs.* IMP2**	**WT/mCherry *vs.* IMP2 - WT *vs.* mCherry (IMP2 genes)**	**WT/mCherry *vs.* IMP2 - WT *vs.* mCherry**
						WT/mCherry *vs.* IMP2
	A = HTRIBE ∩ DEG	**B = HTRIBE ∩ DEG**	**C = HTRIBE ∩ DEG**	**D = HTRIBE(WT *vs.* IMP2 ∪ mCherry *vs.* IMP2) ∩ DEG (WT *vs.* IMP2 ∪ mCherry *vs.* IMP2)**	**E = HTRIBE(WT *vs.* IMP2 ∪ mCherry vs. IMP2 - WT *vs.* mCherry) ∩ DEG (WT *vs.* IMP2 ∪ mCherry *vs.* IMP2 - WT *vs.* mCherry)**	**F = $\frac{E}{D}$**
Site	13.40 (17)	11.31 (7)	4.23 (4)	14.86 (21)	13.16 (9)	0.43
Window	20.25 (32)	12.35 (12)	13.58 (16)	26.18 (46)	16.8 (15)	0.33
CDS/UTR	27.55 (51)	12.26 (16)	17.32 (24)	32.85 (68)	12.62 (13)	0.19
Transcript	39.61 (107)	15.03 (34)	25.62 (52)	45.31 (138)	16.6 (26)	0.19

Table B.3.: Similarity between genes with A2G sites and differentially expressed genes (T=1%, UNION). The overlap between genes with A2G sites from HyperTRIBE analysis and genes with differential expression from DESeq2 analysis was measured using Jaccard Index x 1000 (columns A-E) and is reported as overlapping gene count. Both analyses compared IMP2 to either controls (WT or mCherry in columns A and B), or WT to mCherry (C). Similarly, we computed the overlap between HyperTRIBE and DESeq2 results, specifically for the union sets of WT vs. IMP2 and mCherry vs. IMP2 results (column D) and without genes in WT vs. mCherry result (column E). The last column (F) lists the fractions of genes showing up only in the comparisons with IMP2 samples and not in WT vs. mCherry. Supplementary Figure B.7 illustrates how gene sets were joined. Here, we considered A2G sites/regions detected at 1% average editing percentage and that are present in at least two out of three replicates. The respective gene names are listed in Table III.1 in main text. The results for CDS/UTR in this table and the corresponding gene names are presented in Table III.1 in the manuscript. Supplementary Figure B.6 shows the overlaps between IMP2-gene sets from different genomic regions.

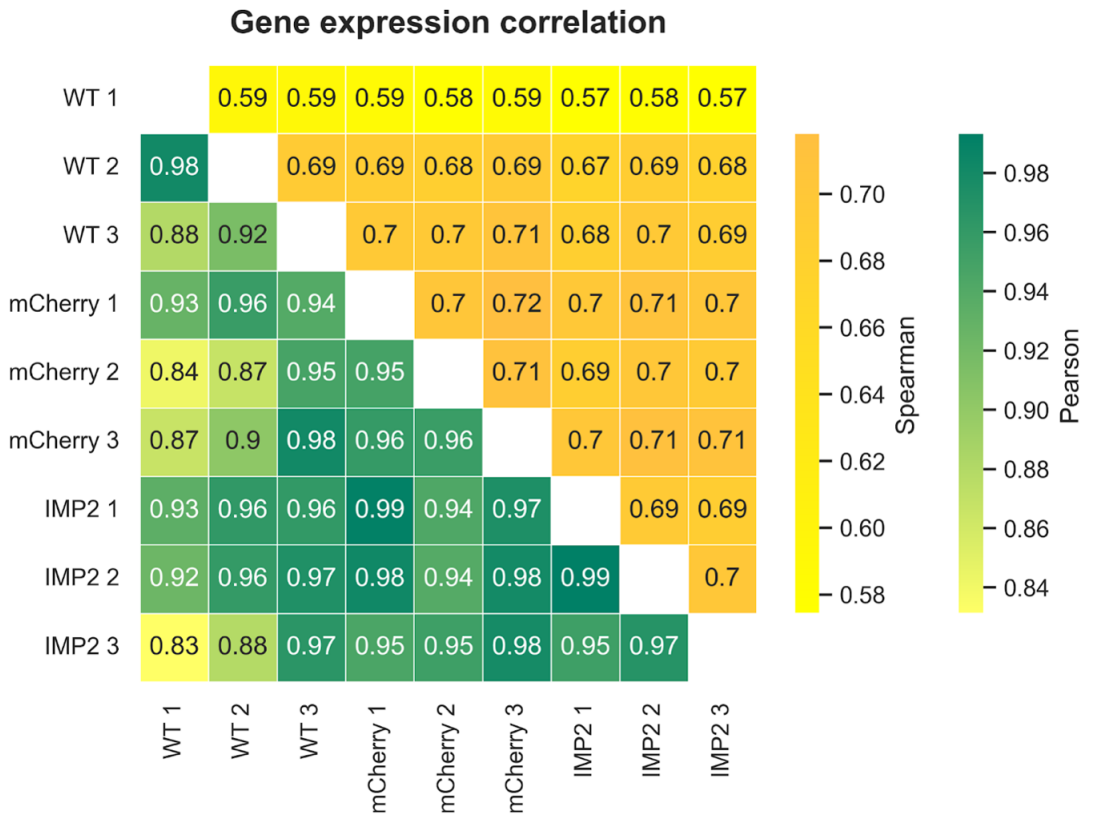

Figure B.2.: Correlation analysis of gene expression across samples. The raw expression levels from different replicates were correlated using either Pearson (lower triangle) or Spearman (upper triangle) correlation. Correlation coefficients are represented by colors in the heatmap.

	WT vs. IMP2	mCherry vs. IMP2	WT vs. mCherry
WT vs. IMP2	-	-	-
mCherry vs. IMP2	0.14	-	-
WT vs. mCherry	0.49	0.13	-

Table B.4.: Similarity between differentially expressed genes (Jaccard indices)

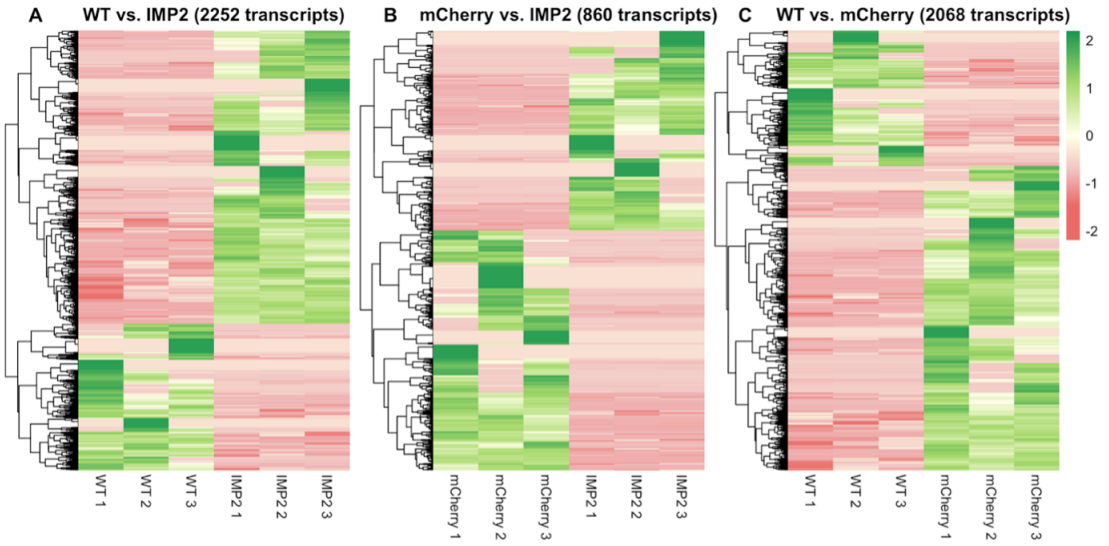

Figure B.3.: Heatmaps for expression levels of differentially expressed genes (DEGs). Using DESeq2, the lists of DEGs specific for WT and IMP2 (A), mCherry and IMP2 (B), and WT and mCherry (C) were determined ($|\text{LFC}| \geq 1$, FDR-adjusted p-value ≤ 0.05). Hierarchical clustering of DEGs shows clusters of up/down-regulated genes in the samples for each pairwise comparison.

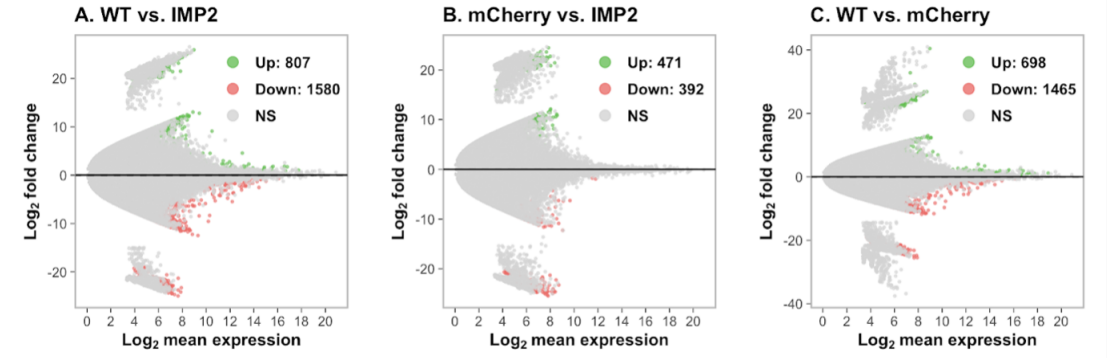

Figure B.4.: MA plots for normalized count data from differential analysis. The gene expression data for each sample was normalized, log-transformed, and compared using DESeq2 in a pairwise manner. The log fold changes (LFCs) of individual genes are plotted against the mean of normalized gene expression in the comparisons between WT and IMP2 (A), mCherry and IMP2 (B), and WT and mCherry (C) samples. Genes with differential expression ($|\text{LFC}| \geq 1$, FDR-adjusted p-value ≤ 0.05) are highlighted in green (up-regulated) or red (down-regulated), while other genes are depicted in gray (NS).

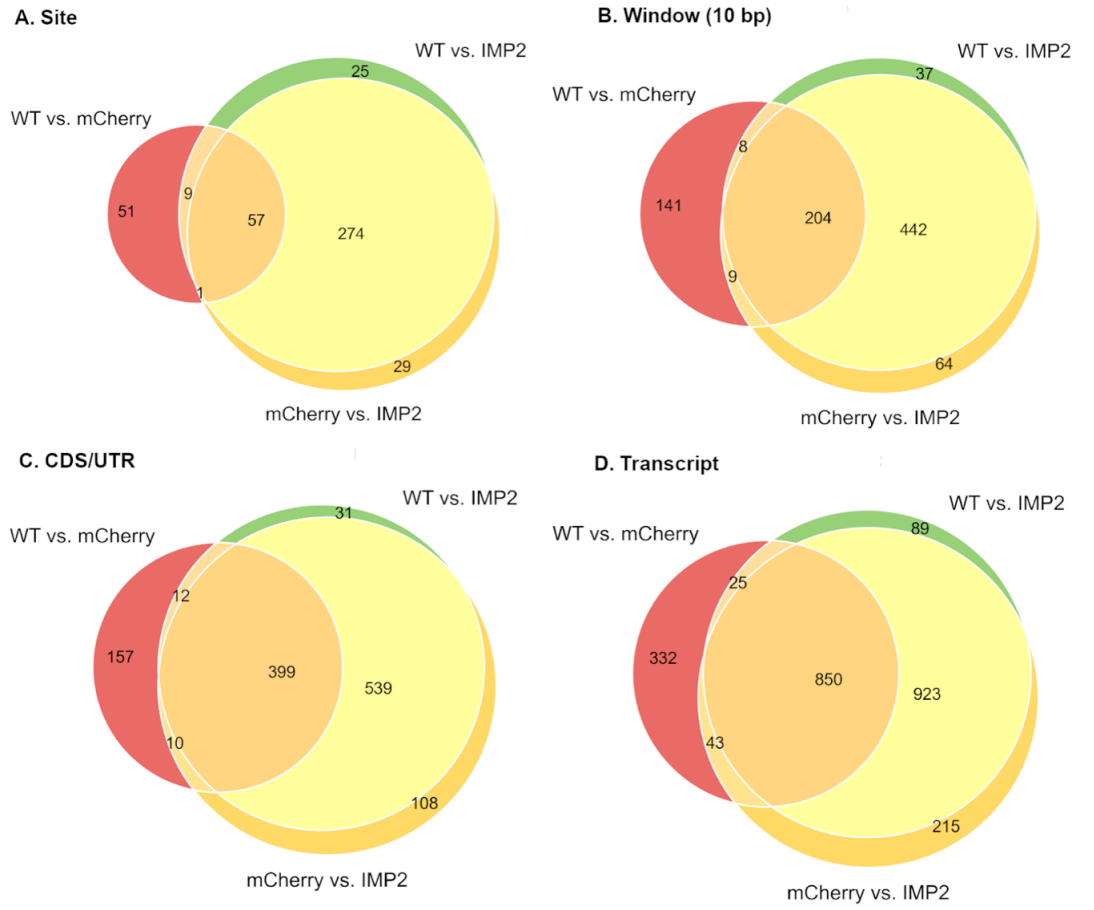

Figure B.5.: Overlaps between the sets of genes containing A2G sites from HyperTRIBE pairwise comparisons. A2G sites were detected at 1% average editing percentage. A gene must contain at least one A2G site present in a specific genome region of at least two out of three replicates to be considered in the result set. In all four investigated regions, comparing the control groups (WT or mCherry) to the IMP2 group resulted in highly similar sets of IMP2 genes (A-D), whereas the IMP2 genes from comparing the control groups take up at least 16.96% from the union of these sets (A).

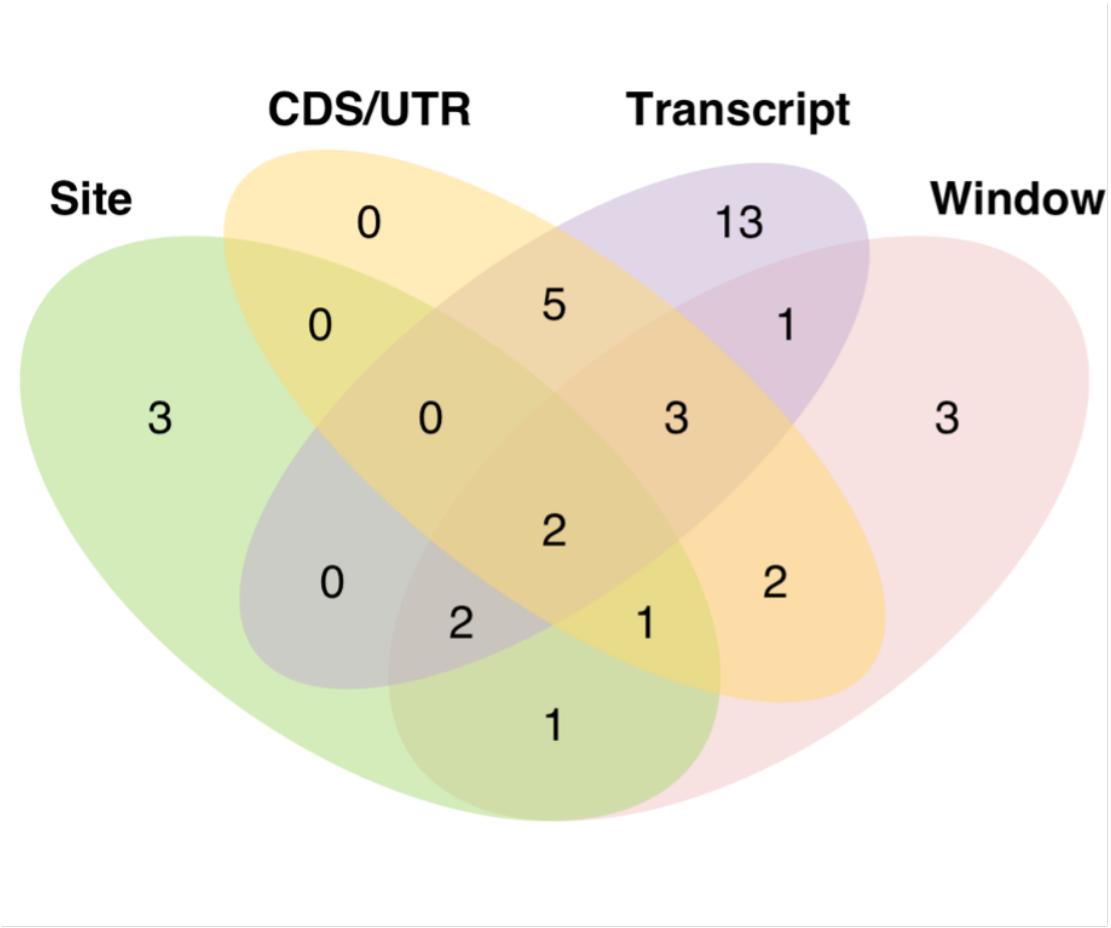

Figure B.6.: Venn diagram of overlapping IMP2 genes based on A2G sites from four different genomic regions. IMP2 genes are genes with A2G sites identified from the comparison between any control group (WT or mCherry) and IMP2 group, but not in the comparison between WT and mCherry.

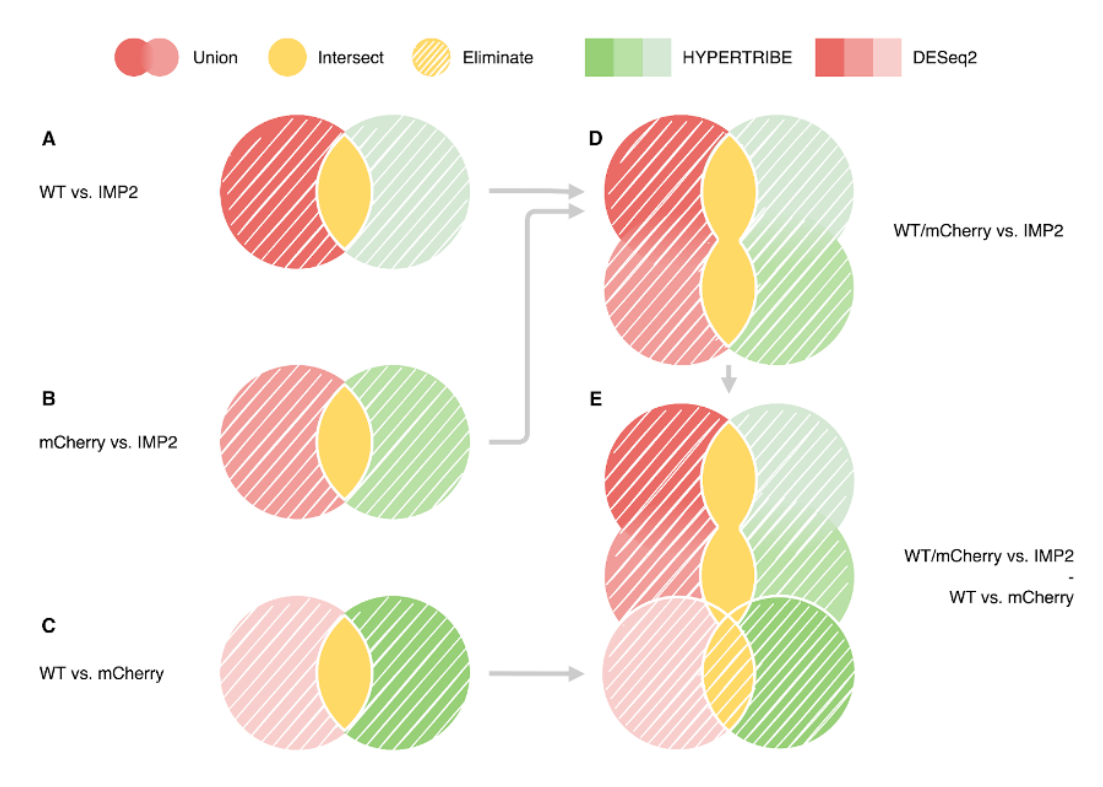

Figure B.7.: Schematic for finding the overlap between HyperTRIBE and DESeq2 gene sets. The genes with both A2G sites and deregulated transcripts are collected for the comparisons between WT or mCherry and IMP2, or between WT and mCherry (A-C). The unions of genes from WT vs. IMP2 and mCherry vs. IMP2 for HyperTRIBE and DESeq2 were overlapped and resulted in WT/mCherry vs. IMP2 gene set (D). From these, "control genes" found in WT vs. mCherry were eliminated, leaving the "IMP2-related genes". The intersection between those defines WT/mCherry vs. IMP2 - WT vs. mCherry group (E - corresponding to Supplementary Table B.3, column E).

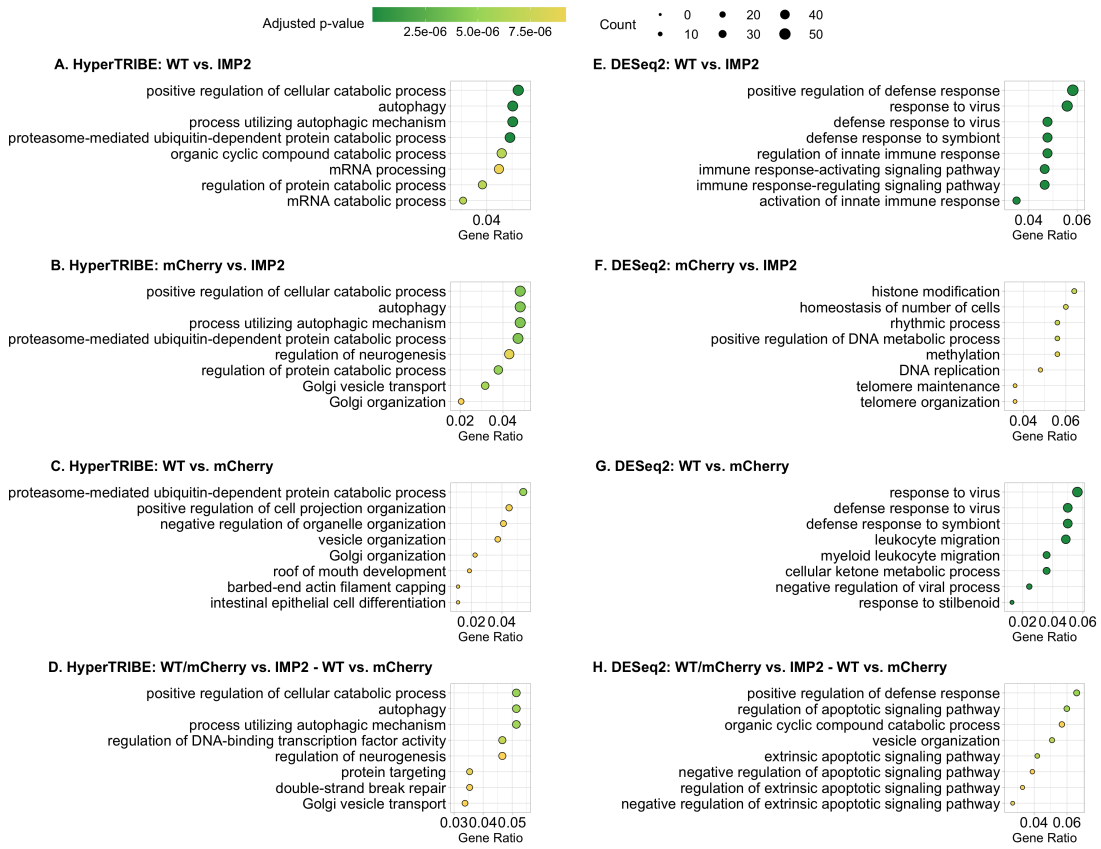

Figure B.8.: Gene ontology enrichment analysis of IMP2 target genes (A-D) and differentially expressed genes (DEGs) (E-H). Shown are the most significantly enriched biological process terms (FDR $\leq$ 0.05) for comparisons between WT and IMP2 (A, E), mCherry and IMP2 (C, G), and for any comparison to IMP2, but not for the comparison between controls (D, H). Enrichments were computed with the R package *clusterProfiler*.

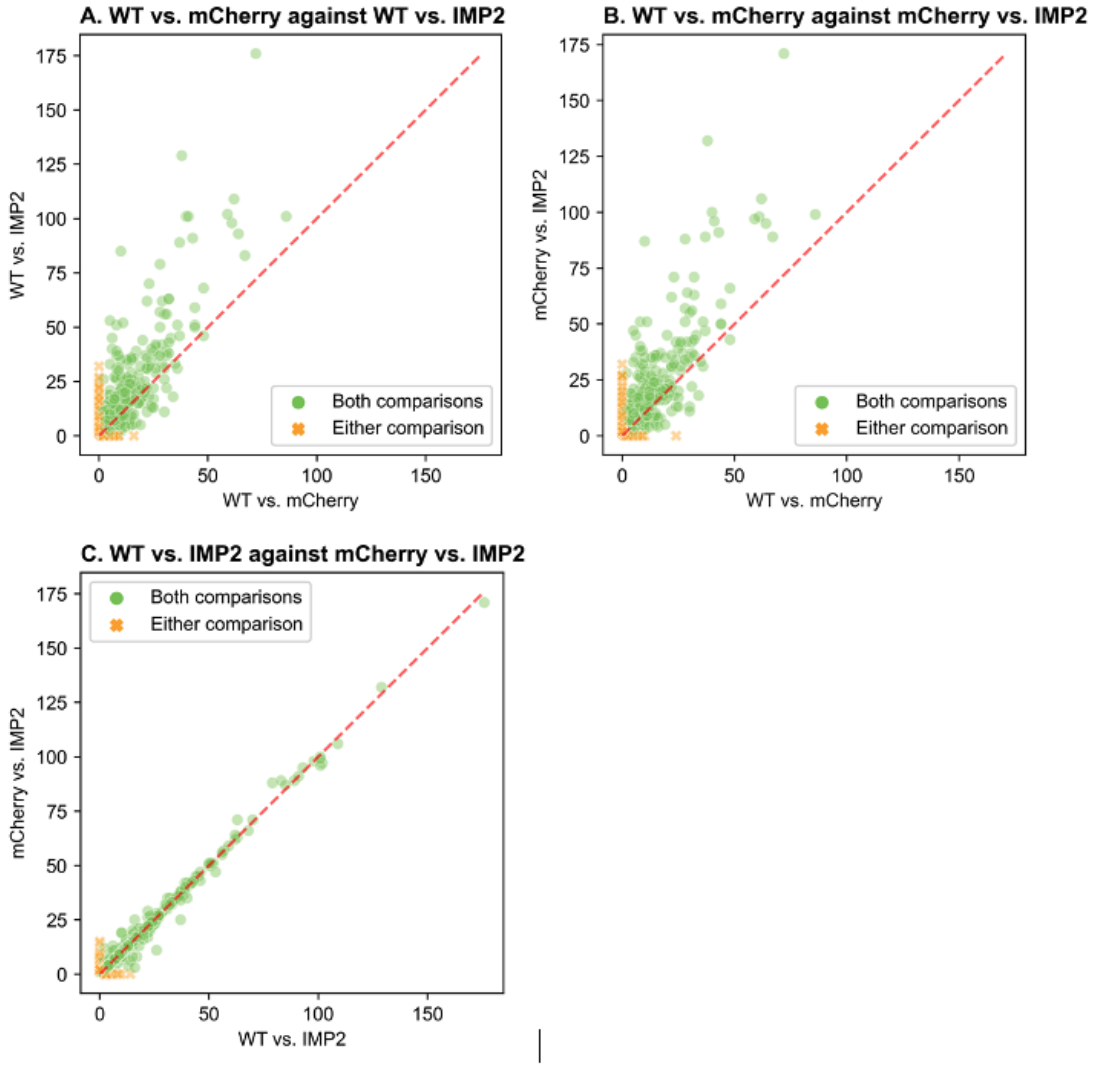

Figure B.9.: Distribution of the number of A2G sites per gene (T=1%, UNION, CDS/UTR). For different pairs of comparisons, the genes with A2G sites are plotted using the number of editing sites identified in each comparison. Genes containing editing sites only in a comparison are represented in yellow, while those with editing sites in both comparisons are depicted in green.

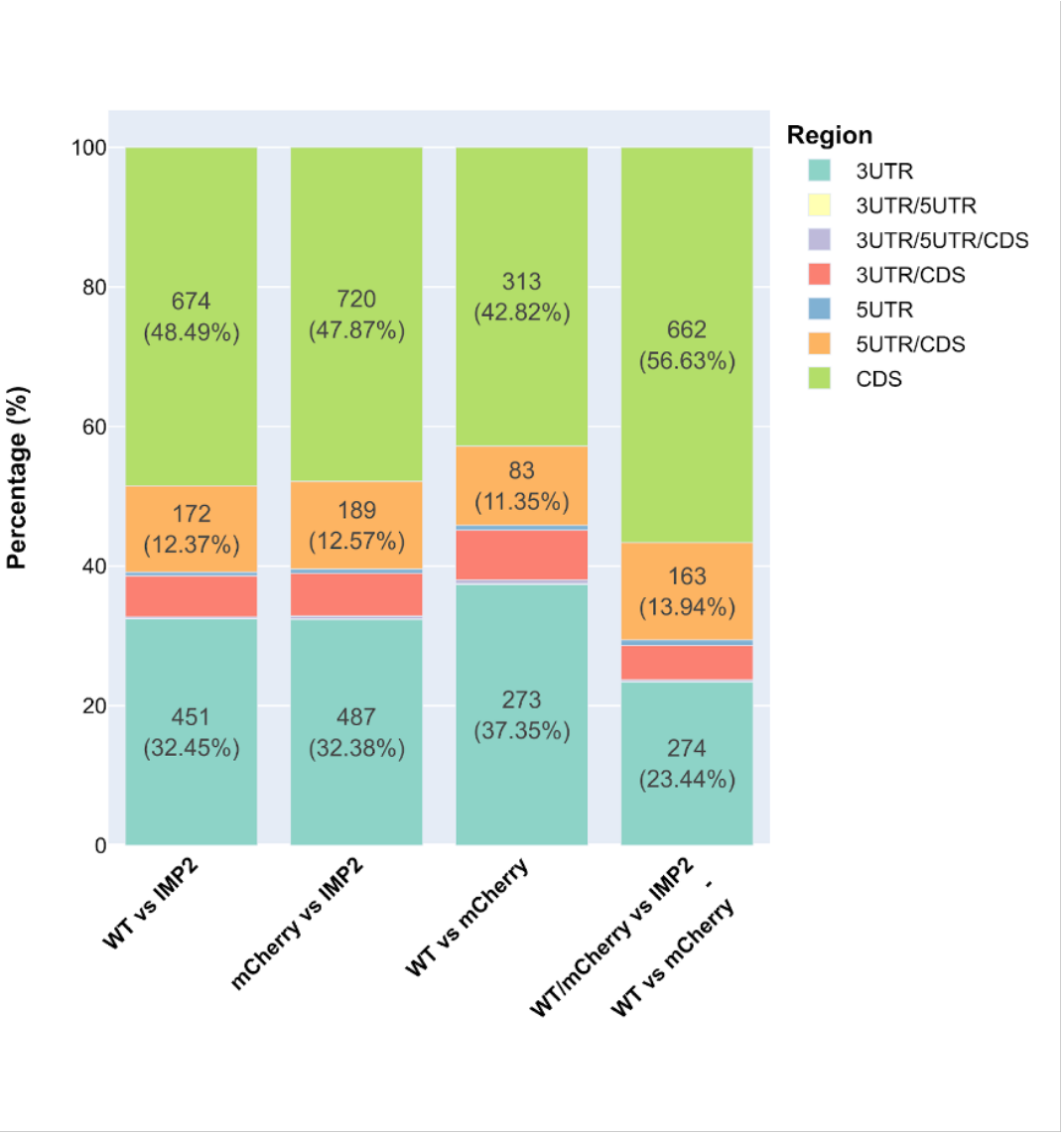

Figure B.10.: Distribution of A2G sites remained for the selected replicate collapsing scheme (T=1%, UNION, CDS/UTR). Using the *Mus musculus* reference genome mm39, regions containing A2G sites could be annotated as CDS, 3'UTR or 5'UTR in a non-exclusive manner. A2G sites detected at 1% average editing percentage and present in CDS or UTR regions of at least two out of three replicates were considered for all comparisons. The group "WT/mCherry vs. - WT vs. mCherry" consists of CDSs or UTRs that do not share any overlap with those in WT vs. mCherry.

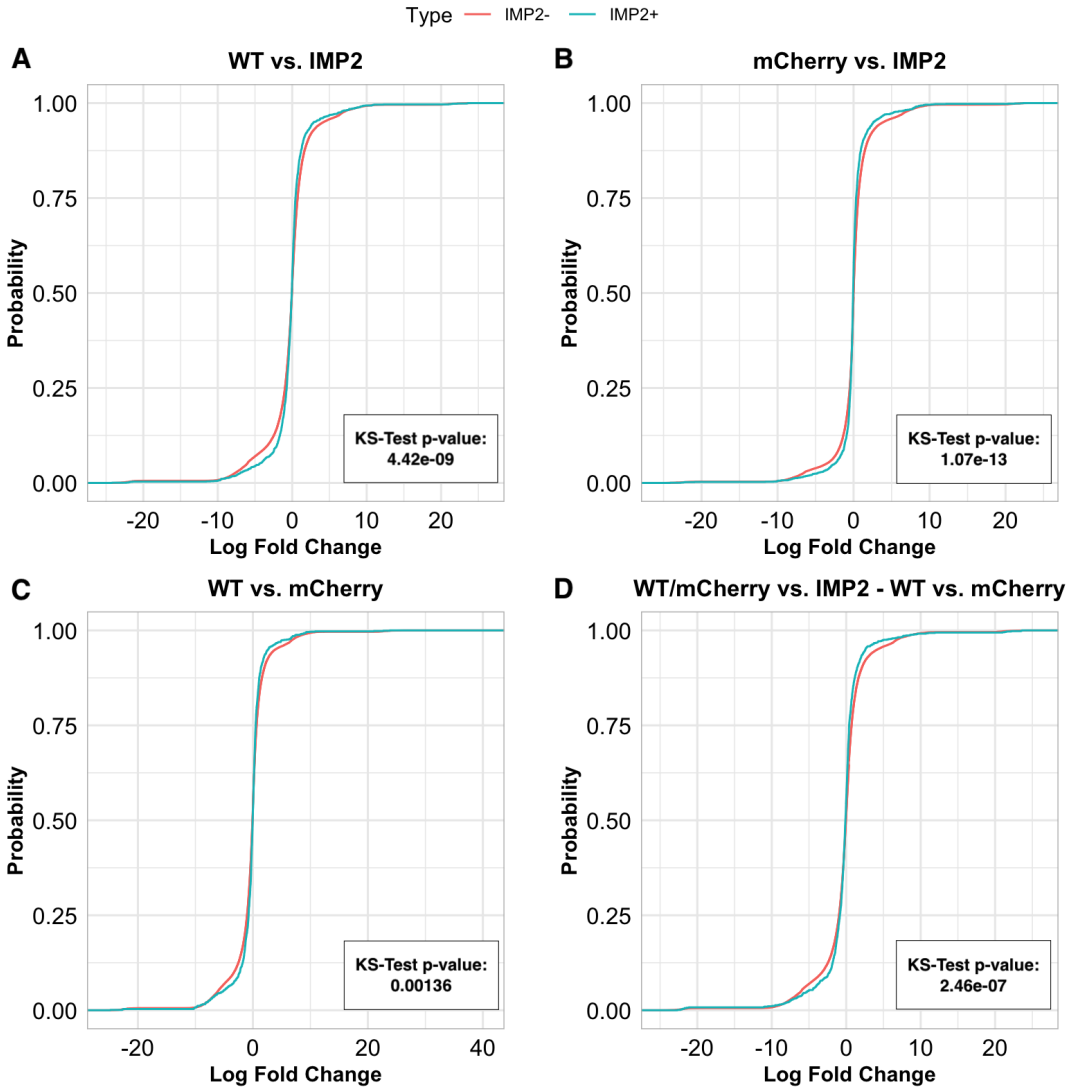

Figure B.11.: Expression profiles and motif analysis of identified IMP2 target genes. (A-D) Cumulative distribution of Log Fold Change (LFC) for IMP2 target genes (T=1%, UNION, CDS/UTR). The cumulative LFC distributions for all genes computed by DESeq2 were plotted separately for genes with and without A2G sites (denoted by IMP2+ and IMP2- with blue and red colors, respectively) that were identified by HyperTRIBE. These LFCs quantify the change in gene expression levels between IMP2 and either WT or mCherry (A and B), or between WT and mCherry (C). (D) shows the IMP2+ set of genes with A2G sites detected in any comparison against IMP2 samples, excluding those belonging to the comparison between controls. Kolmogorov-Smirnow tests were used to compare the cumulative distribution between any IMP2- and IMP2+ set of LFCs.

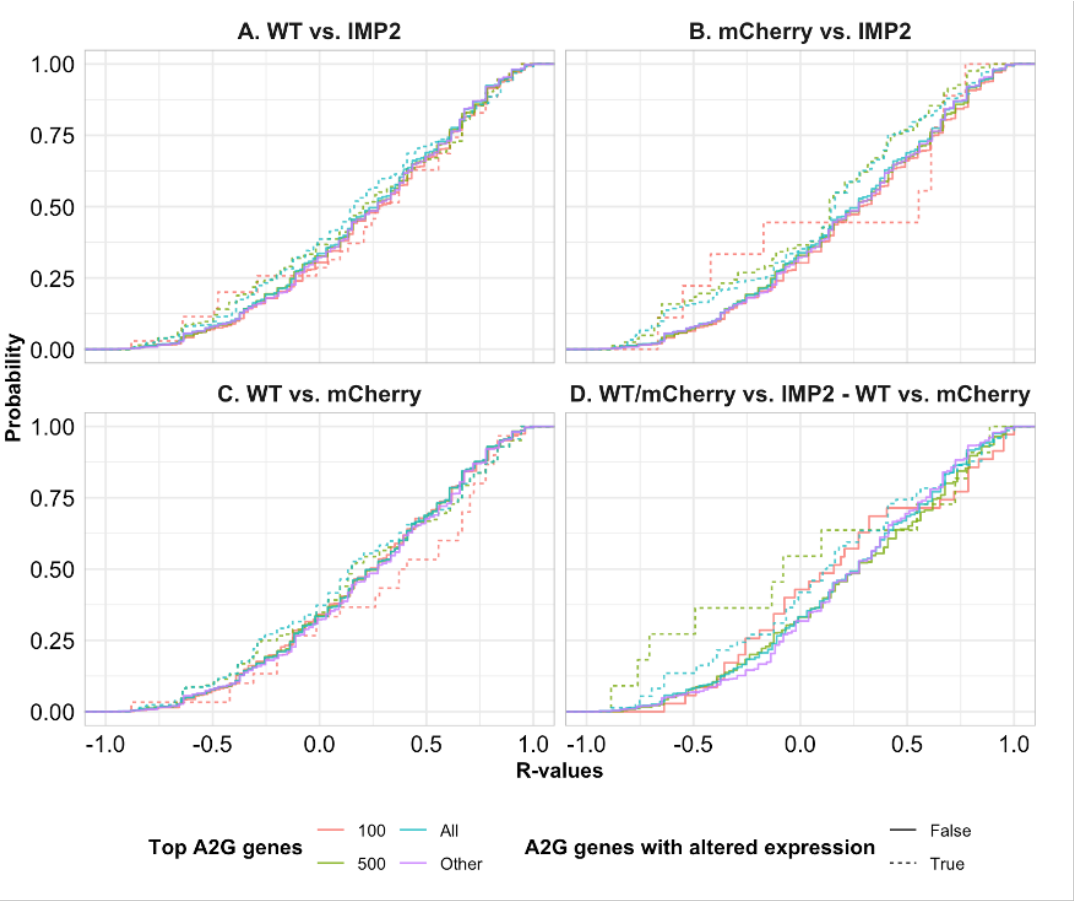

Figure B.12.: Distribution of correlation coefficients in expression level between IMP2 and other genes. Spearman correlation between the expression levels of IMP2 and all other genes was computed and the empirical cumulative distribution of the correlation coefficients was plotted. The genes were categorized into and colored according to four groups, namely all genes with A2G sites (All), top 100 and 500 A2G genes with the highest editing percentage (100 and 500), and genes with no A2G sites (Other). The genes with A2G sites detected by HyperTRIBE are specific for each of the three comparisons, between WT or mCherry and IMP2 or between WT and mCherry. Genes in WT/mCherry vs. IMP2 groups were detected in the WT or mCherry against IMP2 comparison, but not in the WT against mCherry comparison. The distributions for A2G genes with deregulated expression identified by DESeq2 listed in Table III.1 are plotted in dashed lines.

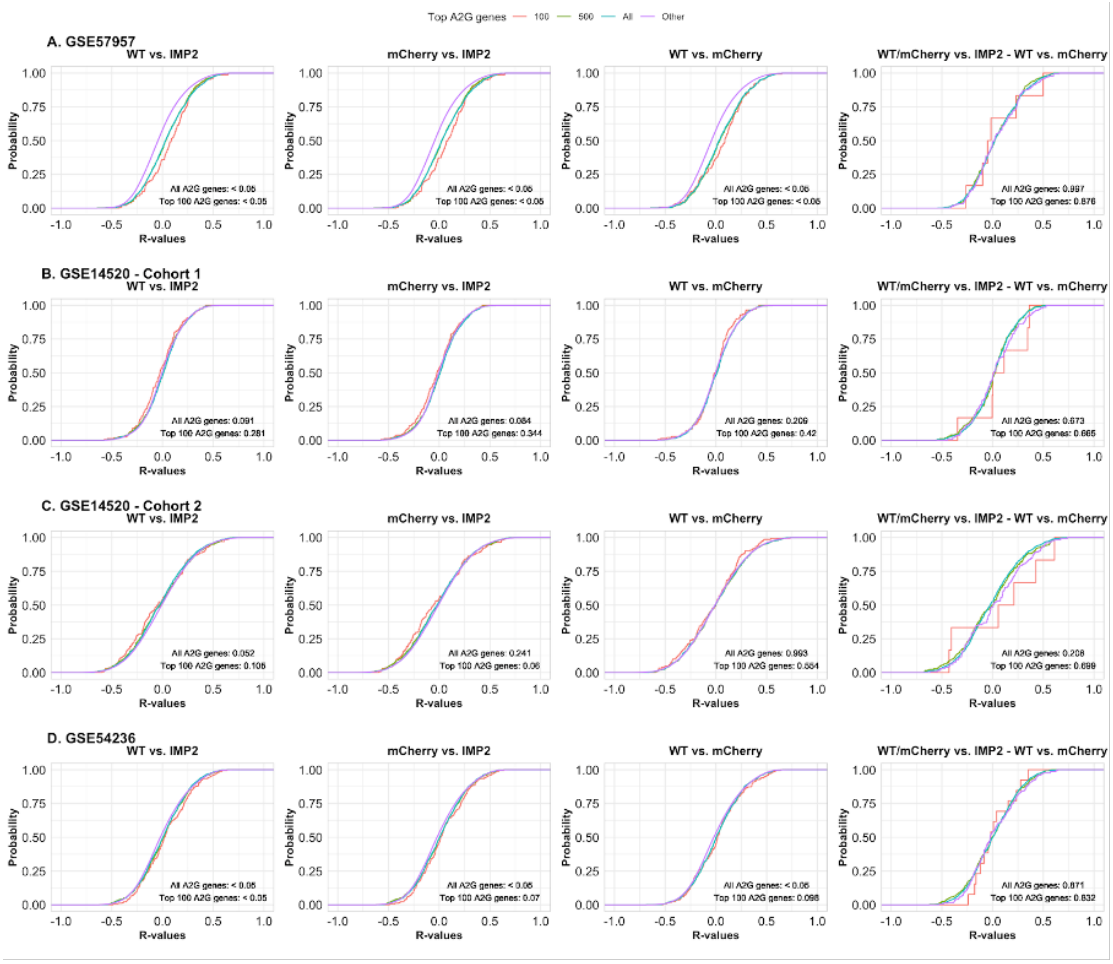

Figure B.13.: Distribution of correlation coefficients in expression level between IMP2 and other genes in public datasets. We retrieved from Gene Expression Omnibus four datasets containing RNA-seq data for normal liver tissues in human patients (GSE57957, GSE14520 cohort 1 and 2, and GSE54236 with 59, 22, 42, and 76 samples, respectively). For each dataset, the Spearman correlation between the expression levels of IMP2 and all other genes was computed and the empirical cumulative distribution of the correlation coefficients was plotted. The genes were categorized into and colored according to four groups, namely all genes with A2G sites (All), top 100 and 500 A2G genes with the highest editing percentage (100 and 500), and genes with no A2G sites (Other). The genes with A2G sites detected by HyperTRIBE are specific for each of the three comparisons, between WT or mCherry and IMP2 or between WT and mCherry. Genes in WT/mCherry vs. IMP2 groups were detected in the WT or mCherry against IMP2 comparison, but not in the WT against mCherry comparison. The Kolmogorow-Smirnow test was performed to compare the distribution between the top 100 A2G genes (100) or all A2G genes (All) to non-A2G genes (Other) and the p-values are reported in the bottom right corner of each panel.

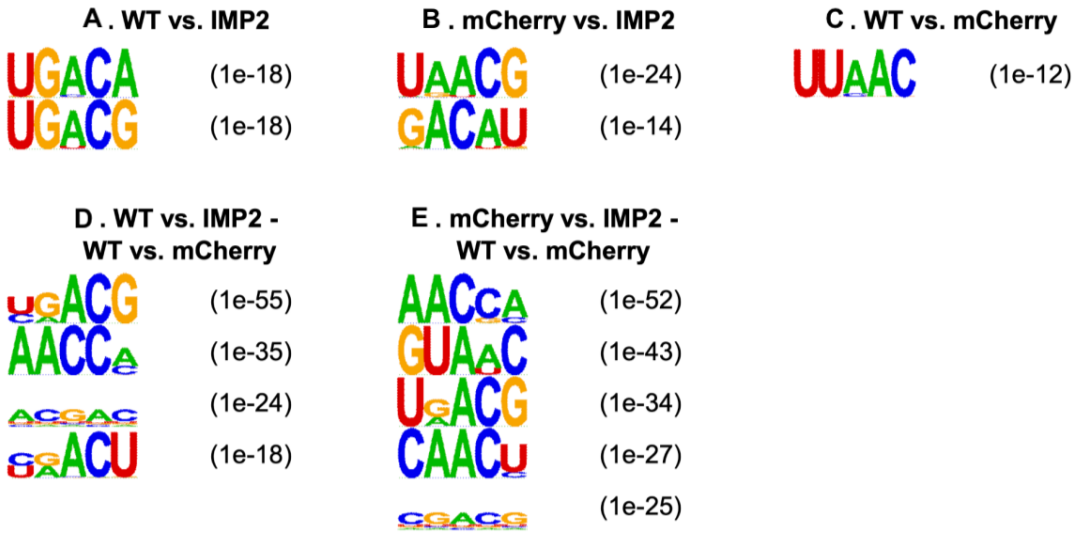

Figure B.14.: m6A-related motifs enriched in IMP2 binding regions were identified using the HOMER Motif Discovery tool. *De novo* motif analysis was performed for sequences in a [-500 bp, +500 bp] region around each A2G site of each comparison group (WT vs. IMP2, mCherry vs. IMP2 and WT vs. mCherry in Subfigure A, B, and C, respectively). In differential motif analysis, all positions identified in the WT vs. mCherry comparison were removed from the WT vs. IMP2 (D) and mCherry vs. IMP2 (E) comparisons. Significantly enriched motifs possess adjusted p-values smaller than 1e-10 and are tabulated in Supplementary File 2. Panels A to E show only motifs similar to RAC/RACH/DRACG variants of the m6A consensus motifs and their adjusted p-values (where D = A/G/U, R = A/G, and H = U/A/C).

Chapter C.

Supplementary Data for Chapter VI

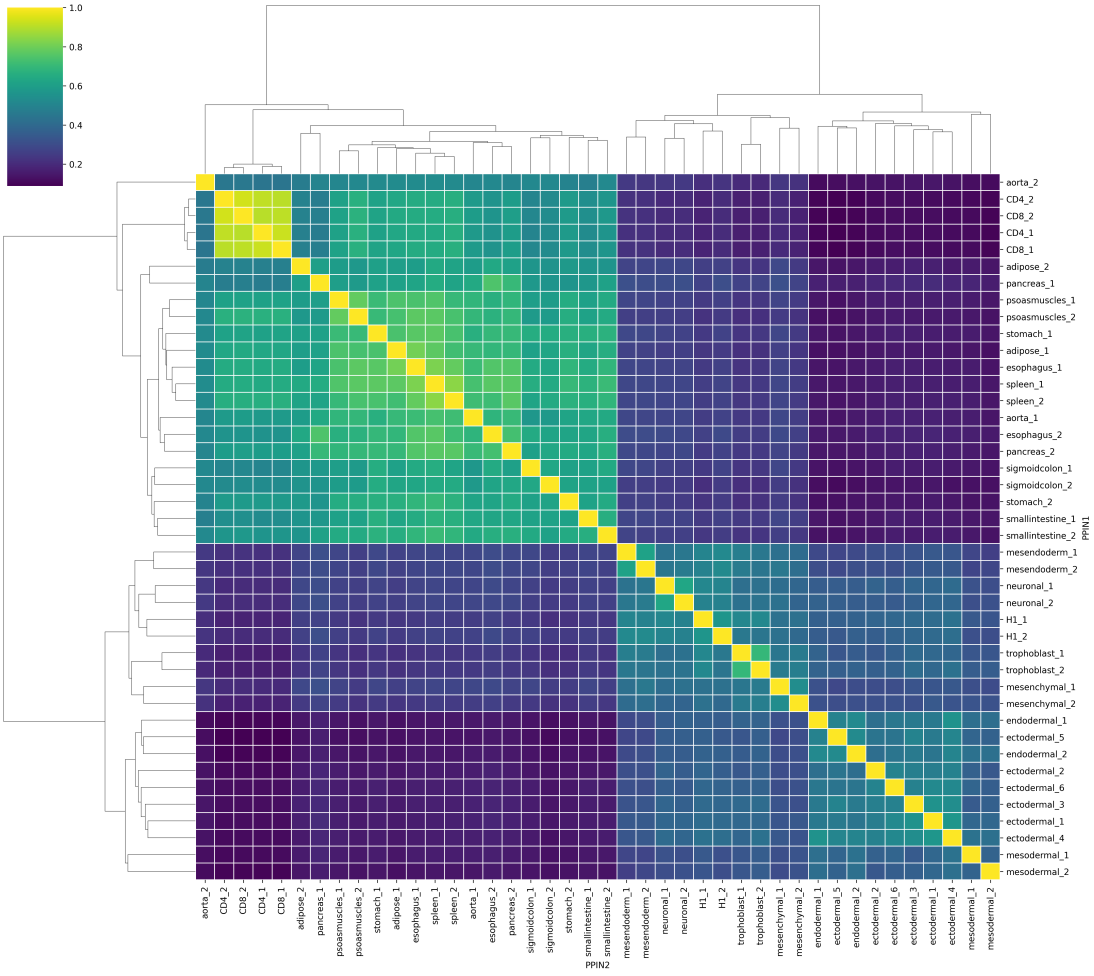

Figure C.1.: Similarity in sample-specific PPIs across 19 Human Epigenomes Atlas cells/tissue. The pairwise distance between any pair of cells/tissues was computed using the Jaccard similarity index based on their shared PPIs.

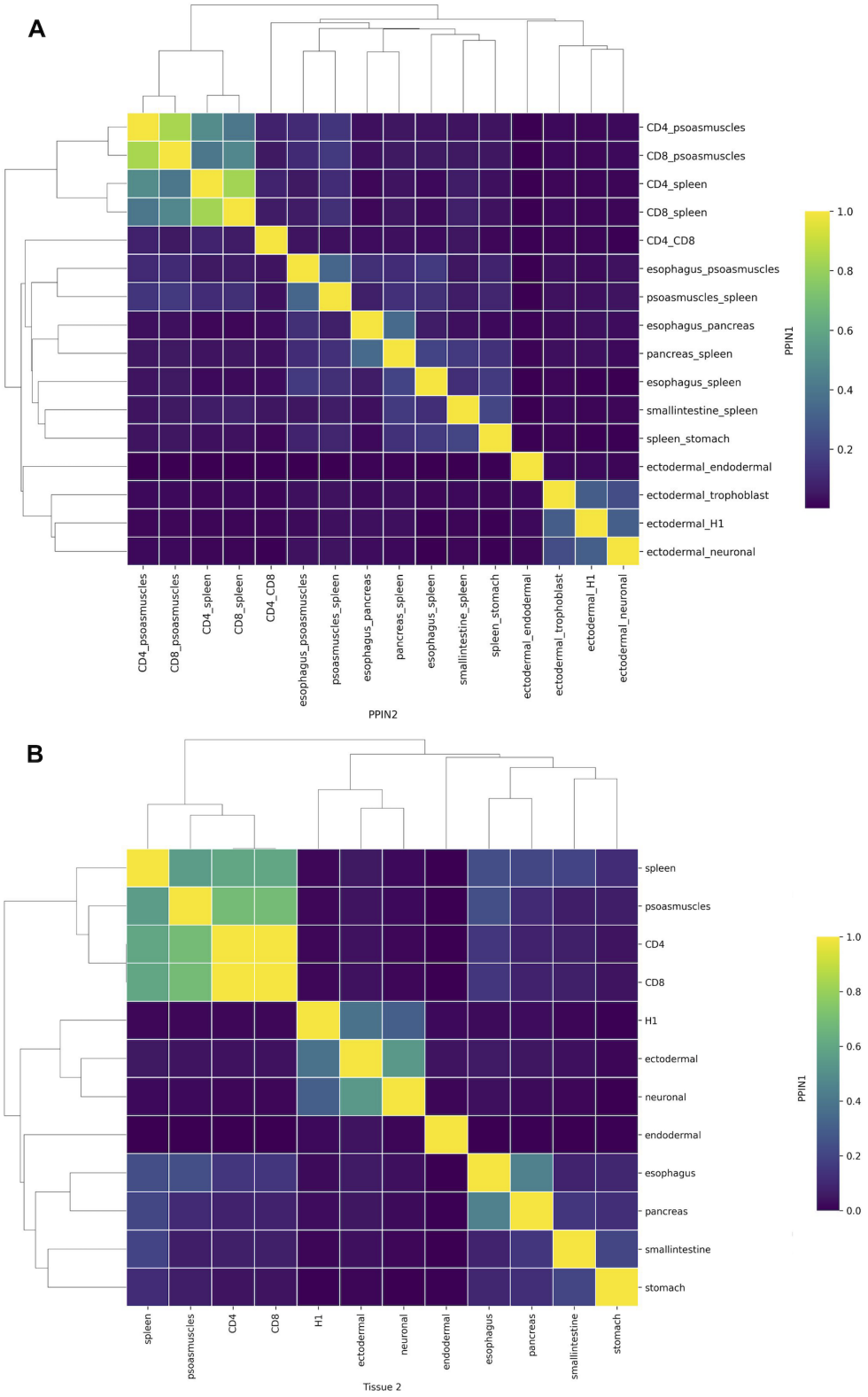

Figure C.2.: Similarity in differential RBPNs across pairwise RBPN comparisons and across cells/tissues of Human Epigenomes Atlas. A. Similarity in differential RBPNs computed by PPICompare. The differential RBPNs were computed for each pair of cells/tissues-dependent RBPNs, which were built using PPIXpress and transcriptomic data from the Human Epigenomes Atlas. *(to be continued)*

(continue) B. Similarity in cell/tissue-specific PPIs from differential RBPNs computed by PPI-Compare. For each cell/tissue, the set of PPIs from all differential RBPNs it involves in was collected. The similarities between differential RBPNs in A and the sets of cells/tissues-dependent PPIs in B were both computed by Jaccard similarity index using the number of overlapping and non-overlapping PPIs between them. Hierarchical clustering was used to cluster the differential RBPNs or cells/tissues-specific PPIs.

Bibliography

[1] G. Adami and L. E. Babiss. "DNA template effect on RNA splicing: two copies of the same gene in the same nucleus are processed differently". In: *EMBO J* 10.11 (1991), pp. 3457–65.

[2] M Aebi and C Weissman. "Precision and orderliness in splicing". In: *Trends in Genetics* 3 (1987), pp. 102–107.

[3] T. Afroz et al. "One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence". In: *Methods Enzymol* 558 (2015), pp. 235–278.

[4] Gael P. Alamancos, Eneritz Agirre, and Eduardo Eyras. "Methods to Study Splicing from High-Throughput RNA Sequencing Data". In: *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*. Ed. by Klemens J. Hertel. Totowa, NJ: Humana Press, 2014, pp. 357–397. ISBN: 978-1-62703-980-2.

[5] Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro, and Martin H. Schaefer. "HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks". In: *Nucleic Acids Research* 45.D1 (Oct. 2016), pp. D408–D414.

[6] M. Allo et al. "Chromatin and alternative splicing". In: *Cold Spring Harb Symp Quant Biol* 75 (2010), pp. 103–11.

[7] Patrick Aloy and Robert B Russell. "Structural systems biology: modelling protein interactions". In: *Nature reviews Molecular cell biology* 7.3 (2006), pp. 188–197.

[8] A. Dehghani Amirabad et al. "Transgenic expression of the RNA binding protein IMP2 stabilizes miRNA targets in murine microsteatosis". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1864.10 (2018), pp. 3099–3108.

[9] S. Anders, P. T. Pyl, and W. Huber. "HTSeq–a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (2015), pp. 166–9.

[10] S. Anders, A. Reyes, and W. Huber. "Detecting differential usage of exons from RNA-seq data". In: *Genome Res* 22.10 (2012), pp. 2008–17.

[11] L. Arribas-Hernández et al. "Principles of mRNA targeting via the m6A-binding protein ECT2". In: *eLife* 10 (2021), e72375.

[12] Madan M Babu. “Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks: 1”. In: *Protein Science* 21 (2012), p. 53.

[13] Gary D Bader and Christopher WV Hogue. “An automated method for finding molecular complexes in large protein interaction networks”. In: *BMC Bioinformatics* 4.1 (Jan. 2003).

[14] Akhilesh Kumar Bajpai et al. “Systematic comparison of the protein-protein interaction databases from a user’s perspective”. In: *Journal of Biomedical Informatics* 103 (2020), p. 103380.

[15] F. E. Baralle and J. Giudice. “Alternative splicing as a regulator of development and tissue identity”. In: *Nat Rev Mol Cell Biol* 18.7 (2017), pp. 437–451.

[16] Ruth Barshir et al. “Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases”. In: *PLoS computational biology* 10.6 (2014), e1003632.

[17] Robert A Barton and Chris Venditti. “Rapid evolution of the cerebellum in humans and other great apes”. In: *Current Biology* 24.20 (2014), pp. 2440–2444.

[18] Omer Basha et al. “The DifferentialNet database of differential protein–protein interactions in human tissues”. In: *Nucleic acids research* 46.D1 (2018), pp. D522–D526.

[19] Alexis Battle et al. “Impact of regulatory variation from RNA to protein”. In: *Science* 347.6222 (2015), pp. 664–667.

[20] T. W. Bebee et al. “The splicing regulators ESRP1 and ESRP2 direct an epithelial splicing program essential for mammalian development”. In: *Elife* 4 (2015).

[21] Rami Bechara et al. “The m6A reader IMP2 directs autoimmune inflammation through an IL-17- and TNFa-dependent C/EBP transcription factor axis”. In: *Science Immunology* 6.61 (2021), eabd1287.

[22] Leslie R Bell et al. “Positive autoregulation of Sex-lethal by alternative splicing maintains the female determined state in Drosophila”. In: *Cell* 65.2 (1991), pp. 229–239.

[23] Bérénice A Benayoun et al. “H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency”. In: *Cell* 158.3 (2014), pp. 673–688.

[24] Alejandro Berna-Erro et al. “STIM2 regulates capacitive Ca2+ entry in neurons and plays a key role in hypoxic neuronal cell death”. In: *Science signaling* 2.93 (2009), ra67–ra67.

[25] Yingtao Bi and Ramana V Davuluri. “NPEBseq: nonparametric empirical bayesian-based procedure for differential expression analysis of RNA-seq data”. In: *BMC bioinformatics* 14 (2013), pp. 1–12.

[26] Eva Bianconi et al. “An estimation of the number of cells in the human body”. In: *Annals of Human Biology* 40.6 (July 2013), pp. 463–471.

[27] Jeetayu Biswas et al. “The structural basis for RNA selectivity by the IMP family of RNA-binding proteins”. In: *Nature Communications* 10.1 (2019), p. 4440.

[28] L.A. Booth et al. “The role of cell signalling in the crosstalk between autophagy and apoptosis”. In: *Cellular Signalling* 26.3 (2014), pp. 549–555.

[29] Alice Bossi and Ben Lehner. “Tissue specificity and the human protein interaction network”. In: *Molecular Systems Biology* 5.1 (Jan. 2009).

[30] Daniel Bottomly et al. “Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays”. In: *PloS one* 6.3 (2011), e17820.

[31] Gregory D. Bowman and Michael G. Poirier. “Post-Translational Modifications of Histones That Influence Nucleosome Dynamics”. In: *Chemical Reviews* 115.6 (2015), pp. 2274–2295.

[32] Onn Brandman et al. “STIM2 is a feedback regulator that stabilizes basal cytosolic and endoplasmic reticulum Ca2+ levels”. In: *Cell* 131.7 (2007), pp. 1327–1339.

[33] Nicolas L Bray et al. “Near-optimal probabilistic RNA-seq quantification”. In: *Nature biotechnology* 34.5 (2016), pp. 525–527.

[34] Marija Buljan et al. “Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks”. In: *Molecular Cell* 46.6 (June 2012), pp. 871–883.

[35] Stephen K Burley et al. “RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences”. In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D437–D451.

[36] Matthew J. Burney et al. “An epigenetic signature of developmental potential in neural stem cells and early neurons”. In: *Stem Cells* 31.9 (2013), pp. 1868–1880.

[37] Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. “Mentha: a resource for browsing integrated protein-interaction networks”. In: *Nature methods* 10.8 (2013), pp. 690–691.

[38] Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. "mentha: a resource for browsing integrated protein-interaction networks". In: *Nature Methods* 10.8 (July 2013), pp. 690–691.

[39] Seth Carbon et al. "The Gene Ontology resource: enriching a GOld mine". In: *Nucleic Acids Research* 49.D1 (Dec. 2020), pp. D325–D334.

[40] Maiwen Caudron-Herger et al. "RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions". In: *Nucleic acids research* 49.D1 (2021), pp. D425–D436.

[41] George Chambers et al. "Proteomics: a new approach to the study of disease". In: *The Journal of pathology* 192.3 (2000), pp. 280–288.

[42] Natali L Chanaday et al. "Presynaptic store-operated Ca2+ entry drives excitatory spontaneous neurotransmission and augments endoplasmic reticulum stress". In: *Neuron* 109.8 (2021), pp. 1314–1332.

[43] J.A. Chao et al. "ZBP1 recognition of beta-actin zipcode induces RNA looping". In: *Genes Dev* 24 (2010), pp. 148–158.

[44] Chris Cheadle et al. "Stability Regulation of mRNA and the Control of Gene Expression". In: *Annals of the New York Academy of Sciences* 1058.1 (2005), pp. 196–204.

[45] Lu Chen et al. "Transcriptional diversity during lineage commitment of human blood progenitors". In: *Science* 345.6204 (Sept. 2014).

[46] Yeon Choi et al. "Time-resolved profiling of RNA binding proteins throughout the mRNA life cycle". In: *Molecular Cell* 84.9 (2024), pp. 1764–1782.

[47] Vasek Chvatal. "A greedy heuristic for the set-covering problem". In: *Mathematics of operations research* 4.3 (1979), pp. 233–235.

[48] Benjamin Cieply and Russ P Carstens. "Functional roles of alternative splicing factors in human disease". In: *Wiley Interdisciplinary Reviews: RNA* 6.3 (2015), pp. 311–326.

[49] A.E. Conway et al. "Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival". In: *Cell Rep* 15 (2016), pp. 666–679.

[50] M. P. Creyghton et al. "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21931–21936.

[51] Michael E Cusick et al. "Interactome: gateway into systems biology". In: *Human molecular genetics* 14.suppl_2 (2005), R171–R181.

[52] Beate Czepukojc et al. "IGF2 mRNA binding protein 2 transgenic mice are more prone to develop a ductular reaction and to progress toward cirrhosis". In: *Frontiers in medicine* 6 (2019), p. 179.

[53] C. Dahlem et al. “First small-molecule inhibitors targeting the RNA-binding protein IGF2BP2/IMP2 for cancer therapy”. In: *ACS Chem. Biol.* 17 (2022), pp. 361–375.

[54] P. Danecek et al. “Twelve years of SAMtools and BCFtools”. In: *Gigascience* 10 (2021).

[55] Basile Darbellay et al. “STIM1L is a new actin-binding splice variant involved in fast repetitive Ca2+ release”. In: *Journal of Cell Biology* 194.2 (2011), pp. 335–346.

[56] Carrie A. Davis et al. “The Encyclopedia of DNA elements (ENCODE): data portal update”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D794–D801.

[57] Jeremy Davis-Turak et al. “Genomics pipelines and data integration: challenges and opportunities in the research setting”. In: *Expert review of molecular diagnostics* 17.3 (2017), pp. 225–237.

[58] Manuel De la Mata, Celina Lafaille, and Alberto R Kornblihtt. “First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal”. In: *Rna* 16.5 (2010), pp. 904–912.

[59] N. Degrauwe et al. “The RNA Binding Protein IMP2 Preserves Glioblastoma Stem Cells by Preventing let-7 Target Gene Silencing”. In: *Cell Reports* 15.7 (2016), pp. 1634–1647.

[60] Paolo Di Tommaso et al. “Nextflow enables reproducible computational workflows”. In: *Nature biotechnology* 35.4 (2017), pp. 316–319.

[61] S. Djebali et al. “Landscape of transcription in human cells”. In: *Nature* 489.7414 (2012), pp. 101–8.

[62] A. Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29 (2013), pp. 15–21.

[63] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[64] Markus Draaken et al. “Involvement of the WNT and FGF signaling pathways in non-isolated anorectal malformations: sequencing analysis of WNT3A, WNT5A, WNT11, DACT1, FGF10, FGFR2 and the T gene”. In: *International Journal of Molecular Medicine* 30.6 (2012), pp. 1459–1464.

[65] Beatrice Dyring-Andersen et al. “Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin”. In: *Nature Communications* 11.1 (Nov. 2020).

[66] Fredrik Edfors et al. “Gene-specific correlation of RNA and protein levels in human cells and tissues”. In: *Molecular systems biology* 12.10 (2016), p. 883.

[67] Jonathan D Ellis et al. “Tissue-specific alternative splicing remodels protein-protein interaction networks”. In: *Molecular cell* 46.6 (2012), pp. 884–892.

[68] Jonathan D. Ellis et al. “Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks”. In: *Molecular Cell* 46.6 (June 2012), pp. 884–892.

[69] Ainara Elorza et al. “Huntington’s disease-specific mis-splicing unveils key effector genes and altered splicing factors”. In: *Brain* 144.7 (2021), pp. 2009–2023.

[70] Frank Emmert-Streib and Matthias Dehmer. “Understanding statistical hypothesis testing: The logic of statistical inference”. In: *Machine Learning and Knowledge Extraction* 1.3 (2019), pp. 945–962.

[71] S. Enroth et al. “Combinations of histone modifications mark exon inclusion levels”. In: *PLoS One* 7.1 (2012), e29911.

[72] Steffen Erkelenz et al. “Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms”. In: *Rna* 19.1 (2013), pp. 96–102.

[73] Philip Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–3048.

[74] Philip A Ewels et al. “The nf-core framework for community-curated bioinformatics pipelines”. In: *Nature biotechnology* 38.3 (2020), pp. 276–278.

[75] Y.-J. Fan and W.-X. Zong. “The cellular decision between apoptosis and autophagy”. In: *Chinese Journal of Cancer* 32.3 (2013), pp. 121–129.

[76] Jianxing Feng et al. “Identifying ChIP-seq enrichment using MACS”. In: *Nature protocols* 7.9 (2012), pp. 1728–1740.

[77] Robert D Finn et al. “iPfam: a database of protein family and domain interactions found in the Protein Data Bank”. In: *Nucleic acids research* 42.D1 (2014), pp. D364–D373.

[78] Robert D. Finn et al. “Pfam: the protein families database”. In: *Nucleic Acids Research* 42.D1 (Nov. 2013), pp. D222–D230.

[79] Marta Florio et al. “A single splice site mutation in human-specific ARHGAP11B causes basal progenitor amplification”. In: *Science advances* 2.12 (2016), e1601941.

[80] Marta Florio et al. “Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion”. In: *Science* 347.6229 (2015), pp. 1465–1470.

[81] Zhishuang Gao et al. “N6-methyladenosine-modified USP13 induces pro-survival autophagy and imatinib resistance via regulating the stabilization of autophagy-related protein 5 in gastrointestinal stromal tumors”. In: *Cell Death & Differentiation* 30.2 (2023), pp. 544–559.

[82] Daniel Garijo et al. “Quantifying reproducibility in computational biology: the case of the tuberculosis drugome”. In: *PloS one* 8.11 (2013), e80278.

[83] S. Gerstberger, M. Hafner, and T. Tuschl. “A census of human RNA-binding proteins”. In: *Nat Rev Genet* 15 (2014), pp. 829–845.

[84] Mohamed Ali Ghadie et al. “Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing”. In: *PLoS computational biology* 13.8 (2017), e1005717.

[85] M. Girvan and M. E. J. Newman. “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences* 99.12 (June 2002), pp. 7821–7826.

[86] Madalina Giurgiu et al. “CORUM: the comprehensive resource of mammalian protein complexes—2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D559–D563.

[87] Elena O Gracheva et al. “Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats”. In: *Nature* 476.7358 (2011), pp. 88–91.

[88] Markus Hafner et al. “Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP”. In: *Cell* 141.1 (2010), pp. 129–141.

[89] Le Han et al. “IGF2BP2 regulates MALAT1 by serving as an N6-methyladenosine reader to promote NSCLC proliferation”. In: *Frontiers in molecular biosciences* 8 (2022), p. 780089.

[90] S. Heinz et al. “Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities”. In: *Molecular Cell* 38.4 (2010), pp. 576–589.

[91] M.W. Hentze et al. “A brave new world of RNA-binding proteins”. In: *Nature Reviews Molecular Cell Biology* 19 (2018), pp. 327–341.

[92] Jorge A Holguin-Cruz, Leonard J Foster, and Jörg Gsponer. “Where protein structure and cell diversity meet”. In: *Trends in Cell Biology* 32.12 (2022), pp. 996–1007.

[93] Markus Hollander et al. “Detecting rewiring events in protein-protein interaction networks based on transcriptomic data”. In: *Frontiers in Bioinformatics* 1 (2021), p. 724297.

[94] Kirsty M Hooper et al. “V-ATPase is a universal regulator of LC3-associated phagocytosis and non-canonical autophagy”. In: *Journal of Cell Biology* 221.6 (2022), e202105112.

[95] Kevin L Howe et al. “Ensembl 2021”. In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D884–D891.

[96] Q. Hu, C. S. Greene, and E. A. Heller. “Specific histone modifications associate with alternative exon selection during mammalian development”. In: *Nucleic Acids Res* 48.9 (2020), pp. 4709–4724.

[97] Q. Hu et al. “Histone posttranslational modifications predict specific alternative exon subtypes in mammalian brain”. In: *PLoS Comput Biol* 13.6 (2017), e1005602.

[98] Yin Hu et al. “DiffSplice: the genome-wide detection of differential splicing events with RNA-seq”. In: *Nucleic acids research* 41.2 (2013), e39–e39.

[99] H. Huang et al. “Recognition of RNA N-methyladenosine by IGF2BP proteins enhances mRNA stability and translation”. In: *Nat. Cell Biol.* 20 (2018), pp. 285–295.

[100] Huilin Huang, Hengyou Weng, and Jianjun Chen. “The Biogenesis and Precise Control of RNA m6A Methylation”. In: *Trends in Genetics* 36.1 (2020), pp. 44–52.

[101] Zohar Itzhaki et al. “Evolutionary conservation of domain-domain interactions”. In: *Genome Biology* 7.12 (Dec. 2006).

[102] Michael Jackson, Kostas Kavoussanakis, and Edward WJ Wallace. “Using prototyping to choose a bioinformatics workflow management system”. In: *PLoS computational biology* 17.2 (2021), e1008622.

[103] Aishwarya G Jacob and Christopher WJ Smith. “Intron retention as a component of regulated gene expression programs”. In: *Human genetics* 136.9 (2017), pp. 1043–1057.

[104] Mahdi Jalili et al. “Unveiling network-based functional features through integration of gene expression into protein networks”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1864.6 (June 2018), pp. 2349–2359.

[105] M. Janiszewska et al. “Imp2 controls oxidative phosphorylation and is crucial for preserving glioblastoma cancer stem cells”. In: *Genes Dev* 26 (2012), pp. 1926–1944.

[106] Ronald Jansen, Dov Greenbaum, and Mark Gerstein. “Relating Whole-Genome Expression Data with Protein-Protein Interactions”. In: *Genome Research* 12.1 (Jan. 2002), pp. 37–46.

[107] Philip Jones et al. “InterProScan 5: genome-scale protein function classification”. In: *Bioinformatics* 30.9 (Jan. 2014), pp. 1236–1240.

[108] A. Kalsotra and T. A. Cooper. “Functional consequences of developmentally regulated alternative splicing”. In: *Nat Rev Genet* 12.10 (2011), pp. 715–29.

[109] Minoru Kanehisa et al. “KEGG: integrating viruses and cellular organisms”. In: *Nucleic Acids Research* 49.D1 (Oct. 2020), pp. D545–D551.

[110] Evans Kataka et al. “Edgetic perturbation signatures represent known and novel cancer biomarkers”. In: *Scientific reports* 10.1 (2020), p. 4350.

[111] S. Ke et al. “A majority of m6A residues are in the last exons, allowing the potential for 3’ UTR regulation”. In: *Genes & Development* 29.19 (2015), pp. 2037–2053.

[112] Xi-Song Ke et al. “Global profiling of histone and DNA methylation reveals epigenetic-based regulation of gene expression during epithelial to mesenchymal transition in prostate cells”. In: *BMC genomics* 11.1 (2010), pp. 1–15.

[113] Olga Kelemen et al. “Function of alternative splicing”. In: *Gene* 514.1 (2013), pp. 1–30.

[114] Sandra Kendzia et al. “A combined computational and functional approach identifies IGF2BP2 as a driver of chemoresistance in a wide array of preclinical models of colorectal cancer”. In: *Molecular Cancer* 22.1 (2023), p. 89.

[115] T. S. Keshava Prasad et al. “Human Protein Reference Database–2009 update”. In: *Nucleic Acids Research* 37.Database (Jan. 2009), pp. D767–D772.

[116] S.M. Kessler et al. “IGF2 mRNA binding protein p62/IMP2-2 in hepatocellular carcinoma: antiapoptotic action is independent of IGF2/PI3K signaling”. In: *Am J Physiol Gastrointest Liver Physiol* 304 (2013), G328–G336.

[117] S.M. Kessler et al. “IMP2/p62 induces genomic instability and an aggressive hepatocellular carcinoma phenotype”. In: *Cell Death Dis.* 6 (2015), e1894.

[118] Daehwan Kim et al. “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature biotechnology* 37.8 (2019), pp. 907–915.

[119] Yul Kim, Bumki Min, and Gwan-Su Yi. “IDDI: integrated domain-domain interaction and protein interaction analysis system”. In: *Proteome science*. Vol. 10. BioMed Central. 2012, pp. 1–9.

[120] E. de Klerk and P. A. t Hoen. “Alternative mRNA transcription, processing, and translation: insights from RNA sequencing”. In: *Trends Genet* 31.3 (2015), pp. 128–39.

[121] Mona L Knapp et al. “A longer isoform of Stim1 is a negative SOCE regulator but increases cAMP-modulated NFAT signaling”. In: *EMBO reports* 23.3 (2022), e53135.

[122] C. M. Koch et al. “The landscape of histone modifications across 1% of the human genome in five human cell lines”. In: *Genome Res* 17.6 (2007), pp. 691–707.

[123] Gavin C. K. W. Koh et al. “Analyzing Protein-Protein Interaction Networks”. In: *Journal of Proteome Research* 11.4 (Mar. 2012), pp. 2014–2031.

[124] S.M. Korn et al. “Structures and target RNA preferences of the RNA-binding protein family of IGF2BPs: An overview”. In: *Structure* 29.9 (2021), pp. 787–803.

[125] G. Koscielny et al. “ASTD: The Alternative Splicing and Transcript Diversity database”. In: *Genomics* 93.3 (2009), pp. 213–20.

[126] Johannes Köster and Sven Rahmann. “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19 (2012), pp. 2520–2522.

[127] A. Kouskouti and I. Talianidis. “Histone modifications defining active genes persist after transcriptional and mitotic inactivation”. In: *Embo j* 24.2 (2005), pp. 347–57.

[128] Kevin Kruse et al. “N-cadherin signaling via Trio assembles adherens junctions to restrict endothelial permeability”. In: *Journal of Cell Biology* 218.1 (Nov. 2018), pp. 299–316.

[129] Izabella Krystkowiak, Jean Manguy, and Norman E Davey. “PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants”. In: *Nucleic Acids Research* 46.W1 (June 2018), W235–W241.

[130] Manfred Kunz et al. “RNA-seq analysis identifies different transcriptomic types and developmental trajectories of primary melanomas”. In: *Oncogene* 37.47 (2018), pp. 6136–6151.

[131] S. Laggai et al. “The IGF2 mRNA binding protein p62/IGF2BP2-2 induces fatty acid elongation as a critical feature of steatosis”. In: *J Lipid Res* 55 (2014), pp. 1087–1097.

[132] Arash Latifkar et al. “IGF2BP2 promotes cancer progression by degrading the RNA transcript encoding a v-ATPase subunit”. In: *Proceedings of the National Academy of Sciences* 119.45 (2022), e2200477119.

[133] D. Lee et al. “Epigenome-based splicing prediction using a recurrent neural network”. In: *PLoS Comput Biol* 16.6 (2020), e1008006.

[134] Sung-Hun Lee and Young Zoon Kim. “DNA and Histone Methylation in Brain Cancer”. In: *DNA and Histone Methylation as Cancer Targets* (2017), pp. 347–376.

[135] W. van Leeuwen et al. “Identification of the stress granule transcriptome via RNA-editing in single cells and”. In: *Cell Rep Methods* 2 (2022), p. 100235.

[136] Ben Lehner and Andrew G Fraser. “A first-draft human protein-interaction map”. In: *Genome Biology* 5.9 (Aug. 2004).

[137] Rasko Leinonen et al. “The sequence read archive”. In: *Nucleic acids research* 39.suppl_1 (2010), pp. D19–D21.

[138] Bo Li and Colin N Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC bioinformatics* 12 (2011), pp. 1–16.

[139] H. Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–9.

[140] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[141] Ning Li et al. “IGF2BP2 modulates autophagy and serves as a prognostic marker in glioma”. In: *ibrain* 10.1 (2024), pp. 19–33.

[142] Luana Licata et al. “MINT, the molecular interaction database: 2012 update”. In: *Nucleic Acids Research* 40.D1 (Nov. 2011), pp. D857–D861.

[143] Susan J Lindsay et al. “HDBR expression: a unique resource for global and individual gene expression studies during early human brain development”. In: *Frontiers in neuroanatomy* 10 (2016), p. 86.

[144] H. Liu et al. “Histone modifications involved in cassette exon inclusions: a quantitative and interpretable analysis”. In: *BMC Genomics* 15 (2014), p. 1148.

[145] M. Liu et al. “TRIBE Uncovers the Role of Dis3 in Shaping the Dynamic Transcriptome in Malaria Parasites”. In: *Front Cell Dev Biol* 7 (2019), p. 264.

[146] Yunshan Liu et al. “The dual roles of MAGE-C2 in p53 ubiquitination and cell proliferation through E3 ligases MDM2 and TRIM28”. In: *Frontiers in Cell and Developmental Biology* 10 (2022), p. 922675.

[147] John Lonsdale et al. “The Genotype-Tissue Expression (GTEx) project”. In: *Nature Genetics* 45.6 (May 2013), pp. 580–585.

[148] Tiago J. S. Lopes et al. “Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases”. In: *Bioinformatics* 27.17 (July 2011), pp. 2414–2421.

[149] 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.

[150] 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 (2014), pp. 1–21.

[151] 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 (2014), p. 550.

[152] R. F. Luco et al. “Epigenetics in alternative pre-mRNA splicing”. In: *Cell* 144.1 (2011), pp. 16–26.

[153] R. F. Luco et al. “Regulation of alternative splicing by histone modifications”. In: *Science* 327.5968 (2010), pp. 996–1000.

[154] V. de M. Ramos, A.J. Kowaltowski, and P.A. Kakimoto. “Autophagy in Hepatic Steatosis: A Structured Review”. In: *Frontiers in Cell and Developmental Biology* 9 (2021), p. 657389.

[155] Q. Ma et al. “Wnt/β-catenin signaling pathway-a versatile player in apoptosis and autophagy”. In: *Biochimie* 211 (2023), pp. 57–67.

[156] Wenji Ma, Craig McAnulla, and Lusheng Wang. “Protein complex prediction based on maximum matching with domain-domain interaction”. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1824.12 (Dec. 2012), pp. 1418–1424.

[157] Luciano E Marasco and Alberto R Kornblihtt. “The physiology of alternative splicing”. In: *Nature Reviews Molecular Cell Biology* 24.4 (2023), pp. 242–254.

[158] G. Mariño et al. “Self-consumption: the interplay of autophagy and apoptosis”. In: *Nature Reviews Molecular Cell Biology* 15 (2014), pp. 81–94.

[159] M. Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet J.* 17 (2011), p. 10.

[160] A. Gregory Matera and Zefeng Wang. “A day in the life of the spliceosome”. In: *Nature Reviews Molecular Cell Biology* 15.2 (2014), pp. 108–121.

[161] Maxime Mazille et al. “Stimulus-specific remodeling of the neuronal transcriptome through nuclear intron-retaining transcripts”. In: *The EMBO journal* 41.21 (2022), e110192.

[162] A.C. McMahon et al. “TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins”. In: *Cell* 165 (2016), pp. 742–753.

[163] Arfa Mehmood et al. “Systematic evaluation of differential splicing tools for RNA-seq studies”. In: *Briefings in Bioinformatics* 21.6 (2020), pp. 2052–2065.

[164] Huaiyu Mi et al. “Large-scale gene function analysis with the PANTHER classification system”. In: *Nature protocols* 8.8 (2013), pp. 1551–1566.

[165] Giovanni Micale et al. “SPECTRA: an integrated knowledge base for comparing tissue and tumor-specific PPI networks in human”. In: *Frontiers in Bioengineering and Biotechnology* 3 (2015), p. 58.

[166] Anna-Maria Miederer et al. “A STIM2 splice variant negatively regulates store-operated calcium entry”. In: *Nature communications* 6.1 (2015), p. 6899.

[167] Tarjei S. Mikkelsen et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells”. In: *Nature* 448.7153 (2007), pp. 553–560.

[168] A. A. Mironov, J. W. Fickett, and M. S. Gelfand. “Frequent alternative splicing of human genes”. In: *Genome Res* 9.12 (1999), pp. 1288–93.

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

[170] Felix Mölder et al. “Sustainable data analysis with Snakemake”. In: *F1000Research* 10 (2021), p. 33.

[171] Andrei Molotkov et al. “Distinct Requirements for FGFR1 and FGFR2 in Primitive Endoderm Development and Exit from Pluripotency”. In: *Developmental Cell* 41.5 (2017), 511–526.e4.

[172] 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.

[173] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. “Detecting overlapping protein complexes in protein-protein interaction networks”. In: *Nature Methods* 9.5 (Mar. 2012), pp. 471–472.

[174] G. Nicastro et al. “Direct m6A recognition by IMP1 underlays an alternative model of target selection for non-canonical methyl-readers”. In: *Nucleic Acids Research* 51.15 (2023), pp. 8774–8786.

[175] J. Nielsen et al. “A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development”. In: *Mol Cell Biol* 19 (1999), pp. 1262–1270.

[176] Barbara A Niemeyer. “Changing calcium: CRAC channel (STIM and Orai) expression, splicing, and posttranslational modifiers”. In: *American Journal of Physiology-Cell Physiology* 310.9 (2016), pp. C701–C709.

[177] T. W. Nilsen and B. R. Graveley. “Expansion of the eukaryotic proteome by alternative splicing”. In: *Nature* 463.7280 (2010), pp. 457–63.

[178] N.A. O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Res.* 44 (2016), pp. D733–D745.

[179] TI Orban and E Olah. “Emerging roles of BRCA1 alternative splicing”. In: *Molecular Pathology* 56.4 (2003), p. 191.

[180] Sandra Orchard et al. “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases”. In: *Nucleic acids research* 42.D1 (2014), pp. D358–D363.

[181] Rose Oughtred et al. “The <scp>BioGRID</scp> database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions”. In: *Protein Science* 30.1 (Nov. 2020), pp. 187–200.

[182] Yosuke Ozawa et al. “Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions”. In: *BMC Bioinformatics* 11.1 (June 2010).

[183] Philipp Pagel et al. “The MIPS mammalian protein-protein interaction database”. In: *Bioinformatics* 21.6 (Nov. 2004), pp. 832–834.

[184] Sharmistha Pal et al. “Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development”. In: *Genome research* 21.8 (2011), pp. 1260–1272.

[185] Qun Pan et al. “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing”. In: *Nature Genetics* 40.12 (Nov. 2008), pp. 1413–1415.

[186] V.L. Patel et al. “Spatial arrangement of an RNA zipcode identifies mRNAs under post-transcriptional control”. In: *Genes Dev* 26 (2012), pp. 43–53.

[187] R. Patro et al. “Salmon provides fast and bias-aware quantification of transcript expression”. In: *Nat. Methods* 14 (2017), pp. 417–419.

[188] Rob Patro et al. “Salmon provides fast and bias-aware quantification of transcript expression”. In: *Nature methods* 14.4 (2017), pp. 417–419.

[189] E. Pawlowska et al. “NF-κB-Mediated Inflammation in the Pathogenesis of Intracranial Aneurysm and Subarachnoid Hemorrhage. Does Autophagy Play a Role?” In: *International Journal of Molecular Sciences* 19.4 (2018), p. 1245.

[190] Mihaela Pertea et al. “CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise”. In: *Genome Biology* 19.1 (Nov. 2018).

[191] A. Petiot. “Development of the mammalian urethra is controlled by Fgfr2-IIIb”. In: *Development* 132.10 (2005), pp. 2441–2450.

[192] Janet D Pierce et al. “Understanding proteomics”. In: *Nursing & health sciences* 9.1 (2007), pp. 54–60.

[193] Allison Piovesan et al. “Human protein-coding genes and gene feature statistics in 2019”. In: *BMC Research Notes* 12.1 (June 2019).

[194] O. Podlaha et al. “Histone modifications are associated with transcript isoform diversity in normal and cancer cells”. In: *PLoS Comput Biol* 10.6 (2014), e1003611.

[195] Elena Popugaeva et al. “STIM2 protects hippocampal mushroom spines from amyloid synaptotoxicity”. In: *Molecular neurodegeneration* 10 (2015), pp. 1–13.

[196] Murali Prakriya and Richard S Lewis. "Store-operated calcium channels". In: *Physiological reviews* (2015).

[197] John Quackenbush. "Computational analysis of microarray data". In: *Nature reviews genetics* 2.6 (2001), pp. 418–427.

[198] A. R. Quinlan and I. M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–2.

[199] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[200] Reazur Rahman et al. "Identification of RNA-binding protein targets with HyperTRIBE". In: *Nature protocols* 13.8 (2018), pp. 1829–1849.

[201] Girish Ramesh et al. "A short isoform of STIM1 confers frequency-dependent synaptic enhancement". In: *Cell Reports* 34.11 (2021).

[202] Deepthi Ramesh-Kumar and Sonia Guil. "The IGF2BP family of RNA binding proteins links epitranscriptomics to cancer". In: *Seminars in Cancer Biology*. Vol. 86. Elsevier. 2022, pp. 18–31.

[203] Anshul Rana et al. "Alternative splicing converts STIM2 from an activator to an inhibitor of store-operated calcium channels". In: *Journal of Cell Biology* 209.5 (2015), pp. 653–670.

[204] L. Regué et al. "Liver-specific deletion of IGF2 mRNA binding protein-2/IMP2 reduces hepatic fatty acid oxidation and increases hepatic triglyceride accumulation". In: *J Biol Chem* 294 (2019), pp. 11944–11951.

[205] L. Regué et al. "RNA m6A reader IMP2/IGF2BP2 promotes pancreatic β-cell proliferation and insulin secretion by enhancing PDX1 expression". In: *Molecular Metabolism* 48 (2021), p. 101209.

[206] Arno Reich et al. "Fas/CD95 Regulatory Protein Faim2 Is Neuroprotective after Transient Brain Ischemia". In: *The Journal of Neuroscience* 31.1 (Jan. 2011), pp. 225–233.

[207] Yan Ren et al. "Long read isoform sequencing reveals hidden transcriptional complexity between cattle subspecies". In: *BMC genomics* 24.1 (2023), p. 108.

[208] Robert Riley et al. "Inferring protein domain interactions from databases of interacting proteins". In: *Genome Biology* 6.10 (Sept. 2005).

[209] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[210] Consortium Roadmap Epigenomics et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–30.

[211] 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.

[212] Thomas Rolland et al. "A Proteome-Scale Map of the Human Interactome Network". In: *Cell* 159.5 (Nov. 2014), pp. 1212–1226.

[213] Barbora Salovska et al. "Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation". In: *Molecular systems biology* 16.3 (2020), e9170.

[214] L. Salwinski. "The Database of Interacting Proteins: 2004 update". In: *Nucleic Acids Research* 32.90001 (Jan. 2004), pp. 449D–451.

[215] T. Schneider et al. "Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3". In: *Nat Commun* 10 (2019), p. 2266.

[216] I. E. Schor et al. "Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing". In: *Proc Natl Acad Sci U S A* 106.11 (2009), pp. 4325–30.

[217] Björn Schwanhäusser et al. "Global quantification of mammalian gene expression control". In: *Nature* 473.7347 (2011), pp. 337–342.

[218] Alexandre Segelle et al. "Histone marks regulate the epithelial-to-mesenchymal transition via alternative splicing". In: *Cell Reports* 38.7 (2022), p. 110357.

[219] Z. Shao et al. "MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets". In: *Genome Biol* 13.3 (2012), R16.

[220] D. Sharma et al. "A review of the tortuous path of nonviral gene delivery and recent progress". In: *Int. J. Biol. Macromol.* 183 (2021), pp. 2055–2073.

[221] 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.

[222] Anita Siller et al. "β2-subunit alternative splicing stabilizes Cav2. 3 Ca2+ channel activity during continuous midbrain dopamine neuron-like activity". In: *Elife* 11 (2022), e67464.

[223] Yvette Simon et al. "Elevated free cholesterol in a p62 overexpression model of non-alcoholic steatohepatitis". In: *World journal of gastroenterology: WJG* 20.47 (2014), p. 17839.

[224] Malvindar K Singh-Bains et al. "Cerebellar degeneration correlates with motor symptoms in Huntington disease". In: *Annals of neurology* 85.3 (2019), pp. 396–405.

[225] Maurice A Smith, Jason Brandt, and Reza Shadmehr. “Motor disorder in Huntington’s disease begins as a dysfunction in error feedback control”. In: *Nature* 403.6769 (2000), pp. 544–549.

[226] T. Smith, A. Heger, and I. Sudbery. “UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy”. In: *Genome Res.* 27 (2017), pp. 491–499.

[227] Xiao Song et al. “Post-transcriptional regulation of cancer/testis antigen MAGEC2 expression by TRIM28 in tumor cells”. In: *BMC cancer* 18 (2018), pp. 1–10.

[228] Natalia Soshnikova and Denis Duboule. “Epigenetic Temporal Control of Mouse Hox Genes in Vivo”. In: *Science* 324.5932 (2009), p. 1320.

[229] Rory Stark, Gordon Brown, et al. “DiffBind: differential binding analysis of ChIP-Seq peak data”. In: *R package version* 100.4.3 (2011), pp. 2–21.

[230] Peter B Stathopulos, Le Zheng, and Mitsuhiko Ikura. “Stromal interaction molecule (STIM) 1 and STIM2 calcium sensing regions exhibit distinct unfolding and oligomerization kinetics”. In: *Journal of Biological Chemistry* 284.2 (2009), pp. 728–732.

[231] François Stricher et al. “HSPA8/HSC70 chaperone protein: structure, function, and chemical targeting”. In: *Autophagy* 9.12 (2013), pp. 1937–1954.

[232] Suya Sun et al. “Reduced synaptic STIM2 expression and impaired store-operated calcium entry cause destabilization of mature spines in mutant presenilin mice”. In: *Neuron* 82.1 (2014), pp. 79–93.

[233] Gustav N Sundell et al. “Proteome-wide analysis of phospho-regulated <scp>PDZ</scp> domain interactions”. In: *Molecular Systems Biology* 14.8 (Aug. 2018).

[234] Damian Szklarczyk et al. “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. In: *Nucleic acids research* 47.D1 (2019), pp. D607–D613.

[235] 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 (Nov. 2020), pp. D605–D612.

[236] Magdalena Szumilas. “Explaining odds ratios”. In: *Journal of the Canadian academy of child and adolescent psychiatry* 19.3 (2010), p. 227.

[237] Yining Tao et al. “Alternative splicing and related RNA binding proteins in human health and disease”. In: *Signal transduction and targeted therapy* 9.1 (2024), p. 26.

[238] Elie Marcel Teyssonnière et al. “Species-wide quantitative transcriptomes and proteomes reveal distinct genetic control of gene expression variation in yeast”. In: *Proceedings of the National Academy of Sciences* 121.19 (2024), e2319211121.

[239] Cole Trapnell et al. “Differential analysis of gene regulation at transcript resolution with RNA-seq”. In: *Nature biotechnology* 31.1 (2013), pp. 46–53.

[240] Cole Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. In: *Nature protocols* 7.3 (2012), pp. 562–578.

[241] A. Trocoli and M. Djavaheri-Mergny. “The complex interplay between autophagy and NF-B signaling pathways in cancer cells”. In: *American Journal of Cancer Research* 1.6 (2011), pp. 629–649.

[242] Elisabeth Tybl et al. “Overexpression of the IGF2-mRNA binding protein p62 in transgenic mice induces a steatotic phenotype”. In: *Journal of hepatology* 54.5 (2011), pp. 994–1001.

[243] Jernej Ule and Benjamin J Blencowe. “Alternative splicing regulatory networks: functions, mechanisms, and evolution”. In: *Molecular cell* 76.2 (2019), pp. 329–345.

[244] Uniprot. “UniProt: the universal protein knowledgebase in 2023”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D523–D531.

[245] E.L. Van Nostrand et al. “A large-scale binding and functional map of human RNA-binding proteins”. In: *Nature* 583 (2020), pp. 711–719.

[246] 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.

[247] Amrita Verma et al. “In silico comparative analysis of LRRK2 interactomes from brain, kidney and lung”. In: *Brain Research* 1765 (Aug. 2021), p. 147503.

[248] Jukka-Pekka Verta and Arne Jacobs. “The role of alternative splicing in adaptation and evolution”. In: *Trends in Ecology & Evolution* 37.4 (2022), pp. 299–308.

[249] Vladimir A Vigont et al. “STIM2 mediates excessive store-operated calcium entry in patient-specific iPSC-derived neurons modeling a juvenile form of huntington’s disease”. In: *Frontiers in cell and developmental biology* 9 (2021), p. 625231.

[250] Christine Vogel and Edward M Marcotte. “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses”. In: *Nature reviews genetics* 13.4 (2012), pp. 227–232.

[251] K. Wächter et al. “Subcellular localization and RNP formation of IGF2BPs (IGF2 mRNA-binding proteins) is modulated by distinct RNA-binding domains”. In: *Biol Chem* 394 (2013), pp. 1077–1090.

[252] Dongxue Wang et al. “A deep proteome and transcriptome abundance atlas of 29 healthy human tissues”. In: *Molecular systems biology* 15.2 (2019), e8503.

[253] E. T. Wang et al. “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221 (2008), pp. 470–6.

[254] Jinyan Wang, Lijuan Chen, and Ping Qiang. “The role of IGF2BP2, an m6A reader gene, in human metabolic diseases and cancers”. In: *Cancer cell international* 21.1 (2021), p. 99.

[255] Qingqing Wang et al. “PQBP1, a factor linked to intellectual disability, affects alternative splicing associated with neurite outgrowth”. In: *Genes & development* 27.6 (2013), pp. 615–626.

[256] John N Weinstein et al. “The cancer genome atlas pan-cancer analysis project”. In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[257] S. M. Weyn-Vanhentenryck et al. “Precise temporal regulation of alternative splicing during neural development”. In: *Nat Commun* 9.1 (2018), p. 2189.

[258] M. E. Wilkinson, C. Charenton, and K. Nagai. “RNA Splicing by the Spliceosome”. In: *Annu Rev Biochem* 89 (2020), pp. 359–388.

[259] 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.

[260] Thorsten Will and Volkhard Helms. “Rewiring of the inferred protein interactome during blood development studied with the tool PPICompare”. In: *BMC Systems Biology* 11.1 (2017), pp. 1–19.

[261] Richard T Williams et al. “Identification and characterization of the STIM (stromal interaction molecule) gene family: coding for a novel class of transmembrane proteins”. In: *Biochemical Journal* 357.3 (2001), pp. 673–685.

[262] Rebecca Woelfle, Andrea L. D’Aquila, and David A. Lovejoy. “Teneurins, TCAP, and latrophilins: roles in the etiology of mood disorders”. In: *Translational Neuroscience* 7.1 (Jan. 2016), pp. 17–23.

[263] Tianzhi Wu et al. “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data”. In: *The Innovation* 2.3 (2021), p. 100141.

[264] Jiansheng Xie et al. “Identification of a STIM1 splicing variant that promotes glioblastoma growth”. In: *Advanced Science* 9.11 (2022), p. 2103940.

[265] Lei Xing et al. “Expression of human-specific ARHGAP11B in mice leads to neocortex expansion and increased memory flexibility”. In: *The EMBO journal* 40.13 (2021), e107093.

[266] Wen Xu, Rashedul Islam, and Michael Rosbash. “Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein”. In: *RNA* 24.2 (2018), pp. 173–182.

[267] X. Xu et al. “Up-regulation of IGF2BP2 by multiple mechanisms in pancreatic cancer promotes cancer proliferation by activating the PI3K/Akt signaling pathway”. In: *Journal of Experimental & Clinical Cancer Research* 38 (2019), p. 497.

[268] Cheng Yang et al. “LncADeep: anab initiolncRNA identification and functional annotation tool based on deep learning”. In: *Bioinformatics* 34.22 (May 2018). Ed. by Inanc Birol, pp. 3825–3834.

[269] Xinping Yang et al. “Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing”. In: *Cell* 164.4 (Feb. 2016), pp. 805–817.

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

[271] Esti Yeger-Lotem and Roded Sharan. “Human protein interaction networks across tissues and diseases”. In: *Frontiers in Genetics* 6 (Aug. 2015).

[272] 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.

[273] Sora Yoon et al. “Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2”. In: *Nucleic Acids Research* 46.10 (Mar. 2018), e60–e60.

[274] G. Yu et al. “clusterProfiler: an R package for comparing biological themes among gene clusters”. In: *OMICS* 16 (2012), pp. 284–287.

[275] Liping Zeng and Bailong Zhang. “Exploring the Genomic Landscape: An In-depth ChIP-seq Analysis Protocol for Uncovering Protein-DNA Interactions”. In: *Current Protocols* 3.10 (2023), e909.

[276] F. Zhao et al. “Insulin-like growth factor 2 mRNA-binding protein 2-regulated alternative splicing of nuclear factor 1 C-type causes excessive granulosa cell proliferation in polycystic ovary syndrome”. In: *Cell Proliferation* 55.1 (2022), e13216.

[277] Shanrong Zhao et al. “Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion”. In: *Scientific Reports* 8.1 (2018), p. 4781.

[278] Z. Zheng et al. “A computational method for studying the relation between alternative splicing and DNA methylation”. In: *Nucleic Acids Res* 44.2 (2016), e19.

[279] H. L. Zhou et al. “Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms”. In: *Nucleic Acids Res* 42.2 (2014), pp. 701–13.

[280] Y. Zhou, Y. Lu, and W. Tian. “Epigenetic features are significantly associated with alternative splicing”. In: *BMC Genomics* 13 (2012), p. 123.

[281] Ying Zhu et al. “NF-κB is involved in the regulation of autophagy in mutant p53 cells in response to ionizing radiation”. In: *Cell Death Discovery* 7.1 (2021), p. 159.

[282] Yazeed Zoabi and Noam Shomron. “Processing and analysis of RNA-seq data from public resources”. In: *Deep Sequencing Data Analysis* (2021), pp. 81–94.