UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Ciências ULisboa

A systems approach to the mechanisms of neurodegeneration

Doutoramento em Biologia
Especialidade de Biologia de Sistemas

Marina Luque García-Vaquero

Tese orientada por:

Prof. Doutora Margarida Henriques da Gama Carvalho
Doutor Javier De Las Rivas

21 de dezembro de 2022
Lisboa - San Sebastián - Salamanca

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

A systems approach to the mechanisms of neurodegeneration

Doutoramento em Biologia
Especialidade de Biologia de Sistemas

Marina Luque García-Vaquero

Tese orientada por:
Proffesora Doutora Margarida Henriques da Gama Carvalho e
Doutor Javier De Las Rivas

Documento especialmente elaborado para a obtenção do grau de doutor

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

A systems approach to the mechanisms of neurodegeneration

Doutoramento em Biologia
Especialidade de Biologia de Sistemas

Marina Luque García-Vaquero

Tese orientada por:

Proffesora Doutora Margarida Henriques da Gama Carvalho e

Doutor Javier De Las Rivas

Júri:
Presidente:

Doutor Rui Manuel dos Santos Malhó, Professor Catedrático e Presidente do Departamento de Biologia Vegetal, da Faculdade de Ciências da Universidade de Lisboa

Vogais:

Doutora Anais Baudot, Investigadora Coordenadora Faculté de Médicine da Université Aix-Marseille

Doutora Ana Rita Fialho Grosso, Professora Auxiliar Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Doutor Clévio David Rodrigues Nóbrega, Professor Auxiliar Faculdade de Medicina e Ciências Biomédicas da Universidade do Algarve

Doutor Nuno Luis Barbosa Morais, Professor Associado Convidado Faculdade de Medicina da Universidade de Lisboa

Doutora Margarida Henriques da Gama Carvalho, Professora Auxiliar Faculdade de Ciências da Universidade de Lisboa, Orientadora

Fundação para a Ciência e Tecnologia do Ministério da Educação e Ciência,

PD/BD/128109/2016

2022

Acknowledgments

To my dearest supervisors: Margarida and Javier have been wonderful mentors from all perspectives. Your hard work, integrity and patience in teaching are not that common. These words are equally true for Francisco who invested countless hours teaching me the basics of R and discussing with me the design details of my analyses. It is has been a great luck to learn from you to become a 'young scientist'. You gave me the time to err and learn from my mistakes while enjoying the amazing time I had in Lisbon.

And speaking of Lisbon, what to say about that place? It has been six years of a great adventure. I would like to thank my helpful mates at RNA Systems Biology lab, my partners in crime in BioSYS PhD program and the lovely folks from C2-HORTA FCUL. Outside the campus and mainly during the weekends, I met more friends who made those years truly special. Filipa, Gonçalo, Nocas, Paolo, Giovanni, Ana Rita or Ricardo are some of their names.

Coming back home. To my mom Alicia and dad Javier who gave me the opportunity to study Biology without putting up too much resistance. From them I inherited the creativity and humor that I consider myself to have. To my brother Pedro, from instilled in me a bottomless curiosity. Thanks to my hometown mates, for their absurd sense of humor and predilection for drinking beers on the pier, to my girlfriends for the sisterhood and affection since early years at school, and to 'Hipoparemesishhh' for the creative ways to waste time during the first years at the college. 'Elmer-cuentro' are my anchor to Madrid and made my decision to come back to the capital much easier. Salamanca is a special place too. There, we few shared pandemic days together. Alberto, Santi and Angi, thanks for keeping my emotional stability afloat.

A most special gratitude is to Marcin, an extraordinary person with whom I shared this and many other adventures. We have enjoyed the sunny days to the fullest and we have supported each other in the hardest hours. Thank you all.

Agradecimientos

Aos meus caros supervisores, Margarida e Javier foram mentores maravilhosos. Seu trabalho duro, integridade e paciência em me ensinar não são tão comuns. Essas palavras são também para Francisco, que passou incontáveis horas me ensinando o básico de R e discutindo comigo os desenho dos nossos análises. Foi uma grande oportunidade aprender com vocês para me tornar uma 'jovem cientista'. Vocês me deram tempo para cometer erros e aprender com eles enquanto aproveitava o tempo maravilhoso que passei em Lisboa.

E o que posso dizer de Lisboa? Foram seis anos de uma grande aventura. Gostaria de agradecer aos meus companheiros no laboratório de RNA Biologia de Sistemas, meus parceiros no crime no programa de doutoramento BioSYS e ao pessoal adorável do C2-HORTA FCUL. Fora do campus e principalmente nos finais de semana, conheci mais amigos que tornaram aqueles anos realmente especiais Filipa, Gonçalo, Nocas, Paolo, Giovanni, Ana Rita or Ricardo são alguns de seus nomes.

De vuelta a casa. A mi madre Alicia y a mi padre Javier quienes me dieron la oportunidad de estudiar Biología si oponer mucha resistencia. De ellos he heredado la creatividad y humor que considero tener. A mi hermano Pedro, del cual he recibido una curiosidad inagotable. Gracias a mis amigos de Donosti por vuestro absurdo sentido del humor y predilección por tomar cervezas en el puerto, a mis amigas por la sororidad y cariño desde los primeros años en el colegio y a 'Hipoparemesishhh' por nuestras creativas formas de perder el tiempo en Leioa. 'Elmer-cuentro' son mi anclaje a Madrid e hicieron que mi decisión de volver a la capital haya sido mucho más fácil. Salamanca es un lugar especial también. Allí hemos compartido los duros días de pandemia. Gracias Alberto, Santi y Angi por mantener a flote mi estabilidad emocional.

El agradecimiento mas especial es para Martín, una persona extraordinaria con la cual he compartido esta y otras muchas aventuras. Hemos disfrutado de los días al sol y nos hemos acompañado en las horas más duras. Obrigada/Gracias a todos.

Preface

The work presented in this thesis is the result from the research founded by my PhD fellowship (ref. PD/BD/128109/2016) from Fundação para a Ciência e Tecnología (FCT, Portugal) and funding assigned to Fly-SMALS project, an EU Joint Programme in Neurodegenerative Disease Research supported through the following funding organizations under the aegis of JPND (France/ANR; Germany/BMBF; Spain/ ISCIII; Portugal/FCT). The PhD project was integrated in the BioSYS PhD Programme coordinated by BioISI – Biosystems and Integrative Sciences Institute (FCT/PD/00065/2012) of the Faculty of Sciences, University of Lisboa (FCUL, Lisbon, Portugal). The PhD research was carried out between 2016 and 2022 at the RNA Systems Biology Lab, in BioISI-FCUL and at the Bioinformatics and Functional Genomics Research Group in Cancer Research Center (CiC-IBMCC, Salamanca, Spain), under the joint supervision of Prof. Dr. Margarida Gama-Carvalho (FCUL) and Dr. Javier De Las Rivas (CiC-IBMCC). This work is a direct continuation of the research line we began to explore in the context of my master thesis, in 2016, under the supervision of Prof. Dr. Francisco R. Pinto and Prof. Dr. Margarida Gama-Carvalho (FCUL), which has actively collaborated in the conceptualization of the works presented in this thesis.

Prefácio

Marina Luque García-Vaquero foi bolseira de doutoramento no âmbito do programa doutoral BioSys - Sistemas Biológicos, Genómica Funcional & Integrativa (FCT/PD/00065/2012) da Faculdade de Ciências da Universidade de Lisboa, financiada pela Fundação para a Ciência e Tecnologia do Ministério da Educação e Ciência, Bolsa PD/BD/128109/2016.

Resumo

As doenças dos neurónios motores (DNMs) caracterizam-se pela deterioração progressiva dos neurónios motores (NMs). Os NMs são um tipo específico de neurónios implicado na comunicação do sistema nervoso central com os músculos e outros sistemas periféricos, como glândulas e órgãos, controlando movimentos voluntários e involuntários. As DNM abrangem um espectro de condições degenerativas dos NMs, associadas a inúmeras alterações genéticas. A esclerose lateral amiotrófica (ELA) e a atrofia muscular espinhal proximal (AME) são os tipos de DNM mais frequentes, atraindo assim os maiores esforços de investigação. A ELA é uma doença neurodegenerativa predominantemente esporádica de adultos, enquanto a AME é uma condição hereditária que se manifesta durante os primeiros anos de vida. As últimas pesquisas indicam que a incidência média da ELA nos países europeus está entre 2,1 e 3,8 por 100.000 pessoas anualmente, enquanto que a incidência de AME é de 1 em 5,000 a 10,000 nados-vivos.

A degeneração dos NMs leva aos primeiros sintomas associados a fraqueza muscular indolor. A fraqueza muscular inicial nos membros estende-se de maneira contígua a outros grupos musculares. Os deficits funcionais progressivos levam a uma perda geral de independência. Nos últimos estádios da doença, a degeneração leva a uma incapacidade fatal para respirar e deglutir. Aproximadamente 50% das pessoas com ELA não sobrevivem mais de 3 anos após os primeiros sintomas e apenas cerca de 10% vivem 10 anos ou mais.

As causas predominantes da ELA hereditária incluem uma expansão hexanucleotídica do gene C9orf72 e mutações missense nos genes SOD1, TARDBP e FUS. Por outro lado, a AME é desencadeada por mutações que afetam a expressão da proteína SMN em aproximadamente 95% dos casos (5q-SMA). No entanto, genes adicionais também foram associados a fenótipos não 5q-SMA. É notável que os genes predominantes em DNMs estão frequentemente envolvidos em funções como metabolismo do RNA, tráfego de vesículas, em particular transporte axonal, mecanismos de degradação de proteínas, metabolismo das mitocôndrias e vias de reparação do DNA. A semelhança observada entre os subtipos de DNMs

sugere que a degeneração seletiva de NMs é desencadeada por mecanismos moleculares comuns. Também é notável que, embora as DNMs se manifestem predominantemente em NMs, uma grande fração dos genes causais seja expressa de forma ubíqua num grande número de tecidos e esteja envolvida em funções básicas para a homeostasia celular.

Embora a comunidade biomédica tenha amplo conhecimento dos genes envolvidos nas DNMs, ainda não conhecemos os eventos fundamentais que desencadeiam a degeneração do NM. A principal razão é que os sistemas biológicos são complexos, ou seja, compostos por um número muito elevado de elementos intimamente interrelacionados. A complexidade dos sistemas celulares dificulta assim a identificação das consequências das alterações genéticas, que podem desencadear respostas variáveis dependendo do contexto celular e genético. Esta limitação é critica, sendo que a falta de uma definição detalhada da patogénese das DNMs impossibilita o desenvolvimento de ferramentas para o diagnóstico precoce e opções terapêuticas efetivas.

A biologia de sistemas visa abordar os desafios biomédicos, capturando, em vez de reduzir, a complexidade do sistema biológico de interesse. Por outras palavras, a biologia de sistemas pretende estudar o modo como a relação entre os elementos de um sistema complexo resulta nos fenómenos biológicos observados na natureza. É assim que o primeiro passo para desenvolver estratégias de investigação biomédica na perspetiva da biologia de sistemas é a construção de modelos biológicos detalhados. Para esse fim, as abordagens bioquímicas de alto rendimento, ou abordagens “ómicas” tem sido fundamentais para gerar bancos de dados de larga escala de livre acesso. Entre estes dados temos os que descrevem as interações físicas entre as proteínas (interatoma), fundamentais para coordenar os processos biológicos. Igualmente, a caracterização dos transcritomas específicos de células e tecidos é critica para restringir os estudos aos genes expressos em cada contexto biológico. Os catálogos públicos de funções e associações a doenças são de grande utilidade para caracterizar e prioritizar os transcritos e proteínas identificados em investigações experimentais ou computacionais.

O estudo de dados “ómicos” requer métodos analíticos especializados. Por exemplo, a representação de interações biomoleculares em redes facilita a integração de dados complexos e a aplicação de conceitos da teoria de redes para prever fenómenos biológicos. Nomeadamente, a topologia das redes biológicas descreve o arranjo das biomoléculas que interagem na rede e fornece informações valiosas para interpretar as propriedades biológicas do sistema, tanto no seu conjunto, como dos elementos que o compõem. Por exemplo, a importância de uma molécula para a atividade de um determinado sistema pode ser inferida a partir do número e tipo de interações que tem em seu redor na rede. Da mesma forma, a caracterização de perturbações hipotéticas na rede biológica permite-nos antecipar resultados patológicos.

Neste contexto, o trabalho desenvolvido neste doutoramento teve como principal objetivo aplicar abordagens computacionais a partir da perspectiva da biologia de sistemas para propor mecanismos transversais explicativos da degeneração de NMs. Em particular, propusemos responder a duas perguntas: 1) como é que o fenótipo comum dos subtipos de DNM surgem da alteração de vias distintas?; 2) Como é que a alteração de proteínas ubiquamente expressas pode afetar apenas a um tipo celular como os NMs?

Com base nestes objetivos, desenvolvemos dois métodos baseados em redes de interação proteína-proteína (IPP).

O primeiro método, chamado BioInt-U, foi desenhado para caracterizar funcionalmente os interactomas específicos de tecido no contexto normal e de doença. O método identifica unidades de Interação Biológica, definidas como grupos de proteínas que interatuam fisicamente e partilham uma anotação funcional biológica (Ontologia). O método foi aplicado em 33 tecidos humanos para identificar os catálogos de funções associadas a cada tecido. Seguidamente, as bibliotecas de funções permitiram identificar propriedades topológicas diferenciais entre as proteínas expressas ubiquamente, daquelas proteínas especificamente expressas em poucos tecidos. Os resultados mostraram que as proteínas ubíquas podem colaborar em processos biológicos básicos para a sobrevivência de qualquer célula,

mas também em funciones especificas de tecido. Finamente, o mapeamento de genes associados a doenças especificas de tecido revelou que as funções que acumulam mais mutações associadas a doença têm maior centralidade nas respetivas redes especificas de tecido.

O segundo método, chamado *Specific-Specific Betweenness* (S2B), foi desenhado para identificar, em redes de interação proteica, proteínas centrais capazes de conetar especificamente qualquer par de conjuntos de genes associados a doenças semelhantes. A qualidade das previsões do método S2B foi avaliada com redes randomizadas e módulos artificiais de doença, desenhados com base nos conceitos mais recentes da medicina de redes. O S2B foi aplicado para prioritizar os candidatos que conetam as proteínas associadas à ELA com as proteínas associadas à AME em redes neuronais humanas e de *Drosophila*. Observamos que muitos candidatos estão envolvidos em funções previamente associadas a DNMs e também a outras doenças neurológicas.

Paralelamente, o nosso laboratório, em colaboração com parceiros internacionais envolvidos no projeto Fly-SMALS (EU-JPND), caracterizou a desregulação transcricional de modelos genéticos de ELA e AME na mosca-da-fruta. Os modelos '*knockdown*' consistiram na redução de expressão, por via de RNA de interferência, dos genes ortólogos de TARDBP, FUS e SMN1. Os analises de sequenciação de RNA (RNA-Seq) revelaram que a inibição da expressão dos genes causais de ELA e AME altera a abundância de um grande grupo de genes em comum, mas também de genes específicos de cada modelo. Para além disso, observámos que os transcritos desregulados se encontram associados a processos biológicos muito variados. O mapeamento dos candidatos identificados na mosca-da-fruta em bibliotecas funcionais obtidas com o método BioInt permitiu encontrar processos biológicos enriquecidos em genes regulados pelos genes causais de ELA e AME. Mais uma vez, os resultados obtidos revelaram que as funções que acumulam a fração mais elevada de candidatos se encontram estreitamente envolvidas na fisiologia dos NMs.

A investigação conclui-se com a combinação dos resultados obtidos em cada estudo para gerar uma visão comum dos mecanismos de DNMs. O uso de bibliotecas funcionais definidas pelo método BioInt permitiu a integração dos candidatos gerados em modelos humanos e de mosca-da-fruta. Os resultados indicaram que as proteínas candidatas em mosca-da-fruta e humano tem marcas biológicas particulares a seu contexto biológico. Além destas diferenças, a combinação dos resultados permitiu prioritizar funções comuns associadas aos diferentes modelos de DNM em humano e mosca-da-fruta e que são afetadas por diferentes redes proteicas.

O presente estudo fornece uma base para continuar a investigar os mecanismos complexos das DNMs em redes de IPP. Como estas, muitas outras doenças têm etiologia complexa e, portanto, consideramos que as estratégias aqui apresentadas também podem ser aplicadas à investigação dos mecanismos biológicos subjacentes a estes contextos.

Palavras-chave

Doenças do neurónio motor; genes associados a doença; função biológica; redes de interação proteína-proteína; topologia

Abstract

Motor neuron diseases (MND) encompass a spectrum of motor neuron (MN) degenerative conditions associated to numerous genetic alterations, the most common of which are ALS and SMA-5q. Despite years of research, the molecular mechanisms underlying MN degeneration remain unclear. The present work aimed to contribute to answer two outstanding questions: 1) how do MND phenotypes arise from changes in distinct cellular pathways; and 2) how can the alteration of ubiquitously expressed proteins generate MN-specific diseases. We made use of network biology principles to investigate the tissue-specific interactomic context of MND genes and elucidate transversal characteristics shared by distinct MNDs. Two novel network-based methods were developed to characterize the functional landscape of tissue-specific interactomes (BioInt-U); and to identify bottleneck proteins connecting pairs of diseases with similar phenotypes (S2B). The application of the BioInt-U method to human PPI networks revealed tissue-specific functional specialization of ubiquitous proteins, with effective prediction of disease phenotypes. The S2B method was applied to prioritize candidates connecting ALS and SMA-linked genes in human and *Drosophila* brain networks, revealing coherent functional roles. RNA-seq data from *Drosophila* models was used to identify neuronal genes that are directly and indirectly regulated by the fly orthologs of the ALS and SMA causal genes TARDBP, FUS and SMN1. This work revealed a phenotypic convergence onto common protein functional modules, albeit through independent targets. In conjugation with the BioInt method, it further provided insights into the origin of MN specific phenotypes. Finally, the candidate genes identified in human and *Drosophila* networks using S2B and *Drosophila* transcriptome data, complemented by a publicly available dataset from ALS patients, were subjected to an integrative analysis by mapping onto BioInt units. Taken together, the work presented here provides novel insights regarding the molecular mechanisms underlying MNDs, while developing computational methods that can be used to address other diseases.

Keywords

Biological process; disease gene; motor neuron diseases; protein-protein interaction network; topology

Table of Contents

List of Figures

List of Tables

List of Abbreviations

APID – Agile Protein Interactomes Dataserver

AP-MS – Affinity Mass Purification

ALS – Amyotrophic Lateral Sclerosis

AS – Alternative Splicing

ASO – Antisense Oligonucleotide Therapy

BiU – Biological Interacting (BioInt) unit

BiU_{TC} – Tissue-Consistent Biological Interacting (BioInt) unit

BiU_{TIC} – Tissue-Inconsistent Biological Interacting (BioInt) unit

BiU_{X} – Not-enriched Biological Interacting (BioInt) unit

BP – Biological Process

CC – Cellular Compartment

cDG – Cross Disease Gene

CNS – Central Nervous System

CO – CORUM

DDR – DNA Damage Response

DE – Differentially Expressed

DEg – Differentially Expressed gene

DGE – Differential Gene Expression

DG – Disease Gene

DGE - Differential Gene Expression

DIOPT – Integrative Ortholog Prediction Tool

dsRNA – double-stranded RNA

ENA – European Nucleotide Archive

ER – Endoplasmic Reticulum

ERAD – Endoplasmic Reticulum Associated Degradation

fALS – Familial Amyotrophic Lateral Sclerosis

FC – Fold Change

FDR – False Discovery Rate

FEA – Functional Enrichment Analysis

FP – False Positive

FPKM – Fragments Per Kilobase of transcript per Million

FTD – Frontotemporal Dementia

GEO – Gene Expression Omnibus

GFP – Green Fluorescent Protein

GO – Gene Ontology

GO-BP – Gene Ontology - Biological Process

GOF – Gain-Of-Function

GSEA – Gene Set Enrichment Analysis

GWAS – Genome-wide Association Studies

HK – Housekeeping

HKu – Housekeeping BioInt unit

HUPO – Human Proteome Organization

HuRI – Human Reference Protein Interactome Mapping Project

IP – Immuno-Precipitation

LOF – Loss-Of-Function

MF – Molecular Function

MN – Motor Neuron

MND – Motor Neuron Disease

mRNA – Messenger RNA

NGS – Next Generation Sequencing

NMJ – Neuromuscular Junction

nonUB – Non Ubiquitous

nonDG – Non Disease Gene

OMIM – Online Mendelian Inheritance in Man

ORA – Over-Representation Analysis

RBP – RNA Binding Protein

RIP – RNA Immuno-Precipitation

RNAi – RNA interference

RNA-seq – RNA sequencing

RNP - Ribonucleoprotein

ROS – Reactive Oxygen Species

rwr – random walk with restart

PPI – Protein-Protein-Interaction

shRNA – short-hairpin RNA

SMA – Spinal Muscular Atrophy

snRNP – Small Nuclear Ribonucleoprotein

sp – shortest path

SRA – Sequence Read Archive

SS – Simpson's Similarity

SS_1 – Specificity Score 1

SS_2 – Specificity Score 2

S2B – Double-Specific Betweenness

TE – Tissue-enriched

TEu – Tissue-enriched BioInt unit

TF – Transcription Factor

TP – True Positive

TS – Tissue-specific

UAS – Upstream Activating Sequence

UB – Ubiquitous

UMLS – Unified Medical Language System

UPR – Unfolded Protein Response

UPS – Ubiquitin Proteasome System

Y2H – Yeast-two-hybrid

1 Introduction

1.1 Motor Neuron Diseases

Motor neuron diseases (MND) include a whole spectrum of disorders particularly affecting motor neurons (MN). Amyotrophic lateral sclerosis (ALS) and spinal muscular atrophy (SMA) are the most frequent MND types, thereby the ones amounting more research efforts. Despite the vast genetic knowledge collected throughout past decades, the pathological mechanisms leading to MN degeneration are still unclear. ALS is a predominantly sporadic, adult-onset neurodegenerative disease, while SMA is an inherited condition that manifests during the first years of life. The disease subtypes of both ALS and SMA present heterogeneous clinical features regarding age of onset, affected body region and disease progression. Likewise, MND patients also reveal diverse mutation profiles. Letting aside the distinctive signatures of the respective subtypes, the most predominant MND disease genes are involved in close molecular processes, suggesting that selective MN degeneration is triggered by common molecular pathomechanisms.

The investigation conducted in this thesis has focused on applying network-based strategies to provide novel hypothesis on the common mechanisms implicated in MN degeneration. Before we delve deeper into the network biology principles exploited in the research, this section will first overview the predominant clinical and genetic traits identified so far in ALS and SMA patients, which sustain the current hypotheses underlying MND pathomechanisms.

1.1.1 Introductory remarks on MND

Healthy neuromuscular system

Motor neurons (MNs) are a specific type of neurons implicated in the communication of the central nervous system (CNS) with muscles and other peripheral systems, such as glands and organs, controlling both voluntary and involuntary movements. Two main types of MNs are recognized according to their anatomic location **(Figure 1.1A).** The cell body of upper MNs is located at the motor cortex and projects to the spinal cord, where they synapse with lower MNs. In turn,

lower MNs are located in the brain stem or spinal cord and project to effector organs and muscles. Visceral lower MNs located at the brain stem are particularly critical for sustaining vital processes since they innervate visceral glands and organs including tongue, esophagus, larynx, lungs, heart, stomach, and intestines. On the other hand, somatic lower MNs innervate skeletal muscle and so control muscle contraction overall. For an extended view, we refer to the sixth edition of Neuroscience book, Chapters 16 and 17 *(Purves et al., 2018).*

The cell body of the MN has the typical morphology of a neuron but is distinguished by a remarkably long axonal projection. The MN axon usually branches to establish many synapses and innervate several muscle fibers. Somatic MNs establish a specialized synapse with muscle fibers, called the neuromuscular junction (NMJ) **(Figure 1.1B).** Vertebrate MNs are cholinergic, that is, employ acetylcholine as primary neurotransmitter. Postsynaptic acetylcholine receptors in muscle cells will activate sodium influx and trigger muscle contraction. Myelination of the MN axon is critical to efficiently transmit the action potential to the synapse. Oligodendrocytes and Schwann glial cells are responsible for axon myelination in the CNS and peripheral nervous system, respectively **(Figure 1.1C).** Likewise, astrocytes are a vital type of glial cells that not only provide metabolic and structural support for neurons, but also influence axonal projection and synaptic signaling. Furthermore, astrocytes are pivotal in regulating immune responses in the CNS.

MND clinical presentation

The clinical phenotype of MNDs encompasses a broad spectrum of symptoms. SMA primarily affects lower MNs while ALS affects both upper and lower MNs **(Figure 1.1A).** SMA is, in most cases, a childhood-onset disease, and infants usually present severe physical disability from the first years of life *(Farrar and Kiernan, 2015).* The age of onset of symptoms in SMA is directly related to the speed of disease progression and life expectancy (extended in next section).

Disregarding the type of MN affected, 75% of ALS cases reveal a spinal onset, being the first symptoms associated with painless muscle weakness. The initial muscular weakness on limbs extends in a contiguous manner from distal to axial

muscles through cell to cell 'domino-like' propagation *(Kanouchi et al., 2012).* Patients experience an increasing fatigue and movement impairment. The progressive functional deficits lead to an overall loss of independence. ALS progression is usually rapid and linear.

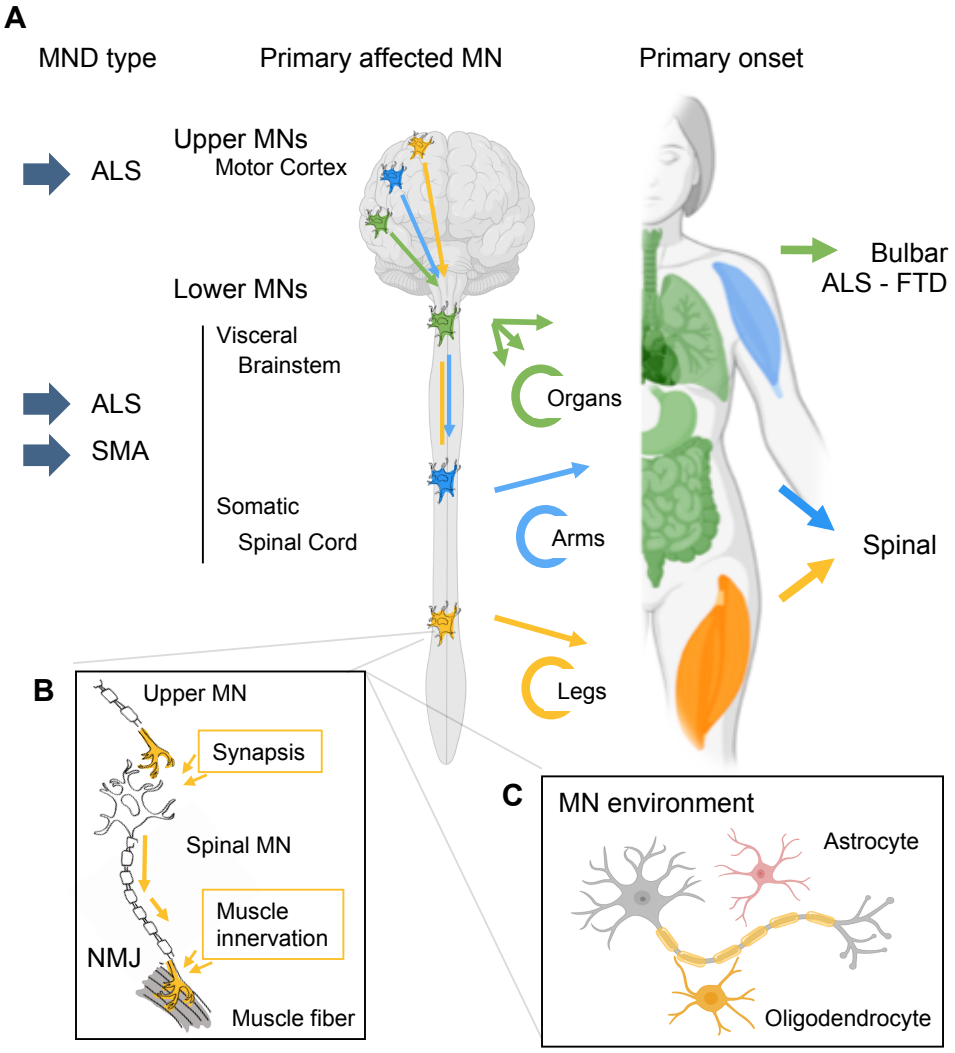


Figure 1.1 Schematic representation of the neuromuscular system and predominant targets of MND

(A) Conceptualization of motor neuron (MN) circuitry and primary targets and onset of ALS and SMA. Depending on the genetic etiology and patient history, MND subtypes can affect various types of MNs (right). Each group of MNs guides a wide variety of both involuntary and voluntary movements. In turn, the degeneration of MNs is initially manifested in different organs or skeletal muscle groups (highlighted in green, blue and yellow, respectively). On this basis, MND types can be broadly classified based on primary onset (left). (B) Sketch of upper and lower MN synapse and muscle innervation in the neuromuscular junction (NMJ). (C) Sketch of most determinant non-MN neuron players in the physiological maintenance of MNs.

The median **age of onset** of ALS is between 51 and 66 years **(Table 1.1)** (Longinetti and Fang, 2019). The **median survival time** of ALS patients from symptom onset to death is between 1 and 2 years. Notwithstanding, 10–20% of patients survive more than ten years *(Jankovska and Matej, 2021).* The rest of ALS patients present with **bulbar onset** characterized by difficulty in speaking or swallowing **(Figure 1.1A).** Cognitive impairment is frequently associated with the bulbar onset cases. The frontotemporal dementia associated with ALS (ALS-FTD) has a significantly worse prognosis that spinal onset ALS *(Jankovska and Matej, 2021).*

MND hereditary patterns and global incidence

SMA incidence is 1 in 5,000-10,000 live births, with the most severe subtype (SMA I) accounting for around 45% of all cases (reviewed by *(Mercuri et al., 2020; Verhaart et al., 2017)* **(Table 1.1).** Although SMA is rare in the population, it is the second most frequent recessive disease following cystic fibrosis, and it is the most common genetic cause of infant mortality.

Table 1.1 Classification criteria of motor neuron diseases (MND)

SMA and ALS subtypes are classified based on genetic mutations and/or clinical features

Motor neuron diseases			MND-associated genes			Clinical features		Frequency	
Type	Subtype		SMN1	SMN2 copies	Other	Onset	Lifespan	Subtype	Type
SMA	SMA-5q	I	Homozygous mutation/deletion of SMN1	2	-	< 6 months	< 2 years	45%	< 30%
		II		3	-	< 18 months	> 25 years (80%)	20%	
		III		4	-	> 18 months	Normal	30%	
		IV		8	-	> 30 years	Normal	<5%	
	Non SMA-5q	-	-	-	> 30	Variable	Variable	<5%	
ALS	Familial	-	-	-	> 50	50-60 years	1-2 years > 10 years (15%)	10%	< 70%
	Sporadic	-	-	-	Unknown			90%	

SMA is in more than 95% of the cases caused by a homozygous deletion/mutation in the SMN1 gene located in chromosome 5q **(Table 1.1).** The SMN protein is encoded by SMN1 and SMN2 genes. While SMN1 is a conserved and essential gene, SMN2 has a silent substitution in exon 7 that impairs splicing and leads the predominant formation of a mRNA isoform without exon 7. This in turn leads to a ~70-80% decrease in the translation of full-length SMN protein. In healthy individuals, SMN1 transcription is sufficient to supply the necessary SMN protein. However, **SMA-5q** or **SMN-related SMA** patients lack normal SMN1 expression. Is at this moment that the translation of SMN protein by the SMN2 alleles becomes a decisive factor for SMA-5q patient survival. Humans can have a variable copy number of SMN2 genes (0-8). As a consequence, SMN2 copy number directly determines the symptom onset and prognosis, which is why it is associated to the four SMA-5q subtypes **(Table 1.1)** (reviewed in *(Mercuri et al., 2020; Ojala et al., 2021).* SMA I and II patients have 2 or 3 copies of SMN2 and are the most severe subtypes, accounting for >60% of total SMA population. These patients are diagnosed during the first months of life and in the most favorable cases, infants with SMA II can survive up to 25 years *(Mercuri et al., 2020).* SMA III patients have more than 3 SMN2 copies and are diagnosed during first years of life but can reach a normal lifespan. SMA IV is the less severe type, with patients diagnosed in their third decade of life. In contrast with SMA-5q, the etiology of SMA cases non-related to SMN1 shows a large genetic heterogeneity. As it will be extended in the following sections, SMA has been associated to mutations in more than 30 genes so far *(Farrar and Kiernan, 2015).*

ALS by contrast is a complex multigenic disease that does not have an apparent hereditary cause in ~90% of the cases **(Table 1.1)** *(Taylor et al., 2016).* Latest surveys indicate that the average ALS incidence in European countries is between 2.1 and 3.8 per 100,000 person annually (reviewed by *(Longinetti and Fang, 2019).* It is also noteworthy that men suffer from ALS at 1.5 times the rate of women *(Jankovska and Matej, 2021)*. Few ALS environmental risk factors have been elucidated so far, including cigarette smoking, traumatic brain injury or intensive physical exercise among the most prominent ones.

Histopathological hallmarks

ALS is the most common form of MND and accounts for about 70% of the cases, out of which 90% are of sporadic etiology. Thus, a large fraction of ALS cases can only be diagnosed when the first symptoms arise. Patients present atrophy in muscle fibers caused by selective degeneration of the MNs involved in muscle innervation. Upon autopsy, the nervous system evinces cell loss in the motor cortex, brainstem and anterior horns of the spinal cord. MN and glial cells reveal a significant accumulation of cytoplasmic inclusions, primarily formed of aggregates containing ubiquitinated TDP-43 protein *(Arai et al., 2006).* An abnormal accumulation of phosphorylated neurofilaments, mitochondria and lysosomes in the proximal axon of large MNs is also frequently found (reviewed in *(Ragagnin et al., 2019)).* Numerous studies also indicate a loss in myelinated axons and neuroinflammation processes *(Komine and Yamanaka, 2015).* Several transgenic models indicate that axonal demyelination and degeneration phenotype occurs prior to MN cell body death *(Dadon-Nachum et al., 2011).* In parallel, degenerative MNs in the spinal cord and motor cortex are frequently surrounded by reactive astrocytes, which concomitantly exhibit intracellular inclusion bodies. However, it is not clear yet whether the activation of astrocytes is a cause or consequence of MN impairment. Subcellular aberrations and organelle abnormalities are frequently found in degenerating MN. Nonetheless, these features are not exclusive to MN death but are commonly found in any degenerating cell. Thus, these hallmarks can be the result of the latest disease stage, once the MN degeneration is irreversible. While it is true that histopathological hallmarks can bring relevant clues, these observations are not sufficient to evince the etiological events causing MN degeneration.

1.1.2 Functional genetics of MND

Although most ALS patients present a sporadic manifestation (sALS), around 10% of patients reveal a traceable hereditary history. The clinical symptoms of sALS and familiar ALS (fALS) are largely indistinguishable, facilitating research into the molecular mechanisms triggering MN degeneration. Even though the biomedical

study of fALS has produced relevant insights into the molecular hallmarks associated with ALS, the ultimate cellular events responsible for MN degeneration remain unclear.

To date, more than 35 genes have been directly linked to ALS **(Figure 1.2A).** Most of these genetic alterations are missense substitutions. However, we find a notable exception in C9orf72 that, in pathological conditions, presents a hexanucleotide expansion that can reach several thousands of repeats. The population suffering from MNDs has not been completely characterized yet and the latest enquiries still return varying conclusions. ALS genetics reveals large differences in European/North American and Asian/Middle East populations *(Mejzini et al., 2019).* Approximately however, the predominant ALS genetic causes include the hexanucleotide expansion of **C9orf72** and missense mutations in **SOD1, TARDBP** and **FUS** genes, representing ~25%, 20%, 5%, <5% of total cases in fALS, and ~10%, 2%, 1% and 1% in sALS, respectively (**Figure 1.2C**, *(Taylor et al., 2016)*). It also is noteworthy that expansion in C9orf72 is particularly related with ALS-FTD spectrum (C9FTD/ALS) in Europe and North America *(Balendra and Isaacs, 2018).* In contrast, SMA is in approximately 95% of the cases triggered by mutations affecting to the expression of the SMN protein. Nonetheless, additional genes have been linked to non 5q-SMA phenotypes (hereinafter referred to as 'SMAs'). Several research groups, including ours have recently revisited the molecular implications of the most frequent MN disease genes (DGs) *(Farrar and Kiernan, 2015; Gama-Carvalho et al., 2017; Mathis et al., 2019; Nguyen et al., 2018; Ojala et al., 2021; Taylor et al., 2016).* **Figure 1.2** summarizes the most consistent MND genetic findings discussed in the abovementioned reviews. The following subsections will cover the most promising pathomechanisms.

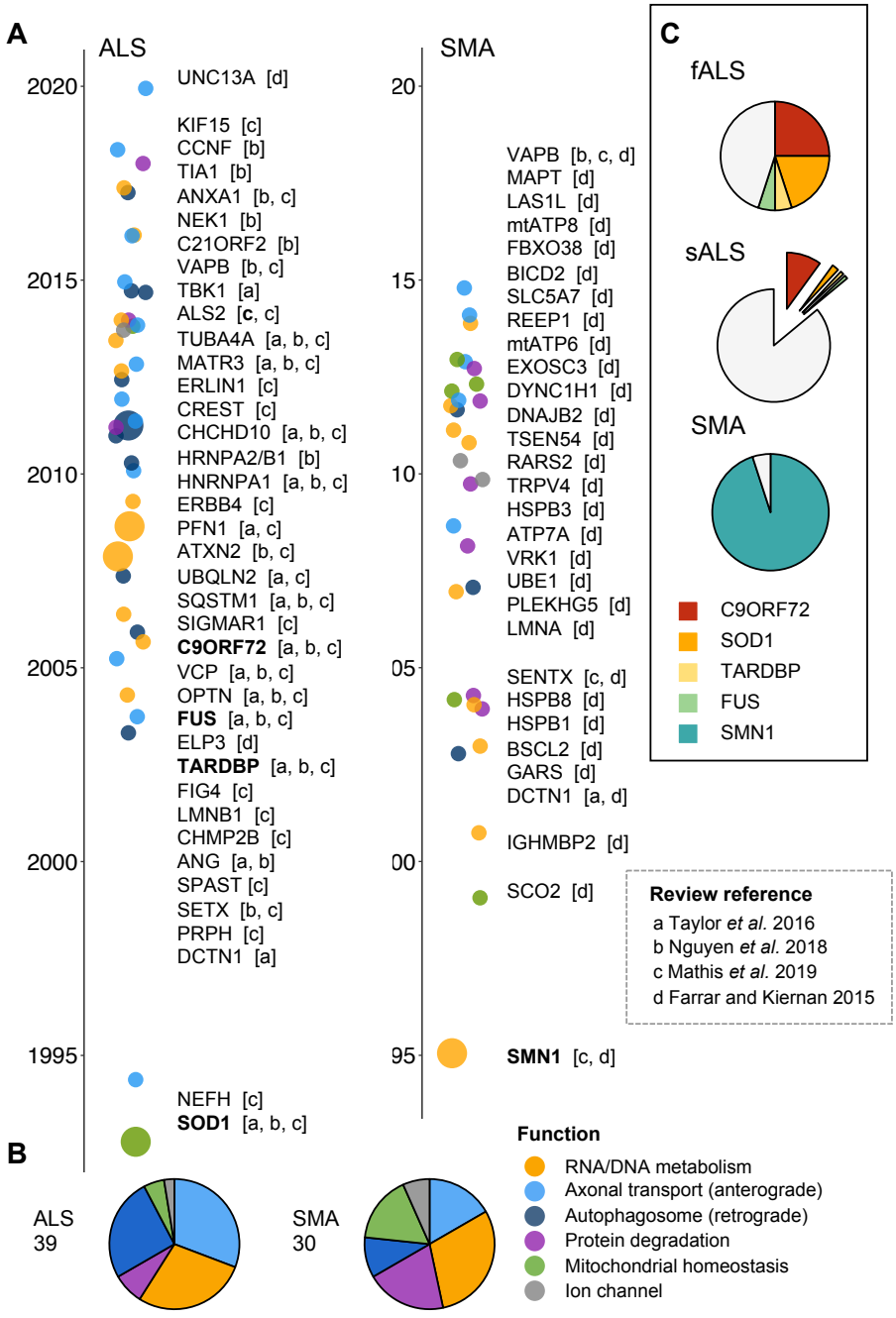


Figure 1.2 Summary of prominent MND disease genes (DGs)

(A) Timeline of MND genetic discoveries discussed in recent reviews (dashed box). Dot color indicates the best-studied roles of the DGs as summarized in panel B. Large dot and bold label points to most frequent DGs in ALS and SMA as illustrated in panel C. (B) Pie charts summarizing the functional roles of DGs. (C) Pie charts summarizing most frequent DGs in ALS and SMA patient groups.

Current molecular hypothesis for MND

Gene mutations can have a wide-ranging impact on gene expression patterns, transcript stability, and protein folding that directly modulate the gene product activity. We can distinguish two broad types of mutations; **Loss-Of-Function (LOF)** mutations inactivate the gene activity and **Gain-Of-Function (GOF)** mutations provide increased wild type functionality or new molecular capabilities. The acquisition of novel functions can be a result from the expression in new tissues, localization in different sub cellular localizations or by the establishment of new interactions *(Li et al., 2019)*. The GOF mutations can in turn become deleterious when generate toxic protein aggregates or impair biological processes. Additionally, mutations affecting the protein localization can be regarded simultaneously as LOF and GOF depending on the cellular compartment that we consider.

The over-stabilization of aberrant protein complexes can trigger cellular stress and degenerate into the aggregation of larger protein-RNA assemblies *(Aulas et al., 2017; Blokhuis et al., 2013).* Protein inclusions can be made up of different proteins and thus can have unpredictable consequences for the cell homeostasis. Protein aggregates are thought to induce cytotoxicity by i) blocking the functions of proteins immobilized in the aggregate, ii) overloading protein degradation systems, and iii) disrupting cell membranes and their associated pathways *(Iuchi et al., 2021).* These events increase oxidative stress, leading to the activation of apoptosis or necrotic pathways.

The toxicity of protein aggregates identified in samples from MND patients is at open debate, as the protein inclusions have **shown neutral, toxic or even protective roles** in different MND models (detailed discussion in *(Hergesheimer et al., 2019; McAlary et al., 2020)*)**.** We find instances of both LOF and GOF mutations in fALS-causing genes *(G. Kim et al., 2020; Taylor et al., 2016)* what once again hinders the identification of transversal disease mechanisms. Regardless to the LOF/GOF mechanisms of the most frequent ALS and SMA-linked mutations in the population, the predominant roles of MN DGs are, according to the overview in **Figure 1.2B;** (i) RNA metabolism, (ii) vesicle trafficking and in particular axonal

transport, (iii) protein degradation, (iv) mitochondria metabolism and (v) DNA damage control. The main goal of this research is to use network-based models that integrate the available molecular information to identify the most determinant players in MND. On this basis, here we will only draw a conceptual briefing of key molecular MND hypotheses and refer to up-to-date reviews for extended discussion **(Figure 1.3).**

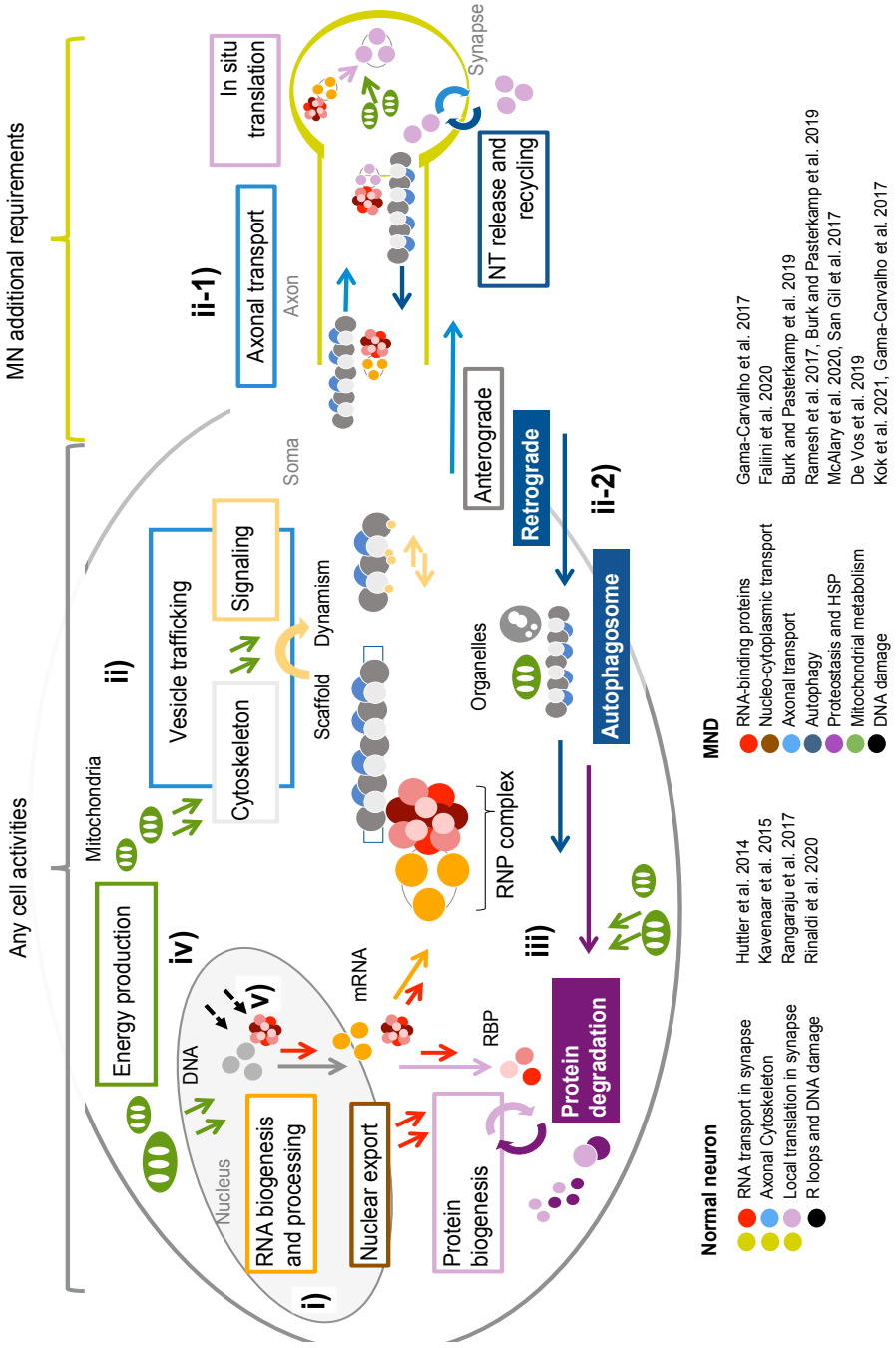


Figure 1.3 Schematic overview of the most consistent MND pathomechanisms

(i) Disturbance of RNA regulation is among the most recurrent MND hallmark. Numerous ALS DGs are RNA-binding proteins (RBPs) implicated in RNA biogenesis and regulation of RNA lifetime and location in the cell. RBPs display numerous PPIs with varying proteins and constitute physical and functional bridges to coordinate biological processes. (ii) Vesicle traffic lies at the center of cell activity being that coordinates the localization of macromolecules to the cell regions where they are required. Vesicle traffic is coordinated by numerous elements but could be broadly divided in scaffold elements made by proteins of the cytoskeleton, and signaling pathways that regulate the cytoskeleton polymerization and movement of motor proteins. (ii-1) MNs have additional functional requirements compared to other cell types. Anterograde vesicle transport is even more critical in MNs given their particular morphology and functional activity. MN axonal projections are particularly long and so MNs require the mobilization of large pools of protein and RNA molecules. Synaptic activity in turn requires of dynamic mRNA pools to produce the proteins necessary for the correct neurotransmitter (NT) release. (ii-2) Retrograde transport is equally essential to recycle and repair the elements deteriorated by normal activity. Vesicles, proteins and organelles are transported to MN soma through the same cytoskeleton scaffold but using distinct signaling pathways. The autophagosome is key for MN physiology as it recycles exhausted mitochondria. (iii) Protein degradation is coordinated by the proteasome and is critical to eliminate potentially harmful proteins and recycle amino acids. External stress and disturbances in protein folding can trigger exacerbated ER stress. At the same time, mutations in the regulators of the proteasome can similarly trigger the MN proteome deregulation. (iv) Energy metabolism is pivotal for any cell. Mitochondria are distributed around the cell to provide ATP to biochemical processes. The production of ATP is inherently related to oxidative stress thus, the cell has control mechanisms to balance reactive oxygen species (ROS). MNs are long living cells with additional energy requirements, which make them more sensitive to oxidative damage. (v) The normal oxidative stress and RNA biogenesis are main sources of DNA damage. However, the alteration of RBPs and energy metabolism in MNDs can notably increase DNA damage and MN genome instability.

i) RNA metabolism

RNA metabolism is tightly orchestrated by hundreds of **RNA-binding proteins (RBPs)** that regulate all the events of RNA biogenesis including transcription, post-transcriptional modifications and translation control. RBPs further shape cellular RNA pools by regulating RNA stability, localization and degradation (mRNA transport regulation in synaptic plasticity reviewed by ***(Hutten et al., 2014).***

Most of the predominant mutations linked to MNDs are located in genes involved in RNA metabolism including SMN1, TDP-43 and FUS that suggest that RNA homeostasis deregulation has key implications in MN degeneration. **TDP-43** (codified by TARDBP) and **FUS** proteins are both RBPs mostly known as regulators of splicing, translation and mRNA stability. In contrast, **SMN** does not belong to the RBP/hnRNP family but collaborates as a chaperone to assemble the RNA-protein complexes necessary to coordinate diverse RNA pathways. The **SMN complex** best-known function is the assembly of small nuclear ribonucleoproteins (snRNPs) that will become the building blocks of the spliceosome machinery. The spliceosome in turn,

is a highly dynamic and large macromolecular complex in charge of the catalysis of RNA splicing but also in the coordination of additional RNA processing steps as the just mentioned. The best-known function of C9orf72 is the regulation of Rab GTPases during autophagy (extended later). Thus, the repeat expansion can block C9orf72 protein through LOF mechanism and affect to vesicle trafficking. However, the most substantiated hypothesis indicates the toxicity of the hexanucleotide expansion is derived from the transcription of RNA foci that sequester RNAs and RBPs including TDP-43 itself (pathological mechanisms of C9orf72 reviewed by (Balendra and Isaacs, 2018). Less frequent MN DGs involved in RNA metabolism include the heterogeneous nuclear ribonucleoparticles **hnRNP A1 and A2/B1,** or the proteins **Matrin 3 (MATR3), Ataxin2 (ATXN2), Angiogenin (ANG), Senataxin (SETX),** or the transcription elongator **ELP3** (RNA regulators implicated in MND reviewed by *(Gama-Carvalho et al., 2017))*

ii) Axonal transport

Neurons have unique functions that require singular morphological adaptations. The main function of neurons is to receive, store and transmit information through complex neuron networks. In turn, the neuron transmits the information pulses through dynamic signaling pathways that require rapid protein expression mechanisms at the synapse. Thus, axonal transport of the elements necessary for protein translation - including RNA and ribosomal macromolecules among others - becomes essential for normal neuronal physiology (local translation in neurons is reviewed in *(Rangaraju et al., 2017).* Added to this, it must be noted that the synapse is typically distant from the neuron cell body. This is especially accentuated in the case of MNs, being that in humans the axon can reach a meter in length. For these reasons, the maintenance of the axonal delivery system is a particularly critical function for MN synaptic activity. In fact, as early as 1990 it was observed that the largest MNs are the first neurons to degenerate in ALS, corroborating that indeed the axonal length is a critical factor in MND.

Axonal transport can be divided in three main processes: 1) the recruitment of the delivery cargo as macromolecules, vesicles or organelles; 2) the polymerization of

the cytoskeleton filaments; and 3) the regulation of cargo transport through the filaments. The cytoskeleton is a highly dynamic network of filamentous proteins that links all regions and components of the cell (reviewed in ***(Hohmann and Dehghani, 2019).*** It maintains cellular organization by giving structural support and mediating communication across the entire cell. There are three major types of cytoskeleton filaments: microtubules, actin filaments, and intermediate filaments. The neuron axon is a distinctive cell extension from the neuronal body with a unique arrangement of cytoskeleton filaments (cytoskeleton organization in axon is reviewed in ***(Kevenaar and Hoogenraad, 2015).*** Neurofilaments are a special type of intermediate filaments that embeds microtubules and determine the radius of the axon and thereby axonal conductance. Both microtubule and microfilaments are constantly undergoing cycles of polymerization and depolymerization to rearrange cytoskeleton organization according to the axon requirements. Microfilaments (actin polymers) provide mechanical support and facilitate the transduction of extracellular mechanical signals that modulate axonogenesis and axon pruning. On the other hand, microtubules (tubulin polymers) coordinate long-distance transport of organelles and vesicles both in anterograde (towards synapse) and retrograde (towards neuron body) directions. Both the microtubules and microfilaments have specific motor proteins that "walk" along them. Myosins move along the microfilaments while dynein and kinesin in microtubules. While one end of the motor protein holds onto the cytoskeleton, the other end binds to the cargo. Motor proteins advance through the cytoskeleton by conformational changes derived from ATP hydrolysis. Microtubule anterograde transport is exerted by kinesins and is essential to deliver necessary cargos to the synapse. The retrograde transport is conducted by dynein motor-proteins to coordinate the recycling of organelles and macromolecules in the cell body.

This is a simplistic vision of cytoskeleton dynamics, and it must be noted that these are complex processes orchestrated by highly dynamic signaling networks including **Rab GTPases and Rab exchange factors (GEFs) proteins** (Rab proteins role in ALS reviewed in ***(Burk and Pasterkamp, 2019).***

ii-1) mRNA axonal transport

The heavy use of axonal transport for synapse in situ protein translation, the extreme polarized morphology of MNs and the high prevalence of DG involved in RNA metabolism and vesicle trafficking strongly point to mRNA transport disturbance as major event for MN. The first step on mRNA transport to the synapse begins in the nucleus. **TDP-43** and **FUS** are well known nucleo-cytoplasmic shuttling RBPs involved in mRNA nuclear export. Along the same line, MND-associated mutations in genes encoding nuclear envelope proteins **Lamin (LMNB1**, in ALS) and **Prelamin (LMNA,** in SMAs) also support the hypothesis that nuclear export is a central event on MN deregulation (nucleo-cytoplasmic traffic in MND is reviewed in *(Fallini et al., 2020).*

Once the RNA cargo is available in the cytoplasm, it must be loaded to the cytoskeleton system to be transported to its final destination. Cytoskeleton constituents as **Tubulin** (**TUB4A**), cytoskeleton regulators as GEF Rab GTPases (**ALS2** and **C9orf72**) and motor proteins as Dynactin (**DCTN1**) are equally essential elements of vesicle trafficking with known association to MNDs (disrupted neuronal trafficking reviewed in *(Burk and Pasterkamp, 2019).* Additional observations indicate the impairment of regulators of microtubule **(SPAST** and **REEP1,** ALS and SMAs DGs, respectively) and microfilament (**PFN1** and **ANXA1**) polymerization. Regulators of neurofilament networks as **NEHF** and **PRPH** have also been linked to ALS. Additional vesicle trafficking elements associated to ALS include **VAPB** that coordinate microtubule and vesicle membrane interaction; **FIG4** that collaborates in phosphoinositide signaling pathway to regulate vesicle trafficking; and **UNC13A,** which is essential for synaptic vesicle maturation. It should be noted that the cargo is not only restricted to mRNA molecules, but neurons also require the transport of ribosomes and organelles, such as mitochondria, in both anterograde and retrograde directions.

ii-2) Autophagy and retrograde transport

Neurons are long-lived and non-dividing cells, and as such particularly susceptible to accumulating aggregates of misfolded proteins and damaged

organelles throughout their lives. Overall, cells have two major protein degradation pathways: the **autophagy–lysosome** pathway and the **ubiquitin proteasome system (UPS).** The UPS is the major proteolytic pathway in the cell and degrades short-lived soluble proteins (extended in next subsection). Autophagy is a lysosome-dependent degradation process that recycles long-lived proteins and cytoplasmic organelles including the ER or mitochondria. Autophagy is tightly dependent on **vesicle retrograde trafficking.** Thus it is not surprising that many MN DGs orchestrating vesicle transport are concomitantly implicated in autophagy deficiency (reviewed in MND *(Ramesh and Pandey, 2017).* For instance, **Dynactin** is an ALS and SMAs DG that coordinates the retrograde axonal transport mediated by the **Dynein** motor protein (also a SMAs-linked DG). Similarly, On the other hand, **Optineurin (OPTN)** is a membrane trafficking protein that regulates the autophagy of damage mitochondria. **VCP** and the **Sequestosome (SQSTM1)** are also autophagy receptors implicated in ALS. Likewise, GEFs of Rab GTPases - as the previously mentioned **C9orf72** and **ALS2** - modulate both cytoskeleton dynamics and autophagy processes. Moreover, **PLEKHG5** is a non 5q-SMA DG with GEF activity that regulates the autophagy of synaptic vesicles. **CHMP2B** is another ALS DG involved in the regulation of endosome and lysosome activities *(Burk and Pasterkamp, 2019).*

iii) Protein homeostasis and ER stress

The best-known histological hallmark of ALS is the accumulation of cytoplasmic protein aggregates. The aggregates are predominantly made up of TDP-43 ubiquitinated protein but additional proteins encoded by DGs implicated in varying functions as **hnRNP A1** and **2B1** or **UBQLN2, OPTN, SQSTM1, VCP** and **PFN1** are also detected in these aggregates. These hallmarks suggest that the mutations in these genes might induce protein aggregation through GOF mechanisms. In parallel though, the cytoplasmic aggregates might also be the result of an exacerbated ER stress and proteasome impairment. For instance, ALS DGs **VCP** and **CCNF** regulate the E3 ubiquitin-protein ligase complex, while **ERLIN1** mediates the ER-associated degradation (ERAD) pathway. On the other hand, **UBQLN2** and **SQSTM1** mediate the proteasomal targeting of misfolded proteins and bridge ERAD and autophagy

pathways (ER stress and proteostasis alterations in ALS reviewed in *(Maharjan and Saxena, 2016; McAlary et al., 2020)).*

In addition to vesicle trafficking, **VAPB** is involved in the ER unfolded protein response (UPR) and its alteration has also been linked to the formation of intracellular protein aggregates. The accumulation of DGs involved in proteasome coordination is even more conspicuous in SMAs sprectrum **(Figure 1.2B).** In particular, it is worth highlighting the mutations in heat shock proteins (**HSPB1, HSPB3, HSPB8**), chaperones (**DNAJB2**) or ubiquitination modifiers (**UBE1** and **FBXO38**) (heat shock response to protein misfolding in neurodegenerative diseases reviewed in *(San Gil et al., 2017)).*

iv) Energy metabolism and mitochondrial stress

The synaptic activity of neurons has intensive energy requirements. The healthy mitochondrial activity necessary to produce chemical energy (ATP) generates reactive oxygen species (ROS) as a byproduct. Thus, cells with such high-energy requirements as neurons are expected to generate more ROS and in turn, be more susceptible to accumulate free radicals. The accumulation of ROS can induce deleterious oxidative modifications on proteins, nucleic acids and lipids. The production of free radicals is compensated by ROS detoxifying enzymes as **SOD1**. The SOD1 enzyme detoxifies superoxide radicals into molecular oxygen and hydrogen peroxide. SOD1 was the first gene to be linked to ALS and latest surveys indicate it is implicated in 20% of fALS. These observations, together with the direct interconnection between cellular stress and protein aggregation, propelled oxidative damage as major potential cause MN degeneration (mitochondria deregulation in ALS is reviewed in *(Smith et al., 2019).* However, accumulating evidence indicates that SOD1-ALS does not exhibit a LOF mechanism. Several ALS-causing mutations in SOD1 do not alter the catalytic activity of the proteins but decrease the protein stability and can augment the protein fibrillation. Further, SOD1 has been detected in protein inclusions in 2% of ALS cases *(McAlary et al., 2020).* While it is evidence that some mutations can induce LOF of SOD1, recent studies show it is not sufficient to cause ALS *(G. Kim et al., 2020).*

Aside the disease mechanism of mutations in SOD1, we still find several ALS DGs to be involved in mitochondrial functions, **CHCHD10** or **NEK1** *(Nguyen et al., 2018).* Mitochondrial dysfunction is even more conspicuous in SMAs (**Figure 1.2B)** being noteworthy the mutations in SCO2, **Seipin (BSCL2), RARS2,** or mitochondrial ATP synthases 6 and 8 **(mt-ATP6, mt-ATP8**) *(Farrar and Kiernan, 2015).* Beyond the alteration of mitochondrial proteins, it is also plausible that energy metabolism is impaired in MNs due to mitochondrial misallocation in the axon terminals and synapses *(Burk and Pasterkamp, 2019; Ramesh and Pandey, 2017).* Nonetheless, it must also be noted that ROS accumulation is a time-dependent cumulative process. Considering that neurons are particularly long-living cells, many authors argue the oxidative stress observed in MND patients is probably a late-stage hallmark and not a triggering event of MN degeneration.

v) DNA damage

DNA damage accumulation has also been a recurrent hypothesis to explain MN degeneration (DNA damage in ALS reviewed in *(Kok et al., 2021).* DNA damage occurs regularly during the normal physiological processes and numerous pathways have evolved to protect and repair DNA. However, certain conditions can aggravate DNA damage or impair pathways involved in DNA repair (DNA damage response, DDR). For instance, it is well described that the oxidative stress derived from ROS accumulation is a determinant factor on DNA damage increase. Likewise, RNA biogenesis involves an inherent risk of potential DNA damage. Thus, the perturbation of RNA metabolism pathways can similarly compromise genome stability. Due to the physical proximity during transcription, DNA and RNA molecules form hybrid structures known as **R-loops**. These are critical to pause RNA polymerase II progression and allow the correct transcription termination. Nonetheless, when transcription is concluded, R-loops have to be untangled to avoid DNA damage (R loops implication in DNA damage reviewed by *(Rinaldi et al., 2021)).* The ALS-causing gene **Senataxin (SETX**) is a DNA/RNA helicase particularly involved in resolving the R-loops. It is not surprising to find that RBPs also display protective roles towards DNA damage and repair processes. RBPs as FUS and TDP-43 are

able to bind RNA, DNA and proteins and so can operate as hubs to coordinate DDR pathways (ALS-linked RBPs in DNA damage is reviewed in *(Gama-Carvalho et al., 2017)).* Similarly, several other genes associated with ALS including **NEK1, SQSTM1, VCP** and **C21ORF2** are known to play a role in DNA repair (reviewed by *(Kok et al., 2021)).* Likewise, in SMAs we may highlight intermediate filament of nuclear membrane **LMNA** and DNA/RNA helicase **IGHMBP2** as they also establish interactions with DNA with potential risks for genome stability.

ALS and SMA genetic overlap

As discussed throughout this section, ALS and SMAs share conspicuous histological and molecular hallmarks. Although mutations in SMN contribute to ~95% of SMA cases (SMA-5q), the remaining SMAs DGs reveal manifest similarities. Overall, the proteins encoded by ALS and SMAs DGs are involved in the same functional processes and oftentimes, physically interact (discussed in *(Gama-Carvalho et al., 2017)*. Overall, it seems likely that MN degeneration can be triggered by a combination of several concomitant processes. In fact, the molecular functions just discussed could be considered a continuum of biological processes and as such, all the suggested pathomechanisms might possibly be snapshots of a larger phenomenon **(Figure 1.3).** Just as the investigation of fALS has revealed significant insights into the disease mechanisms, the integration of SMA and ALS molecular pathways may help to elucidate determinant events on MN degeneration.

1.1.3 Social impact and therapeutic prospects

ALS has a significant economic impact on patients, families, and national health systems. The manifestation of first symptoms commonly forces patients to discontinue working. Late disease stages require families to perform home modifications and, driven by income loss, to frequently undertake voluntary care tasks. The few treatments currently available are expensive and achieve modest improvements. Terminal stages require invasive interventions as tracheostomy and gastrostomy for assisted ventilation and feeding. These decisions cause a great

burden on the patient and families psychoemotional status. Achtert and Kerkemeyer have recently reviewed several studies evaluating the economic costs both at individual and government levels *(Achtert and Kerkemeyer, 2021)*. The national expenditures associated to ALS greatly varied according to the distinct health systems evaluated ranging from €149 million in Canada and €1,329 million in USA per year. Even though ALS is a rare neurodegenerative disease, the study estimated that the national cost per ALS patient is higher than other neurologic diseases like dementia and Parkinson's disease. While the direct annual costs per patient averaged €1,168 in the Spanish health system, the additional medical care options can reach €50.000 per family per year *(Darbà, 2019).*

At present, three treatments for SMN-dependent SMA (SMA-5q) have been approved by the US Food and Drug Administration (FDA). Spinraza® (2016) and Risdiplam® (2020) treatments are aimed to increase SMN protein production using antisense oligonucleotide therapy (ASO) and small molecule to modulate SMN2 splicing, respectively. Zolgensma® (2019) is a vector-based gene therapy to deliver full-length SMN1 cDNA into target MN cells (reviewed by *(Ojala et al., 2021; Schorling et al., 2020)).* These therapies have slowed the progression of SMA-5q subtypes and improved the survival rate of the patients (reviewed by *(Mercuri et al., 2020; Ojala et al., 2021)).* Nonetheless, SMN1 is highly expressed prenatally in most organs, which indicates it could be implicated in organ morphogenesis. In fact, there is mounting evidence that other tissues beyond MN circuitry are vulnerable to SMN deficiency including cardiac, gastrointestinal or endocrine systems *(Ojala et al., 2021; Schorling et al., 2020).* Therefore, more effort needs to be put into prenatal screenings for SMN1 mutations and in the design of additional therapies targeting other tissues than CNS. At the same time, ASO therapies targeting ALS are also being developed. The most promising treatments at present, are designed to lower down the levels of C9orf72 or SOD1 proteins in GOF models (perspectives in ALS therapies are reviewed in *(Masrori and Van Damme, 2020; Xu et al., 2021).* However, the genetic causes of ALS and SMAs are heterogeneous, so gene therapy is not as cost-effective as it is for SMA-5q cases.

The two treatments already approved for ALS treatment are directed to neuroprotection. Riluzole was the first drug approved by the FDA to treat ALS in 1995. Riluzole MN protective roles have been related to its capacity to decrease glutamate release and block MN sodium channels. Likewise, edaravone is a free radical scavenger employed to decrease oxidative stress. Although it is administered in several countries including US, Canada, Japan or South Korea, its use has not being yet approved in the European Union. In any case, riluzole and edaravone benefits are modest and cannot reverse the previous MN degeneration thus, they are not a definite treatment for ALS. Other drug compounds under scrutiny include regulators of protein aggregation, autophagy or the stimulation of muscle growth, among others. Likewise, studies using cell-based therapy are also accumulating promising results (reviewed in *(McAlary et al., 2020; Schorling et al., 2020; Xu et al., 2021)).*

One of the most significant limitations in MND treatment is that the majority of MND cases present as clinical diseases, meaning that can only be diagnosed by the detection of the initial muscle weakness. Unfortunately, first symptoms arise after a vast and rapid MN decay *(Aggarwal and Nicholson, 2002).* The accumulation of potentially toxic cytoplasmic aggregates and vast MN loss prior to the manifestation of the first symptoms underlines the urgency for the identification of early pre-symptomatic biomarkers. While it is true the search for biomarkers at biofluids is a promising enterprise, the latest works still present incongruent results (reviewed by *(Chipika et al., 2020)).* To date, one of the most significant evidences of a presymptomatic trait in ALS patients is the elevation of neurofilament light chain (NfL) in the serum and cerebrospinal fluid a year before the emergence of muscle weakness. However, the rapid MN loss at presymptomatic stages predicts the diagnostic protocols would require much more anticipation to avoid irreversible neuromuscular degeneration.

The lack of a clearly defined MND pathogenesis mechanisms leads to limited diagnosis and therapeutic options. Even with an incomplete view, it remains vital to identify the leading determinants of MN degeneration. The identification of converging

molecular hubs or bottlenecks across SMA and ALS subtypes could potentially enable the design of early diagnostic protocols and unified drug treatments.

1.1.4 *Drosophila* as a model for neuroscience

The fruit fly *Drosophila melanogaster* has been used as an animal model in biological sciences for over 100 years. More notably, in the last decades it has been extensively employed in neurodegenerative research (review in ***(Bolus et al., 2020; van der Voet et al., 2014)***. The fruit fly offers numerous benefits when compared to vertebrate animal models as chicken or rodents. The fly has a short life cycle and high offspring numbers. The development of an adult fly only takes 10 days from fertilization and the female fly can produce up to 1500 eggs in its lifetime. Due to its small size, it is easy to handle in the laboratory and requires low costs for maintenance.

One of the most profitable features of fly as a research model is its easy genetic manipulation. The UAS/Gal4 system is one of the most versatile expression systems developed in *Drosophila* ***(Brand and Perrimon, 1993).*** The UAS/Gal4 system can be used for either transgene overexpression or gene knockdown through RNA interference (RNAi) (Piccin et al., 2001). The RNA interference is conducted by inducing the expression of a double-stranded RNA (dsRNA) that forms short-hairpin structures (shRNA) that interact with target RNA to impede its translation ***(Paddison et al., 2002)***. The transgene or dsRNA is inserted downstream of an upstream activating sequence (UAS). The UAS is recognized by the yeast transcriptional activator Gal4, inducing the strong expression of the downstream sequence. Gal4 expression can in turn be regulated by a wide selection of promoters to achieve time- or tissue-specific expression. Expression modulation is determinant in the investigation of deleterious mutations implicated in adult-onset diseases such as ALS.

To modulate spatial expression, we can use promoters of tissue-specific genes as for instance the Elav fly neuron-specific gene. On the other hand, temporal expression can be modulated by incorporating promoters of hormone-inducible

receptors into the Gal4 system, as done in the GeneSwitch system *(Osterwalder et al., 2001)*. In this way, the researcher modulates transgene expression simply by hormone feeding. Likewise, many researchers use the temperature-sensitive mutant $Gal80^{TS}$ *(McGuire et al., 2003).* $Gal80^{TS}$ is a transcriptional repressor of Gal4 active at 18º C. Then, according to the study design, the researcher can raise the temperature > 25º C to block $Gal80^{TS}$ and permit the transgene or dsRNA expression.

The UAS and Gal4 constructs are typically inserted in different transgenic fly lines with selection markers, allowing the generation of different models by fly crossing procedures. The large collection of inducible tissue and cell-specific promoter elements currently available enables the development of a wide range of highly specific genetic tools. Likewise, the simplicity of Gal4/UAS system greatly facilitates the implementation of high-throughput assays to evaluate genetic modifiers or candidate DGs *in vivo*. Furthermore, the possibility of combining several UAS/GAL4 systems in a single fly model enables the dissection of complex molecular networks.

Fly and human similarities

Despite the nervous system of the fly being anatomically much simpler than the human, both reveal key similarities. The *Drosophila* CNS is composed by a two-lobed brain with two main cell types, glia and neurons. Flies also have a segmented nerve cord similar to the mammalian spinal cord. The simplicity of fly anatomy greatly facilitates the isolation, manipulation and visualization of MNs and their synaptic contact with the muscle fiber at the NMJ. The molecular composition and physiology of the NMJ is most similar to mammalian glutamatergic excitatory synapses, facilitating the examination of the molecular mechanisms underlying synaptic activity. Furthermore, the fly model also enables the monitoring of motility and behavioral deficits namely, memory and learning capabilities. Despite the evolutionary distance between humans and fly, there is a notable conservation in genes and pathways. According to the Ensembl database, human and fly genomes include 20,465 and 13,969 protein-coding proteins, respectively (GRCh38.p13 and BDGP6.32, www.ensembl.org). According to the DIOPT orthology mapping tool, 40.7% of fly

genes are orthologs covering 55.4% of human coding-genome. Inversely, the 53.5% of human genes cover 66.5% of the fly coding-genome (September 2020 version ***(Hu et al., 2011)).*** More important, most human MN DGs are evolutionarily conserved in *Drosophila* (reviewed in ***(Charng et al., 2014; Olesnicky and Wright, 2018; van der Voet et al., 2014)***.

Fly to human translation

Drosophila also has certain limitations that make knowledge transference back to human models more difficult. Decisive pathogenic factors associated to MND, such as the role of the adaptive immune system, are specific to vertebrate species and therefore, cannot be reproduced in *Drosophila* models.

Likewise, mutant fly models do not always display the same phenotypes as those observed in humans. While the human and fly proteomes present a resounding conservation, the protein-interaction (PPI) networks reveal a notable discrepancy ***(Gandhi et al., 2006)***. As expected, the essential interactions between partners in protein complexes are more conserved than transient regulatory interactions ***(Brown and Jurisica, 2007).*** This indicates that fly and human share the fundamental molecular functions but differ in more complex physiological processes. This observation, in turn, might explain why similar fly and human models might result in different phenotypic outcomes. Therefore, even though fly models are really valuable to dissect fundamental biochemical pathways, these studies must always be carefully translated to the human context.

1.2 Systems biology and omics

Reductionism in biology is the idea that biological processes are ultimately defined by the physico-chemical properties of its elements. Thus, a disease phenotype is conceived as the physical malfunction of discrete molecular elements. The development of molecular biology over the past half-century has made methodological reductionism a central approach to study biological systems. In fact, modern biochemistry and cellular biology have been indispensable for the advances in current medicine. Notwithstanding, reductionist approaches fall short when investigating multigenic diseases in such cases as MND. This limitation is mostly due to the fact that biological entities are genuinely complex.

Complex systems are characterized by two main properties; i) they are composed of many discrete elements and ii) their components are highly interconnected in a non-linear fashion. We find examples of complex systems across all levels of biological organization, ranging from the molecular and cellular realm to tissues, organisms, and whole populations or ecosystems **(Figure 1.4A).** The entangled relationship and dependencies between elements hamper the prediction of the system's behavior. In other words, the properties of the system are said to be **emergent phenomena**, only attributable to the collection of relations between the discrete elements *(Anderson P.W., 1972).*

Systems biology aims to address biomedical challenges by capturing, rather than reducing, the complexity of the biological system of interest **(Figure 1.4B).** Systems biology was initially a theoretical science. However, the standardization of high-throughput biotechnologies, together with the development of public data repositories shared through the Internet, has enabled researchers to amass large collections of biological data and construct highly detailed biological models.

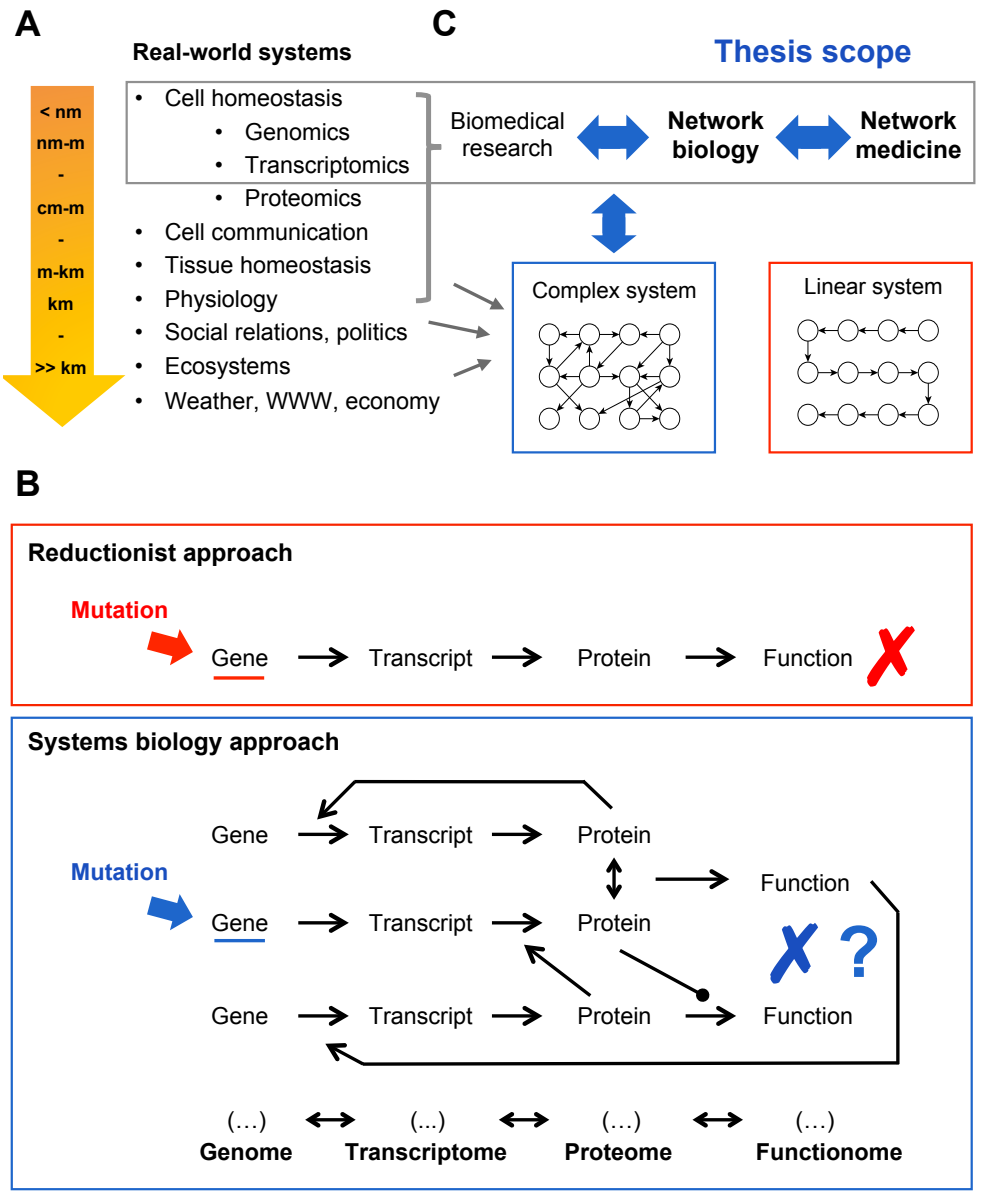


Figure 1.4 Biological systems are genuinely complex

(A) We find examples of complex systems across all scales in nature. Complex systems are made by numerous elements interrelated in a non-linear manner (blue boxes). (B) The complex organization of biomolecular elements hinders the identification of disease triggering events. Many of the limitations in biomedical research are due to applying reductionist approaches that consider that cell malfunction can be explained by mutations in single genes (red boxes). By contrast, systems biology aims to integrate the complex interactions between the biomolecules. (C) The representation of complex systems in networks facilitates their interpretation and modeling. Network theory applied to biomedical research has provided groundbreaking insights towards the mechanisms of cellular organization. Likewise, the principles of network biology can be applied to investigate the etiology of multigenic diseases such as MNDs.

The biomolecular elements that shape living cells are commonly classified into different **'omic'** categories. Best-known categories include genomics, transcriptomics, proteomics or metabolomics. The ulterior objective of omics is to identify and quantify the complete pools of molecular components and their interactions in a given biological system. To this purpose, biochemical approaches - now in a high-throughput fashion - are pivotal for completing the omic databases that will feed the biological models. The “ome” suffix has been extended to the research of other types of large-scale biological information including DNA or protein modifications (epigenome, phosphoproteome), disease-gene associations (diseasome) or functional annotation (functionome). In turn, the investigation of these large datasets requires the establishment of innovative analytical methods. For instance, the representation of biomolecular interactions in networks facilitates the integration of complex data and the application of graph theory concepts to predict biological phenomena.

The research work presented here focused on investigating cell homeostasis and for simplicity; we considered the cell as a whole complex system and obviated the interactions with its environment **(Figure 1.4C).** Protein physical collaboration lies at the center of any cellular activity. Therefore, this work was centered on studying protein-protein interaction (PPI) networks. The construction of the networks required both transcriptional **(Section 1.2.1**) and interactomic **(Section 1.2.3)** data. The methods and outputs generated in this thesis were interpreted using Gene Ontology (GO) functional annotations **(Section 1.2.4)** and disease-gene (DG) associations **(Section 1.2.5).** Overall, the computational methods presented here were designed and evaluated using human data. Additionally, Chapter 3 presents the analysis of in-house transcriptomic sets derived from *Drosophila* MND models and so the output interpretation required the integration of fly orthology data **(Section 1.2.6).** Each of the following subsections briefly describes the technologies most commonly employed to generate omic knowledge, their current limitations and the public repositories where we retrieved the respective datasets. **Section 1.3** will present network biology concepts employed to evaluate the omic datasets introduced in this section.

1.2.1 Transcriptomics

The transcriptome describes the full collection of all RNA molecules in a system (e.g., cellular structure, cell, tissue, organism) in a given moment. There is a plethora of types of RNA molecules depending on their structure and function, but we will only focus on protein-coding messenger RNA (mRNA). The relative quantification of mRNA molecules is commonly employed to address gene expression changes under different experimental conditions. **Microarray and RNA sequencing-based** methods (RNA-seq) are the most popular techniques to collect transcriptomic data.

Microarray platforms were developed in the 1990's and became the first large-scale system to efficiently measure gene expression. Microarrays measure the abundance of a pre-defined set of transcripts by probe hybridization. Affymetrix microarrays were the most popular platforms. Later in the mid 2000s, with the arrival of next generation sequencing (NGS) technologies, **RNA-Sequencing (RNA-seq) methods** emerged as an attractive high-throughput alternative to traditional microarray platforms. In broad terms, RNA-seq uses NGS technology to determine and quantify the sequence of all RNA molecules present in a sample. Thus, one major difference regarding microarray technology is that RNA-seq does not require previous knowledge on the RNA being investigated. Additionally, RNA-seq requires less input RNA (pg Vs. µg) and has a broader detection range i.e., can simultaneously detect RNA species in low and high abundance (reviewed in ***(Lowe et al., 2017)).*** We find several technologies and protocols for RNA-Sequencing (RNA-seq). The Illumina short-read sequencing is the dominant technology employed to date, but third-generation technologies as long-read Oxford Nanopore are firmly taking hold. The protocols have direct implications on the ability of the experiment to address specific biological questions and on the generation of biases in RNA quantification (reviewed ***(Stark et al., 2019)).*** Compared to microarray, RNA-seq methods require additional computational analysis for the reconstruction of the sequence reads into the final RNA transcripts. Thus, RNA-seq presents unique computational challenges namely in data storage and data analysis. In any case, both microarray and RNA-seq provide normalize levels of mRNA abundance that allow the characterization of the

transcriptomic profile of a given sample and perform **differential gene expression (DGE) analysis** between biological samples.

The Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) and European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena) are the largest publicly available repositories of raw high-throughput sequencing data. **Gene Expression Omnibus (GEO)** (ncbi.nlm.nih.gov/geo/) and **BioStudies** (formerly known as ArrayExpress) (ebi.ac.uk/biostudies/) *(Barrett et al., 2013; Sarkans et al., 2021)* are secondary services linked to the SRA and ENA repositories, respectively. These platforms store varied types of biological datasets including microarray and RNA-seq data, and provide comprehensive meta-data about the experiments. Beyond bulk data download, GEO makes available an interactive web tool named **GEO2R** that allows users to compare gene expression profiles from different experiential conditions within a GEO Series.

1.2.2 Proteomics

The proteome is the entire collection of proteins that is expressed in a biological system in a given state and time. High-throughput protein quantification using mass spectrometry-based methods is still a challenging task. Disregarding the time and costs of performing high-throughput quantification, additional drawbacks of mass spectrometry are the raw data processing and analysis complexity. Most important, due to technical limitations, quantitative proteomic data only represents a fraction of the complete proteome. Thus, the current efforts and value of the information provided suffer from significant limitations. To circumvent this, transcriptomic profiling is frequently used to extrapolate the proteome composition of a given condition. Notwithstanding, transcript levels do not necessarily correlate with protein abundance and thus caution must be taken when performing these assumptions. Many researchers, including us, opt to consider protein expression as a binary value by imposing a transcript level cutoff.

The **ProteinAltas** (proteinatlas.org) is web-based database that makes available highly detailed spatial information of protein expression at the subcellular level in distinct cell types and health conditions *(Uhlén et al., 2015).* The ProteinAtlas integrates several types of omic data including mass spectrometry-based proteomics, transcriptomics and antibody-based imaging.

1.2.3 Interactomics

The physical interaction between the molecular constituents of the cell is the driving force of any biological activity. The interactome describes the whole set of molecular interactions but it is often studied in separated networks according to the different biochemical families. The most popular interactome networks address protein-protein interactions (PPI), DNA-protein or RNA-protein interactions (regulatory networks) or enzyme-metabolite interactions (metabolic networks). **Section 1.3** will introduce most relevant properties of biological networks.

Protein-protein interaction networks

The latest efforts of the Human Proteome Organization (HUPO) to gather the complete human proteome generated an encyclopedia of 19,773 protein-coding genes *(Adhikari et al., 2020),* which gives a total of >195.106 possible binary PPIs to evaluate. Although the potential interactome could reach ~200 million PPIs, Venkatesan and colleagues estimated the human complete interactome should be formed by ~130,000 ±30.000 binary interactions *(Venkatesan et al., 2009).* The latest mapping reference recently reported a total of 53.000 PPIs *(Luck et al., 2020),* which would represent - in the most optimistic scenario - ~40% of the interactome. At present, interactome reconstruction endeavors consist on systematically testing all possible physical PPIs. This massive task urges for the standardization of efficient high-throughput PPI detection protocols. The most popular approaches currently available are affinity mass purification and yeast two-hybrid screening.

In a typical **affinity mass purification (AP-MS)** assay, the protein sample is first incubated with the potential interacting target. Then, the proteins interacting with the target are purified in an affinity column and finally identified by mass spectrometry analysis. AP-MS is limited to address the interactions happening between a single target against pools of 400–2000 proteins (reviewed by *(Low et al., 2021).* This feature restricts the build-up of high-throughput protocols to characterize all the possible pairs of interactions across the proteome. On the other hand, **yeast two-hybrid (Y2H)** screening is a protein-fragment complementation assay. The proteins of interest are each covalently linked to two respective fragments of a transcription factor (TF). If the bait and prey physically interact, the fragments of the TF come together and activate the expression of a reporter gene. In contrast to AP-MS, Y2H is the only technology efficient to produce large-scale interactomic data. Indeed, the recent efforts to capture human interactome map required the screening of ~150.106 possible PPIs using Y2H *(Luck et al., 2020).*

Once again, the two techniques have distinct detection biases but combined return complementary views of the interactome. AP-MS is prone to detect protein complexes and Y2H tend to detect binary interactions. A main drawback of AP-MS is that it frequently identifies indirect interactions. On the other hand, Y2H is an artificial system that can only be employed in a synthetic scenario and does not consider whether the proteins are present at the same sub-cellular context or biological condition. Therefore, many of the PPIs detected through Y2H might be not biologically meaningful. To counter back this limitation, numerous authors have proven the benefits of integrating tissue-specific (TS) transcriptomic data to filter potential false positive PPIs and reconstruct TS-interactomes.

As one might expect, PPI techniques are prone to detect more stable PPIs. However, evidence is mounting that stable PPIs constitute a minority of the interactome (Hein et al., 2015). In fact, transient interactions have critical roles in the coordination of complex processes and are as equally fundamental to characterize the cell complex behavior as stable PPIs *(Ghadie and Xia, 2022).* Furthermore, unsurprisingly, early attempts to characterize the interactome were directed to the

characterization of PPIs between proteins with anticipated medical interest. This bias has likely skewed the interactome towards specific protein populations.

APID (apid.dep.usal.es) is a PPI repository that provides a unified version of the protein interactome including the five primary databases of molecular interactions: BioGRID, DIP, HPRD, IntAct and MINT *(Alonso-Lopez et al., 2016; Alonso-López et al., 2019)*. The repository enables to filter PPIs according to the number of experimental validations. It collects the latest interactome datasets for 25 species, which facilitates the design of cross-species analysis protocols. Beyond the identification of binary PPI interactions, the characterization of stable protein complexes, also referred to as to molecular machines, is fundamental to describe the cell's biochemical and mechanical functions. To that purpose, **CORUM** (mips.helmholtz-muenchen.de/corum/) is a gold-standard repository that provides manually curated and experimentally characterized protein complexes *(Giurgiu et al., 2019).*

RNA-protein interaction networks

RNA-binding proteins (RBP) are key regulators of RNA homeostasis. RBPs bind to RNA molecules and varying proteins building large complexes to regulate RNA transcription, stability, splicing and localization, among other activities *(Armaos et al., 2021).* Therefore, RBP-RNA interaction networks provide a comprehensive view of the regulation layers controlling the RNA metabolism and so reveal potential mechanisms of deregulation. Just as there are many types of RNA, many RNA-protein networks can be generated. We only focused on RBP-mRNA networks but it must be noted that non-coding RNAs are particular relevant for mRNA regulation (review of diverse RNA-protein resources *(Yi et al., 2017)).*

In this work we employed a RIP-seq strategy to characterize RNA interactions of RBPs encoded by ALS and SMA genes in *Drosophila* models. The assay consists in immunoprecipitating (IP) the RBPs of interest from a sample followed by sequencing of the bound RNAs. Although RIP-seq confers high-confidence interactions, RBP-IP is not scalable to characterize the complete RNA-protein

interactome. There are varying options to characterize the interactions in a relatively high(er)-throughput manner (reviewed in ***(Hentze et al., 2018)).*** However, a large fraction of the interactions in the public repositories are inferred from algorithms using high-throughput sequencing data to identify RNA-binding motifs and domain (RBD) information ***(Yi et al., 2017)***. Around 1,400 human proteins have been experimentally determined to bind RNA. About half of the proteins in each RNA interactome lacked known RBDs, and hundreds had no known relationship to RNA biology ***(Hentze et al., 2018).*** These observations indicate that state-of-the-art RNA-protein interaction networks are far from complete and, more importantly, we do not yet know how to interpret the functional implications of a large fraction of the RNA-protein interactions.

1.2.4 Functionome

Cataloging the molecular players and their functional capabilities is critical to investigating the different types of cell programs and address pathological conditions. With this objective, the 'functionome' defines the compendium of biological functions a cell can exert, ranging from molecular activities through protein complexes to coordinated cell activities as phagocytosis. Therefore, even though it does not qualify precisely as an omic category, many researchers consider it as such. It must be noted that the functionome was originally built by researchers to classify biological processes and so it presents distinctive features from the canonical omic categories **(Figure 1.5).**

The **Gene Ontology (GO) consortium** (geneontology.org/) provides the framework and the set of concepts for describing the functions of gene products from all organisms ***(Ashburner et al., 2000; Carbon et al., 2021).*** Gene ontology is divided in three independent domains to distinguish the different aspects of a gene product function: molecular function (MF), cellular component (CC) and biological process (BP). MF terms describe the actions or activities of a gene product or molecular machine and CC terms describe their relative localization in the cell. Biological Process (BP) is the most extended category and represents the ultimate biological programs accomplished by multiple molecular activities. Gene functional associations can be addressed using varied evidence protocols. Initially, GO terms

were manually annotated based on published experimental evidence and so were biased towards processes with anticipated research interest. Later, thanks to the accumulation of multivariate biological data, prediction algorithms have notably improved annotation and now are pivotal for electronically inferring GO terms. These assist in providing a broader coverage of the functionome, albeit also introducing a large fraction of false positive annotations *(Yu et al., 2017).*

Beyond the functional characterization of single gene products, most molecular biologists exploit GO annotations to biologically interpret large data outputs derived from high-throughput assays. In broad sense, **functional enrichment analysis (FEA)** is a statistical method to determine the functional traits significantly associated to a given gene or protein list. As recently reviewed by Maleki and colleagues, we can find a plethora of functional enrichment strategies *(Maleki et al., 2020).* The simplest of these tools employ the hypergeometric test (also known as Fisher's exact test). This test calculates the probability of randomly selecting x GO terms in a protein sample with size m when comparing to the distribution in a certain background population n. In short, it calculates the statistical significance of the proportions in a given contingency table. One reason for the extensive use of GO data is its simplicity. FEA can be used without deep understanding of bioinformatics or biostatistics *(Huang et al., 2009).* However, GO annotation is not exempt from biases and limitations that affect FEA outcomes (discussed in *(Gaudet and Dessimoz, 2017).* The two most popular pitfalls when functionally characterizing a gene set are (i) the correct assignment of the background universe and (ii) the management of redundant outputs.

It should be noted that the GO knowledgebase collects the functions associated to all the proteins in the proteome without considering tissue or cell-specific contexts. The inexperienced users frequently dismiss that the default enrichment analysis evaluates the entire GO annotation as the distribution background. However, the statistical enrichment must always be calculated against the experiment-specific gene universe. Another common oversight is the representation of FEA results. Many users opt to summarize the analysis by outlining the top enriched functional terms, either considering the total number of proteins

annotated or the statistical p-value. However, the top GO terms do not necessarily describe the complete functional landscape of a given gene set.

At this point it is important to refer that the GO terms are organized in directed acyclic graphs **(Figure 1.5).** This organization represents the hierarchical relationship between the functions in such a way that the less specific terms connect the increasingly detailed functional annotations. Therefore, each gene is annotated with the most specific term but also with all the functional predecessors. Due to the hierarchical structure of GO data (further discussed in *(Gaudet and Dessimoz, 2017),* most FEA outputs return a considerable fraction of uninformative or redundant functional terms. In order to simplify the FEA outputs, the user can discard uninformative GO terms by applying different strategies.

Semantic similarity algorithms exploit the graph structure of the GO to identify the functions with less descriptive value. The hierarchical structure of GO enables the evaluation of the information content of a functional term according to its position in the graph **(Figure 1.5)**. For instance, top positioned terms describe general functions, while their descendants gradually incorporate more specific information. Likewise, the terms descending in the same branch are restricted to the same function in increasing levels of detail (Further discussed in *(Yu, 2020)).* **Revigo** (http://revigo.irb.hr/) is a web-based tool designed to simplify a predefined functional enrichment by applying varied semantic similarity algorithms *(Supek et al., 2011).* It does not require previous computational expertise and provides appealing visualization options. On the other hand, the **GOSemSim** R package implements numerous semantic algorithms to simplify functional enrichment outputs in R environment *(Yu et al., 2010).*

Although possible, the likelihood that two distinct functions are coordinated by the exact same protein groups is quite remote. While acknowledging that this may not always be the case, it can be assumed that two GO terms are highly redundant if they share a large fraction of the annotated proteins. We can evaluate the similarity in protein composition using **Jaccard's similarity** index **(Figure 1.5)**. The Jaccard index is defined as the size of the intersection (common proteins) divided by the size

of the union of the sample sets (total proteins in the two GO terms). When the two GO term sizes are very unbalanced it is preferable to apply the **Simpson's similarity** index since it divides the protein intersection by the size of the smallest GO term. Once the similarity index is calculated for the complete list, we can reduce the FEA result by merging the GO terms with high similarity.

Besides Gene Ontology, FEA methods can exploit different pathway-centric functional annotations as Reactome, KEGG or WikiPathways (benchmarking analysis of these databases for FEA is discussed in *(Mubeen et al., 2019).* These databases are more stable and contain less false positives than GO. However, ontology annotations give more flexibility and capture the complexity of protein functional relationships better. An additional benefit of GO annotation that we exploited in the last chapter of this work is that it employs a common vocabulary across different species and thus facilitates cross-species knowledge integration.

1.2.5 Disease-gene association databases

The prioritization of disease gene (DG) candidates requires a minimum set of known DGs. Prior to the annotation of disease-gene relations, it is equally necessary to establish a unified classification system to describe diseases. To that purpose, the **Unified Medical Language System (UMLS)** (http://umlsks.nlm.nih.gov) integrates and classifies standard terminology regarding pathological stages in a similar structure as GO *(Bodenreider, 2004).* In turn, UMLS also suffers from exceeding redundancy and benefits from the implementation of similar simplification methods as Jaccard's similarity.

DisGeNET (www.disgenet.org/) is one of the largest publicly available collections of genes and variants associated to human diseases *(Piñero et al., 2020).* It integrates varied types of evidence such as expert curated repositories, genome-wide association studies (GWAS), animal models or text mining of scientific literature. Moreover, the repository offers original metrics to assist the prioritization of genotype-phenotype relationships.

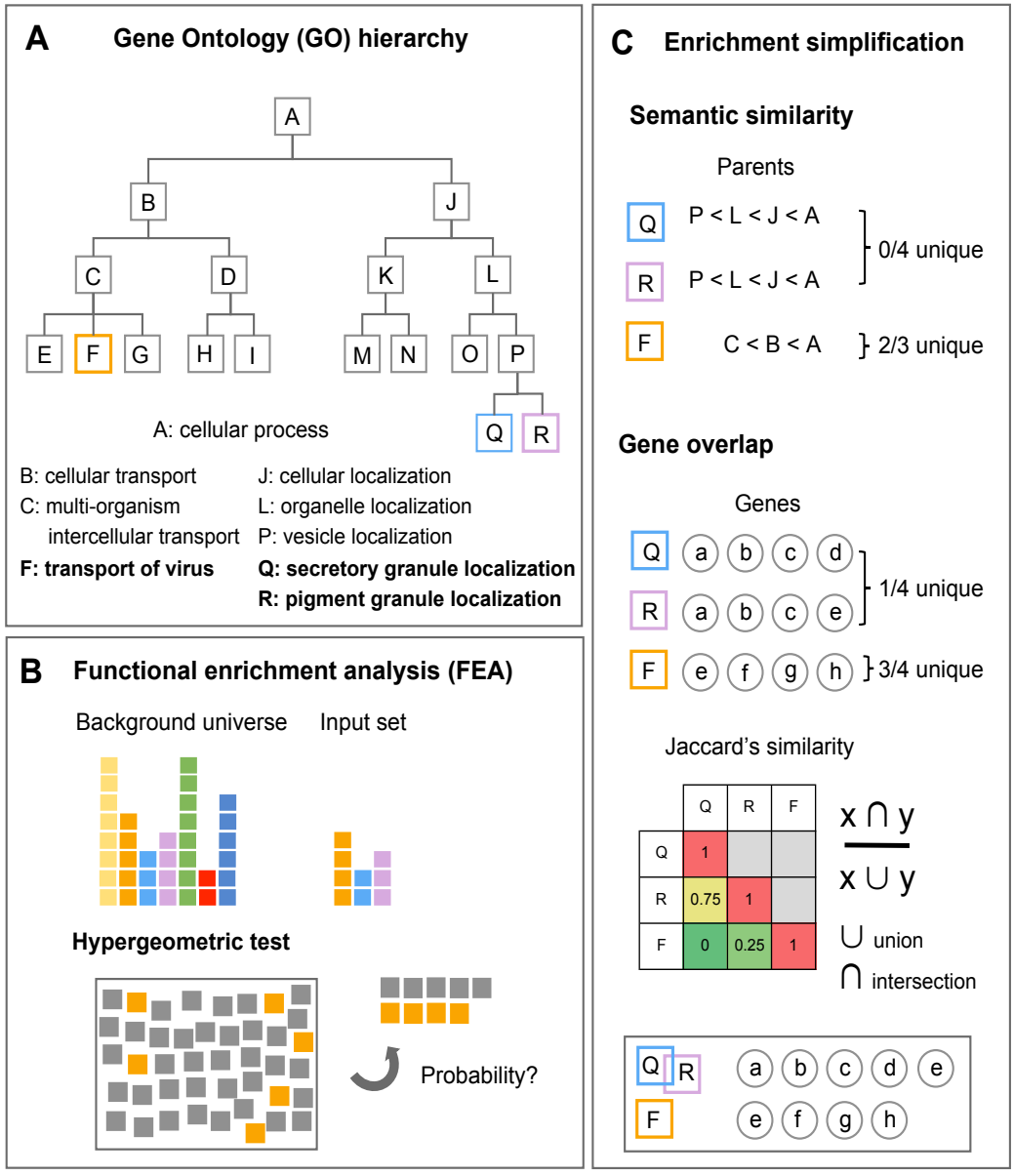


Figure 1.5 Gene Ontology (GO) hierarchy and functional enrichment analysis (FEA)

(A) GO hierarchy. Functional annotations are organized in increasing detail of description. (B) The hypergeometric test is the simplest functional enrichment analysis (FEA) method. It evaluates the probability of retrieving a number elements of a given input set from a known background distribution. (C) FEA might return redundant results. User can apply different simplification strategies such as simplification by semantic similarity or gene overlap. Semantic similarity evaluates the information content in GO hierarchy. Jaccard's similarity coefficient can be applied to evaluate gene intersection. Both methods return a coefficient that the user can adjust to combine redundant functional terms into simplified groups.

On the other hand, Online Mendelian Inheritance in Man (**OMIM,** www.omim.org/) is a knowledge based, manually curated and daily updated repository of human genes and genetic phenotypes *(Amberger and Hamosh, 2017).* Initially focused on monogenic disorders, it nowadays includes information on multifactorial diseases. However, the lack of a controlled vocabulary and consistent annotations hampers the retrieval of information.

1.2.6 *Drosophila* databases

Due to obvious ethical reasons, biomedical research relies on experimentation in animal models. As introduced in **Section 1.1.3,** *Drosophila* melanogaster is one of the most popular invertebrate models in neuroscience research. In order to retrieve valuable insights from fly models, translational researchers require detailed information concerning the relationships between predicted human and fly orthologs. In that sense, the **Integrative Ortholog Prediction Tool (DIOPT)** (flyrnai.org/diopt) is a valuable tool that facilitates the recovery of human and fly orthologs *(Hu et al., 2011).* DIOPT integrates the various ortholog predictors publicly available to calculate orthologous gene-pair relationships. Beyond gene orthology, **FlyBase** (flybase.org/) is the most popular repository of information on experiments conducted in *Drosophila* *(Larkin et al., 2021).* FlyBase gathers literature research, reagent resources and high-throughput data derived from diverse arrays. On the other hand, **FlyAtlas2** (www.flyatlas2.org) collects gene expression data derived from RNA-seq experiments in isolated tissues and developmental stages of *Drosophila* melanogaster (Leader et al., 2018).

1.2.7 Multi-omics integration

Interactions between biomolecules are not restricted to each type of biochemical family. Thus, the integration of the assorted omic data such as genomics, metabolomics, epigenomics or microbiomics into multi-omic networks is essential to reconstruct a more complete vision of the system's organization. However, multi-omics data integration requires complex mathematical and statistical

tools to correlate the different types of biological data *(Subramanian et al., 2020).* Thus, the integration of multi-omics data is a long-term ambition that is beyond the scope of this thesis. Instead, this work takes a first step in evaluating the organization of tissue-specific protein interaction networks. **Table 1.2** collects the public repositories and databases employed in the work presented in this thesis.

Table 1.2 Summary of omic datasets used in the thesis

Brief description of the public repositories used to access the omic data employed in the work presented in the thesis

Biological knowledge resource		Public repository	Repository aim	Repository URL
Omic	Transcriptomics	GEO	Raw and pre-processed transcriptomic datasets	ncbi.nlm.nih.gov/geo
	Proteomics	ProteinAtlas	Multivariate protein expression data	proteinatlas.org
	Interactomics	APID	Compendium of protein-interaction databases	apid.dep.usal.es
		CORUM	Manually curated protein complex interaction data	mips.helmholtz-muenchen.de/corum
Other	Functional-associations	Gene Ontology	Gene-functional associations	www.flyrnai.org/diopt
		REVIGO	Functional enrichment analysis simplification tool	http://revigo.irb.hr
	Disease-associations	UMLS	Unified medical language system	http://umlsks.nlm.nih.gov
		DisGeNET	Disease-gene association database	www.disgenet.org
		OMIM	Disease genomic and phenotypic information	www.omim.org
	Drosophila knowledge	FlyBase	Drosophila-centric multivariate database	www.flybase.org
		FlyAtlas2	Drosophila transcriptomic database	www.flyatlas2.org
		DIOPT	Fly-human ortology database	www.flyrnai.org/diopt

As will be expanded in next section, network-based approaches are essential to explore and interpret the outcomes derived from experimental omic datasets. The characterization of functional and physical interactions occurring between proteins can reveal relevant molecular players in normal and disease cellular states.

1.3 Network biology

Systems biology investigates how the relationship between the elements of a complex system gives rise to the biological phenomena observed in nature. Thus, systems biology moves the study of interest from single molecular elements to the characterization of their interactions. Graphs became cardinal to represent complex interactions and model the outcomes of these systems. A **graph diagram** is a mathematical structure used to represent collections of discrete objects and the relationships between them. The representation of molecular interactions in graphs not only facilitates the integration of complex data but also the application of graph theory concepts to predict biological phenomena. **Graph theory** was first introduced in 1736 by mathematician Leonhard Euler and aims at studying the topological or structural properties of graphs. However, it was not until the early 2000's that Barabási and Olvait realized the potential benefits of implementing graph theory in biomolecular research *(Barabási and Oltvai, 2004).* Although they are not strictly synonyms for simplicity, the terms graph and network will be used in this thesis with the same meaning.

While the previous section presented the conceptual framework of systems biology and discussed the omic approaches employed in this thesis, this section is focused on theoretical and practical notions of network biology. **Section 1.3.1** and **Section 1.3.2** introduce the basic terminology and network properties necessary for a more thorough discussion on the topic. Once the foundations are established, **Section 1.3.3** and **Section 1.3.4** describe the archetype structure of biological networks and its biological interpretation. Next, **Section 1.3.5** refers to the most popular applications of network biology and concludes discussing concepts particularly relevant for the design of the network-based methods presented in **Chapter 2** and **Chapter 3.**

1.3.1 Basic notions of biomolecular networks

A network is a representation of relations between discrete objects within complex systems. In cellular networks, the **nodes** represent macromolecules such as DNA, RNA, proteins or metabolites, and the **edges** depict their functional or physical relations. We find varying types of networks according to the type of biological information we intend to represent **(Figure 1.6A).** The edges can be **directed** or **undirected** depending on whether they represent directional interactions, such as enzyme-substrate **(Figure 1.6b5),** or undirected relations, such as physical protein-protein or RNA-protein interactions **(Figure 1.6b1, b4).** In parallel, the edges can be **weighted** or **unweighted**, depending on whether the interactions incorporate quantitative information. Edge weight can be a discrete variable representing categorical information as "repression" or "activation" (*e.g.,* -1, 1), or a continuous variable to describe interaction strength, stoichiometry relations, or co-expression correlation data, among others **(Figure 1.6b2).** Heterogeneous macromolecular interactions can also be portrayed in multilayered networks. In these networks, each layer is restricted to the interactions between the elements of a single class, while the inter-layer edges represent the relations across the distinct macromolecular entities. The most illustrative example are **metabolic networks,** which usually separate the enzymes from the substrate and products **(Figure 1.b4). Co-expression** or **metabolic networks** frequently incorporate quantitative data relative to expression correlation or stoichiometry data. However, it must be noted that the use of multilayered and/or weighted networks drastically increases the network analytical complexity. Except for mathematical modeling purposes, most studies in network biology, including the ones presented in this thesis, exploit undirected and unweighted single-layer networks.

Network connectivity

A **path** is the list of edges or interactions necessary to connect one node to another. In some cases, one edge might connect the same node itself. These **self-loops** are frequently found in regulatory networks to represent positive or feedback relations. In the case of PPI networks, these can depict homodimer complexes.

However, most analysis using PPI networks discard these self-loops to retrieve a so-called **simple graph**. The **shortest path**, as the name implies, is the minimum number of links one must traverse to move from one node to another **(Figure 1.6C**). The number of edges linking a node is termed **degree.** In directed networks we can further distinguish **out-degree** (edges leading away) and **in-degree** (edges incident to the node).

A network is said to be **connected** if there is always a path to connect any pair of nodes. When a network is unconnected, some subsets of nodes get inaccessible in **subnetworks**. In these cases, the largest connected subnetwork is commonly referred to as **main component (Figure 1.6C).** On the other extreme, the network can be completely connected meaning all elements directly interact with all the nodes in the set. Complete connectivity is rare on large networks and it is most frequently reserved to discrete subnetworks. These special types of subnetworks are named **cliques** and in PPI networks denote the existence of tightly connected protein complexes or molecular machines.

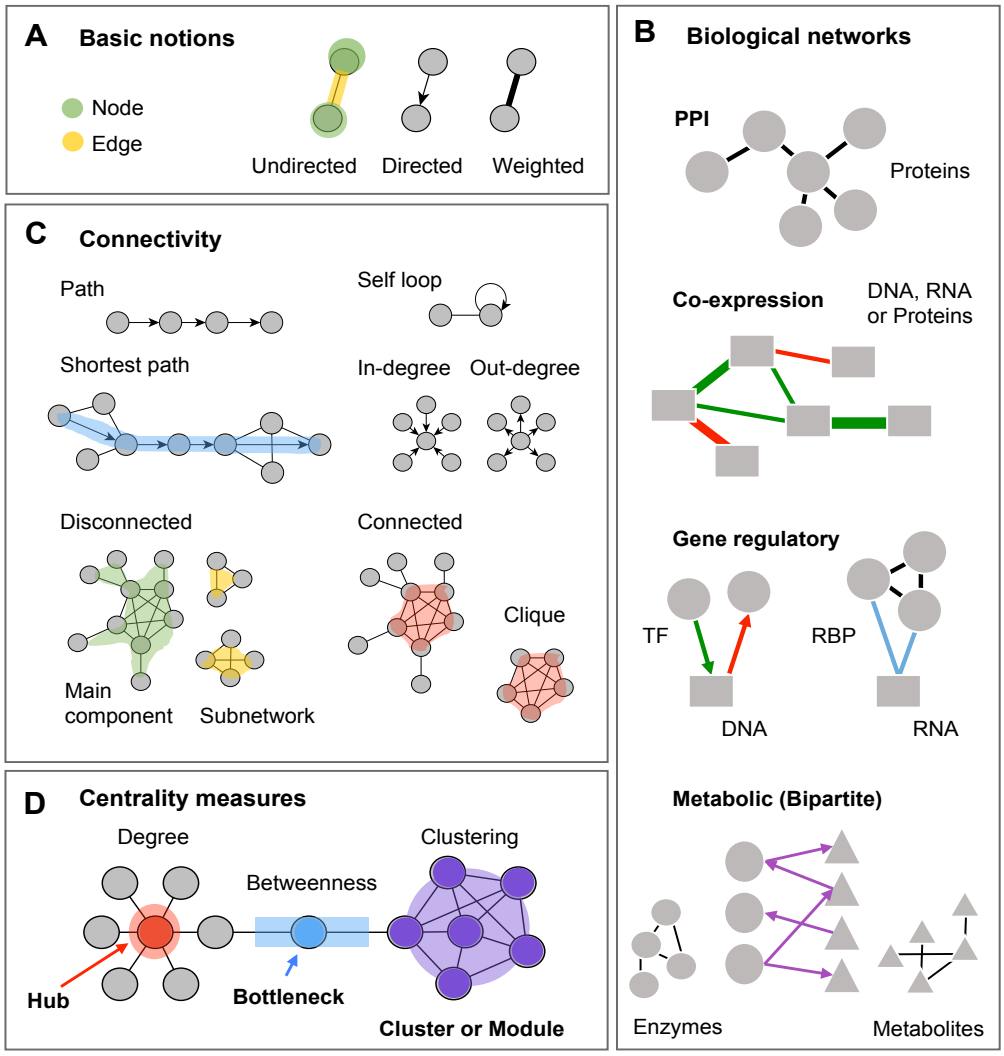


Figure 1.6 Basic notions of network biology

(A) The edges in the network depict (un)directed and (un)weighted interactions between distinct types of biomolecules, which are represented as nodes. (B) We find almost as many forms of networks as of types of existing biological interactions. In the most intuitive biomolecular graph, (b1) Protein-protein interaction (PPI) networks represent undirected physical interactions between proteins. (b2) Co-expression networks are undirected networks that interrelate transcripts/proteins with weighted edges according to their expression correlation profiles. The network can also incorporate heterogeneous molecular entities. (b3) Gene regulatory networks usually depict directed edges starting from the transcription factor (TF) protein to its target gene. (b4) Transcription regulatory networks frequently depict undirected interactions between RNA-binding proteins (RBPs) and their target RNAs. (b5) Metabolic networks are often represented in bipartite biomolecular networks that separate the enzymes from the metabolites, substrates or products. (C) Concepts used in network biology to describe network connectivity properties. Paths refer to discrete subsets of the network connecting two given nodes. In directed networks we distinguish nodes exerting (out-degree) or receiving interactions (in-degree), and nodes interacting with themselves (self-loops). The network is defined as connected when all the nodes can be reached from at least one path. Disconnected networks are divided into subnetworks and main component (largest subnetwork). A clique is a (sub-)network with maximum connectivity, being that all nodes establish an interaction with each other. (D) Most popular centrality measures. Node degree (number of interactions) reveals node connectivity distribution and points to connectivity hubs. Betweenness measures the nodes present in shortest paths and reveals connectivity bottlenecks. Densely connected nodes are identified using clustering algorithms.

1.3.2 Network topology

Network topology describes the arrangement of nodes interacting in the network and, as it will be discussed in the next section, it provides valuable insights to interpret biological properties. Network topology has been traditionally considered from a physical perspective. For instance, in the famous exercise presented by L. Euler in 1736, the *Köningsberg seven bridges problem*, network pathways are presented as bridges to cross a city river. In social networks, the interactions establish an information flow. In biological networks, many authors conceive connectivity as signal propagation, similar to the intercellular calcium signaling waves observed between astrocytes and neurons. Topology can be employed to describe overall network properties and infer mechanistic properties. On a local scale, network topology can be investigated to identify the decisive nodes for maintaining network architecture.

The **degree distribution** is an archetype feature to describe the network topology. The degree distribution of a network is the fraction of nodes in the network with degree k. As it will be discussed next, it is a critical property to define node connectivity. Another determining network property is the **clustering coefficient.** The **local clustering** coefficient measures the ratio between the observed and theoretical maximum degree of a node. It gives an estimation of edge density around a node. In turn, the **global clustering** coefficient is a network-wide measure of the average clustering coefficient, and it shows the tendency of a graph to be divided into clusters. Lastly, a **topological cluster or module** is a subset of nodes that have more edges within the cluster than edges linking to external nodes.

Centrality measures

The essentiality of nodes to maintain local pathways or network connectivity can be inferred from the number of interactions they establish, i.e., their location in the network. Currently we have at our disposal a vast choice of network centrality measures, based on varied node essentiality assumptions. To be concise we will only

describe the measures used in the context of the work presented in this thesis. For a more detailed discussion on the topic refer to *(Jalili et al., 2016).*

The simplest concept around centrality is that nodes with the largest degree (number of interactions) are central because they connect many elements. Therefore, the **degree centrality** of a node can estimate the ability of a node to spread a signal in its neighborhood **(Figure 1.6D).** Nodes with the highest degree are commonly referred to as connectivity **hubs.** On the other hand, it can be argued that a molecular signal would preferentially propagate through the shortest paths. Therefore, it is expected that the most central nodes will be involved in a large number of shortest paths. **Betweenness centrality** counts the number of shortest paths a node is involved in, and estimates the amount of information (as signal waves) that runs through the node. In turn, the nodes with highest betweenness are often regarded as connectivity **bottlenecks**.

1.3.3 Biological interpretation of network topology

Biological processes are exerted through the interaction of molecular components. From this it follows that the macromolecules directly interacting in the network are likely to be involved in similar processes. This assumption is commonly referred to as **'guilt-by-association'** and constitutes the foundation of innumerous network-based methods designed for protein functional prediction.

Biological networks - like many other real-world networks - reveal spontaneous **self-organization** properties, *i.e.,* the elements display non-random local interactions without external instructions. In the simplest model of random graph, the edges are drawn between pairs of nodes uniformly with the same probability *(Erdös and Rényi, 1960)*. The random arrangement of edges in an Erdös-Rényi network predicts that the degree distribution will follow a binomial distribution **(Figure 1.7A).** However, Barabási and Albert found that in many real networks across various fields, the node degrees establish quite different distributions. Most of the nodes display low degree values, while only a few nodes presenting an extremely high degree *(Barabasi and*

Albert, 1999). These networks were named **'scale-free'** because the node degree follows a power-law distribution (also called "*scaling-law*"), *i.e.,* the probability of a node displaying an extremely large degree is low, independently of the network scale. **(Figure 1.7B**). In addition to the non-random degree distribution, the same authors observed that nodes tend to cluster in densely connected modules. At first, the two observations seemed to contradict each other, being that the nodes in networks with high clustering coefficient should present a homogeneous degree **(Figure 1.7C).** However, further topological analysis revealed that the modules are arranged in a hierarchical organization, with small clusters embedded in increasingly larger modules *(Ravasz et al., 2002).* The so-called **hierarchical modularity,** or '*modules-within-modules*' topology, was essential to reconcile in a single architecture both the high modularity and the scale-free degree distribution **(Figure 1.7D).** Most important, this topology became a keystone to interpret some of the elementary properties observed in biological systems.

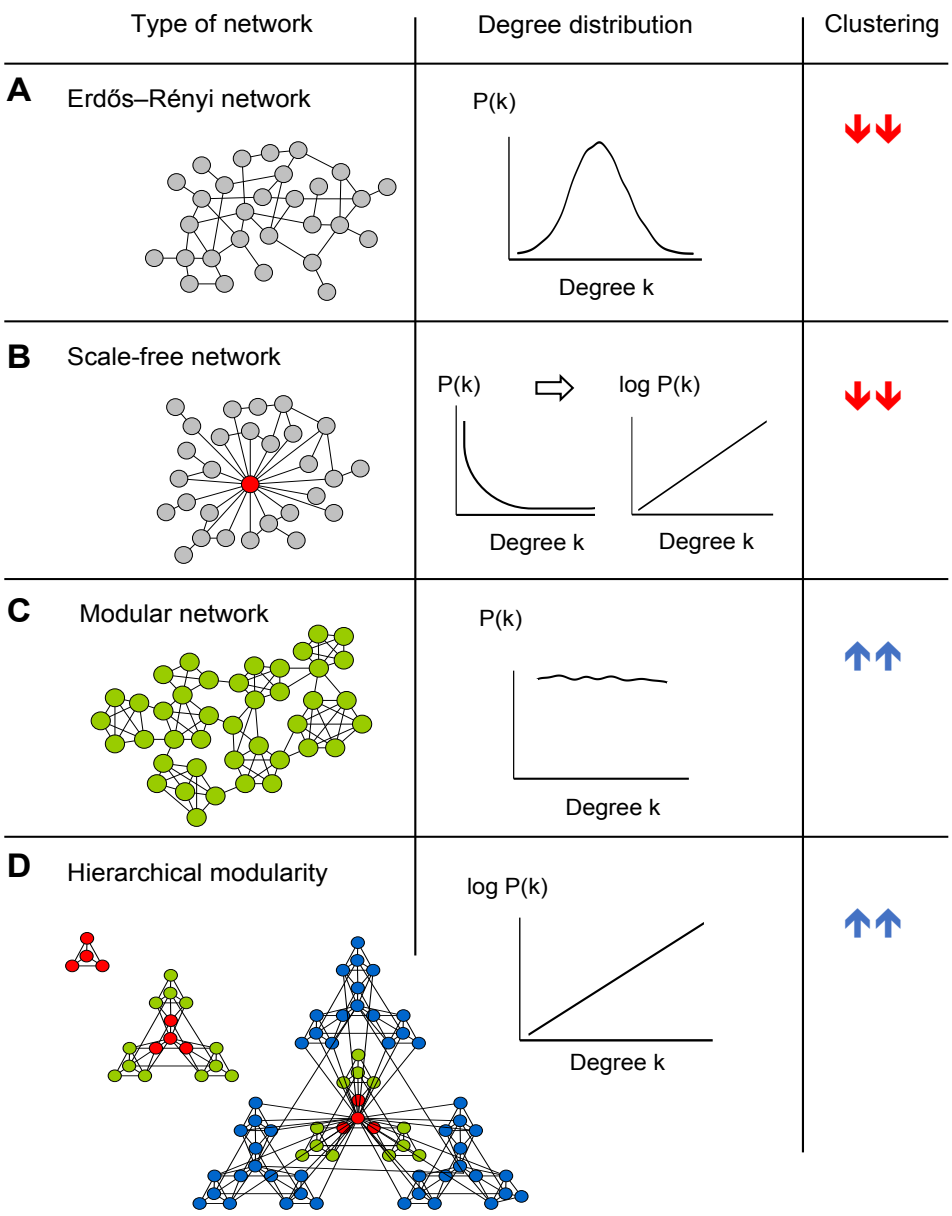


Figure 1.7 Topology of theoretical and real biological networks

(A) Erdös-Rényi network is the simplest model of random network in which degree probability follows a normal distribution and low clustering coefficient. (B) Barabási and Albert in 1999 found that most real-world networks display a scale-free network structure in which most nodes present scarce interactions, and only a few nodes present extremely high degree values. The degree distribution follows a power-law function in which the increase rate in degree is independent of the scale. Theoretically scale-free networks present low clustering coefficient, however biological networks present both scale-free topology and high clustering coefficient. (C) Conversely, modular topology theoretically presents high clustering coefficient but a homogeneous degree distribution (D) Hierarchical modular topology reconciles the two properties observed in real biological networks. Small clusters are integrated in increasingly larger clusters. The nodes at the center of the "cluster-within-cluster" network present extremely high degree values and, at the same time, the network maintains high clustering coefficient. Network in panel D is adapted from (Ravasz et al., 2002).

First, the scale-free topology drastically improves the network connectivity. With just a small fraction of highly connected nodes (hubs), most proteins can reach any node in the network in a few links. This feature is commonly referred to as **'small-world'** property. Second, modularity confers high robustness to the network, being that node depletion can remain restricted to certain network areas without affecting the overall connectivity. Furthermore, the robustness given by modular topology is determinant for biological systems to evolve and achieve high functional complexity (see early discussions on the topic by *(Barabási and Oltvai, 2004; Kitano, 2004).* In addition to the benefits this structure confers, biologists also speculate how these topologies might arise in nature. It is commonly accepted that gene duplication is the leading mechanism to generate the hierarchical modularity found in biological networks **(Figure 1.8A).** Initially, the product of a duplicated gene retains its properties and remains in the same interactome neighborhood, which in turn will increase the cluster density. Then, the low selective pressure on the duplicated macromolecule will enable the gene to mutate and acquire novel properties in the original module and, in a long evolutionary scale, even generate novel functional modules.

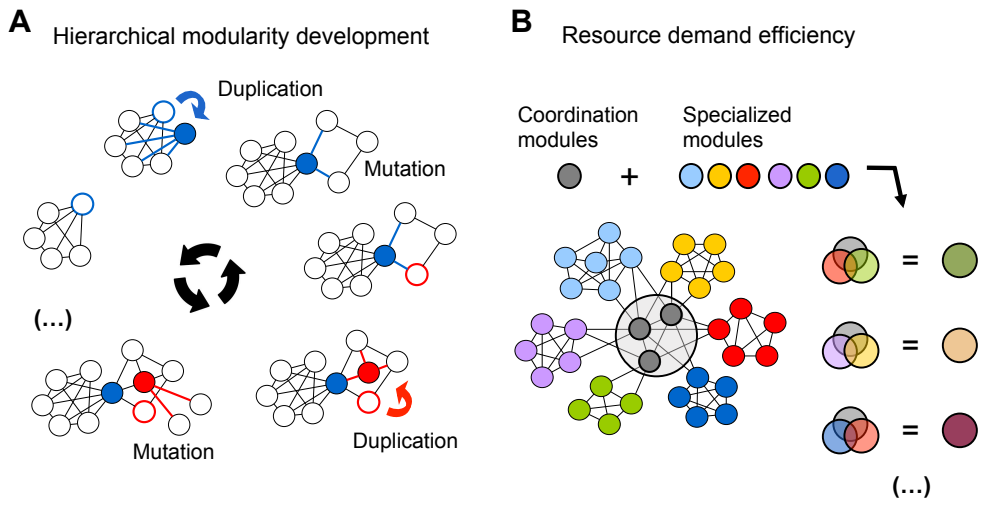


Figure 1.8 Biological insights inferred from hierarchical modularity topology

(A) Hierarchical modularity arises from gene duplication events. Gene duplication decreases selective pressure on the duplicated genes. The duplicated gene maintains the original interactions but can mutate to establish new ones. At an evolutionary scale, the duplication-mutation cycles increase network modularity while enabling the organization of new modules. (B) The organization in hierarchical modules reduces the demand for resources. Discrete modules can exert transversal functions while the specialization of a few nodes can link different modules to coordinate increasingly complex responses.

Together with the small-world property, the hierarchical organization of biological networks is an advantageous mechanism to coordinate complex functions and **reduce resource demands (Figure 1.8B).** Making use of the 'guilt-by-association' principle, it can be assumed that each module is specialized in the regulation of discrete molecular functions, while the inter-cluster interactions will regulate the coordination of more complex responses. In this sense, the inter-cluster connectivity will enable the maintenance and reuse of basic biochemical tools in different pathways or complex processes. For instance, the transcription of a novel gene would only require the customization of a novel regulatory route, while the transcription machinery could remain invariant. From this rationale, it also translates that module size and position in the network can be related with its level of functional specialization, such that the most essential processes will be more densely connected and located at the center. Nonetheless, it must be remarked that this is an oversimplified view of biomolecular systems and, therefore, of the current notions in network biology. Indeed, most of the fundamental network biology principles are still under heated debate, with ongoing research often generating conflicting observations.

1.3.4 Network modularity at the center of debate

Network modularity is still at the center of a vivid discussion for two main reasons. Modularity can be defined using different types of biological information returning non-consistent modules. In the same way, biological information is likely to have biases that distort the actual shape of the modules.

Clustering analysis based on node degree directly returns the archetypical **topological clusters or modules**. However, these are not the only clusters one can find in a biological network. For instance, based on the 'guilt-by-association' principle, proteins involved in the same biological process will tend to interact more densely and thus form **functional modules.** These modules are usually identified exploiting GO functional enrichment strategies (discussed previously in **Section 1.2.4)** to select

groups of interactors. The same applies for **disease modules** in which the products of DGs tend to densely interact.

The second point of discussion is that the biological relevance of topological modularity might be overestimated. The high clustering coefficient observed in seminal studies of network biology might be a topological distortion due to the technical biases in PPI detection towards stable interactions *(Acuner Ozbabacan et al., 2011)*. It remains nonetheless evident that molecular machines and stable macromolecular protein complexes are essential for cell biochemical and biomechanical functions *(Ghadie and Xia, 2022).* However, we might be overestimating their relevance when interpreting physiological outcomes. In fact, current lines of investigation argue that transient interactions (and so, more sparsely connected elements) are indeed central to coordinate even the most primordial functions. Of note, during the past decades, prior to the standardization of high-throughput methods, protein interaction studies primarily focused on proteins with anticipated clinical interest *(Skinnider et al., 2018).* Thus, PPI networks originally grew around DGs, placing them at the center. Furthermore, researchers have preferentially studied the closest neighborhood of DGs and this might have artificially exaggerated the DG clustering into modules.

Protein multifunctionality and module overlap

At this point, it is imperative to highlight a serious limitation when evaluating the topology of PPI networks: they are static representations of cellular dynamic processes. As matter of fact, PPI networks represent the likelihood of a certain physical interaction to happen. However, they cannot inform which is the collection of interactions happening at a specific location and time. To this limitation, it is important to add the fact that the body of evidence collected so far soundly supports the view that most proteins are multifunctional.

This notion is intuitive when considering that biological functions are not exerted by single molecules but, in contrast, are collective properties of the system. As we have stated in the previous subsection, biological systems will tend to minimize

resource efforts and so the combination of different proteins from the same proteome set can give rise to different biological processes **(Figure 1.8B).** From this it translates that proteins can be simultaneously involved in several functional modules and, in turn, that the modules will tend to overlap. This observation is particularly determinant when using canonical clustering algorithms because, by definition, they are designed to identify isolated non-overlapping clusters. While it is true that substantial efforts have been made to adapt clustering algorithms to current biological challenges, it is no less true that the main goal of these algorithms is still to discretize protein connectivity and therefore, that they may still draw biologically incomplete conclusions *(Alcalá-Corona et al., 2021).* In fact, it is very likely that topological, functional and disease modules represent different facets of the same biological reality. Thus, this discussion is fundamental to correctly interpret the self-organization patterns observed in biomolecular networks. Furthermore, in practical terms, these notions directly shape the conceptual design of functional and DG prediction methods.

1.3.5 Network medicine and network biology applications

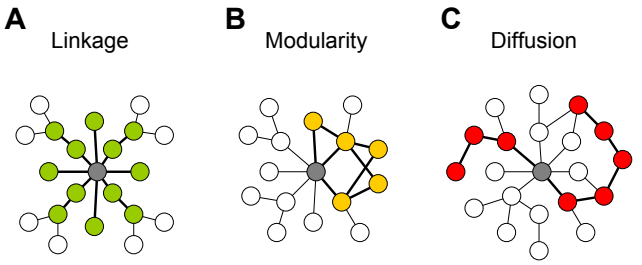
Diseases are considered to be the phenotypic manifestation of cell homeostasis perturbations, which in turn are triggered by molecular alterations disrupting the healthy functional routes. The identification of network vulnerabilities and the understanding of node inter-dependencies is prime to correctly predict the deleterious outcomes of hypothetical perturbations in the network. On this basis, we can design strategies that exploit network topology to predict DGs, biomarkers and drug targets (network medicine applications reviewed by *(Lee and Loscalzo, 2019; Silverman et al., 2020)).*

Network-based DG prioritization methods

Although still an open debate, it is commonly accepted that DGs often lie in a close neighborhood in the network *(Goh et al., 2007; Menche et al., 2015a)*. Thus, the rationale to design DG prioritization methods is quite similar to the approaches aiming at predicting gene product functions. In particular, in the case of DG

prioritization methods, we find an extensive range of strategies among which three main types can be highlighted **(Figure 1.9).**

Figure 1.9 Popular network-based gene prioritization strategies



Most DG prioritization methods are founded on the guilt-by-association principle and so aim to prioritize the nodes closest to known DGs. The proximity of nodes to DGs can be interpreted from varying assumptions. (A) Linkage-based methods exploit centrality measures and prioritize nodes central to reach DGs. (B) Modularity-based methods assume that topological, functional and disease modules overlap and so apply clustering methods to identify modules around DGs. (C) Diffusion-based methods consider that the impact of a gene alteration can be modeled as a finite signal propagation, similarly to a liquid diffusing in a pipe system. Diffusion-based methods model the propagation of the alteration from DGs to prioritize the nodes most frequently found within the impact area.

Linkage-based methods rely on centrality measures to identify direct interactors and critical connectors of previously known DGs **(Figure 1.9A).** In fact, the **S2B method** presented in Chapter 3 is a DG prioritization strategy based on betweenness centrality. Its novelty lies in that, instead of prioritizing DGs of a single disease, it aims at identifying common elements associated to a pair of diseases. **Modularity-based** methods assume there is a high correlation between topological, functional and disease modules, and so rely on clustering algorithms to identify disease modules around known DGs **(Figure 1.9B).** For instance, **DIAMOnD** applies the hypergeometric test to detect candidates significantly enriched in interactions with known DGs *(Ghiassian et al., 2015).* Lastly, **diffusion-based** methods simulate how DGs can propagate the alteration through the network **(Figure 1.9C).** These methods assume the signal has a finite propagation throughout a network, much as the diffusion of a volume of liquid in a pipe system. Thus, signal diffusion can be used to quantitatively estimate network connectivity and the proximity of proteins to certain source of alteration. We find many adaptations to diffusion analysis methods, depending on the biological rationale (reviewed in *(Di Nanni et al., 2020)*). One popular diffusion-based algorithm is commonly known as **random walker.** It simulates the random walk of a walker through the links of a network. The starting point of the walk is a known DG and the walker is enabled to walk a certain number

of links. After several simulated walks, the nodes more visited by the walker are considered more likely to be involved in the disease.

Network vulnerability in MND

From the modular organization of biological networks, it translates that the disturbance of a peripheral node or module should have a restricted impact on overall connectivity. On the opposite scenario, the perturbation of bottlenecks or hubs can trigger a cascade of failures with catastrophic consequences for cell homeostasis. Both from the evolutionary and functional perspective, most central modules will be implicated in the primordial and vital biological processes. On this basis, it might be intuitive to think that DGs preferentially locate at the most vulnerable network points, however, this assumption is likely excessive. As Barabasi and colleagues pointed out in their seminal work introducing network medicine, the removal of central nodes would have such catastrophic impact that the individual would not reach the last developmental stages preceding childbirth *(Barabási et al., 2011).* This notion is even more conspicuous for complex diseases with adult-onset and degenerative manifestations. At least in these scenarios, it is more likely that DGs are preferentially located in the most specialized functional modules at the network periphery. However, it is surprising to find that many genes associated with MND are particularly involved in essential processes for cellular homeostasis. For instance, as discussed in **Section 1.1.2,** the most common type of SMA, 5q-SMA, is caused by mutations in the SMN gene, which is a critical chaperone for spliceosome assembly, among other things. Similarly, numerous proteins coded by ALS DGs, such as FUS or TDP-43 are directly implicated in RNA biogenesis and transport. **So, an outstanding research question is how can these seemingly contradictory observations be reconciled.**

1.4 Thesis objectives and rationale

The main goal of this thesis is to **explore network-based approaches to unveil the molecular events governing MNDs**. As discussed throughout the introduction, MNDs encompass a spectrum of complex diseases with multigenic etiology. Although we currently have at our disposal an extensive list of genes related to MNDs, the community still faces many unanswered questions. This thesis focuses in exploring two compelling observations, namely:

- *How do MND phenotypes arise from the alteration of distinct pathways?*
- *How can the alteration of ubiquitously expressed proteins distinctively target specific cell types as MNs?*

On one hand, MND patients show mutations in genes involved in diverse cell activities but intriguingly, these alterations are manifested with similar clinical hallmarks. This implies that different molecular alterations can converge towards the disturbance of similar molecular modules. On the other hand, the different sensitivity of human tissues towards mutations in MN DGs indicates that the same protein can exert distinct functions depending on the specific cellular environment. These two observations are patent evidence of the complex protein interaction networks underlying cellular organization. In order to attempt to contribute to answering to these questions, the thesis work was structured into the following specific aims:

Chapter 2: Identification of network mechanisms underlying tissue functional diversification, and characterization of the network properties of functions prone to accumulate DGs in tissue- and disease-specific contexts.

Chapter 3: Identification of common molecular players between ALS and SMA predicted disease modules in human protein-protein interaction networks.

Chapter 4: Identification of common molecular players in *Drosophila* knockdown models of MND gene orthologs generated by a collaborative consortium led by the host laboratory.

Chapter 5: Integration of the knowledge retrieved from the human and *Drosophila* predictions to propose a unifying hypothesis of MND pathomechanisms.

Chapter 6: Discussion and final remarks on the investigation conducted in the thesis.

2 Biological Interacting Units identified in human protein networks reveal tissue-functional diversification and its impact on disease

Data presented in this chapter was included in the following work:

García-Vaquero ML., Gama-Carvalho M., Pinto FR., De Las Rivas J. (2022). Biological Interacting Units identified in human protein networks reveal tissue-functional diversification and its impact on disease. Comput. Struct. Biotechnol.

Author contributions:

MG-V conceptualized the study, developed the method, performed the analysis and wrote the manuscript. **FRP** assisted in the design of the study and reviewed the manuscript. **MG-C** supervised the study and reviewed the manuscript. **JDLR** supervised the study and reviewed the manuscript.

2.1 Abstract

Protein-protein interactions (PPI) play an essential role in the biological processes that occur in the cell. Therefore, the dissection of PPI networks becomes decisive to model functional coordination and predict pathological de-regulation. Cellular networks are dynamic and proteins display varying roles depending on the tissue-interactomic context. Thus, the use of centrality measures in individual proteins fall short to dissect the functional properties of the cell. For this reason, there is a need for more comprehensive, relational, and context-specific ways to analyze the multiple actions of proteins in different cells and identify specific functional assemblies within global biomolecular networks. Under this framework, we define *Biological Interacting units* (BioInt-U) as groups of proteins that interact physically and are enriched in a common Gene Ontology (GO). A search strategy was applied on 33 tissue-specific (TS) PPI networks to generate *BioInt* libraries associated with each particular human tissue. The cross-tissue comparison showed that housekeeping assemblies incorporate different proteins and exhibit distinct network properties depending on the tissue. Furthermore, disease genes (DGs) of tissue-associated pathologies preferentially accumulate in units in the expected tissues, which in turn were more central in the TS networks. Overall, the study reveals a tissue-specific functional diversification based on the identification of specific protein units and suggests vulnerabilities specific of each tissue network, which can be applied to refine protein-disease association methods.

2.2 Introduction

Cell physiology, defined as the ability to exert biological functions, emerges from the dynamic interactions in protein networks. Likewise, pathological manifestations arise from genetic alterations that result in protein interaction failure and network malfunction *(Barabási et al., 2011; Wang et al., 2012; Zhong et al., 2009)*. While great progress has been made towards the characterization of protein interactions (PPIs) and disease genes (DGs) *(Low et al., 2021; Piñero et al., 2020),* the relation between protein network connectivity and phenotypic manifestation is still poorly understood. The majority of diseases with restricted histological hallmarks are known to be triggered by DGs with wide tissue expression *(Hekselman and Yeger-Lotem, 2020)*. In that sense it still an open debate how mutations in housekeeping (ubiquitously expressed) genes can distinctively affect to the physiology only on certain tissues.

One fundamental reason for this knowledge gap is that biological networks are complex. Protein networks include large numbers of participating elements and these hold a large range of interchangeable partners. For instance, the complete human protein network available in the APID repository *(Alonso-Lopez et al., 2016; Alonso-López et al., 2019)* included in April 2021 more than 17,000 proteins, with each one being able to interact with more than 30 partners on average (http://bioinfow.dep.usal.es/apid/). In fact, the combinatorial range of PPIs is an eminent force for tissue functional diversification *(Deeds et al., 2012; Greene et al., 2015).* The same protein may establish different interactions and exert varied functional roles depending on the context *(Espinosa-Cantú et al., 2020).* As a consequence, the proteins will be localized at different positions in the network depending on their active functional partners in the considered tissues. On this basis, one could argue the same protein might acquire varying topological properties across TS-networks that distinctively resonate in TS-physiology. Indeed, DGs do not locate at random positions in the PPI networks but tend to display more TS-PPI in the disease tissue than in the unaffected tissues *(Barshir et al., 2014).* This observation suggests the idea that different TS-network may have distinct vulnerable spots, and strongly supports that the characterization of topological properties underlying tissue functional diversity might be critical to understand the emergence of TS patho-phenotypes.

Protein-protein interactions are a strong indicative of functional collaboration. Network connectivity measures, such as clustering coefficient, degree and betweenness centrality, are well-established predictors of protein essentiality and so of potential vulnerabilities in cell physiology ***(Barabási et al., 2011).*** Based on these notions, a variety of PPI network-based strategies have been proposed to identify densely connected modules, recently reviewed by ***(C. Liu et al., 2020).*** While these methods are valuable to predict functional collaboration and DG candidates, they are not suitable to characterize their topological context.

In this study we define and characterize *Biological Interacting units* (referred to as: BioInt-U), identified as biological modules found in PPI networks using tissue-specific mapping and topological interactomic analysis. In this way, BioInt units are found using a network-based framework to define topologically unbiased functional PPI consortia in multiple tissue-specific (TS) interactomes. These units represent an intermediate level of PPI functional coordination in TS networks, which allow the characterization of topological properties of normal and disease-targeted cell processes. A search for these BioInt units was performed within an extensive catalog of human tissues yielding 33 TS-BioInt libraries. Disease impact was assessed by mapping known disease genes (DGs) in BioInt libraries. The cross-tissue and cross-disease mapping revealed distinctive topological properties on the BioInt units, suggesting new explanatory insights into the occurrence of pathologies affecting specific tissues.

The benefits of using BioInt-U are illustrated, as an example study, by its integration with publicly available gene expression profiles (RNA-seq) derived from patients affected by two diseases: psoriasis and pulmonary fibrosis. Our analysis revealed that proteins corresponding to differentially expressed transcripts/genes (DEg) collaborate in the same BioInt units in expected disease tissues. Furthermore, these BioInt units were involved in biological processes previously considered critical in the development of these diseases (fibrosis and psoriasis), providing new potential research targets or candidate proteins to be modulated in these diseases.

2.3 Methods

2.3.1 Computational pipeline to define BioInt units

Reconstruction of TS networks. RNA-seq datasets representing 33 major tissues and organs were retrieved from Uhlén and colleagues work ***(Uhlén et al., 2015)***. The datasets were filtered to only evaluate transcripts expressed above 1 FPKM (Fragments Per Kilobase of transcript per Million). The dataset was TMM-normalized (Trimmed mean of M values) using the limma R package ***(Ritchie et al., 2015).*** Biological replicates were combined calculating the average transcript expression. Next, human physical PPI data reported at least in two experiments was retrieved from the APID repository in April 2021 ***(Alonso-López et al., 2019).*** The tissue-naive PPI network was filtered to create a TS network including only interactions between proteins coded by transcripts expressed in each TS transcriptome. The TS networks were simplified to remove self-loops and isolated proteins using the igraph R package ***(Csárdi and Nepusz, 2006).***

Functional enrichment of TS networks. The GOfuncR R package was used to functionally characterize TS networks in comparison to the unspecific network using Gene Ontology Biological Process (GO-BP), hyper-geometric test, FDR = 0.1 on 500 randomizations ***(Grote, 2020).*** Functional enrichment was simplified into functional groups by collapsing terms with more than 0.9 Jaccard's similarity coefficient, defined as the number of common elements between two sets, divided by the union set size. When GO-BPs are collapsed, the new functional group functional description with fewest characters.

Generation of TS-BioInt libraries The functional enrichment of TS-networks was used to identify the BioInt units, which consist of groups of proteins physically interacting and annotated under the same enriched GO-BP term. The inconsistencies across high-throughput PPI data and the constant PPI data growth in multiple repositories suggest the human interactomic data is still far from complete. Knowing this, we enabled BioInt units to be formed by non-connected subnetworks. The

isolated clusters were discarded only when the main component (largest subnetwork) represented more than 90% of the total BioInt unit. On the other hand, proteins can display transient and varied PPIs. Additionally, most proteins are multifunctional and are frequently annotated with several GO-BP terms. In order to recapitulate protein multifunctionality and the network dynamics, we enabled proteins to be involved in several units simultaneously. The BioInt units were classified in 28 functional categories by performing a direct text mining of key words found in the description of functional units. The list of key terms is available at **Supplementary Data 2.2.** From the total 28, we selected the 22 functional categories associated with sufficient BioInt units. The network topological analysis was focused on betweenness, degree and clustering coefficient measures that were evaluated using the igraph R package *(Csárdi and Nepusz, 2006).*

2.3.2 CORUM protein complex intersection

The molecular machines described at the CORUM repository *(Giurgiu et al., 2019)* were used as gold standard to evaluate the ability of BioInt-U method to identify already established protein functional complexes. Curated 'core' complexes were retrieved from CORUM database in March 2021. The dataset was filtered to only evaluate CORUM complexes including at least 3 distinct proteins. In order to assess the protein overlap between CORUM and BioInt-U, we first combined TS BioInt libraries into a unified version. BioInt units annotated to the same GO-BP term along different tissues were collapsed to include all TS proteins. The full description of unified and TS BioInt units is available in **Supplementary Data 2.4**. The average size of the BioInt units was >17 times larger than CORUM protein complexes. Due to the wide difference in size, the overlap analysis between CORUM complexes and BioInt units was performed applying Simpson's similarity (SS) coefficient, defined as the number of common elements between two sets, divided by the minimum set size (complete analysis available in **Supplementary Data 2.3**).

2.3.3 Disease-gene association

Disease gene (DG) associations were retrieved from the DisGeNET repository in December 2020 *(Piñero et al., 2020).* Disease references annotated as 'Symptom', 'Finding, 'Injury or poisoning' and 'Individual Behavior' were discarded. DGs with a

confidence score lower than 0.1 were discarded. Only diseases including 10 to 200 genes were evaluated. Similar to functional enrichment, disease list was simplified by collapsing terms with more than 0.9 Jaccard's similarity coefficient. When diseases are collapsed, the new disease group includes all genes associated to each pathology but is assigned to the disease description with fewest characters. In order to evaluate the performance of the BioInt-U framework, we created a list of 463 diseases with 11 presumable tissue-specific phenotypes. To generate the TS disease list, we used the same text mining approach as for the functional classification of BioInt units. The DG list and disease classification is available in **Supplementary Data 2.5**.

2.3.4 Gene expression profiles from public repositories

Two independent RNA-Seq transcriptomic profiles characterizing gene expression changes in samples derived from patients affected with psoriasis (GSE166388) and idiophatic pulmonary fibrosis (GSE24206) were downloaded from the Gene Expression OmniBus *(Barrett et al., 2013).* Differential gene expression analysis was performed using the GEO2R tool available through the GEO platform. Transcripts with fold change (FC) values of |log2FC|>2 and |log2FC|>1.5 and p-value <0.05 were selected as differentially expressed genes (DEg) in fibrosis and psoriasis datasets, respectively. The DEg datasets were mapped in the BioInt units to calculate the % of DEg by BioInt unit. Then, the BioInt units including DEg above the 3rd Quartile (0.9% and 1.3%) were considered the most potentially altered functional processes in fibrosis and psoriasis profiles, respectively.

BioInt-U method and output availability

All the analyses presented in this work were performed in R studio environment and figures were generated using ggplot2 and ComplexHeatmap R packages *(Gu et al., 2016; R Core Team, 2020; RStudio Team, 2016; Wickham, 2016).* The framework can be employed for other species and only requires PPI and TS transcriptomic data. The R functions necessary to generate additional BioInt units are available in Github repository **https://github.com/GamaPintoLab/BioInt-U.**

2.4 Results

2.4.1 Framework for dissecting functionally meaningful interactions: BioInt-U

The BioInt-U method was designed to identify groups of interacting proteins collaborating in the same biological processes (**Figure 2.1A**), *i.e.,* biologically interacting modules hereafter referred to as BioInt units. We first (i) reconstructed 33 tissue-specific (TS) networks by mapping TS transcripts/genes identified from TS RNA-seq profiles in a tissue-naive PPI network. Next (ii) the TS networks were functionally characterized by evaluating the Gene Ontology Biological Process (GO-BP) annotated in the network; and (iii) by applying functional enrichment analysis of GO-BP terms. The enriched GO-BP terms were then used to dissect BioInt units. The BioInt units did only retain the proteins enriched in the GO-BP that, at the same time were physically interacting in the TS network (iv) (**Figure 2.1A**). Assuming that some GO terms are very general and define too large and fuzzy functional groups, which are quite uninformative, we only considered BioInt units including less than 200 proteins. The use of BioInt-U in the TS networks returned 33 independent TS functional libraries, each including between 200 and 350 BioInt units (256 on average) (**Figure 2.1B**), each including a mean of 103 proteins (**Figure 2.1C**).

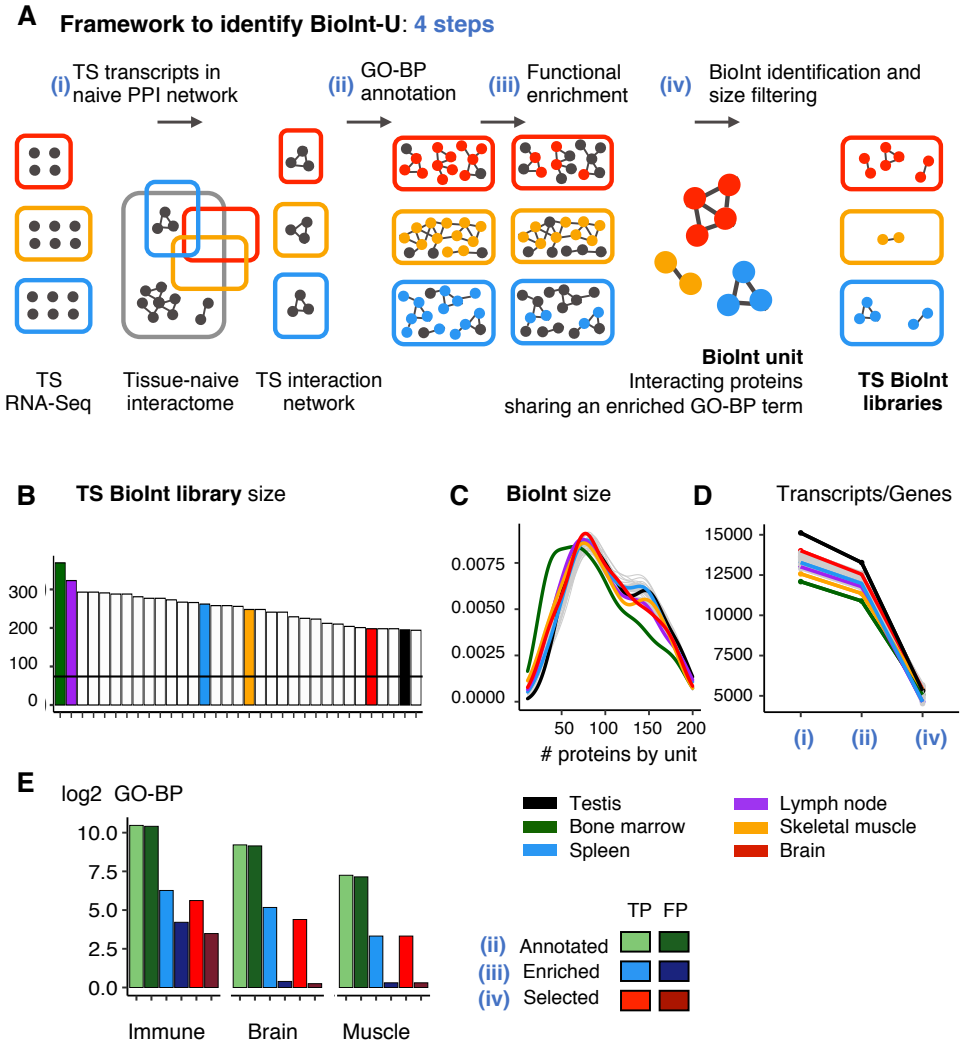


Figure 2.1 BioInt-U framework performance overview.

(A) Schematic illustration of BioInt-U framework. (i) First, we reconstruct tissue-specific (TS) protein-interaction (PPI) network by mapping TS RNA-seq profiles from 33 human tissue samples to tissue-naive PPI data. (ii) TS networks are functionally annotated using Gene-Ontology Biological Process (GO-BP) terms. (iii) TS networks are functionally enriched to keep only functions characteristic of each tissue compared to tissue-naive network. (iv) BioInt units are generated from the lists of enriched functions in each tissue context. The BioInt units are made by groups of proteins physically interacting and annotated by the same enriched GO-BP term. Only BioInt units including less than 200 proteins are selected to construct the 33 TS BioInt libraries. (B) Bar plot summarizing total number of BioInt units identified in each TS library. (C) Density plot describing the number of proteins incorporated in each BioInt unit in each TS library. (D) Line plot describing the transcript/gene recovery along the framework. X-axis points represent three of the steps defined in the framework (in panel A): (i) total transcripts/genes in TS RNA-Seq profiles; (ii) proteins in TS networks annotated with at least one GO-BP; and (iv) final number of proteins in the selected BioInt units. Step (ii) and (iii) returned very similar protein coverage (not shown for clarity). Colored bars and lines in panel B, C and D point to six illustrative specific libraries from tissues: testis, bone-marrow, spleen, lymph-node, muscle and brain. (E) Bar plot summarizing BioInt-U performance at identifying tissue-consistent BioInt units in three representative tissues: immune, brain, muscle. TP: Number of functions correctly annotated (ii), enriched (iii) or selected (iv) in the expected tissues. FP: Number of tissue-specific functions assigned to other than the expected tissues.

2.4.2 TS BioInt libraries recap functional landscape of TS transcriptomes

An ultimate goal in constructing BioInt libraries is to dissect how the interactome is coordinated into TS functional consortia. On this basis, we first corroborated that TS BioInt libraries recapitulate the functional landscape of TS transcriptomes resembling well-established biological properties.

We assessed TS transcriptome coverage at each step of the framework. A tissue-naive human PPI network retrieved from APID *(Alonso-López et al., 2019)* was found to incorporate > 90% of genes identified in each TS transcriptome *(Uhlén et al., 2015)* (see (i) in **Figure 2.1D**). BioInt units are generated from Gene Ontology annotations, so the performance directly relies on the characterization state of the proteins. We found that > 80% of proteins incorporated in TS networks are functionally annotated (see (ii) in **Figure 2.1D**). Of note, the statistical functional enrichment did not affected TS transcriptome coverage (step (iii) omitted in **Figure 2.1D**).

In order to minimize shallow functional terms, only BioInt units with less than 200 proteins were selected for the BioInt libraries. This filtering step discarded ~50% of the enriched GO-BP terms and reduced the TS transcriptome coverage down to 40% (see (iv) in **Figure 2.1D**). **Supplementary Data 2.1** summarizes the properties of each tissue set in the successive steps. Knowing that vague and general terms of GO-BP tend to be associated with many genes, it is likely that large BioInt units, including numerous genes, are not functionally very informative. Therefore, we interpret that the genes/proteins we missed by filtering by size were only annotated with shallow terms (i.e., superficial in the Gene Ontology and rather general), and so are assigned to still poorly defined functions.

Despite the sharp decrease, we confirmed in the forthcoming analysis that the filtering of large BioInt units did not exclude tissue-specific annotations. To assess the ability of BioInt-U to characterize tissue-specific (TS) and housekeeping (HK) processes, we classified the BioInt units into 22 broad "functional categories" (**Supplementary Data 2.2**). For three representative tissues, we calculated the number of these "functional categories" correctly enriched to the expected tissue

(considered as true positive cases, TP) and the number of tissue-specific functions assigned to other than the expected tissues (false positive cases, FP) (**Figure 2.1E**). We confirmed that the functional enrichment was crucial to discard FP annotations, especially in brain and muscle libraries (**Figure 2.1E**, *blue bars*). Moreover, we confirmed that size filtering does not affect the selection of tissue-specific BioInt units (**Figure 2.1E**, *red bars*).

2.4.3 BioInt units represent functional assemblies beyond molecular machines

Molecular machines are commonly defined as "assemblies of molecular components that are designed to perform machine-like movements" ***(Balzani et al., 2000).*** The main components of many molecular machines are proteins/polypeptides, such as the proteasome, spliceosome, respiratory chain complexes, etc. Being that molecular machines lie at the center of every biological process, we expect BioInt units to incorporate them. To address this issue, we took advantage from the CORUM repository as gold standard of curated molecular machines ***(Giurgiu et al., 2019),*** and evaluated the degree of overlap with TS BioInt libraries. We first confirmed that ~90% of proteins involved in molecular machines are actually mapped in TS PPI networks and functionally annotated in GO-BPs **(Figure 2.2A**). CORUM protein coverage dropped when filtering-out BioInt units with more than 200 proteins, indicating that a fraction of CORUM-annotated molecular machines were only incorporated in the largest BioInt units. It is noteworthy though that the decrease in coverage was less pronounced than when considering overall transcripts possibly indicating that CORUM complexes are also incorporated in smaller BioInt units (**Figure 2.2B**). As expected, due to their central roles in cellular activity, we observed that proteins involved in molecular machines tend to be more ubiquitously expressed than proteins not identified as being part of any molecular machine in the CORUM repository (Wilcoxon Rank Sum test, p-value < 10^{-4}, **Figure 2.2C**).

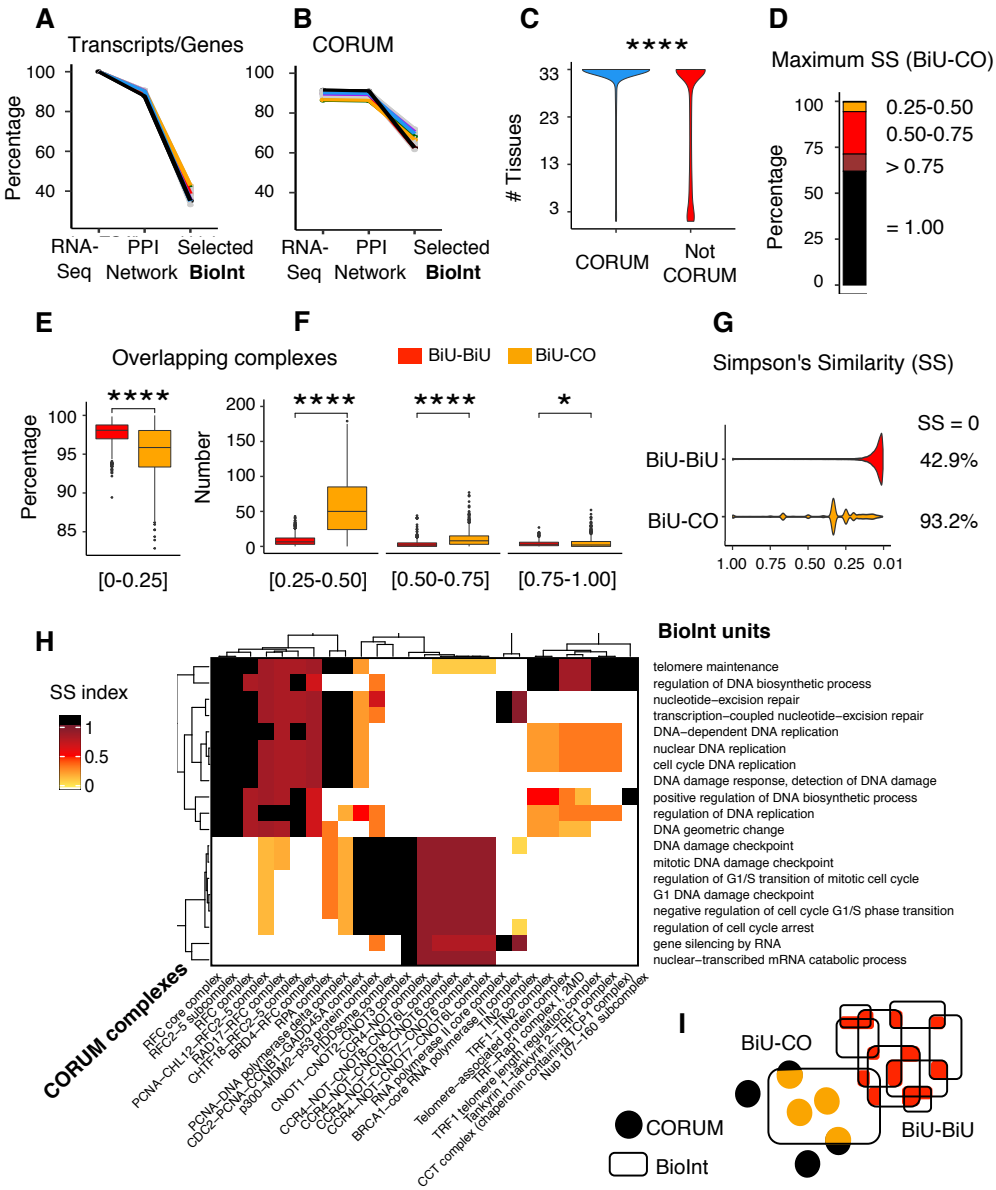


Figure 2.2 Mapping of molecular machines from CORUM repository to BioInt units.

Line plots describing mapping % of transcripts and proteins retrieved from each TS-RNA-Seq profile (A) or CORUM repository (B), in the TS PPI networks annotated with at least one GO-BP, and in the selected BioInt units. (C) Violin plot comparing the tissue expression distribution of proteins identified in CORUM complexes (blue) and transcripts only identified in TS-networks (red). Wilcoxon Rank Sum test, p-value <10-16. (D) Staked barplot summarizing the maximum Simpson's similarity (SS) index found when mapping BioInt units to CORUM complexes. Box plots describing the % (E) and total number (F) of BioInt units sharing proteins - i.e., overlap - with other BioInt units (BiU-BiU) and CORUM complexes (BiU-CO) at increasing SS index intervals. Wilcoxon Rank Sum test; p-value **** <0.0005 and * <0.05. (G) Violin plots comparing SS distribution between BioInt units and CORUM complexes. The distributions do not include pair comparisons with no overlap (SS=0). (H) Heatmap representing SS index between an illustrative subset of BioInt and CORUM complexes (BiU-CO). (I) Schematic picture of the predominant types of overlap found in the comparisons between BioInt units (BiU-BiU) and between units and complexes (BiU-CO).

In order to assess the protein overlap between BioInt units and CORUM complexes (BiU-CO pairs), we first combined the TS BioInt units into a unified library. The combination of all BioInt units identified along the 33 tissues returned a unified Bioint library consisting on 728 unique BioInt units including 7765 proteins overall. Due to the size imbalance between BioInt and CORUM complexes, the pair-wise overlap was addressed using Simpson's similarity (SS) index (see Methods). For each BioInt unit, we calculated the maximum SS index found with at least one CORUM complex (**Figure 2.2D**). Next, we compared the percentage and total number of overlapping complexes at increasing SS index intervals (**Figure 2.2E,F**). Lastly, we evaluated the SS index distribution along all pairs of complexes sharing at least one protein (**Figure 2.2G**). We first confirmed that all BioInt units partially intersected with numerous molecular machines (SS index > 0.25, **Figures 2.2D,E**). Further, we found that more than half of the BioInt units can partially incorporate up to 50 molecular machines (SS index 0.25-0.50 **Figure 2.2F**). Most notably, more than 60% of BioInt units displayed a SS index higher than 0.75 with at least 5 CORUM complexes in average (**Figures 2.2D,E**). **Figure 2.2H** illustrates several examples of BioInt units incorporating complete or close to complete molecular machines. We found that numerous CORUM molecular machines incorporated into single BioInt units were related to DNA and RNA metabolism. This is in good agreement with the fact that ribonucleic acid biogenesis and processing is exerted through successive biochemical processes that require the collaboration of multiple molecular machines. Overall, these results indicate that BioInt units can recapitulate how multiple molecular machines collaborate in more complex biological processes. **Supplementary Data 2.3** provides the full results of the pairwise SS analysis between BioInt units and CORUM complexes in human. **Supplementary Data 2.4** provides all properties and relevant information regarding the BioInt units generated from the analysis.

BioInt units are not redundant and recapitulate protein multifunctionality

We also addressed the overlap between BioInt units (BiU-BiU pairs) to evaluate functional redundancy and protein multifunctionality. We found that almost every BioInt unit slightly overlapped with at least one additional BioInt unit (SS index < 0.25, **Figure 2.2E**). Furthermore, ~57% out of the > 10^6 possible BiU-BiU combinations shared at least one protein indicating that BioInt units frequently overlap (**Figure 2.2G**). Notwithstanding, the SS index was consistently lower than the one observed for BiU-CO pairs. Likewise, the number of complexes overlapping with a SS index > 0.25 was significantly lower than when considering the BiU-CO overlap (Wilcoxon Rank Sum test, **Figure 2.2E**). The low but consistent overlap suggests that most proteins tend to be involved in varied functional consortia. Thus, the overlap analysis confirmed that BioInt libraries are not exceedingly redundant but rather recapitulate protein multifunctionality. Conversely, BioInt units incorporate complete or close to complete molecular machines characterizing the molecular activities at the center of biological processes (schematic interpretation in **Figure 2.2I**).

2.4.4 The functional landscape of tissue-specific BioInt libraries is consistent with the characteristic functions of each tissue

We next investigated whether TS BioInt libraries recapitulate the functional landscape expected for each tissue. To do so, the BioInt units incorporated in each TS BioInt library were first assigned to 22 broad functional groups (see Methods). Then, we evaluated the distribution of these 22 functional classes along the 33 reference tissues. The analysis corroborated that TS processes such as muscle or neuron-related processes are distinctively enriched in the expected tissues (hyper-geometric test, p-value <0.05, **Figure 2.3A**). This is shown, for example, for: neuron and brain, mitosis and testis, or muscle and heart. Conversely, transversal processes as signaling, DNA, RNA or protein metabolism were consistently identified across all the tissues, corroborating that these BioInt units are actually reflecting housekeeping (HK) functions. Notwithstanding, multiple HK functional classes were significantly enriched in particular organs. Direct inspection of these cases, however, reveals striking

agreement with known organ and tissue biology. Examples include the enrichment of signaling-related BioInt units in lung tissue, lipid metabolism processes enrichment in liver, or mitosis overrepresentation in testis. This result is in conformity with the general conception that different tissues rely more heavily on certain basal processes than others. Furthermore, the analysis revealed that the function types considered as HK can be divided in two subgroups based on their distribution across the tissues (**Figure 2.3B**). The majority of BioInt units related to RNA, mitochondria, organelle trafficking, protein metabolism and localization were essentially detected across all tissues (blue plots in **Figure 2.3B)**. In contrast, many functional groups as signaling, mitosis or cytoskeleton incorporated BioInt units with mixed expression patterns (purple plots in **Figure 2.3B**).

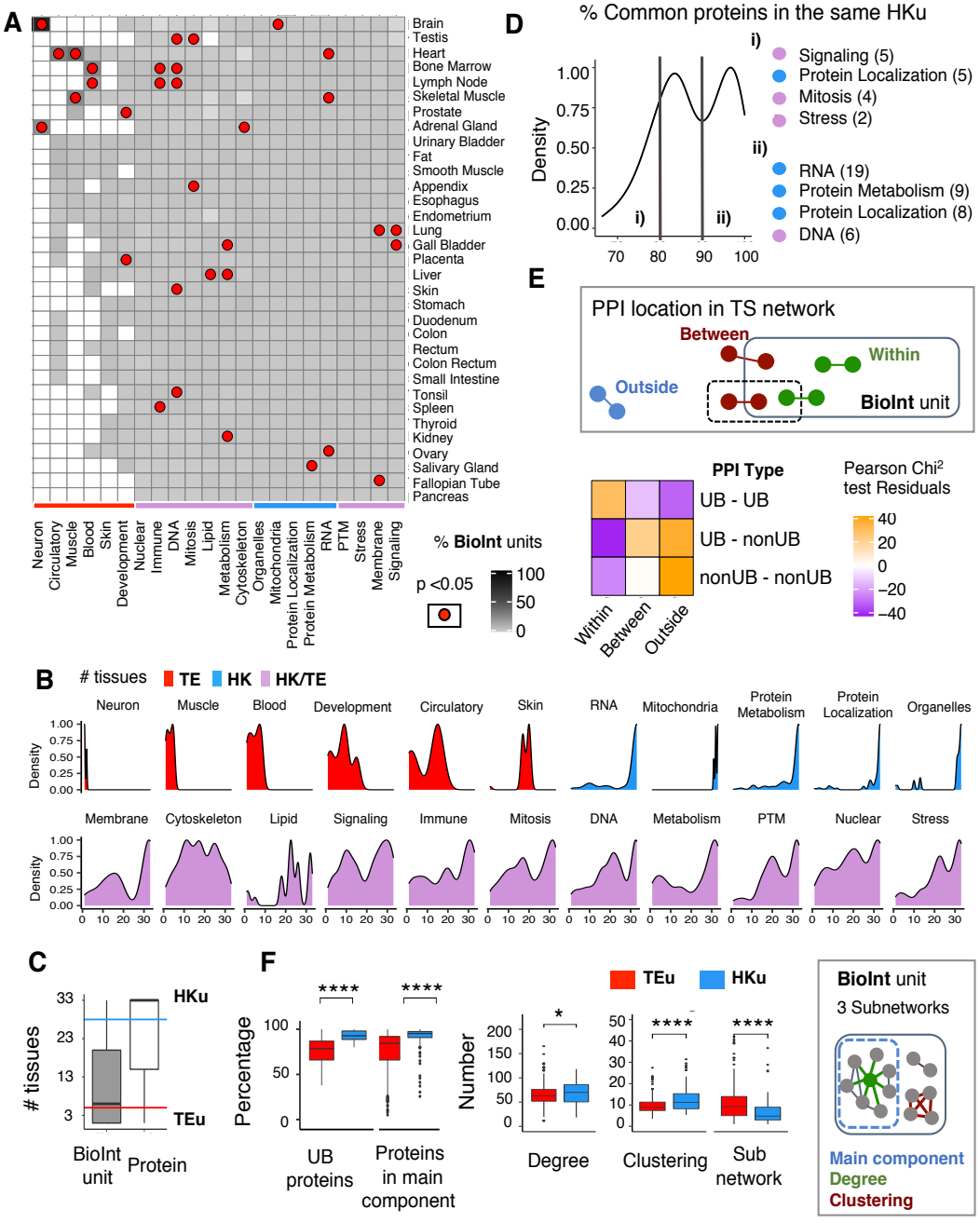


Figure 2.3 Analysis and comparison of functional and topological features of BioInt units with distinct tissue distributions

(A) Heatmap representing the percentage of BioInt units associated to each functional category (columns) along the tissues. BioInt units were classified into 22 general functional categories by text mining key words in the BioInt unit description (see Methods). Red dots point statistically significant enrichments, hyper-geometric test; p-value <0.05. Bottom column color divides function classes according to density distribution in panel B. (B) Density plots describing the tissue distribution of BioInt units assigned to each functional category. Density plots are filled in red, blue and purple to point functional classes with tissue enriched (TE), housekeeping (HK) and mixed HK/TE expression, respectively. (C) Box plot comparing tissue distribution of

transcripts and BioInt units. Colored lines separate the TE and HK units identified in less than 5 tissues or in more than 28 tissues, respectively. (D) Density plot describing the percentage of common proteins identified for the same HK unit along the tissues. Vertical lines highlight the two classes of BioInt units according to protein expression heterogeneity. (E) Heatmap representing the statistical association between the type of PPI and their location along BioInt units (illustrated in top box). Color scale represents the Pearson's residuals obtained from Chi2 test p-value <10-4. (F) Box plots comparing the network properties of HK units (HKu) and TE units (TEu) (illustrated in right box). Left to right: Percentage of UB proteins in each BioInt unit and total proteins in the main component (largest connected subnetwork) of the unit. Average protein degree, average protein global clustering coefficient in TS-network and number of subnetworks by BioInt unit. Wilcoxon Rank Sum test; p-value **** <10-4 and * <0.05.

2.4.5 Dissection of BioInt units brings insight into the mechanisms underlying tissue functional diversity: Ubiquitous (UB) and non-ubiquitous proteins collaborate in HK and TE functions

Excluding the transcriptomic profiles of sexual tissues, we found that a large fraction of the gene transcripts (9,686 expressed genes) to be ubiquitously (UB) expressed across the TS transcriptomes. However, the distribution of BioInt units across tissues drew a notably distinct pattern when compared to transcript expression (**Figure 2.3C**). We found 357 BioInt units annotated in less than 5 tissues (hereafter-called tissue enriched units, TEu) and 122 units annotated in more than 28 tissues (housekeeping units, HKu). While all TS networks incorporated ~70% of UB proteins on average, the percentage of housekeeping units dropped to 17.3% (**Figure 2.3C**). These trends are likely justified by the observation that both HKu and TEu incorporated a mixed composite of UB and nonUB proteins (**Figure 2.3F**). In particular, we found that TE units incorporated a large percentage of UB proteins and HK BioInt units also included a small fraction of nonUB proteins.

Being that different proteins can exert similar biochemical activities, we hypothesized the same HK functional unit might incorporate varying proteins depending on the tissue of context. Additionally, we sought to assess whether the percentage of protein variability in HKu could be associated to their functional roles. We calculated the heterogeneity of each HKu as the percentage of proteins found in common along all tissues (**Figure 2.3D**). Protein variability analysis generated a bimodal density plot in which two major groups can be distinguished: i) heterogeneous HKu with more than 20% of protein variability and ii) highly consistent HKu with tissue variability below 10%. Similarly when evaluating BioInt unit distribution profile across tissues (**Figure 2.3B**), we found that heterogeneous HK units are more frequently

associated to functional classes with mixed expression patterns as signaling, mitosis or stress-related processes; while monotonous HK units are distinctively related to RNA, DNA or protein metabolism and localization. Nonetheless, these trends might indicate that functions considered as "mixed HK" are less characterized, at PPI and/or Gene Ontology level, than functions classified as consistent HK units. Either way, these results support the notions that, ubiquitous and non-ubiquitous proteins collaborate in TE or HK processes and some HK functions could acquire additional relevance in certain tissues.

2.4.6 Network characterization of ubiquitous and tissue-specific BioInt units

Protein localization at the global network can provide valuable information on the coding-gene evolutionary history and current functional essentiality. At the same time, the PPI location in the BioInt units can also indicate whether the protein interactions play a core role in a given function or, rather, coordinate complex functional mechanisms. On this basis, we evaluated the position of HK and TE units in the TS networks using standard network connectivity measures (right box in **Figure 2.3F**). We found that HK units frequently incorporated more proteins and these predominantly located in the main component (*i.e.,* the largest connected subnetwork of the BioInt unit, **Figure 2.3F**). Moreover, the proteins collaborating in the HK units displayed significantly larger average degree (lager number of interactions per protein) and global clustering coefficients (larger interaction density in the protein neighborhood), indicating that HK units hold central positions in TS networks (with a significant difference according to the Wilcoxon tests, **Figure 2.3F**).

To further explore the biological implications of the collaboration between UB and nonUB proteins, we addressed the frequency of homotypic (UB-UB or nonUB-nonUB) and heterotypic (UB-nonUB) PPI interactions at distinct locations in the TS networks: (1) outside units, (2) within the same unit or (3) between two units or one unit to outside (**Figure 2.3E**). We applied the Chi-square test to evaluate the statistical association between the type of PPI and its location in the network (p-value $<10^{-4}$) and used Pearson's residuals to describe the positive or negative association between the conditions. As expected given the ratio of UB and nonUB proteins along BioInt units,

we found that UB-UB interactions are more frequently located within BioInt units while UB-nonUB hetero and nonUB-nonUB homotypic interactions are significantly located outside the Biointunits. Most notably, heterotypic interactions also appeared frequently connecting the BioInt units with proteins outside the units. Overall, these results indicate that HK units are central in the TS networks and further; UB-UB PPIs lie at the center of BioInt units. On the other hand, the heterotypic interactions between nonUB-UB proteins seem to be key to link the functions in the network.

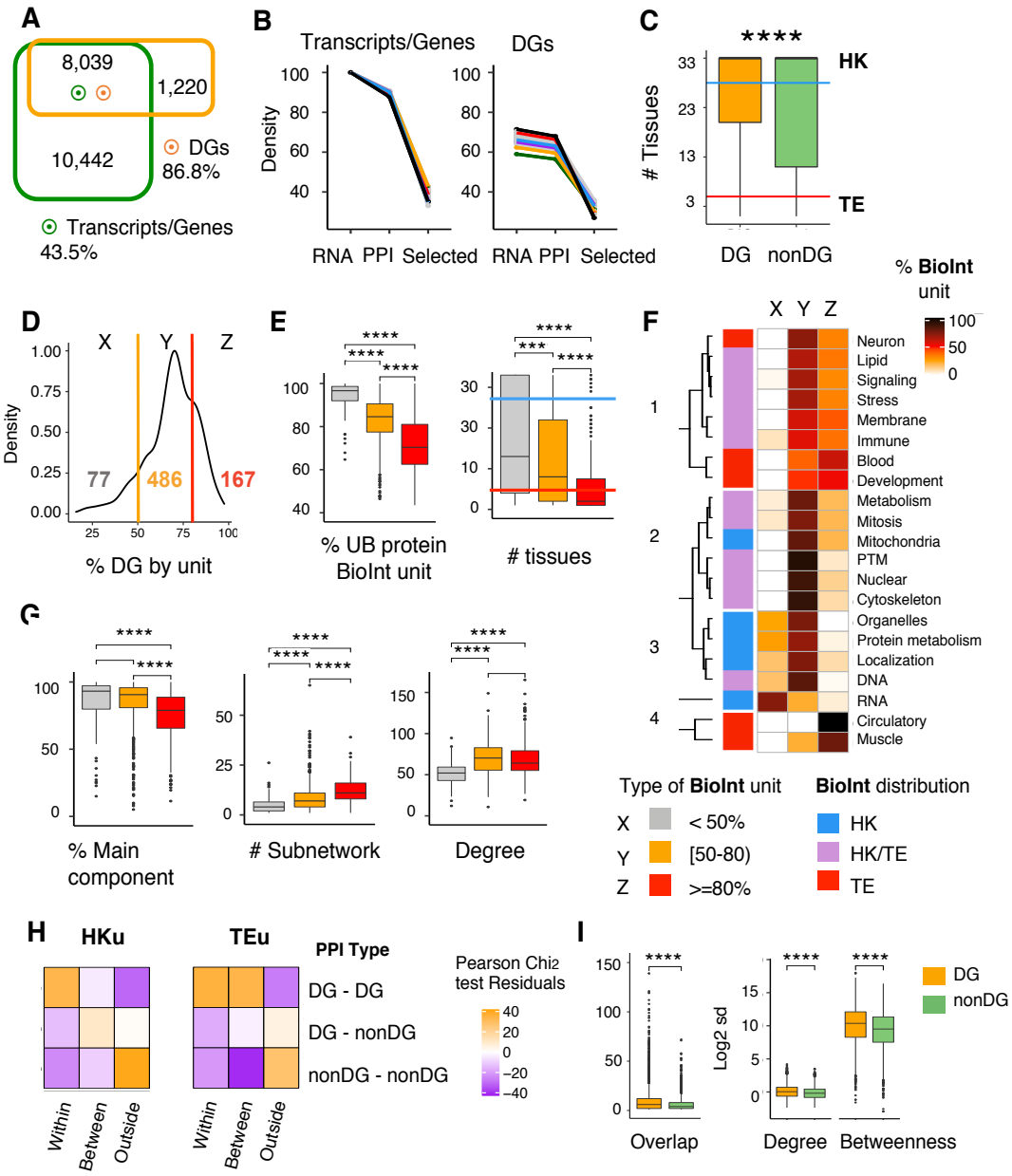


Figure 2.4 Systematic mapping of DGs into TS BioInt libraries.

(A) Venn diagram illustrating the overlap between our unified transcriptome and DGs from the DisGeNET repository. (B) Line plots showing the % of DGs and transcripts mapped to TS RNA-Seq profile, in the TS networks annotated with at least one GO-BP and in the selected BioInt units. (C) Box plots comparing tissue distribution of proteins encoded by nonDG (green) and DGs (orange) (D) Density plot describing the DG % identified in each BioInt unit overall. Vertical lines separate Biointunits in three groups based in DG %. (E) Box plots comparing the DG accumulation (X, Y and Z groups as defined in panel D) with the % of UB protein per unit tissue distribution. (F) Heatmap representing the broad functional classes (rows) assigned to the BioInt units including an increasing % of DGs (columns). Left dendogram and clusters result from a complete-linkage clustering using Euclidean distance. Left column summarizes the functional classes according to tissue expression patterns observed in Figure 3B. Tissue-enriched (TE, red), ubiquitously-expressed (HK, blue) and mixed HK/TE functions (purple). (G) Box plots comparing the protein % incorporated in the largest subnetwork of the BioInt unit (main component), the number of subnetworks by BioInt unit, and average protein degree. (H) Heatmaps representing the statistical association between the type of PPI and their location along HK and TE units. Color scale represents the Pearson's residuals obtained from Chi2 test p-value <10-4. (I) Box plots comparing the total overlap along BioInt units including or not including DGs and the standard deviation (sd) of degree and betweenness coefficients of DGs and nonDGs across TS networks (C, E, G and I) (Wilcoxon Rank Sum test; p-value **** <10-4 and *** <10-3).

2.4.7 Systematic mapping of disease genes (DG) in BioInt-U reveals potential large-scale topological vulnerabilities: DGs are widely expressed but accumulate in TEu

The preferential location of disease-associated genes (DGs) in the TS networks may bring valuable insights into sensitive points in network connectivity. To explore this, we collected 9,259 DG associations for 1,948 pathologies from the DisGeNET repository *(Piñero et al., 2020)*. Our global transcriptome covered 86.8% of DGs and conversely, 43.5% of transcripts (expressed genes) were associated to at least to one disease (**Figure 2.4A**). The DG coverage was barely affected when considering the proteins in the TS-networks but notably dropped in the selected BioInt libraries (**Figure 2.4B**). This indicates once again that a fraction of DGs is only incorporated in units including more than 200 proteins. Furthermore, we found that more than 55% of total DGs were ubiquitously expressed and overall, displayed a broader expression profile than nonDG proteins (Wilcoxon Rank Sum test, p-value $<10^{-4}$, **Figure 2.4C**).

However, when evaluating the DG% by BioInt units, we found that DGs preferentially acumulate in TE units with lower percentage of UB proteins (**Figure 2.4D, E**). In fact, the BioInt Units that incorporate the highest percentage of DGs are almost exclusively annotated in >5 tissues. Unexpectedly though, BioInt units accumulating >50% of DGs are more sparsely connected (i.e., exhibited an smaller main component

and more subnetworks), but incorporate proteins with more central positions in the TS network (**Figure 2.4G**). To evaluate whether the DGs tend to accumulate in any particular type of function, we took profit from the functional classification retrieved in **Figure 2.3B**, and found that BioInt units including 50-80% DGs were implicated in most types of functions (column Y in **Figure 2.4F**). However, BioInt units accumulating more than 80% of DGs were found to be more frequently related to TE or mixed HK/TE processes. Concordantly, the less targeted BioInt units were distinctively associated to HK biological processes.

2.4.8 Interaction of proteins encoded by DGs predominantly located between highly overlapping TEu

Having confirmed that tissue-enriched BioInt units (TEu) tend to incorporate more DGs, we next questioned whether the PPIs between DGs would present distinctive positions in the HK units or in the TE units, that might indicate topological vulnerabilities. We found that DG-DG interactions were more frequently located within BioInt units while nonDG-nonDG interactions were more frequent outside (**Figure 2.4H**). More important, DG-DG interactions were also notably located between TE units. We also found that DGs tend to be found in BioInt units presenting a larger overlap when compared to nonDGs (**Figure 2.4I**).

The connectivity properties of any protein directly depend on the range of PPI available at each TS network. This feature might be crucial to understand the variable impact the same DG can have in different tissues. We found that the variation (in terms of standard deviation) of betweenness and degree coefficients were significantly larger in DGs than other proteins not associated to any disease (**Figure 2.4I**). This observation may provide critical insights into the mechanisms underlying TS disease-phenotypes linked to DGs with wide distribution.

2.4.9 Genes associated with TS diseases accumulate significantly in BioInt units characteristic of the target tissue

To further explore the mechanisms underlying the emergence of TS patho-phenotypes, we next proceed to evaluate the DG mapping at disease-specific and tissue-specific levels. From the 1,948 diseases annotated in DisGeNET, we identified 463 diseases unambiguously associated with 11 tissues (for example, "nefrotic failure" is a kidney dysfunction or "T-cell lymphoma" is associated to alterations in the immune system). The complete list of disease-tissue associations and DGs is available in **Supplementary Data 2.5**. It is reasonable to assume that the functions with most critical roles for a given tissue will accumulate more DGs found in the patient population. Likewise, it is also reasonable that a functional unit will only be efficient when a large fraction of its components are available in normal standards. Of note, the DG associations in DisGeNET do not only refer to causal mutations but also to biomarkers or de-regulated genes. Thus, it is plausible we could find several DGs simultaneously altered in the same patient. On this basis, we estimated the potential impact of each disease in the TS functions by addressing the overrepresentation of disease-specific DGs in each BioInt unit (hyper-geometric test, p-value <0.05).

Most diseases exhibit tissue-specific phenotypes from which it follows that DGs should accumulate in certain tissues in particular (hereafter referred to as "tissue-consistent" impact). We have also discussed in previous section that different cell types might specialize in certain functions. Based on this, we speculate the DGs of tissue-consistent diseases might accumulate in functional classes characteristic of given tissue physiology. The BioInt units enriched in tissue-consistent DGs were homogeneously related to almost all types of functions (**Figure 2.5A**). Thus, to increase the analysis resolution, we only considered Bioint units enriched in DGs for at least 10 tissue-consistent diseases (corresponding to top 3rd Quartile) (**Figure 2.5B**). We found that the BioInt units enriched in tissue-consistent DG lists are accordingly involved in functions specific to the tissues in consideration. This trend was most conspicuous in TE BioInt units related to immune, muscle or neuron functions, wich are predominantly enriched in DGs from tissue-congruent diseases. Likewise, the clustering analysis corroborated that tissues associated to the same broad histological groups (colored rows in **Figure 2.5B**) significantly accumulate DGs in units involved in

the same functional classes (columns in **Figure 2.5B**). This analysis further enabled to discern that several function classes considered as HK processes were distinctively altered in different tissues. For example, DNA-related functions appeared to be more frequently altered in female-organ diseases than in gastro-intestinal disorders. Likewise, stress and signaling-related functions were preferentially altered in immune system-related disorders.

2.4.10 BioInt units enriched in tissue-consistent DGs (BiU_{TC}) exhibit distinctive network properties

From the total 8,285 of BioInt units identified in the 25 TS libraries considered for DG mapping (**Supplementary Data 2.5**), 60% were significantly enriched in DGs of tissue-consistent pathologies (set named BiU_{TC} in the schematic representation in **Figure 2.5C**). Nonetheless, a 35.8% of these BioInt units were also enriched in DGs associated to diseases specific to other tissues (referred to as tissue-inconsistent, and set BiU_{TIC} in **Figure 2.5C**). This indicates that the enrichment in DGs is not sufficient to justify the emergence of the pathology. Thus, we speculated that the BioInt units that are really decisive to trigger pathomechanisms must hold distinctive properties in the network topology. To explore this hypothesis, the 25 TS libraries enriched in at least one tissue-consistent pathology were grouped in 11 major organ groups (**Supplementary Data 2.5**) to then compare the topological properties of: BioInt units enriched in tissue-consistent diseases (named BiU_{TC}); BioInt units enriched in tissue-inconsistent DGs (BiU_{TIC}); and BioInt units not enriched in any DGs (BiU_X, **Figure 2.5D**).

The network analysis at disease-specific level revealed the same trends observed for the systematic DG mapping (**Figure 2.4**). BioInt units accumulating DGs of tissue-consistent diseases (set BiU_{TC} in red in **Figure 2.5C,D**) tend to be expressed in fewer tissues and incorporate less UB proteins than BioInt units not affected by any disease or enriched in tissue-inconsistent DGs (sets BiU_{TIC} and BiU_X in **Figures 2.5C,D)**. Furthermore, the comparative analysis revealed that the BioInt units in set BiU_{TC} frequently included proteins with higher degree and betweenness coefficients in the global TS PPI networks. Most remarkably, BioInt units enriched in DGs of tissue-consistent diseases displayed a larger overlap with additional functional units.

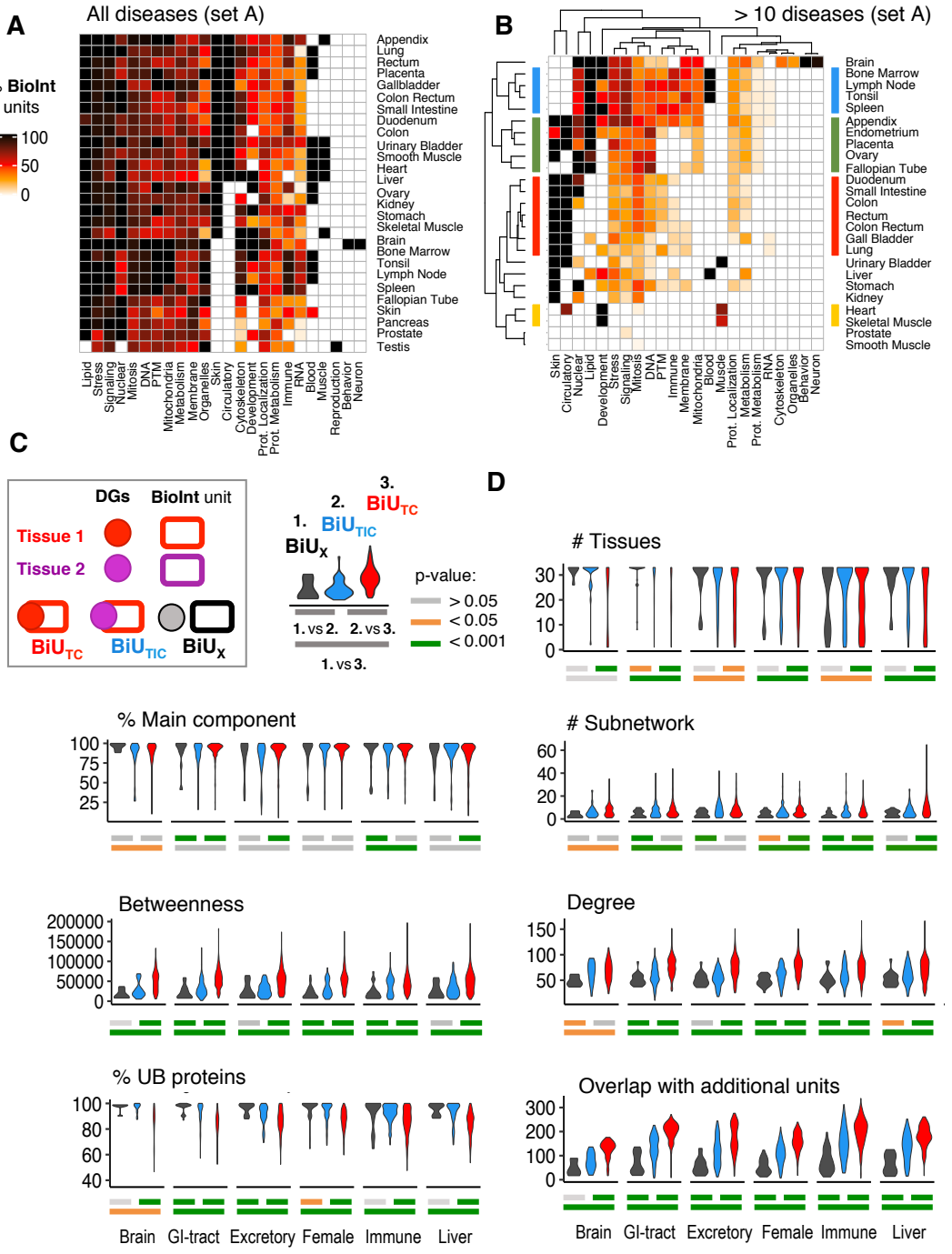


Figure 2.5 Comparison of topological properties of BioInt units accumulating tissue-consistent DGs.

Heatmaps representing the % of Bioint units by functional class enriched in at least one tissue-consistent disease (A) or in at least 10 tissue-consistent diseases (B). Dendograms and clusters result from a complete-linkage clustering using Euclidean distance. (C) Schematic picture illustrating the types of BioInt unit evaluated in panel D. For each TS BioInt library we can distinguish: BioInt units enriched in DGs of tissue-consistent diseases (BiUDT); BioInt units enriched in DGs of diseases not expected at the tissue (BiUD); and

BioInt units not significantly enriched by any DG set (BiUX). (D) Violin plots comparing the network properties of these three types of BioInt units defined above. (Bottom bars indicate the statistical significance, Wilcoxon Rank Sum test p-value; grey >0.05, orange <0.05 and green <0.001).

In fact, the DGs assigned to tissue-consistent diseases are 1.5 times more frequently located at the intersection between BioInt units than proteins encoded by nonDG (Wilcoxon test, p-value $<10^{-4}$). However, an unexpected observation is that the percentage of proteins in main component is similar but BioInt units in set $\mathbf{BiU_{TC}}$ exhibit a larger number of disconnected subnetworks according to our current map of protein interactions.

2.4.11 A case study: Mapping of differentially expressed genes to BioInt units predicts most vulnerable tissues and functions in pulmonary fibrosis and psoriasis

The dissection of the molecular mechanisms underlying complex diseases is still an open challenge. One of the most widely used strategies to investigate pathological conditions is the identification of differential expressed genes (DEg) in RNA-Seq profiles from patient-derived samples. However, the most popular algorithms for DEg analysis assess the expression of each gene independently, thus DEg datasets frequently include a large number of transcripts/proteins disconnected from the PPI network. Likewise, gene expression is highly dynamic and so DEg datasets characterizing the same disease often give different profiles. All this makes the DEg data difficult to integrate and interpret. The integration of DGE profiles with functional enrichment analysis in protein interaction networks has been recently proposed to assist in the prioritization of disease-relevant targets *(Nadeau et al., 2021)*. In a similar argumentative line and to test the analytical procedure presented in this work, we next illustrate how the mapping of disease-related DEg profiles into BioInt libraries can improve the prioritization of potential functional targets.

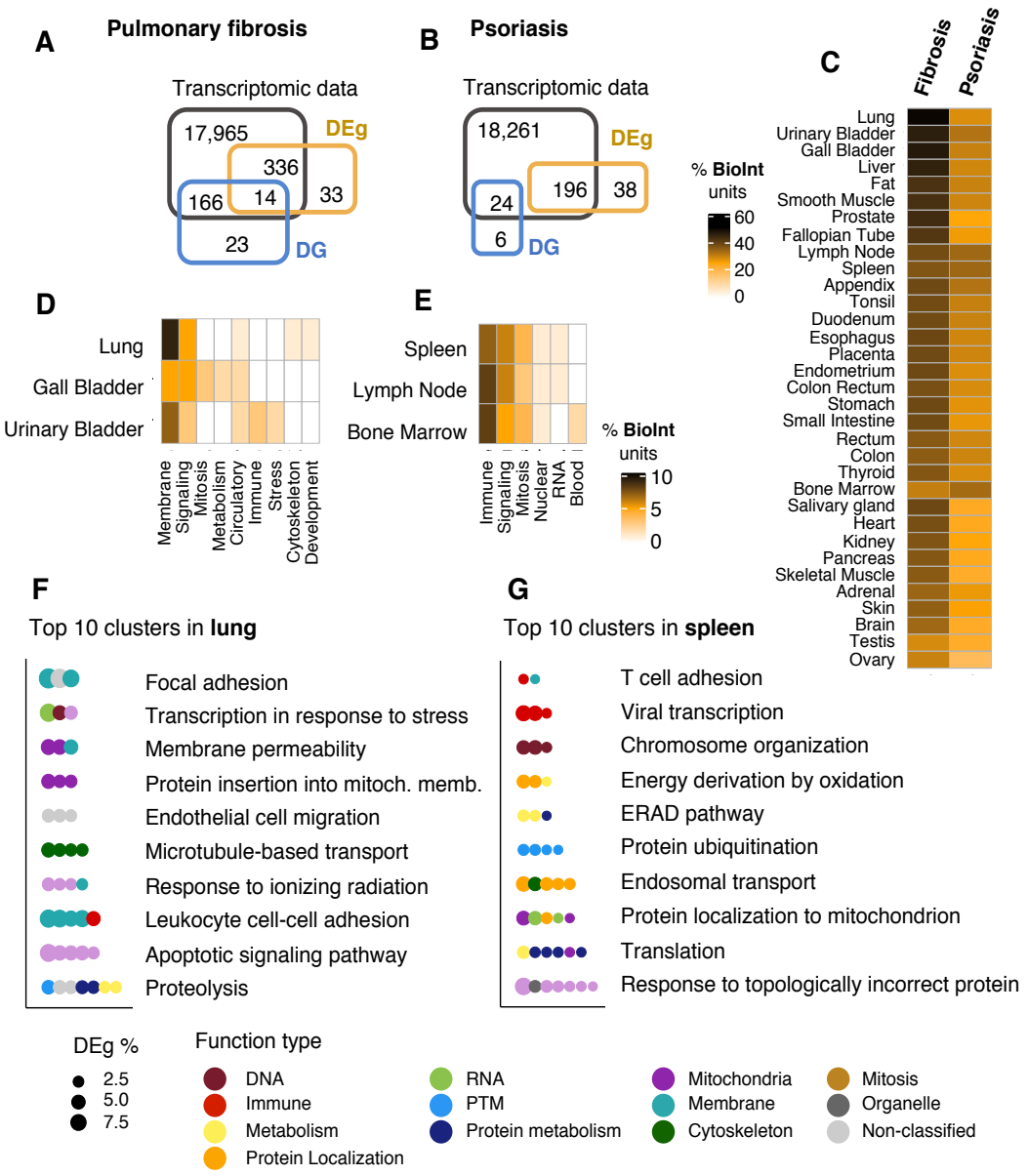


Figure 2.6 Mapping of pulmonary fibrosis and psoriasis RNA-Seq gene expression profiles into TS BioInt libraries

(A, B) Venn diagrams summarizing the overlap between unified transcriptome, DGs collected from DisGeNET and differentially expressed genes (DEg) identified in RNA-Seq profiles of patients suffering from pulmonary fibrosis and psoriasis, respectively. (C) Heatmap representing the percentage of BioInt units enriched in DEg across TS libraries. (D, E) Heatmaps representing the types of functions accumulating most BioInt units enriched in DEg in the top 3 TS libraries in panel C. (F, G) Dot plots summarizing the BioInt units enriched in DEg identified in fibrosis and psoriasis profiles in lung and spleen tissues, respectively. Each dot represents a single BioInt unit and functions with > 0.6 Wang semantic similarity are grouped in clusters (Y-axis). The figure only includes top 10 largest clusters. The text describes the parent GO-BP in common to the BioInt units grouped together. Dot color indicates the type of function (bottom legend) and size the % of DEg.

We selected and analyzed two independent transcriptomic profiles characterizing gene expression changes in patients suffering from psoriasis and idiopathic pulmonary fibrosis *(Meltzer et al., 2011; Qiu et al., 2021).* Within these datasets, 91% of fibrosis-related and 83.5% of psoriasis-related DEg were mapped in our unified transcriptome dataset, respectively (**Figures 2.6A,B**).

We next collected DGs associated with psoriasis and fibrosis in DisGeNET. Despite of the large number of DGs already associated to fibrosis (203), only 6.9% were found to be DEg in the transcriptome profile (**Figure 2.6A**). On the other hand, we only identified 30 DGs associated to psoriasis and none was DEg (**Figure 2.6B**). Similar to our previous analysis, we calculated the overrepresentation of DEg in each BioInt unit across all tissues (hyper-geometric test, p-value <0.05) and selected the 25% most affected units. Interestingly, we found that the tissues including more functional units significantly enriched in DEg were precisely those in which the symptoms are commonly observed (**Figure 2.6C**). Furthermore, the functional types accumulating highest percentage of BioInt units enriched in DEg were also related to functions suspected to be critical in the diseases (**Figures 2.6D,E**). Finally, **Figures 2.6F,**G summarize the functional signatures associated to the BioInt units enriched in DEg from fibrosis and psoriasis profiles in lung and spleen tissues, respectively. To simplify the analysis, we collapsed the BioInt units (dots) presenting a Wang's Semantic similarity coefficient > 0.6 into functional clusters (top 10 largest clusters are arranged in Y axes). In particular, the BioInt units most targeted by fibrosis-related DEg in lung included membrane permeability, proteolysis and apoptosis signaling-related functions *(Sharma et al., 2021).* In the case of psoriasis, stress-related protein folding and degradation-regulatory pathways were consistently altered in immune-related organs *(Wang and Jin, 2019).* The analysis confirmed that DE genes preferentially accumulate in biological processes already involved with fibrosis and psoriasis. Therefore, our analysis illustrates how BioInt units can provide additional insight into why these functions are more vulnerable and also suggest new DG candidates for further evaluation.

2.5 Discussion

The topological characterization of TS-networks is crucial to dissect the mechanisms underlying tissue functional diversity and identify potential vulnerabilities, namely those related to genetic disorders. However, to our best knowledge, most investigations have focused on characterizing the topology of individual proteins and DGs without considering their functional context (recently reviewed by *(Hekselman and Yeger-Lotem, 2020; C. Liu et al., 2020; Yeger-Lotem and Sharan, 2015).* However, it should be noted that PPI networks are static representations of all the physically possible interactions, and these may not be always biologically meaningful. We advocate that the integration of proteins within their functional context can improve the assessment of network properties relevant for cell physiology. On this basis, we designed a network-based strategy to characterize functionally collaborating TS PPI consortia. We applied this framework on 33 human TS networks and conducted a systematic study of the topology patterns associated to distinct normal and pathological states. This analysis revealed how the topological properties of functional units may elucidate the mechanisms of TS functional diversity and deregulation (hypothesis illustration in **Figure 2.7**).

As the very name implies, housekeeping (HK) functions are essential for the survival of any type of cell and are mostly exerted by ubiquitous (UB) proteins expressed in all tissues. Evolutionary selection has favored proteins involved in these functions and so UB proteins dominate TS network composition, accumulate more PPIs and locate at central positions in TS networks *(Barshir et al., 2014; Bossi and Lehner, 2009; Dezső et al., 2008; Lin et al., 2009)*. Beyond the characterization of individual proteins, the systematic analysis of TS BioInt libraries further supported an in-depth comparison between HK and TE functions. We corroborated that HK units are related to core functions such as organelle trafficking, RNA or protein metabolism and are mostly made up of UB proteins with significantly larger degree and betweenness coefficients than proteins exclusively involved in TE functions. Most HK units included a small percentage of nonUB proteins that varied across the TS networks. In parallel, TE units also incorporated a large percentage of UB proteins (**Figure 2.7A**). While the extensive re-use of UB proteins in TS functions is well

described *(Bossi and Lehner, 2009; Chapple et al., 2015; Podder et al., 2009),* the role of nonUB proteins in HK functions is less studied. Our analysis corroborated that UB–UB PPIs are frequently located within functional units highlighting their fundamental roles at the core of the biological processes (**Figure 2.7C**). Conversely, we found that heterotypic nonUB–UB interactions preferentially connect functional units with other proteins outside the network. Our observations are in line with a recent investigation showing that cell-specific interactions link protein complexes in the TS interactome *(Huttlin et al., 2021)* and underscore that nonUB proteins are critical players in the coordination of both HK and TE functions.

It is reasonable to assume that the characterization of mechanisms underlying tissue functional diversity will bring insights into the events triggering TS diseases. The pioneer studies characterizing the DGs topology suggested that deleterious proteins tend to display TS expression *(Goh et al., 2007; Lage et al., 2008).* Currently though, we find innumerous instances of UB proteins involved in diseases with tissue-restricted phenotypes. This indicates that TS protein expression is not sufficient to explain the emergence of TS diseases *(Hekselman and Yeger-Lotem, 2020).* Barshir and colleagues found that DGs tend to display tissue-exclusive PPIs in the tissue where the disease is manifested *(Barshir et al., 2014).* Lee and colleagues reached a similar conclusion when exploring the topology of neuron-related TS networks and hypothesized this might be key to understand the high prevalence of neurological diseases *(Lee et al., 2020).* The systematic mapping of DGs onto BioInt units corroborated that the transcript products of DGs tend to be more widely expressed than those coded by nonDGs (**Figure 2.7E**). Interestingly though, DGs tended to accumulate in functional units annotated in fewer tissues and DG–DG interactions and were more frequently located at the interface of TE and HK functions or connecting TE functions to other proteins outside in the network (**Figure 2.7G**). These results suggest that the impairment of TE functional coordination might be a key feature to spread TS homeostasis deregulation and overcome the threshold to trigger TS pathophenotypes (**Figure 2.7D,H).** A more thorough analysis of TS diseases revealed that DGs accumulate more frequently in functional units found in the disease-target tissues. This observation was particularly apparent in TS functions related to muscle, immune function or neuron physiology.

Nonetheless, many DGs accumulated in BioInt units in tissues other than the expected, indicating that DG enrichment is not the only event accounting for disease manifestation. In our view, this observation illustrates why the functional characterization of DGs might fall short to understand pathological mechanisms. Proteins are multifunctional and collaborate both in HK and TE functions. In turn, proteins establish dynamic PPIs and, as suggested in this work, acquire varying relevance depending on their TS interactome context (**Figure 2.7D**). In particular, our topological analysis reiterated that the functional units accumulating tissue-consistent DGs were actually TE BioInt units with significantly larger overlap with additional units (**Figure 2.7E**). Although multifunctional proteins have been previously associated to pathological events, our analysis brings further evidence towards this from a TS functional perspective.

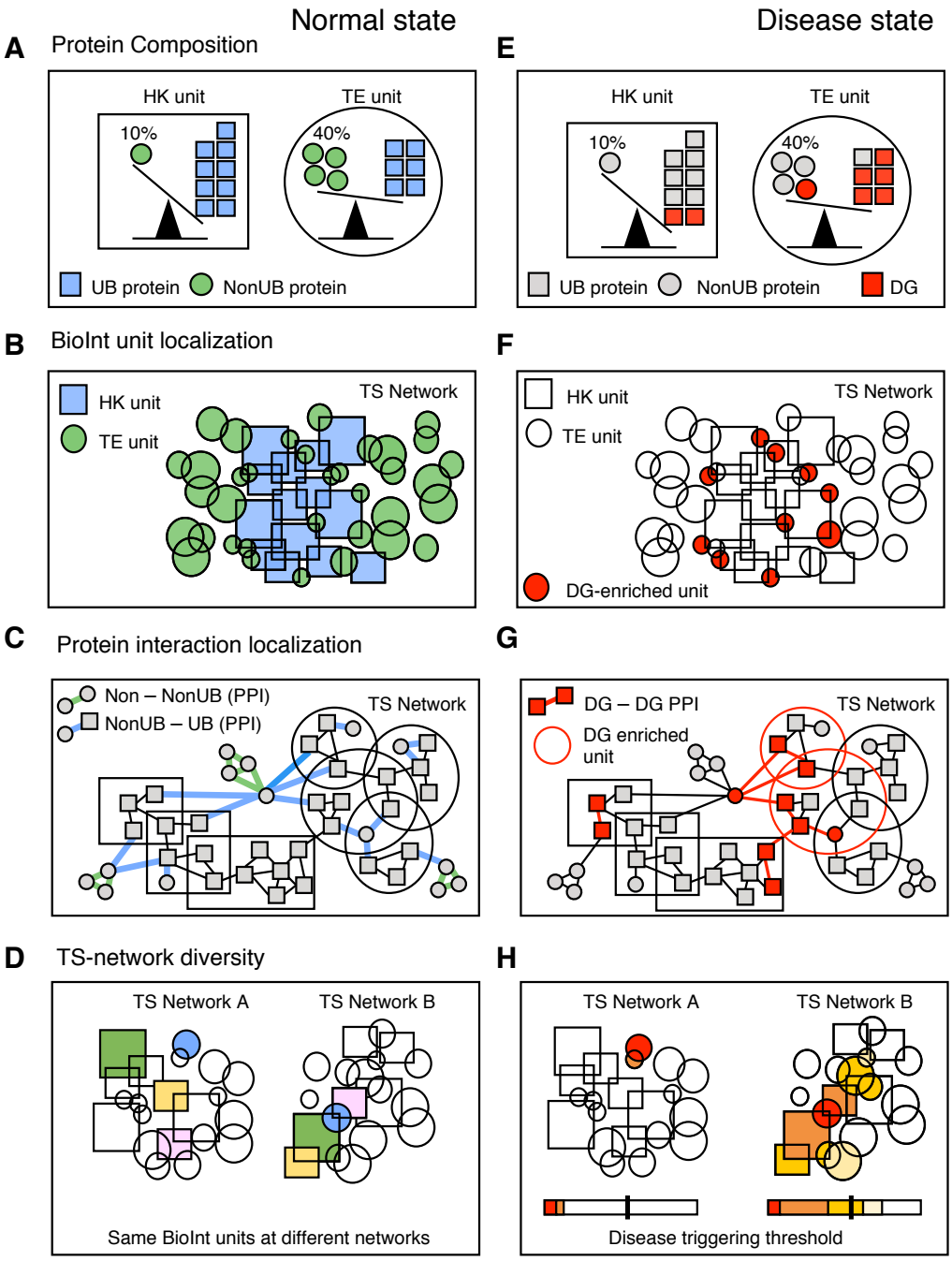


Figure 2.7 Mechanisms underlying functional diversity and tissue vulnerability linked to TS protein networks.

The main conclusions retrieved form this work can be applied in normal (left column) and disease (right column) conditions. First column: (A) HK units are mostly made of ubiquitous (UB) proteins while tissue-enriched (TE) units incorporate a balanced percentage of UB and nonUB proteins. (B) Housekeeping (HK) units include proteins with significantly larger degree and betweenness coefficients when comparing to proteins in TE functions. (C) Homotypic UB protein-interactions (PPIs) are frequently located within functional

units, homotypic nonUB PPIs lay outside functional units and heterotypic interactions preferentially connect the functional units with other units or proteins outside in the tissue-specific (TS) network. (D) Proteins can establish different interactions according to their interactomic neighborhood and so the TS networks can show distinct topological rearrangements. Second column: (E) Disease genes (DGs) are more widely expressed than nonDGs but significantly accumulate in TE units. (F) TE units amassing most DGs also incorporate proteins with larger degree and overlap. (G) DG–DG interactions are frequently located within BioInt units. Particularly in TE units, DG–DG interactions do frequently connect TE units with additional proteins or units outside in the network. These observations suggest that the distinctive disease impact observed for TE units might be triggered by DGs with additional roles in functional coordination. (H) Overall, the TS connectivity patterns might be key to understand why the impact of UB-DGs could distinctively trigger the degeneration of particular tissues.

In parallel, we made the unexpected observation that BioInt units accumulating most DGs are more sparsely connected, while tending to incorporate significantly more central proteins in the TS network. Dynamic interactions are known to play critical roles in the regulation and coordination of protein function. However, high-throughput PPI detection techniques preferentially detect stable interactions and thus, are more likely to dismiss transient PPIs. Although caution must be taken until the advent of more sensitive technologies, this observation suggests that the most vulnerable functions tend to include numerous transient interactions not yet identified. Our conjecture is aligned with previous results indicating that biological and disease modules do not necessarily coincide with topological clusters *(Agrawal et al., 2018; Ghiassian et al., 2015; Wang and Zhang, 2007).* If confirmed, this observation would question the pivotal role of clustering algorithms in the design of network-based methods for biomedical research.

To illustrate the benefits of the BioInt framework in a real case problem, we took advantage of two public transcriptome profiles from patients suffering from psoriasis and pulmonary fibrosis. The scarcity of already known psoriasis-causal genes together with the low overlap between fibrosis DGs and the corresponding DE transcriptome reflects the need for additional research bridging the molecular and pathophenotypic observations. The analysis presented here demonstrates the ability of our method to independently identify the most afflicted tissues and functions and thus bring novel insights to refine DG prioritization methods.

The BioInt-U framework sets the stage for novel approaches to explore the functional relevance of TS topological properties. Nonetheless, it also has limitations.

The identification of BioInt units relies on PPI and Gene Ontology datasets, which are known to be over fitted by proteins of significant research interest. Until a more comprehensive characterization of the interactome and functionome, our investigation is likely to underestimate poorly characterized players. On the other hand, the analysis exploits static networks and ignores cell-specific temporal information of the particular tissue. The integration of dynamic and quantitative expression data could surely benefit network-centered investigations. Notwithstanding, is worth recalling that the use of quantitative data would also increase the analytical complexity. To compensate the lack of spatiotemporal data, we enabled functional units to overlap. In this way, we could evaluate all the possible combinations of functional consortia.

Overall, the work presented here showcases the relevance of evaluating network topology from the functional perspective. The large-scale topological vulnerabilities inferred from our analysis could contribute to the refinement of network-based methods for DG candidate prioritization. Likewise, the evaluation of the topological context of DGs across tissues could facilitate the identification of the most critical drug targets while avoiding unpredicted off-targets.

2.6 Supplementary Data

R code and input data to reproduce BioInt-U method is available in
https://github.com/GamaPintoLab/BioInt-U

Supplementary data files are available in
https://github.com/GamaPintoLab/MLG_PhDThesis_SupData

Supplementary Data S2.1 Descriptive summary of relevant parameters at each step of the BioInt-U framework. Organ samples are summarized in broader tissue groups. The transcriptome datasets were mapped into APID protein-protein interaction (PPI) dataset to generate TS-networks. Next, the enriched Gene Ontology Biological Process (GO-BP) terms were used to define the Bioint units. Only BioInt units with a size of 10 to 200 were selected. Transcript retention in selected BioInt units was calculated when compared to the total transcripts in TS networks. Median, 1st and 3rd Quartile sizes were calculated for each TS Bioint library. Finally, the coefficient of variation across tissues was evaluated for all the parameters in the table.

Supplementary Data S2.2 Summary of key words used for the functional classification of BioInt units. The Bioint units were classified in 24 functional categories (first column) by performing a direct text mining of key words found in the description of Bioint units (second column).

Supplementary Data S2.3 Complete Simpson's similarity analysis of BioInt and CORUM complexes. Table including all the pair-wise Simpson's similarity indexes between BioInt and CORUM complexes

Supplementary Data S2.4 Complete description of functional and topological parameters of BioInt units. Table including all the topological parameters employed throughout this study. The combination of all BioInt units identified along the 33 tissues (second sheet) returned a unified Bioint library consisting on 728 unique BioInt Units (first sheet).

Supplementary Data S2.5 Complete list of disease gene association with tissue-consistent pathologies. First sheet includes a summary of the selected 463 diseases related to 11 tissue-specific phenotypes. Second sheet details the list of DGs annotated for the total 1948 diseases described in DisGeNET and, when available, the their tissue-consistent classification.

3 Searching the overlap between network modules with specific betweenness (S2B) and its application to cross-disease analysis

Data presented in this chapter was included in the following work:

García-Vaquero ML., Gama-Carvalho M., De Las Rivas J. & Pinto FR., Searching the overlap between network modules with specific betweeness (S2B) and its application to cross-disease analysis. *Sci Rep* (2018).

Author contributions:

MG-V conceptualized the study, developed the method, performed the analysis and wrote the manuscript. **FRP** conceptualized the study, developed the method, performed the analysis and wrote the manuscript. **MG-C** supervised the study and reviewed the manuscript. **JDLR** supervised the study and reviewed the manuscript.

3.1 Abstract

Discovering disease-associated genes (DG) is strategic for understanding pathological mechanisms. DGs form modules in protein interaction networks and diseases with common phenotypes share more DGs or have more closely interacting DGs. This prompted the development of Specific Betweenness (S2B) to find genes associated with two related diseases. S2B prioritizes genes frequently and specifically present in shortest paths linking two disease modules. Top S2B scores identified genes in the overlap of artificial network modules more than 80% of the times, even with incomplete or noisy knowledge. Applied to Amyotrophic Lateral Sclerosis and Spinal Muscular Atrophy, S2B candidates were enriched in biological processes previously associated with motor neuron degeneration. Some S2B candidates closely interacted in network cliques, suggesting common molecular mechanisms for the two diseases. S2B is a valuable tool for DG prediction, bringing new insights into pathological mechanisms. More generally, S2B can be applied to infer the ovelap between other types of network modules, such as functional modules or context-specific subnetworks. An R package implementing S2B is publicly available at https://github.com/frpinto/S2B.

3.2 Introduction

Disruption of a gene sequence may cause the dysfunction of the encoded protein, which can trigger the onset of a disease. Such genes are defined as disease causal genes. Nevertheless, a disease is a pathologic phenotype resulting from synergic disruptions of varied cellular functions caused by both genetic and environmental factors *(Naylor and Chen, 2010).* Consequently, disease associated genes (hereinafter called Disease Genes (DGs)) are not necessarily causal. They can be modifiers, that modulate disease severity, or phenotypical, unable to influence the disease course but responsible for disease phenotypes. Genes associated with a disease are more prone to interact with each other than with non-disease related genes, establishing network disease modules *(Oti et al., 2006; del Sol et al., 2010).* Disease modules are neighborhoods of the full interactome network containing all disease associated proteins *(Ghiassian et al., 2015).* As interactomic maps are still incomplete *(Menche et al., 2015b)* and the number of known DGs is limited *(Brunner and van Driel, 2004),* the identification of DGs remains an important issue, contributing to decipher molecular mechanisms of disease and to discover biomarkers and therapeutic options.

Efforts to complete protein interactions networks include not only high throughput experimental approaches *(Rolland et al., 2014),* but also computational predictive methods, recently reviewed by Kotlyar et al. The latter can be based in sequence features, conservation across species, protein domains, 3D structure, interaction network topology, or a combination of several of the previous data types *(Kotlyar et al., 2017).* To expand the list of known DGs, information systems, like DisGeNet *(Piñero et al., 2015),* Open Targets *(Koscielny et al., 2017)* or DISEASES *(Pletscher-Frankild et al., 2015),* integrate and weight heterogeneous evidence sources linking genes with diseases, including text-mining approaches.

Network-based DG prioritization methods aim to recover complete disease modules, using network interactions of known DGs to predict new DG candidates. One such method, DIAMOnD *(Ghiassian et al., 2015),* starts from the set of known DGs and iteratively adds one node to the disease module. The added node is the

more statistically enriched in DGs among its direct neighbors. Other DG prioritization algorithms are based on random walks *(Köhler et al., 2008; Vanunu et al., 2010)* or diffusion algorithms *(Valentini et al., 2014).*

Diseases sharing phenotypes exhibit alterations in similar functional pathways, and their disease modules are more likely to overlap *(Goh et al., 2007; Menche et al., 2015b).* Based on this similarity, researchers have identified common functions among the network neighbors of genes associated with Alzheimer's and Parkinson's diseases *(Calderone et al., 2016),* and looked for common neighbors of proteins associated with autism spectrum disorders *(Sakai et al., 2011).*

However, to our knowledge, there is currently no network-based algorithm aiming to directly predict genes simultaneously associated with two diseases. These can provide hypotheses to explain molecular mechanisms of pathophenotypes shared between two diseases. In addition, these candidates can suggest new therapeutic targets, or provide grounds to repurpose current therapies from one disease to the other. With this aim, we propose a network-based approach called S2B (double specific-betweenness). S2B relies on the assumption that interactors more commonly found on shortest paths linking proteins encoded by genes associated to two diseases must appear in the disease modules overlap. To identify and rank these proteins, S2B employs a specific version of betweenness centrality, which measures how many times a node is involved in a shortest path, focusing specifically on shortest paths linking proteins associated with the two diseases.

A similar network approach has been recently proposed to identify the mediator pathways between DGs and genes differentially expressed between healthy and disease samples *(Park et al., 2017).* Parallel application of this method to related diseases identified common mediator pathways. S2B approaches this problem from a different perspective, as it aims to identify individual proteins that are directly involved in the mechanisms of both diseases simultaneously.

We applied S2B to Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA), two fatal Motor Neuron degenerative Diseases (MND). The most

common form of SMA is caused by recessive mutations in the SMN1 gene, encoding the SMN protein. Numerous causal genes have been reported for ALS, involved in multiple functions such as oxidative stress control (SOD1) *(Chen et al., 2013),* vesicle trafficking (ALS2, FIG4, OPTN, VABP, CHMP2B) or proteasomal functions (UBQLN2, VCP) *(Menzies et al., 2015).* However, RNA metabolism is the function with the largest subset of MND causal genes (TARDBP, FUS, SETX, ATXN2, HNRNPA1, HNRNPA2/B1, ELP3 in ALS, and SMN1 in SMA) *(Carrì et al., 2015; Siddique and Siddique, 2008).* While under debate, protein aggregation and RNA metabolism deregulation are the most accepted hypotheses to explain the MND phenotypes. However, it is very intriguing how such critical events could distinctively affect Motor Neuron (MN) physiology

Although ALS and SMA present distinct clinical features, they show great phenotypic and molecular similarities, implying a common etiology. Indeed, recent work from our group revealed that key MND causal genes SMN, FUS, TDP43 and SETX show tight physical and functional relationship *(Gama-Carvalho et al., 2017).* In the same vein, this paper shows that S2B predicts cross-disease genes (cDGs), providing new insights into the molecular mechanisms of MND.

3.3 Methods

We considered the prediction of cDGs analogous to the problem of finding the overlap between two network modules when information about module composition is incomplete: consider an undirected graph *G* with two overlapping connected subgraphs *A* and *B*. However, we only know subsets *a* and *b* (seeds) that compose *A* and *B*, respectively. With this incomplete information, we cannot define the set of nodes in the overlap between *A* and *B*. We developed a method that knowing the sets of seeds *a* and *b*, predicts which nodes of *G* are more likely part of *A* and *B* simultaneously. This method is based in the computation of the Double Specific Betweenness score (*S2B*) presented in **equation (3.1).**

$$S2B(k,G,a,b)=\frac{\sum_{i}^{i\in a,i\neq k}\sum_{j}^{j\in b,j\neq k} sp(k,i,j,G)\bullet t(i,j,G)}{\sum_{i}^{i\in a}\sum_{j}^{j\in b} t(i,j,G)} \quad \textbf{(3.1)}$$

Equation (3.1) computes auxiliary functions sp(k,i,j,G) **(equation (3.2))** and t(i,j,G) **(equation (3.3)).**

$$sp(k,i,j,G)=\begin{cases} 1 & \text{if} \quad d(i,j,G)=d(i,k,G)+d(k,j,G) \\ 0 & \text{if} \quad d(i,j,G)\neq d(i,k,G)+d(k,j,G) \end{cases} \quad \textbf{(3.2)}$$

$$t(i,j,G)=\begin{cases} 1 & \text{if} \quad d(i,j,G)\leq avgd(G) \\ 0 & \text{if} \quad d(i,j,G)>avgd(G) \end{cases} \quad \textbf{(3.3)}$$

In both **equations (3.2) and (3.3),** *d(i,j,G)* is the length of the shortest path between the i^{th} and the j^{th} nodes of *G*. *sp(k,i,j,G)* is an indicator function with value 1 if node *k* is part of a shortest path between nodes *i* and *j*. *t(i,j,G)* is an indicator function with value 1 if the length of the shortest path between nodes *i* and *j* is equal or lower than the average shortest path length of *G (avgd(G))*. This path length filter is important to avoid the influence of nodes that are loosely related with the other

module. Altogether, it means that *S2B(k,G,a,b)* is the fraction of shortest paths linking a node in *a* to a node in *b* that contain node *k*, with length smaller than the average path length of *G*. Before applying **equation (3.1),** nodes present in *a* and *b* simultaneously are discarded as these, by definition, belong to the overlap between *A* and *B*. Therefore, shortest paths starting from these nodes diverge from the overlap, increasing the chances of crossing with other shortest paths outside the overlap region.

We observed that only a small number of nodes in the network achieved high *S2B*. If we plot *S2B* against *1-quantile(S2B)*, we typically observe an L-shaped curve. To define the threshold value that separates high *S2B* from low *S2B* we apply **equation (3.4).** This equation finds the *S2B* that minimizes the distance to the origin in the referred L-shaped curve.

$$S2B_t = \underset{S2B(k,G,a,b)}{\arg\min}\left(\left(\frac{S2B(k,G,a,b)}{\max_k\left(S2B(k,G,a,b)\right)}\right)^2 + \left(1-quantile\left(S2B(k,G,a,b)\right)\right)^2\right) \quad (3.4)$$

Besides considering only nodes with high *S2B*, we also implemented two specificity scores **(equations (3.5) and (3.6)).**

$$SS_1 = P\left(S2B(k,G,a,b) \geq S2B(k,G,a_R,b_R)\right) \quad (3.5)$$

$$SS_2 = P\left(S2B(k,G,a,b) \geq S2B(k,G_R,a,b)\right) \quad (3.6)$$

SS_1 is the probability that the S2B of node k with seeds a and b is equal or higher than the same score computed with random seed sets a_R and b_R. A high SS_1 means that the S2B is specific for the initial seed sets. SS_2 is the probability that the S2B of node k in graph G is equal or higher than the same score computed with a random graph G_R, were nodes maintain their degree but edges are randomly shuffled. A high SS_2 means that the S2B is specific for the connectivity patterns in G and is not a consequence of the high centrality of k. To compute each specificity score, 200 random seed sets, or randomized networks were employed. Each

randomization contributes to the score of all nodes simultaneously. The computation of S2B and specificity scores took around 22 minutes in a 2.8 GHz Intel Core i7 processor and 8 GB of RAM when using a network with 12424 nodes, 90333 edges, 197 ALS and 48 SMA DGs. A description of the use of S2B method to prioritize cDGs is presented in the **Supplementary text.**

3.4 Supplementary methods

Retrieval of Disease Genes We retrieved all the ALS and SMA disease genes (DGs) described on OMIM (Online Mendelian Inheritance in Man; (www.nlm.nih.gov/mesh/MBrowser.html) ***(Hamosh et al., 2002)*** and DisGeNET; (http://www.disgenet.org) ***(Piñero et al., 2015)*** databases in September 2016 **(Supplementary Data S3.1).** In both cases, we took all the available associations without any quantitative filtering. We also merged associations for available subtypes of each disease.

PPI data collection and network construction Human physical Protein-Protein Interaction (PPI) data was extracted from HuRI (Human Reference Protein Interactome Mapping Project) (interactome.baderlab.org) ***(Rolland et al., 2014; Rual et al., 2005; Venkatesan et al., 2009; Yang et al., 2016; Yu et al., 2011)*** and APID (Agile Protein Interaction Dataserver) (apid.dep.usal.es/) ***(Alonso-Lopez et al., 2016)*** databases (accessed in February 2017). We constructed undirected and unweighted networks using igraph R-package ***(Csárdi and Nepusz, 2006).*** Loop and multiple edges were eliminated and only the main component of the network was selected. Finally, ALS and SMA DGs were labeled as seed nodes.

Artificial disease modules Three different types of modules were used, based on distinct hypothesis for the spread of disease-causing perturbations across cellular networks. Shell modules ***(Ghiassian et al., 2015)*** are composed by a seed node and all other nodes in the network at distance of 2 or lower. These artificial modules assume that the perturbation spreads homogeneously through the network. Connectivity modules ***(Ghiassian et al., 2015)*** are built iteratively around a seed node,

adding at each step the node most significantly enriched in links to previous module members. These modules assume that disease perturbations affect predominantly nodes that are specifically linked to causal genes. Random walk with restart (rwr) ***(Köhler et al., 2008)*** modules simulate the path of an imaginary walker that, at each time step, moves to a randomly chosen direct neighbor or, with a given restart probability, returns to the seed node. The nodes with higher probability of being visited by the walker constitute the model. These modules assume that disease perturbations spread more easily to nodes with multiple and shorter paths linking to the causal nodes. Real disease modules can be a mixture of these and other module types, as the disease perturbation pattern along the network may depend on the type of molecular function of each protein and the nature of each protein-protein interaction.

Artificial disease modules were constructed using the APID3 protein interaction network. Proteins with a degree between 19 and 22 were selected as possible causal seeds for the artificial modules. Each seed originated three artificial modules with different topological properties. Shell modules were composed by the seed and proteins at distance 1 or 2 in the network. Only shell modules with more than 200 and less than 400 proteins were kept. Connectivity modules were composed by the seed and 249 proteins added iteratively. In each step, all proteins out of the growing module were tested for having a higher than expected number of links to proteins in the module using a hypergeometric test. The protein with the smallest p-value was added to the module. Random walk with restart modules were composed by the 250 proteins with higher occupancy probability in the random walk stationary distribution initiated in the seed node with a restart probability of 0.75. The stationary distribution was determined numerically as described in ***(Köhler et al., 2008).*** Within each topology type, existence of overlap between all possible module pairs was evaluated. Only module pairs where the overlap contained between 50 and 125 proteins were used to test S2B performance.

Functional enrichment comparison We performed a comparison of Functional Enrichment Analyses (FEAs) of MND-Disease genes (MND-DGs) set and S2B candidate genes. The initial gene sets entailed 370 MND-DGs (295 ALS and 93 SMA genes, being 18 common to both diseases) and 232 S2B candidate genes.

Functional enrichment of Disease Genes gathers only the GO terms that were associated to at least one ALS and SMA gene simultaneously. Both sets were functionally enriched for Gene Ontology Terms (GO) Biological Process (BP) using EnrichGO R-package ***(Yu et al., 2012).*** Results were filtered using an adjusted p-value ≤ 0.05 and human genome background frequency ≤ 0.10. To better reflect the statistical significance of the results; we calculated the fold enrichment of each GO term (ratio between the frequency of the GO in the gene list and the frequency of the same GO in the human genome background gene list). The FEAs returned 461 and 718 enriched GO terms for MND-DGs and candidates respectively. Due to the Gene Ontology (GO) hierarchical structure, when a GO term is enriched it is likely that some of its ancestors are also enriched, increasing the results size and redundancy. In order to facilitate the analysis of the results, we applied a simplification workflow (Supplementary Figure S3.1B). We firstly created GO groups of GO terms showing gene co-occurrence (at least an overlap of 70% of associated genes) and semantic similarity (GO terms that presented a Lin's semantic similarity score ≥ 0.70) (**Supplementary Figure S3.1B1).** This approach finds "hidden" commonalities between apparently different GO terms, thus it was applied jointly for both FEAs. When a GO group was formed, it retained the 3rd quartile of fold enrichment and the sum of gene frequencies of the merged GO terms for candidate and Disease gene sets respectively (**Supplementary Figure S3.1B1).**

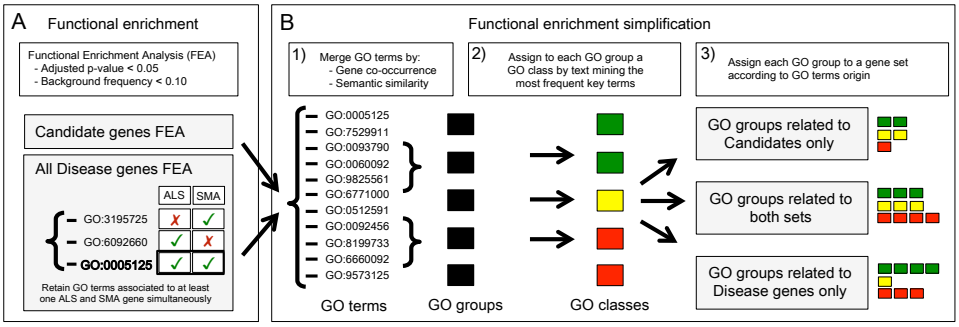


Figure S 3.1 Functional enrichment analysis (FEA) simplification and comparison workflow

(A) FEA of S2B candidate and Disease Genes (MND-DGs) sets. MND-DGs FEA gathers only the GO terms that were associated to at least one ALS and SMA gene simultaneously. **(B)** Functional simplification; 1) Resulting GO terms are merged into GO groups by gene co-occurrence and semantic similarity, 2) GO groups are classified based on the most recurrent key term and 3) GO groups are assigned to a final set according to the genes associated to each GO term in the respective GO group.

Due to the heterogeneity of biological processes retrieved, we manually created 15 major functional classes (GO classes) defined each one by a set of key words described in **Table S3.1**. Then, each GO group was assigned to the GO class most represented in the contained GO terms' descriptions **(Supplementary Figure S3.1B2).** Finally, GO groups were divided in three sets according to; if they had GO terms associated only to candidate genes, to MND-DGs or to both initial gene sets **(Supplementary Figure S3.1B3).**

Table S 3.1 GO classes and corresponding key terms used to define themby text mining of GO terms

	GO class name	Key terms
1	Nervous system	neuron, synaptic, axon, microglial, glial, neural, neuromuscular, neurogenesis, nervous
2	Immune system	immune, host, pathogen, interferon-beta, cytokine, fungus, interleukin-2, interleukin-1, leukocyte
3	Muscle	Muscle
4	Stress	stress, heat, oxidative, UV, X-ray, superoxide
5	Folding	aggregation, folding
6	Apoptosis	apoptosis, apoptotic, autophagy
7	Cytoskeleton	cytoskeleton, microtubule, actin
8	RNA processing	RNA, processing, mRNA, spliceosomal, splice
9	Transcription	transcription, chromatin, histone
10	DNA repair	DNA, repair
11	Protein degradation	degradation, proteolysis, ubiquitination, deubiquitination, ERAD
12	Cell cycle	cycle, mitotic, cytokinesis
13	Protein export/import	localization, transport, import, export, targeting
14	Signaling	transduction, cascade, signaling, signal
15	Development	development, developmental, differentiation, embryo, embryonic, morphogenesis

Analysis of shortest path clusters in S2B candidate interaction network

The physical interactions between S2B candidates were retrieved form the APID3HuRI interactome, generating an S2B candidate interaction network. We generated clusters of candidates that tend to be part of the same shortest paths linking seed proteins. First, we gathered the list of shortest paths used in S2B computation and containing each candidate. For each pair of candidate proteins, we computed a jaccard coefficient evaluating the ratio of the number of shortest paths where both candidates were present together over the number of shortest paths where at least one of the candidates was present. Pairs of candidates with a jaccard coefficient greater than 0.25 were linked in a network. The clusters were expanded to include all the candidates that were present in 75% or more of the shortest paths containing the initial cluster members. Connected components with more than 3 candidates or isolated cliques with 3 members were selected to generate a candidate cluster.

3.5 Results

3.5.1 S2B performance with artificial modules

S2B was applied to random seeds from overlapping artificial modules. Then, the precision and recall in the retrieval of nodes in the overlap region was evaluated. For three different types of artificial modules (see **Supplementary Results),** the probability of belonging to the overlap between the two modules decreased for lower S2B **(Figure 3.1A). Figure 3.1A** also confirms that discarding seeds known to be part of the overlap enhances S2B ability to identify top candidates.

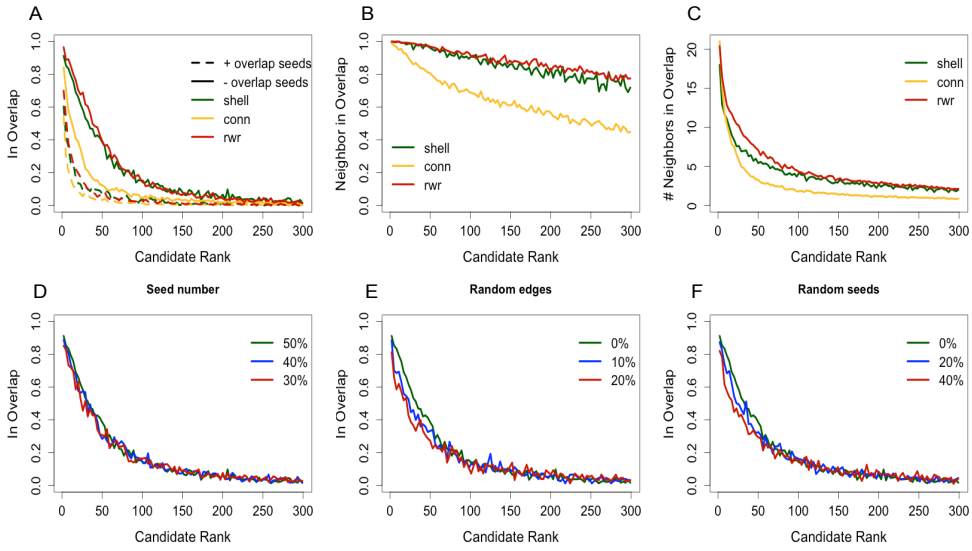


Figure 3.1 S2B performance with artificial disease modules

(A) Fraction of candidates that were in the overlap between modules as a function of S2B decreasing rank. (B) Fraction of candidates that are direct neighbors of proteins in the overlap (C) Recall as a function of S2B decreasing rank. Recall is the fraction of proteins in the overlap between the two modules that have an S2B rank lower or equal to the candidate rank ploted. In A, B and C three models of disease modules were tested: shell, connectivity (conn) and random walk with restart (rwr) based modules. The impact on method performance of excluding seeds known to be part of both modules was evaluated in A and C. Hereafter, results were computed excluding seeds known to be part of both modules. (D) S2B robustness upon reduction of the fraction of module proteins used as seeds. (E) S2B robustness upon randomly rewiring a fraction of network edges. (F) S2B robustness upon replacing a fraction of input seeds by random proteins. In plots A, B, D, E and F, values are averages of S2B candidates in three consecutive ranks. In A, B and C, 95 pairs of shell modules, 355 pairs of conn modules and 200 pairs of rwr modules were evaluated. In D, E and F, 50 pairs of shell modules were used. Shell modules have between 200 and 400 nodes, while conn and rwr modules have 250 nodes. The overlap between two modules is always between 50 and 125 nodes. In A, B, C, E and F, a 50% random sample of each module was used as seeds.

The probability of being in the overlap decays rapidly for lower S2B. However, as shown in Figure 3.1B, candidates maintain a high probability of being direct neighbors of proteins in the overlap for a wider range of S2B ranks. S2B also correlates with the expected number of direct neighbors in the overlap (**Supplementary Figure S3.6A).** Conversely, recall, that is the fraction of all the nodes in the overlap that are correctly predicted in the top ranked S2B candidates, grows almost linearly in the best 50 candidates, and then converges more slowly to its maximum plateau **(Figure 3.1C). Figure 3.1A-C** show that S2B performs better for random walk with restart (rwr) modules, followed closely by shell modules, both in terms of precision and recall. Performance in connectivity modules is weaker, although maintaining similar trends. S2B performance is similar knowing 50% or only 30% of the proteins involved in disease **(Figure 3.1D and S3.6B).** We also assessed the impact of false edges in the network **(Figure 3.1E and S3.6C)** confirming an expected decrease in performance, mainly among the 50 top-ranked candidates. But even when 20% of the network edges are randomly shuffled, prediction quality is not strongly affected. Lastly, **Figure 3.1F** and **Figure S3.6D** show that S2B performance is only slightly decreased by inclusion of up to 40% random seeds. Overall, S2B is robust to changes in module topology, incomplete disease characterization, and false positive edges and disease-gene associations.

3.5.2 Comparing S2B with single disease prioritization methods

To our knowledge, there is currently no other method to predict proteins simultaneously associated with two related diseases (cDGs). However, there are several methods to prioritize genes associated with one disease. We considered applying one of these methods to the seeds of two diseases separately as an S2B alternative. Proteins in the intersection of the two prediction sets would be candidates for simultaneous association with both diseases. We tested this hypothesis with the DIAMOnD algorithm *(Ghiassian et al., 2015).* For each module, 250 iterations were computed and the intersection between the two sets of 250 proteins was compared with the known overlap, estimating DIAMOnD precision **(Table 3.1).**

Table 3.1 Precision of DIAMOnD and S2B predictions of proteins in the overlap between pairs of artificial modules

Predictions are matched relatively to the number of candidates generated by DIAMOnD for the same pair of modules. 50 module pairs of each type were evaluated.

Module type	# Candidates retrieved by DIAMOND (equal to # top S2B candidates) median [1stQ-3rdQ]	Precision median [1stQ-3rdQ]	
		DIAMOnD	S2B
Shell	4 [1-9]	0.00 [0.00-0.18]	1.00 [0.75-1.00]
Connectivity	135 [104-149]	0.60 [0.54-0.73]	0.18 [0.16-0.22]
RWR	8 [1-26]	0.13 [0.00-0.25]	1.00 [0.88-1.00]

DIAMOnD predicts many candidates for connectivity modules with moderate precision, while for shell and rwr modules the number of candidates is generally small and precision low. A better performance of DIAMOnD with connectivity modules was expected, as these are generated with the same algorithm used by DIAMOnD to make predictions. For each pair of artificial modules tested, we selected from the top S2B candidates the same number of candidates predicted by DIAMOnD. The matched S2B precisions are higher than DIAMOnD's for shell and rwr modules, but lower for connectivity modules **(Table 3.1).** For this type, the number of DIAMOnD candidates is large and, as shown in **Figure 3.1A,** S2B precision for connectivity modules decays quickly with candidate rank. S2B predictions would have a median precision of 0.60 (similar to DIAMOnD) if the top 20 candidates were considered. In conclusion, although DIAMOnD is a good approach for connectivity type modules, S2B provides a good performance for every type of module tested.

3.5.3 Identification of common MND genes using S2B

To evaluate the potential of S2B, we focused on the Motor Neuron Diseases (MND) Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA). DGs (seeds) of ALS and SMA (available in supplementary material) were identified from OMIM (*Hamosh et al., 2002)* and DisGeNET (*Piñero et al., 2015*). Human protein interaction networks from two different origins were used. APID (Agile Protein

Interaction DataAnalyzer) *(Alonso-Lopez et al., 2016)* gathers literature reported protein interactions, while HuRI (Human Reference Protein Interactome Mapping Project) results from unbiased large scale screens for binary interactions *(Rolland et al., 2014; Rual et al., 2005; Venkatesan et al., 2009; Yang et al., 2016; Yu et al., 2011)*. Literature-based protein interaction networks are densely connected around proteins of biomedical interest, while large scale experimental techniques may fail to detect interactions between certain types of proteins, such as membrane proteins *(Brito and Andrews, 2011*). In a comparative analysis of S2B results with these networks (supplementary text, **Figure S3.3**), it was observed that the fraction of common S2B candidates grows with the level of confidence of protein interactions retrieved from the literature. A mixed APID/HuRI network also shows a high fraction of candidates in common with the separate analysis of the two networks (**Figure S3.3**). Finnally, we opted to merge HuRI with APID interactions reported in a minimum of 3 independent experiments (APID3). This maximizes global interactome and DG coverage while avoiding poor quality interactions. Analysis of 197 ALS and 48 SMA DGs (**Supplementary Data S3**) within the APID3HuRI network returned 232 candidate proteins with a S2B higher than S2Bt and both SS1 and SS2 higher than 0.90 (**Supplementary Data S3**).

3.5.4 Comparative FEA of S2B candidates and DGs

S2B candidates should be involved in processes associated with both ALS and SMA DGs (MND-DGs). To assess this hypothesis we performed a comparative Functional Enrichment Analysis (FEA) of Gene Ontology (GO) biological processes associated with S2B candidates and MND-DGs sets. For the latter, only enriched GO terms associated with both ALS and SMA DGs were considered. MND-DGs and S2B candidates were enriched in 853 and 1110 GO terms respectively. S2B terms contained 43% (392) of the MND-DGs terms. Among the 232 S2B candidates are 5 SMA seeds, 19 ALS seeds and 2 DGs associated with both ALS and SMA (not used as seeds but selected as candidates). Common GO terms could be due to the presence of these seeds among S2B candidates. To evaluate this hypothesis, we performed a randomization test, repeating the FEA with 1000 random sets of 232

proteins extracted from the interaction network, ensuring that 5 SMA DGs, 19 ALS DGs and 2 DGs associated with both ALS and SMA were selected. None of the GO terms enriched in the S2B candidate set was randomly enriched in more than 3.6% of the random sets, showing that S2B GO terms are not significantly biased. Additionally, the fraction of GO terms enriched in the random sets also associated with MND-DGs was significantly lower than the observed for the S2B candidates ($p<0.001$, randomization test).

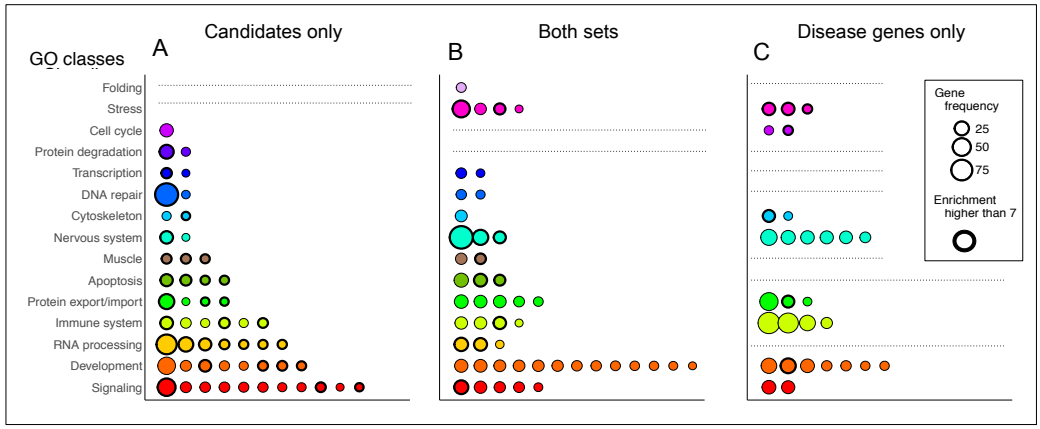


Figure 3.2 Comparison of functional enrichments between S2B candidates and Disease Genes (MND-DGs) sets

Two independent Functional Enrichment Analyses (FEAs) were performed for S2B candidates and DG sets. FEA results were simplified by merging GO terms into GO groups by gene co-occurrence (if they have 70% of associated genes in common) and semantic similarity (if they have a Lin similarity score higher than 0.70). To further simplify the results, each GO group was assigned to a single GO class by counting the key words most frequent in GO terms descriptions (supplementary text). 67 GO groups were not related to any GO class and therefore were discarded. (A) GO groups related only to S2B candidates genes. (B) GO groups related both with S2B candidates and with MND-DGs. (C) GO groups related only with MND-DGs. Each dot represent a single GO group characterized by the sum of gene frequencies (dot size). GO groups with a 3rd quartile fold enrichment higher that 7 are highlighted with bold border.

Among biological processes uniquely enriched in S2B candidates or in MND-DGs there were still similar processes. Therefore, we applied a simplification workflow **(Supplementary Methods)** minimizing redundancy by merging them as GO groups (according to overlap between gene sets and to semantic similarity). We further simplified the results by assigning GO groups to functional classes. Finally, we divided GO groups into three sets; GO groups containing only MND-DGs, S2B candidates or both **(Figure 3.2).**

Functional simplification generated 131 GO groups, 48 common to both S2B candidates and MND-DGs sets (**Figure 3.2B**), representing 62% of the MND-DGs GO groups and covering 13 out of the 15 GO classes. Removing term redundancy further increased the recovery of MND-DGs processes by S2B candidates. There are still many GO groups that belong to unique sets (**Figure 3.2A,C**), but most belong to GO classes that are represented in both S2B candidate and MND-DGs sets. The exceptions are two groups of the 'Protein Degradation' class, which are only enriched in S2B candidates. Interestingly, protein degradation is a relevant pathway for neurodegeneration and has been previously associated with ALS *(Lin et al., 2017).* S2B candidate GO groups have higher fold enrichments (ratio between frequency of GO term in the gene list and frequency of the same GO term in the background (the human genome)) than MND-DGs unique GO groups (bold border dots in **Figure 3.2A,C**). Although MND-DGs set gathers the highest number of nervous system-related groups (**Figure 3.2C**), these have lower fold enrichment when compared with those present in both S2B candidates and MND-DGs sets (**Figure 3.2B**). S2B stronger associations are possible due to the higher specificity of processes enriched in the candidate set.

Overall, FEA of S2B candidates identifies biological processes similar to those found simultaneously in ALS and SMA DGs. However, S2B has a higher capacity to uncover specific processes linked to MND phenotypes. S2B candidates are also significantly enriched in genes associated with neurological, mental and muscular diseases (**Supplementary Results**). This association is an independent observation supporting S2B ability to identify genes in disease module overlaps.

3.5.5 S2B candidates are enriched in DGs simultaneously associated with ALS and SMA identified from different sources

To further validate S2B predictions, we searched for different evidence sources from which DGs for ALS and SMA could be retrieved. We collected DGs from Open

Targets *(Koscielny et al., 2017)* and Diseases *(Pletscher-Frankild et al., 2015)* and filtered out DGs that were in common with DisGeNet or OMIM, or that were not mapped in the APID3HuRI interactome. Open Targets, DISEASES and DisGeNet have text mining approaches and some experimental information sources in common, but resulting disease associations are not extensively overlapping. To complement the list of ALS and SMA DGs not used as input for S2B, we performed a pubmed abstract search for all proteins in the APID3HuRI interactome that were not associated with ALS or SMA through DisGeNet or OMIM. The intersection of these novel DGs sets and the S2B candidate list is reported in **Table 3.2**. S2B candidates are significantly enriched for ALS and SMA DGs obtained from the three sources. Particularly relevant, and in agreement with S2B rationale, is the fact that our candidates have a stronger enrichment for DGs associated simultaneously with both diseases. Overall, we found independent evidences that 99 S2B candidates (out of the 206 not previously associated) are associated with ALS or SMA, 37 of which have evidences for association with both diseases (**Supplementary Data S3).**

Table 3.2 Enrichment of S2B candidates in ALS and SMA DGs from diferent evidence sources.

Open Targets and DISEASES platforms were queried for ALS and SMA DGs. For the Pubmed abstracts category, a gene was considered associated with a disease if at least one abstract contained the gene symbol and the disease name ("Amyotrophic Lateral Sclerosis" or "Spinal Muscular Atrophy"). Abstract search was performed with the reutils R package. S2B candidates and interactome network nodes that were DGs identified through DisGeNet or OMIM were excluded from this analysis. p-values were computed with an hypergeometric test. S2B candidates that are DGs according to these sources and the pmid of the associated abstracts are available in supplementary data.

DGs not present in DisGeNet or OMIM		S2B candidates (206 proteins)	APID3HuRI network (10991 proteins)	Fold Enrichment	p-value
Open Targets	ALS	44	1242	1.89	$<10^{-5}$
	SMA	8	152	2.8	0.005
	Both	6	72	4.45	0.001
DISEASES	ALS	4	77	2.77	0.043
	SMA	3	13	12.31	0.017
	Both	1	1	53.35	$<10^{-6}$
Pubmed abstracts	ALS	72	1482	2.59	$<10^{-6}$
	SMA	48	641	3.99	$<10^{-6}$
	Both	37	413	4.78	$<10^{-6}$

3.5.6 S2B candidate interaction network highlights molecular connections between ALS and SMA

Seeking mechanistic hypothesis explaining MND phenotypes, we explored the physical interactions between S2B candidates (**Figure 3.3**) recovered from the APID3HuRI interactome. Out of the 232 candidates linking ALS and SMA, 215 are connected in a network component through 603 interactions.

With the S2B candidate subnetwork we aim to demonstrate that our method output is not only a ranked list of proteins. Using the knowledge about the interaction between S2B candidates, we can search for groups of proteins that may be stronger candidates together than individually. We followed two approaches to identify structurally coherent subgroups within S2B candidates. First, we identified cliques (groups in which every protein interacts directly with all other members of the group) with more than 3 elements. The high connectivity of cliques may identify functional complexes. Second, we clustered proteins that co-ocurred in the shortest paths used by S2B (**Supplementary Results**). These clusters highlight pathways linking ALS and SMA DGs, suggesting common MND triggering factors.

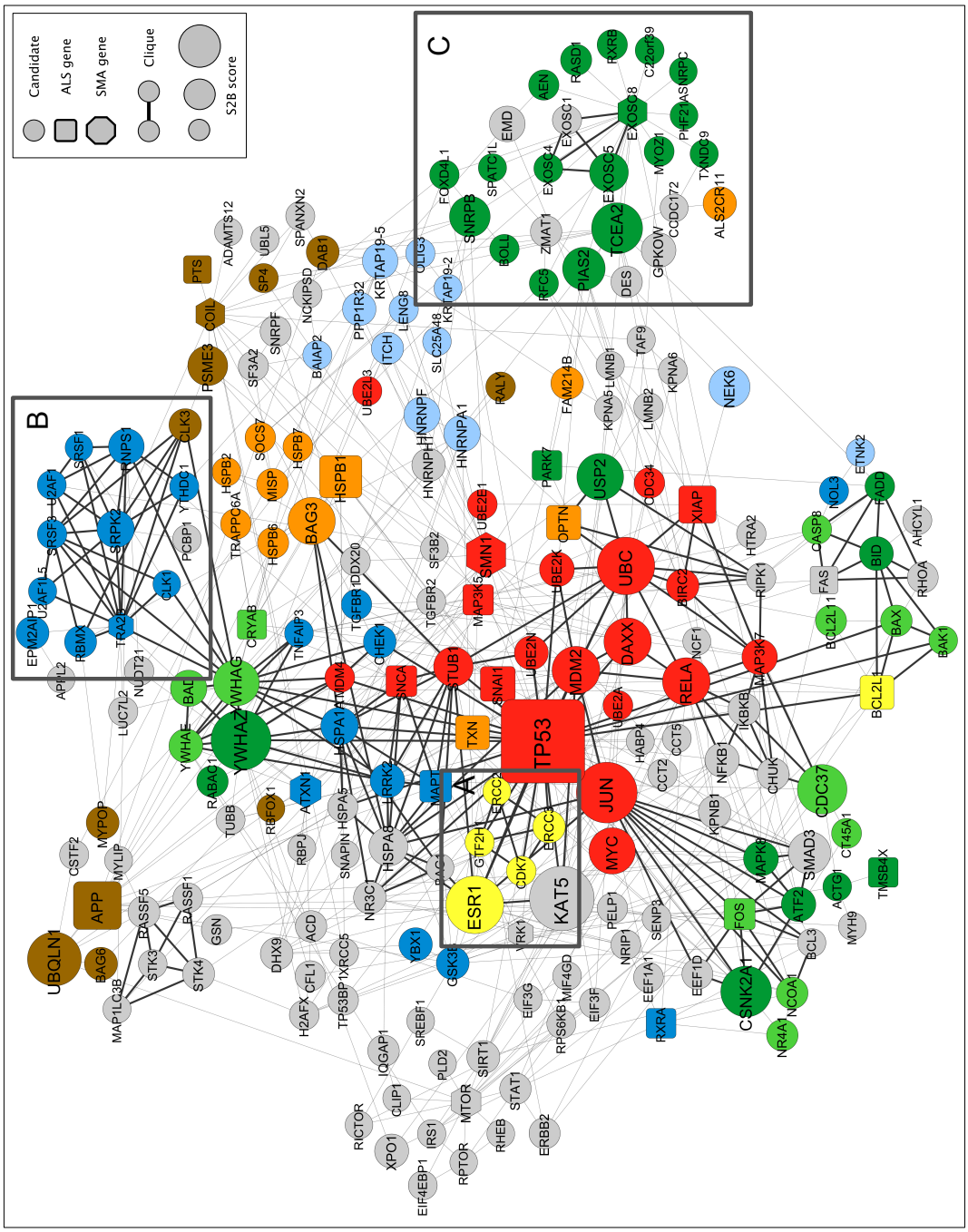


Figure 3.3 S2B candidate interaction network.

Edges represent direct physical interactions between S2B proteins retrieved from the APID3HuRI interactome. Cliques of at least 4 proteins are highlighted with black edges. Clusters formed by proteins that appear frequently together in the shortest paths used by the S2B method (supplementary text) are labeled by node color. **A, B** and **C boxes** outline examples in which cliques and clusters overlap. S2B candidates simultaneously identified as ALS or SMA Disease Genes are denoted by node square shape. Node size is proportional to the S2B score.

The first approach returned 75 cliques divided in three connected components (black edges in **Figure 3.3**). The overlap between most cliques demonstrates the high density of interactions among candidates. The second approach returned 8 clusters (labeled by node colors in **Figure 3.3**) with an average size of 17 proteins (ranging from 6 to 33). Interestingly, identified cliques and clusters display frequent overlap, which would be expected if S2B candidates link ALS and SMA disease modules through discrete molecular pathways.

The most coherent overlap is found around the yellow cluster (**Figure 3.3A),** which captures four of the ten subunits of transcription factor TFIIH complex, involved in RNA polymerase II (Pol II) dependent transcription and the DNA Nucleotide Excision Repair (NER) pathway. The TFIIH core complex is formed by 7 subunits, including the ERCC2 and ERCC3 DNA helicases, which help to create the transcription bubble *(Tirode et al., 1999).* The activity of RNA polymerase II (Pol II) is induced by anchoring the CDK-activating kinase complex (CAK) to the TFIIH core complex. The CAK subcomplex is composed of MAT1, cyclin H and CDK7. The cluster further contains the GFH2H1 gene encoding the TFIIH-core complex p62 subunit, primarily involved in NER pathway *(Wu et al., 2013).*

A relation between neurodegeneration and DNA damage has been proposed *(Madabhushi et al., 2014).* This connection assumed particular relevance for MND with the discovery of mutations causing a juvenile form of ALS (ALS4) and autosomal dominant proximal spinal muscular atrophy (AOA2) in the gene encoding senataxin (SETX) *(Chen et al., 2004; Moreira et al., 2004)*. Senataxin is a DNA-RNA helicase involved in RNA metabolism and DNA integrity maintenance *(Skourti-Stathaki et al., 2011).* Strikingly, Senataxin and SMN protein have been found to collaborate in resolving DNA/RNA hybrids (R-loops), a process that requires tight balance to keep a commitment between correct RNA transcription and DNA damage control *(Zhao et al., 2016).* Recently, a growing number of reports point to R-loops and DNA damage as a key commonality between ALS and SMA *(Farg et al., 2017; Hill et al., 2016; Jangi et al., 2017; Salvi and Mekhail, 2015; Wang et al., 2013).* It is thus quite striking that proteins central to the transcription coupled repair and NER pathways have been selected as top candidates by S2B.

A second cluster highlighted in **Figure 3.3B** also displays a large overlap with a clique group. This group is dominated by splicing-related proteins such as SR proteins (SRSF1, SRSF3), SR-regulating kinases (SRPK2, CLK1, CLK3), general splicing factors (U2AF1, U2AF1L5) and splicing auxiliary components (YTHDC1, RNPS1). The group further includes RBMX and TRA2B (SFRS10), two RNA splicing regulators. Splicing is one of the critical functions that has been proposed to be altered in SMA, since the best known role for the SMN protein is the biogenesis of the splicing machinery. The SMN protein is further involved in generating the core machinery for other RNA-metabolism related functions including histone mRNA processing and cytoplasmic mRNA turnover *(Li et al., 2014).* The connection to splicing was also observed in ALS, as two of the most well studied disease causing mutations involve the TDP-43 and Fus proteins, which both act as splicing regulators *(Gama-Carvalho et al., 2017).* Splicing regulation relies heavily on multifunctional proteins that tend to establish self-regulatory interaction to control their expression levels. RBMX (also called hnRNPG) and TRA2B are able to act as either activators or repressors of splicing (Nasim et al., 2003). Interestingly, RBMX has been shown to act together with TRA2B to regulate splicing of the main SMA modifier gene, SMN2 *(Hofmann and Wirth, 2002).*

RNA binding proteins have also been shown to be closely involved in the maintenance of genome integrity and in the response to DNA damage *(Shkreta and Chabot, 2015).* This seems to involve both the establishment of direct interactions with nascent transcripts to prevent genomic instability, and the regulation of splicing of DNA repair, cell cycle and apoptosis genes. Within the members of this cluster; SRSF1, SRSF3, SRPK2, CLK1, U2AF1, RNPS1, RBMX and TRA2B have all been implicated in this process *(Shkreta and Chabot, 2015)*. These candidates may thus highlight novel elements that disturb RNA processing networks critical for in MND phenotypes.

A third cluster-clique overlap is centered on the RNA exosome complex components EXOSC4, EXOSC5 and EXOSC8 **(Figure 3.3C).** The RNA exosome is a conserved multi-protein complex located in the nucleus and the cytoplasm and is critical for both processing and degradation of various RNAs *(McIver et al., 2016).*

Several tissue-specific diseases and complex disorders have been linked to mutations in exosome complex proteins *(Morton et al., 2018)*. In fact, EXOSC8 is an SMA associated gene *(Boczonadi et al., 2014).* Interestingly, this clique is integrated in a cluster that captures the SNRPB, SNRPC, PHF21A, and TCEA2 genes, among others.

The SNRPB gene encodes the Sm B/B' protein, a component of the spliceosomal U1, U2, U3 and U5 small nuclear ribonucleoproteins (snRNPs), the building blocks of the spliceosome. Sm proteins are recognized by the SMN complex, which assembles them in a ring-like structure around the snRNAs, a function that is compromised in SMA leading to changes in the relative proportions of snRNP complexes *(Wu et al., 2013).* The interaction between EXOSC8 and SNRPB (**Figure 3.3C**) goes in line with previous studies reporting that the Sm complex is required for the processing of small non-coding RNAs by the exosome *(Coy et al., 2013).* In contrast to SNRPB, SNRPC encodes a U1snRNP-specific accessory protein. U1snRNP complex interactions have recently been highlighted as an important link between ALS and SMA *(Gama-Carvalho et al., 2017).*

PHF21A (BHC80) also interacts with EXOSC8 (**Figure 3.3C**). It is a component of histone deacetylase BHC complex and mediates transcriptional repression of neuron-specific genes in non-neuronal cells *(Iwase et al., 2004)*. Conversely, PHF21A protein recognizes H3K4 specific methylation states, an histone that is associated to neurodevelopmental diseases such as Autism Spectrum Disorders *(Vallianatos and Iwase, 2015).* It is known that histone biogenesis disturbance may contribute to the etiology of SMA since low levels of SMN affect U7snRNP biogenesis and, in consequence, histone mRNA processing *(Tisdale et al., 2013)*. This cluster reveals that MND phenotypes might be also influenced by tissue-specific chromatin deregulation events.

The cluster surrounding EXOSC8 further includes the transcription elongation factor TFIIS encoded by TCEA2. TFIIS is a critical factor for efficient transcription elongation and interestingly, a top 10 ranked S2B candidate (**Figure 3.3C**). TFIIS directly binds Pol II to stimulate its release from promoter proximal positions and

thereby produce full length transcripts *(Guo and Price, 2013).* Thus, this cluster reveals strong links between RNA transcription, processing and turnover. On the other hand, recent results highlight important functions for the nuclear exosome in the response to DNA damage, including direct interactions with the Senataxin protein, which acts as an exosome co-factor for sites of transcription-induced DNA damage *(Richard et al., 2013).*

The examples used for detailed exploration of the S2B candidate network (**Figure 3.3A-C**) were selected based solely on structural reasons. However, they outlined a tight relationship between RNA homeostasis (transcription, splicing and degradation) and DNA damage repair that, together with the previous knowledge about ALS and SMA DGs, supports its implications on MND etiology. We believe this analysis demonstrates S2B usefulness to predict protein candidates linking ALS and SMA and furthermore, suggest potential mechanisms that explain the molecular relation between the two diseases.

3.6 Supplementary Results

Double specific-betweenness (S2B) is a network analysis method tailored to take advantage of diseases known to have common phenotypes and predict novel cross-disease associated genes (cDGs). The principle behind S2B is that network paths connecting a protein associated with one disease to a protein associated with the other disease should go through proteins in the overlap between disease modules. Therefore, if we analyze all the known shortest paths linking one disease module to the other, the more frequent members of those shortest paths are very likely in the overlap between disease modules.

The S2B method main inputs are protein interaction networks and lists of Disease Genes (DGs) known to be associated with the two diseases (seeds) (**Supplementary Figure S3.2A**). The core of the method is the computation of a version of Betweenness centrality measure - number of times a protein is part of a shortest path - that is specific for the lists of DGs (Fig S2B). For each node in the network, S2B counts the number of times the node is part of a shortest path between

proteins encoded by Disease A Genes to Disease B Genes. Shortest paths longer than the networks average path length are not included to avoid the influence of proteins loosely related to one of the diseases (yellow nodes in **Supplementary Figure S3.2B**).

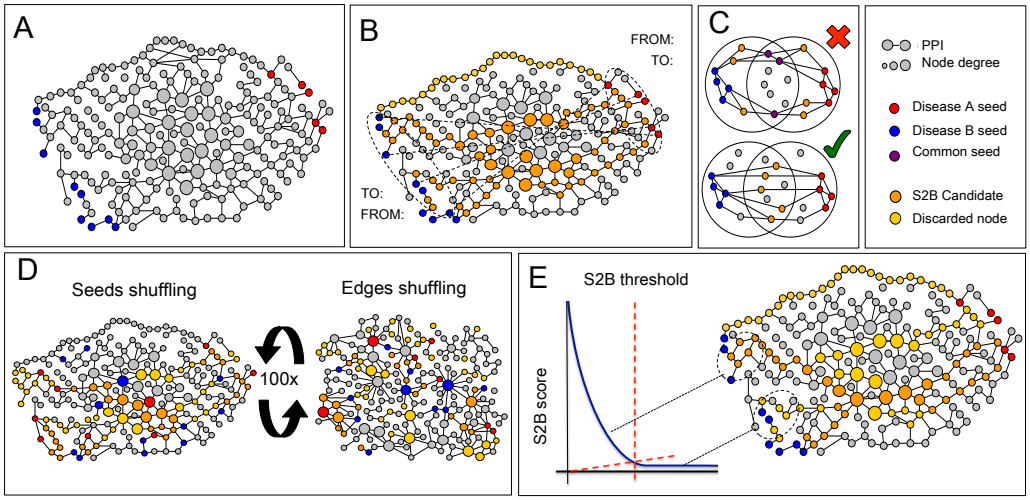


Figure S 3.2 Overview of the S2B method

(A) Human interactome network construction. A global interactome network is constructed using physical protein interaction (PPI) data. Seeds of Disease A and B are identified using gene-disease association data. **(B)** Specific Betweenness Count. The S2B method exploits a novel version of Betweenness centrality measure that counts the number of times a node is involved in a shortest path linking Disease A to Disease B seeds. Shortest paths longer than the network average path length are excluded to avoid the influence of loosely related proteins (yellow nodes). **(C)** Seed filtering. In order to improve S2B performance, genes associated simultaneously to both diseases are discarded. **(D)** Specificity scores **(SS)**. The S2B method includes two specificity scores derived from two types of randomizations that measure how many times a node has a higher specific *S2B* in the original interactome than in randomized networks. The first randomization consist on shuffling the identity of seeds while preserving network structure (1D). In the second, all network edges are shuffled maintaining the degree of nodes in the network (2D). **(E)** S2B candidates selection. First, *S2B* is normalized by dividing it by the number of shortest paths, shorter than the average path length, linking seeds in the network. Then, the *S2B* threshold is defined as the point at which ranked *S2B* decrease rate shifts upwards (described in methods). Final S2B candidates are those proteins that have both SS higher than 0.90 and overcome the *S2B* threshold.

Proteins associated to both diseases are also discarded as these proteins, by definition, belong to the disease modules overlap (**Supplementary Figure S3.2C**). Therefore, shortest paths starting from these proteins diverge from the overlap, increasing the chances of crossing with other shortest paths outside the overlap region. A second layer of specificity is introduced by evaluating if some nodes have high specific *S2B* just because they are very central in the network. To detect these

nodes, *S2B* is recomputed in randomized networks (**Supplementary Figure S3.2D**). If random *S2B* values have similar or higher values than the original *S2B*, then the node is not specifically linking the two sets of seeds. Although these nonspecific nodes can be part of the disease module overlap, they would probably have high *S2B* for many different diseases or if the seeds were random sets of proteins. Specificity Scores (SS) are measured as the fraction of randomized networks yielding lower *S2B* when compared with the original network. Two sorts of network randomization are employed, either by randomly permuting seed protein identity or by shuffling network edges maintaining node degree (number of incident edges). The first method allow us to ask if nodes with high *S2B* are specific for the seeds used, while the second method asks if high *S2B* values are specific for particular pathways in the network.

To enhance the comparability of *S2B* values across different networks or input seed sets, we compute a normalized *S2B* that results from dividing *S2B* values by the total number of shortest paths between seed nodes smaller than the average path length. During the method development we observed that the distribution pattern of *S2B* across the nodes in the network is invariant. If *S2B* are plotted in decreasing order, an L-shape is observed (**Supplementary Figure S3.2E**). This means that there is a small fraction of nodes with high *S2B* while most of the network nodes have very small scores. We define an *S2B* threshold that divides the L-shaped curve in two parts, finding the point that is closest to the origin of the plot (described in methods). To the left of that point we find the set of nodes in the network that accumulate the highest *S2B*. S2B candidates are required to have both SS higher than 0.90 and a *S2B* higher than the *S2B* threshold (orange nodes in **Supplementary Figure S3.2E**).

Identification of common Motor Neuron Disease genes using S2B To evaluate the potential of application of the S2B method, we decided to focus on the Motor Neuron Diseases (MND) Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA) as a case-study. There are numerous ALS and SMA Disease Genes (DG), known to be involved in closely related functions. The genotypic and phenotypic similarities between MND suggest that the ALS and SMA disease modules overlap. The S2B method could therefore help to further define the MND

molecular landscape and possibly identify key elements responsible for triggering MN degeneration.

The first step of the S2B method is to map known MND-DGs (seeds) onto interaction networks (**Supplementary Data S3.1**). Considering that different networks are currently available for the human interactome, we first began by assessing how S2B predictions can be influenced by the source type and quality of the interaction data used. Thus, the S2B method was applied to human protein interaction networks from two different origins. The APID (Agile Protein Interaction Dataserver) repository *(Alonso-Lopez et al., 2016)* gathers protein interactions reported in the literature, while the HuRI (Human Reference Protein Interactome Mapping Project) database is the result of unbiased large scale screens for binary interactions between human proteins *(Rolland et al., 2014; Rual et al., 2005; Venkatesan et al., 2009; Yang et al., 2016; Yu et al., 2011).* Literature-based protein interaction networks are more densely connected around proteins of biomedical interest, while large scale experimental techniques may fail to detect interactions between certain types of proteins, such as membrane proteins (Brito and Andrews, 2011). Both kinds of biases may condition S2B candidate selection. In the case of APID interaction data, three networks with increasing degree of confidence were assembled by only including interactions described in a minimum of two (APID2), three (APID3) or four (APID4) independent experiments. We compared the fraction of common seeds and S2B candidates among the four networks, taking into account the different network properties and intersections (**Supplementary Figure S3.3**).

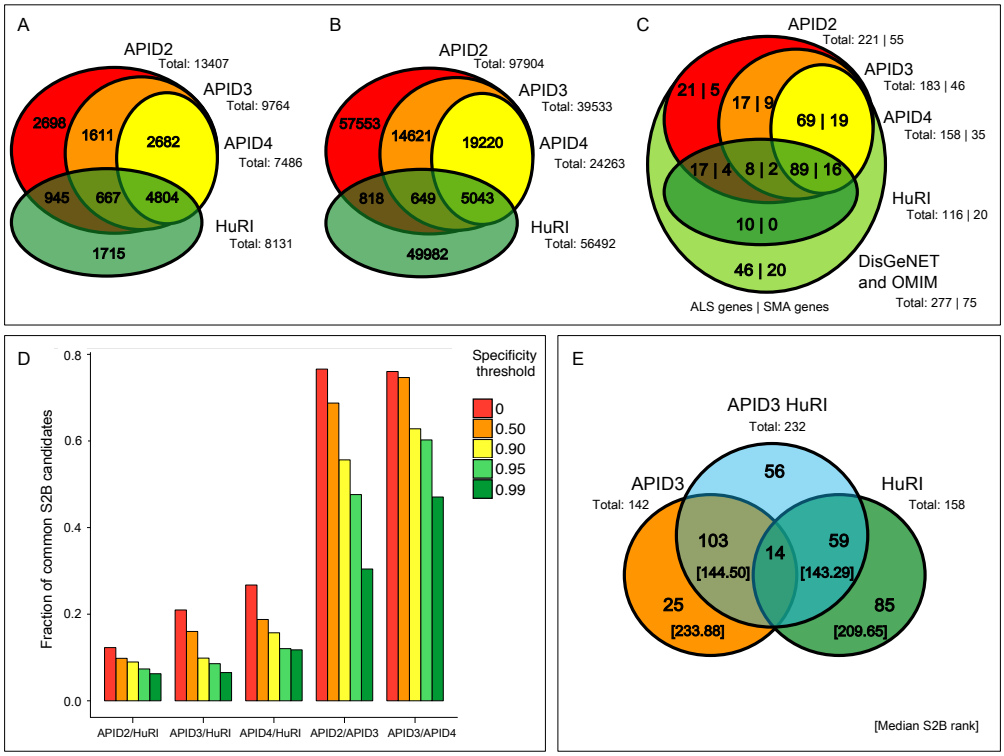


Figure S 3.3 Impact of network characteristics on S2B method predictions

Four protein interaction networks were constructed using different source and quality data. HuRI is constructed with unbiased high throughput-derived binary interactions whereas APID networks combine literature-derived and experimentally validated data. APID2, APID3 and APID4 include interactions described in a minimum of two, three and four independent experiments, respectively. **(A)** Intersection of nodes between networks. **(B)** Intersection of edges between networks. **(C)** Intersection of mapped ALS and SMA DGs between networks. DGs associated to ALS and SMA simultaneously were discarded. **(D)** Fraction of S2B candidates in common among input interaction networks and according to the specificity threshold used. The fraction of common S2B candidates among networks was computed using the smallest network candidate list as a reference. **(E)** Intersection between S2B candidates obtained using APID3, HuRI or the merged network APID3HuRI. The median rank of the S2B candidates found only in APID3 or HuRI were compared against those simultaneously selected using the APID3HuRI network.

APID networks of higher confidence are completely contained in APID networks of less quality (**Supplementary Figure S3.3A-B**). The intersection between APID networks and HuRI is high in terms of proteins but low when edges are considered, which reflects the dissimilarities between different approaches for PPI detection. The presence of a higher number of MND-DGs in APID versus the HuRI network (**Supplementary Figure S3.3C**) is in agreement with the underlying network generation procedures – literature based for APID *versus* unbiased screening for HuRI.

As expected, the increase on S2B Specificity threshold always induces a decrease on the fraction of S2B candidates common to the compared networks (**Supplementary Figure S3.3D**). S2B candidates may differ between networks because the input lists of MND-DGs vary between networks. Likewise, different nodes and edges may reroute shortest paths between disease proteins. Even if for some nodes the shortest paths are conserved across networks, changes in one network context may lower specificity scores and change candidate selection.

The APID2/APID3 and APID3/APID4 network pairs display the highest number of common candidate genes in the absence of a specificity threshold (**Supplementary Figure S3.3D**). Interestingly, the APID3/APID4 overlap is more robust to increasing specificity thresholds, likely reflecting the greater interaction quality of the underlying networks. Likewise, despite of the intrinsic dissimilarities between the HuRI and APID networks, the fraction of HuRI S2B candidates in common with APID increases with interaction quality for all specificity thresholds (**Supplementary Figure S3.3D**). These results suggest that the removal of less reliable interactions has a positive impact on the S2B method capacity to identify the best candidates. On the other hand, higher quality networks are smaller, leading to lower number of mapped DGs (**Supplementary Figure S3.3C**) and the consequent loss of input information for the method.

To maximize global interactome and DG coverage while avoiding poor quality interactions, we opted to merge HuRI and APID3 networks for subsequent analysis. Moreover, the sizes of APID3 and HuRI networks are more similar, which allows a balanced mix of data derived from high-throughput experiments and literature knowledge. The S2B method applied to 197 ALS and 48 SMA DGs within the APID3HuRI network returned 232 candidate proteins potentially related with both diseases simultaneously (**Supplementary Data S3.2**). 82% of the S2B candidates identified with APID3 alone are also found with the merged network APID3HuRI (**Supplementary Figure S3.3E**). Though APID3HuRI candidates captured only 46% of the ones obtained with HuRI alone, the S2B candidates only identified in HuRI or APID3 separately have lower median S2B scores than those found simultaneously in APID3HuRI. Therefore, the use of a combined network returns the most robust

candidates of the individual network analysis. Additionally, APID3HuRI also identifies new S2B candidates (**Supplementary Figure S3.3E**), showing that the combination of both networks produced different shortest paths that uncovered possibly relevant proteins.

S2B candidates are associated with related diseases According to our hypotheis, S2B candidates may be causal, modifiers or directly involved in the phenotypes common to both diseases. It is then logical to expect that some of these candidates may also be associated with other diseases that share phenotypic fetures or affected pathways. According to the DisGenet database, 146 out of the 232 S2B MND candidates are also associated with at least one disease or pathological phenotype (gene-disease associations supported only by text mining were discarded). This set of candidates is actually statistically enriched in associations with 61 diseases (Hypergeometric test, FDR<0.05, complete list in **Supplementary Data S3.5**). The disease enrichment is dominated by 27 cancer related conditions. This may be explained by an intrinsic bias in gene-disease association databases, but also by our previous observation that candidate proteins are enriched in cancer related processes like DNA repair and cell cycle. More interestingly, S2B candidates are enriched in 6 neurological, 2 mental and 3 muscular disorders. This prompted us to analyze the interactions of candidate genes and these three types of disease by building a bipartite network with two distinct types of nodes (genes and diseases), where edges only connect nodes of different types (**Supplementary Figure S3.4**).

Out of a total of 232 S2B candidates, 93 have at least one association with neurological, mental or muscular diseases. In 1000 random sets of 232 genes from the interactome (including 5 SMA seeds, 19 ALS seeds and 2 genes associated with both ALS and SMA, mimicking the composition of the S2B candidate set) the number of genes associated with these types of disease was always significantly lower lower than this (median of 57, 95% confidence interval: [45, 69]). This indicates that S2B candidates are significantly enriched ($p<0.001$) in genes associated with neurological, mental and muscular diseases.

Mental disorders present the highest number of S2B gene-associations (Schizofrenia (28), Depressive disorders (21) and Bipolar disorder (14)). A large diversity of neurological diseases is represented in the network, including neurovascular related (Brain Ischemia (11), Cerebral Hemorrage (7), Cerebral Infarction (6)) and neurodegenerative diseases (Parkinson(11), Alzheimer (8), Cerebellar Ataxia (8), Demyelinating diseases (5), Spinocerebellar Ataxia (4), Ceroid Lipofuscinosis (3) and Pontocerebellar Hypoplasia (3), besides ALS (12) and SMA (3)). Muscular related disorders are represented by Muscular Atrophy (10), Muscular Dystrophy (6), Limb-girdle Muscular Dystrophy (4) and Myopathy (4). Gene-wise, APP (11), GSK3B (9), MTOR (9), BAX (7), DAG1 (7), ERCC2 (7) and SIRT1 (7) have the higher number of disease associations. More interestingly, BAG3, CCT5, DAG1, GSK3B, HNRNPA1, HSP1B, MTOR and OPTN are simultaneously associated with neurological and muscular diseases.

The association of a high number of S2B candidates with other diseases related with ALS and SMA is an independent observation that supports the ability of the S2B method to identify functionally relevant genes in disease module overlaps.

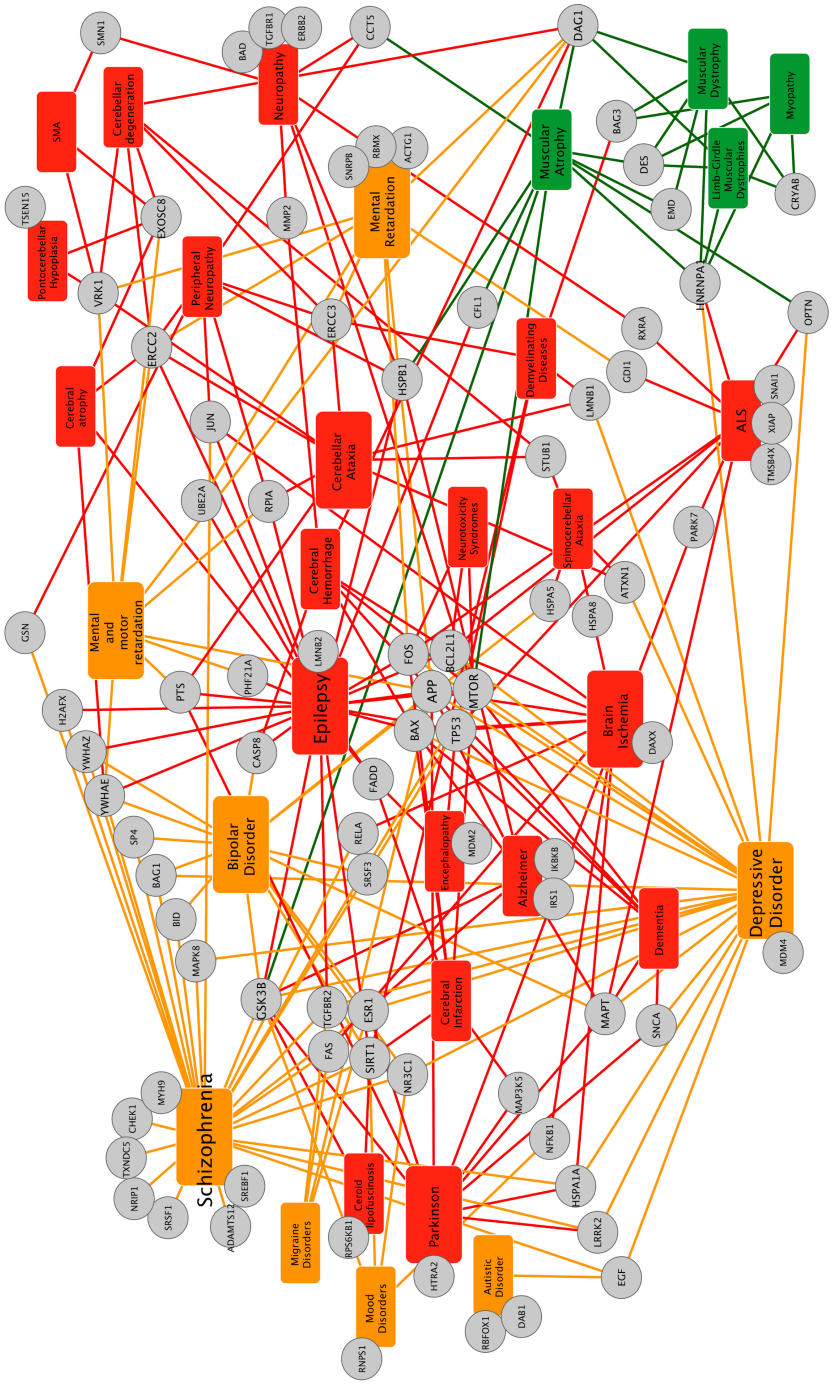


Figure S 3.4 S2B candidate-disease interactions

Diseases associated with S2B candidate proteins (gray nodes) were retrieved from the DisGenet database. Only disease associations with a score higher than 0.08 were used, which discards association based solely on text-mining evidence. All retrieved disease associations are available in Supplementary File S6. Only associations with Mental (orange nodes), Neurological (red nodes) or Muscular (green nodes) related diseases are represented. Some disease denominations were manually edited to merge subtypes of the same disease. Only diseases associated with at least 3 S2B candidates were included in the network.

3.7 Discussion

S2B is built upon the hypothesis that disease genes tend to interact in cellular networks within disease modules and that related diseases have an overlap between their modules. The frequency with wich nodes belong to shortest paths between nodes associated with two related diseases (cDGs) allows the detection of specific nodes bridging disease modules. S2B performance with artificial modules shows that nodes with high *S2B* have a high likelihood of belonging to the overlap between modules. Moreover, this predictive capacity is robust to changes in module topology, to the quantity and quality of the input DGs and network interactions. Our results with artificial modules also support the use of S2B to predict the overlap between network modules of varied type, such as functional modules or context-specific subnetworks.

In the artificial module analysis, we generated and controlled the complete composition of each module, and selected for analysis pairs of modules with overlap. In this selection, we did not control for the presence of network hubs in the overlap. For this reason, applying the specificity thresholds in the analysis of artificial modules should not bias the method performance. Concordantly, it can be observed in **Supplementary Figure S3.5** that proteins with higher S2B values are not biased to pass the filters for both specificity scores.

Network hubs can indeed be part of the overlap between real disease modules and have a significant role connecting the mechanisms of both diseases. However, they are not interesting candidates for follow up studies, since they tend to be unspecific and simultaneously related with many different cellular processes. Therefore, specificity score filtering is important for the analysis of real disease seed sets.

In the study of ALS and SMA, S2B successfully returned candidates involved in processes known to be part of motor neuron degeneration mechanisms, such as apoptosis, DNA repair, RNA processing, protein transport or cytoskeleton organization *(Gama-Carvalho et al., 2017).* More specifically, S2B candidates were enriched for DGs simultaneously associated with ALS and SMA through different information sources and not used as input for S2B predictions.

Some of the cliques and clusters in the candidate interaction network were involved in several of these processes, which suggests that disease proteins tend to be located at the interface between functional modules and corroborates that disease modules do not overlap perfectly with functional and topological network modules *(Barabasi, 2007; Ghiassian et al., 2015).* Many of the S2B candidates were already associated with multiple diseases, some of them closely related with ALS and SMA. Together with the observation that most candidates interact in a densely connected network, these results reinforce the hypothesis that DGs tend to interact with other DGs, specially if the two diseases are related through similar causes or phenotypes *(Goh et al., 2007).*

S2B can be applied to uncover common molecular mechanisms shared by various diseases. Its discovery potential can be amplified through the use of different networks types, such as signaling and gene regulatory networks, and by integrating genome scale molecular data characterizing healthy and disease states. In summary, this work provides a novel approach to predict the overlaps between network modules, which can uncover disease mechanisms through network exploration for pathologies with phenotypic similarity. Its application to the motor neuron diseases SMA and ALS identified several novel genes as potentially involved in critical pathomechanisms, opening new hypothesis for experimental exploration.

3.8 Supplementary Data

R code and input data to reproduce BioInt-U method is available in

https://github.com/GamaPintoLab/S2B

Supplementary data files are available in

https://github.com/GamaPintoLab/MLG_PhDThesis_SupData

Supplementary Data S3:

Sheet.1 List of ALS and SMA seeds

Sheet 2 List of S2B candidates

Sheet 3 List of enriched GO terms

Sheet 4 List of enriched GO groups

Sheet 5 List of enriched diseases

Sheet 6 List of disease-gene associations

Sheet 7 Composition of cliques

Sheet 8 Composition of clusters

4 Analysis of pre-symptomatic *Drosophila* models for ALS and SMA reveals convergent impact on functional protein complexes linked to neuro-muscular degeneration

Data presented in this chapter was included in the following work:

Garcia-Vaquero M.L., M Heim., Flix B., Pereira M., Palin L., Marques T.M., Pinto F.R., De Las Rivas J., Voigt A., Besse F., and Gama-Carvalho M. (2022). Analysis of pre-symptomatic *Drosophila* models for ALS and SMA reveals convergent impact on functional protein complexes linked to neuro-muscular degeneration. Pre-print in bioRxiv, under revision.

Author contributions:

MG-C, AV, FB, JW, JS and **JDLR** conceptualized the research approach and supervised the research work; **BF** developed the fly-lines and generated the RNAi samples; **MH** generated and characterized the GFP-tagged lines and did the RNA-IP assays; **MP** and **TMM** did the RNA-seq data analysis; **LP** assessed the functionality of GFP-tagged lines; **MG-V** developed the methods and performed the network-based analysis; **FRP** contributed to the conceptualization and development of the network analysis. **MG-C, AV, FB** and **MG-V** wrote the manuscript draft.

4.1 Abstract

Background: Spinal Muscular Atrophy and Amyotrophic Lateral Sclerosis share both phenotypic and molecular commonalities, including the fact that they can be caused by mutations in genes encoding proteins involved in RNA metabolism, namely Smn, TDP-43 and Fus. Although this suggests the existence of common disease mechanisms, there is currently no model to explain the converging motoneuron dysfunction caused by changes in the expression of these ubiquitous genes.

Methods: In this work we generated a parallel set of *Drosophila* models for adult-onset RNAi and tagged neuronal expression of the orthologues of SMN1, TARDBP and FUS (Smn, TBPH and Caz, respectively). We profiled nuclear and cytoplasmic bound mRNAs using a RIP-seq approach and characterized the transcriptome of the RNAi models by RNA-seq. To unravel the mechanisms underlying the common functional impact of these proteins on neuronal cells, we devised a computational approach based on the construction of a tissue-specific library of protein functional modules, selected by an overall impact score measuring the estimated extent of perturbation caused by each gene knockdown.

Results: Our integrative approach revealed that although each disease-associated gene regulates a poorly overlapping set of transcripts, they have a concerted effect on a specific subset of protein functional modules, acting through distinct targets. Most strikingly, functional annotation reveals these modules to be involved in critical cellular pathways for neurons and in particular, in neuromuscular junction function. Furthermore, selected modules were found to be significantly enriched in orthologues of human genes linked to neuronal disease.

Conclusions: This work provides a new model explaining how mutations in SMA and ALS-associated disease genes linked to RNA metabolism functionally converge to cause motoneuron dysfunction. The critical functional modules identified represent interesting biomarkers and therapeutic targets given their identification in asymptomatic disease models.

4.2 Background

Motor neuron diseases (MNDs) are characterized by a progressive and selective degeneration and loss of motor neurons accompanied by an atrophy of innervated muscles. Although MNDs encompass heterogeneous groups of pathologies with different onset and genetic origins, a number of MND-causing mutations have been identified in RNA-associated proteins, leading to a model in which alteration of RNA metabolism may be a key, and potentially common, driver of MND pathogenesis *(Achsel et al., 2013; Gama-Carvalho et al., 2017; Ling et al., 2013; Taylor et al., 2016; Zaepfel and Rothstein, 2021)*. This has become particularly clear in the context of two well-studied pathologies: spinal muscular atrophy (SMA) and amyotrophic lateral sclerosis (ALS), which have both been linked to mutations in conserved RNA binding proteins (RBPs). SMA, the most common early-onset degenerative neuromuscular disease, is caused in 95% of patients by a loss of the SMN1 gene, which encodes a protein with chaperone functions essential for the assembly of both nuclear and cytoplasmic ribonucleoprotein (RNP) complexes *(Li et al., 2014; Price et al., 2018).* The best-characterized role of SMN is to promote the assembly of spliceosomal small nuclear ribonucleoprotein complexes (snRNPs) *(Boulisfane et al., 2011; Workman et al., 2012),* but it has also been involved in the assembly of other nuclear sRNPs required for 3'end processing *(Tisdale et al., 2013),* as well as cytoplasmic RNP complexes essential for long-distance mRNA transport *(Donlin-Asp et al., 2017, 2016).* Consistent with these functions, and with additional reported roles in transcription regulation, inactivation of Smn was shown to result in alternative splicing defects and defective axonal RNA targeting *(Fallini et al., 2016, 2011)* To date, how these changes in gene expression account for the full spectrum of symptoms observed in SMA patients and disease models remains unclear. ALS, on the other hand, is the most-common adult-onset MND and has mostly sporadic origins. Remarkably, however, disease-causing mutations in two genes encoding RNA binding proteins, Fus and TDP-43 (alias gene symbol of TARDBP), have been identified in both genetic and sporadic forms of the disease *(Da Cruz and Cleveland, 2011; Gama-Carvalho et al., 2017).* These proteins shuttle between the nucleus and the cytoplasm and regulate different aspects of RNA metabolism, ranging from transcription and pre-mRNA splicing to mRNA stability and axonal targeting *(Birsa et*

al., 2020; Ederle and Dormann, 2017; Ratti and Buratti, 2016). ALS-causing mutations were described to have pleiotropic consequences, compromising both the nuclear and cytoplasmic functions of FUS and TDP-43, and resulting in their accumulation into non-functional cytoplasmic inclusions ***(Ling et al., 2013; Zbinden et al., 2020)***. Whether ALS pathogenesis primarily originates from a depletion of the nuclear pool of these RBPs, or rather from a toxic effect of cytoplasmic aggregates, has remained unclear ***(Fernandes et al., 2018; Li et al., 2013).*** Thus, SMA and ALS are not only connected by pathogenic commonalities ***(Bowerman et al., 2018),*** but also appear to both originate from alterations in RBP-mediated regulatory mechanisms. Further strengthening the possibility that these two MNDs may be molecularly connected, recent studies have suggested that SMN, FUS and TDP-43 belong to common molecular complexes and also exhibit functional interactions ***(Cacciottolo et al., 2019; Chi et al., 2018; Groen et al., 2013; Perera et al., 2016; Sun et al., 2015; Tsuiji et al., 2013; Yamazaki et al., 2012).*** Together, these results have raised the hypothesis that SMN, FUS and TDP-43 may control common transcriptional and/or post-transcriptional regulatory steps, the alteration of which may underlie MND progression ***(Achsel et al., 2013).*** Comparative transcriptomic studies performed so far, however, did not clearly identify classes of transcripts that may be co-regulated by the three MND RBPs ***(Gama-Carvalho et al., 2017; Kline et al., 2017; Lagier-Tourenne et al., 2012),*** letting open the question of common molecular regulatory mechanisms and targets.

A major difficulty in comparing available transcriptomic studies is that datasets were obtained from heterogeneous, and often late-stage or post-mortem samples, preventing robust comparisons and identification of direct vs. indirect targets. Another challenge associated with the identification of relevant regulated mRNAs is that SMN, FUS and TDP-43 are multifunctional and may exhibit distinct sets of target RNAs in the nucleus and the cytoplasm, raising the need for compartment-specific studies. To overcome previous limitations and unambiguously assess the existence of transcripts commonly regulated by SMN, FUS and TDP-43, we decided in this study to systematically identify the direct and indirect neuronal RNA targets of these proteins. For this purpose, we defined a strategy involving the establishment of parallel schemes for tagged-protein expression to perform RNP complex purification,

alongside with gene inactivation, using *Drosophila*, a model organism that expresses functional orthologs of SMN (Smn), FUS (Caz) and TDP-43 (TBPH).

Highlighting the conservation of protein functions from fly to human, expression of human FUS and TDP-43 proteins was shown to rescue the lethality induced upon inactivation of the corresponding fly genes ***(Wang and Marcotte, 2010).*** Furthermore, *Drosophila* models based on expression of mutant human or *Drosophila* proteins have been previously established, that recapitulate the hallmarks of SMA and ALS, in particular motor neuron disabilities and degeneration ***(Aquilina and Cauchi, 2018; Liguori et al., 2021; McGurk et al., 2015; Olesnicky and Wright, 2018; Spring et al., 2019; Voigt et al., 2010)***. Several of these models have been successfully used for large-scale screening and discovery of genetic modifiers ***(Chang et al., 2008; Kankel et al., 2020; Liguori et al., 2021).***

Our study was performed on pre-symptomatic flies, starting from head samples. RNA immunoprecipitation sequencing (RIP-seq) experiments were performed to identify the cytoplasmic and nuclear transcripts bound by each protein. These assays were complemented with gene-specific down-regulation followed by RNA sequencing (RNA-seq) to identify transcripts with altered expression levels and/or splicing patterns. The RIP-seq analysis showed that Smn, Caz and TBPH proteins bind to largely distinct sets of RNA targets, whether in the nucleus or in the cytoplasm. The steady state level of this group of transcripts was not particularly affected by the knockdown of Smn, Caz and TBPH, which collectively altered the expression and/or splicing profile of a limited, albeit significant set of common transcripts. However, the functional enrichment analysis of the differentially expressed genes did not reveal any consistent signatures.

These observations suggested that the common physiological processes regulated by the three proteins may be altered at a higher order level. To unravel the functional relationship between the transcripts regulated by Smn, Caz and TBPH, we designed a strategy to map functionally collaborating protein modules in the context of the neuronal interactome. This approach revealed that despite the limited coherence of the transcripts affected by the knockdown of the three proteins, Smn,

Caz and TBPH converge on the regulation of common biological processes. Among these, we identify seven functional units directly implicated in neuro-muscular junction (NMJ) development. Noteworthy, although these modules were selected based on the joint degree of impact from all the knockdowns, they were found to be enriched in transcripts identified in RIP-seq experiments as bound by Smn, Caz and/or TBPH, as well as in proteins whose orthologs have been associated with MNDs. In summary, our work provides a new conceptual model to explain how changes in three ubiquitous proteins involved in RNA metabolism converge into molecular functions critical for MN processes, thereby leading to overlapping disease phenotypes.

4.3 Methods

Fly lines The fly stocks used were obtained from the Bloomington *Drosophila* Stock Center (BDSC) and the Vienna *Drosophila* Resource Center (VDRC), or were generated using the *Drosophila* Embryo Injection Service form BestGene (http://www.thebestgene.com). BDSC stocks #39014 (expressing shRNA targeting TBPH), #55158 (expressing shRNA targeting Smn) and #32990 (expressing shRNA targeting caz) were used for the transcriptome profiling assays along with the VDRC strain #13673 (expressing dsRNA targeting always early). Transgenic lines used for neuronal expression of GFP-tagged variants of Smn (CG16725, fly Smn1), TBPH (CG10327, fly TDP-43), caz (CG3606, fly FUS) were generated by site directed integration into the same attP landing site (VK00013, BDSC#9732).

Smn, Caz and TBPH coding sequences were PCR-amplified from ESTs LD23602, UASt-Caz plasmid (gift from C. Thömmes) and EST GH09868, respectively, using the primers listed in **Table 4.1.** Smn and Caz PCR products were subcloned into pENTR-D/TOPO vector (Life Technologies), fully sequenced, and recombined into a pUASt-EGFP-attB Gateway destination vector to express N-terminally-tagged proteins. The TBPH PCR product was double digested with NotI and XhoI and ligated into a NotI/XhoI digested pUASt-EGFP-attB plasmid (gift from S. Luschnig).

Table 4.1 Primer sequences

Primer name	Sequence (5'-3')
Smn_fwd	CACCATGTCCGACGAGACGAACG
Smn_rev	GATGGAATTACTTCTTGGGTGTC
Caz fwd	CACCATGGAACGTGGCGGTTATGGTG
Caz_rev	TTAATATGGTCTCGAGCGCATGC
NotI_TBPH-fwd	AAAAGCGGCCGCCATGGATTTCGTTCAAG
XhoI_TBPH_rev	AAAACTCGAGTTAAAGAAAGTTTGACTTCTCCGC

Fly crosses dsRNA expression was induced using the GeneSwitch system. Mifepristone was dissolved in 80% ethanol and pipetted on the surface of regular fly food (final concentration of 0,1 mg/cm2). Vehicle-only treated fly vials served as control. Vials were prepared 24 hours prior to use to allow evaporation of ethanol. Crosses performed for knock-down analyses were as follows: virgins carrying the ubiquitous daughterless GeneSwitch driver (daGS) were crossed with males carrying the UAS:shRNA constructs. In the progeny, male daGS/UAS:shRNA flies were collected one day post eclosion (1 dpe) and exposed to food containing mifepristone (replaced every 2 days). After 10 days, flies were collected, snap frozen in liquid nitrogen and stored at -80°C until further use.

For RIP-seq experiments, males carrying UASt-GFP-fusions (or sole EGFP) were crossed en masse with elav-Gal4; tub-Gal80ts virgins. elav-Gal4/Y/+; tub-Gal80ts/UAS-GFP-Smn (or TBPH or Caz) flies were raised at 18°C, switched to 29°C upon eclosion and aged for 5 to 7 days before being collected in 50 mL Falcon tubes and snap frozen.

Immuno-histochemistry and Western-blotting For analysis of GFP-fusion distribution, brains were dissected in PBS and immuno-stained using anti-GFP antibodies (1:1,000; Molecular Probes, A-11122), as described previously ***(Vijayakumar et al., 2019).*** Samples were imaged on an inverted Zeiss LSM710 confocal microscope. For analysis of GFP-fusion expression, heads were smashed into RIPA buffer (15 heads for 100 mL RIPA) and lysates directly supplemented with SDS loading buffer (without any centrifugation). Total protein extracts or RIP extracts were subjected to SDS Page electrophoresis, blotted to PVDF membranes, and probed with the following primary antibodies: rabbit anti-GFP (1:2,500; #TP-401; Torey Pines); mouse anti-Tubulin (1:5,000; DM1A clone; Sigma) and mouse anti-Lamin (1:2,000; ADL 67.10 and ADL 84.12 clones; DHSB).

RNA Immunoprecipitation assays Falcon tubes half-filled with frozen flies were chilled in liquid nitrogen, extensively vortexed so as to separate heads, legs and wings from the main body. Head fractions were collected at 4°C, through sieving on 630 µm and 400µm sieves stacked on top of each other. 1 mL of heads was used per condition, except for GFP-Smn, where 2 mL of heads were used. For the GFP

control, 500 µL of heads were mixed with 500 µL of w1118 heads so as to normalize the amount of GFP proteins present in the initial lysate.

Adult *Drosophila* heads were grinded into powder with liquid nitrogen pre-chilled mortars and pestles. The powder was then transferred to a prechilled 15 mL glass Dounce Tissue Grinder and homogenized in 8.5 mL of Lysis buffer (20mM Hepes pH 8, 125mM KCl, 4mM MgCl2, 0.05% NP40, 1mM dithoithreitol (DTT), 1:100 Halt™ Protease & Phosphatase Inhibitor Cocktail, Thermo Scientific, 1:200 RNasOUT™, Invitrogen). Cuticle debris were eliminated by two consecutive centrifugations at 100 g for 5 minutes at 4°C. Nuclear and cytoplasmic fractions were then separated by centrifugation at 900 g for 10 minutes at 4°C. The supernatant (cytoplasmic fraction) was further cleared by two consecutive centrifugations at 16,000 g for 20 minutes. The pellet (nuclear fraction) was washed with 1 mL of Sucrose buffer (20 mM Tris pH 7.65, 60 mM NaCl, 15 mM KCl, 0.34 M Sucrose, 1 mM dithoithreitol (DTT), 1:100 Halt™ Protease & Phosphatase Inhibitor Cocktail, Thermo Scientific, 1:200 RNasOUT™, Invitrogen), centrifuged at 900 g for 10 minutes at 4°C and resuspended in 2 mL of Sucrose buffer. 800 µL of High salt buffer (20 mM Tris pH 7.65, 0.2 mM EDTA, 25% Glycerol, 900 mM NaCl, 1.5 mM MgCl2, 1 mM dithoithreitol (DTT), 1:100 Halt™ Protease & Phosphatase Inhibitor Cocktail, Thermo Scientific, 1:200 RNasOUT™, Invitrogen) were then added to reach a final concentration of 300 mM NaCl. After 30 minutes incubation on ice, the nuclear fraction was supplemented with 4.7 mL of Sucrose buffer to reach a concentration of 150 mM NaCl and with CaCl2 to reach a 1 mM CaCl2 concentration. RNAse free DNase I (Ambion™, Invitrogen) was added (0.1 mM final concentration) and the sample was incubated for 15 minutes at 37°C with gentle agitation. 4 mM (final) EDTA was added to stop the reaction and the digested fraction was centrifuged at 16,000 g for 20 minutes (4°C) to obtain soluble (supernatant; used for immuno-precipitation) and insoluble (pellet) fractions.

Cytoplasmic and nuclear fractions were incubated for 30 minutes at 4°C under agitation with 120 µL of control agarose beads (ChromoTek, Germany). Pre-cleared lysates were collected by a centrifuging 2 min at 400 g (4°C). Immuno-precipitations were performed by addition of 120 µL of GFP-Trap®_A beads (ChromoTek, Germany) to each fraction and incubation on a rotator for 1.5 hours at 4°C. Tubes were then centrifuged for 2 minutes at 2,000 rpm (4°C) and the unbound fractions

(supernatants) collected. Beads were washed 5 times with Lysis buffer, resuspended in 100 µL of Lysis buffer supplemented with 30 µg of proteinase K (Ambion) and incubated at 55°C for 30 minutes. Eluates (bound fractions) were then recovered and further processed. At least three independent immune-precipitations were performed for each condition.

RNA extraction, Library preparation and RNA sequencing RNA from IP eluates or frozen fly heads (50 flies aprox/genotype) was extracted using Trizol according to the manufacturer's instructions. RIP-Seq libraries were prepared in parallel and sequenced at the EMBL Genomics core facility. Briefly, libraries were prepared using the non-strand-specific poly(A)+ RNA Smart-Seq2 protocol (Nextera XT part). Following quality control, cDNA libraries were multiplexed and sequenced through single-end 50 bp sequencing (HiSeq 2000, Illumina).

RNA-seq libraries for RNAi analysis were prepared and sequenced at the Genomics Facility, Interdisziplinäres Zentrum für Klinische Forschung (IZKF), RWTH Aachen, Germany. Libraries were generated using the Illumina TrueSeqHT library protocol and ran on a NextSeq machine with paired-end sequencing and a read length of 2x76nt. The 47 raw fastq files of the RNA-seq data generated for this study have been submitted to the European Nucleotide Archive under the umbrella project FlySMALS, with accession numbers PRJEB42797 and PRJEB42798.

RNA-seq data analysis Following quality assessment using FastQC version 0.11.5 (https://www.bioinformatics. babraham.ac.uk/projects/fastqc/), all raw sequencing data was processed with in-house perl scripts to filter out reads with unknown nucleotides, homopolymers with length ≥50 nt or an average Phred score < 30, and trim the first 10 nucleotides (Amaral et al., 2014). Remaining reads were aligned to the BDGP D. melanogaster Release 6 genome assembly build ***(dos Santos et al., 2015)*** using the STAR aligner version 2.5.0 ***(Dobin et al., 2013)*** with the following options: –outFilterType BySJout –alignSJoverhangMin 8 –alignSJDBoverhangMin 5 –alignIntronMax 100000 –outSAMtype BAM SortedByCoordinate –twopassMode Basic –outFilterScoreMinOverLread 0 –outFilterMatchNminOverLread 0 –outFilterMatchNmin 0 –outFilterMultimapNmax 1 –limitBAMsortRAM 10000000000 –quantMode GeneCounts. Gene counts were

determined using the htseq-count function from HTseq (version 0.9.1) in union mode and discarding low quality score alignments (–a 10), using the Flybase R6.19 annotation of gene models for genome assembly BDGP6.

For RIP-seq data analysis, gene counts were normalized and tested for DE using the DESeq2 (Love et al., 2014) package of the Bioconductor project (Huber et al., 2015), following removal of genes with less than 10 counts. mRNAs associated with each protein were identified by performing a differential expression analysis (DEA) for each condition vs the corresponding control pull-down. Transcripts with a positive log2 FC and an adjusted p value for DEA lower that 0.05 were considered to be bound by the target protein.

DEA for RNA-Seq gene counts was performed with the limma Bioconductor package (Ritchie et al., 2015) using the voom method (Law et al., 2014) to convert the read-counts to log2-cpm, with associated weights, for linear modelling. The design formula (~ hormone + Cond, where hormone = treated or non-treated and Cond = Caz, Smn or Tbph RNAi) was used to consider hormone treatment as a batch effect. Differential gene expression analysis was performed by comparing RNAi samples for each target protein to control (always early RNAi) samples. Genes showing up or down-regulation with an adjusted p value <0.05 were considered to be differentially expressed.

Altered splicing analysis (ASA) was performed on the RNA-seq aligned data using rMATS version 4.0.2 ***(Shen et al., 2014)*** with flags -t paired --nthread 10 --readLength 66 --libType fr-firststrand. For the purpose of the downstream analysis, the union of all genes showing any kind of altered splicing using the junction count and exon count (JCEC) analysis with a FDR <0.05 in the comparison between each target gene RNAi versus control RNAi was compiled as a single dataset.

Normalized RNA-Seq data of adult fly brain tissue was retrieved from FlyAtlas2 database in November 2020 (www.flyatlas2.org; ***(Leader et al., 2018)).*** Neuronal transcripts were filtered applying an expression threshold of >1 FPKM (Fragments Per Kilobase per Million). This gene set was then used to filter the final gene lists from RIP-seq, DEA and ASA. The full universe of 8,921 neuronal genes is annotated in **Supplementary Data S4.5**. Clustering analysis was performed using the heatmap function from ggplot2 R package ***(Wickham, 2016)*** (default parameters) and correlation plots were generated using lattice R package. Intersection analyses of

RNA-Seq and RIP-seq datasets were performed using UpSetR and SuperExactTest R packages ***(Gehlenborg, 2019; Wang et al., 2015).***

Network analysis and generation of the library of functional modules

Drosophila physical Protein-Protein Interaction (PPI) data reported at least in one experiment was retrieved from APID repository (http://apid.dep.usal.es; ***(Alonso-López et al., 2019)*** in December 2019. The original unspecific network was filtered to include only interactions between proteins expressed in adult fly brain tissue as described in previous section. The neuronal network was then simplified to remove self-loops and isolated proteins using the igraph R package ***(Csárdi and Nepusz, 2006).*** Bioconductor GOfuncR R package was used to evaluate the functional enrichment of brain network as compared to the unspecific network - Gene Ontology Biological Process, hyper-geometric test, FDR = 0.1 on 1000 randomizations- ***(Grote, 2020)).*** Finally, the functional modules were defined by selecting groups of physically interacting proteins annotated under the same enriched term. It should be noted that not all the proteins collaborating in the same process must physically interact (e.g., as in the case of cell signaling, the membrane receptor does not bind to its downstream transcription factor). Based on this, we enabled modules to be formed by non-connected subnetworks. The isolated clusters were discarded only when the largest subnetwork represented more than 90% of the total module. The same protein might be annotated with several terms and therefore might be involved in several modules simultaneously. Conversely, we are aware that the use of GO data may return functionally redundant modules. Prior any further analysis, module redundancy was evaluated to check that modules do not exceedingly overlap nor represent redundant biological processes. Based on this analysis, a module size from 10 up to 100 proteins was defined as optimal to minimize redundancy.

4.4 Results

4.4.1 Caz, Smn and TBPH proteins do not share common mRNA targets

We hypothesized that the existence of shared RNA targets for Caz, Smn and TBPH might underlie the observed phenotypic commonalities between SMA and ALS. To test this hypothesis, we performed RIP-seq to identify neuronal mRNAs present in the RNP complexes formed by each of these proteins in adult *Drosophila* neurons. To facilitate cross-comparisons and ensure reproducible and cell-type specific purification, we generated three independent transgenic lines with GFP-tagged constructs expressed under the control of UAS sequences inserted into the same chromosomal position. To specifically characterize the neuronal RNA interactome, GFP-fusion proteins were expressed in adult neuronal cells using the pan-neuronal elav-GAL4 driver. The ectopic expression of Caz, Smn and TBPH has been reported to induce toxicity *(Cragnaz et al., 2014; Grice and Liu, 2011; Xia et al., 2012)*. For this reason, we used the TARGET method *(McGuire et al., 2003)* to express GFP-fusion proteins specifically in adult neurons within a limited time window (5-7 days post-eclosion). The TARGET system relies on the temperature-sensitive GAL80 protein, which inhibits GAL4 at low temperature, enabling temporal regulation of UAS constructs. When expressed in neuronal cells, GFP-Caz and GFP-TBPH robustly accumulated in the soma, showing a predominant, although not exclusive nuclear accumulation (**Supplementary Figure S14.A and S4.1C).** As expected, GFP-Smn was found mainly in the cytoplasm, sometimes accumulating in foci **(Supplementary Figure S4.1B).** Despite the same insertion site and promotor sequence, GFP-Smn protein was consistently expressed at lower levels **(Supplementary Figure S4.1D).**

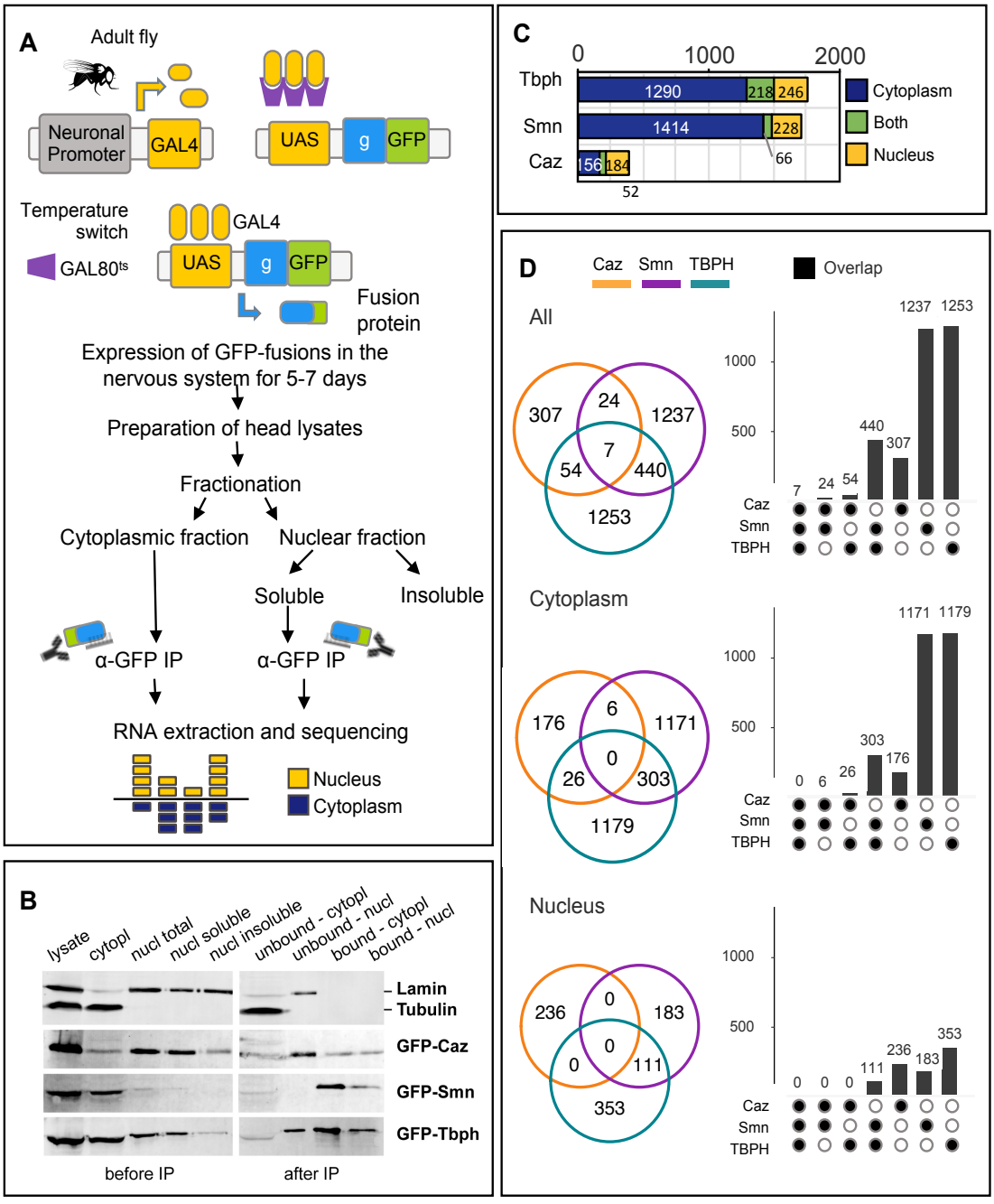


Figure 4.1 RIP-Seq identification of mRNA molecules in Caz, Smn and TBPH complexes in adult Drosophila neurons

(A) Schematic representation of the RIP-Seq procedure. GFP-fusion proteins were expressed conditionally in the adult nervous system via the Gal4/Gal80/UAS system. Lysates were prepared from heads and fractionated into cytoplasmic and nuclear fractions (see Methods). Nuclear proteins were further solubilized with high salt buffer and recovered in the nuclear soluble fraction. The cytoplasmic and nuclear soluble fractions were used for immuno-precipitation with GFP-trap beads. Co-immunoprecipitated mRNAs were extracted and sequenced. **(B)** Western Blots performed on the different fractions recovered along the RIP procedure. Lamin and Tubulin were used as markers of the nuclear and cytoplasmic fractions, respectively.

Note that Tubulin is depleted from the nuclear fraction while Lamin is depleted from the cytoplasmic one. Also note that GFP-Caz, Smn and TBPH differentially distribute in the nuclear and cytoplasmic fractions. **(C)** Bar graph showing the number of neuronal mRNAs co-immunoprecipitated with Caz, Smn or TBPH from cytoplasmic (blue) or nuclear (yellow) lysates. Transcripts present in complexes found in both compartments are shown in green. **(D)** Venn diagrams and corresponding bar graphs showing the overlap between the total (top panel), cytoplasmic (middle panel) or nuclear (lower panel) mRNA interactomes of Caz, Smn and TBPH. The seven transcripts found in the overlap of the top diagram ("All") correspond to mRNA molecules present in distinct nuclear and cytoplasmic complexes.

Since Caz, Smn and TBPH are multifunctional proteins involved in both nuclear and cytoplasmic regulatory functions, we separately characterized their RNA interactome in each cellular compartment. For this purpose, cellular fractionations were performed prior to independent anti-GFP immunoprecipitations, thus generating paired nuclear and cytoplasmic samples **(Figure 4.1A).** As shown in **Figure 4.1B,** relatively pure nuclear and cytoplasmic fractions were obtained from head lysates and GFP-tagged proteins could be efficiently immuno-precipitated from each fraction. For each paired nuclear and cytoplasmic pull-down, co-precipitated RNAs were extracted and used to prepare mRNA-seq libraries for single-end Illumina sequencing. Extracts from flies expressing GFP were used as control. Three independent replicate datasets were generated for each protein, except for GFP-Caz, for which one nuclear pull-down sample did not pass quality control for library generation. The raw sequencing dataset, composed of 23 libraries containing between 17.7 and 64.6 million total reads **(Supplementary Data S4.1),** was submitted to the European Nucleotide Archive (ENA) with the study accession code PRJB42798.

Following quality filtering, alignment to the *Drosophila* melanogaster reference genome and quantification of gene counts, RIP-seq datasets were analyzed to identify mRNA molecules enriched in GFP-fusion versus GFP control pull-downs. An average of 13,500 genes (>0 counts) were detected across all samples, ranging from 10,640 to 15,557 genes **(Supplementary Data S4.1).** As expected, the sequencing datasets clustered primarily depending on the nuclear versus cytoplasmic natures of the extract, and secondly depending on the protein used for pull-down **(Supplementary Figure S24.A).** DESeq2 (*Love et al., 2014)* was used to perform differential expression analysis (DEA) between each of the six pull-down sample groups and the control GFP pull-down. Transcripts displaying positive enrichment with an adjusted p value below 0.05 when compared to the control were considered as associating with the target protein **(Supplementary Data S4.2).**

Although Caz, Smn and TBPH fusion proteins were expressed specifically in neurons via the elav promotor, a certain degree of RNP complex re-association may occur in head lysates during the different experimental steps, as previously described *(Mili and Steitz, 2004).* To discard any non-neuronal transcripts that may have co-precipitated with target proteins, the dataset resulting from the DEA was filtered to include only genes with reported expression in the adult fly brain (see Methods), corresponding on average to 70% of the enriched transcripts (see **Supplementary Figure S4.3).**

These analyses revealed that Smn and TBPH associate with a large fraction of the neuronal transcriptome (1,708 and 1,754 mRNAs in total, respectively), and that most of their identified mRNA targets associate in the cytoplasm rather than in the nucleus **(Figure 4.1C).** A much smaller number (208) of mRNAs were found to associate with Caz in the cytoplasm, with 236 mRNAs detected as enriched in the pull-downs from nuclear fractions. Although this may partly reflect the higher heterogeneity of the Caz pull-down samples (**Supplementary Figure S4.2A),** it is in good agreement with the low abundance of GFP-Caz protein found in the cytoplasm compared to GFP-Smn and GFP-TBPH **(Figure 4.1B).** Of note, the percentage of transcripts simultaneously bound by the same protein on both compartments averaged only 22%, with TBPH displaying a much larger overlap than Smn for a similarly sized set of target mRNAs **(Figure 4.1C).** This observation is in agreement with the current model of mRNP complex remodeling between the nucleus and the cytoplasm, with the compartment-specific set of mRNA bound proteins being influenced both by their relative affinities and abundance **(Mili and Steitz, 2004).**

We next addressed the existence of common RNA targets, which could provide insights in a potential common MN degenerative mechanism in a context of Smn, Caz and TBPH deficiency in humans. Overlap analysis of the mRNA interactomes of Caz, Smn, and TBPH revealed a striking absence of transcripts bound by all three RBPs in the cytoplasmic or nuclear fractions **(Figure 4.1D).** This result does not exclusively result from the small number of RNAs bound by Caz, as a poor overlap was also observed between the large sets of cytoplasmic mRNAs bound by TBPH and Smn. Considering that the universe of protein-associated transcripts was defined exclusively based on the adjusted p value, without imposing a minimal enrichment threshold, this observation is particularly surprising. Together, our

RIP-seq experiments thus uncovered that Caz, Smn and TBPH do not share common RNA targets.

4.4.2 Gene expression changes in response to reduced levels of Caz, Smn and TBPH have significant commonalities but lack a clear functional signature

In addition to regulatory roles associated with mRNA binding activity, Caz, Smn, and TBPH have been shown to have both direct and indirect roles as transcriptional, translational and splicing *regulators (Fiesel et al., 2010; Morera et al., 2019)*. It is thus possible that, despite associating to non-overlapping sets of mRNAs, these proteins may coordinate common gene expression programs through other molecular mechanisms. To address this hypothesis, we used shRNA-expressing fly lines to knock-down the expression of caz, Smn, and TBPH in adult flies by RNA interference (RNAi) and characterized the resulting changes in neuronal gene expression using RNA-seq (**Figure 4.2A).** After identification of fly lines displaying a robust silencing of each target gene, we used the GeneSwitch (GS) system to induce ubiquitous, adult-onset RNAi. This system relies on the feeding of flies with the hormone mifepristone (RU486), which activates GAL4-progesterone-receptor fusions, thus driving transgene expression (**Figure 4.2A).** The system has been reported to display some leakage in the absence of the hormone *(Law et al., 2014)* what supports a linear modeling strategy for differential expression analysis. Exploratory analysis of the normalized RNA-seq dataset revealed that the samples clustered primarily according to genotype, followed by treatment (**Supplementary Figure S4.2B and S4.2C),** an observation consistent with the expected leakage from the siRNA locus (Scialo et al., 2016). Notwithstanding, principal component analysis revealed that hormone-treated samples exhibited a better separation than the corresponding untreated controls, as expected from shRNA-expressing samples (**Supplementary Figure S4.2C,** top left vs right). Of note, hormone treatment seemed to induce common changes across all sample types, explaining up to 7% of the variance in the dataset (**Supplementary Figure S4.2C**, bottom). Based on these observations, differential gene expression (DE) analysis was performed between

hormone treated Caz, Smn and TBPH shRNA-expressing target and control fly lines, using a linear model that considered hormone treatment as a batch effect (see **Supplementary Data S4.3)**. Confirming the robustness of our dataset and DE analysis, the specific shRNA target genes were found to be significantly down-regulated exclusively in the corresponding fly line (**Figure 4.2B).** The highest log2 FC and most significant adjusted p values were observed for caz, followed by Smn, and finally TBPH. Given that these three proteins are known to regulate mRNA processing, we also analyzed the data to identify alternative splicing (AS) changes that occurred as a consequence of the gene knock-down. For this purpose, we used the rMATS, a statistical framework to identify alternative splicing events in datasets of replicate samples. This tool supports the analysis of five major types of AS events (alternative 5' and 3' splice sites, exon skipping, intron retention and mutually exclusive exons) based on reads mapping to annotated exon junctions and neighboring exons **(Supplementary Data S4.4).** However, for the aim of the present study, all AS changes identified in each siRNA line were combined and transcripts defined as either alternatively spliced, or not affected. **Supplementary Data 5** provides the final annotated list of all neuronal genes detected in the different fly models and experiments.

Taking into consideration that RNA-seq was performed using samples isolated from fly heads, the list of transcripts showing significant DE or AS changes in response to caz, Smn or TBPH knock-down was filtered as previously described to exclude non-neuronal genes (**Supplementary Figure S4.3). Figure 4.2C** summarizes the overall results of the RNA-seq analysis. More than 2,200 genes and roughly 450 transcripts were found to be differentially expressed (DE) or alternatively spliced (AS) after caz silencing, respectively. In the case of TBPH silencing, RNA-seq analysis revealed about 1,600 DE and more than 250 AS genes. Silencing of Smn had the mildest detectable effect, with less than 1,400 DE genes and only 213 AS transcripts detected. These results are in agreement with the observed knock-down efficiency and sample heterogeneity **(Figure 4.2B and Supplementary Figure S4.2B),** suggesting that these differences more likely reflect our experimental set-up than a specific characteristic of the gene expression programs regulated by each protein. Of note, the proportion of up- and down-regulated genes within the DE gene set (~50%) was similar in all conditions **(Figure 4.2C).** Furthermore, only a relatively

small fraction of the deregulated transcripts in response to the RNAi was found to be bound by the corresponding protein **(Supplementary Data S4.5).** Caz-regulated transcripts showed minimal direct association with Caz protein (4.6%), whereas ~22% of the genes showing altered expression in response to Smn or TBPH RNAi were found to be enriched in the corresponding RIP-seq assays. Interestingly, this fraction goes up to ~40% when considering only the transcripts displaying alternative splicing changes in response to Smn or TBPH knock-down **(Supplementary Data S4.5).**

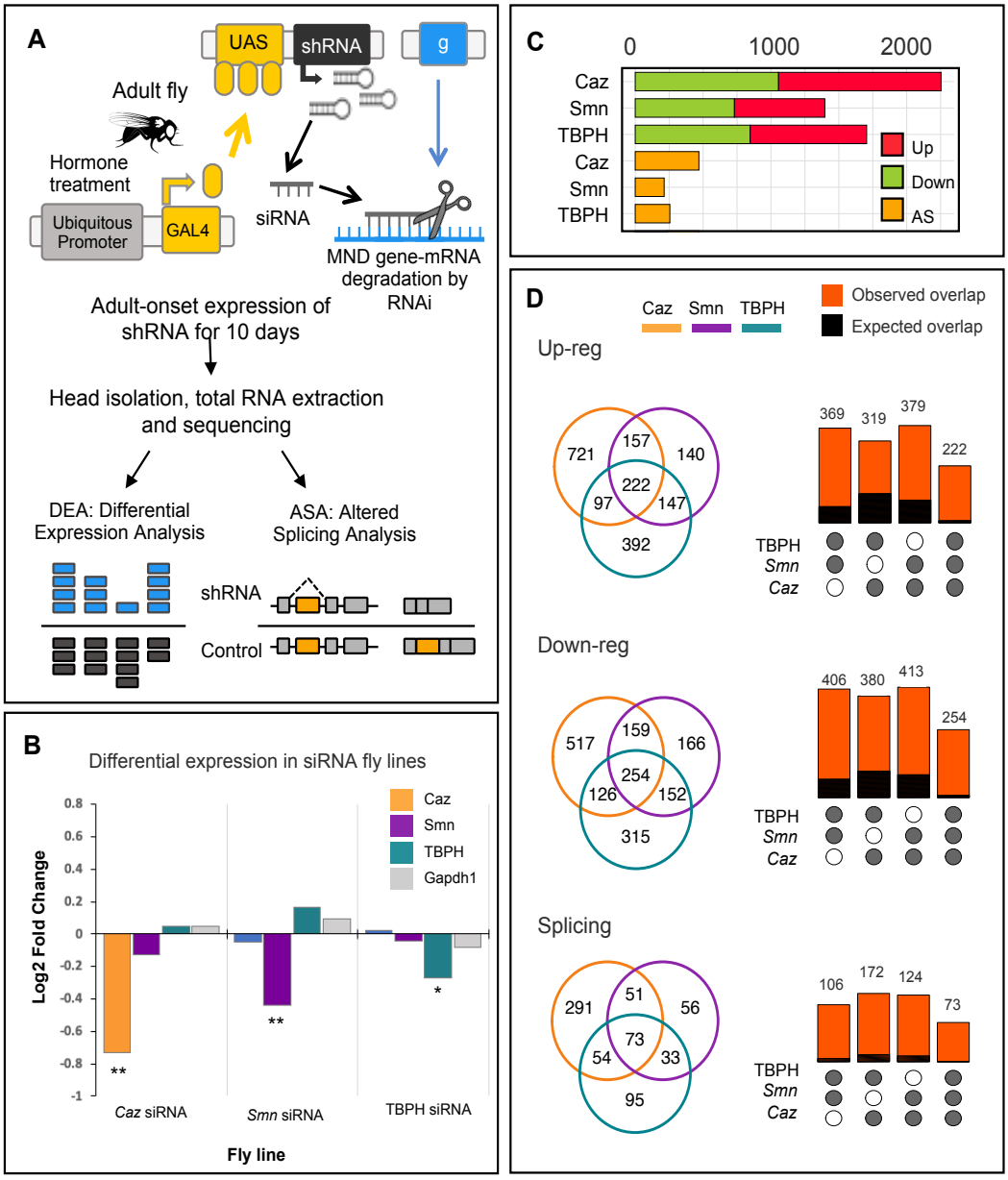


Figure 4.2 RNA-Seq identification of Caz, Smn and TBPH-dependent neuronal transcripts upon adult-induced RNAi knockdown

(A) Schematic representation of the experimental set-up. Hormone-dependent, adult-onset expression of short hairpin (sh) RNA was used to induce RNAi-mediated gene silencing of *caz*, *Smn* or *TBPH*. RNA was prepared from fly heads five to seven days post induction of shRNA expression, quality checked and subjected to mRNA-seq. Fly lines with shRNA against the *always early* (ae) embryonic gene served as control. **(B)** Bar graph showing the RNA-seq log2 fold change of the siRNA target genes plus the Gapdh housekeeping gene in each RNAi fly line. Statistically significant differences to the *ae* siRNA fly line are indicated as ** (adjusted p value < 0.05) and * (adjusted p value = 0.06 and p value < 0.02). **(C)** Bar graph showing the number of upregulated (red), downregulated (green) and differentially spliced mRNAs (yellow) in each RNA-seq dataset. Note that the kind of splicing change was not considered for this analysis. **(D)** Venn diagrams and corresponding bar graphs showing the overlap in upregulated (top), downregulated

(middle) or differentially spliced (bottom) mRNAs in flies with RNAi-mediated silencing of *caz*, *Smn* or *TBPH.* The bar color compares the expected (black) and observed (red) overlap given the total transcripts altered in response to the silencing of *caz*, *Smn* or *TBPH,* respectively. The expected ratios were calculated using *SuperExactTest* R package.

We next asked whether the transcriptome changes induced by the silencing of each target gene displayed any commonalities. A summary of the number of genes displaying common changes in expression as a consequence of the shRNA knockdowns, considering the type of effect (up-regulation, down-regulation, or alternative splicing), is depicted in **Figure 4.2D**. The overlap analysis of these gene sets reveals that a significant number of genes exhibits similar changes in response to all knockdowns, ranging from 16% of the genes identified as alternatively spliced in the caz shRNA fly line (73 out of 469), to 35% of the significantly downregulated genes in the Smn knock-down (254 out of 731) **(Figure 4.2D).** This is well above the overlap expected by random chance, with an estimated p value close to zero $P < 1e-16$, according to the hyper-geometric function for multi-set intersection analysis. Thus, despite the total lack of common RIP-seq targets, the down-regulation of Caz, Smn and TBPH protein expression elicited a partly coherent transcriptome response.

Next, we performed a functional enrichment analysis to identify biological processes linked to the commonly affected genes. Surprisingly, almost no Gene Ontology (GO) terms were enriched in the subset of ~500 common DE genes (**Supplementary Data S4.6)**. This result is in stark contrast with the strong functional signature that was observed for GO enrichment analysis of the subsets of mRNAs captured in the RIP-Seq assays. Of note, the DE/AS genes identified in the individual knockdowns of Caz, Smn or TBPH shared few common GO terms, suggesting the possibility of a synergistic effect on the same cellular pathways. To obtain insights into these potential connections, we proceeded to a more in-depth network-based analysis.

4.4.3 Network-based approaches identify commonly affected neuronal functional modules

Biological processes are dynamic and complex phenomena that emerge from the interaction of numerous proteins collaborating to carry out specialized tasks.

Thus, a biological process can be impacted to similar levels by changes in distinct proteins that contribute to the same regulatory function.

To understand whether the phenotypic commonalities observed in ALS and SMA might result from the deregulation of distinct, but functionally connected target proteins, we used a computational network-based approach. First, we generated a library of tissue-specific "functional modules" comprised of physically interacting and functionally collaborating neuronal proteins **(Figure 4.3A).** To do so, we began by reconstructing the entire *Drosophila* neuronal interaction network using protein-protein interaction (PPI) and adult fly brain RNA-seq datasets available in the APID and FlyAtlas2 repositories, ***respectively (Alonso-López et al., 2019; Leader et al., 2018).*** Notably, 45.5% of the 5 353 proteins found in this neuronal network are encoded by transcripts whose levels and/or splicing were altered in response to caz, Smn and/or TBPH knockdowns. Next, we defined functional modules in the neuronal network by selecting groups of physically interacting proteins annotated under the same enriched functional term. Of the 232 modules with associated GO terms, we focused on the subset of 122 modules composed of 10 to 100 proteins (**Supplementary Data S4.7).** These modules retained 1541 proteins in total, maintaining the high percentage of Caz, Smn and/or TBPH-dependent genes found in the original network (43.7%).

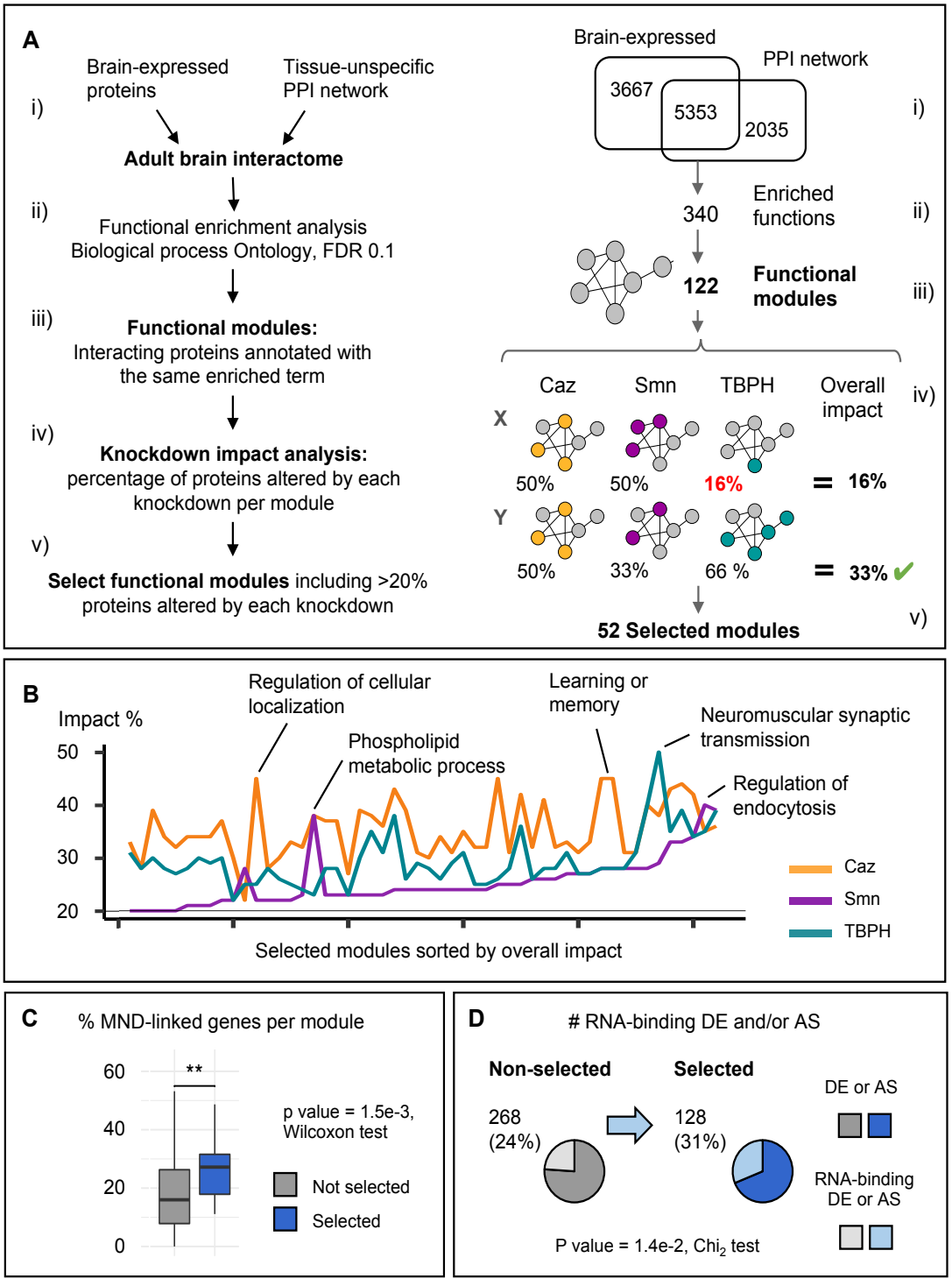


Figure 4.3 Characterization of functional modules impacted by Caz, Smn and TBPH knockdown

(A) Left panel: workflow used to generate and select functional modules. The adult brain interactome was obtained from the APID protein-interaction network after filtering for proteins expressed in the adult

Drosophila brain (**i,** see Methods). Functional enrichment analysis of the resulting interactome was performed to retrieve overrepresented GO Biological Processes **(ii)**. Note that the functional enrichment returns all the proteins annotated in each overrepresented term. The modules were generated from the functional enrichment by retaining the proteins annotated and simultaneously interacting in the brain network **(iii)**. Finally, the impact of caz, Smn and TBPH knockdown was evaluated for each module **(iv)** to select modules with > 20% of transcripts altered in each individual knockdown **(v)**. Right panel: summary of the workflow outputs. "Overall impact" calculation is exemplified for two modules (X/Y) with the impact score indicated on the right. Only the Y module would be selected, as the overall impact of module X is below the defined threshold. **(B)** Line plot comparing the impact of individual knockdowns on selected modules, sorted by increasing overall impact. Modules with the highest impact for each protein are indicated by their short name. **(C)** Box plots showing the percentage of proteins with MND-linked orthologs in each module class. Selected modules (blue) are significantly enriched in proteins with MND-linked orthologs compared to non-selected modules (grey) (p value = 1.5e-3, Wilcoxon test). **(D)** Pie charts representing the fraction of transcripts with altered expression (DE) or splicing (AS) in response to a given protein knockdown that are simultaneously found in RNP complexes bound by the same protein. The high percentage of DE/AS transcripts (selected modules, blue pie chart) is significantly related to a higher frequency of DE/AS transcripts involved in RBPs bound by the same proteins (p-value =1.4e-2, Chi2 test of independence).

To evaluate the impact of each of the three proteins on individual functional modules, we calculated the percentage of nodes belonging to the DE or AS categories. To focus on modules simultaneously affected by the downregulation of caz, Smn and TBPH, we assigned to each module an "overall impact" score, defined as the minimal percentage of transcripts showing altered expression in any given knockdown **(Figure 4.3A).** 52 modules with an overall impact score of ≥ 20% were identified. These modules were selected for further analysis, as they seem to be under the common control of all three proteins, although not necessarily through regulation of the same target genes.

Consistent with the potential functional relevance of the selected modules, associated functional terms were found to comprise a range of biological processes relevant in a MND context. These include general cellular processes such as kinase signal transduction pathways, regulation of the actin cytoskeleton, regulation of endocytosis, as well as neuron-specific processes such as learning and memory, and regulation of synapse assembly **(Supplementary Data S4.7).** Interestingly, differences in the impact of individual gene knockdowns were observed when comparing modules, which we propose to reflect some degree of functional specialization of the two ALS-related genes and the single SMA-associated gene **(Figure 4.3B).** For example, the module related to "learning and memory" functions was strongly impacted by caz down-regulation, but to a lower extent by Smn or TBPH silencing. In contrast, the module "neuromuscular synaptic transmission" was strongly impacted by TBPH, followed by caz, and less so by Smn knockdown. Finally,

some modules, like the one linked to "regulation of endocytosis" tended to be similarly impacted by all three knockdowns. Overall, the impact profiles of TBPH and Caz knockdowns on functional modules are much more similar to each other than to Smn, which generally displays lower impact scores, with a few exceptions including "regulation of endocytosis" **(Figure 4.3B).** This observation is quite striking considering that Caz and TBPH are associated to the same disease. To determine the relevance of the selected modules to the pathophysiology of MNDs, we calculated for each module the percentage of proteins with human orthologs already linked to MNDs (according to the DisGeNET repository *(Piñero et al., 2020)).*

Remarkably, a strong enrichment in the proportion of proteins with MND-linked human orthologs was observed for the selected modules when compared to those that did not pass the defined "overall impact" threshold (p value = 1.5e-3, Wilcoxon test) (**Figure 4.3C).** This result suggests that we were able to identify novel disease-relevant interactions based on the convergent analysis of Caz, Smn and TBPH-dependent functional modules in *Drosophila*.

As the selected modules represent core biological functions regulated by the three proteins, we looked at the prevalence of direct targets (i.e., mRNAs identified by RIP-seq) among the genes that encode proteins belonging to these modules and show DE and/or AS changes upon caz, Smn and/or TBPH knock-down. We observed that 31% of the 411 DE/AS transcripts associated to selected modules are also bound by at least one of the three MND proteins. This percentage decreases significantly to 24% of the 1119 DE/AS transcripts associated to non-selected (low impact) functional modules (p value = 1.4e-2, **Figure 4.3D),** being even lower in transcripts that are not part of any module (18% of 2280 transcripts, p value = 3.1e-8; **Supplementary Data S4.7).** Together, these results suggest that our integrated data analysis approach was able to identify key functional processes that are commonly and directly regulated by the three proteins. The results obtained point to a convergent functional impact that occurs through the regulation of distinct individual targets. The connection to the identified biological processes is mediated by functional protein networks enriched in molecules with already known links to MNDs.

Further exploration of the selected networks may thus provide relevant information to understand MND pathophysiology.

4.4.4 Convergent disruption of neuromuscular junction processes by altered Caz, TBPH or Smn protein levels

Pairwise comparison of the 52 selected modules revealed a high number of shared genes between many of them (see **Supplementary Figure S4.4**). To generate a non-redundant map of the common functional networks established by Caz, Smn and TBPH, we coalesced groups of highly interconnected modules into larger but more condensed "super-modules" (**Figure 4.4).** This resulted in seven super-modules named after their core functional association: signaling, traffic, cytoskeleton, stress, behavior, synaptic transmission, and neuro-muscular junction (NMJ) (**Supplementary Data S4.7).**

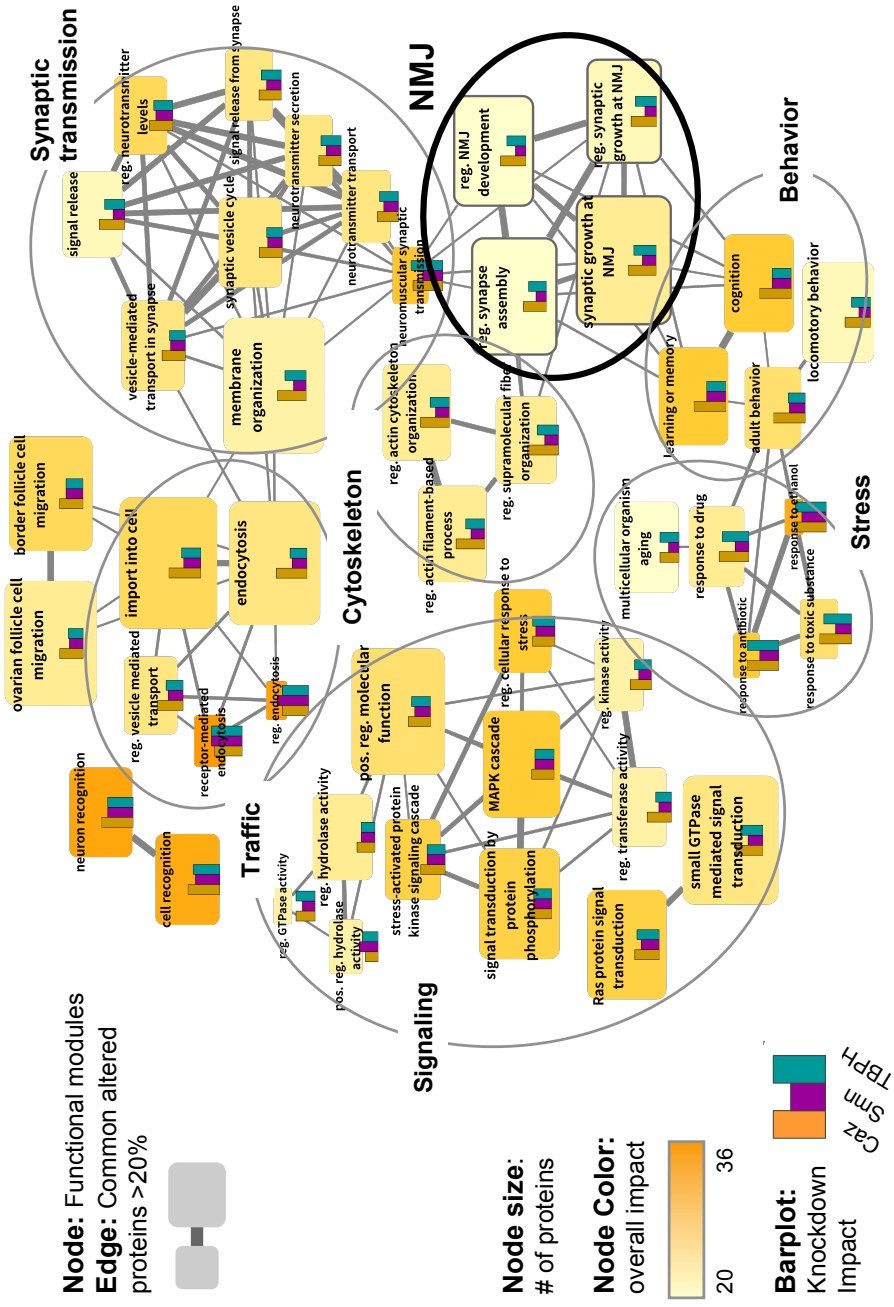


Figure 4.4 Identification of functional super-modules through protein overlap analysis

Network representation of the selected functional modules. Nodes represent the selected modules designated by the original name of the gene ontology term. Node size indicates the number of proteins incorporated in the module and gradient color the overall impact, i.e., minimum % of transcripts altered by each knockdown. The bar plots within the nodes indicate the impact of each knockdown on the module. Edge width indicates the number of commonly altered transcripts between two modules. Module overlaps (edges) below 20% were discarded which led six unconnected modules (not represented in the network). Modules were manually grouped into 7 "super-modules" (circles) based on edge density (common altered transcripts) and functional similarity of module names.

These super-modules range in size from 77 to 259 nodes, with a maximum and minimum overlap between any two super-modules of 87 and 1 out of 673 nodes in total, respectively (**Supplementary Figure 4.5).** We next determined the presence of MND-associated gene orthologues in the different super-modules (MND-linked, **Figure 4.5**, left panel). We further mapped the distribution of DE transcripts that are direct targets of Caz, Smn and TBPH (RNA-binding, **Figure 4.5**, middle panel); and of transcripts showing altered splicing (Altered Splicing, **Figure 4.5**, right panel). This analysis revealed a distinctive distribution of these characteristics in the groups of modules that were coalesced into super-modules, which is particularly evident regarding the percentage of transcripts displaying altered splicing or with potential roles in MND. In particular, the super-modules related to behavior, neuro-muscular junction (NMJ) and cytoskeleton incorporated the largest fraction of MND-linked and AS transcripts. Given the critical link between MNDs and the physiology of NMJs, we focused on the NMJ super-module for a more in-depth analysis.

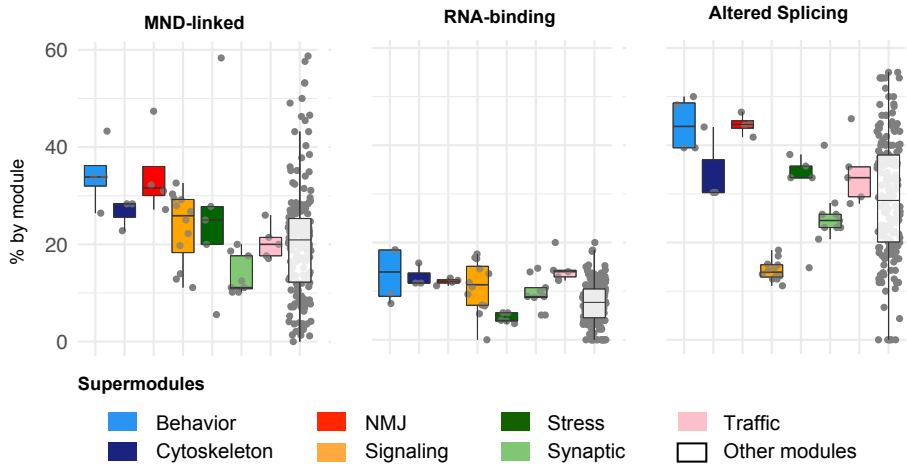


Figure 4.5 Analysis of super-module features

Box plots showing the distribution of the percentage of annotated proteins in the different super-modules (colored boxes) and across all other (non-selected) library modules (grey boxes). Grey dots represent the percentage in each individual module that is part of the super-module group. (Left) Percentage of proteins encoded by MND-linked gene orthologs according to the DisGeNET repository. (Middle) Percentage of proteins encoded by DE transcripts that are direct RNA-binding targets of caz, Smn and TBPH. (Right) Percentage of proteins encoded by transcripts with altered splicing patterns.

The NMJ super-module comprises 104 proteins, of which 49% (51 nodes) are encoded by genes differentially expressed and/or displaying altered splicing in at least one knockdown condition (**Supplementary Data S4.8).** 38 of these genes establish direct interactions, forming the subnetwork represented in Figure 6A. To assess the degree to which the NMJ "super module" functionally interacts with Caz, Smn and TBPH in vivo, we cross-referenced it to genetic modifiers of *Drosophila* Smn, Caz or TBPH mutants identified in genome-wide screens for modulators of degenerative phenotypes using the Exelixis transposon collection ***(Chang et al., 2008; Kankel et al., 2020; Sen et al., 2013)***. Interestingly, 21 nodes (~20%) of the NMJ "super module" were identified as either suppressors or enhancers of these models of neurodegeneration (**Supplementary Data S4.8).** Given that the reported percentage of recovered modifiers in these screens ranged between 2% and 5%, this result highlights the biological relevance of the functional modules identified through our approach. Detailed analysis of the FlyBase annotations for the genes within the NMJ subnetwork represented in **Figure 4.6A** provides interesting insights into the potential mechanisms causing neuronal dysfunction in the context of MNDs.

First, essential genes are highly overrepresented in the module. While about 30% of *Drosophila* genes are expected to be essential for adult viability ***(Spradling et al., 1999)***, more than 75% of genes present in the NMJ super-module have a lethal phenotype (**Figure 4.6B).** Exceptions are CASK, liprin-γ, Nlg2, metro, dbo and nwk. For RhoGAP92B and Nrx-1, it is so far not entirely clear whether mutant alleles would cause lethality. We next asked whether the human orthologs of these genes are linked to neurological disorders. TBPH (TDP-43), unc-104 (KIF1A, B, C), Ank2 (Ank2), futsch (MAP1A/B), sgg (GSK3A/B), Src64B (FYN/SRC) and Nrx-1 (Nrx-1-3) have been implicated in MNDs (hexagonal nodes in network). In addition to these, a high number of genes have human orthologs linked to other neuronal dysfunctions or diseases. For example, human orthologs to fly genes CASK (CASK), Mnb (DYRK1A), Rac1 (RAC1), Dlg-1 (DLG1), Cdc42 (CDC42), Fmr1 (FMR1, FXR1/2), trio (TRIO), Nedd4 (NEDD4L/NEDD4) and CamKII (CAMK2A/B/D) have been linked to intellectual disability. Epilepsy has been associated with mutations in the human gene orthologs of cac (CACNA1A/B/E), alpha-Spec (SPTAN1) and slo (KCNMA1). In addition, human psychiatric diseases like schizophrenia or bipolar disorder can be caused by

alterations in genes with high similarity to Pak (PAK1/2/3) and dbo (KLHL20 indirect, via regulation of Pak, *(Wang et al., 2016).*

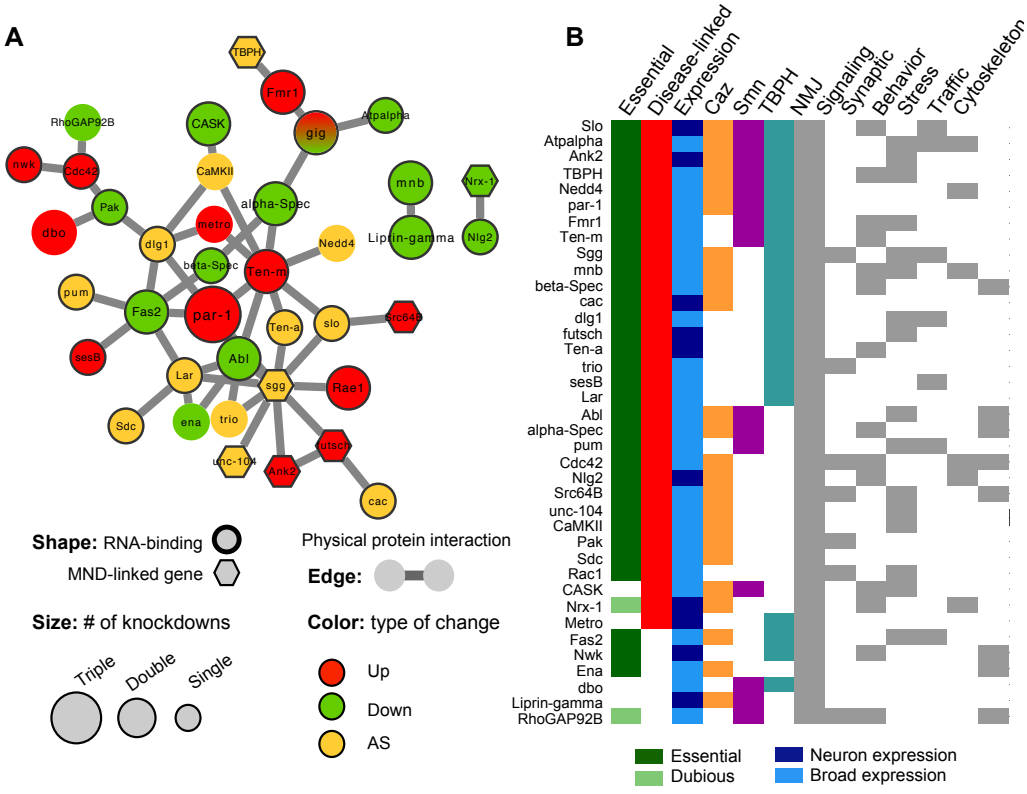


Figure 4.6 Detail of the core protein-interaction network of the neuromuscular junction (NMJ) super-module

(A) Protein-interaction subnetwork of NMJ super-module nodes that are encoded by transcripts altered in at least one knockdown model. Only proteins with direct interaction with other proteins encoded by DE/AS transcripts are represented. Node size indicates the number of knockdown models in which the transcript revealed altered expression (DE) and/or splicing (AS). Several transcripts are both DE and AS; yellow nodes indicate transcripts only showing AS. Bold outline highlights proteins encoded by transcripts found in Caz, Smn and TBPH RNP complexes. Hexagons highlight proteins whose human orthologs are MND-linked. (B) Categorical heat map summarizing FlyBase annotations for the proteins in the network represented in A. Essential proteins were defined according to the FlyBase repository. Proteins labeled as "Dubious" display a lethal phenotype after induction of RNAi. Thus, it is likely that flies homozygous for amorphic mutations would result in lethality during development. However, since this might result from off-target effects, they were not considered essential. MND-associations were retrieved from the DisGeNET repository. Caz, Smn and TBPH columns indicate in which knockdown models the corresponding transcripts were found altered. Last 7 columns indicate whether the protein is also found in other super-modules.

Alterations in the human gene coding Teneurin Transmembrane Protein 4 (TENM4, shares high homology with fly Ten-a and Ten-m) are known to cause hereditary essential tremor-5, while human neuroligins NLGN1, NLGN3 and NLGN4X were linked to autism/Asperger syndrome and encode orthologs to fly Nlg2. Finally, alterations in human orthologs to fly Pum (PUM1/2), beta-Spec (SPTBN1/2) and Ank2 (ANK1/2/3) have been associated with Ataxia-like phenotypes and mental retardation. In total, we were able to find direct associations to human MN or neurological disorders for 32 out of the 38 represented genes. Thus, although most of the genes captured in our analysis are not exclusively expressed in neurons, their mutations are somehow associated to abnormal neuroanatomy and function. Interestingly, this holds true for the non-essential genes as well. It is also noteworthy that, in spite of the relatively limited overlap between the different super-modules, all the proteins that constitute this core NMJ network are common to at least another super-module, and on average to more than half of them (**Figure 4.6B).**

Altogether, these observations imply that the proteins encoded by the NMJ super-module genes fulfill relevant functions in NMJ maintenance and that their alteration could eventually contribute to MND. Our results reveal that Caz, Smn and TBPH act in concert to regulate biological processes linked to NMJ maturation and function by altering the expression of transcripts encoding distinct, yet physically and functionally interacting proteins. We propose that the functional complexes established by these proteins may represent important players in disease progression, emerging as potential common therapeutic targets rather than the individual proteins that compose them.

4.5 Discussion

SMA and ALS are the most common MNDs and are characterized by a progressive degeneration of motor neurons and loss of skeletal muscle innervation. Although both diseases share many pathological features, including selective motor neuron vulnerability, altered neuronal excitability, as well as pre- and post-synaptic NMJ defects *(Bowerman et al., 2018),* their very different genetic origins and onset led them to be classified as independent, non-related diseases. This view has been challenged by recent studies demonstrating that disease-causing proteins (Smn for SMA, Fus and TDP-43 for ALS) are connected through both molecular and genetic interactions (reviewed by *(Gama-Carvalho et al., 2017).*

Furthermore, the increasing number of functions attributed to these proteins converges onto common regulatory processes, among which control of transcription and splicing in the nucleus, as well as mRNA stability and subcellular localization in the cytoplasm. Despite the observed convergence in the molecular function of Smn, Fus and TDP-43, transcripts co-regulated by these three proteins, and thus central to SMA and ALS pathophysiology, have not been identified by previous transcriptomic analyses. In this study, we used the power of *Drosophila* to systematically identify, on one hand the mRNA repertoires bound by each protein in the nucleus and cytoplasm of adult neurons and, on the other hand, the mRNA populations undergoing significant alterations in steady-state levels or splicing as a consequence of the knockdown of each protein. This approach revealed a striking absence of mRNAs commonly bound by the three proteins and a small, albeit significant, number of commonly altered transcripts. Notwithstanding, and contrary to the simplest model that explains shared disease phenotypes, this subset of shared transcripts did not present any functional signature linking it to biological pathways related to disease progression.

Considering that functional protein complexes are at the core of all critical cellular mechanisms, an alternative model posits that shared phenotypes may arise through convergent effects on independent elements of such complexes. To investigate this possibility, we mapped the de-regulated transcripts identified in our transcriptomic analysis onto a comprehensive and non-biased library of neuronal physically interacting and functionally collaborating protein consortia. This library was

generated by integrating publicly available information from *Drosophila* PPI networks, neuronal gene expression and gene ontology annotations. This approach led to the identification of a set of 52 functional modules significantly impacted by all three proteins through the regulation of distinct components (**Figure 3**). Of note, although we used as selection criterium the presence of a minimum of 20% of module elements displaying altered gene expression in each knock-down model, we found that modules passing this cut-off were significantly enriched in direct RNA binding targets of Smn, Caz and TBPH compared to non-selected modules (**Figure 4.3D**). Considering that only a very small proportion of these targets are common to the three proteins, this observation underscores our hypothesis of convergent regulation of functional complexes through distinct individual elements. Furthermore, the enrichment of RIP targets in the selected modules establishes a direct mechanistic link between changes in the levels of Smn, Caz and TBPH and changes in the steady state expression of module components. It is possible that the steady-state levels of transcripts encoding other proteins that belong to the same complex will vary as part of homeostatic feed-back processes. This could justify the presence of a relatively large number of DE/AS genes that are common to the three knockdown models, but whose transcripts are not found as direct protein targets in our RIP-seq data.

The functional classification of the 52 selected modules revealed a striking connection with critical pathways for MND. Particularly relevant, mapping of the human orthologues of the different module components revealed a high number of genes with reported association to MNDs. This observation provides support to the relevance of our approach, which uses *Drosophila* as a model for uncovering molecular interactions underlying human disease. It is noteworthy that the enrichment in disease-associated orthologues was not homogeneous across the super-modules generated by coalescing highly related modules into a smaller number of larger functional protein consortia **(Figure 4.5).** Interestingly, we found that a super-module related to NMJ function was among the highest scoring regarding both enrichment in MND associated genes and presence of alternatively spliced/direct RNA binding targets. The subset of DE/AS genes present in this module forms a highly interconnected network and the analysis of FlyBase annotations for this focused subset provided interesting insights into potential mechanisms that may underlie neuronal disfunction. An unusually large number of DE/AS genes within the NMJ

super-module was found to correspond to essential genes, indispensable for the development of adult flies. Alterations in the abundance and/or function of these genes have been linked in several cases to a disturbance of nervous system function. This is reflected by an alteration in stress response and/or abnormal behavior in either embryos, larvae or adult flies. Strikingly, even the non-lethal genes captured in this super-module have been shown to impact nervous system development and cause abnormal neuroanatomy when mutated/silenced.

The essential function of most of the selected genes obviously prohibits the analysis of loss-of-function phenotypes in the adult organism. In neurons, classical forward and reverse genetics of essential genes is not possible and, according to the post mitotic nature of neurons, clonal analysis is impossible. This is the reason why there is little genetic data on gene products involved in neuronal maintenance. Conditional knockouts and spatiotemporal control of RNAi-mediated gene silencing (like the approach used here) is a way to overcome this limitation. We can only speculate whether a neuron specific, adult-onset knockdown of the individual genes within the super-module will impair adult neuron integrity. However, taking all the data together, it is reasonable to assume that the collective deregulation of this set of genes within the super-module is incompatible with proper neuronal function. This assumption is particularly sound if the encoded proteins and their associated functional complexes are found to contribute to cellular processes critical for neurons, as indeed we find in this case. In fact, for almost all proteins encoded by the NMJ sub-network, synaptic functions have been reported. Interestingly, the other identified super-modules are also functionally annotated to cellular mechanisms that are especially important in neurons, like signaling, cytoskeletal dynamics, traffic and transport. Thus, an attractive model emerges for SMA and ALS MN dysfunction that states that convergent functional impacts can emerge from the independent, subtle deregulation of a group of proteins that are part of a set connected, neuronal functional modules. A persisting impairment in critical neuronal processes could initiate a self-reinforcing cycle of detrimental events, eventually resulting in neuronal decline. Especially in the case of sporadic, late-onset ALS, this model would comply with the events observed in disease progression.

Conclusions

In conclusion, our work reveals common functional hubs that are under the control of the SMA and ALS disease-associated genes Smn, TBPH and Caz, through independent target genes and transcripts that encode proteins which collaborate in neuronal functional consortia. These common hubs are deregulated in pre-symptomatic disease models and are primarily composed of ubiquitously expressed genes, suggesting that they may serve as a starting point for the discovery of novel disease biomarkers. Furthermore, the identification of common molecular dysfunctions linked to distinct MNDs and disease-associated genes suggests that common therapeutic strategies to help slowdown disease progression or improve symptoms may be amenable in spite of different genetic backgrounds.

4.6 Supplementary Figures

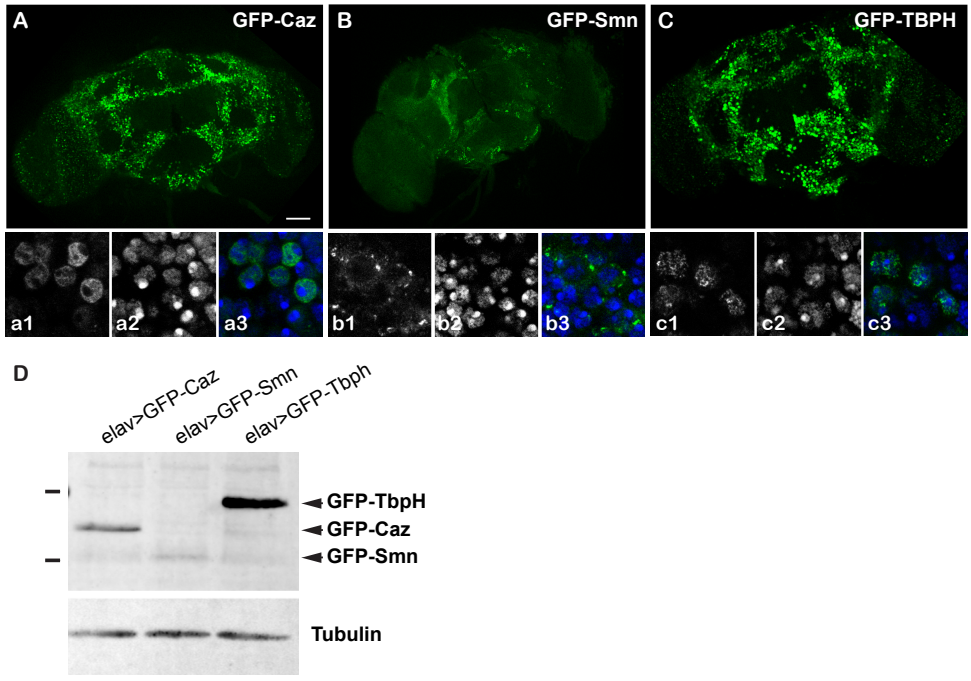


Figure S 4.1 Characterization of the UAS-GFP-Caz, UAS-GFP-Smn and UAS-GFP-TBPH fly models

(A-C) Adult brains dissected from elav>GFP-Caz **(A)**, elav>GFP-Smn **(B)** and elav>GFP-TBPH **(C)** flies 5-7 days after expression. The GFP signal is shown in green. Insets in a1-a3, b1-b3 and c1-c3 show the sub-cellular distribution of GFP-Caz, GFP-Smn and GFP-TBPH respectively. GFP signals are shown in white (left) or green (overlay, right). DAPI signals are shown in white (middle) or blue (overlay, right). Scale bars: 50 µm. Complete genotypes: elav-Gal4/Y; tub-Gal80ts/UAS-GFP-Caz **(A)**, elav-Gal4/Y; tub-Gal80ts/UAS-GFP-Smn **(B)** and elav-Gal4/Y; tub-Gal80ts/UAS-GFP-TBPH **(C)**. **(D)** Western blot performed on lysates from adult elav>GFP-Caz (left), elav>GFP-Smn (middle) and elav>GFP-TBPH (right) brains. Anti-GFP antibodies were used to detect GFP fusions. Tubulin was used as a loading control.

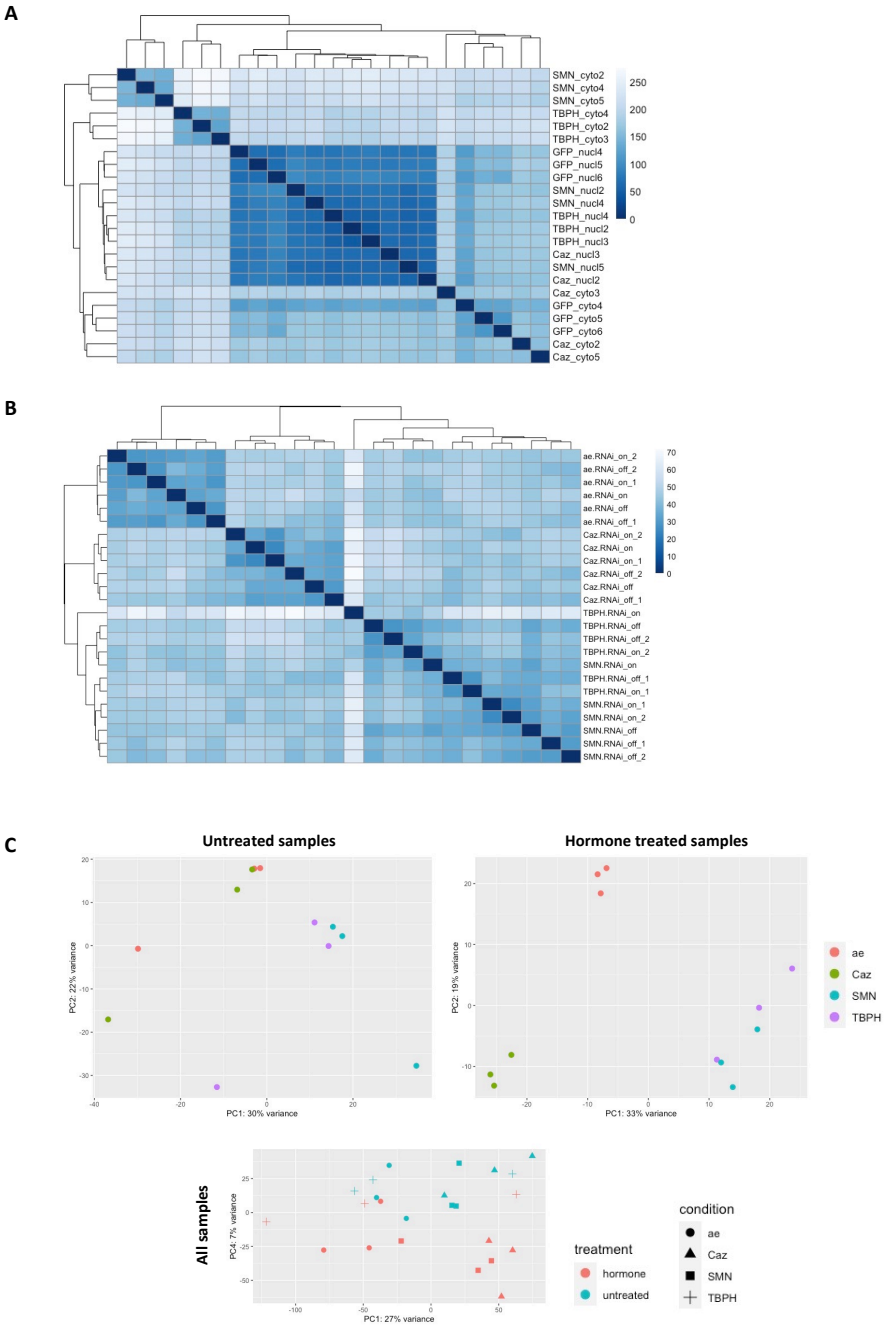


Figure S 4.2 Overview of RIP-seq and mRNA-seq data

(A) Sample-to-sample distance heatmap for the RIP-seq dataset revealing overall similarities and dissimilarities between dataset samples based on Euclidean distance. **(B)** Sample-to-sample distance heatmap for the mRNA-seq dataset revealing overall similarities and dissimilarities between dataset samples based on Euclidean distance. **(C)** Principal component analysis for mRNA-seq datasets. Top panels: analysis of dataset according to treatment status (left - untreated; right - hormone treated), samples colored by fly line. Bottom panel: full dataset, samples colored by treatment, symbols indicate fly line (condition).

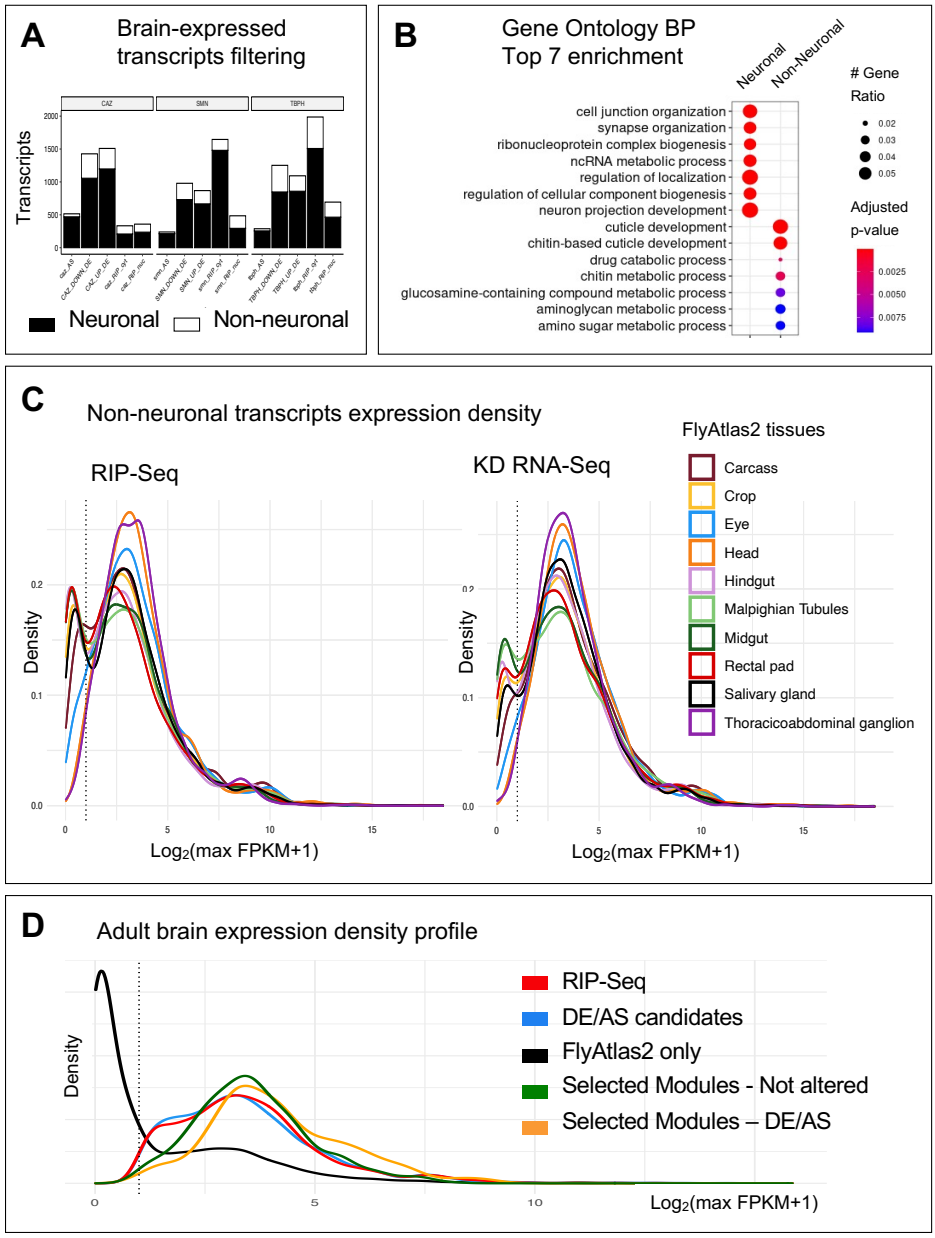


Figure S 4.3 Coverage of RIP-Seq and mRNA-seq experiments in FlyAtlas tissue-specific RNA-Seq profiles

(A) Normalized RNA-Seq data of adult fly brain tissue was retrieved from the FlyAtlas2 database (see methods). The total 9020 transcripts were filtered using an expression threshold of > 1 FPKM. From the total 7369 transcripts identified in the RIP-Seq and knockdown experiments, 5511 were also detected in this dataset, and will be referred to as "neuronal" transcripts hereafter. Bar graph shows the number of transcripts identified in each experiment. **(B)** clusterProfiler R package was used to compare the functional enrichment of the 5511 “neuronal” and 1858 “non-neuronal” transcripts identified in RIP-Seq and knockdown experiments using Gene Ontology Biological Process, hyper-geometric test, adjusted p-value 0.05. From 824 enriched terms in neuronal transcripts, 92 include at the description the following key words: "synap", "axon", "neuro", "dendrite", "nervous", "button", "glial" or "cortex". Non-neuronal

transcripts were enriched in 19 terms, none of them related to neuronal processes. Figure summarizes the top 7 functions enriched in each set. **(C)** 67.4% of the 1858 transcripts "non-neuronal" identified in the experiments were detected in 10 additional tissues available at FlyAtlas2 and displayed highest expression densities on head, thoracicoabdominal ganglion and eye tissues. Figure shows density plots of log2-transformed FPKM values. **(D)** Density plot of log2-transformed FPKM values of “neuronal” transcripts from the FlyAtlas2, RIP-Seq, DE/AS, and selected functional modules subsets, revealing an enrichment of our datasets in transcripts with medium to high expression levels in neurons, particularly for the transcripts with altered expression retained in the selected modules.

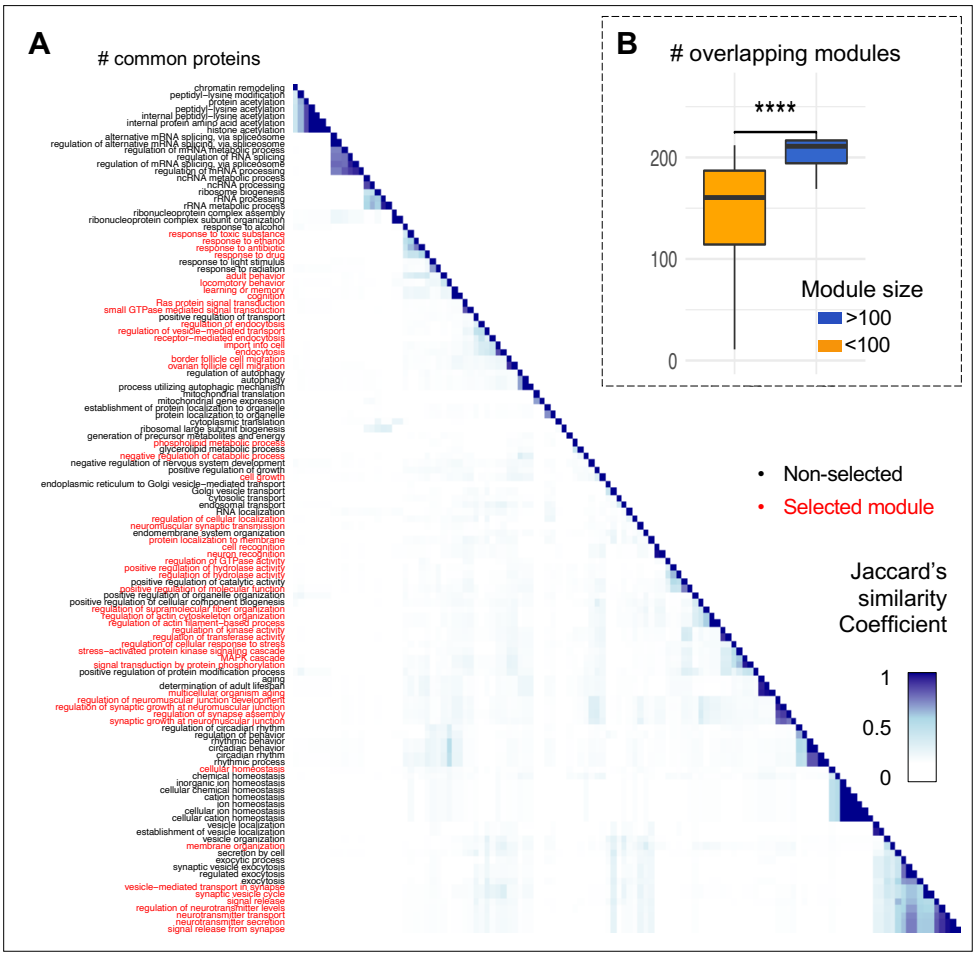


Figure S 4.4 Evaluation of protein redundancy across functional modules

(A) Complete-linkage hierarchical clustering using Jaccard’s similarity coefficient for the 122 modules with a size between 10 to 100 proteins. The 52 modules passing the overall impact cut-off of >20% of transcripts altered in at least one knockdown are labeled in red. **(B)** Box plots describing the number of modules sharing at least one protein when comparing modules including less or more than 100 proteins Wilcoxon test, p value 2.2×10-16.

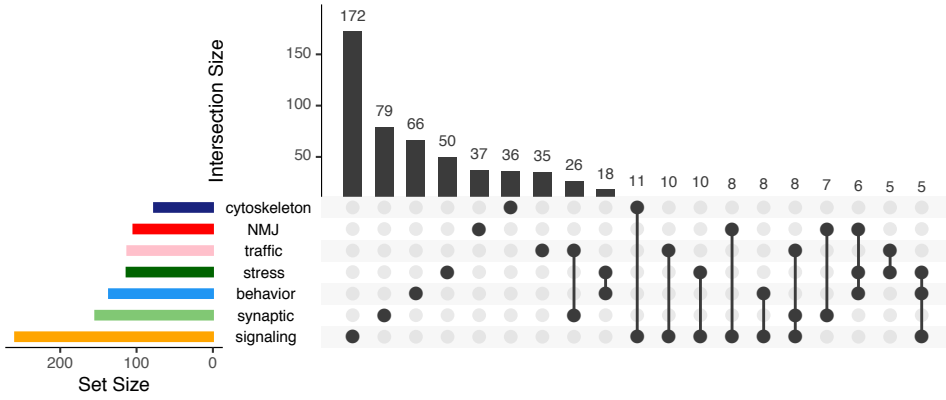


Figure S 4.5 Bar plot indicating the number of proteins found in common between different super-modules

Colored bars indicate total number of proteins in each super-module. Black bars indicate the overlap between super-modules. Only overlap sets including at least 5 proteins are shown.

4.7 Supplementary Data

The datasets generated and/or analyzed during the current study are available in the European Nucleotide Archive repository under the umbrella study FlySMALS, with accession numbers:

PRJEB42797 (https://www.ebi.ac.uk/ena/browser/view/PRJEB42797),

PRJEB42798 (https://www.ebi.ac.uk/ena/browser/view/PRJEB42798).

Supplementary data files are available in:

https://github.com/GamaPintoLab/MLG_PhDThesis_SupData

Supplementary Data S4.1 Sequencing library statistics

Supplementary Data S4.2 List of RIP-Seq enriched transcripts

Supplementary Data S4.3 RNA-Seq DE transcripts

Supplementary Data S4.4 Alternative splicing analysis results

Supplementary Data S4.5 Annotation of all FlyAtlas "neuronal" genes regarding the presence in the different DE/AS/RIP data subsets

Supplementary Data S4.6 FEA of target gene-dependent transcripts

Supplementary Data S4.7 Functional Module annotation

5 Integration of cross-species MND insights

5.1 Introduction

The pivotal goal of the research presented in this thesis was to investigate the molecular events common to the MND spectrum. With this aim, we developed Double-Specific Betweenness (S2B), a network-based method to identify bottleneck proteins connecting ALS and SMA disease modules in the human Protein-Protein Interaction (PPI) network (**Chapter 3**). In parallel, we characterized the deregulation in gene expression induced by the knockdown of Tbph (TARDBP), Caz (FUS) and SMN (SMN1) *Drosophila* orthologs of ALS and SMA-genes in fly models (**Chapter 4).** Initially in **Chapter 3,** the biological functions of S2B candidates were characterized by performing a straightforward functional enrichment analysis. On the other hand, the neuronal-specific role of the three disease associated genes in *Drosophila* models was inferred using the strategy presented in **Chapter 2,** based on mapping of Differentially Expressed genes (DEgs) onto fly brain BioInt units.

Chapters 3 and 4 returned valuable insights into the cellular processes potentially critical for MN survival. However, the two studies are founded on very different strategies. The S2B method was designed to prioritize candidates according to their network centrality, while the DG functional profiles were inferred from broad changes in the steady state transcriptome. Consequently, the two strategies are expected to reveal different facets of MND pathomechanisms. For simplicity, from here on, all the proteins selected by the S2B method or identified as being DG targets are commonly referred to as 'MND candidates'. In this final analysis, we aim to integrate the insights gained throughout the work presented in this thesis, to propose a unified understanding of the early events in MN degeneration.

Finding functional homology between proteins from different animal models to human is the first step to interpret the results retrieved from biomedical research. Unfortunately, orthology mapping between human and invertebrates such as *Drosophila* is highly complex due to the large number of paralog genes that diversified in human. The large imbalance in the total number of genes and the interaction rewiring between paralog proteins directly influences the network connectivity patterns, which in turn is decisive for the coordination of cellular functions ***(Shou et al., 2011).***

For this purpose, several approaches that incorporate information from network architecture have been proposed. One of the most straightforward strategies consists in comparing protein complex membership. However, it falls short to identify the conservation of interactions coordinating the protein complexes. On the other hand, global network alignment methods (recently reviewed in *(Ma and Liao, 2020))* notably increase the analysis dimensionality and, therefore, fell outside the scope of the chapter. As an intermediate solution, many researchers opt to compare topological network properties - namely node degree, betweenness or clustering coefficient - to identify overall connectivity patterns. However, the human PPI network is twice the size of its fly counterpart, so the connectivity comparison would require prior use of size normalization techniques *(Biran et al., 2019).* To our best knowledge though, these algorithms are not available in R environment yet.

On this basis, we opted to take advantage of BioInt units to address which functional assemblies accumulated more MND candidates. BioInt units are reconstructed from Tissue-specific (TS) networks, so they incorporate TS functional information. At the same time, GO is a universal catalog what allowed the direct comparison between the BioInt units generated across different specie-specific networks. Of note, the results presented below are discussed in a closing section that integrates the conclusions extracted throughput the thesis.

5.2 Methods

Tissue-specific networks Human and *Drosophila* TS networks were reconstructed using tissue-naive PPI data recovered from the APID repository on April 2021 ***(Alonso-López et al., 2019)*** and publicly available RNA-seq profiles derived from human brain samples ***(Uhlén et al., 2015)*** and from the FlyAtlas2 repository (Leader et al., 2018) (see methods in **Chapter 2 and 4**, respectively).

Identification of S2B candidates S2B candidates were established using the brain PPI network, and ALS and SMA disease-associated genes retrieved from the DisGeNET repository ***(Piñero et al., 2020)*** on May 2021. In the case of fly S2B candidates, the human disease genes were converted to their ortholog genes using the DIOPT tool ***(Hu et al., 2011)*** accessed on September 2020. S2B candidates were filtered using the default S2B score, threshold, and specificity cutoffs > 0.9, in 100 randomizations each, as described in **Chapter 3.**

Identification of differentially expressed genes (DEgs) The human ALS-Differential Gene Expression (DGE) profile was established from a publicly available RNA-seq dataset on spinal cord samples from control and ALS patients (GEO accession number GSE76220) ***(Barrett et al., 2013).*** The authors of the study collected adult spinal motor neurons (MNs) using laser capture microdissection (LCM) of lumbar spinal cord sections from 13 sporadic ALS patients and 9 controls. Total RNA was sequenced on Illumina GA II platform ***(Krach et al., 2018).*** The DGE analysis was performed in the GREIN web platform (***Mahi et al., 2019)*** that applies negative binomial generalized linear model as implemented in edgeR ***(Robinson et al., 2010).*** The ALS-DE profile was filtered using an absolute log2 fold change cutoff > 0.5 and adjusted p-value < 0.05. The *Drosophila* MND DEG counterparts were defined from the differential expression analysis performed in **Chapter 4**. In this case, independent RNA-seq profiles were generated for adult fly knockdown models of Tbph (TARDBP), Caz (FUS) and SMN (SMN1) genes. The analysis in **Chapter 5** includes all the transcripts found altered in at least one knockdown. **Table 1** summarizes the methods employed in MND candidate prioritization and the outcomes returned from the analyses.

Table 5.1 Overview of the methodology and data employed to generate the MND candidate gene sets

The S2B method requires as input data both ALS and SMA DGs and a brain PPI network. Fly S2B candidates are identified from a *Drosophila* brain PPI network and MN-DG orthologs. Differentially expressed genes were identified from RNA-Seq experiments using distinct statistic algorithms and thresholds. MND candidates were mapped to human and fly brain BioInt libraries. The total human orthologs mapped from *Drosophila* candidates are noted in parentheses. The cross-species analysis only considered the BioInt units significantly enriched in MND candidates from at least one set.

	S2B method							
	Disease genes (seeds)		Brain network		Candidate list	Candidates in brain network	Candidates in BioInt units	BioInt units enriched in candidates
	ALS	SMA	Nodes	Edges				
Human	189	32	12538	185659	264	264	176	40
Fly	78	14	5171	34994	108	108 (92 orthologs)	70	24
	Differential Expression analysis							
	Experimental set		Algorithm					
	Genetics	Sample	Method	Filter				
Human	ALS patients	Spinal Cord	GEO2R	adj.p < 0.05 FC > abs(0.5)	596	470	157	14
Fly	TBPH, Caz, Smn knockdown	Head	DESeq2	adj.p < 0.05	2382	1098 (1164 orthologs)	412	16

Cross-species MND candidate integration Human and *Drosophila* BioInt libraries were generated using the same criteria as indicated in **Chapter 2. Supplementary Data 5.1** collects the BioInt units reconstructed in *Drosophila* TS PPI networks. MND candidates were mapped on BioInt unit libraries. Human and/or fly BioInt units enriched in at least one MND candidate set - hyper-geometric test, p-value < 0.05 - and sharing > 0.6 Wang's semantic similarity were aggregated in functional groups.

5.3 Results

The cross-species analysis of MND molecular mechanisms integrates four candidate gene lists. The human S2B candidates and fly Differential Gene Expression (DGE) profiles were derived from the work presented in **Chapters 3 and 4,** respectively. To strengthen the cross-species data integration, we incorporated two complementary gene sets; fly S2B candidates and human ALS-DEGs (see Methods, **Figure 5.1A).** The candidates were mapped onto BioInt libraries (generated in **Chapter 2**) and the BioInt units enriched in MND candidates from at least one out of the four candidate gene sets were selected as potential MND pathways (**Figure 5.1B).** The BioInt libraries were generated from the same Gene Ontology (GO) catalog thus enabling the BioInt units of the two species to be analyzed together (**Figure 5.1C**). To identify higher-order common hallmarks in the BioInt units enriched in human and fly libraries, they were grouped into broader functional terms. By doing so, the functional groups can incorporate BioInt units from human or fly libraries, or both. Finally, we evaluated whether a functional group includes BioInt units enriched in various MND candidate sets.

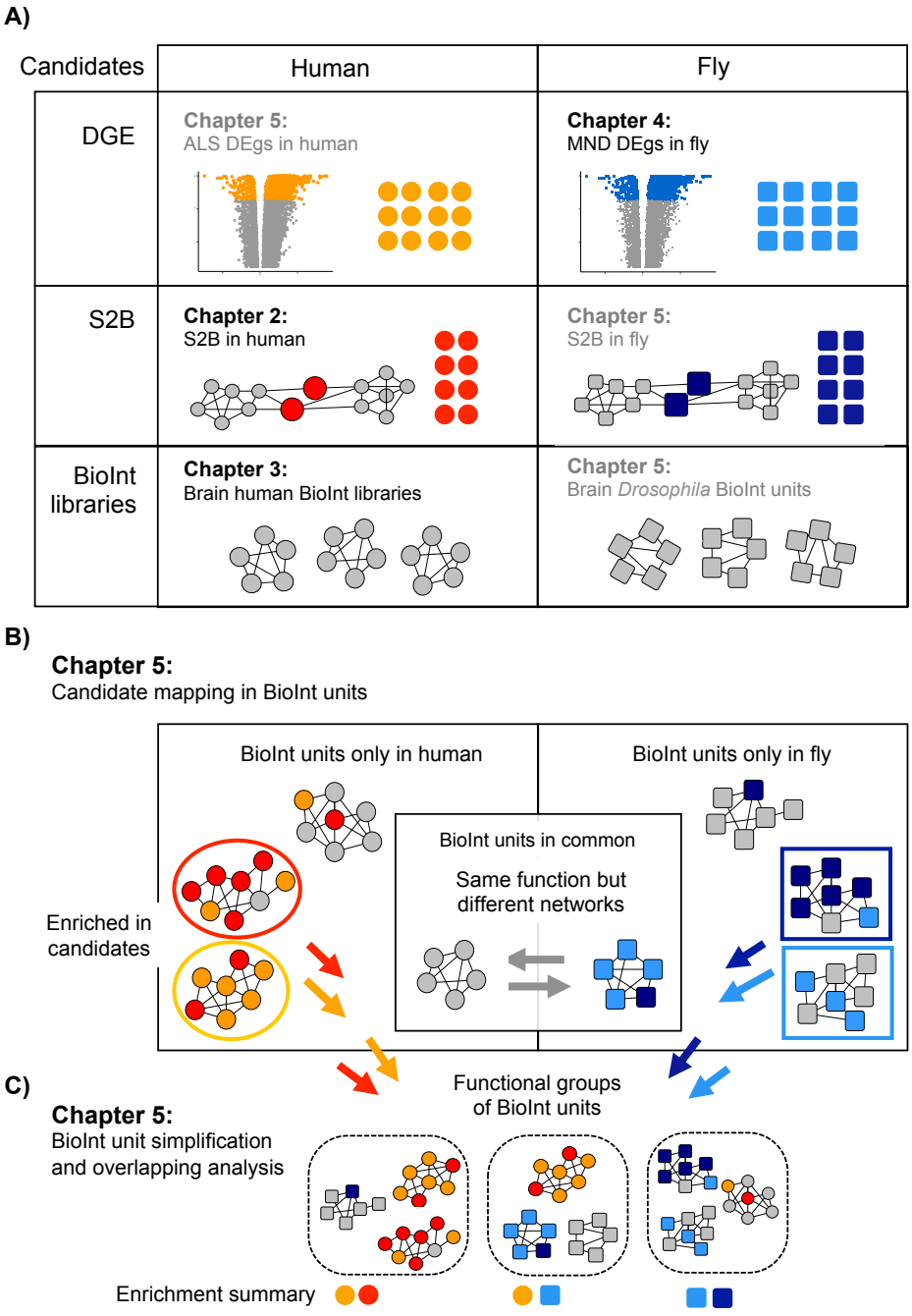


Figure 5.1 Diagram summarizing the thesis workflow

(A) Human and fly BioInt libraries, differential gene expression (DGE) and S2B candidate identification. The table indicates the chapters in which each type of candidate was retrieved. **(B, C)** Sketches to summarize the analysis workflow in Chapter 5. **(B)** The functional characterization of MND gene candidates was performed by mapping DEGs and S2B candidates into BioInt units. The BioInt units enriched in MND gene candidates (hyper-geometric test, p-value <0.05) were defined as functions likely affected by MND molecular alterations. **(C)** The cross-specie knowledge integration was performed by simplifying (Semantic similarity of Gene Ontology annotations) the candidate BioInt units into broader functional groups.

5.3.1 The differences between human and *Drosophila* BioInt libraries may reflect the species' functional complexity

Prior to the cross-species analysis, we assessed the global properties of the tissue-specific (TS) PPI networks and BioInt libraries from the two species **(Figure 5.2)**. First, we addressed the total percentage of human and fly orthologs identified in the TS gene profiles (**Figure 5.2A**). As expected, human dataset included larger fraction of species-specific genes. The TS gene profiles were then mapped onto PPI networks. For each TS-network, we evaluated the total number of edges or nodes and percentage of transcripts ubiquitously expressed (**Figure 5.2B,C**). Ortholog coverage in TS PPI networks indicates the % of *Drosophila* proteins with at least one known ortholog in human (red) and vice-versa (blue) (**Figure 5.2C**). Compared to overall ortholog coverage in **Figure 5.2A**, the human TS PPI networks lost a fraction of human-specific genes. The particular loss of human-specific proteins in TS PPI networks indicates that their interactions a poorly characterized.

Human networks were larger and incorporated higher fraction of non-ubiquitous proteins. This likely illustrates that human networks incorporate a large fraction of specialized and evolutionarily more recent proteins. When comparing networks from the same species, the human brain network was among the networks with the lowest % of ubiquitous proteins but the highest coverage of orthologs (black dot in **Figure 5.2C**). This observation could indicate that, regardless of human specialization, many brain functions are rooted in evolutionarily ancestral genes. On the other hand, human BioInt libraries included on average twice as many BioInt units as fly counterparts (**Figure 2D**). Human brain BioInt library was one of the smallest (in terms of number of BioInt units), while the fly brain library was one of the largest.

We also evaluated the distribution of ubiquitous genes along the BioInt units in the two species' libraries. The difference between human and fly libraries is similar to that of networks (**Figure 5.2C,D**). However, the median % of ubiquitous proteins in BioInt units is greater than in overall TS PPI networks, indicating that ubiquitous proteins collaborate in several units.

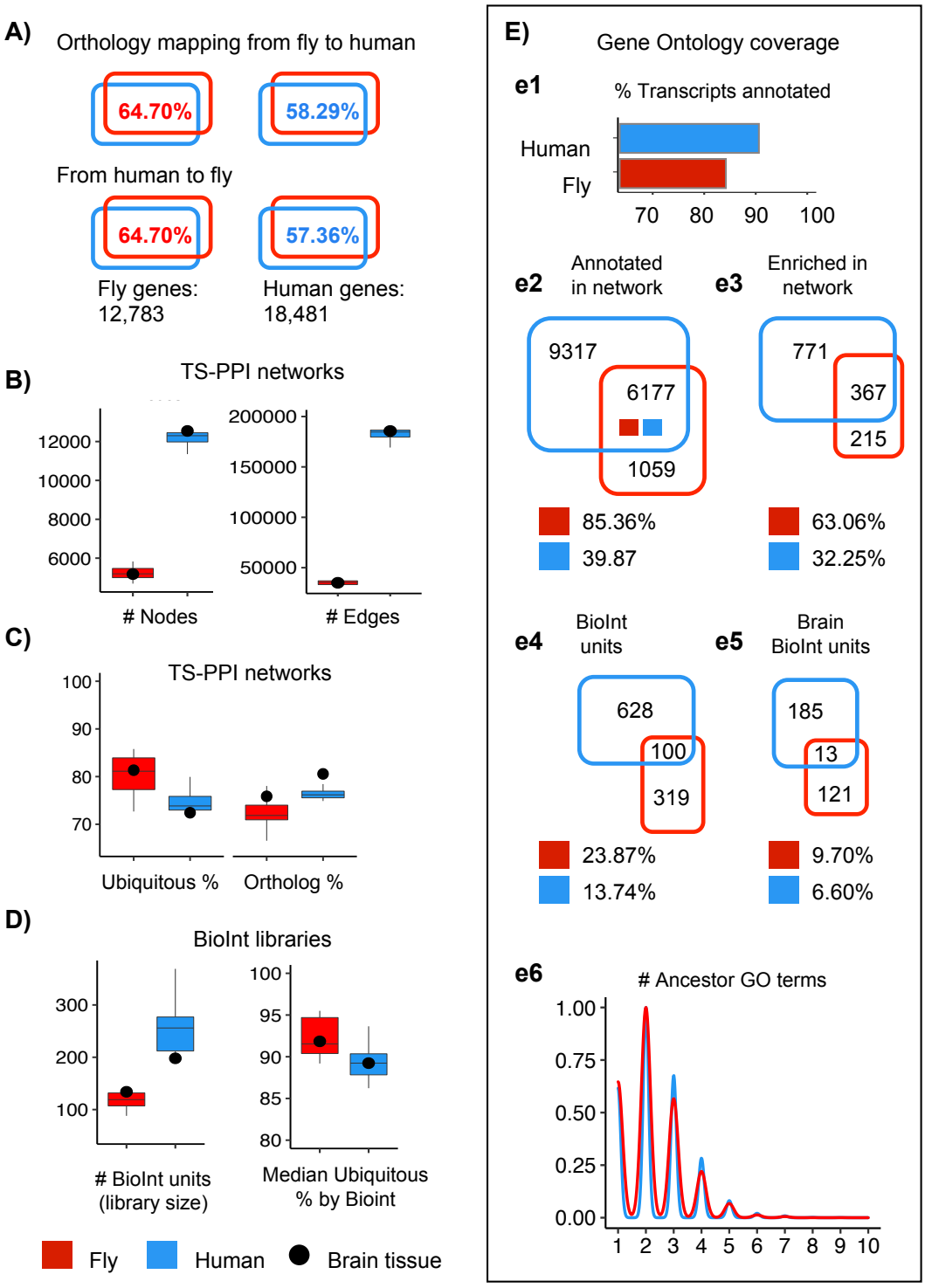


Figure 5.2 Comparison of the properties of tissue-specific (TS) protein-interaction (PPI) network and TS-BioInt libraries in fly and human models

(A) Venn diagrams summarizing the % of fly and human genes in the complete transcriptomes associated to orthologs in the counterpart species. **(B, C)** Boxplots comparing total number of nodes and edges (B), % of ubiquitous proteins and orthologs in the TS-networks reconstructed in human and fly models (C). Black dot indicates the specific value of brain network. Ortholog coverage indicates the % of *Drosophila* proteins with at least one known ortholog in human (red) and vice-versa (blue). **(D)** Number of BioInt units identified in each TS-network and median percentage of ubiquitous proteins by unit by TS-network. **(E)** Coverage of Gene Ontology (Go-BP) along the construction of BioInt libraries. **(e1)** Bar plots comparing the % of transcripts annotated with at least one Go-BP. **(e2-e4)** Venn diagrams representing the intersection of Go-BP annotated (e2), significantly enriched (e3) or defined as BioInt units (e4 - assemblies of PPIs enriched in Go-BP including up to 200 proteins) in all the TS-networks combined. (e5) Indicates the total intersection of BioInt units identified in human and fly brain networks. The colored boxes in d2-d5 indicate the % of common Go-BPs based on human or fly total set sizes. **(e6)** Density plot representing the number of ancestors associated with each Go-BP annotated in the TS-networks (corresponding to the sets in e2).

Next, we determined the coverage of Gene Ontology-Biological Process (GO-BP) annotations at each step of the workflow to reconstruct BioInt units **(Figure 5.2E).** We found that more than 84% of transcripts in the two species are functionally annotated (**Figure 5.2e1).** However, the total number of functional terms in the human network was twice the one found in fly (**Figure 5.2e2**). Despite the large difference in absolute numbers, we found that more than 85% of fly annotations were common to human. As we proceeded through the steps of the BioInt workflow - TS PPI reconstruction (**Figure 5.2e1**), GO-BP annotation (**Figure 5.2e2**), GO-BP enrichment (**Figure 5.2e3**) and selection of BioInt units including up to 200 proteins (**Figure 5.2e4**)-, we lost more than 98% of common annotations. Although it was a notable loss, in **Chapter 2** we reached to the conclusion that functional enrichment and size filtering was key to remove false positive and/or unspecific annotations. Finally, the brain libraries employed in this chapter included a similar number of BioInt units but a low % of common units between human and fly (**Figure 5.2e5**).

The consistent difference in total size between human and fly GO-BP and BioInt unit sets could either reflect an inequity in functional annotation efforts or indicate that human tissues have a richer catalog of functional options. As an estimation of functional annotation depth, we took advantage of the ontology hierarchy and compared the number of ancestors of the terms annotated in fly and human sets (**Figure 5.2d2** step). The density plot in **Figure 5.2e6** shows that the annotations of the two species displayed almost the same distribution in number of ancestors. Thus, the smaller number of annotations in *Drosophila* datasets is not only due to shallow ontology.

Tis observation, together with the smaller size of *Drosophila* PPI network and its lower annotation in GO-BPs reinforce the fact that the fly is a less complex organism than the human.

5.3.2 MND candidates display broad tissue expression patterns but accumulate in tissue-specific BioInt units

From the research presented in **Chapter 2** we concluded that DGs expressed in the same tissues can trigger distinct alterations depending on the TS-network contexts. To corroborate this observation, we re-evaluated the properties of the TS-BioInt libraries in both species. **Figure 5.3A** shows that - with the exception of a few tissues such as bone marrow in human, and ovary or testis in Drosophila - BioInt libraries displayed similar size distribution profiles. In accordance with the total number of MND candidate genes identified in S2B and DGE analyses (**Figure 5.3B**), Drosophila BioInt units accumulated a larger % of DEgs, overall. On the other hand, the fraction of MND candidates per BioInt unit displayed a similar distribution profile across all the tissues, including brain (**Figure 5.3B**). Thus, as previously observed in **Chapter 2**, the accumulation (total fraction) of MND protein candidates in BioInt units might not be the only determinant feature to induce the TS pathological phenotypes.

We next evaluated the distribution of MND gene/protein candidates in the TS PPI networks and their significant enrichment (hyper-geometric test, p-value >0.05) in BioInt units across the TS BioInt libraries (top and bottom pie charts in **Figure 5.3C,** respectively). TS RNA-seq datasets covered 27 and 11 non-sexual tissues in human and fly species, respectively. Fly datasets lacked from samples of heart, muscle and tracheal-respiratory tissues. However, the large difference in total tissue libraries is mostly due to the fact that *Drosophila* is organized in simpler organ systems. For instance, human dataset included 5 tissue samples involved in immune functions fly organism does not have to compare against.

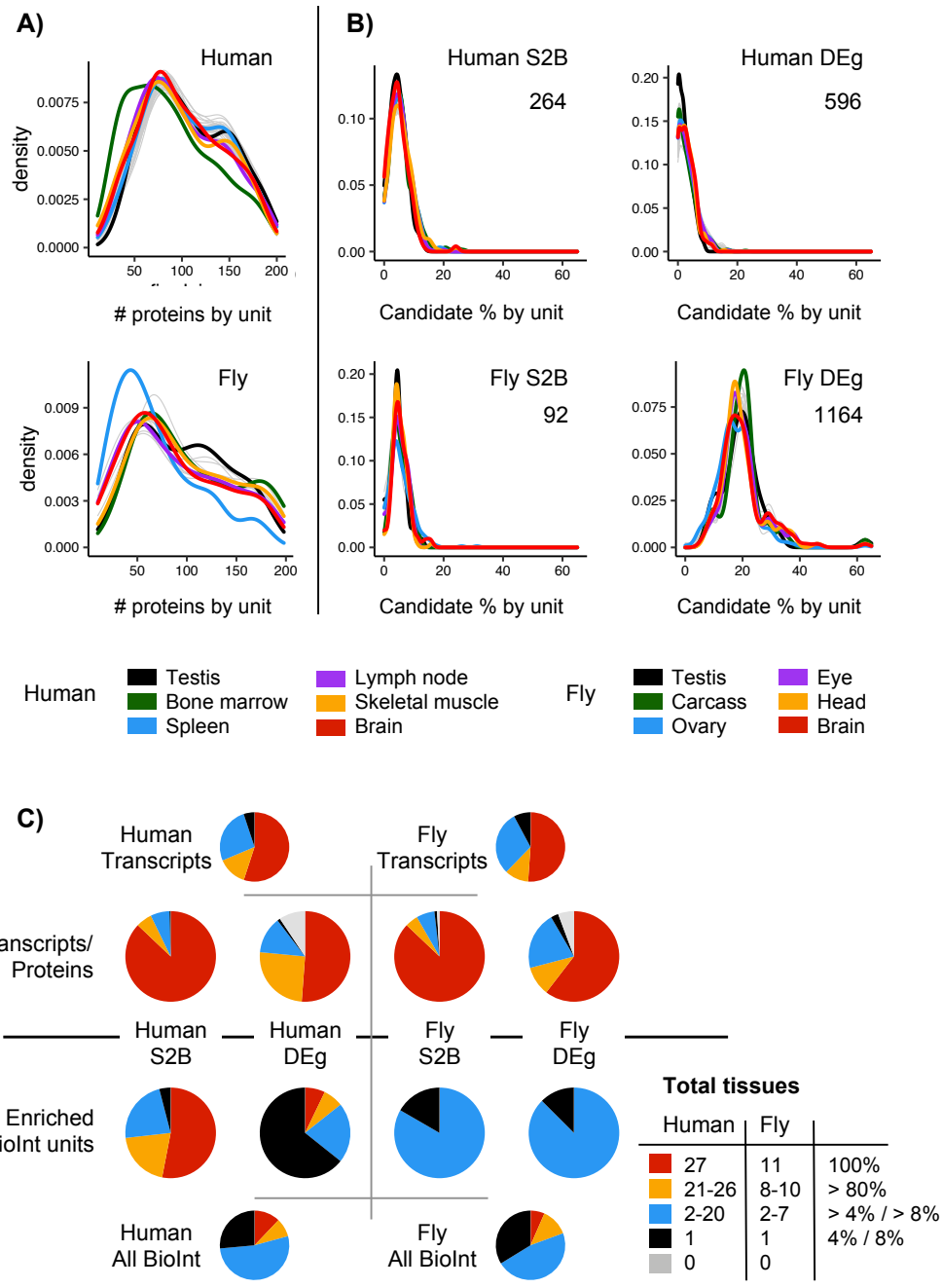


Figure 5.3 Comparison of the properties of tissue-specific (TS) BioInt libraries in fly and human models

(A) Density plots summarizing the number of proteins incorporated in each BioInt unit in each TS-library. **(B)** Density plot summarizing the % of MND candidates incorporated in each BioInt unit across the TS-libraries. Colored lines in A and B point to six illustrative tissues. **(C)** Pie charts summarizing the tissue expression patterns of transcripts (top half) and BioInt units enriched in MND candidates (bottom half) in human and fly universe (left and right panels, respectively). Color indicates the % of tissues in which the transcript or BioInt unit was identified -excluding sexual tissues.

Compared to the overall transcriptome profiles, we found that, with exception of human DEgs, the MND candidates have markedly broad expression, with most candidates being expressed in more than 80% of the tissues (first and second rows in **Figure 5.3C**). ALS-DEgs were identified in MNs from spinal cord samples. Therefore, the 9.4% of DEgs not mapped in brain transcriptome likely encode proteins restricted to MN interactome. By contrast, fly MND-DEgs were identified in head samples, which represents a much broader universe of transcripts. On the other hand, the distribution of BioInt units enriched in candidate genes drew a notably distinct pattern (third and fourth rows in **Figure 5.3C**). As expected from the results in **Chapter 2**, the overall distribution of BioInt units is more tissue specific than that of individual transcripts in the tissue transcriptomes. However, with exception of human S2B, it is noteworthy that the BioInt units enriched in MND candidates are notably more tissue specific. While the tissue distribution of BioInt units enriched in *Drosophila* candidates was very similar, the profiles of human BioInt units were very distinct. The BioInt units enriched in ALS-DEgs are the most tissue specific while the BioInt units enriched in human S2B candidates, the most transversal.

5.3.3 The simplification of BioInt units into functional groups reveals common functional hallmarks associated to human and fly MND candidates

Once we identified the BioInt units enriched in MND protein candidates (hyper-geometric test, p-value <0.05), we next sought to integrate the functional hallmarks derived from the two models. The overlap analysis of individual MND protein candidates showed relatively small intersection between the four sets. With the exception of the fly DEgs, each set included more than 82% of unique candidates (**Figure 5.4A).** As expected, due to the low intersection between human and fly brain libraries, there were no common BioInt units enriched in candidates from the two models. Likewise, the intersection of BioInt units enriched in DEg and S2B candidates was also very low (**Figure 5.4B**). To overcome the initial lack of commonalities, we took advantage of the semantic information of Gene Ontology hierarchies to group BioInt units in broader functional concepts. In this way, the functional groups can

incorporate BioInt units from fly, human or both organisms. The functional groups were generated from BioInt units enriched in at least one candidate set. Overlap analysis in **Figure 5.4C** indicates the number of functional groups including BioInt units enriched in varying candidate sets.

We next evaluated the hallmarks of the 38 functional groups incorporating BioInt units significantly enriched in at least one candidate set (hyper-geometric test, p-value >0.05, **Figure 5.4C,D**). We found six functional groups including BioInt units enriched in all the four MND candidate gene sets, which will be explored in more detail in next section. Aside the commonalities between the four sets, it was noteworthy that mitochondrion homeostasis or mRNA processing-related processes were significantly enriched in both human and fly S2B candidates. Furthermore, human S2B candidates were also enriched in functions related to DNA metabolism, nuclear import, or microtubule-cytoskeleton processes. These trends suggest that the S2B method captures functions involved in transversal activities of the cell.

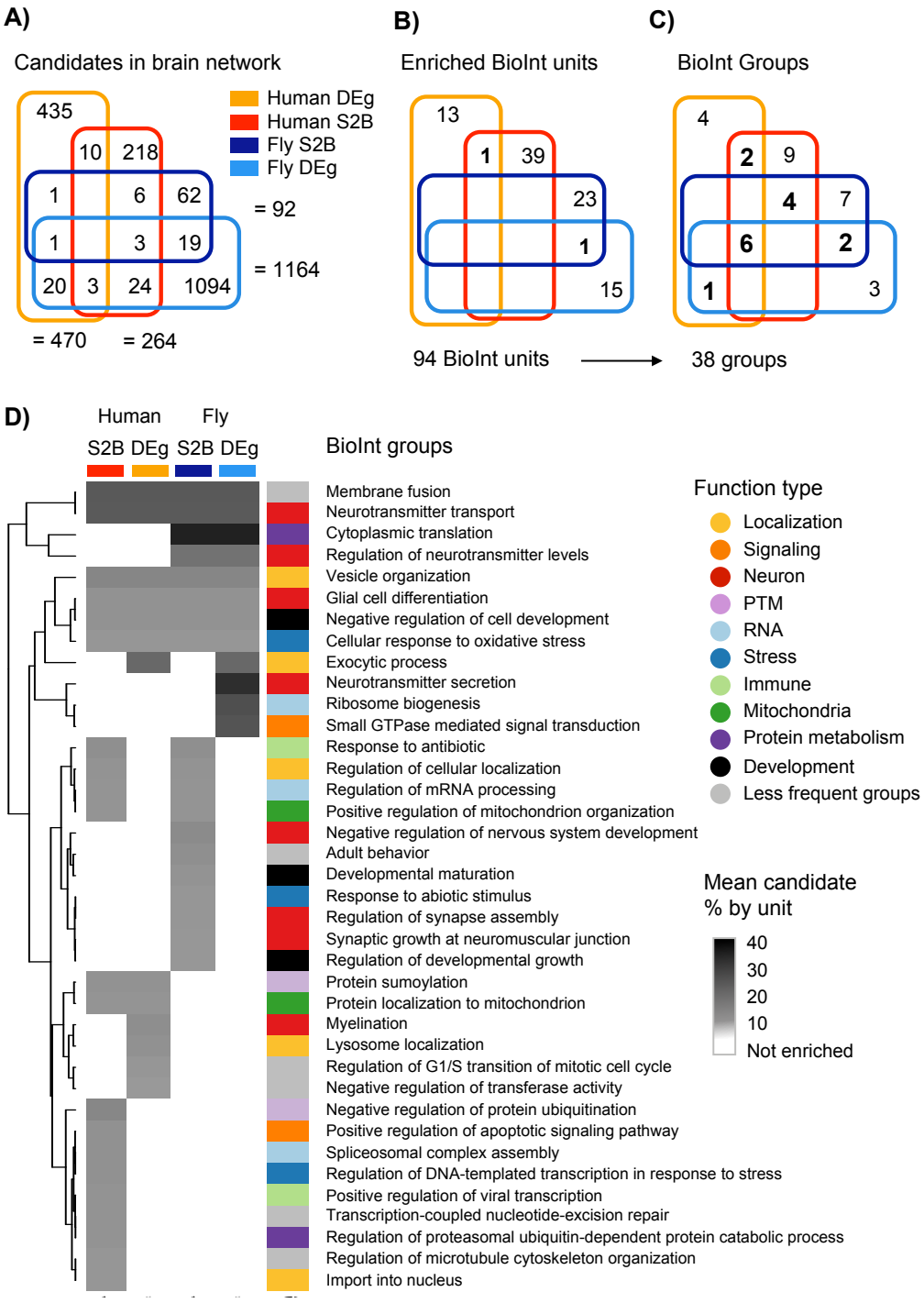


Figure 5.4 Overlap analyses of proteins involved in functions enriched in MND candidates from the four sets simultaneously

(A) Intersection analysis of all proteins (A) and MND candidates (B) involved in the respective BioInt units.

5.3.4 The integration of common MND functional groups in the core-PPI network reveals potential players linking fly and human MND-pathomechanisms

The six functional groups enriched in MND candidates from all four sets were related to neurotransmitter transport, vesicle organization, membrane fusion, glia cell differentiation, cell development and oxidative stress regulation. As introduced in **Chapter 1** and as discussed throughout this thesis, these functions have an evident relevance for the physiology of MNs. The intersection analysis in **Figure 5.5** indicated that functional groups could be simplified in three major processes, hereafter referred to as, vesicle transport, glia differentiation and oxidative stress.

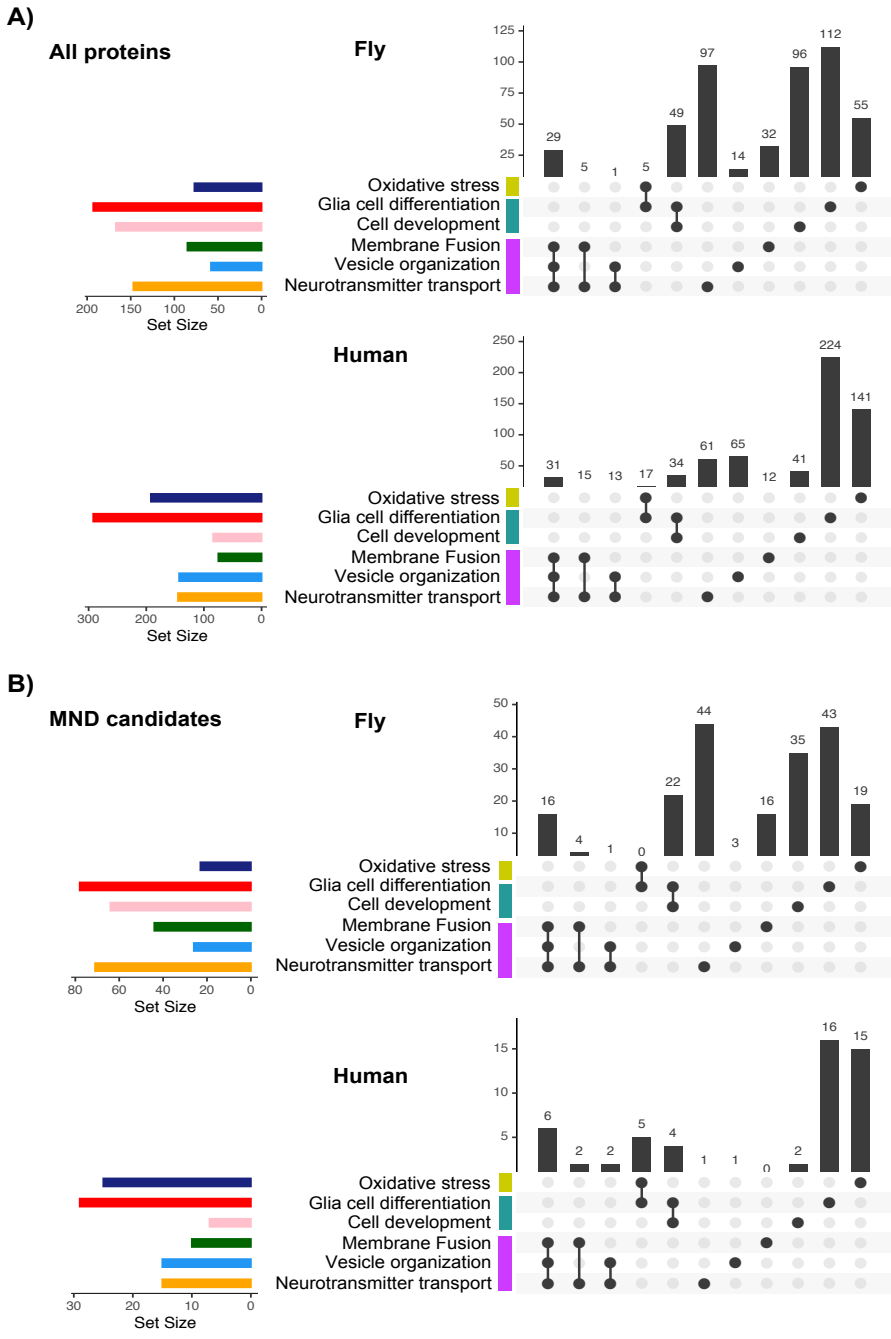


Figure 5.5 Core-PPI network linking human and *Drosophila* orthologs identified as MND candidates

Node shape indicates the specie while color indicates the type of MND candidate. Dashed edges represent ortholog relation between fly and human genes. Edge width indicates the number of interactions each ortholog presents in each function. Node size of proteins in functional groups indicates the number of interactions within the same function subnetwork.

The six functional groups included a total of 700 and 537 human and fly proteins, respectively. The goal of this Chapter was however, to emphasize the core molecular interactions linking the cross-species MND phenotypes. To this end, **Figure 5.6** only depicts the protein interactions between ortholog pairs simultaneously identified as MND candidates in the two species, especially in oxidative stress, vesicle transport and glia differentiation-related processes. Non-candidate proteins were excluded from the network and the interaction of core orthologs with the proteins in the functional groups was simplified into single edges. Likewise the interactions between proteins associated to the same function were represented by the size of the node. DEgs were much more abundant than S2B candidates, especially in fly functions. Thus, DEgs with less than 5 interactions - 12 in the case of fly vesicle transport and glia differentiation - were excluded from the figure. The full list of 94 BioInt units enriched in human and fly MND gene candidates is available in **Supplementary Data 5.2.**

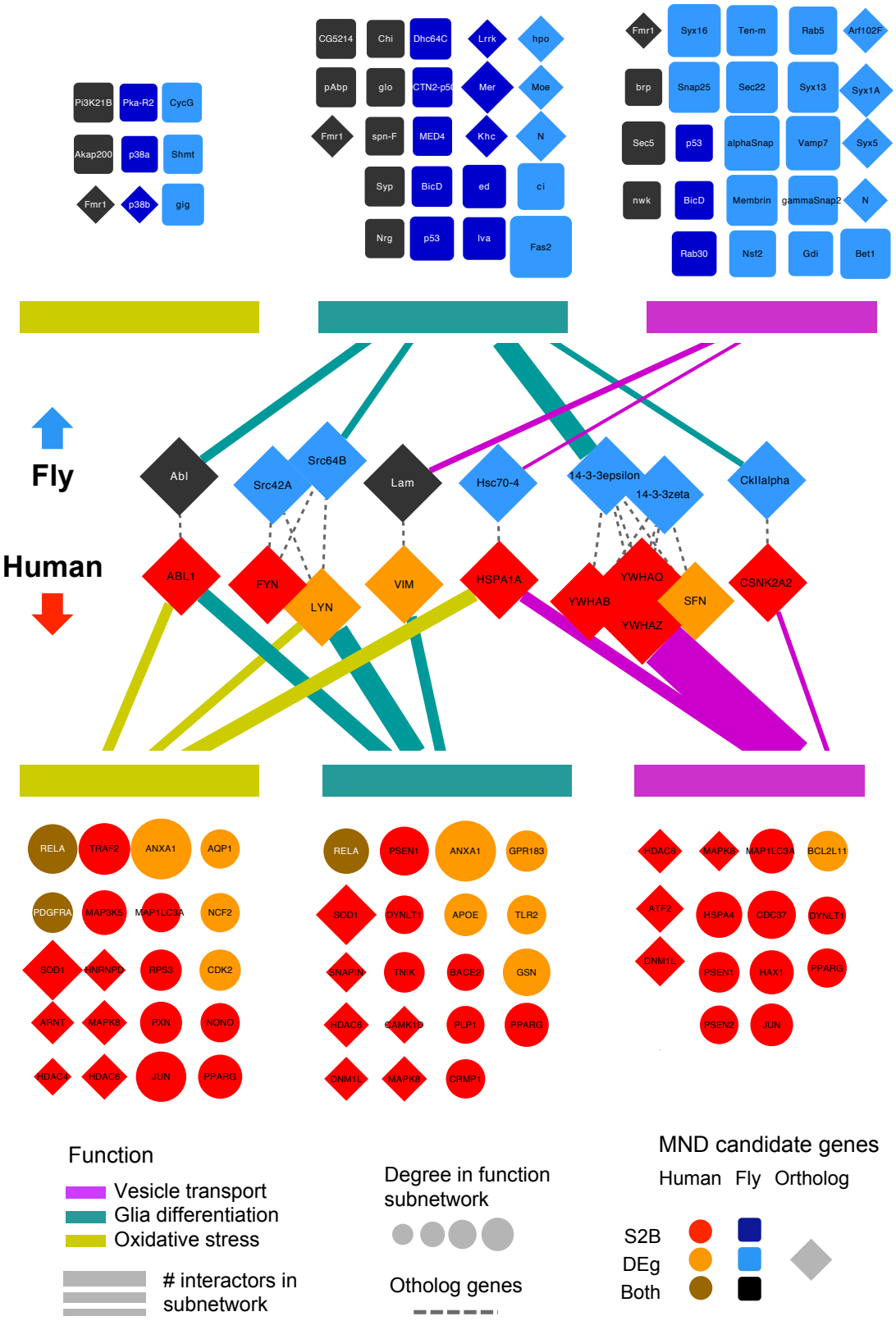

Pi3K21B
Pka-R2
CycG
Akap200
p38a
Shmt
Fmr1
p38b
gig
CG5214
Chi
Dhc64C
Lrrk
hpo
pAbp
glo
CTN2-p5
Mer
Moe
Fmr1
spn-F
MED4
Khc
N
Syp
BicD
ed
ci
Nrg
p53
lva
Fas2
Fmr1
Syx16
Ten-m
Rab5
Arf102F
brp
Snap25
Sec22
Syx13
Syx1A
Sec5
p53
alphaSnap
Vamp7
Syx5
nwk
BicD
Membrin
gammaSnap2
N
Rab30
Nsf2
Gdi
Bet1
Fly
Human
Abl
Src42A
Src64B
Lam
Hsc70-4
14-3-3epsilon
14-3-3zeta
CkIIalpha
ABL1
FYN
LYN
VIM
HSPA1A
YWHAB
YWHAQ
YWHAZ
SFN
CSNK2A2
RELA
TRAF2
ANXA1
AQP1
PDGFRA
MAP3K5
MAP1LC3A
NCF2
SOD1
HNRNPD
RPS3
CDK2
ARNT
MAPK8
PXN
NONO
HDAC4
HDAC6
JUN
PPARG
RELA
PSEN1
ANXA1
GPR183
SOD1
DYNLT1
APOE
TLR2
SNAPIN
TNIK
BACE2
GSN
HDAC6
CAMK1D
PLP1
PPARG
DNM1L
MAPK8
CRMP1
HDAC6
MAPK8
MAP1LC3A
BCL2L11
ATF2
HSPA4
CDC37
DYNLT1
DNM1L
PSEN1
HAX1
PPARG
PSEN2
JUN
Function
Vesicle transport
Glia differentiation
Oxidative stress
interactors in subnetwork
Degree in function subnetwork
Otholog genes
MND candidate genes
Human Fly Ortholog
S2B
DEg
Both

Figure 5.6 Core-PPI network linking human and *Drosophila* orthologs identified as MND candidates

Node shape indicates the specie while color indicates the type of MND candidate. Dashed edges represent ortholog relation between fly and human genes. Edge width indicates the number of interactions each ortholog presents in each function. Node size of proteins in functional groups indicates the number of interactions within the same function subnetwork.

Vesicle transport, glia differentiation and oxidative stress are molecularly interrelated. Glia encompasses several types of non-neuronal cells including astrocytes, microglia, oligodendrocytes, among others. Both glia and neuron cells exhibit a coordinated morphogenesis and exert a close communication. Astrocytes and microglia are closely implicated in neuron synapse development and maintenance (reviewed by *(Stogsdill and Eroglu, 2017).* Microglia, the resident macrophage of the CNS, reacts to lesions of the nervous system by releasing pro-inflammatory factors. The overstimulation of innate immune response and inflammatory cascades directly affects to the neurons' physiology aggravating ongoing pathological states. In particular, activated microglia can induce overproduction of ROS and increase oxidative stress in neighboring neurons. Vesicle trafficking is pivotal for axonogenesis, synaptogenesis and synapse activity. Cell-to-cell communication and chemotaxis are modulated by dynamic intracellular signaling pathways that require rapid protein expression mechanisms at the synapse. Thus, axonal transport of the elements necessary for protein translation - including RNA and ribosomal macromolecules among others - becomes essential for normal neuronal physiology *(Rangaraju et al., 2017).* On this basis, it is not surprising to find that our results highlight proteins interrelating glia differentiation, synapse vesicle traffic and oxidative stress. Of note, we found that "glia differentiation" group covers many proteins involved in axonogenesis and synaptic traffic.

Most important, we found that several of the ortholog pairs in the core networks are involved in **heterogeneous intracellular signaling pathways. 14-3-3 proteins (YWHA)** form complexes that bind to a multitude of proteins providing a dynamic PPI hub to regulate signaling processes related to neurogenesis, and synaptogenesis (reviewed in *(Cornell and Toyo-oka, 2017)).* Human **SFN** (14-3-3 sigma) also collaborates in cell cycle and DNA damage response pathways. In *Drosophila*, 14-3-3 proteins have also been implicated in neuronal differentiation.

Non-receptor tyrosine-protein kinases **Src-FYN/LYN** and **Abl-ABL1** ortholog pairs regulate varying intracellular signaling pathways. **FYN Src kinase** mainly targets membrane receptors for regulating synaptic traffic *(Nygaard et al., 2014).* Among **ABL1** substrates we find signaling adaptors, other kinases, cytoskeletal proteins, TFs and chromatin remodelers implicated in cell cycle and DNA damage control *(Wang, 2014)*. On the other hand, the orthologs of the catalytic subunit of Casein kinase II **(CSNK2A2-CkIIalpha)** were also identified as fly and human MND candidates. CKII phosphorylates a large number of substrates implicated in varied processes involved in synaptic plasticity and proteostasis including chaperones, cell cycle regulators and apoptosis mediators **(Castello et al., 2017).**

It was remarkable that the fly orthologs just mentioned are distinctively involved in synaptic vesicle trafficking, while, at the same time, their human orthologs were additionally implicated in cellular stress and DNA damage responses.

Among the *Drosophila* MND candidates (top half, **Figure 5.6)** we found several Syntaxin proteins **(Syx5, Syx1A, Syx13 and Syx16).** Syntaxins are SNAP-receptors (SNARE proteins) implicated in secretory pathways including neurotransmitter (NT) release (reviewed in *(Quiñones-Frías and Littleton, 2021)).* Likewise, the network highlighted several SNAP proteins implicated in ER-Golgi and endosome-phagosome vesicular transport **(snap25, alphaSnap, gammaSnap2)** and the vesicle-fusing ATPase 2 **(Nsf2).** Additional proteins decisive for axonal guidance and synapse formation in *Drosophila* were **Nrg** and **Fas2** cell-adhesion proteins and **Ten-m** transmembrane protein that interacts with cytoskeleton-binding proteins as **Nwk** *(Mosca, 2015).* Nwk, in turn, is a cytoskeleton-remodeling protein that localizes in presynaptic zones to regulate endosomal cargoes.

In parallel, the counterpart human orthologs revealed direct interactions with human MND candidates involved in **cellular stress control.** **HSPA1A-Hsc70-4** orthologs are heat shock protein members of Hsp70 protein family. Hsp70 are cytosolic molecular chaperones essential for cell homeostasis including protein folding and degradation. Fly Hsc70 has been involved in clathrin-mediated endocytosis while human HSAP1A has been implicated in UPR response to oxidative

stress *(Shah et al., 2017)* and chromatin remodeling to DNA damage *(Shukla and Tekwani, 2020)* by interacting with several MAP kinases and histone deacetylases - identified as MND candidates too -. In turn, **MAPK8** targets multiple TFs also identified as MND candidates namely, **JUN** stress-related signaling pathway and **RELA**/NK-kappa B pathways. NF-KB is a transcription regulator essential for many inflammatory factors. At the same time, platelet-derived growth factor receptor A **(PDGFRA)** is essential in nervous system development and adult neuronal maintenance or neuroprotection *(Funa and Sasahara, 2014).*

Pathways related to **MAPK8, JUN or NFKB-RELA** have been already pointed as potential therapeutic targets for neuroprotection in Parkinson's Disease *(Rai et al., 2021).* However, the neuroinflammation observed across all neurodegenerative diseases - including MNDs - can be induced by varying factors and cell cross-talks (reviewed in *(Z. Liu et al., 2020)).* Likewise, as observed for NFKB activation, the release of pro-inflammatory factors can have both beneficial or detrimental roles depending on the disease stage *(Ouali Alami et al., 2018).* Beyond treatments to decrease neuroinflammation, inhibitors of **Src** kinase and drug targets of **CKII** have are being investigated in Alzheimer's disease and other psychiatric disorders *(Castello et al., 2017; Nygaard et al., 2014).* The gene expression deregulation of **YWHA** proteins is being investigated for indirect prognosis of pre-symptomatic psychosis too *(Demars et al., 2020).*

5.4 Discussion

The goal of the work presented in the last chapter was to integrate the knowledge extracted throughout the thesis to generate a transversal hypothesis underlying MND in fly and human species. We found that a large fraction of human and fly DEgs and S2B candidates exhibited broad tissue expression. Likewise, the mapping of DEgs and S2B candidates in BioInt units across the TS libraries did not show a distinctive accumulation in brain tissue. However, within brain BioInt library, it was noteworthy that the BioInt units enriched in MND candidates were actually more tissue specific. These trends were previously observed in **Chapter 2** substantiating that, disregarding the high number of housekeeping proteins, disease-associated proteins are particularly involved in TS biological processes.

Nonetheless, the tissue distribution of BioInt units enriched in fly and human MND candidates was notably distinct. The difference between BioInt units enriched in DEgs (DEg-BioInt units) is justified by the use of different types of samples in human and *Drosophila* (MN from spinal cord samples and fly head lysates, respectively). On the other hand, the disparity between S2B-BioInt units probably indicates differences in the topology of the species' networks. Considering that the goal of the S2B method is to identify network bottlenecks, the wide tissue distribution of human S2B-BioInt units reinforces that S2B candidates are involved in cellular processes essential to any tissue. However, fly S2B candidates were prioritized using orthologs from ALS and SMA human DGs. Thus, the unexpected tissue specificity of fly S2B-BioInt units suggests the topology of disease modules generated from ortholog DGs is artificially distorted. On the other hand, the GO-BP coverage analysis across human and *Drosophila* datasets indicated that the small GO-BP overlap between TS PPI networks is mainly due to biological differences inherent to the species. We showed that the combination of BioInt units and semantic similarity information improves the integration of cross-specie knowledge.

The core-network reconstructed from the combination of BioInt units enriched in human and fly MND protein candidates revealed a close molecular relation between **signaling pathways and vesicle trafficking to coordinate glia-MN**

communication and synaptogenesis. It was patent the protein candidates in MND core-network are involved in heterogeneous processes and so may have a wide impact in cell homeostasis. Accordingly, most of these proteins were effectively prioritized by the S2B method. Therefore, they likely constitute molecular bottlenecks with potential application in the design of therapeutic treatments.

The implication of glia deregulation in MND has gained much prominence during recent years *(T. Kim et al., 2020; Komine and Yamanaka, 2015).* Our findings also suggested that MN and glial communication is at the center of the pathological hallmarks identified in human and fly species. On the other hand, the imbalanced number of candidates involved in oxidative stress and synaptic vesicle traffic between human and *Drosophila* may reflect the physiological differences between the two species. Human MNs are long living cells and so are more susceptible to accumulate oxidative stress that fly MNs. Thus, it is plausible that human MNs require tighter control of ROS species and DNA damage that fly MNs. The observation that the same orthologs pairs are involved in varying functions could manifest that these human MND candidates have acquired additional roles in human.

5.5 Supplementary Data

R code to reproduce BioInt-U method is available in:

https://github.com/GamaPintoLab/BioInt-U

Supplementary data files are available in:

https://github.com/GamaPintoLab/MLG_PhDThesis_SupData

Supplementary Data S5.1 *Drosophila* TS BioInt units

Supplementary Data S4.2 Selected Human and *Drosophila* BioInt units

6 Integrated discussion

The ultimate objective of this PhD research was to identify common molecular mechanisms of MNDs. The inherent complexity of biological systems hinders the identification of critical molecular players of multigenic diseases. Nonetheless, the characterization of biomolecular networks can expose underlying mechanisms of biological organization and therefore, provide information on the triggering events of pathological phenotypes.

Today, we can access a large volume of cell- and tissue-specific transcriptomic studies in public repositories. These in turn have been employed to systematically reconstruct tissue-specific interactomes from tissue-naive PPI datasets *(Kotlyar et al., 2016).* The BioInt-U method made use of these tissue-specific networks to identify biologically interacting units, i.e., groups of interacting proteins associated to the same enriched GO-BP terms. We next conducted a systematic comparison of the topological properties of proteins and functional units across 33 normal human tissues. In most cases, the topological characterization of tissue-specific networks has been restricted to disease-specific contexts *(Karimizadeh et al., 2019; Marín et al., 2019; Sircar and Parekh, 2015; Will and Helms, 2016).* To our best knowledge though, few studies have aimed to characterize the normal coordination of biological processes along the distinct tissue-specific interactomes. From our perspective, the dissection of the normal tissue functionomes is pivotal to distinguish cell type specific functions from those essential to any type of cell. This knowledge in turn, is critical to understand the mechanisms by which mutations in housekeeping genes can trigger tissue-specific disease manifestations.

The analysis of human TS BioInt libraries supported an in-depth comparison between HK and TE functions. We corroborated that HK units are related to core functions such as organelle trafficking, RNA or protein metabolism and are mostly made up of UB proteins with significantly larger degree and betweenness coefficients than proteins exclusively involved in TE functions. Likewise, our results suggested that nonUB proteins are critical players in the coordination of both HK and TE functions. The systematic mapping of DGs in the BioInt units indicated that more than 55% of total DGs - including MNDs - were ubiquitously expressed and overall, displayed a broader expression profile than nonDG proteins. We found that the UB proteins

encoded by DGs displayed different interaction profiles depending on the TS network. In particular, proteins associated to TS diseases displayed higher degree and betweenness coefficient in the corresponding tissue networks. This observation indicates that the HK proteins might acquire more central roles in the affected tissues. In parallel, the analysis evinced that nonUB proteins associated to TS diseases are critical for the functional coordination and so could similarly alter core functions.

We are aware that BioInt-U only recapitulates functional coordination at the protein level and so it lacks from crucial information such as gene regulation interactions *(Sonawane et al., 2017).* Nonetheless, we opted for a strategy that only requires GO-BP ontology, transcriptomic and interactomic data to be easily adapted to less well characterized organisms, or to take advantage of RNA-seq and interactomic datasets from novel tissue samples or single cell studies. We find several methods with similar goals to ours *(Basha et al., 2020; Greene et al., 2015; Jung et al., 2021; Vella et al., 2018).* However the data requirements and algorithms necessary to apply these strategies are much more complex that the one we propose. As recently pointed out by Zolotareva and Kleine, from the ~100 functional enrichment methods published since 2002, only 34% were currently accessible (Zolotareva and Kleine, 2019). Thus, it is not surprising to find that, despite their drawbacks and limitations, ORA (hyper-geometric test) and GSEA (pre-ranked gene set enrichment) are the most extensively used functional enrichment strategies.

Among the most similar methods to ours, we may highlight MTGO *(Vella et al., 2018)*, a clustering method that integrates GO and interactomic information to reconstruct functional complexes. However, due to the computational complexity of network clustering, the computing time of MTGO often required impracticable amounts of time. For example, for a network consisting of 2,400 nodes annotated with >13,000 GO terms and ~20,000 interactions, MTGO required > 24 hours, while BioInt-U performed the same analysis in ~30 minutes. Overall, we have shown the benefits of using the BioInt-U method to explore the topological properties of tissue-specific interactomes. Additionally, we demonstrated its potential to functionally characterize complex candidate profiles, as the ones retrieved from RNA-seq

datasets of *Drosophila* MND knockdown models, human ALS, psoriasis, and pulmonary fibrosis.

Cellular activity is the result of the coordination of closely interrelated molecular events. Therefore, the alteration of distinct molecular elements can affect the interconnected pathways and trigger similar pathophenotypes. MNDs are a prime example of the molecular complexity of disease conditions. MNDs encompass a spectrum of MN degenerative conditions associated to numerous genetic alterations. ALS and SMA are the most frequent subtypes of MND and, according to DisGeNET on May 2021, they have been associated to the mutation of >180 and >30 genes, respectively ***(Piñero et al., 2020).*** It is therefore surprising that even though ALS and SMA only share 4.2%-25% of disease-linked genes, the two conditions present overlapping clinical hallmarks. Despite the phenotypic variability across patients ***(Lopate et al., 2010),*** it is evident that the identification of transversal molecular events is critical to understand the emergence of pathological conditions and thus develop effective clinical solutions. To that end, the characterization of biological networks and the prioritization of key players in molecular communication is a meaningful first step. On this basis, we made use of network biology principles to investigate the cross-sectional characteristics of MND subtypes. The results retrieved from the analysis of BioInt libraries pointed that the DGs actually display higher centrality in TS networks. This observation reinforces the S2B method usefulness to prioritize proteins involved in pathological conditions. We showed the benefits of the S2B method in identifying bottleneck candidates specific to connect ALS and SMA disease modules.

The functional characterization of MND candidates corroborated that both the proteins specifically connecting ALS and SMA disease modules and the proteins deregulated in the human and fly disease models are frequently involved in RNA metabolism, vesicle trafficking, intracellular signaling and stress. The RNA transcription, cytoskeleton organization, vesicle traffic and signaling pathways are noticeably interrelated and converge onto the coordination of axonal transport in neurons. Axonal transport is essential for the local protein translation at the synapse and so for the communication of any neuron. However, MNs establish extraordinarily

long axonal projections and are therefore more sensitive to disturbances in the axonal transport. Additionally, the deregulation of axonal transport also affects to the retrograde transport pathways that coordinate autophagy and protein recycling. Thus, the deregulation of axonal transport is expected to deregulate organelle, protein and RNA clearance and so trigger cellular stress *(Houghton et al., 2022).* In turn, the cellular stress has a direct impact on DNA damage *(Konopka and Atkin, 2018),* RNA metabolism and protein homeostasis, perpetuating the MN deregulation. On the other way around, additional sources of stress as mitochondria deregulation *(Lau et al., 2018)* or astrocytes overstimulation (Kia et al., 2018) could similarly impact DNA/RNA homeostasis and signaling pathways involved in intracellular trafficking *(Ding et al., 2022).* From a physiological perspective, it is patent that these functions are dependent one on each other thus, the distinct alterations can converge into axonal transport deficiency, synapse malfunction and MN degeneration *(Ragagnin et al., 2019*). In the same line, our results brought additional evidences that the MND candidates can exert varying functions depending on their interactomic context and serve as multifunctional coordinators.

The work presented in the thesis has several limitations, many of which are derived from the lack of experimental data. The transcript and protein interactome greatly vary depending on the subcellular compartment and cellular state. However, the protein interactome is a static representation of all the possible physical interactions between proteins. This limitation is compounded by the fact that most popular high-throughput technologies to characterize PPIs are unable to capture weak interactions. The transient or weak interactions are equally essential as the most stable interactions thus, it is predicted that we still lack a large fraction of functional PPI data *(Ghadie and Xia, 2022).* At the same time, the cellular transcriptome is highly dynamic and RNA-seq methods can only capture the singular state of the sample investigated. Therefore, the information represented in the two types of omic data is largely incomplete and so, the predictions extracted from static network-based analysis do not necessarily resemble to all patients' conditions. Single-cell and spatial transcriptomics and proteomics are already being employed to target specific research questions *(Adil et al., 2021; Lundberg and Borner, 2019).* Equivalent efforts are being made to characterize protein transient interactions through proximity-

labeling methods as APEX or BioID *(Bosch et al., 2021).* These methods bring additional challenges, namely in the design of protocols for normalization, reduction of data dimensionality, and integration in multi-omic models. Regardless, the systematic use of these technologies in a high-throughput fashion will certainly bring groundbreaking advances in network medicine approaches.

The prioritization of disease-targeted functions was addressed using the hyper-geometric test to identify BioInt units significantly accumulating DGs. The decision relied in the assumption that genes related to disease conditions will most often be found in functions essential for cell homeostasis *(Barabási et al., 2011).* Thus, the analysis did not address the effect of the discrete mutations on functional coordination. Traditionally, most network studies have considered gene mutations as node removal. However, many MND-causing mutations can increase protein interaction strength and/or promote new interactions with additional proteins. For instance, nearly 200 different mutations in the SOD1 gene have been associated to diverse ALS phenotypes *(Bernard et al., 2020).* The analysis of protein interaction rewiring would require in-depth experimental characterization of the interactomic changes derived from each specific mutation. Once more data is available, modeling the rewiring of BioInt units could reveal additional mechanistic insights to explain tissue-specific vulnerabilities.

The use of animal models such as mouse, zebra fish or *Drosophila* is essential for biomedical research in neurodegeneration. However, the conclusions outlined from animal models do not always translate into humans *(Ferreira et al., 2020).* The most obvious reason is that the species present divergent evolutionary patterns. In other words, the cellular and tissular differences between animal species mostly arise from the distinctive arrangement of molecular interactions *(Fan et al., 2019; Shou et al., 2011)*. Thus, even if several species share a large fraction of functional orthologs, these can trigger very distinct responses. This is likely the reason why the BioInt units enriched in human and *Drosophila* S2B candidates revealed distinct tissue distribution. This result urges further investigation of the topological differences between human and fly orthologs.

The observations drawn from this work emphasized that biological processes are not isolated outcomes but rather a continuum of molecular events. The work presented in this thesis has generated 264 and 108 S2B candidates and, 470 and 1098 candidate genes with altered expression in human and fly models, respectively. Overall, these candidates were involved in interconnected functions and were recurrently associated to central biological processes including RNA metabolism and vesicle trafficking regulation. The disruption of any central pathway is expected to have deleterious impact on cell survival and as we also observed, regulators of DNA and oxidative stress were concomitantly present among the most central MND candidates. In the same line, as previously observed in MND-causal genes, MND candidates displayed broad tissue expression patterns. This observation revives the debate on how mutation of housekeeping proteins primarily affects MN physiology. The functional and interactomic characterization of these candidates across 33 human tissues revealed they frequently present additional interactions in neuronal context. Thus, future studies on the tissue-specific interactions of MND candidates could elucidate the mechanisms underlying the tissue-specificity of MND phenotypes.

Biomedical research must provide short candidate lists for designing early diagnosis and therapeutic options. In that sense, the characterization of the tissue-specific roles of the candidates is valuable information for pharmaceutical research. Broadly expressed protein candidates can be investigated as potential early biomarkers to be traceable in "liquid-biopsies". On the other hand, tissue-specific membrane receptors could be suitable candidates for drug targeting. Notwithstanding, the construction of MND core-network required a very restrictive filtering based on previous notions on the disease. Thus, the core-network is knowledge-biased and emphasizes many observations form previous studies. In order to extract new knowledge regarding the pathology, it would be advantageous to further explore those candidates excluded from the core-network.

In any case, the MND core-network illustrated multiple molecular links that explain the hallmarks found in MNDs and demonstrated the benefits of using BioInt-U and S2B methods in MND research. Additionally, they were useful to delineate common events in the pathophenotypes in human and Drosophila. The present study proposes a conceptual model regarding how distinct DGs can converge onto

common functional modules and provides a first outline to continue investigating complex mechanisms of MND in PPI networks. Likewise, many other diseases have complex etiology and so the methods and insights here presented could also be profitable in other research areas.

7 References

Achsel T, Barabino S, Cozzolino M, Carrì MT. The intriguing case of motor neuron disease: ALS and SMA come closer. Biochem Soc Trans 2013;41:1593–1597.

Achtert K, Kerkemeyer L. The economic burden of amyotrophic lateral sclerosis: a systematic review. Eur J Heal Econ 2021;22:1151–66. https://doi.org/10.1007/s10198-021-01328-7.

Acuner Ozbabacan SE, Engin HB, Gursoy A, Keskin O. Transient proteinprotein interactions. Protein Eng Des Sel 2011;24:635–48. https://doi.org/10.1093/protein/gzr025.

Adhikari S, Nice EC, Deutsch EW, Lane L, Omenn GS, Pennington SR, et al. A high-stringency blueprint of the human proteome. Nat Commun 2020;11:1–16. https://doi.org/10.1038/s41467-020-19045-9.

Adil A, Kumar V, Jan AT, Asger M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. Front Neurosci 2021;15:398. https://doi.org/10.3389/fnins.2021.591122.

Aggarwal A, Nicholson G. Detection of preclinical motor neurone loss in SOD1 mutation carriers using motor unit number estimation. J Neurol Neurosurg Psychiatry 2002;73:199–201. https://doi.org/10.1136/jnnp.73.2.199.

Agrawal M, Zitnik M, Leskovec J. Large-scale analysis of disease pathways in the human interactome. Pacific Symp. Biocomput., vol. 23, NIH Public Access; 2018, p. 111–22. https://doi.org/10.1142/9789813235533_0011.

Alcalá-Corona SA, Sandoval-Motta S, Espinal-Enríquez J, Hernández-Lemus E. Modularity in Biological Networks. Front Genet 2021;12:701331. https://doi.org/10.3389/fgene.2021.701331.

Alonso-Lopez D, Gutierrez MA, Lopes KP, Prieto C, Santamaria R, De Las Rivas J. APID interactomes : providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res 2016;44:529–35. https://doi.org/10.1093/nar/gkw363.

Alonso-López Di, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M, et al. APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. Database 2019;2019. https://doi.org/10.1093/database/baz005.

Amaral AJ, Brito FF, Chobanyan T, Yoshikawa S, Yokokura T, Vactor D, et al. Quality assessment and control of tissue specific RNA-seq libraries of Drosophila transgenic RNAi models. Front Genet 2014;5:43.

Amberger JS, Hamosh A. Searching online mendelian inheritance in man (OMIM): A knowledgebase of human genes and genetic phenotypes. Curr Protoc Bioinforma 2017:1.2.1-1.2.12. https://doi.org/10.1002/cpbi.27.

Anderson P.W. More is Different. Science 1972;177:393–6. https://doi.org/10.1126/science.177.4047.393.

Aquilina B, Cauchi RJ. Modelling motor neuron disease in fruit flies: Lessons from spinal muscular atrophy. J Neurosci Methods 2018;310:3–11.

Arai T, Hasegawa M, Akiyama H, Ikeda K, Nonaka T, Mori H, et al. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochem Biophys Res Commun 2006;351:602–11.

https://doi.org/10.1016/j.bbrc.2006.10.093.

Armaos A, Zacco E, Sanchez de Groot N, Tartaglia GG. RNA-protein interactions: Central players in coordination of regulatory networks. BioEssays 2021;43:2000118. https://doi.org/10.1002/bies.202000118.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000 251 2000;25:25–9. https://doi.org/10.1038/75556.

Aulas A, Fay MM, Lyons SM, Achorn CA, Kedersha N, Anderson P, et al. Stress-specific differences in assembly and composition of stress granules and related foci. J Cell Sci 2017;130:927–37. https://doi.org/10.1242/jcs.199240.

Balendra R, Isaacs AM. C9orf72-mediated ALS and FTD: multiple pathways to disease. Nat Rev Neurol 2018;14:544–58. https://doi.org/10.1038/s41582-018-0047-2.

Balzani V V, Credi A, Raymo FM, Stoddart JF. Artificial Molecular Machines. Angew Chem Int Ed Engl 2000;39:3348–91. https://doi.org/10.1002/1521-3773(20001002)39:19<3348::aid-anie3348>3.0.co;2-x.

Barabasi A-L. Network Medicine — From obesity to the "Diseasome." N Engl J Med 2007;357:404–7. https://doi.org/10.1056/NEJMe078114.

Barabasi A-L, Albert R. Emergence of scaling in random networks. Science 1999;286:509–512. https://doi.org/10.1126/science.286.5439.509.

Barabási A, Gulbahce N, Loscalzo J. Network Medicine: A Network-based approach to human disease. Nat Rev Genet 2011;12:56–68. https://doi.org/10.1038/nrg2918.Network.

Barabási A, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet 2004;5:101–13. https://doi.org/10.1038/nrg1272.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013;41:D991–5. https://doi.org/10.1093/NAR/GKS1193.

Barshir R, Shwartz O, Smoly IY, Yeger-Lotem E. Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases. PLOS Comput Biol 2014;10:e1003632. https://doi.org/10.1371/JOURNAL.PCBI.1003632.

Basha O, Argov CM, Artzy R, Zoabi Y, Hekselman I, Alfandari L, et al. Differential network analysis of multiple human tissue interactomes highlights tissue-selective processes and genetic disorder genes. Bioinformatics 2020;36:2821–8. https://doi.org/10.1093/bioinformatics/btaa034.

Bernard E, Pegat A, Svahn J, Bouhour F, Leblanc P, Millecamps S, et al. Clinical and molecular landscape of ALS patients with SOD1 mutations: Novel pathogenic variants and novel phenotypes. a single ALS center study. Int J Mol Sci 2020;21:1–11. https://doi.org/10.3390/ijms21186807.

Biran H, Kupiec M, Sharan R. Comparative analysis of normalization methods for network propagation. Front Genet 2019;10:4. https://doi.org/10.3389/fgene.2019.00004.

Birsa N, Bentham MP, Fratta P. Cytoplasmic functions of TDP-43 and FUS and their role in ALS. Semin Cell Dev Biol 2020;99:193–201.

Blokhuis AM, Groen EJN, Koppers M, Van Den Berg LH, Pasterkamp RJ. Protein aggregation in amyotrophic lateral sclerosis. Acta Neuropathol 2013;125:777–94. https://doi.org/10.1007/s00401-013-1125-6.

Boczonadi V, Müller JS, Pyle A, Munkley J, Dor T, Quartararo J, et al. EXOSC8 mutations alter mRNA metabolism and cause hypomyelination with spinal

muscular atrophy and cerebellar hypoplasia. Nat Commun 2014;5. https://doi.org/10.1038/ncomms5287.

Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res 2004;32. https://doi.org/10.1093/nar/gkh061.

Bolus H, Crocker K, Boekhoff-Falk G, Chtarbanova S. Modeling neurodegenerative disorders in drosophila melanogaster. Int J Mol Sci 2020;21. https://doi.org/10.3390/ijms21093055.

Bosch JA, Chen CL, Perrimon N. Proximity-dependent labeling methods for proteomic profiling in living cells: An update. Wiley Interdiscip Rev Dev Biol 2021;10:e392. https://doi.org/10.1002/wdev.392.

Bossi A, Lehner B. Tissue specificity and the human protein interaction network. Mol Syst Biol 2009;5:260. https://doi.org/10.1038/msb.2009.17.

Boulisfane N, Choleza M, Rage F, Neel H, Soret J, Bordonné R. Impaired minor tri-snRNP assembly generates differential splicing defects of U12-type introns in lymphoblasts derived from a type I SMA patient. Hum Mol Genet 2011;20:641–648.

Bowerman M, Murray LM, Scamps F, Schneider BL, Kothary R, Raoul C. Pathogenic commonalities between spinal muscular atrophy and amyotrophic lateral sclerosis: Converging roads to therapeutic development. Eur J Med Genet 2018;61:685–698.

Brand AH, Perrimon N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development 1993;118:401–15. https://doi.org/10.1242/dev.118.2.401.

Brito GC, Andrews DW. Removing bias against membrane proteins in interaction networks. BMC Syst Biol 2011;5. https://doi.org/10.1186/1752-0509-5-169.

Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol 2007;8:1–11. https://doi.org/10.1186/gb-2007-8-5-r95.

Brunner HG, van Driel MA. From syndrome families to functional genomics. Nat Rev Genet 2004;5:545–51. https://doi.org/10.1038/nrg1383.

Burk K, Pasterkamp RJ. Disrupted neuronal trafficking in amyotrophic lateral sclerosis. Acta Neuropathol 2019;137:859–77. https://doi.org/10.1007/s00401-019-01964-7.

Cacciottolo R, Ciantar J, Lanfranco M, Borg RM, Vassallo N, Bordonné R, et al. SMN complex member Gemin3 self-interacts and has a functional relationship with ALS-linked proteins TDP-43, FUS and Sod1. Sci Rep 2019;9:18666.

Calderone A, Formenti M, Aprea F, Papa M, Alberghina L, Colangelo AM, et al. Comparing Alzheimer's and Parkinson's diseases networks using graph communities structure. BMC Syst Biol 2016;10:25. https://doi.org/10.1186/s12918-016-0270-7.

Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Res 2021;49:D325–34. https://doi.org/10.1093/nar/gkaa1113.

Carrì MT, Valle C, Bozzo F, Cozzolino M. Oxidative stress and mitochondrial damage: importance in non-SOD1 ALS. Front Cell Neurosci 2015;9:41. https://doi.org/10.3389/fncel.2015.00041.

Castello J, Ragnauth A, Friedman E, Rebholz H. CK2—an emerging target for neurological and psychiatric disorders. Pharmaceuticals 2017;10. https://doi.org/10.3390/ph10010007.

Chang HC, Dimlich DN, Yokokura T, Mukherjee A, Kankel MW, Sen A, et al. Modeling spinal muscular atrophy in Drosophila. PLoS One 2008;3:e3209.

Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C. Extreme multifunctional proteins identified from a

human protein interaction network. Nat Commun 2015;6:7412. https://doi.org/10.1038/ncomms8412.

Charng WL, Yamamoto S, Bellen HJ. Shared mechanisms between Drosophila peripheral nervous system development and human neurodegenerative diseases. Curr Opin Neurobiol 2014;27:158–64. https://doi.org/10.1016/j.conb.2014.03.001.

Chen S, Sayana P, Zhang X, Le W. Genetics of amyotrophic lateral sclerosis: an update. Mol Neurodegener 2013;8:28. https://doi.org/10.1186/1750-1326-8-28.

Chen Y, Bennett CL, Huynh HM, Blair IP, Puls I, Irobi J, et al. DNA / RNA Helicase gene mutations in a form of Juvenile Amyotrophic Lateral Sclerosis (ALS4). Am J Hum Genet 2004;74:1128–35.

Chi B, O'Connell JD, Iocolano AD, Coady JA, Yu Y, Gangopadhyay J, et al. The neurodegenerative diseases ALS and SMA are linked at the molecular level via the ASC-1 complex. Nucleic Acids Res 2018;46:11939–11951.

Chipika RH, Siah WF, McKenna MC, Li Hi Shing S, Hardiman O, Bede P. The presymptomatic phase of amyotrophic lateral sclerosis: are we merely scratching the surface? J Neurol 2020:Advance online publication. https://doi.org/10.1007/s00415-020-10289-5.

Cornell B, Toyo-oka K. 14-3-3 proteins in brain development: Neurogenesis, neuronal migration and neuromorphogenesis. Front Mol Neurosci 2017;10:318. https://doi.org/10.3389/fnmol.2017.00318.

Coy S, Volanakis A, Shah S, Vasiljeva L. The Sm Complex Is Required for the Processing of Non-Coding RNAs by the Exosome. PLoS One 2013;8:e65606. https://doi.org/10.1371/journal.pone.0065606.

Cragnaz L, Klima R, Skoko N, Budini M, Feiguin F, Baralle FE. Aggregate formation prevents dTDP-43 neurotoxicity in the Drosophila melanogaster eye. Neurobiol Dis 2014;71:74–80.

Da Cruz S, Cleveland DW. Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. Curr Opin Neurobiol 2011;21:904–19. https://doi.org/10.1016/j.conb.2011.05.029.

Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal 2006;Complex Sy:1695.

Dadon-Nachum M, Melamed E, Offen D. The "dying-back" phenomenon of motor neurons in ALS. J Mol Neurosci 2011;43:470–7. https://doi.org/10.1007/s12031-010-9467-1.

Darbà J. Current status and direct medical cost of amyotrophic lateral sclerosis in the region of Catalonia: A population-based analysis. PLoS One 2019;14. https://doi.org/10.1371/journal.pone.0223772.

Deeds E, Krivine J, Feret J, Danos V, Fontana W. Combinatorial complexity and compositional drift in protein interaction networks. PLoS One 2012;7. https://doi.org/10.1371/JOURNAL.PONE.0032032.

Demars F, Kebir O, Marzo A, Iftimovici A, Schramm C, Amado I, et al. Dysregulation of peripheral expression of the YWHA genes during conversion to psychosis. Sci Rep 2020;10:1–9. https://doi.org/10.1038/s41598-020-66901-1.

Dezső Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, et al. A comprehensive functional analysis of tissue specificity of human gene expression. BMC Biol 2008;6:1–15. https://doi.org/10.1186/1741-7007-6-49.

Ding Q, Kesavan K, Lee KM, Wimberger E, Robertson T, Gill M, et al. Impaired signaling for neuromuscular synaptic maintenance is a feature of Motor Neuron Disease. Acta Neuropathol Commun 2022;10:61. https://doi.org/10.1186/s40478-022-01360-5.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

Donlin-Asp PG, Bassell GJ, Rossoll W. A role for the survival of motor neuron protein in mRNP assembly and transport. Curr Opin Neurobiol 2016;39:53–61.

Donlin-Asp PG, Fallini C, Campos J, Chou CC, Merritt ME, Phan HC, et al. The Survival of Motor Neuron Protein Acts as a Molecular Chaperone for mRNP Assembly. Cell Rep 2017;18:1660–1673.

Ederle H, Dormann D. TDP-43 and FUS en route from the nucleus to the cytoplasm. FEBS Lett 2017;591:1489–1507.

Erdös P, Rényi A. On the evolution of random graphs. Publ Math Inst Hung Acad Sci 1960;5:17–61. https://doi.org/10.2307/1999405.

Espinosa-Cantú A, Cruz-Bonilla E, Noda-Garcia L, DeLuna A. Multiple Forms of Multifunctional Proteins in Health and Disease. Front Cell Dev Biol 2020;8:451. https://doi.org/10.3389/fcell.2020.00451.

Fallini C, Donlin-Asp PG, Rouanet JP, Bassell GJ, Rossoll W. Deficiency of the Survival of Motor Neuron Protein Impairs mRNA Localization and Local Translation in the Growth Cone of Motor Neurons. J Neurosci 2016;36:3811–3820.

Fallini C, Khalil B, Smith CL, Rossoll W. Traffic jam at the nuclear pore: All roads lead to nucleocytoplasmic transport defects in ALS/FTD. Neurobiol Dis 2020;140:104835. https://doi.org/10.1016/j.nbd.2020.104835.

Fallini C, Zhang H, Su Y, Silani V, Singer RH, Rossoll W, et al. The survival of motor neuron (SMN) protein interacts with the mRNA-binding protein HuD and regulates localization of poly(A) mRNA in primary motor neuron axons. J Neurosci 2011;31:3914–25. https://doi.org/10.1523/JNEUROSCI.3631-10.2011.

Fan J, Cannistra A, Fried I, Lim T, Schaffner T, Crovella M, et al. Functional protein representations from biological networks enable diverse cross-species inference. Nucleic Acids Res 2019;47:e51–e51. https://doi.org/10.1093/nar/gkz132.

Farg MA, Konopka A, Soo KY, Ito D, Atkin JD. The DNA damage response (DDR) is induced by the C9orf72 repeat expansion in amyotrophic lateral sclerosis. Hum Mol Genet 2017;26:2882–96. https://doi.org/10.1093/hmg/ddx170.

Farrar MA, Kiernan MC. The Genetics of Spinal Muscular Atrophy: Progress and Challenges. Neurotherapeutics 2015;12:290–302. https://doi.org/10.1007/s13311-014-0314-x.

Fernandes N, Eshleman N, Buchan JR. Stress Granules and ALS: A Case of Causation or Correlation? Adv Neurobiol 2018;20:173–212.

Ferreira GS, Veening-Griffioen DH, Boon WPC, Moors EHM, van Meer PJK. Levelling the translational gap for animal to human efficacy data. Animals 2020;10:1–13. https://doi.org/10.3390/ani10071199.

Fiesel FC, Voigt A, Weber SS, Haute C, Waldenmaier A, Görner K, et al. Knockdown of transactive response DNA-binding protein (TDP-43) downregulates histone deacetylase 6. EMBO J 2010;29:209–221.

Funa K, Sasahara M. The roles of PDGF in development and during neurogenesis in the normal and diseased nervous system. J Neuroimmune Pharmacol 2014;9:168–81. https://doi.org/10.1007/s11481-013-9479-z.

Gama-Carvalho M, Garcia-Vaquero ML, Pinto FR, Besse F, Weis J, Voigt A, et al. Linking Amyotrophic Lateral Sclerosis and Spinal Muscular Atrophy through RNA-transcriptome homeostasis: a genomics perspective. J Neurochem 2017;38:42–9. https://doi.org/10.1111/jnc.13945.

Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, et al.

Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 2006;38:285–93. https://doi.org/10.1038/ng1747.

Gaudet P, Dessimoz C. Gene ontology: Pitfalls, biases, and remedies. Methods Mol. Biol., vol. 1446, Humana Press Inc.; 2017, p. 189–205. https://doi.org/10.1007/978-1-4939-3743-1_14.

Gehlenborg N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets 2019.

Ghadie MA, Xia Y. Are transient protein-protein interactions more dispensable? PLOS Comput Biol 2022;18:e1010013. https://doi.org/10.1371/journal.pcbi.1010013.

Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLOS Comput Biol 2015;11:e1004120. https://doi.org/10.1371/JOURNAL.PCBI.1004120.

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res 2019;47:D559–63. https://doi.org/10.1093/NAR/GKY973.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. PNAS 2007;104:8685–90. https://doi.org/10.1073/pnas.0701361104.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet 2015;47:569–76. https://doi.org/10.1038/ng.3259.

Grice SJ, Liu JL. Survival motor neuron protein regulates stem cell division, proliferation, and differentiation in Drosophila. PLoS Genet 2011;7:1002030.

Groen EJN, Fumoto K, Blokhuis AM, Engelen-Lee JY, Zhou Y, van den Heuvel DMA, et al. ALS-associated mutations in FUS disrupt the axonal distribution and function of SMN. Hum Mol Genet 2013;22:3690–704. https://doi.org/10.1093/hmg/ddt222.

Grote S. GOfuncR: Gene ontology enrichment using FUNC 2020.

Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016.

Guo J, Price DH. RNA polymerase II transcription elongation control. Chem Rev 2013;113:8583–603. https://doi.org/10.1021/cr400105n.

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM): a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2002;30:52–5.

Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, et al. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell 2015;163:712–23. https://doi.org/10.1016/j.cell.2015.09.053.

Hekselman I, Yeger-Lotem E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat Rev Genet 2020;21:137–50. https://doi.org/10.1038/s41576-019-0200-9.

Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol 2018;19:327–41. https://doi.org/10.1038/nrm.2017.130.

Hergesheimer RC, Chami AA, De Assis DR, Vourc'h P, Andres CR, Corcia P, et al. The debated toxic role of aggregated TDP-43 in amyotrophic lateral sclerosis: A resolution in sight? Brain 2019;142:1176–94. https://doi.org/10.1093/brain/awz078.

Hill SJ, Mordes DA, Cameron LA, Neuberg DS, Landini S, Eggan K, et al. Two familial ALS proteins function in prevention/repair of transcription-associated DNA damage. Proc Natl

Acad Sci 2016;113:e7701–9. https://doi.org/10.1073/pnas.1611673113.

Hofmann Y, Wirth B. hnRNP-G promotes exon 7 inclusion of survival motor neuron (SMN) via direct interaction with Htra2-beta1. Hum Mol Genet 2002;11:2037–49.

Hohmann, Dehghani. The Cytoskeleton—A Complex Interacting Meshwork. Cells 2019;8:362. https://doi.org/10.3390/cells8040362.

Houghton OH, Mizielinska S, Gomez-Suaga P. The Interplay Between Autophagy and RNA Homeostasis: Implications for Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. Front Cell Dev Biol 2022;10:959. https://doi.org/10.3389/fcell.2022.838402.

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 2011;12. https://doi.org/10.1186/1471-2105-12-357.

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1–13. https://doi.org/10.1093/nar/gkn923.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 2015;12:115–121.

Hutten S, Sharangdhar T, Kiebler M. Unmasking the messenger. RNA Biol 2014;11:992–7. https://doi.org/10.4161/rna.32091.

Huttlin E, Bruckner R, Navarrete-Perea J, Cannon J, Baltier K, Gebreab F, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell 2021;184:3022-3040.e28. https://doi.org/10.1016/J.CELL.2021.04.011.

Iuchi K, Takai T, Hisatomi H. Cell death via lipid peroxidation and protein aggregation diseases. Biology (Basel) 2021;10. https://doi.org/10.3390/biology10050399.

Iwase S, Januma A, Miyamoto K, Shono N, Honda A, Yanagisawa J, et al. Characterization of BHC80 in BRAF–HDAC complex, involved in neuron-specific gene repression. Biochem Biophys Res Commun 2004;322:601–8. https://doi.org/10.1016/j.bbrc.2004.07.163.

Jalili M, Salehzadeh-Yazdi A, Gupta S, Wolkenhauer O, Yaghmaie M, Resendis-Antonio O, et al. Evolution of centrality measurements for the detection of essential proteins in biological networks. Front Physiol 2016;7:375. https://doi.org/10.3389/fphys.2016.00375.

Jangi M, Fleet C, Cullen P, Gupta S V, Mekhoubad S, Chiao E, et al. SMN deficiency in severe models of spinal muscular atrophy causes widespread intron retention and DNA damage. Proc Natl Acad Sci U S A 2017;114:e2347–56. https://doi.org/10.1073/pnas.1613181114.

Jankovska N, Matej R. Molecular pathology of ALS: What we currently know and what important information is still missing. Diagnostics 2021;11. https://doi.org/10.3390/diagnostics11081365.

Jung S, Singh K, Del Sol A. FunRes: Resolving tissue-specific functional cell states based on a cell-cell communication network model. Brief Bioinform 2021;22:1–10. https://doi.org/10.1093/bib/bbaa283.

Kankel MW, Sen A, Lu L, Theodorou M, Dimlich DN, McCampbell A, et al. Amyotrophic Lateral Sclerosis Modifiers in Drosophila Reveal the Phospholipase D Pathway as a Potential Therapeutic Target. Genetics 2020;215:747–766.

Kanouchi T, Ohkubo T, Yokota T. Can regional spreading of amyotrophic lateral sclerosis motor symptoms be explained by prion-like propagation? J Neurol Neurosurg Psychiatry 2012;83:739–45.

https://doi.org/10.1136/jnnp-2011-301826.

Karimizadeh E, Sharifi-Zarchi A, Nikaein H, Salehi S, Salamatian B, Elmi N, et al. Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis. BMC Med Genomics 2019;12. https://doi.org/10.1186/S12920-019-0632-2.

Kevenaar JT, Hoogenraad CC. The axonal cytoskeleton: From organization to function. Front Mol Neurosci 2015;8. https://doi.org/10.3389/fnmol.2015.00044.

Kia A, McAvoy K, Krishnamurthy K, Trotti D, Pasinelli P. Astrocytes expressing ALS-linked mutant FUS induce motor neuron death through release of tumor necrosis factor-alpha. Glia 2018;66:1016–33. https://doi.org/10.1002/glia.23298.

Kim G, Gautier O, Tassoni-Tsuchida E, Ma XR, Gitler AD. ALS Genetics: Gains, Losses, and Implications for Future Therapies. Neuron 2020;108:822–42. https://doi.org/10.1016/j.neuron.2020.08.022.

Kim T, Song B, Lee IS. Drosophila glia: Models for human neurodevelopmental and neurodegenerative disorders. Int J Mol Sci 2020;21:1–38. https://doi.org/10.3390/ijms21144859.

Kitano H. Biological robustness. Nat Rev Genet 2004;5:826–37. https://doi.org/10.1007/978-3-7643-7567-6_10.

Kline RA, Kaifer KA, Osman EY, Carella F, Tiberi A, Ross J, et al. Comparison of independent screens on differentially vulnerable motor neurons reveals alpha-synuclein as a common modifier in motor neuron diseases. PLoS Genet 2017;13:e1006680.

Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for Prioritization of Candidate Disease Genes. Am J Hum Genet 2008;82:949–58. https://doi.org/10.1016/j.ajhg.2008.02.013.

Kok JR, Palminha NM, Souza CDS, El-Khamisy SF, Ferraiuolo L. DNA damage as a mechanism of neurodegeneration in ALS and a contributor to astrocyte toxicity. Cell Mol Life Sci 2021;78:5707. https://doi.org/10.1007/S00018-021-03872-0.

Komine O, Yamanaka K. Neuroinflammation in motor neuron disease. Nagoya J Med Sci 2015;77:537–49.

Konopka A, Atkin JD. The emerging role of DNA damage in the pathogenesis of the C9Orf72 repeat expansion in amyotrophic lateral sclerosis. Int J Mol Sci 2018;19. https://doi.org/10.3390/ijms19103137.

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res 2017;45:D985–94. https://doi.org/10.1093/nar/gkw1055.

Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: Tissue-specific view of the human and model organism interactomes. Nucleic Acids Res 2016;44:D536–41. https://doi.org/10.1093/nar/gkv1115.

Kotlyar M, Rossos AEM, Jurisica I. Prediction of Protein-Protein Interactions. Curr Protoc Bioinforma 2017;60:8.2.1-8.2.14. https://doi.org/10.1002/cpbi.38.

Krach F, Batra R, Wheeler EC, Vu AQ, Wang R, Hutt K, et al. Transcriptome–pathology correlation identifies interplay between TDP-43 and the expression of its kinase CK1E in sporadic ALS. Acta Neuropathol 2018;136:405–23. https://doi.org/10.1007/s00401-018-1870-7.

Lage K, Hansen N, Karlberg E, Eklund A, Roque F, Donahoe P, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci U S A 2008;105:20870–5. https://doi.org/10.1073/PNAS.0810772105.

Lagier-Tourenne C, Polymenidou M, Hutt KR, Vu AQ, Baughn M, Huelga SC, et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nat

Neurosci 2012;15:1488–97. https://doi.org/10.1038/nn.3230.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati P V., et al. FlyBase: Updates to the Drosophila melanogaster knowledge base. Nucleic Acids Res 2021;49:D899–907. https://doi.org/10.1093/nar/gkaa1026.

Lau DHW, Hartopp N, Welsh NJ, Mueller S, Glennon EB, Mórotz GM, et al. Disruption of ER-mitochondria signalling in fronto-temporal dementia and related amyotrophic lateral sclerosis. Cell Death Dis 2018;9:1234567890. https://doi.org/10.1038/s41419-017-0022-7.

Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:R29.

Leader DP, Krause SA, Pandit A, Davies SA, Dow JATT. FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. Nucleic Acids Res 2018;46:809–815. https://doi.org/10.1093/nar/gkx976.

Lee CE, Singleton KS, Wallin M, Faundez V. Rare Genetic Diseases: Nature's Experiments on Human Development. IScience 2020;23. https://doi.org/10.1016/J.ISCI.2020.101123.

Lee LYH, Loscalzo J. Network Medicine in Pathobiology. Am J Pathol 2019;189:1311–26. https://doi.org/10.1016/j.ajpath.2019.03.009.

Li DK, Tisdale S, Lotti F, Pellizzoni L. SMN control of RNP assembly: from post-transcriptional gene regulation to motor neuron disease. Semin Cell Dev Biol 2014;0:22–9. https://doi.org/10.1016/j.semcdb.2014.04.026.SMN.

Li Y, Zhang Y, Li X, Yi S, Xu J. Gain-of-Function Mutations: An Emerging Advantage for Cancer Biology. Trends Biochem Sci 2019;44:659–74. https://doi.org/10.1016/j.tibs.2019.03.009.

Li YR, King OD, Shorter J, Gitler AD. Stress granules as crucibles of ALS pathogenesis. J Cell Biol 2013;201:361–372.

Liguori F, Amadio S, Volonté C. Fly for ALS: Drosophila modeling on the route to amyotrophic lateral sclerosis modifiers. Cell Mol Life Sci 2021;78:6143–6160.

Lin G, Mao D, Bellen HJ. Amyotrophic Lateral Sclerosis Pathogenesis Converges on Defects in Protein Homeostasis Associated with TDP-43 Mislocalization and Proteasome-Mediated Degradation Overload. Curr Top Dev Biol 2017;121:111–71. https://doi.org/10.1016/bs.ctdb.2016.07.004.

Lin W-H, Liu W-C, Hwang M-J. Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. BMC Syst Biol 2009;3. https://doi.org/10.1186/1752-0509-3-32.

Ling SC, Polymenidou M, Cleveland DW. Converging mechanisms in als and FTD: Disrupted RNA and protein homeostasis. Neuron 2013;79:416–38. https://doi.org/10.1016/j.neuron.2013.07.033.

Liu C, Ma Y, Zhao J, Nussinov R, Zhang YC, Cheng F, et al. Computational network biology: Data, models, and applications. Phys Rep 2020;846:1–66. https://doi.org/10.1016/j.physrep.2019.12.004.

Liu Z, Cheng X, Zhong S, Zhang X, Liu C, Liu F, et al. Peripheral and Central Nervous System Immune Response Crosstalk in Amyotrophic Lateral Sclerosis. Front Neurosci 2020;14:575. https://doi.org/10.3389/fnins.2020.00575.

Longinetti E, Fang F. Epidemiology of amyotrophic lateral sclerosis: An update of recent literature. Curr Opin Neurol 2019;32:771–6. https://doi.org/10.1097/WCO.0000000000000730.

Lopate G, Baloh RH, Al-Lozi MT, Miller TM, Fernandes Filho JA, Ni O, et al. Familial ALS with extreme phenotypic variability

due to the I113T SOD1 mutation. Amyotroph Lateral Scler 2010;11:232–6. https://doi.org/10.3109/17482960902898069.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

Low TY, Syafruddin SE, Mohtar MA, Vellaichamy A, A Rahman NS, Pung YF, et al. Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. Cell Mol Life Sci 2021;78:5325–39. https://doi.org/10.1007/s00018-021-03856-0.

Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol 2017;13. https://doi.org/10.1371/journal.pcbi.1005457.

Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8. https://doi.org/10.1038/s41586-020-2188-x.

Lundberg E, Borner GHH. Spatial proteomics: a powerful discovery tool for cell biology. Nat Rev Mol Cell Biol 2019;20:285–302. https://doi.org/10.1038/s41580-018-0094-y.

Ma CY, Liao CS. A review of protein–protein interaction network alignment: From pathway comparison to global alignment. Comput Struct Biotechnol J 2020;18:2647–56. https://doi.org/10.1016/j.csbj.2020.09.011.

Madabhushi R, Pan L, Tsai L-H. DNA Damage and Its Links to Neurodegeneration. Neuron 2014;83:266–82. https://doi.org/10.1016/J.NEURON.2014.06.034.

Maharjan N, Saxena S. ER strikes again: Proteostasis Dysfunction In ALS. EMBO J 2016;35:798–800. https://doi.org/10.15252/embj.201694117.

Mahi N Al, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. Sci Rep 2019;9:1–9. https://doi.org/10.1038/s41598-019-43935-8.

Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene Set Analysis: Challenges, Opportunities, and Future Research. Front Genet 2020;11:654. https://doi.org/10.3389/fgene.2020.00654.

Marín N, Esteban F, Ramírez-Rodrigo H, Ros E, Sáez-Lara M. An integrative methodology based on protein-protein interaction networks for identification and functional annotation of disease-relevant genes applied to channelopathies. BMC Bioinformatics 2019;20. https://doi.org/10.1186/S12859-019-3162-1.

Masrori P, Van Damme P. Amyotrophic lateral sclerosis: a clinical review. Eur J Neurol 2020;27:1918–29. https://doi.org/10.1111/ene.14393.

Mathis S, Goizet C, Soulages A, Vallat JM, Masson G Le. Genetics of amyotrophic lateral sclerosis: A review. J Neurol Sci 2019;399:217–26. https://doi.org/10.1016/j.jns.2019.02.030.

McAlary L, Chew YL, Lum JS, Geraghty NJ, Yerbury JJ, Cashman NR. Amyotrophic Lateral Sclerosis: Proteins, Proteostasis, Prions, and Promises. Front Cell Neurosci 2020;14. https://doi.org/10.3389/fncel.2020.581907.

McGuire SE, Le PT, Osborn AJ, Matsumoto K, Davis RL. Spatiotemporal Rescue of Memory Dysfunction in Drosophila. Science 2003; 302:1765–8.

McGurk L, Berson A, Bonini NM. Drosophila as an In Vivo Model for Human Neurodegenerative Disease. Genetics 2015;201:377–402.

McIver SC, Katsumura KR, Davids E, Liu P, Kang Y-A, Yang D, et al. Exosome complex orchestrates developmental signaling to balance proliferation and

differentiation during erythropoiesis. Elife 2016;5. https://doi.org/10.7554/eLife.17877.

Mejzini R, Flynn LL, Pitout IL, Fletcher S, Wilton SD, Akkari PA. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? Front Neurosci 2019;13. https://doi.org/10.3389/fnins.2019.01310.

Meltzer EB, Barry WT, D'Amico TA, Davis RD, Lin SS, Onaitis MW, et al. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: Proof-of-principle. BMC Med Genomics 2011;4. https://doi.org/10.1186/1755-8794-4-70.

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science 2015;347:841. https://doi.org/10.1126/science.1257601.

Menzies FM, Fleming A, Rubinsztein DC. Compromised autophagy and neurodegenerative diseases. Nat Rev Neurosci 2015;16:345–57. https://doi.org/10.1038/nrn3961.

Mercuri E, Pera MC, Scoto M, Finkel R, Muntoni F. Spinal muscular atrophy — insights and challenges in the treatment era. Nat Rev Neurol 2020 1612 2020;16:706–15. https://doi.org/10.1038/s41582-020-00413-4.

Mili S, Steitz JA. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. RNA 2004;10:1692–1694.

Moreira M-C, Klur S, Watanabe M, Németh AH, Le Ber I, Moniz J-C, et al. Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. Nat Genet 2004;36:225–7. https://doi.org/10.1038/ng1303.

Morera AA, Ahmed NS, Schwartz JC. TDP-43 regulates transcription at protein-coding genes and Alu retrotransposons. Biochim Biophys Acta Gene Regul Mech 2019;1862:194434.

Morton DJ, Kuiper EG, Jones SK, Leung SW, Corbett AH, Fasken MB. The RNA exosome and RNA exosome-linked disease. RNA 2018;24:127–42. https://doi.org/10.1261/rna.064626.117.

Mosca TJ. On the Teneurin track: A new synaptic organization molecule emerges. Front Cell Neurosci 2015;9. https://doi.org/10.3389/fncel.2015.00204.

Mubeen S, Hoyt CT, Gemünd A, Hofmann-Apitius M, Fröhlich H, Domingo-Fernández D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet 2019;10:1203. https://doi.org/10.3389/fgene.2019.01203.

Nadeau R, Byvsheva A, Lavallée-Adam M. PIGNON: a protein–protein interaction-guided functional enrichment analysis for quantitative proteomics. BMC Bioinformatics 2021;22:1–22. https://doi.org/10.1186/s12859-021-04042-6.

Di Nanni N, Bersanelli M, Milanesi L, Mosca E. Network Diffusion Promotes the Integrative Analysis of Multiple Omics. Front Genet 2020;11:106. https://doi.org/10.3389/fgene.2020.00106.

Nasim MT, Chernova TK, Chowdhury HM, Yue B-G, Eperon IC. HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. Hum Mol Genet 2003;12:1337–48.

Naylor S, Chen JY. Unraveling human complexity and disease with systems biology and personalized medicine. Per Med 2010;7:275–89. https://doi.org/10.2217/pme.10.16.

Nguyen HP, Van Broeckhoven C, van der Zee J. ALS Genes in the Genomic Era and their Implications for FTD. Trends Genet 2018;34:404–23. https://doi.org/10.1016/j.tig.2018.03.001.

Nygaard HB, Van Dyck CH, Strittmatter SM. Fyn kinase inhibition as a novel therapy for Alzheimer's disease. Alzheimer's Res Ther 2014;6:1–8.

https://doi.org/10.1186/ALZRT238/TABLES/2.

Ojala KS, Reedich EJ, DiDonato CJ, Meriney SD. In Search of a Cure: The Development of Therapeutics to Alter the Progression of Spinal Muscular Atrophy. Brain Sci 2021;11:1–39. https://doi.org/10.3390/BRAINSCI11020194.

Olesnicky EC, Wright EG. Drosophila as a model for assessing the function of RNA-binding proteins during neurogenesis and neurological disease. J Dev Biol 2018;6. https://doi.org/10.3390/jdb6030021.

Osterwalder T, Yoon KS, White BH, Keshishian H. A conditional tissue-specific transgene expression system using inducible GAL4. Proc Natl Acad Sci U S A 2001;98:12596–601. https://doi.org/10.1073/PNAS.221303298.

Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. J Med Genet 2006;43:691–8. https://doi.org/10.1136/jmg.2006.041376.

Ouali Alami N, Schurr C, Olde Heuvel F, Tang L, Li Q, Tasdogan A, et al. NF-κB activation in astrocytes drives a stage-specific beneficial neuroimmunological response in ALS. EMBO J 2018;37. https://doi.org/10.15252/embj.201798697.

Paddison PJ, Caudy AA, Bernstein E, Hannon GJ, Conklin DS. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev 2002;16:948–58. https://doi.org/10.1101/gad.981002.

Park J, Hescott BJ, Slonim DK. Pathway centrality in protein interaction networks identifies functional mediators of pulmonary disease. BioRxiv 2017:171942. https://doi.org/10.1101/171942.

Perera ND, Sheean RK, Crouch PJ, White AR, Horne MK, Turner BJ. Enhancing survival motor neuron expression extends lifespan and attenuates neurodegeneration in mutant TDP-43 mice. Hum Mol Genet 2016;25:4080–4093.

Piccin A, Salameh A, Benna C, Sandrelli F, Mazzotta G, Zordan M, et al. Efficient and heritable functional knock-out of an adult phenotype in Drosophila using a GAL4-driven hairpin RNA incorporating a heterologous spacer. Nucleic Acids Res 2001;29:55–55. https://doi.org/10.1093/nar/29.12.e55.

Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015;2015. https://doi.org/10.1093/database/bav028.

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48:D845–55. https://doi.org/10.1093/NAR/GKZ1021.

Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease–gene associations. Methods 2015;74:83–9. https://doi.org/10.1016/j.ymeth.2014.11.020.

Podder S, Mukhopadhyay P, Ghosh TC. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. Gene 2009;439:11–6. https://doi.org/10.1016/j.gene.2009.03.005.

Price PL, Morderer D, Rossoll W. RNP Assembly Defects in Spinal Muscular Atrophy. Adv Neurobiol 2018;20:143–171.

Purves D, Augustine GJ, Fitzpatrick D, Hall WC, LaMantia A-S, Mooney RD, et al. Neuroscience. 6th ed. New York: Sinauer Associates; 2018.

Qiu X, Zheng L, Liu X, Hong D, He M, Tang Z, et al. ULK1 Inhibition as a Targeted Therapeutic Strategy for Psoriasis by Regulating Keratinocytes and Their Crosstalk With Neutrophils. Front Immunol 2021;12:3096.

https://doi.org/10.3389/fimmu.2021.714274.

Quiñones-Frías MC, Littleton JT. Function of Drosophila Synaptotagmins in membrane trafficking at synapses. Cell Mol Life Sci 2021;78:4335–64. https://doi.org/10.1007/s00018-021-03788-9.

R Core Team. R: A Language and Environment for Statistical Computing 2020.

Ragagnin AMG, Shadfar S, Vidal M, Jamali MS, Atkin JD. Motor neuron susceptibility in ALS/FTD. Front Neurosci 2019;13. https://doi.org/10.3389/fnins.2019.00532.

Rai SN, Tiwari N, Singh P, Mishra D, Singh AK, Hooshmandi E, et al. Therapeutic Potential of Vital Transcription Factors in Alzheimer's and Parkinson's Disease With Particular Emphasis on Transcription Factor EB Mediated Autophagy. Front Neurosci 2021;15. https://doi.org/10.3389/fnins.2021.777347.

Ramesh N, Pandey UB. Autophagy dysregulation in ALS: When protein aggregates get out of hand. Front Mol Neurosci 2017;10:263. https://doi.org/10.3389/fnmol.2017.00263.

Rangaraju V, tom Dieck S, Schuman EM. Local translation in neuronal compartments: how local is local? EMBO Rep 2017;18:693–711. https://doi.org/10.15252/embr.201744045.

Ratti A, Buratti E. Physiological functions and pathobiology of TDP-43 and FUS/TLS proteins. J Neurochem 2016;138 Suppl:95–111.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–5. https://doi.org/10.1126/science.1073374.

Richard P, Feng S, Manley JL. A SUMO-dependent interaction between Senataxin and the exosome, disrupted in the neurodegenerative disease AOA2, targets the exosome to sites of transcription-induced DNA damage. Genes Dev 2013;27:2227–32. https://doi.org/10.1101/gad.224923.113.

Rinaldi C, Pizzul P, Longhese MP, Bonetti D. Sensing R-Loop-Associated DNA Damage to Safeguard Genome Stability. Front Cell Dev Biol 2021;8:1657. https://doi.org/10.3389/fcell.2020.618157.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47–e47. https://doi.org/10.1093/nar/gkv007.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616.

Rolland T, Tas M, Sahni N, Yi S, Lemmens I, Fontanillo C, et al. Resource a proteome-scale map of the human interactome network. Cell 2014;159:1212–26. https://doi.org/10.1016/j.cell.2014.10.050.

RStudio Team. RStudio: Integrated Development for R. Inc, Boston, MA 2016.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005;437:1173–8. https://doi.org/10.1038/nature04209.

Sakai Y, Shaw CA, Dawson BC, Dugas D V., Al-Mohtaseb Z, Hill DE, et al. Protein Interactome Reveals Converging Molecular Pathways Among Autism Disorders. Sci Transl Med 2011;3:86ra49. https://doi.org/10.1126/scitranslmed.3002166.

Salvi JS, Mekhail K. R-loops highlight the nucleus in ALS. Nucleus 2015;6:23–9. https://doi.org/10.1080/19491034.2015.1004952.

San Gil R, Ooi L, Yerbury JJ, Ecroyd H. The heat shock response in neurons and astroglia and its role in neurodegenerative diseases. Mol Neurodegener 2017;12:1–20. https://doi.org/10.1186/s13024-017-0208-6.

dos Santos G, Schroeder A, Goodman J, Strelets V, Crosby M, Thurmond J, et al. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res 2015;43:D690–697.

Sarkans U, Füllgrabe A, Ali A, Athar A, Behrangi E, Diaz N, et al. From ArrayExpress to BioStudies. Nucleic Acids Res 2021;49:D1502–6. https://doi.org/10.1093/nar/gkaa1062.

Schorling DC, Pechmann A, Kirschner J. Advances in Treatment of Spinal Muscular Atrophy - New Phenotypes, New Challenges, New Implications for Care. J Neuromuscul Dis 2020;7:1–13. https://doi.org/10.3233/JND-190424.

Scialo F, Sriram A, Stefanatos R, Sanz A. Practical Recommendations for the Use of the GeneSwitch Gal4 System to Knock-Down Genes in Drosophila melanogaster. PLoS One 2016;11:161817.

Sen A, Dimlich DN, Guruharsha KG, Kankel MW, Hori K, Yokokura T, et al. Genetic circuitry of Survival motor neuron, the gene underlying spinal muscular atrophy. Proc Natl Acad Sci U S A 2013;110. https://doi.org/10.1073/pnas.1301738110.

Shah SZA, Zhao D, Hussain T, Yang L. The role of unfolded protein response and mitogen-activated protein kinase signaling in neurodegenerative diseases with special focus on prion diseases. Front Aging Neurosci 2017;9:120. https://doi.org/10.3389/fnagi.2017.00120.

Sharma P, Alizadeh J, Juarez M, Samali A, Halayko AJ, Kenyon NJ, et al. Autophagy, apoptosis, the unfolded protein response, and lung function in idiopathic pulmonary fibrosis. Cells 2021;10. https://doi.org/10.3390/cells10071642.

Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. PNAS 2014;111:E5593-601. https://doi.org/10.1073/pnas.1419161111.

Shkreta L, Chabot B. The RNA Splicing Response to DNA Damage. Biomolecules 2015;5:2935–77. https://doi.org/10.3390/biom5042935.

Shou C, Bhardwaj N, Lam HYK, Yan KK, Kim PM, Snyder M, et al. Measuring the evolutionary rewiring of biological networks. PLoS Comput Biol 2011;7:e1001050. https://doi.org/10.1371/journal.pcbi.1001050.

Shukla S, Tekwani BL. Histone Deacetylases Inhibitors in Neurodegenerative Diseases, Neuroprotection and Neuronal Differentiation. Front Pharmacol 2020;11:537. https://doi.org/10.3389/fphar.2020.00537.

Siddique N, Siddique T. Genetics of Amyotrophic Lateral Sclerosis. Phys Med Rehabil Clin N Am 2008;19:429–39. https://doi.org/10.1016/j.pmr.2008.05.001.Genetics.

Silverman EK, Schmidt HHHW, Anastasiadou E, Altucci L, Angelini M, Badimon L, et al. Molecular networks in Network Medicine: Development and applications. Wiley Interdiscip Rev Syst Biol Med 2020;12:e1489. https://doi.org/10.1002/wsbm.1489.

Sircar S, Parekh N. Functional characterization of drought-responsive modules and genes in Oryza sativa: A network-based approach. Front Genet 2015;6:256. https://doi.org/10.3389/fgene.2015.00256.

Skinnider MA, Stacey RG, Foster LJ. Genomic data integration systematically biases interactome mapping. PLoS Comput Biol 2018;14.

https://doi.org/10.1371/journal.pcbi.1006474.

Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human Senataxin Resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-Dependent termination. Mol Cell 2011;42:794–805. https://doi.org/10.1016/j.molcel.2011.04.026.

Smith EF, Shaw PJ, De Vos KJ. The role of mitochondria in amyotrophic lateral sclerosis. Neurosci Lett 2019;710:132933. https://doi.org/10.1016/j.neulet.2017.06.052.

del Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. Curr Opin Biotechnol 2010;21:566–71. https://doi.org/10.1016/j.copbio.2010.07.010.

Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, et al. Understanding Tissue-Specific Gene Regulation. Cell Rep 2017;21:1077–88. https://doi.org/10.1016/j.celrep.2017.10.001.

Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, et al. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics 1999;153:135–177.

Spring AM, Raimer AC, Hamilton CD, Schillinger MJ, Matera AG. Comprehensive Modeling of Spinal Muscular Atrophy in Drosophila melanogaster. Front Mol Neurosci 2019;12:113.

Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet 2019;20:631–56. https://doi.org/10.1038/s41576-019-0150-2.

Stogsdill JA, Eroglu C. The interplay between neurons and glia in synapse development and plasticity. Curr Opin Neurobiol 2017;42:1–8. https://doi.org/10.1016/j.conb.2016.09.016.

Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights 2020;14. https://doi.org/10.1177/1177932219899051.

Sun S, Ling S-C, Qiu J, Albuquerque CP, Zhou Y, Tokunaga S, et al. ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. Nat Commun 2015;6:6171. https://doi.org/10.1038/ncomms7171.

Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 2011;6:e21800. https://doi.org/10.1371/journal.pone.0021800.

Taylor JP, Brown RH, Cleveland DW. Decoding ALS: From genes to mechanism. Nature 2016;539:197–206. https://doi.org/10.1038/nature20413.

Tirode F, Busso D, Coin F, Egly J-M. Reconstitution of the Transcription Factor TFIIH: Assignment of Functions for the Three Enzymatic Subunits, XPB, XPD, and cdk7. Mol Cell 1999;3:87–95. https://doi.org/10.1016/S1097-2765(00)80177-X.

Tisdale S, Lotti F, Saieva L, VanMeerbeke JP, Crawford TO, Sumner CJ, et al. SMN is essential for the biogenesis of U7 Small nuclear ribonucleoprotein and 3'-end formation of Histone mRNAs. Cell Rep 2013;5:1187–95. https://doi.org/10.1016/j.celrep.2013.11.012.

Tsuiji H, Iguchi Y, Furuya A, Kataoka A, Hatsuta H, Atsuta N, et al. Spliceosome integrity is defective in the motor neuron diseases ALS and SMA. EMBO Mol Med 2013;5:221–34. https://doi.org/10.1002/emmm.201202303.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science 2015;347. https://doi.org/10.1126/science.1260419.

Valentini G, Paccanaro A, Caniza H, Romero AE, Re M. An extensive analysis of

disease-gene associations using network integration and fast kernel-based gene prioritization methods. Artif Intell Med 2014;61:63–78. https://doi.org/10.1016/j.artmed.2014.03.003.

Vallianatos CN, Iwase S. Disrupted intricacy of histone H3K4 methylation in neurodevelopmental disorders. Epigenomics 2015;7:503–19. https://doi.org/10.2217/epi.15.1.

Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 2010;6:e1000641. https://doi.org/10.1371/journal.pcbi.1000641.

Vella D, Marini S, Vitali F, Di Silvestre D, Mauri G, Bellazzi R. MTGO: PPI Network Analysis Via Topological and Functional Module Identification. Sci Rep 2018;8:5499. https://doi.org/10.1038/s41598-018-23672-0.

Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al. An empirical framework for binary interactome mapping. Nat Methods 2009;6:83–90. https://doi.org/10.1038/nmeth.1280.

Verhaart IEC, Robertson A, Wilson IJ, Aartsma-Rus A, Cameron S, Jones CC, et al. Prevalence, incidence and carrier frequency of 5q–linked spinal muscular atrophy – a literature review. Orphanet J Rare Dis 2017;12. https://doi.org/10.1186/S13023-017-0671-8.

Vijayakumar J, Perrois C, Heim M, Bousset L, Alberti S, Besse F. The prion-like domain of Drosophila Imp promotes axonal transport of RNP granules in vivo. Nat Commun 2019;10:2593.

van der Voet M, Nijhof B, Oortveld MAW, Schenck A. Drosophila models of early onset cognitive disorders and their clinical applications. Neurosci Biobehav Rev 2014;46:326–42. https://doi.org/10.1016/j.neubiorev.2014.01.013.

Voigt A, Herholz D, Fiesel FC, Kaur K, Müller D, Karsten P, et al. TDP-43-Mediated Neuron Loss In Vivo Requires RNA-Binding Activity. PLoS One 2010;5:e12247. https://doi.org/10.1371/journal.pone.0012247.

Wang JYJ. The Capable ABL: What Is Its Biological Function? Mol Cell Biol 2014;34:1188–97. https://doi.org/10.1128/mcb.01454-13.

Wang M, Chen PY, Wang CH, Lai TT, Tsai PI, Cheng YJ, et al. Dbo/Henji Modulates Synaptic dPAK to Gate Glutamate Receptor Abundance and Postsynaptic Response. PLoS Genet 2016;12:e1006362.

Wang M, Zhao Y, Zhang B. Efficient Test and Visualization of Multi-Set Intersections. Sci Rep 2015;5:16923.

Wang PI, Marcotte EM. It's the machine that matters: Predicting gene function and phenotype from protein networks. J Proteomics 2010;73:2277–89. https://doi.org/10.1016/j.micinf.2011.07.011.Innate.

Wang W-Y, Pan L, Su SC, Quinn EJ, Sasaki M, Jimenez JC, et al. Interaction of FUS and HDAC1 regulates DNA damage response and repair in neurons. Nat Neurosci 2013;16:1383–91. https://doi.org/10.1038/nn.3514.

Wang WM, Jin HZ. Heat shock proteins and psoriasis. Eur J Dermatology 2019;29:121–5. https://doi.org/10.1684/ejd.2019.3526.

Wang X, Wei X, Thijssen B, Das J, Lipkin S, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 2012;30:159–64. https://doi.org/10.1038/NBT.2106.

Wang Z, Zhang J. In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 2007;3:1011–21. https://doi.org/10.1371/journal.pcbi.0030107.

Wickham H. ggplot2: Elegant Graphics for Data Analysis 2016.

Will T, Helms V. PPIXpress: Construction of condition-specific protein interaction networks based on transcript expression. Bioinformatics 2016;32:571–8. https://doi.org/10.1093/bioinformatics/btv620.

Workman E, Kolb SJ, Battle DJ. Spliceosomal small nuclear ribonucleoprotein biogenesis defects and motor neuron selectivity in spinal muscular atrophy. Brain Res 2012;1462:93–99.

Wu Q, Beland FA, Chang C-W, Fang J-L. Role of DNA Repair Pathways in Response to Zidovudine-induced DNA Damage in Immortalized Human Liver THLE2 Cells. Int J Biomed Sci 2013;9:18–25.

Xia R, Liu Y, Yang L, Gal J, Zhu H, Jia J. Motor neuron apoptosis and neuromuscular junction perturbation are prominent features in a Drosophila model of Fus-mediated ALS. Mol Neurodegener 2012;7:10.

Xu X, Shen D, Gao Y, Zhou Q, Ni Y, Meng H, et al. A perspective on therapies for amyotrophic lateral sclerosis: can disease progression be curbed? Transl Neurodegener 2021;10:1–18. https://doi.org/10.1186/s40035-021-00250-5.

Yamazaki T, Chen S, Yu Y, Yan B, Haertlein TC, Carrasco MA, et al. FUS-SMN Protein Interactions Link the Motor Neuron Diseases ALS and SMA. Cell Rep 2012;2:799–806. https://doi.org/10.1016/j.celrep.2012.08.025.

Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell 2016;164:805–17. https://doi.org/10.1016/j.cell.2016.01.029.

Yeger-Lotem E, Sharan R. Human protein interaction networks across tissues and diseases. Front Genet 2015:257. https://doi.org/10.3389/FGENE.2015.00257.

Yi Y, Zhao Y, Huang Y, Wang D. A brief review of RNA-Protein interaction database resources. Non-Coding RNA 2017;3. https://doi.org/10.3390/ncrna3010006.

Yu G. Gene ontology semantic similarity analysis using GOSemSim. Methods Mol. Biol., vol. 2117, Humana, New York, NY; 2020, p. 207–15. https://doi.org/10.1007/978-1-0716-0301-7_11.

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010;26:976–8. https://doi.org/10.1093/bioinformatics/btq064.

Yu G, Lu C, Wang J. NoGOA: Predicting noisy GO annotations using evidences and sparse representation. BMC Bioinformatics 2017;18. https://doi.org/10.1186/s12859-017-1764-z.

Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.

Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, et al. Next-generation sequencing to generate interactome datasets. Nat Methods 2011;8:478–80. https://doi.org/10.1038/nmeth.1597.

Zaepfel BL, Rothstein JD. RNA Is a Double-Edged Sword in ALS Pathogenesis. Front Cell Neurosci 2021;15:708181.

Zbinden A, Pérez-Berlanga M, Rossi P, Polymenidou M. Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force. Dev Cell 2020;55:45–68.

Zhao DY, Gish G, Braunschweig U, Li Y, Ni Z, Schmitges FW, et al. SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. Nature 2016;529:48–53. https://doi.org/10.1038/nature16469.

Zhong Q, Simonis N, Li Q, Charloteaux B, Heuze F, Klitgord N, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol 2009;5. https://doi.org/10.1038/MSB.2009.80.

Zolotareva O, Kleine M. A Survey of Gene Prioritization Tools for Mendelian and Complex Human Diseases. J Integr Bioinform 2019;16. https://doi.org/10.1515/jib-2018-0069.