Elucidating TDP-43 Cryptic Exons and Short Tandem Repeat Expansions: Molecular Mechanisms and Epidemiological Insights in Neurodegenerative Disease

Doctoral thesis

Ph.D. in Neuroscience

Student:

Matteo Zanovello

[redacted]

Supervisors:

Prof Pietro Fratta

Prof Linda Greensmith

Prof Pedro Machado

Department of Neuromuscular Diseases

UCL Queen Square Institute of Neurology

I, Matteo Zanovello, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Mors ubi gaudet succurrere vitae

Acknowledgements

How do I start this? The journey has been fairly long, but never too harsh. When I arrived, I was completely unaware of everything about lab work, and I had a very shallow sense of bioinformatics. I'm not sure how much better I am now, but I'm sure that I could never have got here without the help and support from the fantastic people around me.

First, I have to mention the Admin team who could answer all my questions and help me cope with my bureaucratic anxiety: thank you, Kully, Debbie, Irina, Adrian, and everyone. To Queen Square House security, despite the abuse you face every day, you were always there when I needed you. Thank you!

Thank you to the Neurogenetics team (7th and 9th floors, particularly to Kristina, Annarita, Clarissa, Elisa, Delia and Valentina) for the helpful discussion and for helping me understand the genetics of neurological disorders a bit more. I would never have discovered what doing science means without the discussions I had with Gipi, James, and Linda's labs. A particular vote of thanks to "el Papi" Jobert, one of the most creative scientists I met. I'm sure you're going to do great in science. To the awesome people at QS Brain Bank, and the endless times someone had to get me in because I didn't have access. To the Crick, always providing pleasant surroundings for making science, and the lovely people I met there.

To the fantastic collaborators, including but not limited to Michael, Andy, Emanuele, Len, Jan, Bryan, Jack, Jenny, Shawn, and Steven, I learned from you all.

To Gianni, without whom I wouldn't be here. I hope to absorb your clinical sense, to inform my clinical research and push translational research further.

And now, the 4th floor, or I should say, home. I spent days and nights (not many to be fair) there, and you were there to host my bike when I forgot my lock. My experience at UCL wouldn't be the same without you, Lizzy. I greatly appreciated the wise suggestions, the chance to attend fancy dining, and your book on how to write a thesis - I hope I followed it. To Adrian's lab, especially Carmelo. You've become a scientific partner in crime, and I'm sure our paths will cross again.

And now, the FrattaLab! To Martha, for being there from day 0, and patiently

teaching me very basic wet lab stuff. To AL, Sam, and more recently, Alla, for all the dry lab stuff and for your patience in replying kindly to my very dumb questions. To the Kennedy's disease team, Annalucia and Wenanlan, even if we never worked together, you've always been supportive. To Puja, thank you for all the help with medical questions and English pronunciation. To Eugeni, to be the wildest and most fun person ever. To Becca and Pete, for all the pub adventures (and Guinness and Negronis and Vermut...). To Sara, I truly enjoyed discussing science with you, even if you joined the lab only recently. To Oscar, Max, Lea, and Auri, for teaching me so much about those four quite important letters (could it be AGCT?). To Flami, for bearing me and my stupid ideas. To Ari, my favourite Maltese in the world, and my private pathology teacher. To Michela and Ale, the Superchicche, I will miss you a lot (but will host you whenever you want). To Dario, I can never be grateful enough for all the discussions about RNA! To Nicol and Fra, official representatives for the Republic of Venice in London, I will miss the pizza and beer in the office, the puns, and every moment spent with you. To Simo, I have known you for only a year, but I consider you almost a brother, and I hope to share more fun and science in the future.

Thanks, Luca, mayor of London, for always being there for me, and for teaching me much more than you can imagine. To Linda and Pedro, Ash, Bilal, and Matt, for being the core support of the TDP-43 project and my whole journey here.

To Arianna, for being unofficially my supervisor, mentor, and coffee-break partner. I struggle to find words to describe how much you have helped me, from checking my calculations and stats to providing an energising breakfast to survive a day of paper writing.

To Pietro, for being officially my supervisor, mentor, and opponent in ping pong. I would probably need more words than I'm allowed for the whole thesis to express my gratitude. Thank you for teaching me when to speak and when not, how to write convincingly, which experiments to pursue, which ones to ask for more expert help, and which ones to avoid.

To all the friends in Italy, London, and around the world, for making this adventure brighter.

To my amazing parents, Elisa and Marcello, for teaching me what love can do.

To my siblings, Francesco, Marco, Sofia, and Angela. Even if our lives have taken different paths, we will always be connected by all the experiences we had. You taught me that life is a mess, but with the right team, you're sound.

Last, to my grandmas, Adriana and Tatiana, for being the two wise giants whose shoulders I stand on.

Abstract

TDP-43 cytosolic aggregation and nuclear depletion occur in amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD) and other neurodegenerative diseases, collectively called TDP-43 proteinopathies. Nuclear loss of TDP-43 causes the derepression of cryptic exons (CEs), often leading to transcript degradation through nonsense-mediated decay (NMD). While CEs play a pathogenic role in ALS/FTD, a thorough understanding of the biology of TDP-43 CEs and their relevance for other neurodegenerative diseases is still lacking.

Short tandem repeat (STR) expansions are responsible for several neurodegenerative diseases, with some of them exhibiting TDP-43 pathology. Notwithstanding their role in neurodegeneration, their prevalence is unknown.

In the first chapter, by leveraging novel bioinformatic tools on large whole-genome sequencing datasets, I uncover a surprisingly high occurrence of STR expansions in the general population and then estimate the prevalence and instability features of STR diseases by mathematical modelling.

In the second chapter of this thesis, I characterise the biology of TDP-43 CEs by experimental and computational approaches. First, by inhibiting NMD, I demonstrate that while some CEs evade NMD and could serve as biomarkers, others are revealed only upon NMD inhibition. Later, I analyse how TDP-43 loss affects CE expression. I orthogonally validate these findings by iCLIP and postmortem RNA-sequencing.

Sporadic inclusion body myositis (sIBM) is a muscle disease whose pathogenesis consists of inflammatory and degenerative features. In the third chapter, I examine the role of TDP-43 and its CEs in vivo by transcriptomics and proteomics analysis of sIBM biopsies, where TDP-43 aggregates in the skeletal muscle, linking CEs expression to adaptive immune responses.

Overall, this work advances understanding of TDP-43 CE biology and identifies novel disease effectors and potential biomarkers for TDP-43 proteinopathies. Moreover, it sheds light on the prevalence of STR and related diseases, with important implications for genetic and clinical medicine.

Impact statement

Amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) are two incurable neurodegenerative diseases with shared genetics and pathology. Notwithstanding the advances in the field in recent decades, there is still a lack of understanding of the molecular biology underpinning these illnesses. On the other hand, short tandem repeat (STR) expansion disorders are also responsible for a spectrum of neurological manifestations; however, their prevalence estimates have been based on small cohort studies, while their prevalence in the general population is unknown. This thesis aims to examine the consequences of TDP-43 loss of function and to assess the prevalence of STRs in the general population, addressing several gaps in the field of neurodegenerative diseases.

Using in vitro models, clinical samples, and large biomedical repositories, and leveraging both molecular laboratory techniques and bioinformatics tools, I gained insights into the molecular functions and population genomics of a spectrum of neurodegenerative diseases. Overall, this work uncovers new candidate therapeutics and disease markers for TDP-43 proteinopathies and sets the stage for a more modern and comprehensive assessment of STRs. These findings have broad applications, including the development of novel biomarkers, the opportunity to modulate the immune response in neurodegeneration, and improved genetic counselling for neurological diseases. Additionally, these insights can be transferred to other neurological diseases, where the molecular processes described here also occur. In particular, it will be essential to perform a thorough characterisation of CEs in Alzheimer's disease and other rare disorders with TDP-43 pathology, as well as to implement the use of large datasets for other diseases.

From an academic perspective, three peer-reviewed papers have been published from this work. As of the 30th July 2025, *"TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A"* has been cited 390 times, *"Unexpected frequency of the pathogenic AR CAG repeat expansion in the general population"* has been cited 22 times, and *"Increased frequency of repeat expansion mutations across different populations"* has been cited 32 times. Based on the last two chapters

of this thesis, two manuscripts are currently under preparation. From the work described in this thesis, four new sets of RNA-seq (Chapter 4) and a novel RNA-seq and proteomics dataset (Chapter 5) have been generated. Upon publication of the related papers, this data will be made available to the scientific community. I'm pleased that the sequencing data generated for this thesis has been used in other theses and preprints. Most of the bioinformatic analyses used in this thesis have been made publicly available and can be found at the end of the Methods section. This suggests that these studies are capable of addressing key crucial questions and that these analyses may have a significant impact on the field.

Overall, this work offers insight into molecular alterations in neurodegeneration and the prevalence of genetic mutations in the general population. These have not only contributed valuable tools and knowledge to the field but have also laid the groundwork for future discoveries.

UCL Research paper declaration form(s)

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1038/s41586-022-04436-3

   (c) **Where was the work published?** Nature

   (d) **Who published the work?** Springer Nature

   (e) **When was the work published?** 23 February 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Anna-Leigh Brown, Oscar G. Wilkins, Matthew J. Keuss, Sarah E. Hill, Matteo Zanovello, Weaverly Colleen Lee, Alexander Bampton, Flora C. Y. Lee, Laura Masino, Yue A. Qi, Sam Bryce-Smith, Ariana Gatt, Martina Hallegger, Delphine Fagegaltier, Hemali Phatnani, NYGC ALS Consortium, Jia Newcombe, Emil K. Gustavsson, Sahba Seddighi, Joel F. Reyes, Steven L. Coon, Daniel Ramos, Giampietro Schiavo, Elizabeth M. C. Fisher, Towfique Raj, Maria Secrier, Tammaryn Lashley, Jernej Ule, Emanuele Buratti, Jack Humphrey, Michael E. Ward & Pietro Fratta

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** Yes. https://www.biorxiv.org/content/10.1101/2021.04.02.438170v1.full.pdf
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

If 'Yes', please please give a link or doi:

(c) **Where is the work intended to be published?**

(d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** Conceptualization: A.-L.B., O.G.W., M.J.K., S.E.H., J.H., M.E.W. and P.F. Data curation: A.-L.B., O.G.W., M.Z., W.C.L. and S.B.-S. Formal analysis: A.-L.B., O.G.W., M.J.K., M.Z., W.C.L., S.B.-S., A.B., M.H., E.K.G., S.S., J.F.R., S.L.C. and D.R. Funding acquisition: P.F., M.E.W. and E.B. Investigation: A.-L.B., O.G.W., M.J.K., S.E.H., M.Z., W.C.L., F.C.Y.L., L.M., Y.A.Q., S.B.-S., A.B., A.G., M.H., E.K.G., S.S., J.F.R., S.L.C., D.R., E.K.G. and S.L.C. Methodology: A.-L.B., O.G.W., M.J.K., S.E.H., J.H., M.E.W. and P.F. Project administration: P.F. and M.E.W. Resources: H.P., T.L., E.B., D.F. and J.N. Software: A.-L.B., O.G.W., M.Z., S.B.-S., J.H. and M.J.K. Supervision: P.F., M.E.W., J.H., J.U., M.S., T.R., T.L., E.M.C.F., G.S. and T.F. Visualization: A.-L.B., O.G.W., M.J.K., W.C.L. and S.E.H. Writing, original draft: A.-L.B., O.G.W., M.J.K., M.E.W. and P.F. Writing, review and editing: S.E.H., W.C.L., E.B., J.U., J.H., M.E.W., P.F. A.-L.B., O.G.W. M.J.K. and S.E.H. contributed equally; therefore each may place their name first in author order when referencing this manuscript in personal communications.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 1

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Matteo Zanovello
Date: 03/09/2024

Supervisor/Senior Author signature (where appropriate)**:** Pietro Fratta
Date: 03/09/2024

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Unexpected frequency of the pathogenic AR CAG repeat expansion in the general population

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1093/brain/awad050

   (c) **Where was the work published?** Brain

   (d) **Who published the work?** Oxford University Press

   (e) **When was the work published?** 17 February 2023

   (f) **List the manuscript's authors in the order they appear on the publication:** Matteo Zanovello, Kristina Ibáñez, Anna-Leigh Brown, Prasanth Sivakumar, Alessandro Bombaci, Liana Santos, Joke J F A van Vugt, Giuseppe Narzisi, Ramita Karra, Sonja W Scholz, Jinhui Ding, J Raphael Gibbs, Adriano Chiò, Clifton Dalgard, Ben Weisburd, The American Genome Center (TAGC) consortium, Genomics England Research Consortium, Project MinE ALS Sequencing Consortium, The NYGC ALS Consortium, Michael G Hanna, Linda Greensmith, Hemali Phatnani, Jan H Veldink, Bryan J Traynor, James Polke, Henry Houlden, Pietro Fratta, Arianna Tucci

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** No
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☒ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**
   If 'Yes', please please give a link or doi:

   (c) **Where is the work intended to be published?**

(d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** Conceptualization: M.Z., K.I., P.F. and A.T. Data curation: M.Z., K.I., A-L.B., P.S., A.B., J.J.F.A.v.V., G.N., R.K., B.W. and A.T. Formal analysis: M.Z., K.I., A-L.B., P.S., J.J.F.A.v.V., G.N., R.K., B.W. and A.T. Funding acquisition: M.G.H., L.G., J.H.V., B.J.T., P.F. and A.T. Investigation: M.Z., K.I., A.-L.B., P.S., A.B., L.S., J.J.F.A.v.V., G.N., R.K. and B.W. Methodology: M.Z., K.I., P.F. and A.T. Project administration: P.F. and A.T. Resources: S.W.S., J.D., J.R.G., A.C., C.D., B.W., H.P., J.H.V., B.J.T., J.P., H.H., P.F. and A.T. Software: M.Z., K.I., A-L.B., P.S. and A.T. Supervision: P.F. and A.T. Validation: M.Z., K.I. and A.T. Visualization: M.Z. and K.I. Writing, original draft: P.F. Writing, review and editing: M.Z., P.F. and A.T.
4. **In which chapter(s) of your thesis can this material be found?** Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Matteo Zanovello
Date: 03/09/2024

Supervisor/Senior Author signature (where appropriate)**:** Pietro Fratta
Date: 03/09/2024

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Increased frequency of repeat expansion mutations across different populations

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1038/s41591-024-03190-5

   (c) **Where was the work published?** Nature Medicine

   (d) **Who published the work?** Springer Nature

   (e) **When was the work published?** 01 October 2024

   (f) **List the manuscript's authors in the order they appear on the publication:** Kristina Ibañez, Bharati Jadhav, Matteo Zanovello, Delia Gagliardi, Christopher Clarkson, Stefano Facchini, Paras Garg, Alejandro Martin-Trujillo, Scott J Gies, Valentina Galassi Deforie, Anupriya Dalmia, Davina J. Hensman Moss, Jana Vandrovcova, Clarissa Rocca, Loukas Moutsianas, Chiara Marini-Bettolo, Helen Walker, Chris Turner, Maryam Shoai, Jeffrey D Long, EUROSCA network, Pietro Fratta, Douglas R Langbehn, Sarah J Tabrizi, Mark J Caulfield, Andrea Cortese, Valentina Escott-Price, John Hardy, Henry Houlden, Andrew J Sharp, Arianna Tucci

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** Yes. https://doi.org/10.1101/2023.07.03.23292162
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**
   If 'Yes', please please give a link or doi:

(c) **Where is the work intended to be published?**

(d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** K.I. and A.T. conceptualised the research project. K.I., B.J., M.Z., D.G., C.C., S.F., P.G., L.M., M.S., V.E.P., A.T. and A.J.S. conducted data analysis, interpreted statistical findings and created visual representations of the data. K.I., B.J., M.Z., P.G., A.M.T., S.J.G., V.G.D., D.J.H.M., C.R. and A.T. performed quality check visually inspecting pileup plots corresponding to repeat expansions. J.V. analysed the carrier frequency for RFC1 within the 100K GP. C.M.B., H.W., C.T., S.T., J.D.L. and D.R.L. provided data from HD and DM1 patient registries. A.C. provided valuable insights into the genetics of RFC1. Funding and supervision: A.T., A.J.S., P.F., M.J.C., J.H., H.H. Writing—original draft: K.I., A.T., A.J.S., A.D., D.J.H.M., M.Z., C.R., and M.S. meticulously reviewed and edited the manuscript for clarity, accuracy, and coherence. All authors carefully reviewed the manuscript, offering pertinent feedback that enhanced the study's quality, and ultimately approved the final version.
4. **In which chapter(s) of your thesis can this material be found?** Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Matteo Zanovello
Date: 03/10/2024

Supervisor/Senior Author signature (where appropriate)**:** Arianna Tucci
Date: 03/10/2024

Contents

List of Figures

List of Tables

List of Abbreviations

NDD = NeuroDegenerative Disease
TDP-43 = Transactive response DNA-binding Protein, 43 kDa
STR = Short Tandem Repeat
ALS = Amyotrophic Lateral Sclerosis
FTD = FrontoTemporal Dementia
sIBM = sporadic Inclusion Body Myositis
SBMA = Spinal and Bulbar Muscular Atrophy
SCA = SpinoCerebellar Ataxia
ALS-FRS-R = ALS Functional Rating Scale Revised
CSF = CerebroSpinal Fluid
MRI = Magnetic Resonance Imaging
PET = Positron Emission Tomography
ALS-eci = ALS with evidence of executive dysfunction
ALS-neci = ALS with no executive dysfunction but impairment in other cognitive domains
ALS-bi = ALS with behavioural changes
FTD-MND = FrontoTemporal Dementia with Motor Neurone Disease
PPA = Primary Progressive Aphasia
DPR = Dipeptide Repeat Protein
RBP = RNA Binding Protein
RED = Repeat Expansion Disorder
AD = Autosomal Dominant
AR = Autosomal Recessive
XD = X-linked Dominant
AD = Alzheimer's Disease
HD = Huntington's Disease
MND = Motor Neurone Disease
AR = Androgen Receptor
CNS = Central Nervous System

BBB = Blood-Brain Barrier
PBMC = Peripheral Blood Mononucleated Cell
ERAD = Endoplasmic Reticulum-Associated protein Degradation
H&E = Hematoxylin and Eosin
LATE = Limbic-predominant Age-related TDP-43 Encephalopathy
IBMPDF/ALS = Inclusion Body Myositis with Paget's Disease and Frontotemporal dementia/Amyotrophic Lateral Sclerosis
TARDBP = TransActive Response DNA-Binding Protein
NTD = N-Terminal Domain
NLS = Nuclear Localisation Signal
CTD = C-Terminal Domain
3'-UTR = 3'-UnTranslated Region
CE = Cryptic Exon
PTC = Premature Stop Codon
NMD = Nonsense-Mediated Decay
WGS = Whole Genome Sequencing
EH = ExpansionHunter
QC = Quality Check
100K GP = 100,000 Genomes Project
TOPMed = Trans-Omics for Precision Medicine
AFR = African
AMR = Admixed American
EAS = East Asian
EUR = European
SAS = South Asian
gVCF = genomic Variant Call Format
1K GP3 = 1000 Genomes Project phase 3
FAM = Fluorescein Amidite
DM1 = Myotonic Dystrophy type 1
CHX = CycloHeXimide
RIN = RNA Integrity Number
PSI = Percent Spliced In
HPC = High Performance Cluster
SRA = Sequence Read Archive
NYGC = New York Genome Center
FTD-TDP = FTD with TDP-43 inclusions
FTD-non-TDP = FTD without TDP-43 inclusions

NHNN = National Hospital for Neurology and Neurosurgery
DIA = Data Independent Acquisition
CI = Confidence Interval
ICD-10 = International Statistical Classification of Diseases and Related Health Problems 10th Revision
HPO = Human Phenotype Ontology
HDL2 = Huntington's Disease-Like 2
DRPLA = DentatoRubral-PallidoLuysian Atrophy
CANVAS = Cerebellar ataxia, neuropathy, vestibular areflexia syndrome
PCA = Principal Component Analysis
FACS = Fluorescence-Activated Cell Sorting
CLIP = Cross-Linking and Immune Precipitation
TCR = T-Cell Receptor
pMHC = peptide-Major Histocompatibility Complex

Chapter 1

Introduction

1.1 Hallmarks of neurodegenerative diseases

Neurodegenerative diseases (NDDs) affect millions of people worldwide and consist of several disorders affecting the nervous system, especially the neurons (Wilson et al., 2023). Because these cells are terminally differentiated, their structural and functional impairment results in irreversible motor, sensory, and/or cognitive-behavioural changes. In the last decades, advances in molecular biology and genetics have enabled physicians and researchers to characterise several hallmarks of NDDs, namely pathological protein aggregation, synaptic and neuronal network dysfunction, aberrant proteostasis, cytoskeletal abnormalities, altered energy metabolism, DNA and RNA defects, inflammation, and neuronal cell death (Figure 1).

Due to the breadth of the field, and acknowledging that these hallmarks are interwoven, this work will focus on a subset of them, particularly pathological protein aggregation, DNA and RNA defects, and inflammation. Mechanistically, I analysed a subset of the myriads of pathways that can lead to NDDs: in the first result chapter, I analyse how short tandem repeats (STRs) are distributed in the general population. In the other two result chapters, I assess how the nuclear loss of the transactive response DNA-binding protein, 43 kDa (TDP-43), intrinsically related to its pathological aggregation, leads to defects in RNA splicing, which are also potentially linked to inflammation.

From a clinical/phenotypic perspective, I focus on the following NDDs: amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), sporadic inclusion body myositis (sIBM), Spinal and Bulbar Muscular Atrophy (SBMA), and other diseases caused by STR expansions, including spinocerebellar ataxias (SCAs).

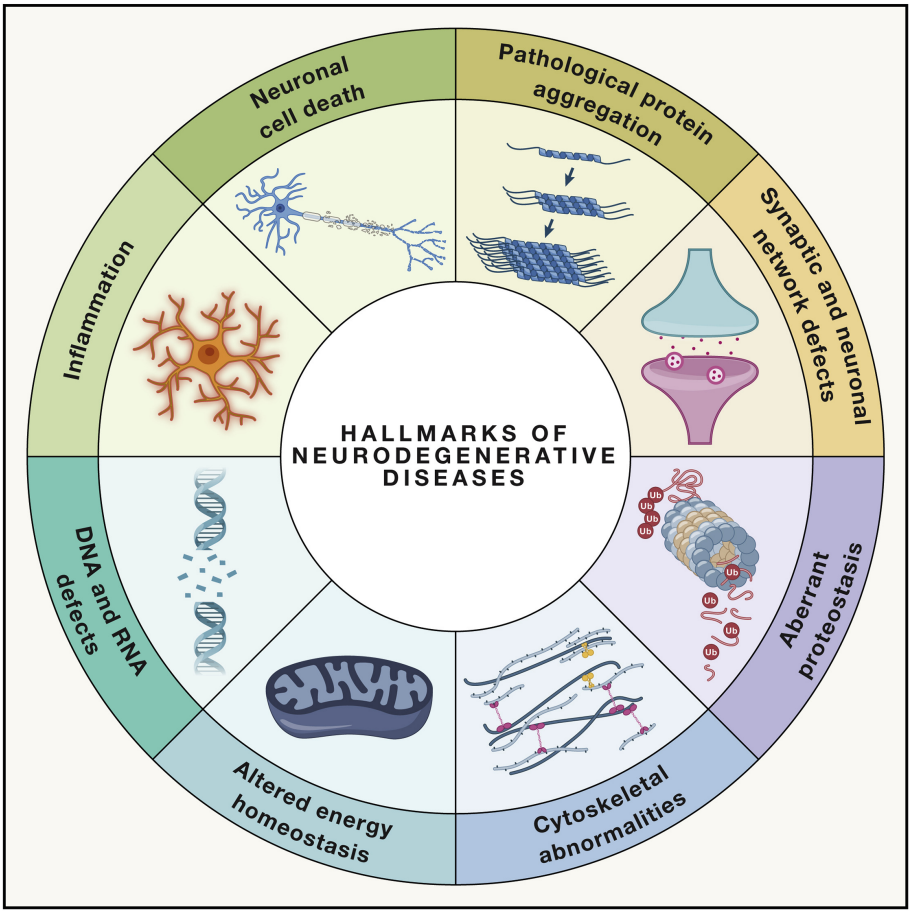


Figure 1: **Hallmarks of neurodegenerative diseases.** *Adapted from Wilson et al., 2023* Based on decades of basic, translational, and clinical research, genetic factors and biochemical pathways underlying many NDDs have been identified, resulting in the identification of eight NDDs hallmarks: pathological protein aggregation, synaptic and neuronal network dysfunction, aberrant proteostasis, cytoskeletal abnormalities, altered energy homeostasis, DNA and RNA defects, inflammation, and neuronal cell death.

The course of ageing, along with genetics and environmental factors, is deemed responsible for the development of NDDs. However, these disorders can manifest in different ways, and, at the same time, different phenotypes can coexist. Moreover, the relationship between the known genetic and environmental factors with disease phenotypes is equivocal, and cases with different genetic mutations can present with similar phenotypes. However, historically, physicians attempted to group similar phenotypes under specific diseases, which has informed neurological subspecialties.

The ALS-FTD spectrum

ALS, also known as motor neurone disease, is an incurable neurodegenerative disorder characterised by the progressive loss of motor neurons that eventually leads to death from respiratory failure. The global prevalence of ALS is 4.5 people per

100,000 (Chiò et al., 2013), with a lifetime risk of developing the disease of 1:350 (Al-Chalabi and Hardiman, 2013). However, these figures are projected to increase, due to the ageing of the population, diagnostics improvement and the widening of the recognised phenotypic manifestations (Arthur et al., 2016).

The traditional clinical hallmarks of ALS are the relentless progressive motor neuron features, eventually affecting limb, bulbar, and respiratory muscles, causing paralysis and difficulties with swallowing and breathing. Clinically, ALS is commonly classified according to the site of onset, progression rate and whether there is a family history of the disease. ALS diagnosis relies on the clinical picture and electromyography testing, but it is mainly a diagnosis of exclusion. Respiratory failure is the most common cause of death (Hobson and McDermott, 2016), and the survival time from diagnosis is between three and five years after the onset (Chiò et al., 2009). Currently, there is no cure for ALS, and the management of patients consists of treating symptoms and providing supportive care. The primary outcome measures in ALS trials are survival and/or rate of decline based on the revised ALS Functional Rating Scale (ALS-FRS-R). However, targeted therapies based on genetics (Korobeynikov et al., 2022; Miller et al., 2022) and novel biomarkers have been proposed, spacing from cerebrospinal fluid (CSF) and blood neurofilaments (Gaiani et al., 2017) to imaging techniques such as magnetic resonance imaging (MRI) (Ferraro et al., 2017) and positron emission tomography (PET) (Pagani et al., 2014; Zanovello et al., 2021).

FTD is the second most common form of pre-senile dementia, with a prevalence ranging between 1 and 26 per 100,000 people (Logroscino et al., 2023). The disease is characterised by behavioural manifestations, including apathy, perseveration and disinhibition (Raaphorst et al., 2012), and cognitive manifestations, particularly deficits in fluency, language, social cognition, executive functions and verbal memory. FTD is still incurable, and the survival time is 6-11 years from symptom onset, although variability between subtypes has been reported (Coyle-Gilchrist et al., 2016). The diagnosis is hampered by the mimic with psychiatric disorders, and the current diagnostic criteria are outdated (Gorno-Tempini et al., 2011; Mesulam et al., 2012; Rascovsky et al., 2011). However, advances in clinical, imaging and molecular characterisation have improved the accuracy of FTD diagnosis, as well as the development of rational therapies and biomarkers (Antonioni et al., 2023; Olney et al., 2017).

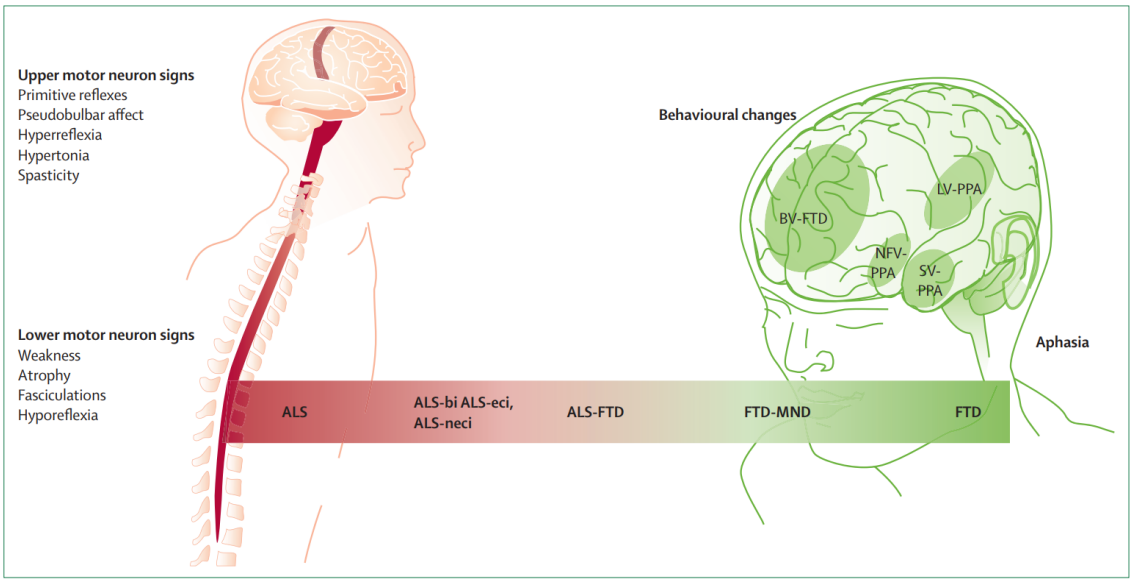


Figure 2: **The ALS-FTD clinical spectrum.** *Adapted from Van Es et al., 2017*
ALS and FTD are the phenotypic extremes on a spectrum disorder (the ALS-FTD continuum), which comprises ALS with some degree of cognitive impairment or behavioural changes (ALS-eci if there is evidence of executive dysfunction, ALS-neci if there is no executive dysfunction but impairment in other cognitive domains, or ALS-bi if behavioural changes are present); ALS-FTD if patients fulfil the diagnostic criteria for both the conditions; FTD-MND if motor neuron involvement develops after the diagnosis of FTD. FTD can be divided into two subtypes, that with time tend to overlap in each patient: behavioural variant and primary progressive aphasia (PPA), which can be further subdivided into three forms: non-fluent variant (left posterior frontal and insular regions), semantic variant (anterior temporal region), and logopenic variant (left temporo-parietal regions). ALS appears to be more closely related to the behavioural variant of FTD than the PPAs. Primitive reflexes include Babinski's sign and Hoffmann's sign.

In the last 20 years, growing evidence has shown a clinical overlap between ALS and FTD (Figure 2) (Burrell et al., 2011; Phukan et al., 2012). Indeed, FTD is a presenting feature in 13.8% of ALS cases (Phukan et al., 2012), and up to 50% of patients with ALS have cognitive and behavioural changes within the spectrum of FTD (Olney et al., 2017). Similarly, 12.5% of patients with behavioural variant FTD develop ALS, and motor neuron involvement is seen in about 40% of patients with FTD (Burrell et al., 2011). This observation, along with the pathological detection of TDP-43 protein aggregates (Arai et al., 2006; Neumann et al., 2006) and the discovery of the causal mechanism involving the *C9orf72* repeat expansion (DeJesus-Hernandez et al., 2011; Renton et al., 2011) in both ALS and FTD, has led to considering the two diseases as parts of a common neurodegenerative spectrum (Abramzon et al., 2020; Burrell et al., 2016).

1.2 ALS-FTD: aetiology and mechanisms

From an aetiological perspective, diseases within the ALS-FTD spectrum are described as multifactorial, dependent on environmental and genetic factors, which interact through a series of yet-to-be-known steps until disease manifestations (Al-Chalabi et al., 2014). More than 40 genes have been linked to ALS (Ghasemi and Brown, 2018) and around 10 to FTD (Greaves and Rohrer, 2019) (Table I). The first gene to be associated with ALS was *SOD1* (Rosen et al., 1993), a regulator of oxidative stress, while the first to be associated with FTD was *MAPT*, which encodes the tau proteins (Hutton et al., 1998).

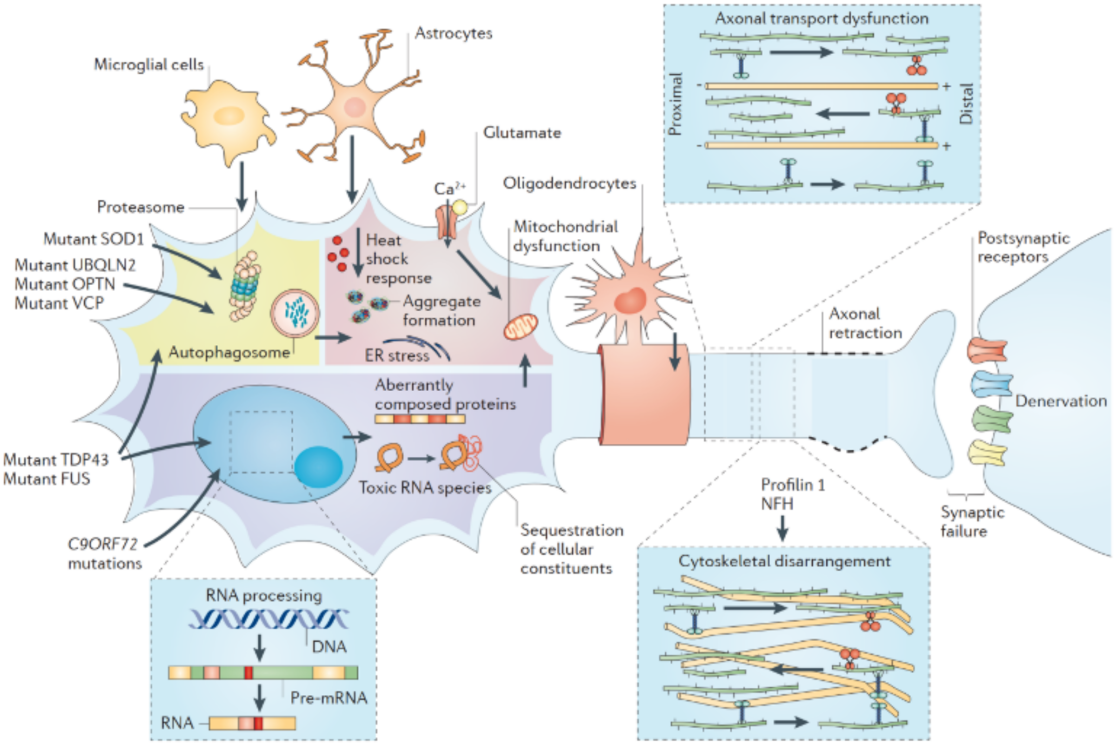


Figure 3: **Cellular and molecular mechanisms responsible for ALS-FTD.** *Adapted from Robberecht and Philips, 2013*
Altered protein homeostasis in ALS (shaded yellow) is due to interference with normal proteasomal and autophagic activity or endoplasmic reticulum-associated protein degradation (ERAD). RNA-binding proteins may self-assemble to form prion-like aggregates. Disturbance of normal DNA and RNA processing (shaded purple), which yields erroneously assembled proteins and toxic RNA species, is caused by mutations in *C9orf72*, *TARDBP* (encoding for TDP-43), and *FUS*. These primary pathogenic changes result in progressive cellular failure (shaded red) that is characterised by protein clumping, aggregate formation, ER stress and Golgi and mitochondrial failure. Axonal architecture (cytoskeleton) and function (transport) fail, and axonal retraction, modified by axonal attraction and repellent systems, results in the denervation of neurons or muscle. Glial cells modify this process through loss and/or the gain of physiological function. Vulnerability factors such as stress response capacity and susceptibility to excitotoxicity determine the sensitivity of neurons to these processes. Some ALS-causing mutant proteins act more downstream in this model.

Gene	Protein	Inheritance
Impaired protein homeostasis		
VCP	Watts et al., 2007, Johnson et al., 2010	AD
OPTN	Maruyama et al., 2010	AD or AR
TBK1	Freischmidt et al., 2015, Pottier et al., 2015, Cirulli et al., 2015	AD
SQSTM1	Fecto et al., 2011	AD
UBQLN2	Deng et al., 2011	XD
VAPB	Nishimura et al., 2004	AD
CHMP2B	Parkinson et al., 2006	AD
FIG4	Chow et al., 2009	AD
SIGMAR1	Luty et al., 2010	AD
ANXA11	Smith et al., 2017	AD
ALS2	Yang et al., 2001, Hadano et al., 2001	AR
SOD1	Rosen et al., 1993	AD or AR
Aberrant RNA metabolism		
C9orf72	DeJesus-Hernandez et al., 2011, Renton et al., 2011	AD
TARDBP	Kabashi et al., 2008, Van Deerlin et al., 2008, Sreedharan et al., 2008, Gitcho et al., 2008	AD
FUS	Kwiatkowski et al., 2009, Vance et al., 2009	AD or AR
TAF15	Ticozzi et al., 2011	AD
EWSR1	Couthouis et al., 2012	AD
SETX	Y. Z. Chen et al., 2004	AD
ANG	Greenway et al., 2006	AD
HNRNPA1	H. J. Kim et al., 2013	AD
HNRNPA2B1	H. J. Kim et al., 2013	AD
MATR3	Johnson et al., 2014	AD
ELP3	Bento-Abreu et al., 2018	Unknown
Altered cytoskeletal dynamics		
DCTN1	Puls et al., 2003	AD
PFN1	C. H. Wu et al., 2012	AD
TUBA4A	Smith et al., 2014	AD
NEFH	Figlewicz et al., 1994	AD
PRPH	Gros-Louis et al., 2004	AD
Other disease mechanisms		
CHCHD10	Bannwarth et al., 2014	AD
NEK1	Cirulli et al., 2015, Brenner et al., 2016, Kenna et al., 2016	AD
C21orf2	Van Rheenen et al., 2016	Unknown
UNC13A	Van Es et al., 2009	Unknown
SPG11	Orlacchio et al., 2010	AR
DAO	Mitchell et al., 2010	AD
ATXN2	Elden et al., 2010	AD
GRN	Baker et al., 2006	AD
KIF5A	Nicolas et al., 2018	AD

Table I: **Genes implicated in ALS-FTD and their main pathogenetic mechanism.** AD, autosomal dominant; AR, autosomal recessive; XD, X-linked dominant

In 2011, it was found that the most common genetic cause of ALS and FTD is a GGGGCC repeat expansion in the *C9orf72* gene (DeJesus-Hernandez et al., 2011; Renton et al., 2011). This causes disease via different mechanisms, such as haploinsufficiency (Haeusler et al., 2016), RNA toxicity mechanisms including the formation of DNA-RNA hybrids (Haeusler et al., 2014) and RNA foci (Cooper-Knock et al., 2014), and repeat-associated translation of dipeptide repeat proteins (DPRs) (Boeynaems et al., 2017).

The other two most commonly mutated genes in ALS-FTD are *FUS* and *TARDBP* (the latter encoding for TDP-43), which are both RNA-binding proteins (RBPs), characterised by the presence of low-complexity domains (Gasset-Rosa et al., 2019).

In more recent years, several other RBPs, as well as genes involved in protein homeostasis and cytoskeletal transport (Peters et al., 2019), were found to be associated with ALS-FTD (Figure 3). The imbalance of these three mechanisms leads to RBP aggregation in the cytoplasm, resulting in a toxic gain of function, with the formation of stress granules (Conicella et al., 2016), and the loss of the normal processing of their target RNAs (Brown et al., 2022).

Repeat expansions in ALS-FTD and neurological diseases

Repeat expansion disorders (REDs) are a heterogeneous group of conditions which mainly affect the nervous system, and include Fragile X syndrome and the commonest ALS-FTD inherited form of amyotrophic lateral sclerosis and frontotemporal dementia (*C9orf72*-ALS-FTD) (Paulson, 2018).

The same underlying mechanism causes REDs: the expansion of short repetitive DNA sequences (1-6 bp), also known as STRs, within their respective genes. The mutational process is gradual: normal alleles are usually passed stably from parent to child, with rare changes in repeat size; intermediate-size repeats are more likely to expand into the disease range, giving rise to pathogenic repeat lengths in the next generation.

REDs are clinically heterogeneous: for example, *C9orf72* expansions can present as either FTD or ALS even within the same family; and one in three patients carrying the repeat expansion in *C9orf72* shows an atypical presentation at onset such as Alzheimer's disease (AD) and Huntington's disease (HD) among others (Gossye et al., 1999; Moore et al., 2020). For many REDs, the variability in repeat lengths underlines the substantial clinical heterogeneity (Van Der Ende et al., 2021); longer repeats cause more severe disease and earlier symptom onset (Langbehn et al., 2004).

REDs are reported to be quite rare; however, taken together, they affect 1 in 3000 people, with the most common being Fragile-X syndrome, *C9orf72*-related

diseases, and Myotonic Dystrophy type 1, followed by SCAs, SBMA, and HD. However, prevalence estimates are based on clinical presentation or family history. Unfortunately, the diverse presentation of these disorders makes it difficult to trace their distribution. Recently, bioinformatics tools to size repeat expansions from WGS data have been developed (Dolzhenko et al., 2017), making large WGS repositories a valuable resource for assessing the prevalence of STRs.

With regards to the ALS-FTD spectrum, repeat expansions in *C9orf72* are the most common genetic cause of these disorders. Moreover, the recent discovery that *ATXN2* intermediate repeats are also linked to ALS raises the question about how this mutational mechanism can lead to the ALS-FTD phenotypes and if it can alter the disease progression (Cooper-Knock et al., 2024; Elden et al., 2010). Another study assessed the prevalence of STRs in ALS and showed an overall increased prevalence of STRs in ALS. However, taken singularly, aside from the two known STRs previously linked to ALS, none of the others examined showed a significant association with the disease (Henden et al., 2023), a finding confirmed by another work on *HTT* expansions in ALS-FTD (Zimmermann et al., 2025).

Spinal and Bulbar Muscular Atrophy

Another RED that manifests as a form of motor neurone disease (MND) is Spinal and Bulbar Muscular Atrophy (SBMA), also known as Kennedy's disease, which occurs when the CAG repeat coding for a polyglutamine tract in exon 1 of the androgen receptor (*AR*) gene expands beyond 37 repeats (La Spada, 2022). SBMA fully manifests only in males, with a mean age at onset of 43 years, which is partially influenced by CAG repeat size (Fratta, Nirmalananthan, et al., 2014) and is characterised by progressive muscular weakness induced by the degeneration of the lower motor neurons and primary muscular damage (La Spada, 2022). Importantly, SBMA is also associated with a variety of non-neurological conditions, including insulin resistance, fatty liver disease, and metabolic syndrome (Manzano et al., 2018).

A role for neuroinflammation

In the last decades, particularly in the AD field, there has been an increased recognition of the role of the immune system in neurodegenerative diseases (Wilson et al., 2023). As the central nervous system (CNS) is protected from circulating antigens and immune cells by the blood-brain barrier (BBB), part of the efforts have been devoted to describing damage to this structure. Moreover, the only resident immune cell type in the brain is the microglia; therefore, research has focused on

this population. However, more recently, other immune cell populations have been linked to diseases along the ALS-FTD spectrum, eliciting a novel interest in immune biomarkers, both via CSF analysis and neuroimaging (Ferraro et al., 2017; Gaiani et al., 2017; Pagani et al., 2014; Zanovello et al., 2021).

Highly differentiated and clonal CD8+ T cell populations have recently been described to be enriched in heterogeneous neurodegenerative and inflammatory conditions, including ALS (Campisi et al., 2022), AD (Gate et al., 2020), and sIBM (Greenberg et al., 2019), among others. These cell populations exist in peripheral blood mononuclear cells (PBMCs) and in disease-relevant tissue such as CSF or skeletal muscle. The presence of an antigen-specific response is suggested by the clonal expansion and increased expression of markers associated with activation and cytotoxicity (Hu et al., 2022). Therefore, it is reasonable to hypothesise that these cell populations could be disease-associated and somehow implicated in the pathogenesis of these conditions. However, no clear antigenic targets have been described, and attempts at finding them have largely been unsuccessful (Dhanwani et al., 2020; Jiang et al., 2023; Ramachandran et al., 2023).

1.3 TDP-43 pathology

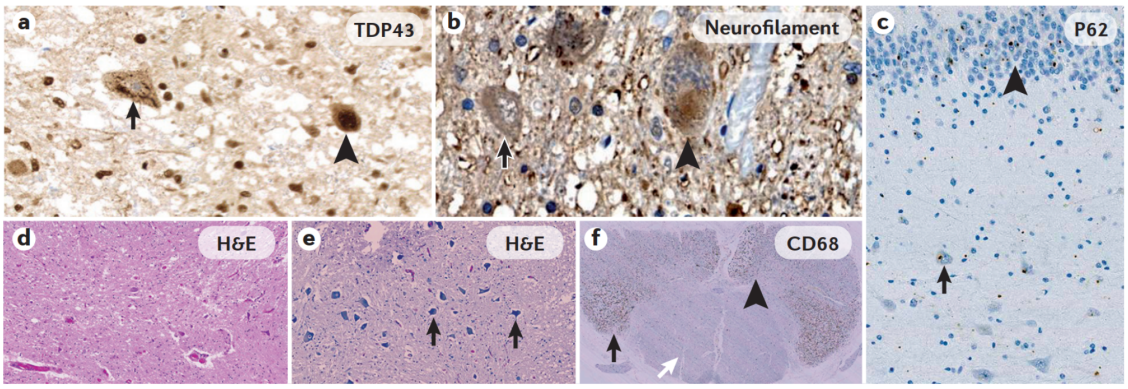


Figure 4: **Different types of pathology underlie the ALS-FTD spectrum.** *Adapted from Hardiman et al., 2017*
a. Normal localization of TDP43 in the nucleus (arrowhead) and aberrant localization in a diseased neuron with loss of nuclear expression and a 'skein-like' inclusion in the cytoplasm (black arrow). b. Normal motor neuron (black arrow) and a hyaline conglomerate inclusion that is labelled for neurofilament (arrowhead) in a patient with *SOD1*-linked ALS. c. P62-positive, TDP-43 negative dipeptide repeat inclusions with a 'stellate' morphology in the pyramidal cells of CA4 (black arrow) and granule cells of the dentate fascia (arrowhead) in the hippocampus of a patient with *C9orf72*-linked ALS. d. Depleted numbers of motor neurons (the absence of arrows) in the ventral horn of the spinal cord in a patient with ALS. e. Motor neurons (arrows) in the spinal cord ventral horn of a healthy individual. f. Marked microglial reactivity (CD68 labelling) in the lateral tracts (black arrow) and ventral horns (arrowhead), with no labelling in the dorsal columns (white arrow) of the spinal cord in a patient with ALS. H&E, haematoxylin and eosin.

The pathological hallmark of diseases along the ALS-FTD spectrum is the aggregation of ubiquitylated proteinaceous inclusions in the cytoplasm (Robberecht and Philips, 2013), which leads to the death of selected neuronal populations. The shedding of protein complexes might induce prion-like cell-to-cell propagation of disease (Polymenidou and Cleveland, 2011), potentially explaining the focal initial pathology and progressive, continuous spread (Braak et al., 2013). Pathogenic protein aggregates include misfolded tau, SOD1, and FUS in cases with mutations in these genes (Figure 4). In all the other cases of ALS (97%) and FTD (around 50%), a hyper-phosphorylated and ubiquitinated form of TDP-43 is the major component of the inclusion bodies in the cytoplasm at autopsy (Arai et al., 2006; Neumann et al., 2006), although mutations in *TARDBP* are a rare cause of ALS and no mutations in *TARDBP* have been linked with FTD. Degeneration of neurons is accompanied by numerous alterations of glial cells: astrocytes exhibit excitotoxicity due to impaired synaptic glutamate reuptake, microglia switch from neuroprotective (M2) to neurotoxic (M1) phenotype (McCauley and Baloh, 2019), and incomplete oligodendrocyte differentiation (Philips et al., 2013) with decreased metabolic support to axons.

Expanding the TDP-43 proteinopathy spectrum

More recently, TDP-43 aggregation has been reported in other brain diseases, including Alzheimer's disease and tauopathies (Arai et al., 2009), as well as in chronic traumatic encephalopathy (McKee et al., 2015). The role of ageing in TDP-43 proteinopathies is suggested by the fact that misfolded TDP-43 is found in the brains of older adults over age 85 with limbic-predominant age-related TDP-43 encephalopathy (LATE) (Nelson et al., 2019).

Outside the CNS, cytoplasmic ubiquitinated TDP-43 inclusions have been found in the muscle of patients affected by several myopathies, including sIBM (Weihl et al., 2008), inclusion body myositis with Paget's disease and frontotemporal dementia/amyotrophic lateral sclerosis (IBMPDF/ALS) caused by mutations in *VCP* (Johnson et al., 2010; Watts et al., 2007), *HNRNPA1/A2B1* (H. J. Kim et al., 2013), *MATR3* (Johnson et al., 2014), and *SQSTM1* (encoding for the protein p62) (Fecto et al., 2011), and even in paraspinal muscles and diaphragm of ALS cases (Table II) (Cykowski et al., 2017; Hernandez Lain et al., 2011). Interestingly, myogranules containing TDP-43 mislocalise in the sarcoplasm during muscle development and healing, suggesting a possible imbalance of this mechanism as a trigger for aggregation of TDP-43 in the disease pathogenesis (Vogler et al., 2018).

Disease	Associated pathology	Associated genes
sIBM (Weihl et al., 2008)	tau, p62	-
Distal myopathy with rimmed vacuoles (Küsters et al., 2009)	p62	*GNE*
Myotinilopathies (Olivé et al., 2009)	-	*MYOT*
Desmilinopathies (Olivé et al., 2009)	-	*DES*
Inflammatory myopathies with mitochondrial pathology (Temiz et al., 2009)	-	-
IBMPDF/ALS (Nalbandian et al., 2011)	p62	*VCP, hnRNPA1, hnRNPA2B1, SQSTM1, MATR3*
Limb-girdle muscular atrophy (Sandell et al., 2016)	-	*DNAJB6*
Distal hereditary motor neuropathy and myofibrillary myopathy (Cortese et al., 2018)	-	*HSPB8*
Hereditary motor and sensory neuropathy (Yamashita et al., 2019)	-	*TFG*
ALS (Cykowski et al., 2017; Hernandez Lain et al., 2011)	p62	*TARDBP, C90rf72*

Table II: **Diseases associated with TDP-43 pathology in non-CNS tissues.**

Inclusion body myositis

sIBM is the most common idiopathic inflammatory myopathy in people > 50 years of age. The disease causes progressive muscle atrophy and weakness, particularly of finger flexors and quadriceps, that lead to disability (Benveniste et al., 2011; Greenberg, 2019). sIBM does not respond to traditional immunomodulators and immunosuppressants, and no therapy exists. Its insidious onset makes the diagnosis difficult, with a median time from onset to diagnosis of around 5 years (Needham et al., 2008; Rose, 2013). The diagnosis is aided by muscle biopsy, which shows endomysial inflammation, mitochondria pathology, rimmed vacuoles and p62-positive protein aggregates, containing TDP-43 and beta-amyloid (Catalán et al., 2014; Snedden et al., 2022; Weihl et al., 2008). There is considerable debate as to the primary cause of sIBM, considering the combination of inflammatory and degenerative pathological features (Benveniste et al., 2015; Greenberg, 2019; Keller et al., 2017; Weihl and Mammen, 2017).

Interestingly, there also exist hereditary forms of inclusion body myositis, caused by mutations in *GNE, VCP, hnRNPA1* and *hnRNPA2B1*. These diseases share the neurodegenerative pathological features with sIBM; however, they have an earlier age of onset and muscle biopsies typically lack inflammation (Britson et al., 2022).

1.4 TDP-43 is an essential RNA-binding protein

TDP-43 molecular structure and functions

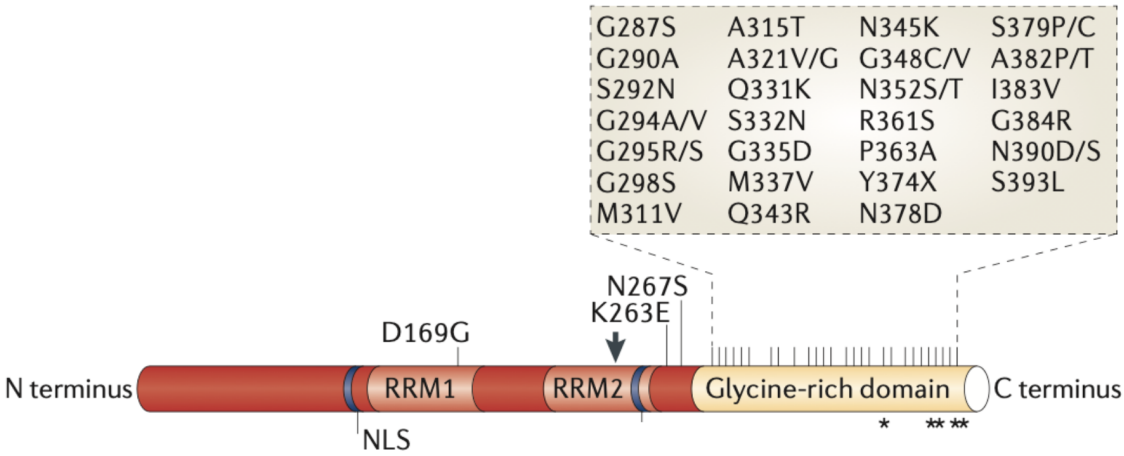


Figure 5: **TDP43 structure highlighting ALS-FTD mutations and phosphorylation sites (*).** *Adapted from Lee et al., 2012*
The graphical representation of TDP-43 protein highlights the presence of several disease-associated mutations, particularly localised in the C-terminal Glycine-rich domain. The asterisks highlight the phosphorylation sites, also located in the C-terminal domain.

The *TARDBP* (transactive response DNA-binding protein) gene, located in human chromosome 1, encodes the protein TDP-43 (Figure 5). Structurally, TDP-43 is 414 amino acids long and consists of several domains, including:

- an N-terminal domain (NTD, residues 1-76) with a well-defined fold that has been shown to form dimers and oligomers (Afroz et al., 2017);
- a nuclear localisation signal (NLS, residues 82–98);
- the 2 highly conserved RNA recognition motifs RRM1 (106-176) and RRM2 (191-259), required to bind target RNA and DNA (Lukavsky et al., 2013; Polymenidou et al., 2011; Sephton et al., 2011; Tollervey et al., 2011);
- an unstructured C-terminal domain (CTD) encompassing residues 274-414, containing a glycine-rich region, involved in protein-protein interactions that are fundamental for TDP-43 splicing repression role (Ling et al., 2015). Interestingly, this domain harbours most of the mutations associated with familial amyotrophic lateral sclerosis (Conicella et al., 2016; Sreedharan et al., 2008; Vega et al., 2019).

Initially, TDP-43 was discovered as a DNA-binding protein that represses HIV-1 transcription by binding to chromosomally integrated viral DNA (TAR sequence)

(Ou et al., 1995). Following studies highlighted TDP-43 as a key player in the regulation of the alternative splicing of the *CFTR* (Buratti and Baralle, 2001) and apoA-II genes (Mercado, 2005), unearthing its fundamental role in RNA regulation.

In fact, TDP-43, by binding to UG-rich sequences in around a third of the transcriptome (Tollervey et al., 2011), plays a role in transcriptional repression, translational regulation and pre-mRNA splicing; these mechanisms ensure the mRNA degradation rate and stability. Moreover, the RNA-binding capacity of TDP-43 is paramount to stabilise stress granules in the cytoplasm and to regulate gene expression, including that of TDP-43 itself, by binding to the 3' untranslated region (3'-UTR).

RNA splicing

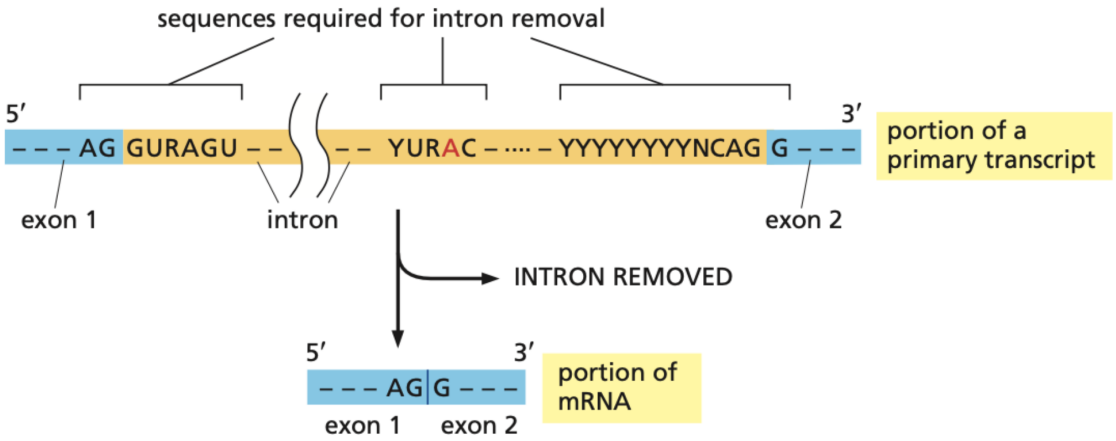


Figure 6: **RNA splicing schematics highlighting consensus sequences.** *Adapted from Alberts et al., 2015*
Blue segments represent exons, while introns are shaded in yellow.

Splicing is an energy-consuming sequential process that consists of the removal of introns, long noncoding intervening sequences, from pre-mRNA, and the joining of exons. The evolutionary reason for splicing is to enable the production of tissue- and development-specific proteins and the emergence of new ones by genetic recombination of common protein domains and random mutations. This balance between accuracy and flexibility, while increasing the eukaryotic coding potential, can cause human disease when pathogenic mutations cause aberrant splicing (Alberts et al., 2015).

Mechanistically, the splicing machinery recognises three portions of the precursor RNA molecule by consensus sequences: the 5' splice site, the 3' splice site, and the branch point in the intron sequence that forms the base of the excised lariat (Figure 6). The choice of splice sites depends on the strength of these three signals,

the co-transcriptional assembly of the spliceosome, chromatin structure, and the "bookkeeping" that underlies exon definition.

Splicing is unusually flexible: a mutation in a nucleotide sequence critical for the splicing of a particular intron does not necessarily prevent the splicing of that intron altogether. Instead, the mutation typically creates a new pattern of splicing, either via exon skipping or by the usage of a cryptic splice junction, with the splicing machinery seeking the next best pattern. This flexibility is also guaranteed by the presence of additional splicing enhancers and repressors, such as TDP-43.

TDP-43 cryptic splicing

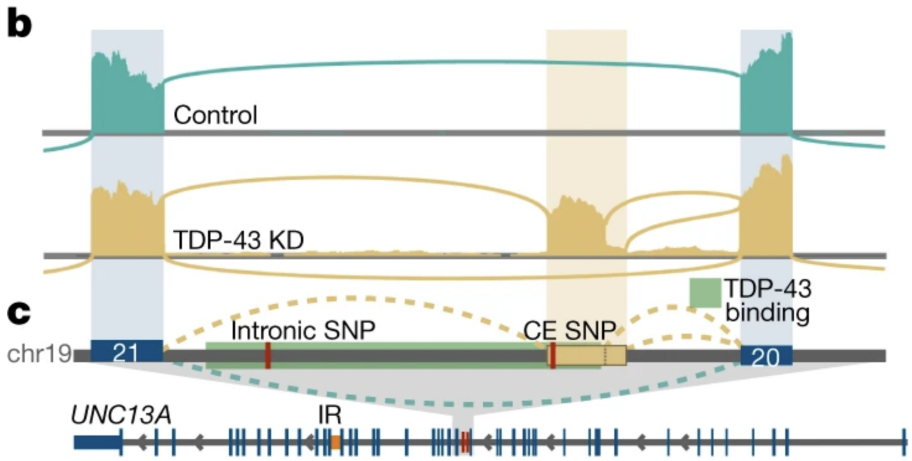


Figure 7: **TDP-43 knockdown causes the inclusion of cryptic exons.** *Adapted from Brown et al., 2022*

What's the role of TDP-43 in splicing? In physiological conditions, TDP-43 represses the splicing of non-conserved sequences and enhances the splicing of conserved exons by binding to intronic UG-rich regions. TDP-43 extrusion from the nuclei (Arai et al., 2006; Neumann et al., 2006) leads to the de-repression of these splicing events and the inclusion of non-conserved cryptic exons (CEs) and skipping of conserved exons in mature mRNA (Figure 7) (Humphrey et al., 2017; Jeong et al., 2017; Ling et al., 2015). Cryptic exons may have originated from a lack of evolutionary pressure on introns, and are highly species- and tissue-specific.

Cryptic exons are usually cassette exons or exon extensions, and often contain premature stop codons (PTCs) or lead to a frameshift exon, which eventually leads to Nonsense-Mediated Decay (NMD) of the transcripts, and loss of mRNA and protein. In-frame CEs, alternative last exon, and exon skipping can potentially be translated and lead to a toxic gain of function (Fiesel et al., 2012; Prpar Mihevc et al., 2016; Roczniak-Ferguson and Ferguson, 2019; Shiga et al., 2012; Tollervey et al., 2011).

All these changes can be detected by means of RNA sequencing, and, since expression of some CEs correlates with the burden of TDP-43 pathology, they could serve as a potential readout of TDP-43 loss of function and proteinopathy (Brown et al., 2022; Prudencio et al., 2020). So far, thousands of these events have been described, both in vitro and in vivo. Interestingly, many of them affect pathways that are disrupted in ALS-FTD, including nuclear export and autophagy, as well as synaptic function and axonal growth (Brown et al., 2022; Klim et al., 2019; Ling et al., 2015; Melamed et al., 2019). More recently, CEs have been described in other diseases with TDP-43 proteinopathy as co-pathology, particularly AD (Estades Ayuso et al., 2023) and sIBM (Britson et al., 2022).

CEs in *STMN2* (Klim et al., 2019; Melamed et al., 2019) and *UNC13A* (Brown et al., 2022; X. R. Ma et al., 2022) have been identified as particularly vulnerable to TDP-43 loss of function, with subsequent loss of functional transcripts and proteins. Their misprocessing contributes to neuronal dysfunction and degeneration, positioning them as promising candidates for therapeutic intervention using antisense oligonucleotides that rescue their physiological splicing (Baughn et al., 2023; Keuss et al., 2024).

On the other hand, TDP-43 CEs that escape NMD have the potential to be translated as additional peptides and lead to protein gain or loss of function (Irwin et al., 2024; Seddighi et al., 2024). Previous works have shown that the CE in *HDGFL2* forms a cryptic peptide included in stable protein isoforms. Due to its stability and detectability in biofluids and tissue by means of ELISA and immunohistochemistry, the *HDGFL2* cryptic peptide has emerged as a potential biomarker of TDP-43 proteinopathies, including ALS, FTD, and AD (Calliari et al., 2024; Irwin et al., 2024). Moreover, since CD8+ T cells can recognise novel peptides derived from non-canonical reading frame translation in cancer cells (Laumont and Perreault, 2018), it is possible that CD8+ T cells can mount a response against cells expressing cryptic epitopes in TDP-43 proteinopathies.

Furthermore, TDP-43 also regulates alternative polyadenylation, influencing the length of 3' UTRs and, consequently, the stability, localisation, and translation of mRNAs (Arnold et al., 2025; Bryce-Smith et al., 2024; Zeng et al., 2024). Disruption of cryptic exon repression or polyadenylation control highlights TDP-43's essential function in maintaining RNA homeostasis within the central nervous system.

However, despite the increasing number of scientific publications about TDP-43-related splicing alterations, outstanding questions remain, including the relevance of each CE to disease pathogenesis, and the characterisation of their biological features,

including stability, expression, and sensitivity to TDP-43 loss. It is also possible that the expression of other RBPs may affect the inclusion of cryptic exons (Appocher et al., 2017; Šušnjar et al., 2022). This is particularly relevant since other RBPs have been shown to repress the inclusion of CEs (Ling et al., 2016).

Nonsense-Mediated Decay

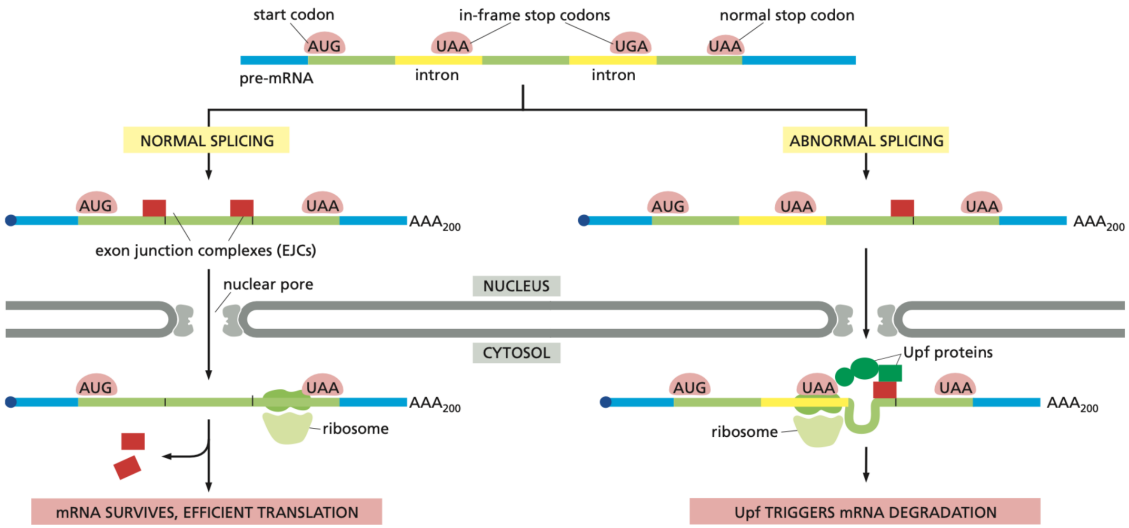


Figure 8: **Nonsense-Mediated Decay schematics.** *Adapted from Alberts et al., 2015*

The control of gene expression relies on several mechanisms, occurring from the DNA to the protein level. At the mRNA level, NMD degrades incorrectly spliced mRNA as soon as they are exported to the cytosol, avoiding the translation of truncated or aberrant proteins (Alberts et al., 2015).

Mechanistically, NMD recognises premature nonsense (stop) codon (UAA, UAG, or UGA) in the mRNA transcript. As the 5' end emerges from a nuclear pore, the mRNA is translated by a ribosome, which displaces exon junction complexes (EJCs) after each exon is translated. Since the normal stop codon lies within the last exon, by the time the ribosome reaches it and stalls, no more EJCs will be bound to the mRNA and the mRNA is considered correct and can be translated (Figure 8). However, if the ribosome reaches a stop codon earlier, when EJCs remain bound, the mRNA molecule is rapidly degraded.

Evolutionarily, NMD allows eukaryotic cells to more easily explore new genes formed by DNA rearrangements, mutations, or alternative patterns of splicing - by selecting only those mRNAs for translation that can produce a full-length protein.

In human diseases caused by nonsense, frameshift, and splice-site mutations, NMD prevents the translation of potentially toxic proteins from mutant copies of the gene.

1.5 Research questions, Aims, and Overview

ALS, FTD, and sIBM are incurable neurodegenerative diseases with overlapping genetics and pathological mechanisms, particularly TDP-43 proteinopathy. Despite recent advances, the molecular underpinnings of these disorders remain incompletely understood. Parallel to this, STR expansion disorders contribute to a range of neurological manifestations, but their true prevalence in the general population has been underexplored due to reliance on small cohort studies.

The overall aim of this thesis is to deepen our understanding of TDP-43 proteinopathies and STR-related neurological disorders by combining molecular, immunological, and population-level approaches.

Specifically, this work addresses three core questions:

1. What is the prevalence of STR expansions in the general population?
2. What is the role of TDP-43 loss of function and NMD in regulating CEs?
3. Are CE-derived peptides present in human disease with TDP-43 pathology, and can they elicit an immune response?

Using in vitro models, clinical tissue samples, and large-scale whole-genome sequencing datasets, I integrated molecular biology and bioinformatics to approach these questions from both mechanistic and translational perspectives. This work provides new insights into CE biogenesis, proposes novel immunogenic markers for TDP-43-associated diseases, and presents population-scale estimates of STR prevalence.

- The first result chapter explores the prevalence of STR expansions, interrogating large WGS datasets, revealing the distribution of pathogenic and intermediate alleles in the general population.
- The second result chapter investigates the biological properties of TDP43 CEs at the transcriptome-wide level, focusing on their relationship to TDP-43 loss of function and on the regulatory role of NMD.
- The third and last result chapter characterises CE peptides in sIBM tissues and demonstrates their potential to stimulate cytotoxic T cell responses, establishing a functional link between TDP-43 pathology and immune activation in the disease context.

Collectively, these findings advance our understanding of the molecular biology and population genetics of neurodegeneration. They support the development of new biomarkers, therapeutic strategies targeting immune pathways, and improved genetic counselling practices. Furthermore, the methodologies and insights generated here have broader implications for other neurological diseases, including Alzheimer's disease and rare disorders with overlapping molecular features.

Chapter 2

Materials and Methods

2.1 Contributions

Parts of the text in this chapter are copied from the works "TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A" (Brown et al., 2022), "Unexpected Frequency of the Pathogenic AR CAG Repeat Expansion in the General Population" (Zanovello et al., 2023), and "Increased Frequency of Repeat Expansion Mutations across Different Populations" (Ibañez et al., 2024), and therefore, other people contributed to ideas, drafting, and writing. Credits for contributions can be found in the UCL Research paper declaration forms at the beginning of the thesis and in the relevant result chapter.

2.2 Genotyping of STRs

Whole Genome Sequencing and AR genotyping

Whole-genome sequencing and cohort characterisation

Appendix Table 1 provides a summary of the age and ethnicity of the cohorts assessed in this study. Whole genome sequencing (WGS) data, including chemistry, read length, coverage, alignment, genome build, and ExpansionHunter (EH) version from each cohort, are summarised in Appendix Table 2.

AR genotyping

EH was used to estimate repeat lengths of the *AR* CAG disease-causing expansions in samples that had undergone WGS. This algorithm has been validated using experimentally confirmed samples carrying pathogenic expansions (Dolzhenko et al.,

2017; Ibañez et al., 2022). Pathogenic alleles in the *AR* gene were defined as those containing 38 or more CAG repeats (La Spada, 2022).

Visual inspection

As previously validated (Ibañez et al., 2022; Roy et al., 2018), EH calls for *AR* CAG repeat underwent a blind quality check process (QC) by visual inspection. The EH calls can be visualised by generating "pileup" graphs, which enable the reviewer to easily evaluate the number of reads and the sequences supporting each call, and therefore assess the length of the repeat expansion, as shown in Figure 9. 486 pileups were checked, of which 282 from 100K GP (≥34 repeats), 67 from NIH (≥34 repeats), 14 from Project MinE (≥37 repeats), and 123 from GnomAD (≥37 repeats). See Appendix Table 1 for EH calls before and after visual QC in each cohort.

Genotyping of STRs from 100k GP and TOPMed

Ethics Statement Inclusion & Ethics

The 100000 Genomes Project is a UK programme to assess the value of WGS in patients with unmet diagnostic needs in rare diseases and cancer. Following ethical approval for the 100000 Genomes Project by the East of England Cambridge South Research Ethics Committee (reference 14/EE/1112), including for data analysis and return of diagnostic findings to the patients, these patients were recruited by health-care professionals and researchers from 13 Genomic Medicine Centres in England and were enrolled in the project if they or their guardian provided written consent for their samples and data to be used in research, including this study. For ethics statements for the contributing TOPMed studies, full details are provided in the original description of the cohorts (Taliun et al., 2021).

Whole genome sequencing datasets

Both the 100,000 Genomes Project (100K GP) and Trans-Omics for Precision Medicine (TOPMed) include WGS data optimal to genotype short DNA repeats: WGS libraries generated using PCR-free protocols, sequenced at 150 base-pair read-length, and with a 35x mean average coverage (Appendix Table 3). For both the 100K GP and TOPMed cohorts, the following genomes were selected: i) WGS from genetically unrelated individuals (see "Ancestry and relatedness inference" below); ii) WGS from people not presenting with a neurological disorder - these people were excluded to avoid overestimating the frequency of a repeat expansion due to individuals recruited due to symptoms related to a RED.

The TOPMed project has generated omics data, including WGS, on over 180,000 individuals with heart, lung, blood, and sleep disorders (see NHLBI Trans-Omics for Precision Medicine WGS-About TOPMed (https://topmed.nhlbi.nih.gov/)). TOPMed has incorporated samples gathered from dozens of different cohorts, each collected using different ascertainment criteria.

Repeat expansion genotyping and visualisation

The EH v3.2.2 software package was used for genotyping repeats in disease-associated loci (Dolzhenko et al., 2017, 2019). EH assembles sequencing reads across a predefined set of DNA repeats using both mapped and unmapped reads (with the repetitive sequence of interest) to estimate the size of both alleles from an individual. REViewer software package was used to enable direct visualisation of haplotypes and corresponding read pileup of the EH genotypes (Dolzhenko et al., 2022).

Ancestry and relatedness inference

For relatedness inference, WGS genomic Variant Call Format (gVCF) for each participant were aggregated with Illumina's agg or gvcfgenotyper (https://github.com/Illumina/gvcfgenotyper). All genomes passed the following QC criteria: cross-contamination <5% (VerifyBamId) (Jun et al., 2012), mapping rate >75%, mean-sample coverage >20, and insert size > 250 bp. No variant QC filters were applied in the aggregated dataset, but the VCF filter was set to 'PASS' for variants which passed GQ (genotype quality), DP (depth), missingness, allelic imbalance, and Mendelian error filters. From here, by using a set of 65,000 high-quality SNPs, a pairwise kinship matrix was generated using the PLINK2 implementation of the KING-Robust algorithm (www.cog-genomics.org/plink/2.0/) (Chang et al., 2015). For relatedness, PLINK2 '–king-cutoff' (www.cog-genomics.org/plink/2.0/) relationship-pruning algorithm 3 was used with a threshold of 0.044. These were then partitioned into 'related' (up to, and including, 3rd-degree relationships) and 'unrelated' sample lists. Only unrelated samples were selected for this study. 1000 Genomes Project phase 3 (1K GP3) data (The 1000 Genomes Project Consortium, 2015) was used to infer ancestry, by taking the unrelated samples and calculating the first 20 PCs using GCTA (Jun et al., 2012). The aggregated data were then projected (100K GP and TOPMed separately) onto 1K GP3 PC loadings, and a random forest model was trained to predict ancestries based on 1) First 8 1K GP3 PCs, 2) setting 'Ntrees' to 400, and 3) train and predict on 1kPG3 five broad super-populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). In total, the following WGS data were analysed: 34,190 individuals in

the 100K GP; 47,986 in TOPMed; 2,504 in the 1K GP3. The demographics can be found in Appendix Table 4.

2.3 AR detection by WGS benchmarking

To assess the performance of WGS to detect the CAG repeat in the *AR* gene, WGS calls were benchmarked against PCR fragment analysis, obtained as follows: First, WGS was obtained from 20 individuals with previously identified pathogenic expansion in *AR* by standard diagnostic PCR testing (i.e. positive control). Furthermore, PCR fragment analysis results for 56 patients recruited to the 100k GP who had been previously tested for the *AR* expansion were obtained (i.e. negative controls). Moreover, 21 DNA samples from patients recruited to the 100k GP, where WGS/EH predicted the presence of an expansion, were assessed by PCR.

PCR

The CAG trinucleotide repeat length in AR was quantified using a PCR method, where AR alleles were amplified by PCR using GoTaq DNA polymerase (Promega) with the forward primer (6FAM-GCCTGTTGAACTCTTCTGAGC) containing a fluorescein amidite (FAM)-label, used to enable fluorescence detection during the fragment analysis, and the reverse primer GCTGTGAAGGTTGCTGTTCCTC (Fratta, Collins, et al., 2014). PCR products were electrophoresed on an ABI 3730xl DNA analyser with a LIZ-500 size standard (Applied Biosystems). Fragment analysis was performed with GeneMapper software (version 5.0, Applied Biosystems), deriving numbers of repeats from a standard curve generated using samples of known repeat size ascertained by Sanger sequencing.

2.4 Statistical analysis

The statistical formulas used to assess the repeat expansion performance dataset have been taken from https://www.medcalc.org/calc/diagnostic_test.php. Considering TN = True Negative; FP = False Positive; TP = True Positive; FN = False negative; PPV = positive predictive value:

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$PPV = \frac{sensitivity * prevalence}{(sensitivity * prevalence) + (1 - specificity) * (1 - prevalence)}$$

The R correlation coefficient was calculated using Pearson's equation. 95% CIs for the X chromosome frequencies were computed using the Wilson score method.

Computation of RED genetic prevalence

The frequency of each repeat size across the 100K GP and TOPMed genomic datasets was determined. Genetic prevalence was calculated as the number of genomes with repeats exceeding the premutation and full mutation cut-offs (Table 14) for autosomal dominant and X-linked REDs. For autosomal recessive REDs, the total number of genomes with monoallelic or biallelic expansions was calculated, compared to the overall cohort. Overall, unrelated and non-neurological disease genomes corresponding to both programmes were considered, breaking down by ancestry.

2.5 Disease prevalence estimation

I tabulated the cumulative distribution of disease onset reported for 983 patients (Laskaratos et al., 2021), binning them in five-year age groups (upper panel of Figure 13). I also plotted the distribution of the general English male population (n=27,827,831) ("Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019", 2020), using the same five-year age group bins (2nd panel of Figure 13). I then multiplied the cumulative distribution of the disease onset by the corresponding general male count for each age group to obtain the distribution of the disease by age group, which I then used to estimate the disease prevalence.

Modelling disease prevalence using carrier frequency

To estimate the expected number of new cases by age group, the age at onset distribution of the specific disease, available from cohort studies or international registries, was used. For *C9orf72*-disease, I tabulated the distribution of disease onset of 811 patients with *C9orf72*-ALS pure and overlap FTD, and 323 patients with *C9orf72*-FTD pure and overlap ALS (Murphy et al., 2017). HD onset was modelled using data derived from a cohort of 2,913 individuals with HD described

by Langbehn et al., 2004, and Myotonic Dystrophy type 1 (DM1) was modelled on a cohort of 264 non-congenital patients derived from the UK Myotonic Dystrophy patient registry (https://www.dm-registry.org.uk/). Data from 157 patients with SCA2 and *ATXN2* allele size equal to or higher than 35 repeats from EUROSCA were used to model the prevalence of SCA2 (http://www.eurosca.org/). From the same registry, data from 91 patients with SCA1 and *ATXN1* allele sizes equal to or higher than 44 repeats and from 107 patients with SCA6 and *CACNA1A* allele sizes equal to or higher than 20 repeats were used to model the disease prevalence of SCA1 and SCA6, respectively.

As some REDs have reduced age-related penetrance, e.g. *C9orf72*-carriers may not develop symptoms even after 90 years of age (Murphy et al., 2017), age-related penetrance was obtained as follows: as regards *C9orf72*-ALS/FTD, it was derived from the red curve in Figure 2 of that paper (data available at https://github.com/nam10/C9_Penetrance) reported by Murphy et al., 2017 and was used to correct *C9orf72*-ALS and *C9orf72*-FTD prevalence by age.

Both the general UK population and the age at onset distribution of each disease were tabulated. After standardisation over the total number, the onset count was multiplied by the carrier frequency of the genetic defect and then multiplied by the corresponding general population count for each age group, to obtain the estimated number of people in the UK developing each specific disease by age group. This estimate was further corrected by the age-related penetrance of the genetic defect, where available (e.g., *C9orf72*-ALS and *C9orf72*-FTD). Finally, to account for disease survival, I performed a cumulative distribution of prevalence estimates grouped by a number of years equal to the median survival length for that disease. The median survival length (n) used for this analysis is: 3 years for *C9orf72*-ALS, 10 years for *C9orf72*-FTD (Glasmacher et al., 2020), 15 years for SCA2 and for SCA1 (Diallo et al., 2018), 15 years for HD (40 CAG repeat carriers); for SCA6 normal life expectancy was assumed. For DM1, since life expectancy is partly related to the age of onset, the mean age of death was assumed to be 45 years for patients with childhood-onset and 52 years for early adult-onset patients (10-30 years) (Mathieu et al., 1999), while no age of death was set for DM1 patients with onset after 31 years. Since survival is approximately 80% after 10 years (Wahbi et al., 2018), I subtracted 20% of the predicted affected individuals after the first 10 years. Then, survival was assumed to proportionally decrease in the following years until the mean age of death for each age group was reached.

The resulting estimated prevalences of *C9orf72*-ALS/FTD, SCA2, and SCA1 by age group were plotted in Figure 17 (dark blue area). The literature-reported

prevalence by age for each disease was obtained by dividing the new estimated prevalence by age by the ratio between the two prevalences, and represented as a light blue area.

To compare the new estimated prevalence to the clinical disease prevalence reported in the literature for each disease, I employed figures calculated in European populations, as it is closer to the UK population in terms of ethnic distribution:

i) *C9orf72*-FTD: the median prevalence of FTD was obtained from studies included in the systematic review by Hogan and colleagues (Hogan et al., 2016) (83.5 in 100,000). Since 4%-29% of FTD patients carry a *C9orf72* repeat expansion (Van Mossevelde et al., 2018), I calculated *C9orf72*-FTD prevalence by multiplying this proportion range by median FTD prevalence (3.3 - 24.2 in 100,000, mean 13.78 in 100,000);

ii) *C9orf72*-ALS: The reported prevalence of ALS is 5-12 in 100,000 (Gossye et al., 1999), and *C9orf72* repeat expansion is found in 30%-50% of individuals with familial forms and in 4%-10% of people with sporadic disease (Zampatti et al., 2022). Given that ALS is familial in 10% of cases and sporadic in 90%, I estimated the prevalence of *C9orf72*-ALS by calculating the [(0.4 of 0.1) + (0.07 of 0.9)] of known ALS prevalence of 0.5-1.2 in 100,000 (mean prevalence is 0.8 in 100,000);

iii) HD prevalence ranges from 0.4 in 100,000 in Asian countries to 10 in 100,000 in Europeans, and the mean prevalence is 5.2 in 100,000. 40-CAG repeat carriers represent 7.4% of patients clinically affected by HD according to the Enroll-HD version 6. Considering an average reported prevalence of 9.7 in 100,000 Europeans, I calculated a prevalence of 0.72 in 100,000 for symptomatic 40-CAG carriers;

iv) DM1 is much more frequent in Europe than in other continents, with figures of 1 in 100,000 in some areas of Japan. A recent meta-analysis has found an overall prevalence of 12.25 per 100,000 individuals in Europe, which I used in this analysis.

Given that the epidemiology of autosomal dominant ataxias varies among countries (De Mattei et al., 2023) and no precise prevalence figures derived from clinical observation are available in the literature, I approximated SCA2, SCA1, and SCA6 prevalence figures to be equal to 1 in 100,000.

2.6 Distribution of repeat lengths in different populations

Analysis of repeat size distribution

The distribution of each of the 16 repeat expansion loci, where the WGS pipeline enabled discrimination between the premutation/reduced penetrance and the full mutation, was analysed across the 100K GP and TOPMed datasets (Figure 18). For each gene, the distribution of the repeat size across each ancestry subset was visualised as a density plot and as a boxplot; moreover, the 99.9th percentile and the threshold for intermediate and pathogenic ranges were highlighted.

Correlation between intermediate and pathogenic repeat frequency

The percentage of alleles in the intermediate and in the pathogenic range (premutation plus full mutation) was computed for each population (combining data from 100K GP with TOPMed) for genes with a pathogenic threshold below or equal to 150 bp. The intermediate range was defined as either the current threshold reported in the literature (La Spada, 2022; Opal and Ashizawa, 2023; Pulst, 2019) (*ATXN1*=36; *ATXN2*=31; *ATXN7*=28; *CACNA1A*=18; *HTT*=27) or as the reduced penetrance/premutation range according to Figure 14 for those genes where the intermediate cutoff is not defined (*AR, ATN1, DMPK, JPH3, TBP*). Genes where either the intermediate or pathogenic alleles were absent across all populations were excluded. Intermediate and pathogenic allele frequencies (percentages), split by population, were displayed as a scatter plot using R and the package tidyverse, and correlation was assessed using Spearman's rank correlation coefficient with the package ggpubr and the function stat_cor (Figure 19).

2.7 Novel RNA-seq data generation

Dox-inducible TDP-43 knockdown

SH-SY5Y and SK-N-BE(2) cells were transduced with SmartVector lentivirus (V3IHSHEG_6494503) containing a shRNA cassette for *TARDBP*, which can be induced by doxycycline. Transduced cells were selected with puromycin (1 μg/ml) for one week. TDP-43-knockdown SH-SY5Y and SK-N-BE(2) cells were plated as single cells and expanded to obtain a clonal population. Cells were grown in DMEM/F12

containing Glutamax (Thermo) supplemented with 10% FBS (Thermo) and 1% PenStrep (Thermo). For induction of shRNA against *TARDBP*, cells were treated for 10 days with increasing amounts of doxycyline hyclate (Sigma), as follows:

- For experiments in SH-SY5Y cells, 12.5 ng/ml, 18.75 ng/ml, 21 ng/ml, 25 ng/ml, and 75 ng/ml.
- For experiments in SK-N-BE(2), 20 ng/ml, 30 ng/ml, 40 ng/ml, 50 ng/ml, 75 ng/ml, 100 ng/ml, and 1000 ng/ml.

NMD inhibition in human cell lines

For the NMD inhibition experiment with cycloheximide (CHX) in SH-SY5Y, 10 days after the induction of shRNA against TDP-43 with 25 ng/ml doxycyline hyclate (Sigma), cells were treated either with 100 μM CHX or DMSO (Pereverzev et al., 2015) for 6 h before isolating the RNA. Overall, 4 conditions were generated:

- Control for TDP-43 knockdown and control for NMD (control/control)
- Control for TDP-43 knockdown and CHX-treated (control/CHX)
- Positive for TDP-43 knockdown and control for NMD (TDP43/control)
- Positive for TDP-43 knockdown and CHX-treated (TDP43/CHX)

Human iPSC culture and CRISPRi

The iPS cells used were from the WTC11 line, derived from a healthy 30-year-old male, and obtained from the Coriell cell repository. Informed consent was obtained from the donor. The iPS cells were engineered (Fernandopulle et al., 2018; Wang et al., 2017) to express mouse or human neurogenin-2 under a doxycycline-inducible promoter, as well as an enzymatically dead Cas9 (+/− CAG-dCas9-BFP-KRAB) (Tian et al., 2019). To achieve knockdown, sgRNAs targeting either a control guide (GTCCACCCTTATCTAGGCTA), *UPF1* (GGCCAGACGCAGACGCCCCC), or *TARDBP* (GGGAAGTCAGCCGTGAGACC) were delivered to iPS cells by lentiviral transduction (Lois et al., 2002), for a total of 4 groups:

- Control for TDP-43 knockdown and control for NMD (control/control)
- Control for TDP-43 knockdown and CHX-treated (UPF1KD/control)
- Positive for TDP-43 knockdown and control for NMD (control/TDP43KD)

- Positive for TDP-43 knockdown and CHX-treated (UPF1KD/TDP43KD)

After lentiviral transduction, cells were plated in tissue culture dishes coated with human embryonic stem cell-qualified Matrigel (Corning). The following morning, cells were washed with PBS and the media was changed to E8 alone or (depending on cell number) E8 supplemented with 10 μM ROCK inhibitor in a 37 °C, 5% CO2 incubator. Cells were passaged with accutase (Life Technologies) for 5–10 min at 37 °C and then washed with PBS before re-plating. Two days after lentiviral delivery, cells were selected overnight with either puromycin (10 μg/ml) or blasticidin (50–100 μg/ml). iPS cells were then expanded for 2 days before initiating neuronal differentiation. To initiate neuronal differentiation, 20–25 million iPS cells per 15 cm plate were individualised using accutase on day 0 and re-plated onto Matrigel-coated tissue culture dishes in N2 differentiation media containing: knockout DMEM/F12 media (Life Technologies Corporation) with N2 supplement (Life Technologies Corporation), 1× GlutaMAX (Thermofisher Scientific), 1× MEM nonessential amino acids (Thermofisher Scientific), 10 μM ROCK inhibitor (Selleckchem) and 2 μg/ml doxycycline (Clontech). The media was changed daily during this stage. On day 3 pre-neuron cells were replated onto dishes coated with freshly made poly-L-ornithine (0.1 mg/ml; Sigma), 6-well dishes (2 million per well), in i3Neuron Culture Media: Brain-Phys media (Stemcell Technologies) supplemented with 1× B27 Plus Supplement (ThermoFisher Scientific), 10 ng/ml BDNF (PeproTech), 10 ng/ml NT-3 (PeproTech), 1 μg/ml mouse laminin (Sigma), and 2 μg/ml doxycycline (Clontech). i3Neurons were then fed three times a week by half media changes. i3Neurons were then collected on day 7 after the addition of doxycycline.

Human muscle samples

Publicly available data from (Britson et al., 2022) (n = 2 controls, n = 2 sIBM) were requested and obtained from the authors (Baltimore cohort). Novel data included a cohort from Padua, including n = 4 controls and n = 4 sIBM, and a cohort from Trieste, including n = 5 controls and n = 7 sIBM. All cases provided written consent for their samples and data to be used in research, including this study; they were selected based on the following inclusion and exclusion criteria:

- sIBM group: Any person diagnosed with sIBM by a qualified clinician, according to internationally accepted diagnostic criteria (Griggs et al., 1995; Rose, 2013).

- Controls: Normal muscle tissue from patients investigated for cramps or fatigue with normal examination, neurophysiology tests and normal histology.

For the Padua cohort, sequencing libraries were prepared with polyA enrichment and sequenced at UCL Genomics with the following specifics: 2 × 150 bp, depth > 40M/sample. For the Trieste cohort, sequencing libraries were prepared with polyA enrichment and sequenced by Novogene with the following specifics: 2x150 bp, depth > 40M/sample.

2.8 RNA analysis

RNA isolation and QC

RNA was extracted from SH-SY5Y, SK-N-BE(2) and i3Neurons with an RNeasy mini kit (Qiagen) following the manufacturer's protocol, including the on-column DNA digestion step. RNA concentrations were measured by Nanodrop, and 500–1000 ng of RNA was used for reverse transcription. Samples undergoing RNA sequencing were assessed for RNA quality on a TapeStation 4200 (Agilent), and bands were quantified with TapeStation Systems Software v3.2 (Agilent). For SH-SY5Y and SK-N-BE(2), the RNA integrity number (RIN) was above 9.4. For i3Neurons, it was above 6.9.

RT-qPCR

Reverse transcription was performed from 500–1000 ng of RNA using RevertAid cDNA synthesis kit (Thermo) using random hexamer primers and following the manufacturer's protocol. Transcript levels were quantified by qPCR (QuantStudio 5 Real-Time PCR system, Applied Biosystems) using the ΔΔCt method (Livak and Schmittgen, 2001). Using RefFinder (Xie et al., 2023), I identified *GAPDH* as the most stable endogenous control across the conditions of interest; primers for *GAPDH* are listed in Table III.

For NMD experiments in SH-SY5Y cells, the *HNRNPL* NMD transcript was used as a positive control since it has been shown to undergo NMD (Humphrey et al., 2020). Primers to amplify *HNRNPL* NMD transcript, *UNC13A* CE, *UNC13B* CE, and *STMN2* CE are found in Table III. Gene expression analysis of *TARDBP* and *UPF1* was assessed by means of qPCR, using the primers listed in Table III.

Primer name	Sequence
GAPDH forward	5′-CACCAGGGCTGCTTTTAACT-3′
GAPDH reverse	5′-GACAAGCTTCCCGTTCTCAG-3′
HNRNPL NMD forward	5′-GGTCGCAGTGTATGTTTGATG-3′
HNRNPL NMD reverse	5′-GGCGTTTGTTGGGGTTGCT-3′
UNC13A CE forward	5′-CAAGCGAACTGACAAATCTGCCGTGTCG-3′
UNC13A CE reverse	5′-CCTGGAAAGAACTCTTATCCCCAGGAACTAGTTTGTTG-3′
UNC13B CE forward	5′-TCCGAGCAGTTACCAAGGTT-3′
UNC13B CE reverse	5′-GAAAAGCGAGGAGCCCTTCAG-3′
STMN2 CE forward	5′-GCTCTCTCCGCTGCTGTAG-3′
STMN2 CE reverse	5′-CTGTCTCTCTCTCTCGCACA-3′
TARDBP forward	5′-GATGGTGTGACTGCAAACTTC-3′
TARDBP reverse	5′-CAGCTCATCCTCAGTCATGTC-3′
UPF1 forward	5′-TCGAGGAAGATGAAGAAGACAC-3′
UPF1 reverse	5′-TCCGTTGCAGAACCACTTC-3′

Table III: Primers used for qPCR.

RNA sequencing, differential gene expression and splicing analysis

Sequencing libraries were prepared with polyA enrichment using a TruSeq Stranded mRNA Prep Kit (Illumina) and sequenced on an Illumina HiSeq 2500 or NovaSeq 6000 machine at UCL Genomics with the following specifics:

- SH-SY5Y cycloheximide and curves: 2 × 100 bp, depth > 40M/sample
- BE(2) curves and UPF1 iPSCs: 2x150 bp, depth > 40M/sample

Raw sequences (in fastq format) were trimmed using Fastp (S. Chen et al., 2018) with the parameter "qualified_quality_phred: 10", and aligned using STAR (v2.7.0f) (Dobin et al., 2013) to the GRCh38 with gene models from GENCODE v40 or v42 (the latter for the muscle biopsy samples) (Frankish et al., 2019). The STAR outputs are BAM files, tab-delimited files containing the counts and coordinates for all splice junctions found in the sample, and alignment metadata including the genomic location where a read is mapped to, read sequence, and quality score. Then, using samtools (H. Li et al., 2009), BAM files were sorted and indexed by the read coordinates location. Trimmed fastqc files were aligned to the transcriptome using Salmon (v1.5.1) (Patro et al., 2017), outputting isoform-specific counts used for differential gene expression, performed using DeSEQ2 (Love et al., 2014) without covariates, using an index built from GENCODE v40 or v42 (the latter for the muscle biopsy samples) (Frankish et al., 2019). The DESeq2 median of ratios, which controls for both sequencing depth

and RNA composition, was used to normalise gene counts. Differential expression was defined at a Benjamini–Hochberg false discovery rate < 0.1. I kept genes with at least 5 counts per million in more than 2 samples. Splicing analysis was performed using Majiq (v2.1) (Vaquero-Garcia et al., 2016) on STAR-aligned and sorted BAMs and the GRCh38 reference genomes. MAJIQ's outputs were used to categorise each of the splicing junctions by the overlap with annotated transcripts and exons, using GTF and DASPER (https://github.com/dzhang32/dasper), classifying them into "annotated", "novel_acceptor", "novel_donor", "novel_exon_skip", "novel_combo" and "ambig_gene" and "none". Cryptic splicing was defined as junctions with Ψ (PSI, percent spliced in) $< 5\%$ in the control samples, $\Delta\Psi > 10\%$ between groups, and the junctions were unannotated in GENCODE v40 (Frankish et al., 2019). In some cases, for a cryptic exon, there was only enough coverage to call one of the two defining junctions significant. QC reports were generated with FASTQC and then summarised using MultiQC (Ewels et al., 2016). The alignment and splicing pipelines are implemented in Snakemake (v5.5.4) (Mölder et al., 2021), a workflow management software that allows reliable, reproducible, and scalable analysis, and runs on the UCL high-performance cluster (HPC). PCA analysis and data visualisation were run on R using custom scripts made publicly available.

Publicly available data

For public RNA-seq datasets, raw fastqs were downloaded from the Sequence Read Archive (SRA) using "fasterq-dump" from the sratoolkit v2.9.6. FACS-sorted frontal cortex neuronal nuclei data were obtained from Gene Expression Omnibus (GSE126543).

Information about gene expression in different tissues was obtained from ASCOT (Ling et al., 2020). Genes of interest (Table IX) were selected and displayed using the tidyverse package in R.

Parsing of CE junctions

NYGC cohort

To check the disease and tissue specificity of CEs found in cells in postmortem brains, I interrogated the New York Genome Center (NYGC) RNA-seq cohort. The present analysis included 1,682 tissue samples from 446 individuals (104 non-neurological controls, 279 ALS, and 63 FTD) from the NYGC ALS dataset. All non-*SOD1/FUS* ALS samples were grouped as "ALS-TDP" in this work for simplicity. FTD samples were classified according to the pathological diagnosis of FTD with TDP-

43 inclusions (FTD-TDP), or without (FTD-non-TDP). Sample processing, library preparation, and RNA-seq quality control have been described before (Prudencio et al., 2020; Tam et al., 2019).

RNA-seq samples were uniformly processed, including adapter trimming with Trimmomatic and alignment to the hg38 genome build using STAR (2.7.2a) (Dobin et al., 2013) with indexes from GENCODE v30 (Frankish et al., 2019). Extensive quality control was performed using SAMtools (H. Li et al., 2009) and Picard Tools (https://broadinstitute.github.io/picard/) to confirm the sex and tissue of origin.

Junctions from each sample were then extracted with RegTools (Cotto et al., 2023) using a minimum of 8 bp as an anchor on each side of the junction and a maximum intron size of 500 kb and clustered together using LeafCutter (Y. I. Li et al., 2018) with relaxed junction filtering (minimum total reads per junction = 30, minimum fraction of total cluster reads = 0.0001). This produced a matrix of junction counts across all samples. A CE was considered detected in a sample if there was at least one uniquely mapped spliced read supporting either the CE acceptor or the CE donor. Cryptic splice junctions masked by nonsense-mediated decay were exported as a BED file that was used as a query to quantify CE expression in the NYGC RNA-seq data. This returned a table with coordinates, gene name, and how many spliced reads in each sample supported that junction, as well as sample code, disease, and tissue. This pipeline uses bedtools (Quinlan and Hall, 2010) and bash parsing with awk and is implemented in Snakemake v5.5.4 (Mölder et al., 2021).

Muscle cohort

To assess the presence of CE peptides (Table IX) in the muscle RNA-seq cohort, I used the pipeline described above. A CE was considered detected in a sample if there was at least one uniquely mapped spliced read supporting either the CE acceptor or the CE donor, and only the higher of the two was considered. Cryptic splice junctions were exported as a BED file used as a query to quantify CE expression in the sIBM RNA-seq data. This returned a table with coordinates, gene name, and the number of spliced reads per sample supporting that junction.

Targeted RNA-seq

To detect and quantify TDP-43 CEs in postmortem brains, I designed a targeted RNA sequencing experiment based on a 3-primer PCR followed by Illumina sequencing. After collecting fresh frozen tissue from the frontal cortices of 10 FTD patients and 4 controls (Table IV) from UCL Queen Square Brain Bank, RNA was extracted

using the miRNeasy mini kit (Qiagen, 217004) and reverse-transcribed using the SuperScript™ IV First-Strand Synthesis System (Thermo Fisher, 18091050) according to the manufacturer protocols. This was followed by two rounds of PCRs using Q5 High-Fidelity 2X Master Mix (New England Biolabs, M0492). In the first round, I used a pool of primers targeting a set of well-characterised CEs Appendix Table 5. The annealing temperature was set to 60 °C for 10 cycles. After clean-up using Mag-Bind TotalPure NGS (Omega Bio-tek, M1378), a second PCR was performed to bind the Illumina adapters and to pool the samples. Pooled samples were purified using Mag-Bind TotalPure NGS (Omega Bio-tek, M1378) and sent for Illumina sequencing at The Francis Crick Institute Genomic STP. Alignment was performed as described in the previous paragraph. To quantify CEs, I computed the fraction of reads covering the CE junction over the total junction number for a given gene. Data visualisation was performed in R 4.3.2, using the ggplot2 package.

Sample name	Disease duration	Age	Sex	Pathology	Mutations
Control1	NA	80	Female	Control	None
Control2	NA	38	Male	Control (minimal amyloid deposition)	None
Control3	NA	79	Female	Control	None
Control4	8	60	Female	AD (PCA)	None
FTD1	5	62	Female	FTD-TDP-A	*C9orf72*
FTD2	3	80	Female	FTD-TDP-C	None
FTD3	15	73	Female	FTD-TDP-C	None
FTD4	10	74	Male	FTD-TDP-C	None
FTD5	8.8	67	Female	FTD-TDP-A	*C9orf72*
FTD6	23	67	Male	FTD-TDP-C	None
FTD7	6	68	Male	FTD-TDP-A	*C9orf72*
FTD8	15	65	Male	FTD-TDP-C	None
FTD9	10	72	Male	FTD-TDP-A	*TBK1*
FTD10	14	78	Male	FTD-TDP-C	None

Table IV: Cohort used for targeted RNA sequencing.

2.9 Protein analysis on human skeletal muscle

Biopsy collection

The use of samples for the present study was approved by Brain UK. Anonymisation of study samples was performed by a member of the clinical care team who was not directly involved in the study. I selected cases based on the following inclusion

and exclusion criteria:

- sIBM group: Any person diagnosed with sIBM by a qualified clinician, according to internationally accepted diagnostic criteria (Griggs et al., 1995; Rose, 2013).
- Controls: Normal muscle tissue from patients investigated for cramps or fatigue with normal examination, neurophysiology tests and normal histology.

Quadriceps muscles from 17 sIBM and 4 control cases were biopsied, flash-frozen in liquid nitrogen, and stored at -80 °C in the Neuropathology department of the National Hospital for Neurology and Neurosurgery (NHNN) in London, UK.

Immunohistochemistry

For immunohistochemistry, 8 μm slides were cut with a cryostat. DAB-staining consisted of 10 minutes of fixation with 4% PFA, followed by blocking endogenous peroxidases with 3% H2O2 in methanol. Preblocking in 10% milk/TBS-T for 30 minutes was followed by a 2-hour incubation with primary antibodies against TDP-43 (mouse monoclonal, Abnova, catalogue number H00023435-M01, final dilution 1:800), p62 (mouse monoclonal, Abcam, catalogue number ab56416, final dilution 1:100), or HDGFL2-CE (rabbit polyclonal, donated by Leonard Petrucelli, final dilution 1:500). After a 30-minute incubation with secondary antibodies, I incubated the slides for 30 minutes with ABC complex and DAB staining for 3 minutes, followed by counterstaining in Mayer's hematoxylin and mounting overnight before visualisation on the microscope and scanning. Quantification of immune infiltration, TDP-43, p62 inclusions, and HDGFL2-CE burden analysis was performed in a blind, semiquantitative fashion (Meyerholz and Beck, 2018), and confirmed by a neuropathologist.

Proteomics

4 25 μm thick flash-frozen muscle biopsies from each sIBM and control case were collected and 20 μg of tissue were dissolved in 100 μl of lysis buffer, consisting of 50 mM Tris-HCl pH 8, 50 mM NaCl, 1% v/v SDS, 1% v/v Triton X-100, 1% v/v NP-40, 1% v/v Tween 20, 1% v/v glycerol, 1% sodium deoxycholate (w/v), 5 mM EDTA pH 8, 5 mM dithiothreitol, 5 KU benzonase and protease inhibitors. Following homogenisation and spinning at 20000g at 4 °C for 15 minutes, cells were snap-frozen. Samples were sent to Andy Qi's facility at CARD, NIH, where they were processed by liquid chromatography and tandem mass spectrometry (LC-MS/MS) and analysed using data-independent acquisition (DIA). For DIA database search,

direct DIA-library was used in Spectronaut (v16) (Martinez-Val et al., 2021) with a customised database containing UniProt-human proteome reference (UP000005640). Trypsin was chosen as a digestion enzyme (allowing semi-tryptic digestion), mass error tolerance was set to 15ppm, and 3 mis-cleavages were allowed. The statistical analysis was conducted in RStudio (v4.3), using the tidyverse package and custom scripts (https://github.com/mattzano/muscle_splicing); a two-sided t-test was used for comparison, and p-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure.

2.10 AlphaFold2 and TCRmodel2 prediction

To assess the potential impact of the inclusion of an additional coding exon to the HDGFL2 protein structure, I employed Alphafold2 (Jumper et al., 2021), using the monomer settings. The input sequences included the native HDGFL2 sequences from UniProt Q7Z4V5, and the one including its cryptic peptide (Table IX). The resulting structure files were visualised on https://molstar.org/viewer/. Sequences from three T-cell receptors (TCRs) (validated by Shawn Chizari in the Jiang Lab at uPenn) were combined with their corresponding peptide-Major Histocompatibility Complex (pMHC) or with one of the other two (Table V). Input sequences were uploaded to the TCRmodel server (https://tcrmodel.ibbr.umd.edu/) (Yin et al., 2023). Best-ranked model iPTM scores (pTM, corresponding to overall topological accuracy, calculated for interchain interfaces) for the interaction between matched and unmatched TCR:pMHC were tested by ANCOVA, with the TCR as a covariate.

TCR	HLA	target	TCR:pMHC	iPTM	iPTM_TCR:pMHC
CE-TCR-14	B0801	HDGFL2	matched	0.888	0.872
CE-TCR-14	A0301	IgLON5	unmatched	0.875	0.867
CE-TCR-14	A0201	pp65CMV	unmatched	0.873	0.857
CE-TCR-30	A0301	IgLON5	matched	0.892	0.864
CE-TCR-30	B0801	HDGFL2	unmatched	0.851	0.804
CE-TCR-30	A0201	pp65CMV	unmatched	0.819	0.751
pp65CMV-TCR	A0201	pp65CMV	matched	0.884	0.863
pp65CMV-TCR	B0801	HDGFL2	unmatched	0.769	0.698
pp65CMV-TCR	A0301	IgLON5	unmatched	0.718	0.622

Table V: TCR modelling data.

2.11 Data and code availability

The alignment, splicing, splice junction parsing pipelines and scripts for downstream analysis are implemented in Snakemake version 5.5.4 and R 4.3.2, and are available at:

- https://github.com/frattalab/rna_seq_snakemake
- https://github.com/frattalab/splicing
- https://github.com/frattalab/bedops_parse_star_junctions
- https://github.com/mattzano/splicing_chx
- https://github.com/mattzano/tdp43_concentration
- https://github.com/mattzano/muscle_splicing

For the 100K GP, full data is available in the Genomic England Secure Research Environment. Access is controlled to protect the privacy and confidentiality of participants in the Genomics England 100,000 Genomes Project and to comply with the consent given by participants for the use of their healthcare and genomic data. Access to full data is permitted through the Research Network (https://www.genomicsengland.co.uk/research/academic/join-research-network).

For TOPMed, a detailed description of the TOPMed participant consents and data access is provided in Taliun et al., 2021. The TOPMed data used in this manuscript are available through dbGaP. A complete list of TOPMed genetic variants with summary-level information used in this manuscript is available through the BRAVO variant browser (bravo.sph.umich.edu). The TOPMed imputation reference panel described in this manuscript can be used freely for imputation through the NHLBI BioData Catalyst at the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/). DNA sequence and reference placement of assembled insertions are available in VCF format (without individual genotypes) on dbGaP under the TOPMed GSR accession phs001974. For the 1000 Genomes Project, the WGS datasets are available from the European Nucleotide Archive under accessions PRJEB31736 (unrelated samples) and PRJEB36890 (related samples).

The following GitHub repositories used in this work are free to access:

- Expansion Hunter v3.2.2 (to estimate the repeat size within defined loci)
- REViewer v0.2.7 (to generate pileup plots for quality check)

- ExpansionHunter_Classifier (March 2024 release, to automatically run quality assessment of EHv322 call)
- gvcfgenotyper (to merge gVCF files, when inferring ancestry across genomes within the 100K GP and TOPMed datasets)
- https://github.com/nam10/C9_Penetrance (to compute survival curve analysis for C9orf72)
- www.cog-genomics.org/plink/2.0/ PLINK2
- (https://github.com/dzhang32/dasper) DASPER
- https://tcrmodel.ibbr.umd.edu/ TCRModel2
- https://www.medcalc.org/calc/diagnostic_test.php MedCalculator

Chapter 3

Genetic prevalence of repeat expansion disorders

3.1 Contributions

Parts of the text in this chapter are copied from the works "Unexpected Frequency of the Pathogenic AR CAG Repeat Expansion in the General Population" (Zanovello et al., 2023) and "Increased Frequency of Repeat Expansion Mutations across Different Populations" (Ibañez et al., 2024), and therefore, other people contributed to ideas, drafting, and writing. Credits for contributions can be found in the UCL Research paper declaration forms. Particularly, EH software was run by Dr. Kristina Ibanez and Dr. Bharati Jadhav. The disease modelling was done in collaboration with Dr. Delia Gagliardi.

3.2 Results

AR CAG repeat expansion prevalence in the general population

The information on the frequency of REDs has relied on epidemiology studies or PCR screening of selected populations. Epidemiological studies report a 1:30303 or less prevalence amongst male populations (Bertolin et al., 2019; Guidetti et al., 2001; Zelinkova et al., 2016), but SBMA is often reported to be underdiagnosed. However, an epidemiological study in the Vasa region of Finland reported 13 cases in a population of 85000 males (1:6538), although this was attributed to a founder effect (Udd et al., 1998); two studies based on PCR sizing in selected populations reported an unexpectedly high frequency of this genetic defect, namely a PCR screening of a European population, which found the mutation frequency to be 1:6888 X

chromosomes (Gardiner et al., 2019); and a meta-analysis of 86 datasets based on PCR sizing reported a population frequency of 1:3703 (Laskaratos et al., 2021).

Cohort	Gender	Phenotype category	Total participants	Total X chromosomes	Total expansions ≥38	X chromosome frequency ≥38 (95% CI)
100k GP	Male	Non–neuro	13 072	13 072	2	1/6536 (1793–23833)
	Female	All	20 400	40 800	11	1/3709 (2071–6642)
gnomAD	Male	All	14 947	14 947	5	1/2989 (1277–6998)
	Female	All	14 116	28 232	11	1/2567 (1433–4596)
NIH	Male	Ctrl	1529	1529	1	1/1529 (271–8661)
	Female	All	5176	10 352	2	1/5176 (1420–18874)
MinE	Male	Ctrl	1272	1272	2	1/636 (175–2319)
	Female	All	3765	7530	3	1/2510 (854–7380)
Summary	Male	–	30 820	30 820	10	1/3082 (1674–5674)
	Female	–	43 457	86 914	27	1/3219 (2213–4683)
	All	–	74 277	117 734	37	1/3182 (2309–4386)

Table VI: **AR repeat expansion frequency.**

Although next-generation sequencing and public genomic data repository technologies have allowed the frequency of single nucleotide variants to be estimated precisely across very large populations (Karczewski et al., 2020), the inability to reliably size STRs from whole-genome sequencing (WGS) has not permitted the same information to be gathered for STR expansions, which are a major cause of neurogenetic disorders including SBMA. Recently developed bioinformatics tools, such as EH, allow the sizing of STRs from WGS data (Dolzhenko et al., 2017).

Given the unexpected findings from population studies (Gardiner et al., 2019; Laskaratos et al., 2021) and considering the limitations of PCR sizing and the use

of selected populations, I sought to investigate the frequency of the genetic variant underlying SBMA in the general population by exploiting WGS and using clinically curated public genomic data repositories. I validated this approach, applied it to the 100k GP cohort (The 100,000 Genomes Project Pilot Investigators et al., 2021) and replicated it on three other large WGS datasets (Table VI).

A sensitive and specific pipeline to detect *AR* CAG expansions

The WGS analysis pipeline to analyse the *AR* expansion combines EH with visual validation of positive results, in accordance with recent guidelines from the American College of Medical Genetics (Figure 9) (Ibañez et al., 2022; Roy et al., 2018). I benchmarked this pipeline against the gold standard diagnostic method, PCR. I used 133 alleles from 97 samples where the WGS pipeline identified PCR-confirmed expanded (n = 38) and normal (n = 94) alleles, resulting in a sensitivity of 100% (95% CI 90.8–100%), specificity 99% (95% CI 94.2–99.7%), and positive predictive value of 97.4% (95% CI 84.4–99.6%) (Table VII). Size estimation correlation yielded R = 0.99 ($P < 2.2 \times 10\text{-}16$), with high accuracy in alleles with less than 38 repeats, whilst larger repeats were determined to be in the pathogenic range, but less accurately sized as previously shown (Figure 10) (Dolzhenko et al., 2019).

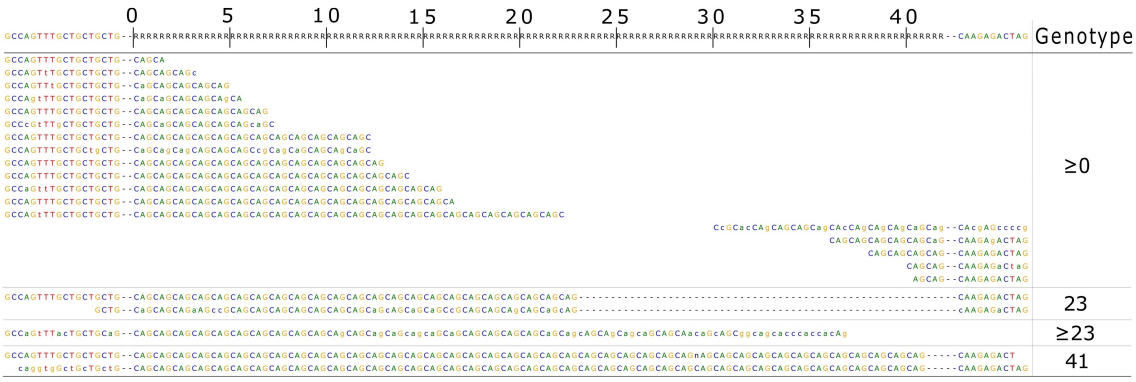


Figure 9: **Visualisation of repeat expansion reads from EH shows reads revealing 23 and 41 CAG repeat alleles.**

Parameter	Value (95% confidence interval (CI))
Sensitivity	100% (90.8–100%)
Specificity	99.0% (94.2–99.7%)
Positive predictive value	97.4% (84.4–99.6%)

Table VII: **Sensitivity, specificity, and positive predictive value for *AR* pathogenic expansion detection.**

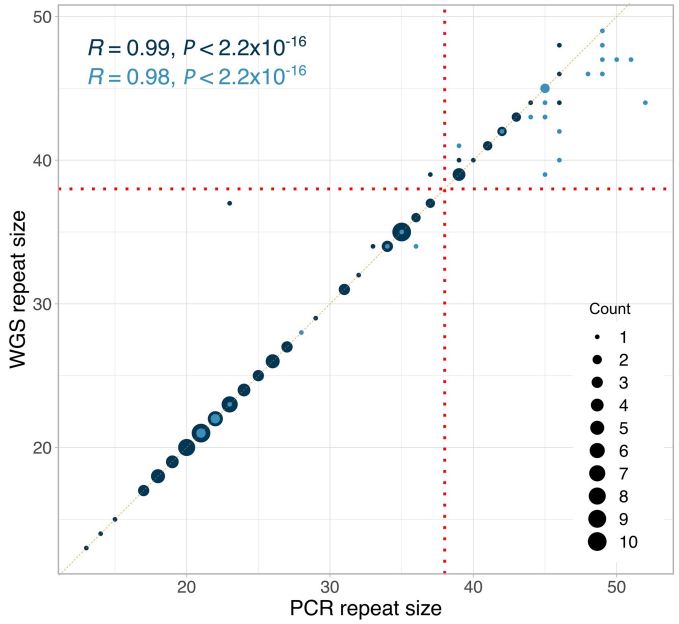


Figure 10: **Comparison of repeat size estimation between WGS pipeline and PCR.** N = 133 alleles. Dark points indicate length confirmed by reads spanning the whole repeat and both the flanking sides; light points indicate length confirmed by reads spanning part of the repeat and one flanking side.

Unexpected frequency of pathogenic *AR* CAG expansions

The 100k GP sequenced the whole genomes of people with a wide range of rare diseases and cancers in the National Health Service in England. Individuals were recruited with their family members where available (The 100,000 Genomes Project Pilot Investigators et al., 2021). The *AR* allele size distribution in 75,035 individuals from this cohort showed a typical bell shape with a peak at 21 repeats (Figure 11). Analysis of 40,412 unrelated individuals within this cohort identified 25 people carrying pathogenic repeats (≥38 repeats), including 11 females and 14 males. Clinical data available for each individual recruited to the 100k GP, including International Statistical Classification of Diseases Revision 10 (ICD-10) codes and Human Phenotype Ontology (HPO) terms, were reviewed. Of the 14 males, seven proved to have a clinically confirmed diagnosis of SBMA, whilst all remaining individuals were under 21 years of age, except for one recruited for retinal disorders (Table VIII). None of the female carriers, who can generally develop mild symptoms, had HPO terms associated with neuromuscular conditions, although some of them had HPO terms associated with other neurological diseases. To estimate the frequency of *AR* pathogenic expansions, the repeat size was assessed in all unrelated female and male individuals. To avoid overestimating the frequency due to individuals being recruited because of SBMA-related symptoms, all males recruited

under 'neurological disorders' were excluded. I found the X chromosome frequency of the pathogenic expansion to be 1:6536 (95% CI 1:1793–1:23 833, n = 13 072) and 1:3709 (95% CI 1:2071–1:6642, n = 40 800) in males and females, respectively (Table VI and Figure 12).

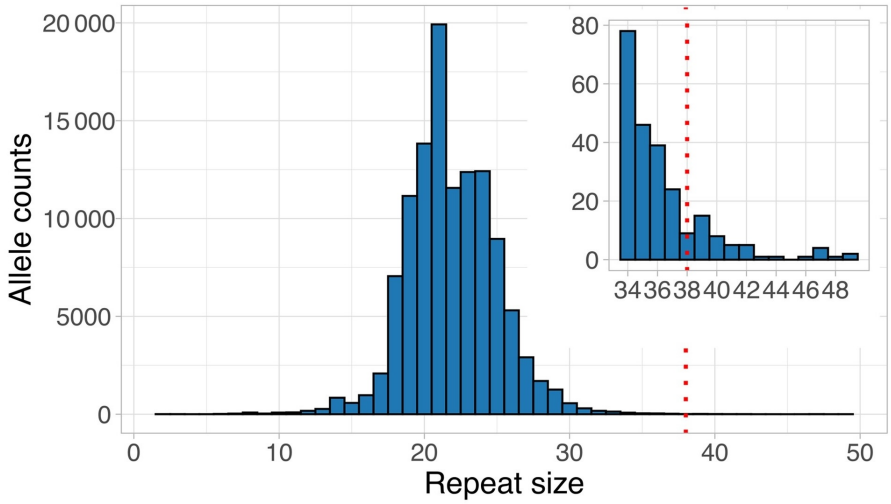


Figure 11: **Allele size distribution across 75,035 100k GP genomes; inset highlights the distribution of alleles containing ≥34 repeats.**

Age group (years)	Gender	Neurology and neurodevelopmental disorders	Other diseases
≦20	Males	5	1 (Cancer)
≦20	Females	2	1 (Ophthalmo)
>20	Males	7*	1 (Ophthalmo)
>20	Females	3	1 (Renal), 1 (Cardio), 3 (Cancer)

Table VIII: **Clinical data for 100K GP samples with an *AR* CAG ≥38 repeats**
*SBMA diagnosis confirmed with local clinician

Multiple large cohorts confirm *AR* CAG expansion frequency

Given the surprisingly high frequency of the *AR* repeat expansion, I sought to carry out the analysis on replication datasets, using North American (NIH and gnomAD) and European (Project MinE) cohorts, where control and neurodegenerative diseases were sequenced with WGS (Project MinE ALS Sequencing Consortium, 2018). The AR expansion frequency was 1:2989 and 1:2567 X chromosomes in all males (n = 14947) and all females (n = 28232), respectively, in the gnomAD cohort, 1:1529 and 1:5176 X chromosomes in control males (n = 1529) and all females (n = 10352),

respectively, in the NIH cohort, and 1:636 and 1:2510 X chromosomes in control males (n = 1272) and all females (n = 7530), respectively, in the MinE cohort (Figure 12). Estimates of *AR* expansion frequency from these cohorts fall within the 95% CI of the frequency estimated in the 100k GP discovery cohort. A pooled analysis resulted in an overall frequency of 1:3182 X chromosomes (95% CI 1:2309–1:4386, n = 117734) (Table VI). Notably, the results with a threshold of 37 repeats, which is known to cause SBMA (Sinnreich et al., 2004), were even higher at 1:1899 X chromosomes (95% CI 1:1482–1:2434) (Appendix Table 1).

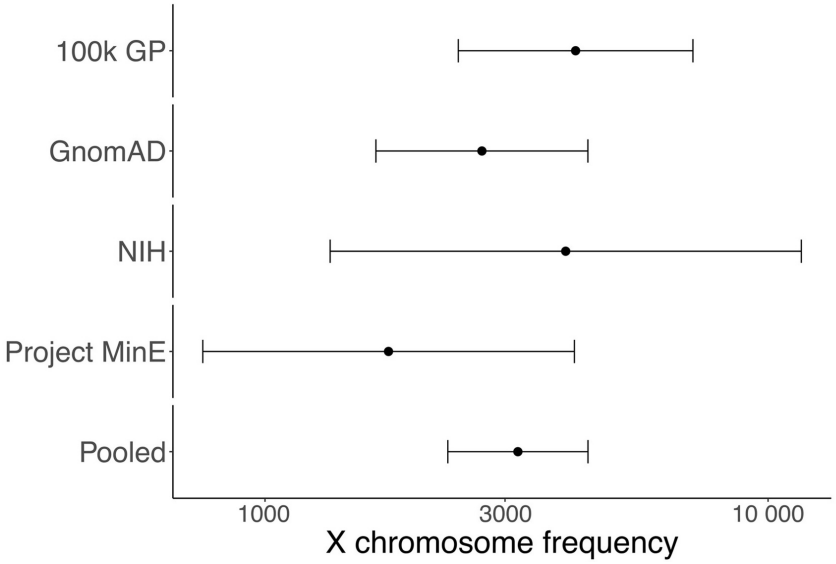


Figure 12: **Frequency estimation of the *AR* CAG expansion.** WGS pipeline detects 1:3182 *AR* expansions in the pooled 100k GP, gnomAD, NIH, and MinE cohorts. Error bars show 95% CI.

Discrepancy between observed and expected disease prevalence

The expected prevalence of the disease is lower than the mutation frequency, as SBMA is an adult-onset disease. I, therefore, used the SBMA age of onset distribution (Laskaratos et al., 2021) and the general English male population age distribution ("Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019", 2020) with the genetic frequency data to estimate disease prevalence (Figure 13). Surprisingly, the results estimated SBMA prevalence at 1:6887 males, more than 4-fold more frequent than previous patient-based epidemiological studies (Bertolin et al., 2019; Guidetti et al., 2001; Zelinkova et al., 2016). To rule out a founder effect, as seen in the Finnish study (Udd et al., 1998), I performed a haplotype analysis on European samples from the 100k GP, which resulted in non-significant associations (Appendix Figure 1).

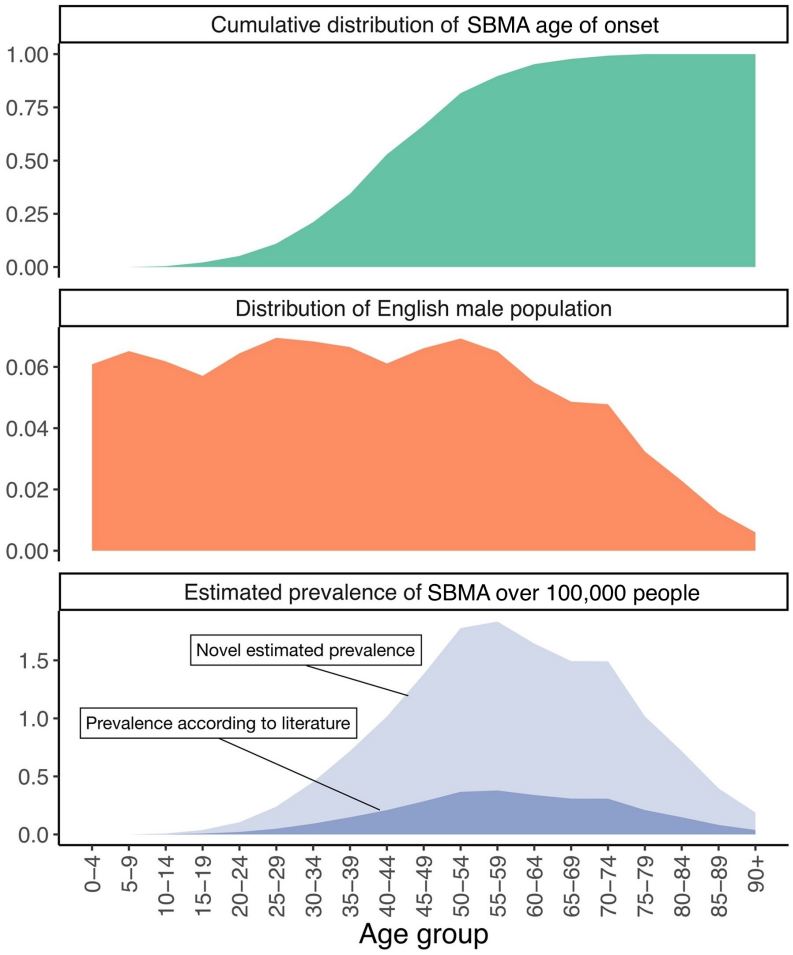


Figure 13: **Disease prevalence modelling.** Top) Cumulative distribution of SBMA age of onset for n = 983 SBMA cases from the most recent SBMA meta-analysis. The y-axis represents the cumulative distribution of cases. Middle) distribution of the English male population (n = 27 827 831). The Y-axis represents the proportion of people for each age group. Bottom) resulting estimated prevalence of SBMA by age group, considering the literature reported male prevalence of 1:30,303 or less (dark area), and the novel estimated prevalence according to the WGS result (1:6887 males, light area). The Y-axis represents the prevalence over 100,000 people.

Assessment of REDs in different populations

Previous studies have estimated that REDs affect 1 in 3,000 people (Ibañez et al., 2022). Despite their broad distribution in human populations, few global epidemiological studies have been performed. In these studies, prevalence estimates are either population-based, in which affected individuals are identified based on clinical presentation, or genetically tested based on the presence of a relative with a RED. Given that one of the most striking features of REDs is that they can present with markedly diverse phenotypes, REDs can remain unrecognised, leading to underestimation of the disease prevalence (Johnson et al., 2021).

While many of the epidemiological studies to date have been conducted in cohorts

of European origin, studies in other ancestries have highlighted population differences at specific RED loci (Baine et al., 2018; De Castilhos et al., 2014; Teive et al., 2019). In Europeans, it is estimated that the prevalence of *C9orf72*-FTD is 0.04-134 in 100,000, and *C9orf72*-ALS is 0.5-1.2 in 100,000 (Gossye et al., 1999). The spinocerebellar ataxias (SCAs) are a group of rare neurodegenerative disorders mainly affecting the cerebellum. They are individually rare worldwide, with largely variable frequencies among populations (Sequeiros et al., 2012), mainly due to founder effects. Overall, the worldwide prevalence of SCAs is 2.7-47 cases per 100,000 (Teive et al., 2019) with SCA3 being the most common form worldwide, followed by SCA2, SCA6 and SCA1 (Schöls et al., 2004). With the advent of disease-modifying therapies for REDs, it is becoming necessary to determine comprehensively the number of patients and the type of RED expected in different populations, so that targeted approaches can be developed accordingly. So far, the largest population study of the genetic frequency REDs involved the PCR-based analysis of 14,196 individuals of European ancestry (Gardiner et al., 2019).

In the past few years, bioinformatic tools have been developed to profile DNA repeats from short-read whole exome (Van Der Sanden et al., 2021) and WGS (Tanudisastro et al., 2024). A recent work has shown that disease-causing repeat expansions can be detected from WGS with high sensitivity and specificity, making large-scale WGS datasets an invaluable resource for the analysis of the frequency and distribution of REDs (Ibañez et al., 2022). As I showed above in this chapter, I applied this pipeline to a large WGS cohort to assess the distribution of repeat expansions in the *AR* gene, which cause SBMA, and found an unexpectedly high frequency of pathogenic alleles, suggesting under-diagnosis or incomplete penetrance of this RED. However, a comprehensive study of REDs in the general population and across different ancestries using WGS has never been performed.

In the following part, I used large-scale genomic databases to address two main questions: i) What is the frequency of RED mutations in the general population? ii) How does the distribution of REDs vary across populations?

Cohort description

For this work, RED loci were assessed in two large-scale medical genomics cohorts with high-coverage WGS and rich phenotypic data: the 100K GP and TOPMed (The 100,000 Genomes Project Pilot Investigators et al., 2021, https://www.genomicsengland.co.uk/initiatives/100000-genomes-project, https://topmed.nhlbi.nih.gov/) First, WGS data were generated using PCR-free protocols and sequenced with paired-end 150 bp reads. To avoid overestimating the frequency of

REDs, individuals with neurological diseases were excluded, as their recruitment was driven by the fact that they had a neurological disease potentially caused by repeat expansion. Relatedness and principal component analyses were performed to identify a set of genetically unrelated individuals and predict broad genetic ancestries based on 1,000 Genomes Project super-populations (1K GP3) (The 1000 Genomes Project Consortium, 2015). The resulting dataset comprised a cross-sectional cohort of 82,176 genomes from unrelated individuals (median age 61, Q1-Q3: 49-70, 58.5% females, 41.5% males; genetically predicted to be of European (n=59,568), African (n=12,786), American (n=5,674), South Asian (n=2,882), and East Asian (n=1,266) descent.

RED mutation frequency

Disease	Locus	Premutation/reduced penetrance	Full Mutation repeat threshold
Spinal and bulbar muscular atrophy	*AR*	35-37	38
Dentatorubral-pallidoluysian atrophy	*ATN1*	35-47	48
Spinocerebellar ataxia 1	*ATXN1*	39-43	44
Spinocerebellar ataxia 2	*ATXN2*	33-34	35
Spinocerebellar ataxia 3	*ATXN3*	45-59	60
Spinocerebellar ataxia 7	*ATXN7*	34-36	37
C9orf72-related Frontotemporal dementia and/or amyotrophic lateral sclerosis 1	*C9orf72*	-	31
Spinocerebellar ataxia 6	*CACNA1A*	19	20
Myotonic dystrophy 1	*DMPK*	35-49	50
Friedreich ataxia	*FXN*	45-65	66
Huntington disease	*HTT*	36-39	40
Huntington disease-like 2	*JPH3*	29-39	40
Oculopharyngodistal myopathy 3; Neuronal intranuclear inclusion disease (NIID)	*NOTCH2NLC*	41-55	56
Spinocerebellar ataxia 12	*PPP2R2B*	32-65	66
Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome	*RFC1*	-	1042 (median number of repeats)[56] IQR=844-1306 range=249-3885
Spinocerebellar ataxia 17	*TBP*	41-48	49

Figure 14: **List of RED loci included in the study, including repeat-size thresholds for reduced penetrance and full mutations.**

To estimate the number of individuals carrying premutation or full mutation alleles (Figure 14), repeats in RED genes (Depienne and Mandel, 2021) for which WGS can accurately discriminate between normal and pathogenic alleles were selected (Ibañez et al., 2022), based on either or both, of the following conditions: the threshold between premutation and full mutation is shorter than the sequencing read-length (and therefore WGS can accurately distinguish between premutation and full mutation), or WGS was validated against the current gold-standard PCR test (Ibañez

et al., 2022). WGS accurately classifies alleles in the normal, premutation, and full mutation range in all loci except *FMR1* (that causes Fragile X syndrome). The accuracy of repeat sizing by WGS is not affected by genetic ancestry by comparing genotypes generated by WGS to PCR from different populations, but it might underestimate the size of large expansions in *FMR1, DMPK, FXN,* and *C9orf72*, as previously described (Ibañez et al., 2022).

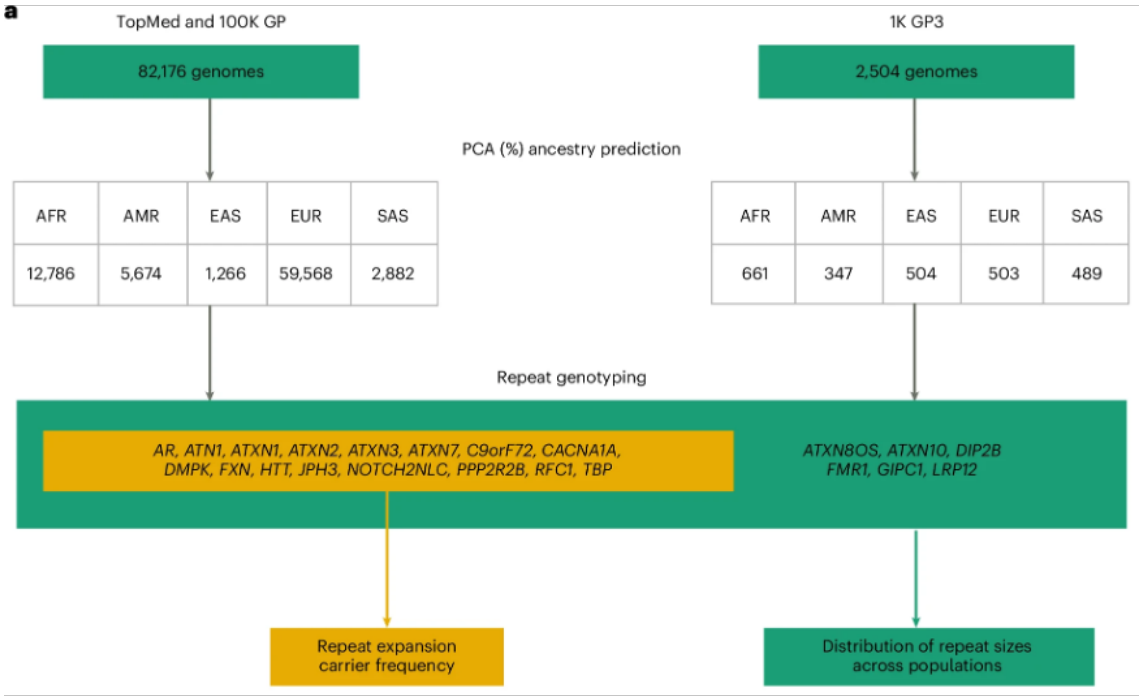


Figure 15: **Technical flowchart.** WGS from the 100K GP and TOPMed datasets were first selected by excluding those associated with neurological diseases. WGS data from the 1K GP3 were also selected by having the same technical specifications (see Methods). After inferring ancestry prediction, repeat sizes for all 22 REDs were computed by using EH v3.2.2. On one hand, for 16 REDs overall carrier frequency, disease modelling, and correlation distribution of long normal alleles were computed in the 100K GP and TOPMed projects. On the other hand, the distribution of repeat sizes across different populations was analysed in the 100K GP and TOPMed combined, and in the 1K GP3 cohorts.

Overall, for premutation and full mutation carrier frequency analysis, I analysed 16 RED loci, representing a broad spectrum of REDs and different modes of inheritance: 1) autosomal dominant: HD, Huntington disease-like 2 (HDL2), DM1, *C9orf72*-ALS/FTD, the spinocerebellar ataxias (SCA1, SCA2, SCA3, SCA6, SCA7, SCA12, SCA17), dentatorubral-pallidoluysian atrophy (DRPLA), and *NOTCH2NLC*, which causes a spectrum of neurological disorders, especially neuronal intranuclear inclusion disease and oculopharyngodistal myopathy; 2) autosomal recessive: Friedreich ataxia and Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS), 3) X-linked SBMA (Figure 14).

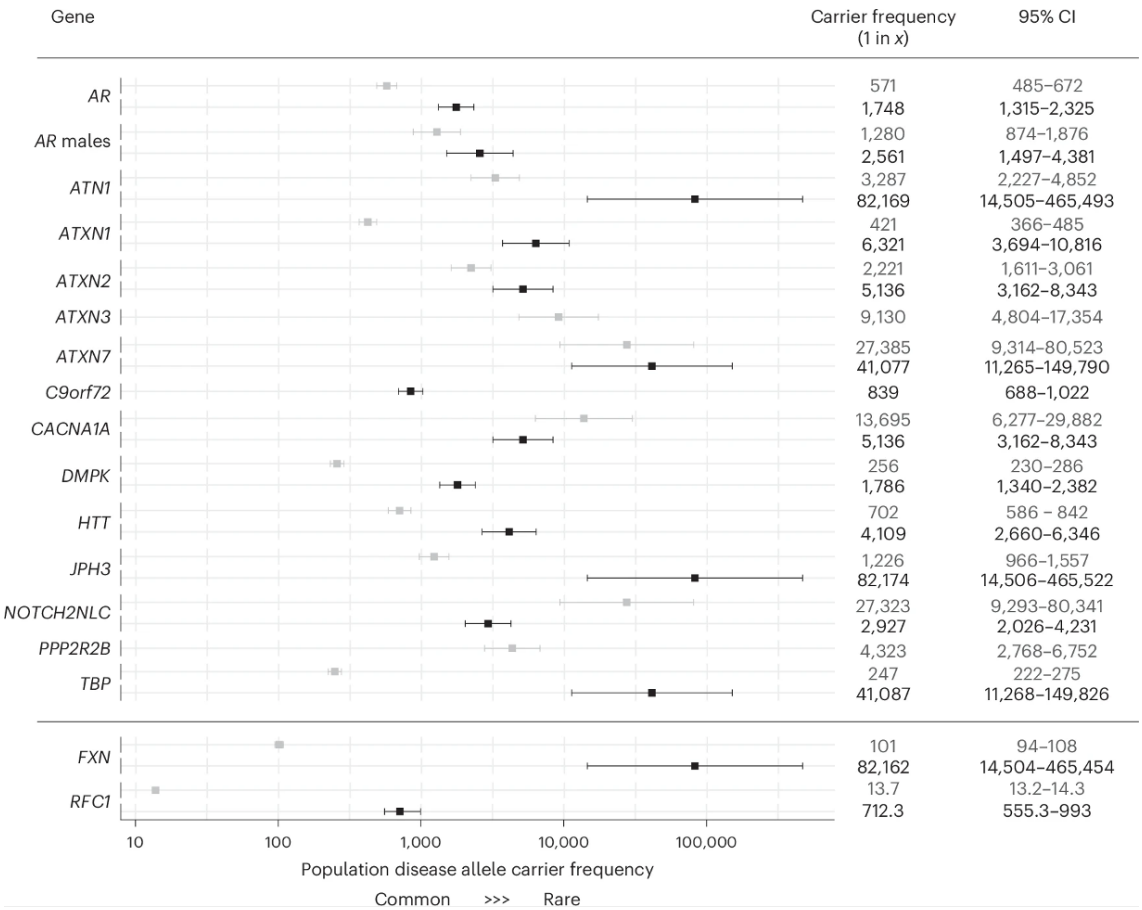


Figure 16: **Forest plot with combined overall disease allele carrier frequency in the combined 100K GP and TOPMed datasets.** N = 82,176 (N individuals may vary slightly between loci due to data quality and filtering). The squares show the estimated disease allele carrier frequency, and the bars show the 95% CI values. Details of the statistical models are described in the Methods section. Grey and black boxes show premutation/reduced penetrance and full mutation allele carrier frequencies for each dominant locus, respectively. Grey and black boxes show mono- and bi-allelic carrier frequencies for recessive loci (*RFC1* and *FXN*), respectively.

The analysis workflow (Figure 15) included profiling each RED locus, followed by quality control of all alleles (employing Expansion Hunter classifier, https://github.com/bharatij/ExpansionHunter_Classifier, and visual inspection of pileup plots as previously described in Dolzhenko et al., 2022) predicted to be larger than the premutation threshold. In total, for autosomal dominant and X-linked REDs, there were 290 individuals carrying one fully expanded repeat and 1,279 individuals carrying one repeat in the premutation range, meaning that the frequency of individuals carrying full-expansion and premutation alleles among this large cohort is 1 in 283 and 1 in 64, respectively.

The most common expansions (in the full-mutation range, Table 14) were those in *C9orf72* (*C9orf72*-ALS/FTD) and *DMPK* (DM1) with a frequency of 1 in 839 and 1 in 1,786, respectively, followed by expansions in *AR* (SBMA: 1 in 2,561 males)

and *HTT* (HD: 1 in 4,109). Surprisingly, many individuals were found to carry expansions in the SCA genes: 1 in 5,136 in *ATXN2* (SCA2), 1 in 5,136 in *CACNA1A* (SCA6), and 1 in 6,321 in *ATXN1* (SCA1). By contrast, expansions in *ATXN7* (SCA7) and *TBP* (SCA17) were present in only two individuals at each locus (1 in 41,077) (Figure 16).

Modelling the expected number of people affected by REDs

REDs have variable age at onset, disease duration, and penetrance (Paulson, 2018). Therefore, the mutation frequency cannot be directly translated into disease frequency (i.e. prevalence). To estimate the expected number of people affected by REDs, by using the mutation frequency of the most common REDs (C9orf72-ALS/FTD, DM1, HD, SCA1, SCA2, SCA6), I modelled the distribution by age of those expected to be affected by REDs in the UK population. For this analysis, I used the data from the Office of National Statistics ("Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019", 2020), and the age of onset, penetrance, and impact on survival of each RED based on either cohort studies or disease-specific registries.

I estimated, on average, a two- to three-fold increase in the predicted number of people with REDs, compared to currently reported figures based on clinical observation, depending on the RED (Figure 17). Since *C9orf72* expansions cause both ALS and FTD, I modelled both diseases separately, providing for *C9orf72*-ALS an expected number of people affected over two times higher than previous estimates: 1.8 per 100,000 versus 0.5-1.2 per 100,000 (Gossye et al., 1999; Zampatti et al., 2022), and for *C9orf72*-FTD 6.5 per 100,000 within the wide reported range (Hogan et al., 2016; Van Mossevelde et al., 2018). For SCA2 and SCA1, the model indicates over three-fold increase in the number of people expected with the disease compared to the reported prevalence (3 and 3.7 per 100,000, respectively, versus the currently reported prevalence of 1 per 100,000) (De Mattei et al., 2023; Opal and Ashizawa, 2023). Overall, these data indicate that REDs are either underdiagnosed or that not all individuals who carry a repeat larger than the established full mutation cut-off develop the condition (i.e. incomplete penetrance).

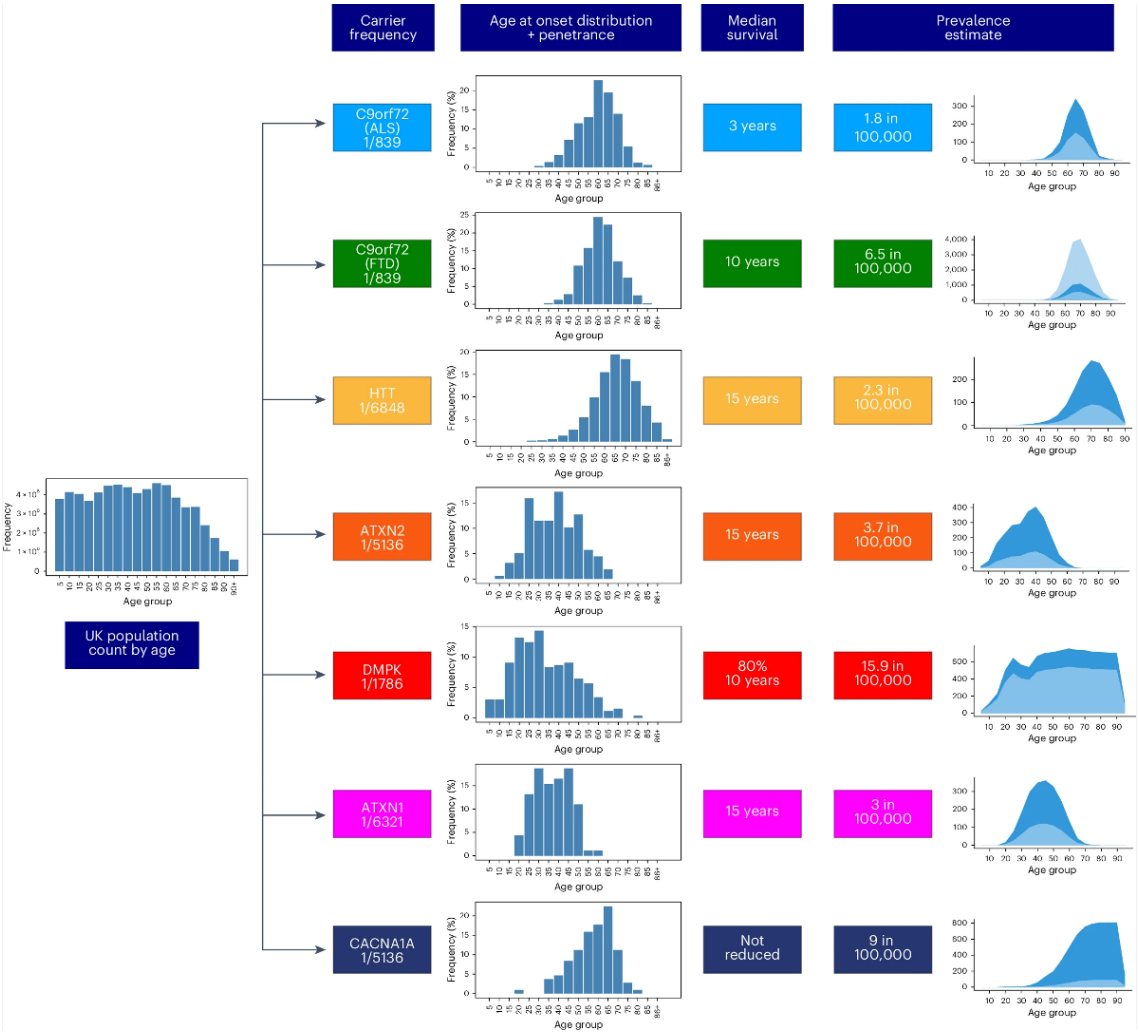


Figure 17: **Flowchart showing the modelling of disease prevalence by age for *C9orf72*-ALS, *C9orf72*-FTD, HD in 40 CAG repeat carriers, SCA2, DM1, SCA1, and SCA6.** UK population count by age is multiplied by the combined disease allele frequency of each genetic defect and the age of onset distribution of each corresponding disease, and corrected for median survival. Penetrance is also taken into account for *C9orf72*-ALS and *C9orf72*-FTD. Estimated number of people affected by REDs (dark blue area) compared to the reported prevalence from the literature (light blue area). Age bins are 5 years each. For *C9orf72*-FTD, given the wide range of the reported disease prevalence, both lower and upper limits are plotted in light blue. The Y-axis for the plots on the right represents the estimated affected people in the UK population.

Distribution of repeat lengths in different populations

REDs are thought to arise from large normal polymorphic repeats (large normal, or "intermediate" range repeats), as they have an increased propensity to further expand upon transmission from parent to progeny, moving into the pathogenic range. The uneven RED prevalence across major populations has been associated with the variable frequency of intermediate alleles (Kay et al., 2018; Takano et al., 1998).

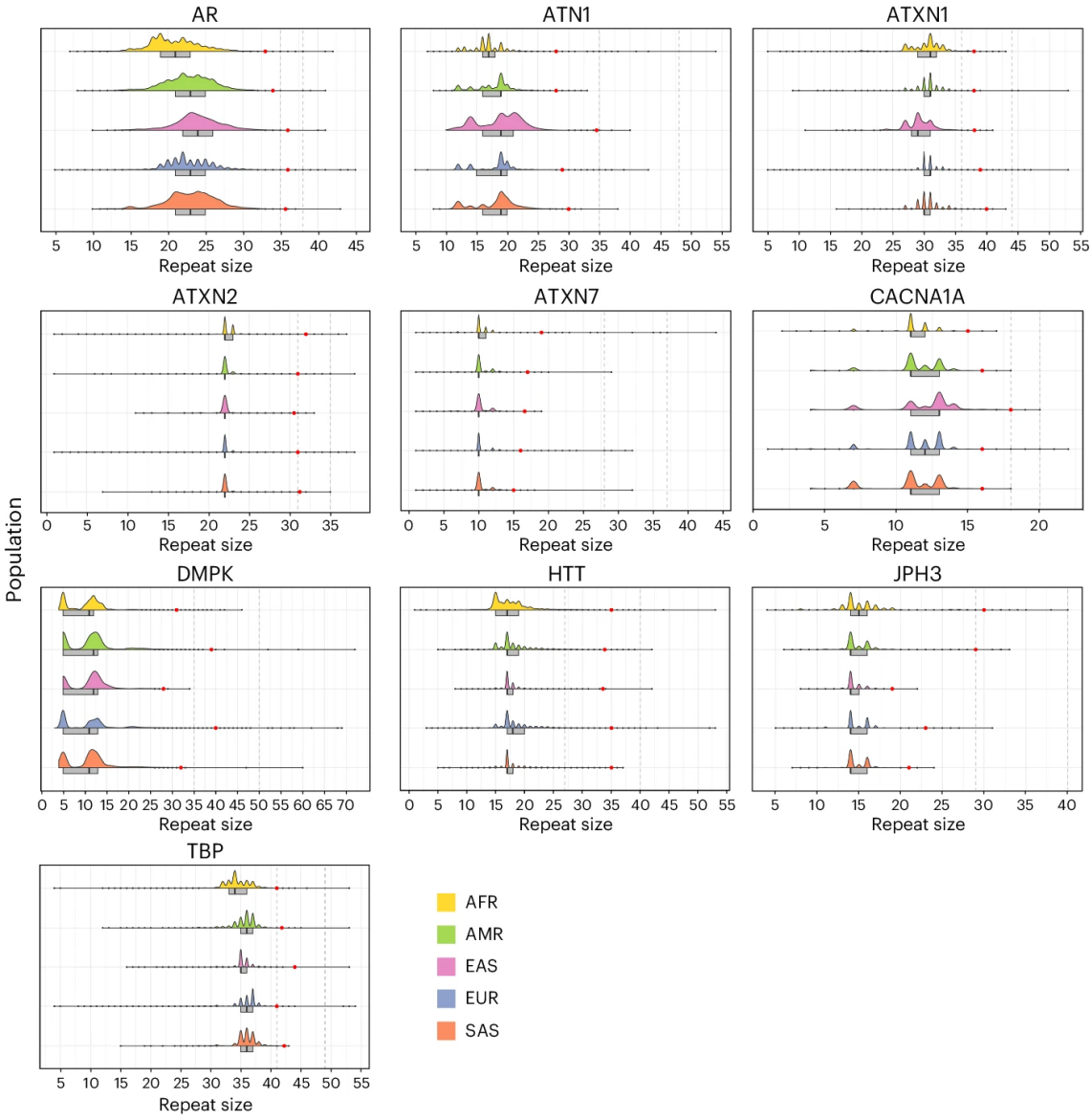


Figure 18: **Distribution of repeat lengths in different populations.** A) Half-violin plots showing the distribution of alleles in different populations (African = 12,786; American = 5,674; East Asian = 1,266; European = 59,568; South Asian = 2,882) for 10 loci from the combined 100K GP and TOPMed cohorts. Box plots highlight the interquartile range and median, and black dots show values outside 1.5 times the interquartile range. Red dots mark the 99.9th percentile for each population and locus. Vertical bars indicate the intermediate and pathogenic allele thresholds.

Therefore, I analysed intermediate allele frequencies for those genes where WGS can accurately size intermediate alleles across populations and confirmed that: i) the overall distribution of repeat lengths varies across populations (Figure 18); ii) overall, the frequency of intermediate alleles varies in each population, and correlates with the frequency of pathogenic alleles; (R = 0.65; p = 3.1x10-7, Spearman correlation) (Figure 19). These data suggest that different distributions of repeat lengths underlie differences in the epidemiology of REDs.

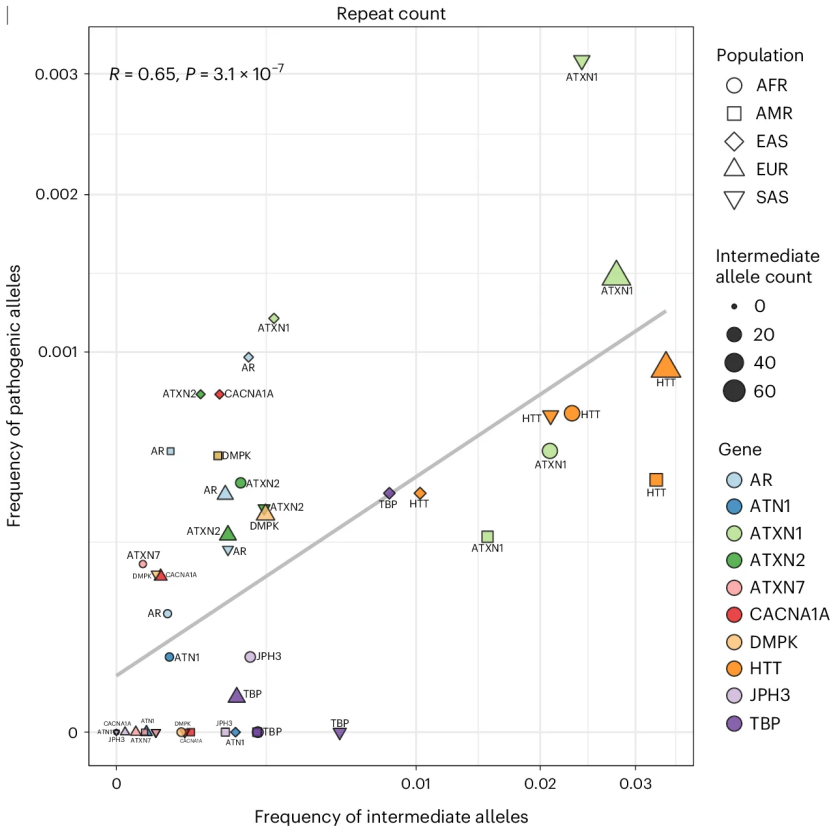


Figure 19: **Correlation between the frequency of intermediate and full repeats.** B) Scatter plot shows the frequency of intermediate allele carriers (x-axis) against the frequency of pathogenic allele carriers (y-axis). Data points are divided by population (n=5) and gene (n=10), and size represents the total number of intermediate alleles. Correlations were computed using the Spearman method and two-tailed p-values.

3.3 Discussion

AR repeat expansion prevalence

Overall, this work identifies an unexpected frequency of the *AR* pathogenic expansion in a UK cohort and confirms this finding using three other large European and North American datasets. Previous findings of an epidemiological study in the Vasa region and a meta-analysis are in line with the present findings. Importantly, the use of WGS data allowed us to curate the dataset for relatedness and perform a haplotype analysis that rules out founder effects.

The discrepancy between patient numbers and the frequency of the genetic defect may be due to (i) underdiagnosis of this neuromuscular condition; (ii) variable disease expressivity/reduced penetrance; (iii) pleomorphic clinical manifestations; or (iv) a combination of these factors. Underdiagnosis of the disease has frequently been suggested, and, whilst the classic disease manifestation with bulbar and limb

weakness, highly elevated creatine kinase levels, and gynaecomastia is very typical, the disease can manifest with only certain symptoms and often with a negative family history due to its X-linked mode of transmission, favouring misdiagnosis (La Spada, 2022; Udd et al., 1998). Different from other STR expansion disorders showing incomplete penetrance for all the repeat lengths (Murphy et al., 2017), SBMA is reported to be incompletely penetrant between 35 and 37 repeats, but fully penetrant from 38 (La Spada, 2022). Moreover, although strong variability in manifestations and severity of SBMA can occur within siblings, reports of incomplete penetrance within families of SBMA patients are lacking. A recent meta-analysis raised the hypothesis that the *AR* CAG repeat is partially penetrant up to 45 repeats (Laskaratos et al., 2021), although the fact that in the 100k GP all the males older than 45 years, with more than 37 repeats, had an SBMA phenotype argues against reduced penetrance as being the main driver of the discrepancy between patient numbers and mutation frequency. Larger numbers and more targeted studies will be needed to fully clarify this. Lastly, SBMA has been associated with a number of common non-neurological disorders such as insulin resistance, non-alcoholic fatty liver disease, and metabolic syndrome (Manzano et al., 2018), and in light of the frequency of the genetic defect, it should likely be considered in people with these conditions.

In conclusion, I identified an unexpectedly high frequency of the SBMA genetic defect in European and North American populations, suggesting SBMA is underdiagnosed and highlighting how testing may be relevant not only to neuromuscular diseases.

Frequency of repeat expansions and their distribution across different populations

By analysing a cross-sectional cohort of 82,176 people, this study provides the largest population-based estimate of disease allele carrier frequency and RED distribution in different populations, showing that: i) the disease allele carrier frequency of REDs is approximately ten times higher than the previous estimates based on clinical observations, and that, based on population modelling, REDs would be predicted to affect, on average, two to three times more individuals than are currently recognised clinically; ii) the different distribution of repeat lengths between populations broadly reflects the known differences in disease epidemiology; iii) an appreciable proportion of the population carries alleles in the premutation range, and are therefore at risk of having children with REDs.

The estimates for DM1 and SCAs match those previously reported using PCR-based approaches to determine the genetic prevalence of REDs (Gardiner et al., 2019; Johnson et al., 2021; Kay et al., 2016), confirming the accuracy of the presented results.

Different factors might explain the discrepancy between the increased number of people carrying disease alleles in this cohort compared with known RED epidemiology.

First, these estimates are based on large admixed cohorts, as opposed to epidemiological studies based on clinically affected individuals in smaller populations. As REDs have variable clinical presentation and age at onset, individuals with REDs may remain undiagnosed in studies in which estimates of disease frequency rely on clinical ascertainment of patients. Notably, the first descriptions of the clinical phenotype of RED were based on families collected for linkage studies with an intrinsic ascertainment bias for more severe disease manifestations, resulting in a lack of very mild cases in the phenotypic spectrum. Because of the wide spectrum of milder phenotypic presentations of REDs, the prevalence of these diseases might have been underestimated. In fact, a large number of individuals carrying repeats in the lower end of the pathogenic range were observed (e.g. *HTT, ATXN2*, and *DMPK*).

Moreover, prevalence studies are only based on individuals with manifest disease, leading to a potential bias in the disease penetrance from those who have not developed the illness. It is believed that the penetrance of REDs is characterised by a threshold effect, with people carrying an allele above a particular repeat length certainly developing the disease, as opposed to those carrying shorter repeats. It is possible that individuals carrying alleles currently classified as fully penetrant (e.g. $\geq$40 CAG repeats in *HTT*) may sometimes remain asymptomatic. In this regard, previously published studies on HD (Langbehn et al., 2004) and SBMA (Laskaratos et al., 2021) have suggested incomplete penetrance of repeats in the lower end of the pathogenic range.

Finally, these individuals may carry genetic modifiers of REDs, such as interspersions. All alleles in the pathogenic range were visually inspected, and no atypical sequence structures were found within the expanded repeat.

The finding that a much larger number of people in the general population carry pathogenic alleles of REDs has important implications both for diagnosis and genetic counselling of RED. For diagnosis, when a patient presents with symptoms compatible with a RED, clinicians should have a higher index of suspicion of these diseases, and clinical diagnostic pathways should facilitate genetic testing for REDs. Currently,

genetic testing for REDs tends to be a PCR-based targeted assay, with clinicians suspecting a RED ordering a test for a specific gene. As REDs are clinically and genetically heterogeneous with a tendency to have overlapping features, REDs may remain undiagnosed. The wider use of WGS and the advent of genetic technologies such as long-read sequencing can potentially address this by simultaneously interrogating an entire panel of RED loci (Miyatake et al., 2022). The broader availability of such diagnostic tools would increase the diagnostic rate for REDs, thus closing the gap between disease incidence rates and estimates based on population genetic sequencing.

As for genetic counselling, when a RED expansion is identified incidentally in an individual clinically unaffected, it would be important to address the potentially incomplete penetrance of the repeat, especially for small expansions. Further studies, both in clinically affected individuals and in large clinical and genomic datasets from the general population, are needed to address the full clinical spectrum and the penetrance by repeat sizes.

These results are concordant with current epidemiological studies about the relative frequency of REDs, with the most common being DM1 and *C9orf72*-ALS/FTD.

Further research is needed to study the potential role of population-specific cis and trans genetic modifiers of repeat expansion mutations that underlie the marked global differences in RED distribution found in the present study.

One limitation of this study is that WGS cannot accurately size repeats larger than the sequencing read length, and it is therefore not possible to accurately estimate the disease allele frequency of all RED loci. Of the 46 RED loci that have been linked to human disease (Depienne and Mandel, 2021), all loci where it is technically possible to address the specific questions were included: 1) to accurately estimate the disease allele carrier frequency over 16 REDs; 2) to analyse the distribution of repeat lengths in different populations in 10 REDs, providing the basis for the different prevalence of REDs in different populations. Many newly discovered REDs are caused by large expansions (Vegezzi et al., 2024): only a broader availability of long-read sequencing technologies will facilitate addressing important questions about the frequency of these mutations.

Both 100K GP and TOPMed datasets are Eurocentric, comprising over 62% of European samples. TOPMed is more diverse, with 24% and 17% of African and American genomes, respectively, which are only present at 3.2% and 2.1% frequency in 100K GP. East and South Asian backgrounds are underrepresented in both datasets, limiting the ability to detect rarer repeat expansions in these populations. Further

analyses on more heterogeneous and diverse large-scale WGS datasets are necessary not only to confirm the present findings but also to shed light on additional ancestries. With regard to this, there are multiple ongoing projects with Asian populations (D. Wu et al., 2019). Countries including China, Japan, Qatar, Saudi Arabia, India, Nigeria and Turkey have launched their own genomics projects during the last decade (Kumar and Dhanda, 2020). Analysing genomes from these programmes will yield more details on the distribution of REDs around the world.

Despite efforts to estimate the frequency of REDs globally and locally, there is uncertainty surrounding their true prevalence, limiting the knowledge of the burden of disease required to secure dedicated resources to support health services, such as the estimation of the numbers of individuals profiting from drug development and novel therapies, or participating in clinical trials. There are currently no disease-modifying treatments for REDs; however, both disease-specific treatments and drugs which target the mechanisms leading to repeat expansions are in development. This study shows that the number of people who may benefit from such treatments is greater than previously thought.

3.4 Overview and future directions

Overall, this chapter demonstrates the strength of using large population cohorts.

There are some limitations to this work. First, the use of the 100k GP cohort, which is enriched for individuals with rare diseases and paediatric conditions, deviates from a truly population-based sample, introducing ascertainment bias. For example, the overrepresentation of paediatric cases may lead to the erroneous categorisation of individuals as "unaffected", although they may still be at risk for adult-onset diseases (like REDs) not yet manifested at the time of sampling.

Moreover, penetrance calculations, that are inherently sensitive to phenotypic information availability, relies on estimates derived from cohorts with known phenotypes, which can lead to over- or underestimation of penetrance, particularly for genetic diseases with pleomorphic manifestation or variable disease expressivity, such as those caused by expansions in *C9orf72* (Murphy et al., 2017). Moreover, the age-dependent penetrance of REDs introduces additional ambiguity and may reduce the precision of genotype-phenotype associations, depending on the demographics of the cohort assessed.

The choice of repeat size cutoffs for defining pathogenic expansions was informed by

prior literature (Rüb et al., 2013; Sone et al., 2019; Wallace and Bean, 2022). However, it is important to acknowledge that the knowledge about REDs is still limited and therefore these thresholds could change in the next years. For several loci, I selected conservative size thresholds based on established diagnostic criteria (e.g. $\geq$38 repeats) CAG repeats in *AR* for SBMA), but these often reflect historical reporting biases and may not fully capture the spectrum of pathogenicity or instability. Future work may benefit from continuous models of repeat size rather than dichotomous classification (expanded vs. not), as well as locus-specific modelling of repeat instability and penetrance. In this study, while discrete cutoffs were necessary for statistical modelling, I acknowledge that this approach simplifies the underlying biology and may obscure gradations of risk and expressivity.

Interestingly, the results for the prevalence of the *AR* CAG repeat expansion in the two different studies fall within the same confidence interval (1/(2309-4386) in the first, and 1/(1315-2325) in the second, respectively) despite the difference in point prevalence (1/3182 and 1/1748, repsectively). Importantly, there is only partial overlap between the two datasets, since cases from the 100k GP have been used for both studies, while 54% of cases from the first study and 58% of cases from the second study come from different datasets (Project MinE, NIH, and gnomAD in the first, and TOPMed in the second, respectively), potentially explaining the variability in the findings.

An increasing number of large datasets are becoming available, and in certain instances, as in the case of UK Biobank, genetic data is paired with extensive biofluids and imaging data. These datasets will allow to address questions that the genetic prevalence findings leave open, such as what the penetrance of disease is.

For example, in the case of SBMA, the disease is accompanied by clear changes in creatine phosphokinase levels and by distinct changes in muscle and liver fat fraction. The availability of datasets such as UK Biobank will allow to address whether carriers, independently of their medical diagnosis, have typical disease features.

Moreover, mathematical modelling can be used not only to predict the prevalence of diseases with variable penetrance and survival from genetic prevalence data but also to generate a model to predict the instability threshold for different repeat expansions.

Chapter 4

Biology of TDP-43 cryptic exons

4.1 Contributions

Parts of the text in this chapter are copied from the work "TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A" (Brown et al., 2022), and therefore, other people contributed to ideas, drafting, and writing. Credits for contributions can be found in the UCL Research paper declaration forms at the beginning of the thesis. Particularly, the alignment, splicing, and parsing pipelines were developed by Dr. Anna-Leigh Brown and Dr. Samuel Bryce-Smith. The double *UPF1/TARDBP* knockdown in iPSC neurons was done by Dr. Matthew Keuss and Dr. Puja Mehta. The dose-response SH-SY5Y curves were generated by Dr. Matthew Keuss.

4.2 Overview

It is well established that upon loss of TDP-43, CEs occur (Ling et al., 2015), but many questions regarding their biology remain unanswered. The fate of CEs and whether they are degraded by NMD has not yet been characterised in a comprehensive manner. Furthermore, how much TDP-43 loss is required for this to occur is yet unknown. This chapter has two main sections addressing these different questions.

In the first section, I generated and analysed two sets of experiments to assess the impact of NMD on CEs, showing that most CEs lead to RNA degradation and protein loss through the NMD machinery. The fact that these novel RNAs are targeted for degradation poses the question of whether they can escape detection using conventional methods, and indeed, I have shown that NMD inhibition allows for the unmasking of hidden CEs. Finally, I used protein-RNA binding data and

postmortem RNA-seq data to validate the binding of TDP-43 to NMD-masked CEs and their detection in patient samples.

In the second part, I addressed the question of how different levels of TDP-43 loss impact cryptic splicing. I generated and analysed two more sets of cells to assess the impact of TDP-43 loss on CEs, highlight how different CEs need different levels of TDP-43 loss to appear, and to what level they are expressed at when TDP-43 is completely lost.

4.3 Results

Investigating the contribution of NMD to CEs

Where do CEs end up? Are CEs degraded, and if so, do some go undetected?

UNC13A and *UNC13B* CEs are sensitive to NMD

In the manuscript on the *UNC13A* CE (Brown et al., 2022), to functionally prove *UNC13A* and *UNC13B* induced NMD of their CE-including transcripts, I treated SH-SY5Y cells with TDP-43 knockdown either with 100 μM CHX (to stall translation and impair NMD) or DMSO for 6 hours. I confirmed this result by treating SH-SY5Y cells with a combination of siRNAs against *TARDBP* and *UPF1*, a key NMD factor.

In both cases, inhibition of the NMD machinery led to the rescue of both *UNC13A* CE and *UNC13B* frameshift exon (that leads to a PTC at the beginning of exon 11); conversely, CE-including *STMN2* transcripts (containing an alternative polyadenylation signal and therefore not predicted to undergo NMD) were not altered by the treatments (Figure 20).

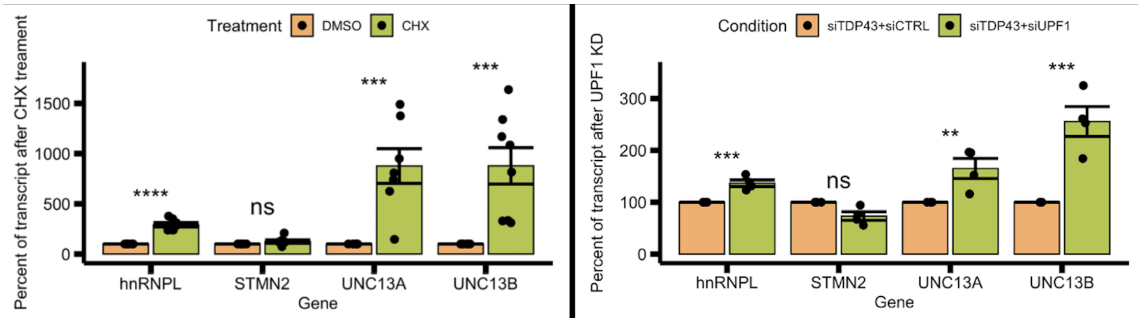


Figure 20: ***UNC13A* and *UNC13B* CEs, but not *STMN2*, are NMD-sensitive.** Transcript expression upon treatment with CHX (left) and *UPF1* siRNA knockdown (right) suggests that *UNC13A* and *UNC13B* CEs, but not *STMN2*, are sensitive to NMD. *HNRNPL* is used as a positive control. Graphs represent the means ± SEM, Left: n = 7 biological replicates (*UNC13A*, *HNRNPL* and *STMN2*) and 8 biological replicates (*UNC13B*), Right: n = 4 biological replicates per condition.

Of note, even at a very mild TDP-43 knockdown, where *UNC13A* CE is not normally visible, CHX treatment enabled us to detect it (Figure 21). This observation highlights how the identification of CEs from whole-cell RNA-seq data may be skewed and supports the idea that the occurrence of *UNC13A* CE is underestimated because of efficient degradation through NMD.

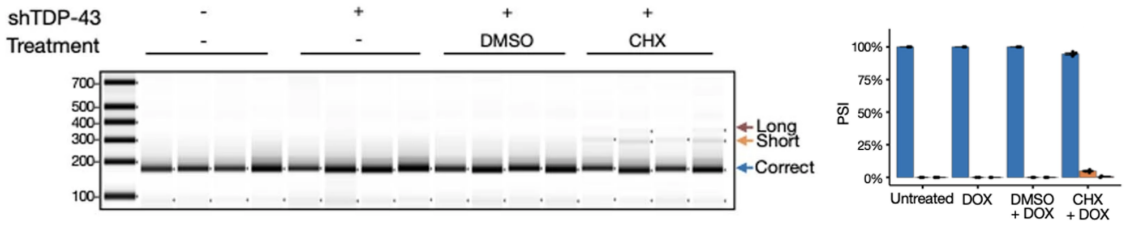


Figure 21: **NMD inhibition unmasked *UNC13A* CE even at a mild TDP-43 knockdown.** Left: RT-PCR products from *UNC13A* in the setting of mild TDP-43 knockdown ("+", as for the mildest knockdown in Figure 30) with the addition of either DMSO (control) or CHX (NMD inhibition). Right: quantification of left graphs represents the means ± SEM, n = 4 biological replicates per condition. PSI refers to the percentage of the band intensity corresponding to the CE compared to the sum of the band intensities for each lane.

Overall, this data suggested that correct pre-mRNA splicing of *UNC13A* and *UNC13B*, necessary for their protein expression, is guaranteed by TDP-43 nuclear function and that NMD is responsible for their degradation upon TDP-43 loss of function.

Inhibition of NMD in TDP-43 knockdown cells with CHX

I used CHX to inhibit NMD in SH-SY5Y cells. Four conditions were generated:

- Control for TDP-43 knockdown and control for NMD (control-control)
- Control for TDP-43 knockdown and CHX-treated (CHX-control)
- Positive for TDP-43 knockdown and control for NMD (control-TDP43KD)
- Positive for TDP-43 knockdown and CHX-treated (CHX-TDP43KD)

After performing RNA-seq and running the alignment and splicing pipelines, I analysed the data. First, I assessed the data distribution and I ran a principal component analysis (PCA) on the normalised counts to identify what was driving the majority of the gene expression differences in the dataset. Each condition clusters in the PCA (Figure 22). CHX treatment determines the majority of changes in PC1, whilst TDP-43 knockdown appears to drive the PC2.

Differential gene expression analysis confirmed efficient knockdown of TDP-43, as well as reduced expression of known CE-bearing genes, including *STMN2*, *UNC13A* and *UNC13B*, independent of the CHX treatment.

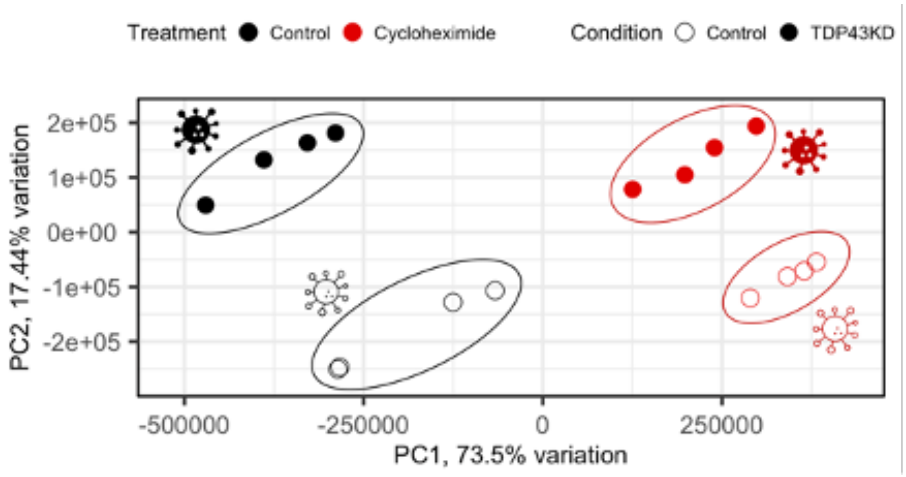


Figure 22: **PCA biplot based on the gene expression for the CHX experiment.** PCA biplot showing that PC1 explains 73.5% of the gene expression variability, mostly driven by CHX treatment, while PC2, driven by the TDP-43 knockdown, explains 17.44% of the variability.

Differential splicing analysis showed known CEs in genes related to pathways proved to be dysregulated in ALS-FTD, including nucleocytoplasmic transport (*NUP188*), autophagy (*ATG4B*), synaptic function (*UNC13A, SYT7*), metabolism (*INSR*), and microtubule stabilisation (*STMN2*). Moreover, two RBPs with a role in mRNA splicing (*CELF5, ELAVL3*) were shown to harbour CEs, suggesting that they can actively participate in the splicing dysfunction downstream of TDP-43 loss of function. Furthermore, a CE in the gene *AARS1*, a tRNA aminoacyl transferase related to CMT2N and whose counterpart *GARS1* is mutated in CMT2D, is predicted not to include FS or PTC, thus possibly coding for a novel protein isoform - hypothesis confirmed in Seddighi et al., 2024. Whether this novel protein isoform contributes to the disease phenotype through a gain-of-function mechanism - the same implicated in the aforementioned diseases - is yet to be established.

Overall, differential splicing analysis led to the discovery of 275 cryptic splice junctions in 194 genes for the control-control vs control-TDP43KD comparison, including known TDP-43 targets such as *STMN2* and *UNC13A*. In the CHX-control vs CHX-TDP43KD comparison, significant cryptic splice junctions were increased to 400 in 253 genes. 188 (39%) cryptic junctions were shared between the two datasets, with 44% of genes uncovered by NMD. The remaining 18% is showing only in the TDP-43 knockdown condition.

RNA-sequencing data allows us to predict whether the presence of a premature

termination codon in the CE or a frameshift-inducing CE would lead to NMD. One can then use different tools to block NMD and test whether, indeed, specific transcripts are rescued.

I tried to quantitatively classify the events into two groups: one group where blocking NMD determined an increase in splice isoform detection of over 5%, and one where blocking NMD did not have a strong impact on the PSI. 69% of CEs underwent an increase in PSI upon CHX treatment, while 31% of CEs were not rescued. In many cases, the rescue was predicted by PTCs or frameshift exons, compared to in-frame and alternative polyadenylation events (Figure 23).

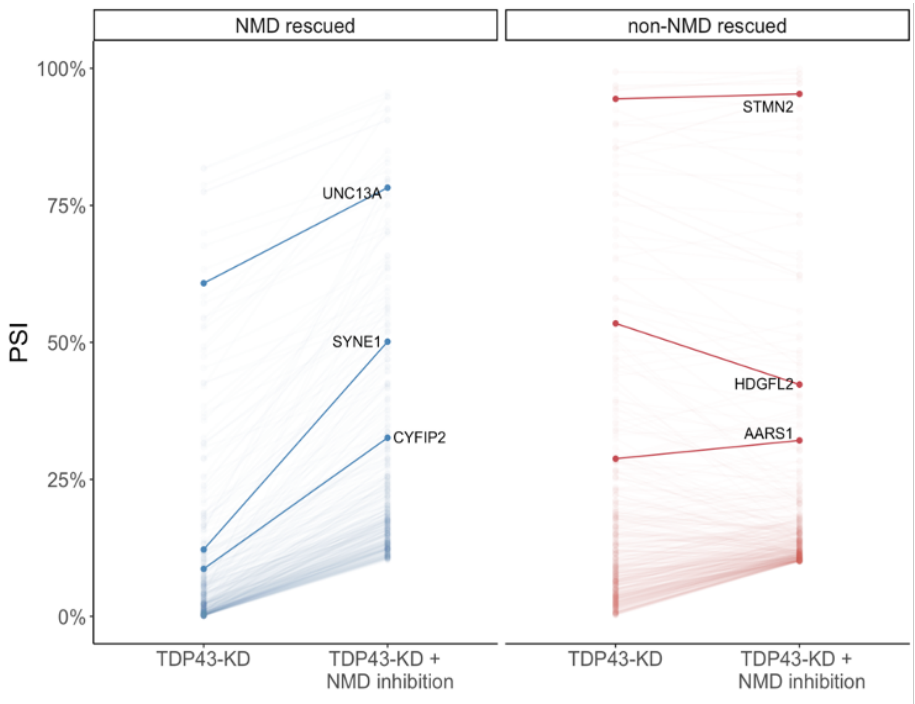


Figure 23: **CEs are rescued by NMD inhibition, as predicted by their sequences.** CE junctions were clustered by the rescue on their PSIs upon CHX treatment. *STMN2*, *HDGFL2*, and *AARS1* CEs, which are not predicted to introduce PTCs or frameshift exons, were not rescued, while *UNC13A* CE, which introduces a PTC, was not. Some events predicted to undergo NMD by their sequence, such as *CYFIP2* and *SYNE1*, and that are not significantly changed upon TDP-43 knockdown alone, are unmasked by CHX treatment.

As anticipated, the majority of transcripts containing PTC or FS were rescued upon NMD inhibition, while alternative polyA and in-frame cryptic exons were not ($\chi^2 = 7.45$, p = 0.00633) (Figure 24).

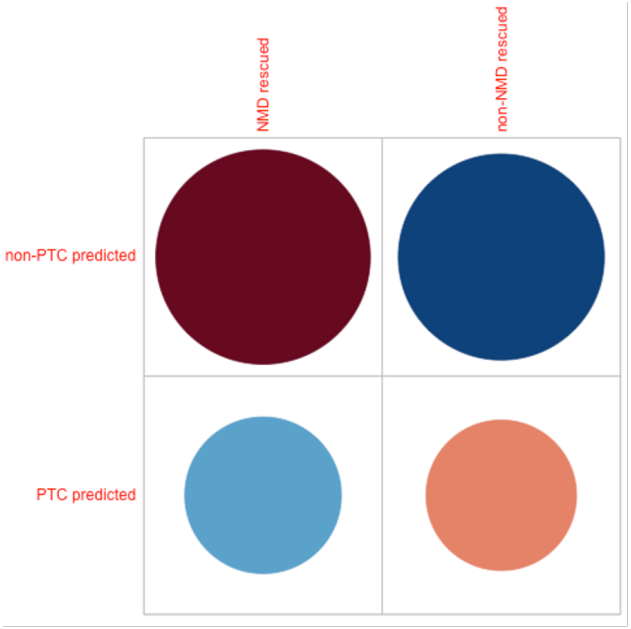


Figure 24: **Changes in CE inclusion upon CHX treatment mirror predicted RNA fate from sequencing.**

Particularly, the PSIs of *STMN2* CE (Klim et al., 2019; Melamed et al., 2019), encoding for an alternative poly(A) and *HDGFL2* (Irwin et al., 2024) CEs, which drive in-frame insertions, were not changed by CHX, while *UNC13A* CE (Brown et al., 2022; X. R. Ma et al., 2022) was rescued by more than 10%. Moreover, other events predicted to be in-frame (*RSF1, ACTL6B, AGRN, DNM1, KCNQ2, MYO18A*) (Seddighi et al., 2024) are not rescued by CHX.

In order to identify CEs that are induced by TDP-43 knockdown, but efficiently cleared by NMD, I focused on ones which were not called significant in the condition with TDP-43 knockdown, but present only when adding CHX.

Among the NMD-masked genes, two caught my attention, namely *SYNE1* and *CYFIP2* (Figure 23), the former being causative of a recessive form of spinocerebellar ataxia and muscular dystrophy (Attali et al., 2009; Fanin et al., 2015; Gros-Louis et al., 2007; Izumi et al., 2013; Noreau et al., 2013; Synofzik et al., 2016; Q. Zhang et al., 2007) and the latter of a neurodevelopmental disorder (Begemann et al., 2021; Y. Zhang et al., 2019; Zweier et al., 2019). Therefore, these two CEs could contribute to the neurodegenerative phenotype of TDP-43 proteinopathies in the brain.

Furthermore, in these two genes, CLIP data show the presence of TDP-43 binding sites in the introns where cryptic splicing happens(Figure 25). Interestingly, other RBPs that have been associated with ALS-FTD (H. J. Kim et al., 2013; Kwiatkowski et al., 2009; Vance et al., 2009) and with cryptic splicing (Ling et al., 2016) have binding sites around or on the CEs.

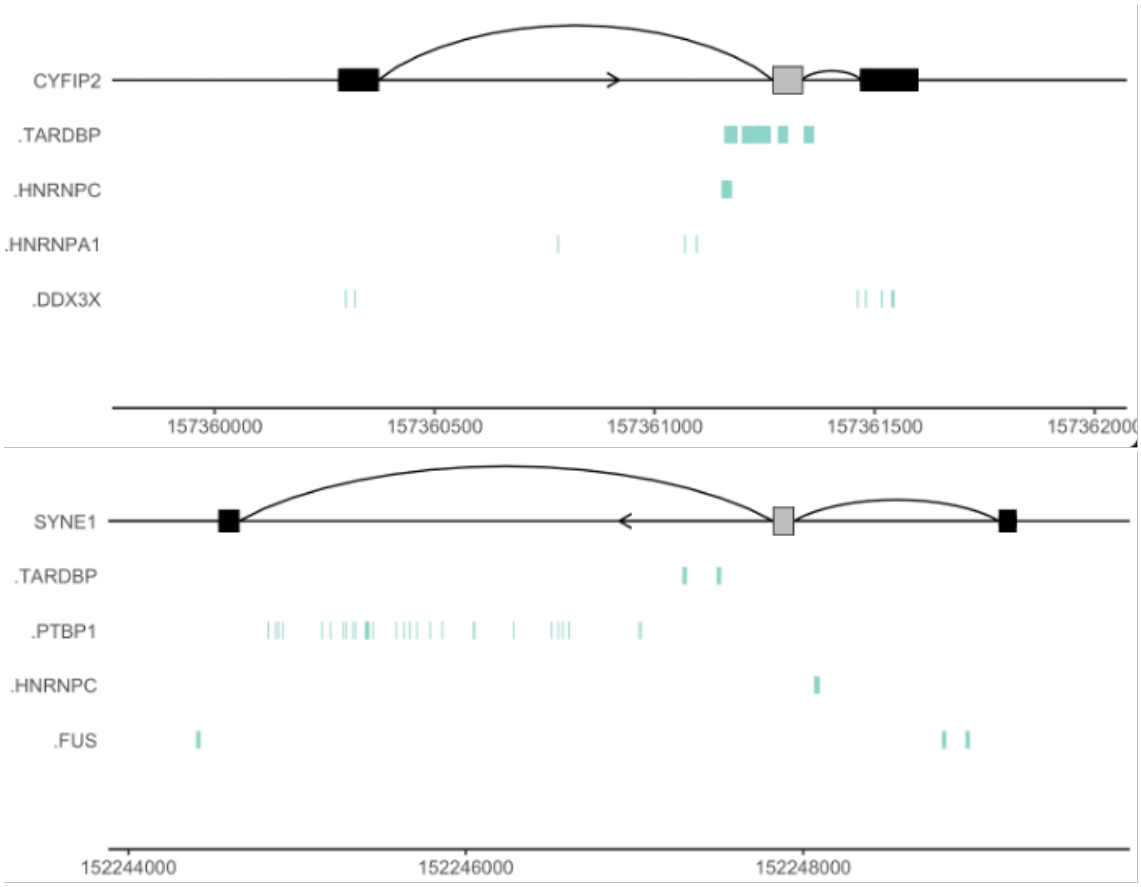


Figure 25: **RBP binding sites around *CYFIP2* and *SYNE1* CEs.** Visualisation of *CYFIP2* (top) and *SYNE1* (bottom) CEs. TDP-43 binding sites from POSTAR (Zhao et al., 2022) are within or around these CEs. Other ALS-FTD-related RBPs bind in the CE-containing intron.

To investigate the presence of NMD-masked CEs in human brains, I interrogated a dataset where nuclei of postmortem FTD patients were sorted using fluorescence-activated cell sorting (FACS) into TDP-43 positive and TDP-43 negative nuclei (Liu et al., 2019). Since the NMD machinery is located in the cytoplasm, it is more likely to pick up the NMD-sensitive CE junctions in nuclei, and indeed, I found both CEs in TDP-43 negative nuclei, as was found in a previous paper (X. R. Ma et al., 2022).

Inhibition of NMD by Double *TARDBP* and *UPF1* knockdown

The drawback of using CHX is that, since it stalls the ribosome, it impacts the transcriptome levels quite abruptly and has numerous secondary effects (Carrard et al., 2023). Therefore, a more elegant way to inhibit NMD is to selectively knock down factors of the NMD machinery, such as *UPF1*. Moreover, novel protocols to differentiate iPSCs into specific cell types, such as cortical neurons (i3Neurons), provide a better model to study human brain diseases (Fernandopulle et al., 2018). Therefore, I analysed the RNA-seq data from i3Neurons transduced with a virus

against *TARDBP* and one against *UPF1*.

The correlation of differential splicing results between the UPF1 KD and CHX experiment showed a strong and significant correlation, although stronger TDP-43 knockdown in SH-SY5Y cells led to higher average PSI in the CHX experiment (Figure 26).

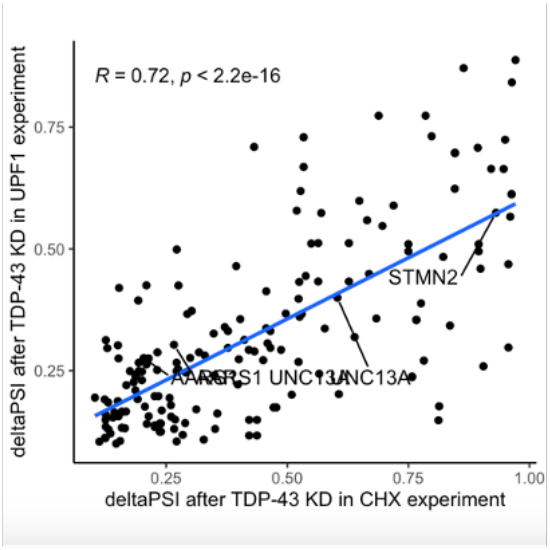


Figure 26: **Correlation of significant differential splicing, expressed as ΔPSI between TDP-43 knockdown and controls, between the two datasets.**

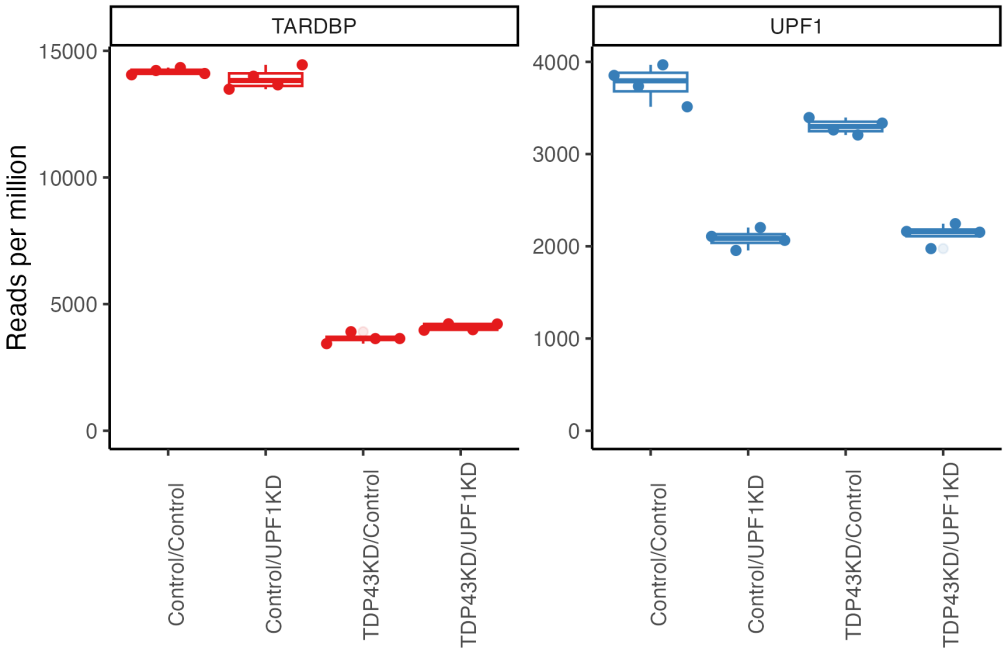


Figure 27: **Levels of *TARDBP* and *UPF1* from RNA-seq in the different experimental conditions.**

Differential gene expression and splicing analysis confirmed the downregulation of TDP-43 (Figure 27), and the inclusion of TDP-43 CEs, independent of *UPF1* knockdown.

However, possibly due to a low knockdown of *UPF1* (around 50%, Figure 27) or to compensating mechanisms (i.e. upregulation of other NMD factors), the events unmasked by NMD were only 12%, compared to 44% in the CHX experiment, and *CYFIP2* and *SYNE1* ΔPSI were not significant, although their expression is downregulated (Figure 28; see also Figure 33).

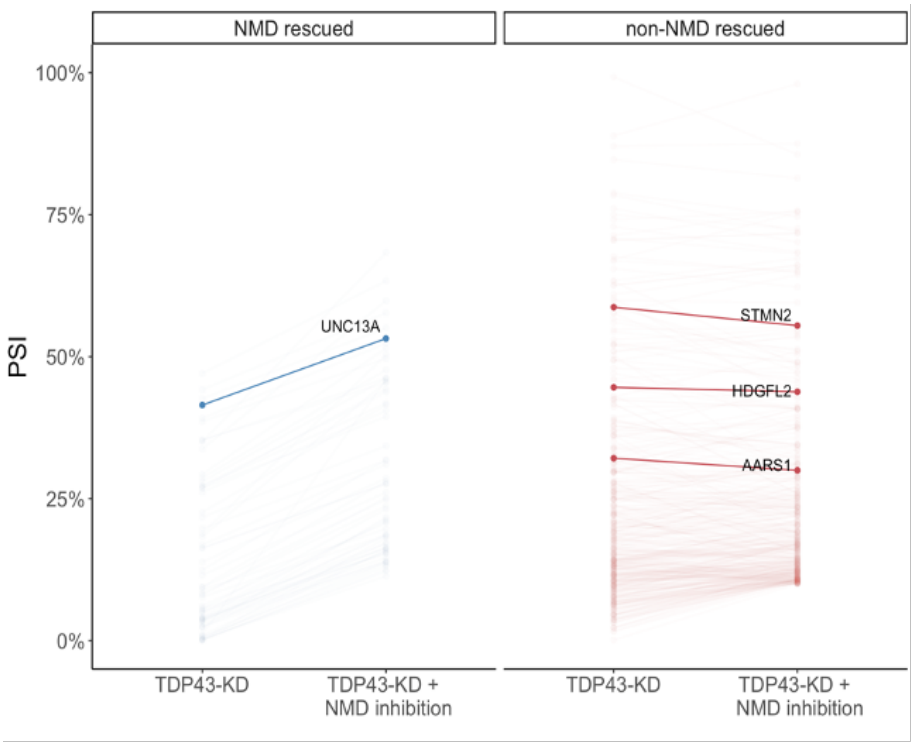


Figure 28: **CEs are rescued by NMD inhibition, as predicted by their sequences.** CE junctions were clustered by the rescue on their PSIs upon *UPF1* knockdown. *STMN2*, *HDGFL2*, and *AARS1* CEs, which are not predicted to introduce PTCs or frameshift exons, were not rescued, while UNC13A CE, which introduces a PTC, was not.

Validation of NMD-masked CEs in patient CNS

In order to validate the NMD-masked cryptic events, I interrogated bulk RNA-seq from postmortem ALS-FTD patients and controls from the NYGC, categorised by tissue of origin and TDP-43 pathology.

Previous works (Brown et al., 2022; Prudencio et al., 2020) have shown that *STMN2* and *UNC13A* cryptic expressions were exclusive to the diseases and tissues with TDP-43 pathology. As a cutoff, cryptic events need to be expressed in at least 10 different TDP-proteinopathy tissue samples and in no more than 5 different non-TDP-proteinopathy tissue samples (Brown et al., 2022).

This selective splicing pattern was confirmed for *SYNE1* and *CYFIP2*, which may be particularly relevant in TDP-43 proteinopathies because of their functions and relevance for other brain diseases. In the dataset, it is clear that the CE events in *SYNE1* and *CYFIP2* are disease- and tissue-specific for TDP-43 pathology (Figure 29).

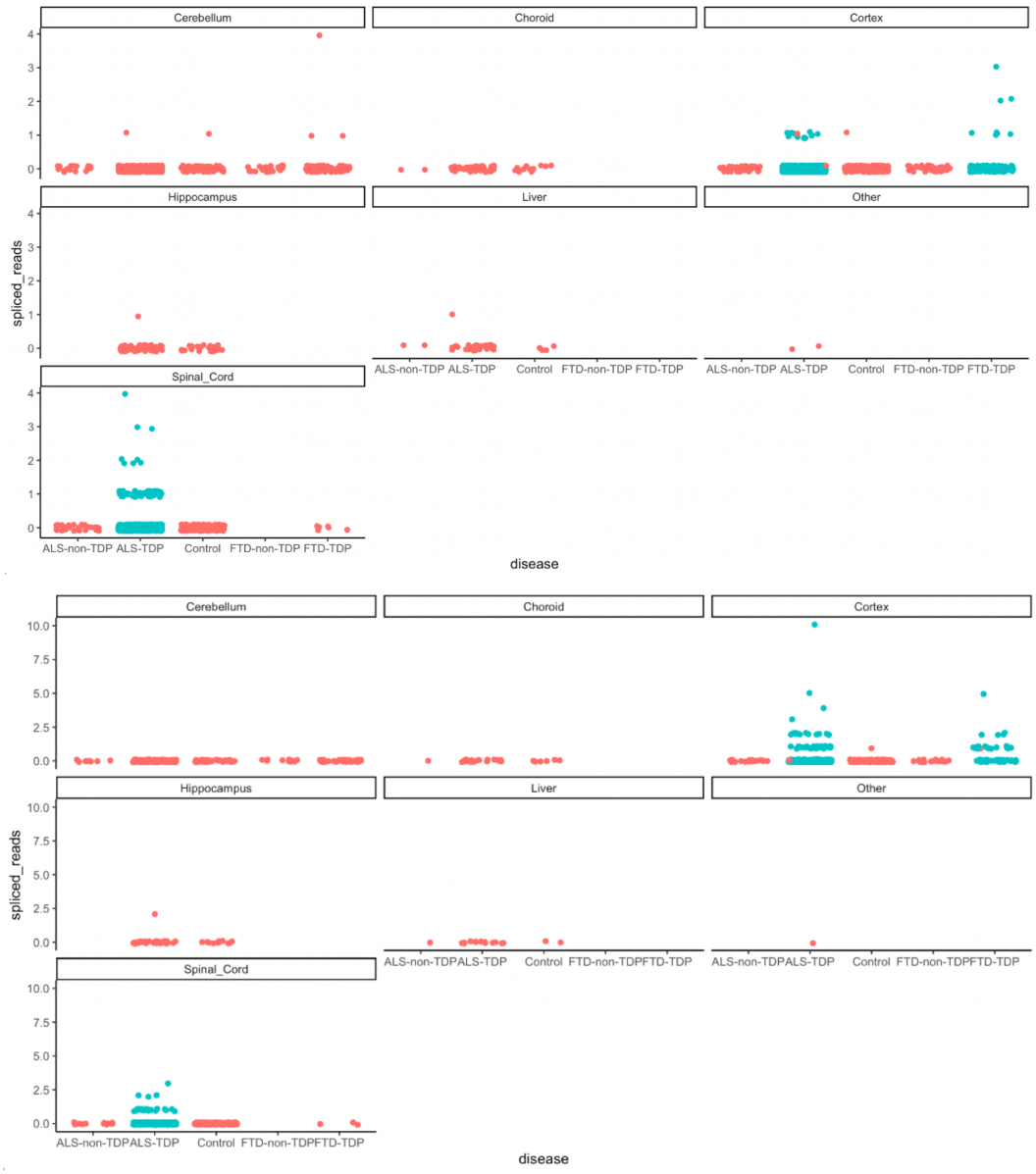


Figure 29: **Validation of *CYFIP2* and *SYNE1* in postmortem RNA-seq.** Parsing of *CYFIP2* (top) and *SYNE1* (bottom) CE junctions in the NYGC cohort showed disease (TDP-path - blue - versus non-TDP-path - pink) and tissue (cortex and spinal cord) specificity for both events.

Investigating the impact TDP-43 dosage has on CEs

How much TDP-43 loss is necessary for CEs to appear?

Analysis of two novel *in vitro* datasets

In a previous manuscript (Brown et al., 2022), SH-SY5Y cells treated with increasing amounts of TDP-43 knockdown showed that splicing and expression changes in *UNC13A* and *UNC13B* paralleled those of TDP-43.

An outstanding question is whether TDP-43-regulated cryptic exons all behave the same way. Which cryptic exons appear first? Which ones later? To answer this question and gain a more comprehensive understanding, I performed RNA-seq on three technical replicates of clonal SH-SY5Y or SK-N-BE(2) with increasing amounts of TDP-43 loss (Figure 30 and 31).

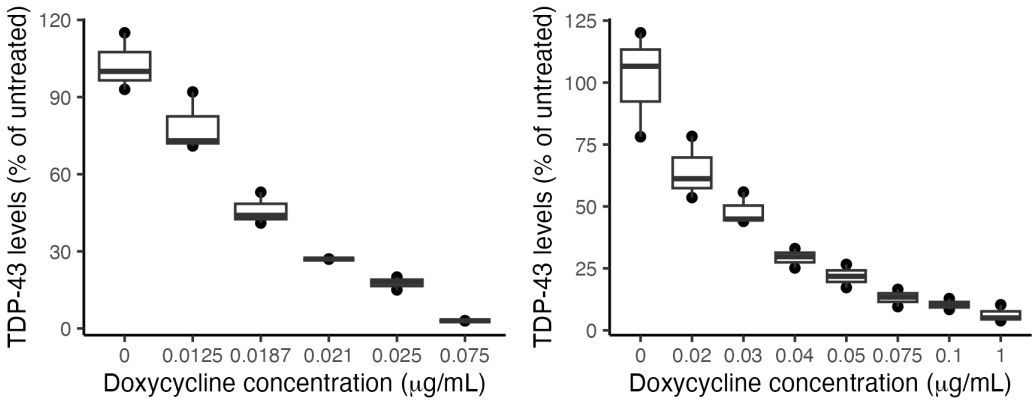


Figure 30: **Levels of TDP-43 in qPCR show a doxycycline-dependent response.** Levels of TDP-43, measured as GAPDH-normalised values on RT-qPCR in both SH-SY5Y (left) and SK-N-BE(2) (right) cells, showed a clear doxycycline-dependent response in both cell lines. Dots represent the value for each replicate, and boxplots highlight the median and first and third quartiles.

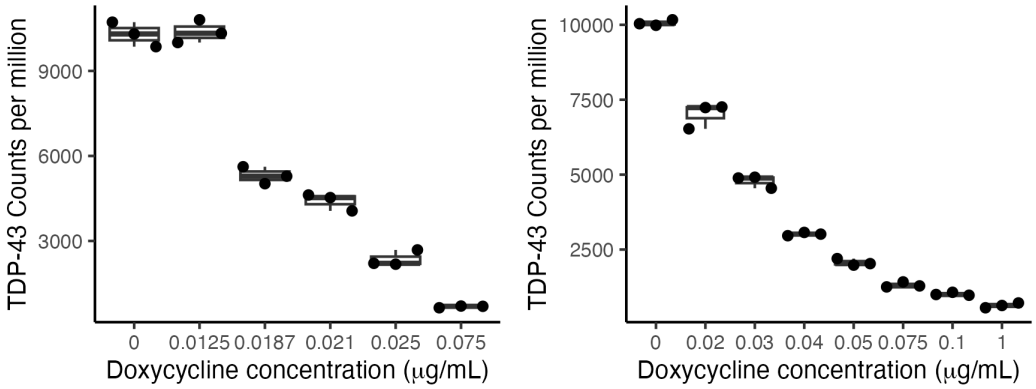


Figure 31: **Levels of TDP-43 from DESeq2 show a doxycycline-dependent response.** Levels of TDP-43, measured as normalised read counts from RNA sequencing in both SH-SY5Y (left) and SK-N-BE(2) (right) cells, showed a clear doxycycline-dependent response in both cell lines. Dots represent the value for each replicate, and boxplots highlight the median and first and third quartiles.

After RNA sequencing and alignment, I performed differential expression and differential splicing analyses using established pipelines in the two datasets.

Differential expression analysis showed significantly increasing up- and down-regulated genes with decreasing TDP-43 levels (Figure 32).

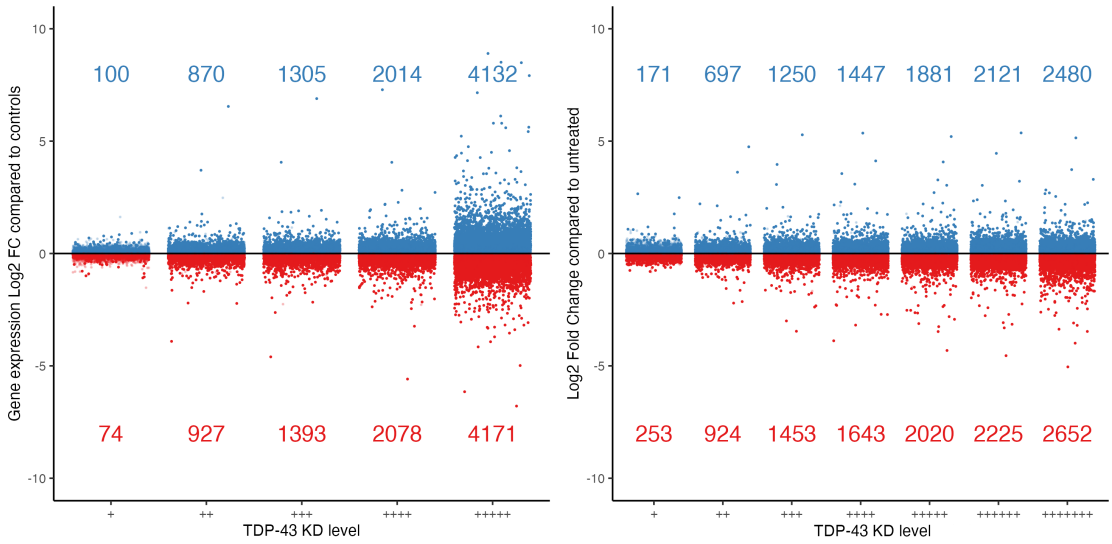


Figure 32: **Increasing TDP-43 knockdown is paralleled by gene expression changes.** Funnel plots showing that increasing TDP-43 knockdown causes more genes to be significantly up- and down-regulated in SY-SY5Y (left) and SK-N-BE(2) cells (right).

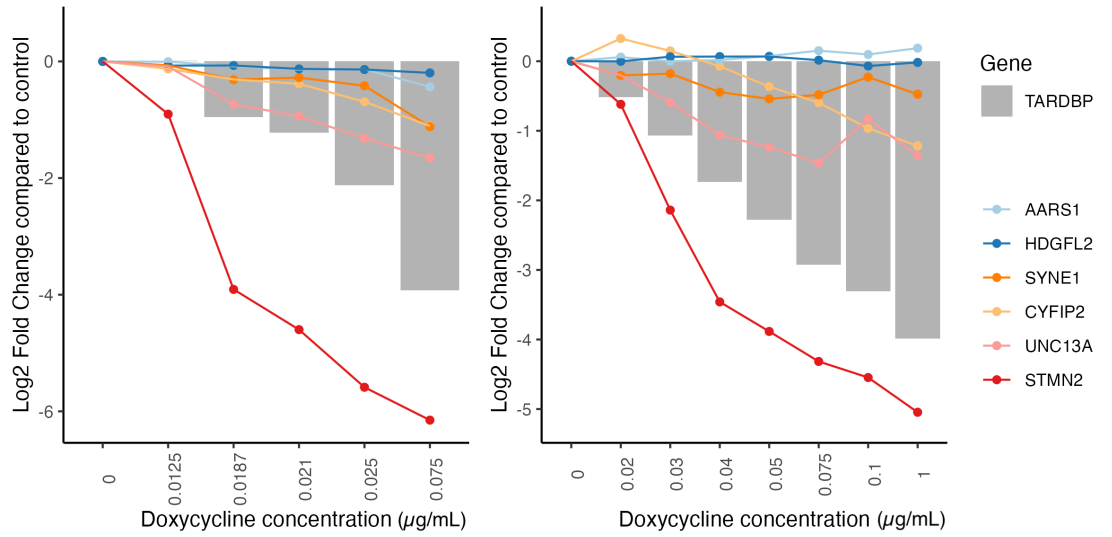


Figure 33: **Genes carrying CEs are responding to TDP-43.** Slope plots show that increasing TDP-43 knockdown (bars) causes changes in the expression of genes containing CEs in SY-SY5Y (left) and SK-N-BE(2) cells (right). Interestingly, also genes containing NMD-masked CEs (*CYFIP2* and *SYNE1*) showed decreased gene expression.

Interestingly, some of the CE-bearing genes were significantly down-regulated

only when a certain level of TDP-43 loss was achieved. Moreover, genes containing NMD-masked CEs showed reduced expression levels, confirming the fact that they are efficiently degraded and contribute to an overall reduction of the transcripts. (Figure 33).

At the splicing level, the trend is confirmed, with increasing doxycycline concentrations leading to the appearance of more CEs in SH-SY5Y cells (Figure 34).

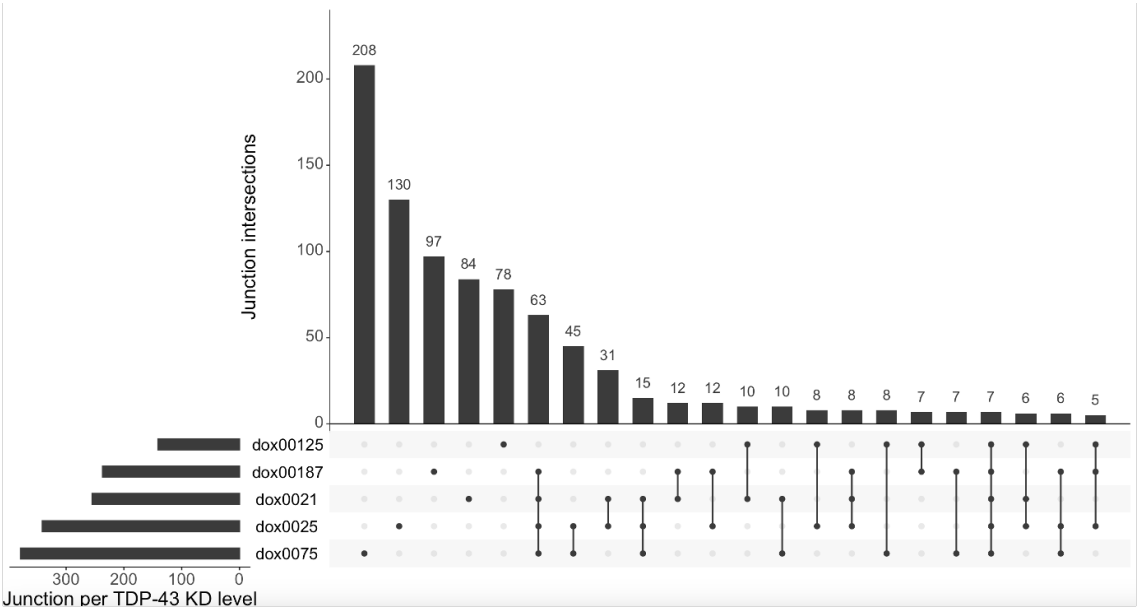


Figure 34: **Number of CEs correlates with increasing TDP-43 knockdown.** Upset plot in SH-SY5Y cells shows the number of differently spliced CE junctions and the co-occurrence of the junctions.

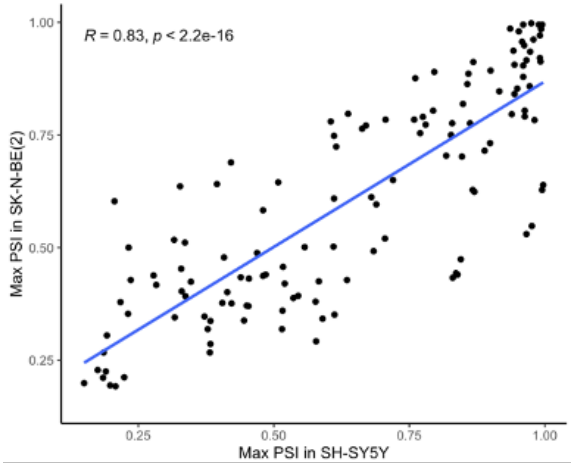


Figure 35: **Differential splicing across the two dose-response experiments.** As for the NMD experiment, the correlation of significant differential splicing between the two datasets, expressed as ΔPSI between TDP-43 knockdown and controls (right), is strongly significant.

Differential splicing analysis between the highest level of TDP-43 knockdown in both cell lines vs its own controls showed a significant correlation (Figure 35).

I then focused my attention on the pattern of appearance of CEs following increasing TDP-43 knockdown. Interestingly, *STMN2* is detected even with the mildest decrease in TDP-43 ("early"), while *UNC13A* and *HDGFL2* can be detected only with stronger levels of TDP-43 knockdown. Other events show up only with even stronger TDP-43 knockdown levels ("intermediate"), including *AARS1*, and, lastly, a set of CEs appears only when TDP-43 is completely lost from cells ("late") (Figure 36).

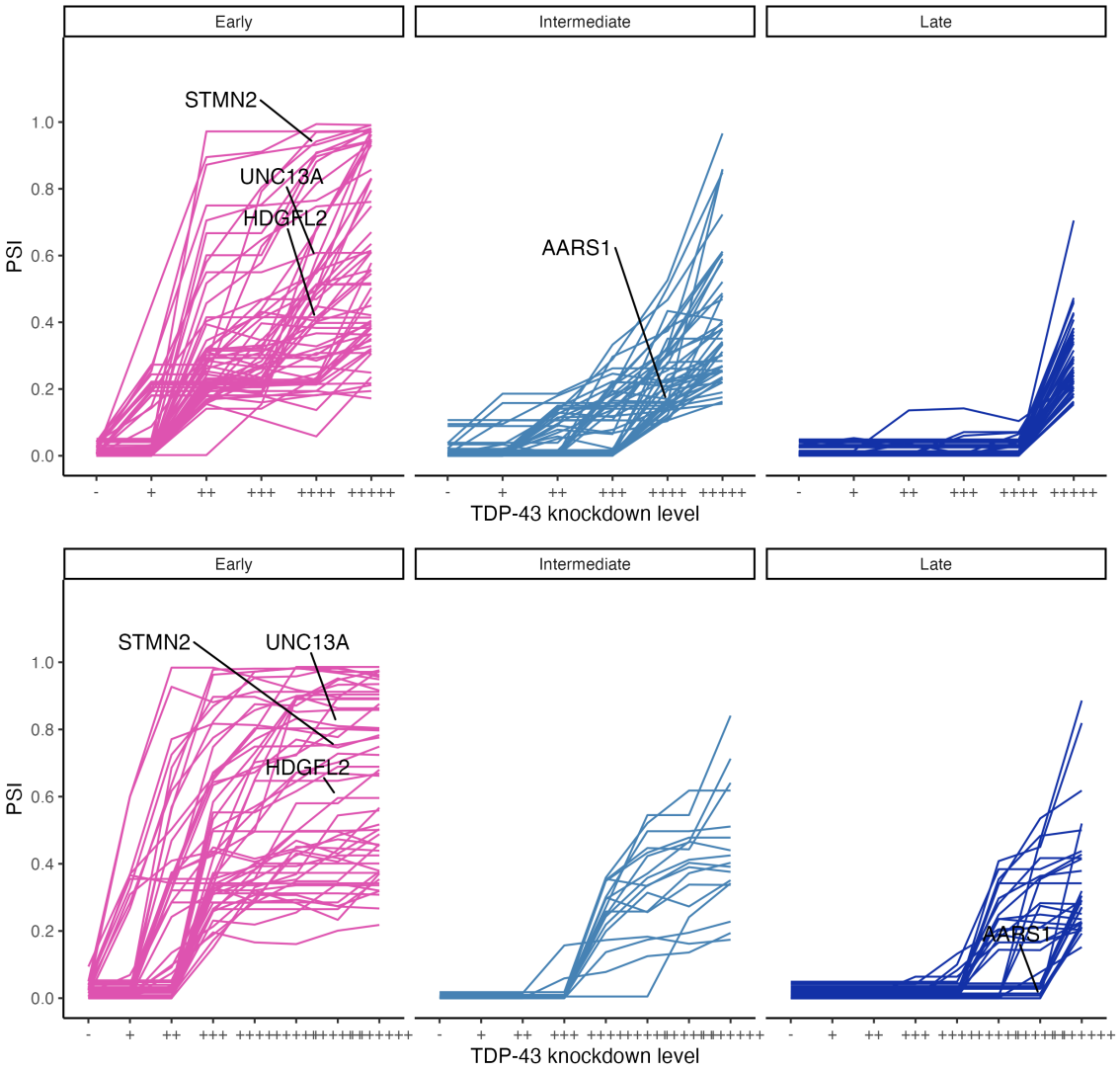


Figure 36: **Patterns of CE expression in response to increasing TDP-43 knockdown.** Spaghetti plots highlighting CE PSI with increasing TDP-43 knockdown levels in SH-SY5Y (top) and SK-N-BE(2) (bottom) cells. *STMN2* CE reaches a high PSI even with the mildest TDP-43 knockdown ("early" event), while *AARS1* CE needs a stronger TDP-43 knockdown to be detected ("intermediate" or "late" event). Interestingly, different CEs reach a different maximum PSI even when TDP-43 is completely depleted.

To further characterise the overlap between the two dose-response experiments, I visualised them through an alluvial plot (Figure 37). This showed that most of the events that are shared by the two cell lines fall into the same categories, although the classification in the SK-N-BE(2) resulted in fewer "intermediate" events, highlighting cell-specific differences in the appearance of TDP-43 CEs.

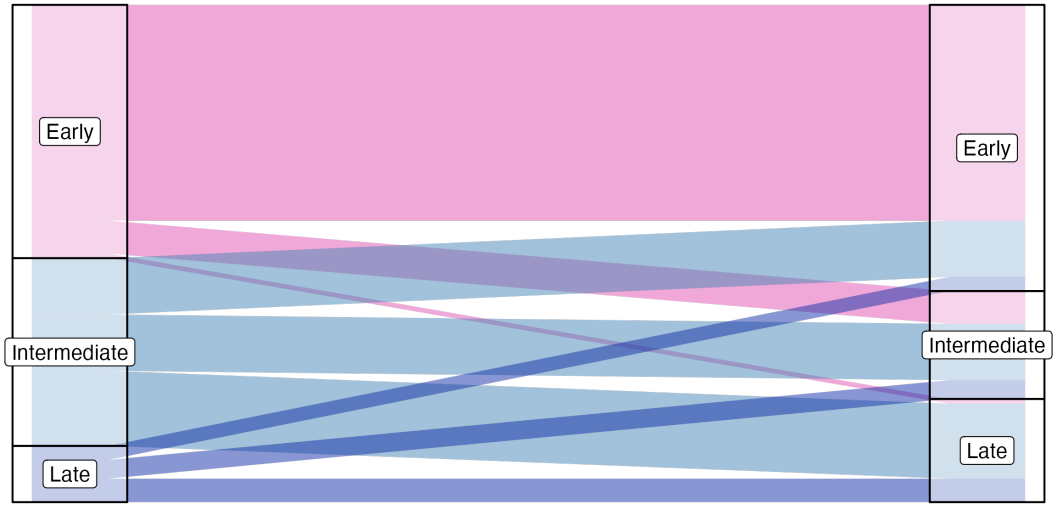


Figure 37: **Relationship between CE categories in the two experiments.** Alluvial plot shows the share of CEs between the three different categories (i.e. "early", "intermediate", and "late") in SH-SY5Y (left) and SK-N-BE(2) (right) cells.

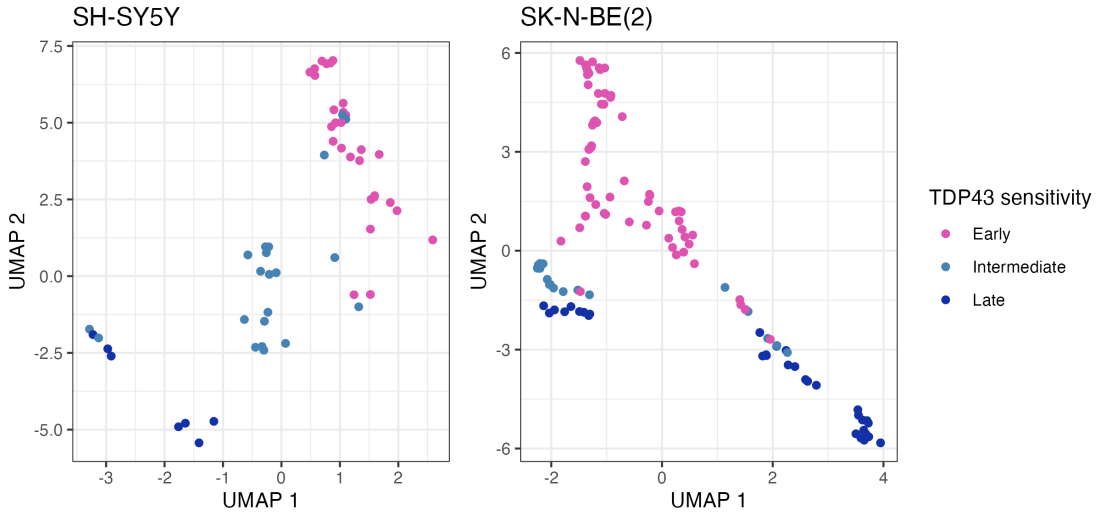


Figure 38: **Unsupervised clustering detects patterns of TDP-43 dose-response.** UMAP clustering of CEs in in SH-SY5Y (left) and SK-N-BE(2) (right) cells.

As the classification into "early", "intermediate", and "late" is arbitrary, I compared these categories with the results from an unsupervised machine learning anal-

ysis, by employing the Uniform Manifold Approximation and Projection (UMAP) algorithm.

In Figure 38, it is evident that the manual categories seem to also be distinct in the UMAP space for both SH-SY5Y and SK-N-BE(2) cells.

Validation of TDP-43 dosage effect on cryptic exons *in vivo*

To assess the relevance of these findings in a disease-relevant context, I performed a targeted RNA sequencing experiment on postmortem brains from FTD and healthy controls. Briefly, by selective PCR amplification of a panel of transcripts with and without CE inclusion followed by Illumina sequencing, I was able to accurately assess the relative expression of CEs in brain tissue When grouping by category and considering the absolute count of reads covering CEs (Figure 39), FTD cases appear to have strikingly more CEs than healthy controls, and there is a trend where early events have more counts than intermediate and late ones.

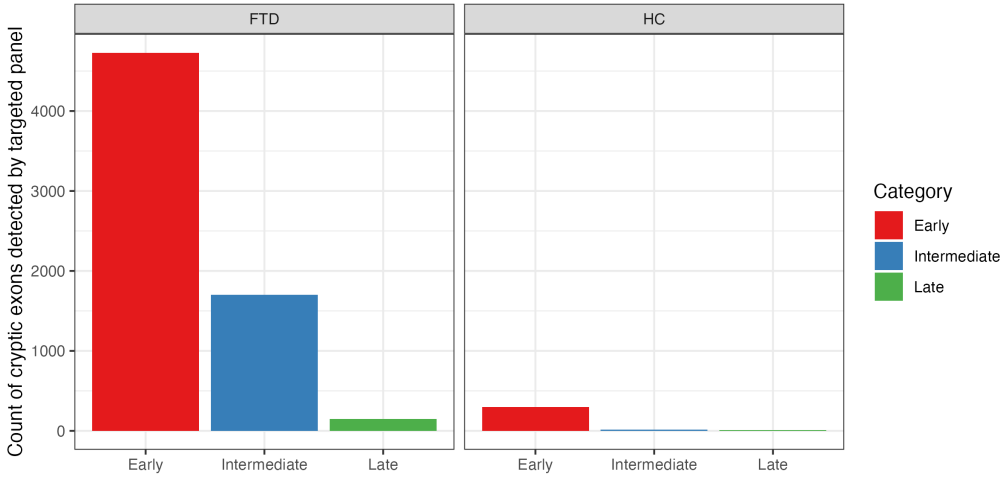


Figure 39: **FTD cases have higher inclusion of CEs than healthy controls.** Pooled counts of CEs detected by targeted panel highlight higher expression of "early" CEs compared to "intermediate" and "late" ones.

4.4 Discussion

While the presence of TDP-43 aggregation in the cytoplasm of ALS and FTD brains has been known for almost 20 years (Arai et al., 2006; Neumann et al., 2006), less has been known about the role of TDP-43 nuclear loss of function in the disease pathogenesis. However, mounting evidence suggests that alterations in TDP-43 function as a splicing regulator are paramount in the cell breakdowns responsible for the disease manifestations. Indeed, several papers suggested that TDP-43 binds to

around a third of the pre-mRNA and that the loss of this crucial protein contributes to neuronal degeneration through missplicing (Polymenidou et al., 2011; Tollervey et al., 2011). Later, in a seminal paper (Ling et al., 2015), the derepression of non-conserved intronic sequences called cryptic exons when TDP-43 is lost has been described.

The relevance of this mechanism in the disease pathogenesis has been confirmed by the following works that focused on two of these CEs, namely *STMN2* and *UNC13A*. While the former is a crucial protein for axonal maintenance, the latter is a key synaptic protein. The CE in the first intron of *STMN2* contains a polyadenylation signal and causes a corresponding decrease in the levels of the native transcript. This leads to the inhibition of axonal outgrowth and regeneration (Klim et al., 2019; Melamed et al., 2019). On the other hand, *UNC13A* has been a known genome-wide association study (GWAS) hit for ALS and FTD (Diekstra et al., 2014; Pottier et al., 2019; Van Es et al., 2009), but only in the last years have researchers unravelled the interplay between the inclusion of a CE in intron 20 of this gene and the presence of the risk SNPs in the same intron (Brown et al., 2022; X. R. Ma et al., 2022). This is responsible for reduced survival of risk allele carriers by facilitating the inclusion of the CE, which includes a premature stop codon with subsequent degradation via NMD and protein loss.

Another mechanism by which cryptic splicing plays a pathogenic role is the generation of in-frame CEs that can escape NMD and be translated to novel peptides. Indeed, these peptides have been found in human CSF and plasma (Calliari et al., 2024; Irwin et al., 2024; Seddighi et al., 2024).

However, while these data demonstrate that CEs play a pathogenic role in ALS-FTD, a thorough understanding of the biology of TDP-43 cryptic splicing is still lacking. Therefore, in this chapter, I aimed to characterise the biology of TDP-43 CEs by experimental and computational approaches. First, by inhibiting NMD, I showed that while some CEs evade NMD and could serve as biomarkers, others are revealed only upon NMD inhibition. Later, I analysed how TDP-43 loss affects CE expression, highlighting differential responses between CEs. I orthogonally validated these findings by iCLIP and RNA-sequencing analysis on postmortem tissues.

First, I showed how inhibition of NMD, either by CHX treatment or *UPF1* knockdown, can rescue the expression of CE-bearing transcripts in *UNC13A* and *UNC13B*, while the expression of *STMN2* is not rescued. This is in agreement with the fact that *STMN2* CE is in-frame and contains an alternative poly-adenylation site, while *UNC13A* CE is predicted to introduce a premature stop codon and *UNC13B*

CE is predicted to introduce a frameshift with a PTC in a downstream exon.

Moreover, I showed how NMD efficiently masks the *UNC13A* CE transcripts by inducing a mild knockdown in SH-SY5Y cells and rescuing their expression by means of CHX treatment.

These experiments also confirmed the relationship between TDP-43 loss of function and CE expression.

Later, by performing whole-transcriptome RNA-seq on the CHX experiment, I visualised how NMD impact the expression of several CEs. Particularly, I confirmed that *UNC13A* CE is rescued upon CHX treatment, while *STMN2* CE is not. Furthermore, I showed that in-frame CEs in *AARS1* and *HDGFL2* are also stable when cells are treated with CHX. Eventually, I uncovered CEs in *SYNE1* and *CYFIP2*, which are completely masked by NMD. *SYNE1* mutations cause recessive spinocerebellar ataxias and muscular dystrophies, and *CYFIP2* mutations cause neurodevelopmental disorders (Attali et al., 2009; Begemann et al., 2021; Fanin et al., 2015; Gros-Louis et al., 2007; Izumi et al., 2013; Noreau et al., 2013; Synofzik et al., 2016; Q. Zhang et al., 2007; Y. Zhang et al., 2019; Zweier et al., 2019). The presence of CEs in *SYNE1* and *CYFIP2* could lead to neuronal defects in brain TDP-43 proteinopathies such as ALS and FTD.

To strengthen the findings from RNA-seq, I made use of CLIP data from the POSTAR database (Zhao et al., 2022), showing that both *CYFIP2* and *SYNE1* CE-containing introns are bound by TDP-43.

Since CHX can lead to cell death, I used a more elegant way, that is, the inhibition of the key NMD factor *UPF1*, to inhibit NMD in iPSC-derived cortical-like neurons. Here, I could confirm the rescue of *UNC13A* CE and the stability of *STMN2, HDGFL2*, and *AARS1* CEs upon NMD inhibition. However, possibly due to the mild knockdown of *UPF1* and/or to a different basal gene expression, I wasn't able to unmask the CEs in *SYNE1* and *CYFIP2*.

Eventually, I looked for the presence of NMD-masked CEs in *SYNE1* and *CYFIP2* in postmortem brain tissues from people affected by ALS and FTD from the NYGC. Here, I was able to see that both CEs are present only in people affected by the diseases and only in tissues where TDP-43 is expected to aggregate.

In the last part of the chapter, I analysed two sets of human neuronal cells, SH-SY5Ys and SK-N-BE(2), with increasing levels of TDP-43 knockdown. I first validated the dose-response curves by means of qPCR and then confirmed them in the RNA-seq data.

I then visualised the impact of different levels of TDP-43 knockdown on gene expression levels, with a clear rise in the number of significantly up- and down-

regulated genes with increasing TDP-43 knockdown.

Furthermore, I showed how increasing TDP-43 loss of function leads to increasing loss of *STMN2, UNC13A, CYFIP2* and *SYNE1* transcripts, while it does not change the expression of *HDGFL2* and *AARS1*, which contain in-frame CEs.

I then clustered the CEs into three categories, namely "early", "intermediate", and "late", based on their appearance in the dose-response curve, and showed a good overlap between the two cell lines. By using an unsupervised clustering algorithm, UMAP (McInnes et al., 2018), I could see that the three categories separate nicely on the UMAP space.

Eventually, by performing a targeted RNA-seq experiment, I could see that the expression of "early" CEs is greater than that of "intermediate" and "late" in postmortem tissue from FTD patients.

A limitation of this study is that the RNA-seq experiments presented were conducted using technical replicates, meaning a single biological sample was sequenced, without independent replicates from distinct cultures or individuals. While technical replicates (e.g. different culture plates, sequencing lanes or library preps from the same RNA source) may reduce stochastic noise, they do not account for biological variability, such as differences in gene expression due to passage number or differentiation state. As a result, differential expression analyses in this context lack statistical power and cannot be subjected to formal hypothesis testing (e.g. adjusted p-values or confidence intervals). Observed differences should therefore be considered qualitative or exploratory, and may highlight candidate pathways or genes, rather than provide definitive evidence of regulation. Future validation with biological replicates (e.g. independent differentiations or donor-derived lines) will be necessary to confirm the robustness and generalizability of these transcriptomic findings.

In this chapter, I generated novel in vitro datasets, which were used to make hypotheses that could be validated in patient tissue. Eventually, when the manuscript related to this chapter is published, these datasets will be made available to the scientific community to foster research on related topics.

The fact that different cell lines showed different patterns of expression of cryptic exons raises interesting questions about the tissue and cell-specificity of TDP-43 CEs, particularly in consideration of the relative expression of TDP-43. With regards to this, it would be important to perform cross-linking and immune precipitation (CLIP) on the differential knockdown cell lines, to show which TDP-43 binding sites are lost at each level of knockdown. Moreover, it will be interesting to use this molecular biology technique to assess which mRNA sites are bound by TDP-43 when

it acts as a monomer and as a multimer. This is particularly interesting in light of a recent paper that showed that TDP-43 condensation properties alter its binding specificity to RNA (Hallegger et al., 2021).

The different sensitivity of CEs to TDP-43 loss of function can be due not only to basal TDP-43 expression and its relative loss but also to other *trans* mechanisms, including the differential expression of other RBPs. Particularly, TDP-43 belongs to the hnRNP family (Bampton et al., 2020), with several members of this family being involved with the pathogenesis of ALS and other TDP-43 proteinopathies. Among these, mutations in hnRNPA1 and hnRNPA2B1 are responsible for genetic forms of TDP-43 proteinopathies (H. J. Kim et al., 2013); hnRNPK has been found to mislocalise and induce cryptic splicing in FTD (Bampton et al., 2021). Moreover, ELAVL3 and CELF5, two RBPs that play crucial roles in neuronal development (Fisher and Feng, 2022), contain CEs and can therefore act downstream of TDP-43 to induce other transcriptomic alterations. Therefore, performing knockdown experiments of other RBPs, including *hnRNPA2B1, hnRNPA1, CELF5*, and *ELAVL3* could provide more insights on their role in the expression of CEs.

Aside from *trans* factors, the mRNA sequence (and structure) itself can act as a *cis* modifier of TDP-43 cryptic splicing. Among the relevant factors in the RNA sequences, it is possible that the length and purity of TDP-43 binding motifs, as well as the distance from these motifs to the CE splice sites, can modulate the avidity of the binding of TDP-43 to the pre-mRNA, and therefore the repression of these events. Moreover, the strength of the splice sites can impact the chances of inclusion of CEs.

Moreover, the sensitivity of the transcripts containing CEs to NMD can alter the detection of these CEs. Indeed, I confirmed the stability of the transcripts containing in-frame CEs when treating the cells with CHX. Moreover, these transcripts are translated into novel peptides, making them appealing as biomarkers and suggesting possible novel roles for them in the disease pathogenesis, including the presentation of neoantigens to the immune system.

Eventually, the different sensitivity of CEs to TDP-43 knockdown can be relevant as a readout of TDP-43 loss of function and response to treatment in clinical trials, as well as staging of TDP-43 proteinopathies, considering the spreading of TDP-43 pathology (Braak et al., 2013). With regard to this, it will be interesting to look at other stressors (including pharmacological drugs) that can increase or decrease the propensity of TDP-43 mislocalization from the nucleus, thus modulating or even inducing the expression of CEs.

Chapter 5

TDP-43 CRYPTIC PEPTIDES IN HUMAN DISEASE: DETECTION AND NOVEL ROLES

5.1 Contributions

Validation of the TCRs (Table V) was done by Shahab Chizari. Independent assessment of immunohistochemistry analysis was performed by Dr. Ashirwad Merve.

5.2 Overview

In the previous chapter, I used very homogeneous and reliable cell models to address specific questions regarding TDP-related RNA misprocessing. However, to assess their impact on TDP-43 proteinopathy pathogenesis, CEs must also be characterised *in vivo*. This is a paramount step in the discovery of novel therapeutic targets and biomarkers for TDP-43 proteinopathies. Therefore, in my second aim, I try to identify specific *in vivo* molecular signatures of TDP-43 loss of function, particularly in the context of muscle diseases.

In order to unearth the relationship between TDP-43, RNA, and neurodegeneration, we can make use of muscle from diseases where TDP-43 has been shown to mislocalise to the cytoplasm. The advantages brought by analysing this tissue are the early disease stage, since biopsies are usually taken at the time of diagnosis (compared to postmortem brain), and the opportunity to combine different techniques including immunohistochemistry, proteomics, and RNA-seq on serial sections, therefore relating the extent of the TDP-43 pathology with the molecular changes driven by TDP-43 loss of nuclear function.

Previous works have characterised the presence of novel peptides derived from TDP-43 CEs in ALS, FTD, and AD, where the TDP-43 pathology happens in the CNS (Calliari et al., 2024; Irwin et al., 2024; Seddighi et al., 2024). However, because of the BBB, the potential for antigen presentation to immune cells is lower.

On the other hand, one of the hallmarks of sIBM pathology is the presence of immune infiltrate, and recent work suggests that features of neurodegeneration, including TDP-43 loss of function and rimmed vacuoles, precede the immune response and are not resolved upon T cell depletion (Britson et al., 2022). The same work described a set of CEs in sIBM, including that in *HDGFL2*; however, characterisation of the presence and the extent of cryptic peptides resulting from TDP-43 proteinopathy in sIBM has not been reported.

Particularly, HDGFL2 is a ubiquitously expressed histone-binding protein (https://www.ncbi.nlm.nih.gov/gene/84717). A TDP-43 induced CE derived from the *HDGFL2* gene is reproducibly detected via transcript analysis and proteomics in various cell lines, including human iPSC-derived neurons, HeLa cells, in addition to CSF samples from patients with ALS-FTD (Calliari et al., 2024; Irwin et al., 2024; Seddighi et al., 2024). Furthermore, two different groups have developed antibodies that bind to the CE-encoded peptide within HDGFL2 (Calliari et al., 2024; Irwin et al., 2024). Irwin et al. developed and used their HDGFL2 CE monoclonal antibody to detect the cryptic peptide in both the CSF and blood of individuals with ALS-FTD, and demonstrated that these effects occur during the presymptomatic and early stage of the disease (Irwin et al., 2024). Similarly, Calliari et al. developed an immunoassay utilising a polyclonal HDGFL2 CE-specific antibody to show that HDGFL2 CE peptide was significantly increased in brain regions with TDP-43 pathology in FTLD-TDP and AD-TDP, compared to non-TDP-43 controls (Calliari et al., 2024). These findings not only indicate that *HDGFL2* CE can be a reliable fluid biomarker, but that its reliable presence in the blood, CSF, and brain suggests clear targets for an adaptive immune response.

5.3 Results

TDP-43 mislocalization and cryptic splicing in human muscle

To assess the presence of CE peptides in skeletal muscle biopsies, I gathered one previously published RNA-seq dataset (Britson et al., 2022) and two newly generated cohorts, comprising a total of 11 controls and 13 sIBM cases. By using the splicing junction coordinates corresponding to the cryptic peptides listed in Table IX, I was

Gene	CE peptide sequence
AARS1	LECSGMITAHCSLNFLGSSDPLPSSWDYR
ACTL6B	ASFELLGSSNPPASASQSAGIIG
ARHGAP22	GGWMNPWMNRWMHGWMHGWIHGCMDEQMWRNEEPRIPETEQKLTSP
	LQKPSDGGAEPEPWADAVPSQAGPRAEGGLAEEAEEHHEELAAALVCA
	AWGSAFLLQGQR*
CELSR3	LRGAG
DNM1	PAPPVRVNCHVDFVLSFR
EPB41L4A	SDIESPYKTEVTKGQAEVCESVCAYV
HDGFL2	EPTIWFGKGHSGMLASEGREAVLTRLHESERVRKQERERDTEERRE
IGLON5	QPRAGPSHSRSPRSSSRPSSLSAWCQLHRLRMPSPHSLALVQPAGHGLQ
	RDDCVGVTEPACCAEVCEWAWAHTCKHALANLSTQCACVDVCVCKG
MYO18A	VKEEDKTLPKPGSPGKEEGA
NECAB2	SFSASFCMGGFGHSLSIFESLLCAGHSAR
PHF2	GKHE*
PTPRZ1	GLTLSPRLECRGTISAHCNLPLPGLTDPPTSASRVARTI
PXDN	VHSDTCDLDKQKWAGETLEKRSSDGLELC
SLC24A3	(E)GRKEGIRKEEWMDGWI
STMN2	GLGRRPSRER*
SYNJ2	AVEHVGQHGV*
XPO4	LAHCWRLTVCVCVW
ZNF423	VLGSAWDSRGSLCGEPCRETT

Table IX: **List of TDP-43 CE peptides assessed in human skeletal muscle.**

able to detect them in a subset of the sIBM cases while being almost undetectable in controls (Figure 40). Particularly, the CE in *HDGFL2* showed the highest specificity (0/11 healthy controls) and sensitivity (11/13 sIBM cases). Conversely, the CE in *MYO18A* was found also in controls (Figure 40).

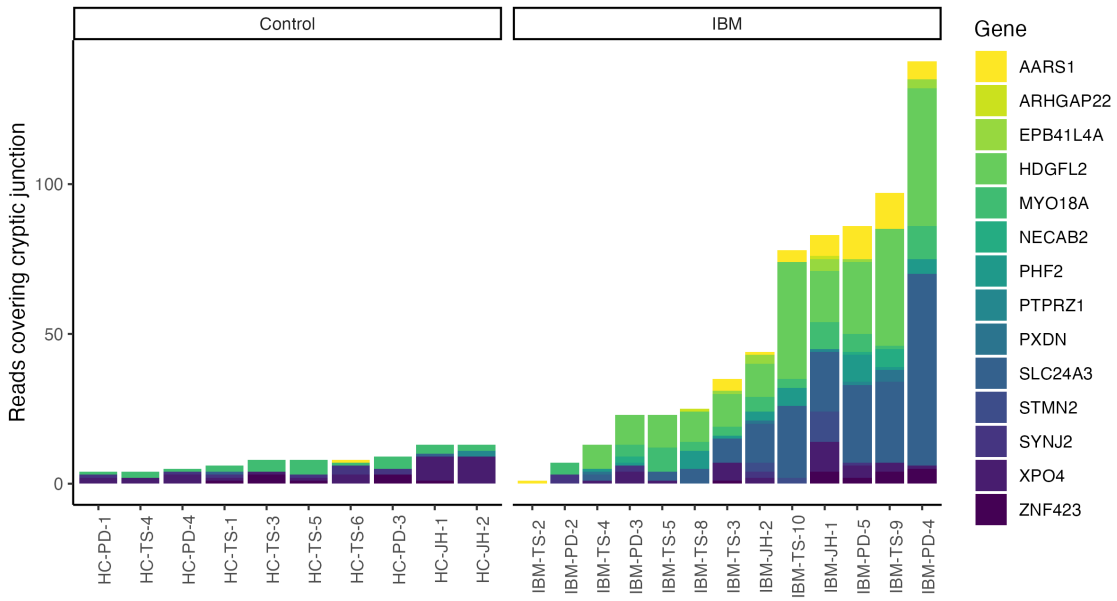


Figure 40: **Raw counts of RNA-seq reads covering CE peptide junctions.** Controls (n = 11), sIBM (n = 13).

Other CEs weren't found in any of the sIBM cases, while they were previously reported in the CSF of people with ALS and FTD (Seddighi et al., 2024). Among the factors responsible for the difference between the two sets of diseases, the most relevant is the difference in gene expression between the two tissues. To assess this, I interrogated the ASCOT database (Ling et al., 2020) (Figure 41).

Particularly, the gene expression in different tissues can explain the absence of CEs in *ACTL6B* and *IgLON5* in the sIBM cohort, as these genes are lowly expressed in the skeletal muscle. Strikingly, I was able to detect the CE in *STMN2* in some sIBM cases (Figure 40), although whether this is due to concomitant nerve pathology or transient expression in regenerating muscle cells is unknown (Vogler et al., 2018). Other factors that can account for tissue-specific differences in the inclusion of CEs are the extent of the TDP-43 loss and the presence and expression levels of other RBPs that can modulate the splicing of CEs.

In summary, this analysis reliably identified cryptic splicing events that are predicted to give rise to cryptic proteins in sIBM muscles.

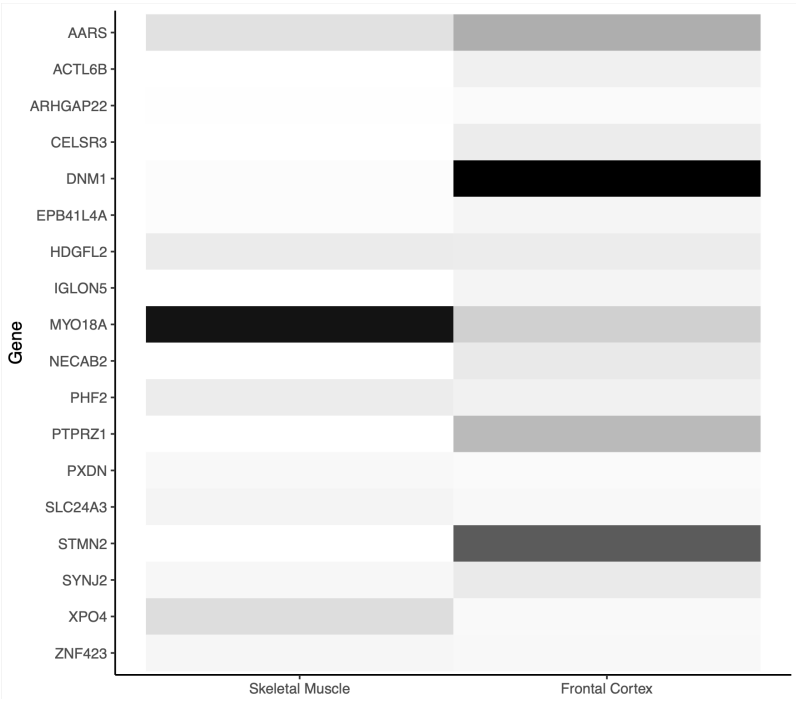


Figure 41: **Expression of genes carrying CE peptides in skeletal muscle and frontal cortex.** Data obtained from ASCOT.

Detection of HDGFL2 cryptic peptide by immunohistochemistry

To be able to visualise cryptic splicing and its relationship with TDP-43 nuclear loss and cytoplasmic accumulation in sIBM, as well as with the immune response, I gathered a novel bioptic cohort consisting of 17 sIBM patients and 4 controls, on which I performed immunohistochemistry and proteomics.

TDP-43 and p62 accumulation is well-established features of sIBM and are present in the majority of cases (Weihl et al., 2008). By means of immunohistochemistry, I confirmed TDP-43 nuclear clearance and cytoplasmic aggregation in this sIBM cohort and its colocalisation in muscle fibres with p62-positive aggregates (Figure 42). Using a newly generated antibody specific for the HDGFL2 cryptic peptide, I was able to test the presence of this product of TDP-43 mis-splicing in IBM muscles. I detected it using immunohistochemistry in 9/10 sIBM cases and 0/4 controls and showed its specificity for diseased fibres (Figure 42 and Table X).

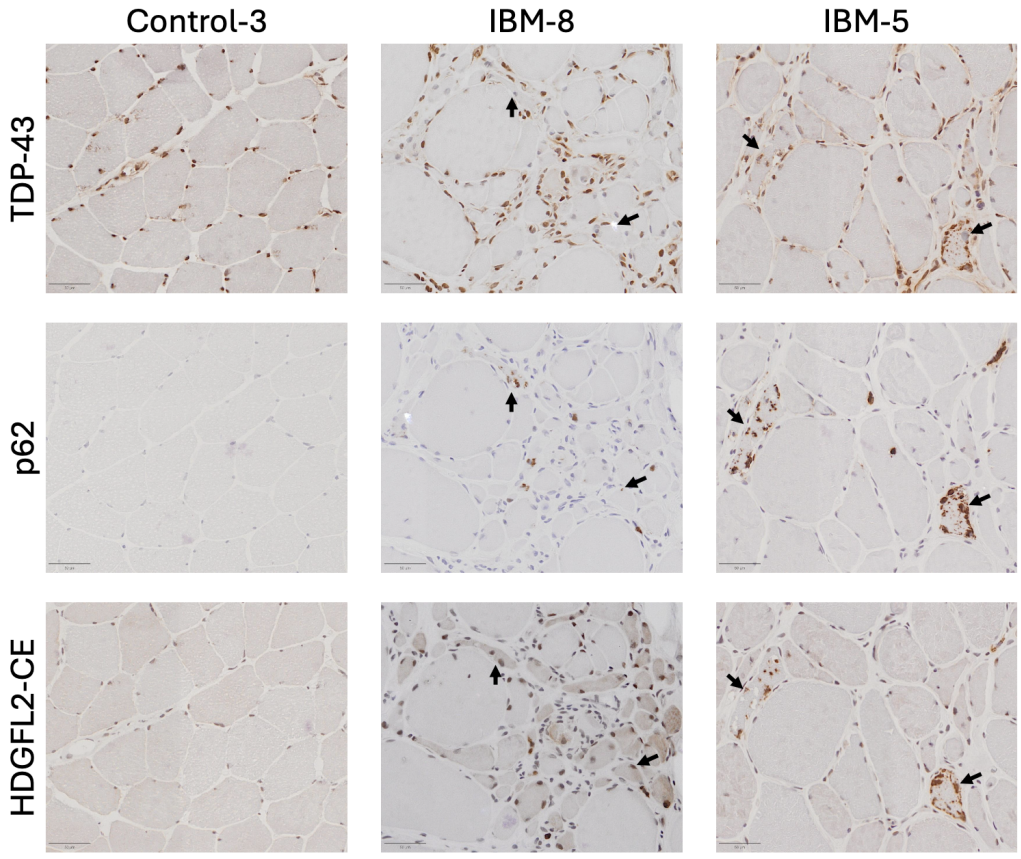


Figure 42: **Immunohistochemistry detection of TDP-43, p62, and HDGFL2 cryptic peptide in muscle.** For each case, arrows show colocalisation of HDGFL2 cryptic peptide in cells with p62 and TDP-43 nuclear depletion and/or aggregation.

In summary, my immunohistochemistry analysis revealed the presence of HDGFL2 cryptic peptide in IBM muscle. Unexpectedly, the cryptic peptide protein was also accumulated in p62-positive and TDP-43-positive inclusions.

Case	Immune infiltrate	TDP-43 nuclear loss	p62	HDGFL2-CE
Control-1	-	NA	-	-
Control-2	-	-	-	-
Control-3	-	-	-	-
Control-4	-	-	-	-
IBM-17	-	-	-	-
IBM-16	+	+	+	+
IBM-2	+	++	++	++
IBM-5	+	++	++	++
IBM-7	++	++	++	++
IBM-9	++	NA	++	++
IBM-10	++	+++	++	++
IBM-8	++	+++	++	+++
IBM-3	++	+++	+++	+++
IBM-13	++	+++	+++	+++

Table X: **Semi-quantitative scores for colocalisation of immune infiltrate, TDP-43 loss, p62 and HDGFL2 cryptic peptide.**
Scores range from 0 (-) to 3 (+++) based on two independent evaluations, of which one from me and one from an expert neuropathologist. NA means that due to technical problems with the staining, it was not possible to quantify the extent of TDP-43 nuclear loss.

Cryptic peptides correlate with immune infiltration in sIBM

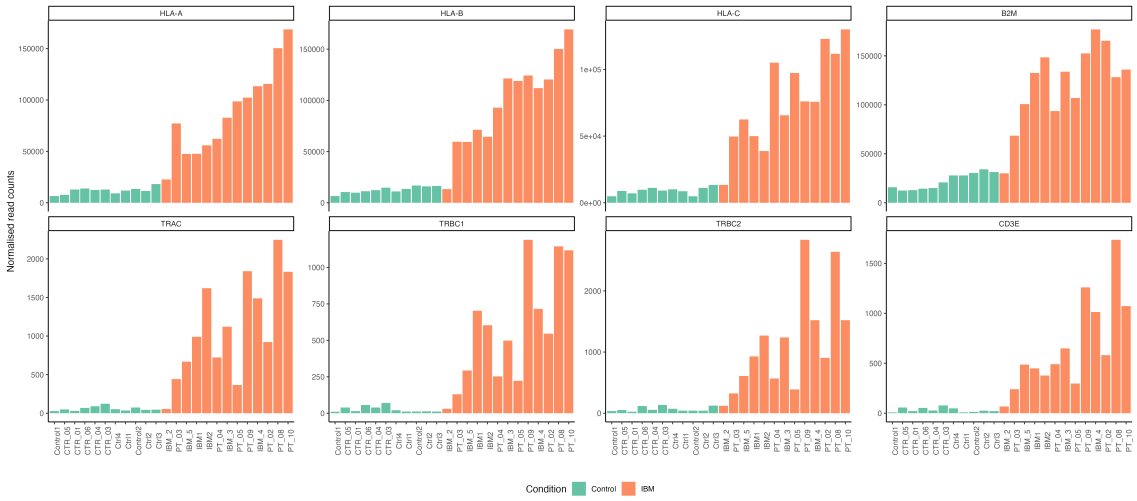


Figure 43: **Expression of T cells and MHC-I genes.** RNA-seq normalised read counts for TCR constant chains (*TRAC, TRBC1*, and *TRBC2*), *CD3E, HLA-A, HLA-B, HLA-C*, and (*B2M*) in controls (green) and sIBM cases (orange).

The immune response in IBM involves an increase in MHC expression by the muscle fibres and resident cells and is also associated with an increase in immune infiltrate. I first assessed antigen presentation genes, including HLA type I and *B2M*, and found a strong upregulation in sIBM cases compared to controls. Similarly, genes encoding TCR subunits and *CD3E*, which represent the infiltration by CD8-positive T cells, are more expressed in sIBM cases than in controls (Figure 43).

Using this gene set as a proxy for immune infiltration, I found that these measures correlated with the expression of TDP-43 CE peptides (Figure 44).

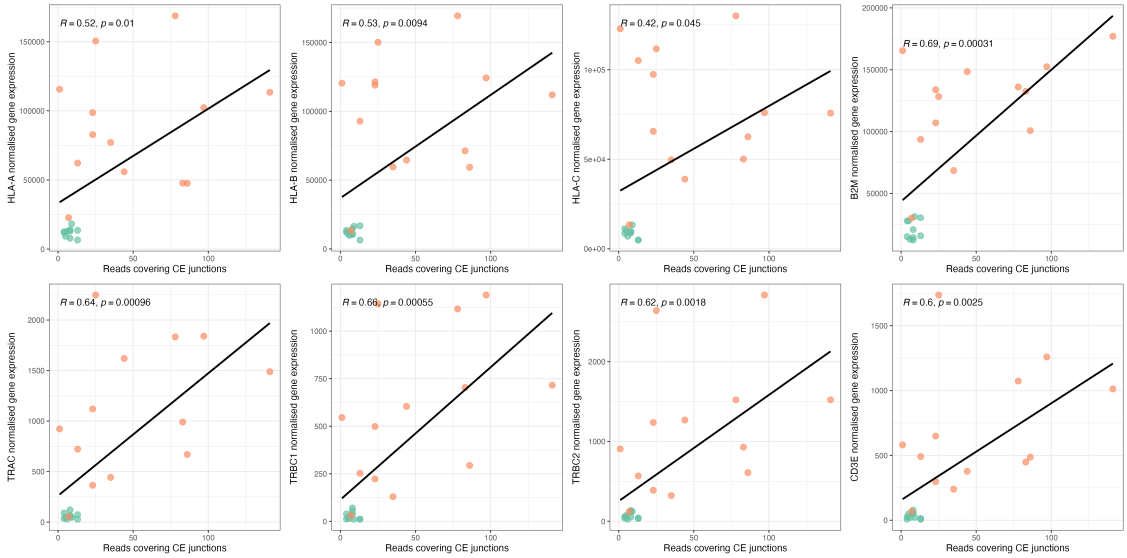


Figure 44: **Correlation of T cell and MHC-I genes to CE burden.** Correlation between RNA-seq normalised read counts for TCR constant chains (TRAC, TRBC1, and TRBC2), CD3E, HLA-A, -B, -C, and beta2-microglobulin (B2M) and CE raw counts in controls and sIBM cases.

In summary, my analysis of RNA-seq from IBM muscles shows that cryptic splicing - which can be considered a proxy for TDP-43 pathology - correlates with antigen presentation and immune infiltrates in sIBM.

HDGFL2 cryptic peptide and immune activation in sIBM

In order to further investigate the relationship between TDP-43-related cryptic splicing and immune infiltration, I used proteomics data combined with immunohistochemistry (Fig. 42 and Table X, and Appendix Table 6 and Appendix Figure 2). Similarly to my approach with transcriptomics data, I used the levels of specific proteins associated with immune infiltration and antigen presentation and investigated their relation with TDP-43 pathology, as assessed by immunohistochemistry (Figure 45). I subdivided the IHC results into high and low pathology based on the

expression of the HDGFL2 cryptic peptide. First, I verified that TDP-43 levels in the proteomics analysis correlated with the expression of the HDGFL2 cryptic peptide (Figure 45). Interestingly, proteomics data demonstrated that CD3E and TCR-B, two of the main protein components of the T cell receptor, were virtually absent from controls, and expressed only in a subset of sIBM cases with increased inclusion of the HDGFL2 cryptic peptide in the immunohistochemistry (Figure 45). I then investigated MHC-I and antigen-processing proteins and found that their expression levels were significantly higher in the group with higher HDGFL2 cryptic peptide expression (Figure 45).

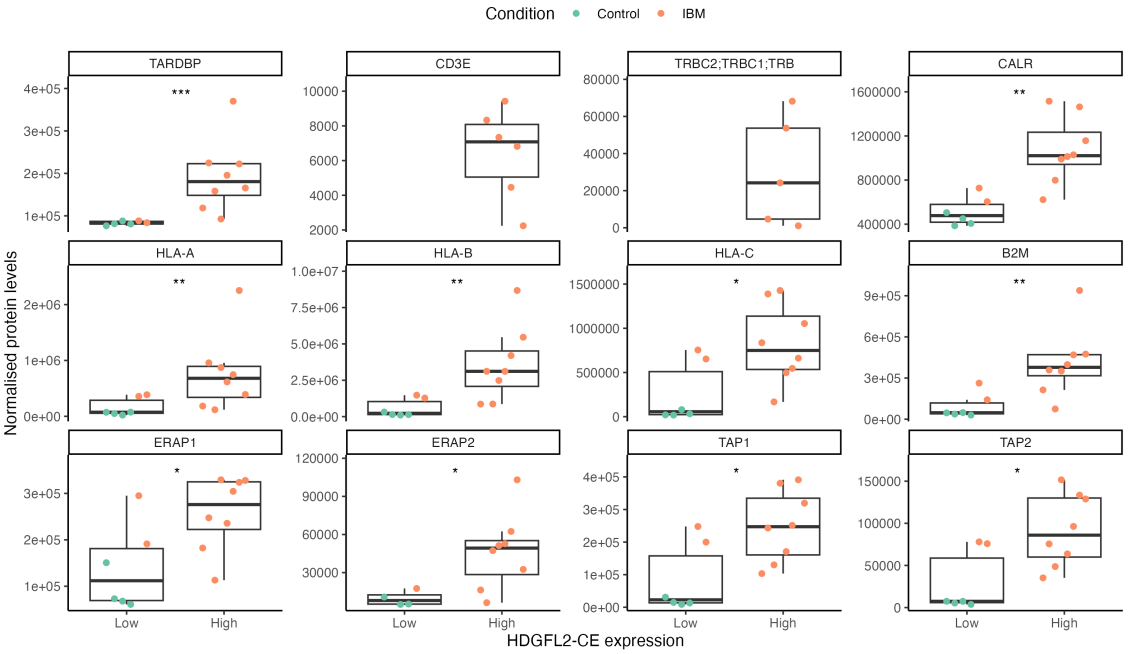


Figure 45: **Expression of HDGFL2 cryptic peptide is linked with increased immune response.** Proteomics normalised read counts for TCR constant chains (TRBC), CD3E, TDP-43, HLA-A, -B, -C, B2M, CALR, TAP1/2, and ERAP1/2 (pMHC processing pathway). Statistics and boxplots are between cases with high or low HDGFL2-CE staining in the immunohistochemistry, while colour is based on disease conditions (green for control, orange for IBM cases).

In summary, the combination of my proteomics and IHC analyses shows agreement with my transcriptomics investigations and demonstrates that immune infiltration correlates with a more cryptic peptide burden in human muscle tissue.

Impact of the HDGFL2 cryptic peptide on protein structure

The TDP-43-dependent cryptic exon in *HDGFL2* generates a 46 amino acid peptide within a 671 amino acid protein. To assess the potential impact on protein structure, I used AlphaFold, a bioinformatic tool to model the structure of proteins

based on their primary sequence (Jumper et al., 2021). By comparing the wild-type isoform with that including the CE peptide, my analysis shows that the inclusion of the CE peptide leads to the inclusion of a novel alpha-helix (Figure 46), raising the possibility that this modification can lead to neomorphic functions of HDGFL2.

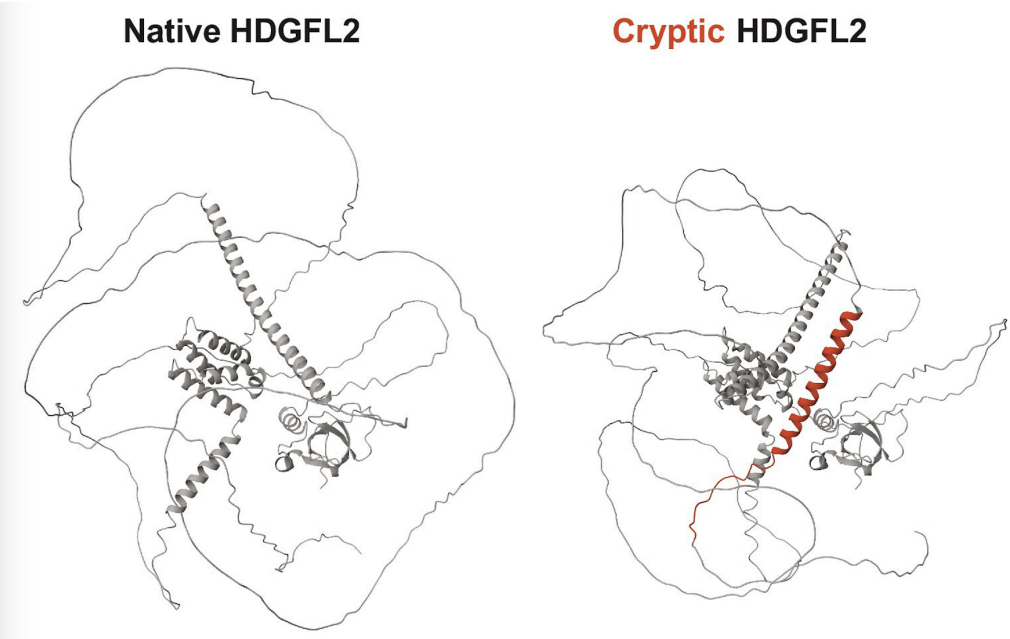


Figure 46: **AlphaFold2 prediction of HDGFL2 cryptic peptide inclusion.** Predicted structure of HDGFL2 using AlphaFold2, annotated with the predicted cryptic peptide (red) induced by TDP-43 depletion.

Modelling the TCR:pMHC interaction

The expression of novel peptides can be useful as a disease biomarker, but also raises the possibility of the immune system recognising these novel peptides and mounting a response. I worked in collaboration with an immunology group at UPenn to identify specific T cells and TCRs directed to cryptic peptides, and my collaborators were able to sequence and validate TCRs specifically directed to neoantigens derived from the HDGFL2 cryptic peptide.

Particularly, they employed a novel single-cell TCR sequencing method (K.-Y. Ma et al., 2021; S.-Q. Zhang et al., 2018), and, by incubating CD8+ T-cells with epitopes derived from cryptic peptides (Table IX) and viral epitopes as controls, they were able to identify specific TCR sequences directed towards TDP-43 CE peptides. They then validated two of them (one against the CE peptide in HDGFL2 and one against that in IGLON5) by overexpressing these TCRs in CD8+ T cells and showing their specific activation using the CD69 activation assay, an established tool in immunology research (Beeler et al., 2008).

Therefore, I used these TCRs and modelled their interaction with matched and unmatched pMHC complexes. I used TCRmodel2 (Yin et al., 2023), a tool derived from AlphaFold2 specifically designed for predicting the TCR:pMHC interaction. My analysis showed that the identified TCRs bind more specifically to matched pMHC than random TCRs, although the results are not significant (Fig. 47 and 48).

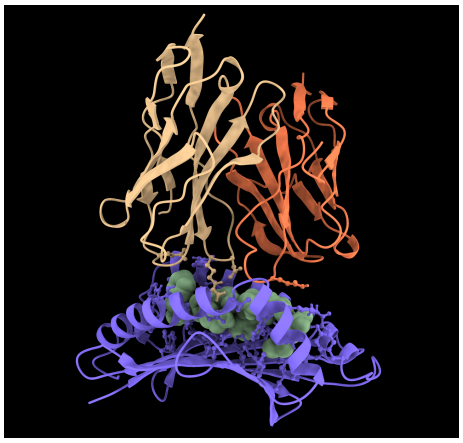


Figure 47: **Example of TCRmodel2 modelling of TCR:pMHC interactions.** The image shows the result of the modelling of an epitope from the cryptic peptide in HDGFL2 (green), the MHC molecule presenting it (purple), and the two subunits of the TCR (TCRalpha in pale yellow and TCRbeta in orange).

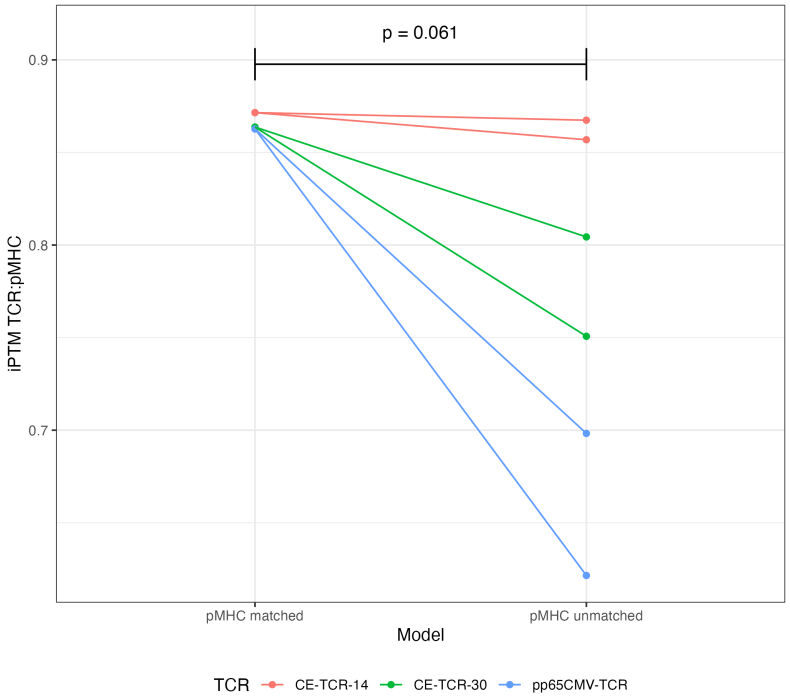


Figure 48: **Quantification of TCRmodel2 iPTMs.**
The point-and-line plot depicts the ipTM (pTM, corresponding to overall topological accuracy, calculated for interchain interfaces) score for TCR matched to their pMHC (left) and unmatched (right). Colours refer to the TCR analysed. Statistics is based on ANCOVA test.

These predictions, along with the data from my collaborators (not shown), support a direct role for cryptic peptides in T cell activation in IBM.

5.4 Discussion

The neuronal mislocalisation of TDP-43 from the nucleus to the cytoplasm is a feature of almost all cases of ALS (except genetic forms due to mutations in SOD1 and FUS), around 50% of FTD cases, and 50% of AD cases (Amador-Ortiz et al., 2007; Arai et al., 2006; Neumann et al., 2006). In addition to these, individuals from familial pedigrees with mutations in *hnRNPA1, hnRNPA2B1*, and *VCP* are affected by ALS, FTD, parkinsonism, IBM, and/or Paget's disease of the bone (Johnson et al., 2010; H. J. Kim et al., 2013; Watts et al., 2007). Therefore, the scientific community recognised a novel entity called multisystem proteinopathy (Evangelista et al., 2016). Moreover, TDP-43 pathology has been found in myocytes in several myopathies and muscular dystrophies (Table II). Despite the advances in the understanding of TDP-43 pathology in different tissues, little is known about the role of TDP-43 loss of function and cryptic splicing in non-neuronal cells and their role in the pathogenesis of these disorders.

Interestingly, sIBM pathogenesis includes inflammatory and degenerative features. Amongst the inflammatory changes in sIBM, myofibers display MHC-I upregulation and highly differentiated CD8+ T cells invade the tissue(Greenberg et al., 2019; Rothwell et al., 2017). With regards to the neurodegenerative features in sIBM, analyses on biopsies show amyloid-like fibrils (Askanas et al., 1994), p62-positive aggregates, and TDP-43 mislocalisation (Weihl et al., 2008). Previous work showed widespread RNA impairment in sIBM (Cortese et al., 2014), and a recent paper highlighted the presence of TDP-43 CEs in sIBM muscle, including that in *HDGFL2* (Britson et al., 2022), showing that cryptic splicing is a sensitive and specific biomarker of the disease. Furthermore, CE peptides have been identified in tissues affected by ALS, FTD, and AD, where TDP-43 pathology occurs in the CNS (Calliari et al., 2024; Irwin et al., 2024; Seddighi et al., 2024). However, a thorough analysis of CE peptides resulting from TDP-43 proteinopathy in sIBM has not been documented. Additionally, since these peptides are expressed only in disease conditions, they could be processed and presented as neoantigens and trigger immune responses, particularly from CD8+ T cells, further contributing to the disease pathogenesis.

To dissect the role of TDP-43 cryptic splicing in non-CNS tissue and its relationship with the immune infiltration in sIBM, I made use of several techniques, including immunohistochemistry, mass spectrometry-based proteomics, RNA sequencing, and

protein structural modelling.

First, by means of RNA-seq on human skeletal muscles, I showed how TDP-43 CEs encoding cryptic peptides are sensitive and specific markers of sIBM. Differences in basal gene expression are likely responsible for the missed detection of some CEs found in neuronal cell lines in the skeletal muscle.

I then performed immunohistochemistry on serial muscle sections to visualise in parallel the loss of TDP-43, the p62-positive aggregates, and the expression of the HDGFL2 cryptic peptide, as well as the extent of the immune infiltrate. I made use of a novel-generated antibody specific for the HDGFL2 cryptic peptide (Calliari et al., 2024; Seddighi et al., 2024) in muscle tissue. While these markers are not present in controls, they are visible in sIBM biopsies, even if there are differences between different affected individuals. Interestingly, I showed that HDGFL2 is not only specific for diseased fibres, but it also aggregates similarly to TDP-43 and p62/SQSTM1.

Later, by means of RNA-seq, I correlated the extent of cryptic splicing with markers of antigen presentation through the MHC-I pathway and of CD8+ T cells.

I confirmed these findings using mass spectrometry-based proteomics, and I found that the expression of the HDGFL2 cryptic peptide in immunohistochemistry can cluster a subset of sIBM cases with greater expression of CD8+ T cell and MHC-I pathway markers.

I eventually made use of AlphaFold-based modelling to visualise the impact of the HDGFL2 cryptic peptide on the protein sequence and to model the binding of antigens derived from this novel peptide to the TCR.

Overall, this chapter offers insights into the prevalence of TDP-43 cryptic peptides in sIBM muscle, which could lead to the development of markers of disease status and progression and response to therapy in this disease. Interestingly, I also found that diseased muscle tissues express *STMN2* CE, although which cell types are expressing it is still unknown. Indeed, muscle from subjects with sIBM exhibits increased numbers of satellite cells and regenerating myofibers in comparison to controls (Wanschitz et al., 2013). This is also interesting because another seminal paper examined the role of TDP-43 mislocalisation during muscle development and response to injury, suggesting that TDP-43 mislocalisation in disease could be the exacerbation of a physiological mechanism (Vogler et al., 2018).

Overall, my work is supportive of TDP-43 cryptic peptides being able to induce a CD8+ T cell response, but follow-up experiments will be needed to better understand and test this mechanism.

Specifically, spatial transcriptomics experiments will allow to look at the phenotype of the T cells infiltrating the tissue together with the extent of TDP-43 cryptic splicing. This technique can also be performed in the spinal cord of patients affected by ALS, enabling us to address the phenotype of the immune infiltrates in ALS, since recent works found the presence of intrathecal CD8+ T cells in AD and ALS (Campisi et al., 2022; Gate et al., 2020).

It will also be useful to characterise the T cell phenotype and HLA genotyping, as well as other immune genes linked to TDP-43 proteinopathies, such as IL18RAP (Eitan et al., 2022), in people affected by TDP-43 proteinopathies, by analysis of the PBMCs of affected individuals.

Beyond the TDP-immune link, future experiments will also allow testing the potential of detecting cryptic peptides in biofluids by means of ELISA-based techniques, similarly to what has been done in ALS individuals (Irwin et al., 2024), to be used as biomarkers for sIBM. Lastly, although detailed structures for CNS TDP proteinopathies are emerging, this still has not been revealed for muscle and cryoEM work on patient tissue (Arseni et al., 2022, 2023, 2024) will be crucial in allowing us to better understand the similarities and differences between CNS and muscle TDP-43 proteinopathies.

Chapter 6

Conclusions

In this thesis, I've assessed molecular and genetic aspects related to the clinical entities ALS, FTD, sIBM, and REDs, the latter including HD, SCAs, and SBMA. These diseases fall under the umbrella of neurodegenerative diseases, for which hallmarks have been described (1).

Alterations in DNA repair and RNA processing are an increasingly recognised mechanism across neurodegeneration. Mutations in *SETX* and the increasingly recognised role for repeat expansions highlight the importance of DNA repair mechanisms in neurodegenerative diseases. Alterations in RBPs in ALS-FTD and their direct consequences on splicing are one example of RNA processing defects that have emerged over the last years (Brown et al., 2022; Klim et al., 2019; X. R. Ma et al., 2022; Melamed et al., 2019; Prudencio et al., 2020).

Other hallmarks of neurodegenerative diseases are altered proteostasis and the presence of protein aggregates. Each neurodegenerative condition is characterised by aggregates of one or more proteins (e.g. tau, TDP-43, alpha-synuclein, DPRs). Furthermore, mutations in proteins responsible for protein homeostasis, such as VCP and UBQLN2 (Deng et al., 2011; Johnson et al., 2010), highlight the direct role of protein clearance in the neurodegenerative cascade. Although neurodegenerative conditions are clinically and pathologically distinct, novel commonalities are emerging, and the accumulation of TDP-43 has been increasingly described to be present across diseases, and has been described in ALS, sIBM, FTD, AD and HD (Arai et al., 2006, 2009; Neumann et al., 2006; Schwab et al., 2008; Weihl et al., 2008)

Lastly, the role of the immune system has been increasingly recognised in NDDs. GWAS results have directly highlighted the importance of immune pathways in AD and PD (Andrews et al., 2023; J. J. Kim et al., 2024), and there has been increasing evidence and clinical trials targeting these mechanisms across conditions (e.g. NCT01409915, NCT01882010, NCT03039673).

My work has touched upon the above categories of events. Other common themes across neurodegeneration are also of relevance, but have not been touched upon by my work. These include cytoskeletal abnormalities, including axonal transport, synaptic and neuronal network dysfunction, and altered energy metabolism.

In the first chapter, my work explored the variability of repeat expansions across neurodegenerative conditions. By leveraging novel bioinformatic tools on large whole-genome sequencing datasets, I uncover a surprisingly high occurrence of STR expansions in the general population and then estimate the prevalence and instability features of STR diseases by mathematical modelling.

In the second chapter, I focused on the RNA processing consequences of TDP-43, by generating novel *in cellulo* models of TDP-43 loss of function and analysing them through bioinformatic tools. Although many of these events have been amply described by the work of many, there are still a lot of outstanding questions. Specifically, the fate of the misprocessed RNA and its degradation and stability are still unknown on a large scale. My work addresses specifically which transcripts undergo NMD - I find that the majority indeed is cleared by NMD, but there is also a substantial proportion that escapes this surveillance pathway. Another important outstanding question is related to the progression of TDP-43 aggregation and loss of function: TDP-43 pathology is not an all-or-nothing switch and occurs gradually: how does this impact the multitude of TDP-43-dependent RNA processing events? My work recreated *in vitro* numerous levels of TDP-43 loss of function and used comprehensive RNA-sequencing and analyses to tease out the early events that respond to even mild TDP-43 loss of function and events that instead require near-complete loss of TDP-43 to occur. This compendium allows us to better understand TDP-43 biology and will also be helpful in prioritising and understanding the relevance of the different changes in relation to biomarker potential and pathogenicity.

My third chapter touches on both the altered proteostasis and inflammation themes in neurodegeneration. Previous work had described how a minority of missplicing events are able to lead to the translation of novel cryptic peptides (Seddighi et al., 2024). I used a novel antibody generated towards one of these peptides to characterise its behaviour in sIBM muscles. My findings highlight how these cryptic peptides can themselves aggregate and therefore contribute to altered proteostasis in a possible feedforward cycle. Furthermore, I went on to show that altered splicing and immune infiltration correlate in diseased muscle and carried out work to start exploring the immunogenic potential of these peptides.

There are a number of strengths in the work I carried out. Firstly, the combination of dry lab and wet lab skills I was able to acquire allowed me to comprehensively address TDP-43-related changes, first in the context of a pure TDP-43 loss of function and then in disease settings, where TDP-43 pathology is also featured. I was able to take advantage of and combine data from very well-defined cell models with large patient datasets. This allowed me to dissect unique molecular mechanisms and to assess the effect of these mechanisms in the context of human disease. Furthermore, I was able to assess the relevance of TDP-43 beyond the most canonically studied CNS, where I validated the findings from the *in cellulo* models by targeted RNA-seq, but I was also able to investigate TDP-43-related changes in skeletal muscle.

Lastly, my work also has a direct impact on the interpretation of genetic data and genetic counselling. The increasing understanding of the frequency and distribution of repeat expansions changes the estimate of their penetrance and affects the interpretation of these findings in clinical settings.

Overall, this work advances our understanding of TDP-43 CE biology and identifies novel disease effectors and potential biomarkers for TDP-43 proteinopathies. Moreover, it sheds light on the prevalence of STR and related diseases, with important implications for genetic and clinical medicine.

Bibliography

Abramzon, Y., Fratta, P., Traynor, B. J., Chia, R., & Conforti, F. L. (2020). The Overlapping Genetics of Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Frontiers in Neuroscience*, *14*(42), 1–10. https://doi.org/10.3389/fnins.2020.00042

Afroz, T., Hock, E.-M., Ernst, P., Foglieni, C., Jambeau, M., Gilhespy, L. A. B., Laferriere, F., Maniecka, Z., Plückthun, A., Mittl, P., Paganetti, P., Allain, F. H. T., & Polymenidou, M. (2017). Functional and dynamic polymerization of the ALS-linked protein TDP-43 antagonizes its pathologic aggregation. *Nature Communications*, *8*(1), 45. https://doi.org/10.1038/s41467-017-00062-0

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2015). *Molecular biology of the cell* (6th). Garland Science, Taylor & Francis Group.

Al-Chalabi, A., Calvo, A., Chiò, A., Colville, S., Ellis, C. M., Hardiman, O., Heverin, M., Howard, R. S., Huisman, M. H., Keren, N., Leigh, P. N., Mazzini, L., Mora, G., Orrell, R. W., Rooney, J., Scott, K. M., Scotton, W. J., Seelen, M., Shaw, C. E., … Pearce, N. (2014). Analysis of amyotrophic lateral sclerosis as a multistep process: A population-based modelling study [Publisher: Elsevier Ltd]. *The Lancet Neurology*, *13*(11), 1108–1113. https://doi.org/10.1016/S1474-4422(14)70219-4

Al-Chalabi, A., & Hardiman, O. (2013). The epidemiology of ALS: A conspiracy of genes, environment and time [Publisher: Nature Publishing Group]. *Nature Reviews Neurology*, *9*(11), 617–628. https://doi.org/10.1038/nrneurol.2013.203

Amador-Ortiz, C., Lin, W.-L., Ahmed, Z., Personett, D., Davies, P., Duara, R., Graff-Radford, N. R., Hutton, M. L., & Dickson, D. W. (2007). TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer's disease. *Annals of Neurology*, *61*(5), 435–445. https://doi.org/10.1002/ana.21154

Andrews, S. J., Renton, A. E., Fulton-Howard, B., Podlesny-Drabiniok, A., Marcora, E., & Goate, A. M. (2023). The complex genetic architecture of Alzheimer's disease: Novel insights and future directions. *eBioMedicine*, *90*, 104511. https://doi.org/10.1016/j.ebiom.2023.104511

Antonioni, A., Raho, E. M., Lopriore, P., Pace, A. P., Latino, R. R., Assogna, M., Mancuso, M., Gragnaniello, D., Granieri, E., Pugliatti, M., Di Lorenzo, F., & Koch, G. (2023). Frontotemporal Dementia, Where Do We Stand? A Narrative Review. *International Journal of Molecular Sciences*, *24*(14), 11732. https://doi.org/10.3390/ijms241411732

Appocher, C., Mohagheghi, F., Cappelli, S., Stuani, C., Romano, M., Feiguin, F., & Buratti, E. (2017). Major hnRNP proteins act as general TDP-43 functional modifiers both in Drosophila and human neuronal cells. *Nucleic Acids Research*, *45*(13), 8026–8045. https://doi.org/10.1093/nar/gkx477

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y., & Oda, T. (2006). TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochemical and Biophysical Research Communications*, *351*(3), 602–611. https://doi.org/10.1016/j.bbrc.2006.10.093

Arai, T., Mackenzie, I. R. A., Hasegawa, M., Nonoka, T., Niizato, K., Tsuchiya, K., Iritani, S., Onaya, M., & Akiyama, H. (2009). Phosphorylated TDP-43 in Alzheimer's disease and dementia with Lewy bodies. *Acta Neuropathologica*, *117*(2), 125–136. https://doi.org/10.1007/s00401-008-0480-1

Arnold, F. J., Cui, Y., Michels, S., Colwin, M. R., Stockford, C. M., Ye, W., Maheswari Jawahar, V., Jansen-West, K., Philippe, J., Gulia, R., Gou, Y., Tam, O. H., Menon, S., Situ, W. G., Cazarez, S. L., Zandi, A., Ehsani, K. C., Howard, S., Dickson, D. W., ... La Spada, A. R. (2025). TDP-43 dysregulation of polyadenylation site selection is a defining feature of RNA misprocessing in amyotrophic lateral sclerosis and frontotemporal dementia. *Journal of Clinical Investigation*, *135*(11), e182088. https://doi.org/10.1172/JCI182088

Arseni, D., Chen, R., Murzin, A. G., Peak-Chew, S. Y., Garringer, H. J., Newell, K. L., Kametani, F., Robinson, A. C., Vidal, R., Ghetti, B., Hasegawa, M., & Ryskeldi-Falcon, B. (2023). TDP-43 forms amyloid filaments with a distinct fold in type A FTLD-TDP. *Nature*, *620*(7975), 898–903. https://doi.org/10.1038/s41586-023-06405-w

Arseni, D., Hasegawa, M., Murzin, A. G., Kametani, F., Arai, M., Yoshida, M., & Ryskeldi-Falcon, B. (2022). Structure of pathological TDP-43 filaments from ALS with FTLD. *Nature*, *601*(7891), 139–143. https://doi.org/10.1038/s41586-021-04199-3

Arseni, D., Nonaka, T., Jacobsen, M. H., Murzin, A. G., Cracco, L., Peak-Chew, S. Y., Garringer, H. J., Kawakami, I., Suzuki, H., Onaya, M., Saito, Y., Murayama, S., Geula, C., Vidal, R., Newell, K. L., Mesulam, M., Ghetti, B., Hasegawa, M., & Ryskeldi-Falcon, B. (2024). Heteromeric amyloid filaments of ANXA11 and TDP-43 in FTLD-TDP type C. *Nature*, *634*(8034), 662–668. https://doi.org/10.1038/s41586-024-08024-5

Arthur, K. C., Calvo, A., Price, T. R., Geiger, J. T., Chiò, A., & Traynor, B. J. (2016). Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. *Nature Communications*, *7*, 1–6. https://doi.org/10.1038/ncomms12408

Askanas, V., Engel, W. K., Bilak, M., Alvarez, R. B., & Selkoe, D. J. (1994). Twisted tubulofilaments of inclusion body myositis muscle resemble paired helical filaments of Alzheimer brain and contain hyperphosphorylated tau. *The American Journal of Pathology*, *144*(1), 177–187.

Attali, R., Warwar, N., Israel, A., Gurt, I., McNally, E., Puckelwartz, M., Glick, B., Nevo, Y., Ben-Neriah, Z., & Melki, J. (2009). Mutation of SYNE-1, encoding an essential component of the nuclear lamina, is responsible for autosomal recessive arthrogryposis. *Human Molecular Genetics*, *18*(18), 3462–3469. https://doi.org/10.1093/hmg/ddp290

Baine, F. K., Peerbhai, N., & Krause, A. (2018). A study of Huntington disease-like syndromes in black South African patients reveals a single SCA2 mutation and a unique distribution of normal alleles across five repeat loci. *Journal of the Neurological Sciences*, *390*, 200–204. https://doi.org/10.1016/j.jns.2018.04.031

Baker, M., Mackenzie, I. R., Pickering-Brown, S. M., Gass, J., Rademakers, R., Lindholm, C., Snowden, J., Adamson, J., Sadovnick, A. D., Rollinson, S., Cannon, A., Dwosh, E., Neary, D., Melquist, S., Richardson, A., Dickson, D., Berger, Z., Eriksen, J., Robinson, T., ... Hutton, M. (2006). Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature*, *442*(7105), 916–919. https://doi.org/10.1038/nature05016

Bampton, A., Gatt, A., Humphrey, J., Cappelli, S., Bhattacharya, D., Foti, S., Brown, A.-L., Asi, Y., Low, Y. H., Foiani, M., Raj, T., Buratti, E., Fratta, P., & Lashley, T. (2021). HnRNP K mislocalisation is a novel protein pathology of frontotemporal lobar degeneration and ageing and leads to cryptic splicing. *Acta Neuropathologica*, *142*(4), 609–627. https://doi.org/10.1007/s00401-021-02340-0

Bampton, A., Gittings, L. M., Fratta, P., Lashley, T., & Gatt, A. (2020). The role of hnRNPs in frontotemporal dementia and amyotrophic lateral sclerosis. *Acta Neuropathologica*, *140*(5), 599–623. https://doi.org/10.1007/s00401-020-02203-0

Bannwarth, S., Ait-El-Mkadem, S., Chaussenot, A., Genin, E. C., Lacas-Gervais, S., Fragaki, K., Berg-Alonso, L., Kageyama, Y., Serre, V., Moore, D. G., Verschueren, A., Rouzier, C., Le Ber, I., Augé, G., Cochaud, C., Lespinasse, F., N'guyen, K., De Septenville, A., Brice, A., ... Paquis-Flucklinger, V. (2014). A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *Brain*, *137*(8), 2329–2345. https://doi.org/10.1093/brain/awu138

Baughn, M. W., Melamed, Z., López-Erauskin, J., Beccari, M. S., Ling, K., Zuberi, A., Presa, M., Gonzalo-Gil, E., Maimon, R., Vazquez-Sanchez, S., Chaturvedi,

S., Bravo-Hernández, M., Taupin, V., Moore, S., Artates, J. W., Acks, E., Ndayambaje, I. S., Agra De Almeida Quadros, A. R., Jafar-nejad, P., ... Cleveland, D. W. (2023). Mechanism of *STMN2* cryptic splice-polyadenylation and its correction for TDP-43 proteinopathies. *Science*, *379*(6637), 1140–1149. https://doi.org/10.1126/science.abq5622

Beeler, A., Zaccaria, L., Kawabata, T., Gerber, B. O., & Pichler, W. J. (2008). CD69 upregulation on T cells as an *in vitro* marker for delayed-type drug hypersensitivity. *Allergy*, *63*(2), 181–188. https://doi.org/10.1111/j.1398-9995.2007.01516.x

Begemann, A., Sticht, H., Begtrup, A., Vitobello, A., Faivre, L., Banka, S., Alhaddad, B., Asadollahi, R., Becker, J., Bierhals, T., Brown, K. E., Bruel, A.-L., Brunet, T., Carneiro, M., Cremer, K., Day, R., Denommé-Pichon, A.-S., Dyment, D. A., Engels, H., ... Rauch, A. (2021). New insights into the clinical and molecular spectrum of the novel CYFIP2-related neurodevelopmental disorder and impairment of the WRC-mediated actin dynamics. *Genetics in Medicine*, *23*(3), 543–554. https://doi.org/10.1038/s41436-020-01011-x

Bento-Abreu, A., Jager, G., Swinnen, B., Rué, L., Hendrickx, S., Jones, A. R., Staats, K. A., Taes, I., Eykens, C., Nonneman, A., Nuyts, R., Timmers, M., Silva, L., Chariot, A., Nguyen, L., Ravits, J., Lemmens, R., Cabooter, D., Van Den Bosch, L., ... Robberecht, W. (2018). Elongator subunit 3 (ELP3) modifies ALS through tRNA modification. *Human Molecular Genetics*, *27*(7), 1276–1289. https://doi.org/10.1093/hmg/ddy043

Benveniste, O., Guiguet, M., Freebody, J., Dubourg, O., Squier, W., Maisonobe, T., Stojkovic, T., Leite, M. I., Allenbach, Y., Herson, S., Brady, S., Eymard, B., & Hilton-Jones, D. (2011). Long-term observational study of sporadic inclusion body myositis. *Brain*, *134*(11), 3176–3184. https://doi.org/10.1093/brain/awr213

Benveniste, O., Stenzel, W., Hilton-Jones, D., Sandri, M., Boyer, O., & Van Engelen, B. G. M. (2015). Amyloid deposits and inflammatory infiltrates in sporadic inclusion body myositis: The inflammatory egg comes before the degenerative chicken. *Acta Neuropathologica*, *129*(5), 611–624. https://doi.org/10.1007/s00401-015-1384-5

Bertolin, C., Querin, G., Martinelli, I., Pennuto, M., Pegoraro, E., & Sorarù, G. (2019). Insights into the genetic epidemiology of spinal and bulbar muscular atrophy: Prevalence estimation and multiple founder haplotypes in the Veneto Italian region. *European Journal of Neurology*, *26*(3), 519–524. https://doi.org/10.1111/ene.13850

Boeynaems, S., Bogaert, E., Kovacs, D., Konijnenberg, A., Timmerman, E., Volkov, A., Guharoy, M., De Decker, M., Jaspers, T., Ryan, V. H., Janke, A. M., Baatsen, P., Vercruysse, T., Kolaitis, R. M., Daelemans, D., Taylor, J. P., Kedersha, N., Anderson, P., Impens, F., ... Van Den Bosch, L. (2017).

Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics [Publisher: Elsevier Inc.]. *Molecular Cell*, *65*(6), 1044–1055.e5. https://doi.org/10.1016/j.molcel.2017.02.013

Braak, H., Brettschneider, J., Ludolph, A. C., Lee, V. M., Trojanowski, J. Q., & Tredici, K. D. (2013). Amyotrophic lateral sclerosis—a model of corticofugal axonal spread. *Nature Reviews Neurology*, *9*(12), 708–714. https://doi.org/10.1038/nrneurol.2013.221

Brenner, D., Müller, K., Wieland, T., Weydt, P., Böhm, S., Lule, D., Hübers, A., Neuwirth, C., Weber, M., Borck, G., Wahlqvist, M., Danzer, K. M., Volk, A. E., Meitinger, T., Strom, T. M., Otto, M., Kassubek, J., Ludolph, A. C., Andersen, P. M., & Weishaupt, J. H. (2016). NEK1 mutations in familial amyotrophic lateral sclerosis. *Brain*, *139*(5), e28–e28. https://doi.org/10.1093/brain/aww033

Britson, K. A., Ling, J. P., Braunstein, K. E., Montagne, J. M., Kastenschmidt, J. M., Wilson, A., Ikenaga, C., Tsao, W., Pinal-Fernandez, I., Russell, K. A., Reed, N., Mozaffar, T., Wagner, K. R., Ostrow, L. W., Corse, A. M., Mammen, A. L., Villalta, S. A., Larman, H. B., Wong, P. C., & Lloyd, T. E. (2022). Loss of TDP-43 function and rimmed vacuoles persist after T cell depletion in a xenograft model of sporadic inclusion body myositis. *Science Translational Medicine*, *14*(628), eabi9196. https://doi.org/10.1126/scitranslmed.abi9196

Brown, A.-L., Wilkins, O. G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C., Bampton, A., Lee, F. C. Y., Masino, L., Qi, Y. A., Bryce-Smith, S., Gatt, A., Hallegger, M., Fagegaltier, D., Phatnani, H., NYGC ALS Consortium, Phatnani, H., Kwan, J., Sareen, D., ... Fratta, P. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature*, *603*(7899), 131–137. https://doi.org/10.1038/s41586-022-04436-3

Bryce-Smith, S., Brown, A.-L., Mehta, P. R., Mattedi, F., Mikheenko, A., Barattucci, S., Zanovello, M., Dattilo, D., Yome, M., Hill, S. E., Qi, Y. A., Wilkins, O. G., Sun, K., Ryadnov, E., Wan, Y., NYGC ALS Consortium, Vargas, J. N. S., Birsa, N., Raj, T., ... Fratta, P. (2024, January). TDP-43 loss induces extensive cryptic polyadenylation in ALS/FTD. https://doi.org/10.1101/2024.01.22.576625

Buratti, E., & Baralle, F. E. (2001). Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator ofCFTR Exon 9. *Journal of Biological Chemistry*, *276*(39), 36337–36343. https://doi.org/10.1074/jbc.M104236200

Burrell, J. R., Halliday, G. M., Kril, J. J., Ittner, L. M., Götz, J., Kiernan, M. C., & Hodges, J. R. (2016). The frontotemporal dementia-motor neuron disease continuum. *The Lancet*, *388*(10047), 919–931. https://doi.org/10.1016/S0140-6736(16)00737-6

Burrell, J. R., Kiernan, M. C., Vucic, S., & Hodges, J. R. (2011). Motor Neuron dysfunction in frontotemporal dementia. *Brain*, *134*(9), 2582–2594. https://doi.org/10.1093/brain/awr195

Calliari, A., Daughrity, L. M., Albagli, E. A., Castellanos Otero, P., Yue, M., Jansen-West, K., Islam, N. N., Caulfield, T., Rawlinson, B., DeTure, M., Cook, C., Graff-Radford, N. R., Day, G. S., Boeve, B. F., Knopman, D. S., Petersen, R. C., Josephs, K. A., Oskarsson, B., Gitler, A. D., ... Petrucelli, L. (2024). HDGFL2 cryptic proteins report presence of TDP-43 pathology in neurodegenerative diseases. *Molecular Neurodegeneration*, *19*(1), 29. https://doi.org/10.1186/s13024-024-00718-8

Campisi, L., Chizari, S., Ho, J. S. Y., Gromova, A., Arnold, F. J., Mosca, L., Mei, X., Fstkchyan, Y., Torre, D., Beharry, C., Garcia-Forn, M., Jiménez-Alcázar, M., Korobeynikov, V. A., Prazich, J., Fayad, Z. A., Seldin, M. M., De Rubeis, S., Bennett, C. L., Ostrow, L. W., ... Marazzi, I. (2022). Clonally expanded CD8 T cells characterize amyotrophic lateral sclerosis-4. *Nature*, *606*(7916), 945–952. https://doi.org/10.1038/s41586-022-04844-5

Carrard, J., Ratajczak, F., Elsens, J., Leroy, C., Kong, R., Geoffroy, L., Comte, A., Fournet, G., Joseph, B., Li, X., Moebs-Sanchez, S., & Lejeune, F. (2023). Identifying Potent Nonsense-Mediated mRNA Decay Inhibitors with a Novel Screening System. *Biomedicines*, *11*(10), 2801. https://doi.org/10.3390/biomedicines11102801

Catalán, M., Selva-O'Callaghan, A., & Grau, J. (2014). Diagnosis and classification of sporadic inclusion body myositis (sIBM). *Autoimmunity Reviews*, *13*(4-5), 363–366. https://doi.org/10.1016/j.autrev.2014.01.016

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, Y. Z., Bennett, C. L., Huynh, H. M., Blair, I. P., Puls, I., Irobi, J., Dierick, I., Abel, A., Kennerson, M. L., Rabin, B. A., Nicholson, G. A., Auer-Grumbach, M., Wagner, K., De Jonghe, P., Griffin, J. W., Fischbeck, K. H., Timmerman, V., Cornblath, D. R., & Chance, P. F. (2004). DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *American Journal of Human Genetics*, *74*(6), 1128–1135. https://doi.org/10.1086/421054

Chiò, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., Traynor, B. G., & On Behalf Of The Eurals Consortium. (2009). Prognostic factors in ALS: A critical review. *Amyotrophic Lateral Sclerosis*, *10*(5-6), 310–323. https://doi.org/10.3109/17482960802566824

Chiò, A., Traynor, B. J., Collins, J., Simeone, J., Goldstein, L., & White, L. (2013). Global Epidemiology of Amyotrophic Lateral Sclerosis: A Systematic Review of the Published Literature. *Neuroepidemiology*, *41*(2), 118–130. https://doi.org/10.1159/000351153

Chow, C. Y., Landers, J. E., Bergren, S. K., Sapp, P. C., Grant, A. E., Jones, J. M., Everett, L., Lenk, G. M., McKenna-Yasek, D., Weisman, L. S., Figlewicz, D. A., Brown, R. H., & Meisler, M. H. (2009). Deleterious Variants of FIG4, a Phosphoinositide Phosphatase, in Patients with ALS. *American Journal of Human Genetics*, *84*(1), 85–88. https://doi.org/10.1016/j.ajhg.2008.12.010

Cirulli, E. T., Lasseigne, B. N., Petrovski, S., Sapp, P. C., Dion, P. A., Leblond, C. S., Couthouis, J., Lu, Y. F., Wang, Q., Krueger, B. J., Ren, Z., Keebler, J., Han, Y., Levy, S. E., Boone, B. E., Wimbish, J. R., Waite, L. L., Jones, A. L., Carulli, J. P., ... Munoz-Blanco, J. L. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*, *347*(6229), 1436–1441. https://doi.org/10.1126/science.aaa3650

Conicella, A. E., Zerze, G. H., Mittal, J., & Fawzi, N. L. (2016). ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure*, *24*(9), 1537–1549. https://doi.org/10.1016/j.str.2016.07.007

Cooper-Knock, J., Demaegd, K., Kernan, A., Vugt, J. V., Harvey, C., Moll, T., O'Brien, D., Gornall, S., Drury, L., Farhan, S., Dion, P., Rouleau, G., Western, A., Parsons, P., Mclean, B., Benatar, M., Berg, L. V. D., Damme, P. V., Dankbaar, J. W., ... Es, M. V. (2024, December). An observational study of pleiotropy and penetrance of amyotrophic lateral sclerosis associated with CAG-repeat expansion of ATXN2. https://doi.org/10.21203/rs.3.rs-5419198/v1

Cooper-Knock, J., Walsh, M. J., Higginbottom, A., Highley, J. R., Dickman, M. J., Edbauer, D., Ince, P. G., Wharton, S. B., Wilson, S. A., Kirby, J., Hautbergue, G. M., & Shaw, P. J. (2014). Sequestration of multiple RNA recognition motif-containing proteins by C9orf72 repeat expansions. *Brain*, *137*(7), 2040–2051. https://doi.org/10.1093/brain/awu120

Cortese, A., Laurà, M., Casali, C., Nishino, I., Hayashi, Y. K., Magri, S., Taroni, F., Stuani, C., Saveri, P., Moggio, M., Ripolone, M., Prelle, A., Pisciotta, C., Sagnelli, A., Pichiecchio, A., Reilly, M. M., Buratti, E., & Pareyson, D. (2018). Altered TDP-43-dependent splicing in *HSPB8* -related distal hereditary motor neuropathy and myofibrillar myopathy. *European Journal of Neurology*, *25*(1), 154–163. https://doi.org/10.1111/ene.13478

Cortese, A., Plagnol, V., Brady, S., Simone, R., Lashley, T., Acevedo-Arozena, A., de Silva, R., Greensmith, L., Holton, J., Hanna, M. G., Fisher, E. M., & Fratta, P. (2014). Widespread RNA metabolism impairment in sporadic

inclusion body myositis TDP43-proteinopathy. *Neurobiology of Aging*, *35*(6), 1491–1498. https://doi.org/10.1016/j.neurobiolaging.2013.12.029

Cotto, K. C., Feng, Y.-Y., Ramu, A., Richters, M., Freshour, S. L., Skidmore, Z. L., Xia, H., McMichael, J. F., Kunisaki, J., Campbell, K. M., Chen, T. H.-P., Rozycki, E. B., Adkins, D., Devarakonda, S., Sankararaman, S., Lin, Y., Chapman, W. C., Maher, C. A., Arora, V., … Griffith, M. (2023). Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nature Communications*, *14*(1), 1589. https://doi.org/10.1038/s41467-023-37266-6

Couthouis, J., Hart, M. P., Erion, R., King, O. D., Diaz, Z., Nakaya, T., Ibrahim, F., Kim, H. J., Mojsilovic-petrovic, J., Panossian, S., Kim, C. E., Frackelton, E. C., Solski, J. A., Williams, K. L., Clay-falcone, D., Elman, L., McCluskey, L., Greene, R., Hakonarson, H., … Gitler, A. D. (2012). Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis. *Human Molecular Genetics*, *21*(13), 2899–2911. https://doi.org/10.1093/hmg/dds116

Coyle-Gilchrist, I. T., Dick, K. M., Patterson, K., Vázquez Rodríquez, P., Wehmann, E., Wilcox, A., Lansdall, C. J., Dawson, K. E., Wiggins, J., Mead, S., Brayne, C., & Rowe, J. B. (2016). Prevalence, characteristics, and survival of frontotemporal lobar degeneration syndromes. *Neurology*, *86*(18), 1736–1743. https://doi.org/10.1212/WNL.0000000000002638

Cykowski, M. D., Powell, S. Z., Peterson, L. E., Appel, J. W., Rivera, A. L., Takei, H., Chang, E., & Appel, S. H. (2017). Clinical Significance of TDP-43 Neuropathology in Amyotrophic Lateral Sclerosis. *Journal of Neuropathology & Experimental Neurology*, *76*(5), 402–413. https://doi.org/10.1093/jnen/nlx025

De Castilhos, R. M., Furtado, G. V., Gheno, T. C., Schaeffer, P., Russo, A., Barsottini, O., Pedroso, J. L., Salarini, D. Z., Vargas, F. R., Lima, M. A. D. F. D. D., Godeiro, C., Santana-da-Silva, L. C., Toralles, M. B. P., Santos, S., Van Der Linden, H., Wanderley, H. Y., De Medeiros, P. F. V., Pereira, E. T., Ribeiro, E., … Jardim, L. B. (2014). Spinocerebellar Ataxias in Brazil—Frequencies and Modulating Effects of Related Genes. *The Cerebellum*, *13*(1), 17–28. https://doi.org/10.1007/s12311-013-0510-y

De Mattei, F., Ferrandes, F., Gallone, S., Canosa, A., Calvo, A., Chiò, A., & Vasta, R. (2023). Epidemiology of Spinocerebellar Ataxias in Europe. *The Cerebellum*, *23*(3), 1176–1183. https://doi.org/10.1007/s12311-023-01600-x

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G.-Y. R., Karydas, A., Seeley, W. W., Josephs, K. A., Coppola, G., Geschwind, D. H., … Rademakers, R. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, *72*(2), 245–256. https://doi.org/10.1016/j.neuron.2011.09.011

Deng, H. X., Chen, W., Hong, S. T., Boycott, K. M., Gorrie, G. H., Siddique, N., Yang, Y., Fecto, F., Shi, Y., Zhai, H., Jiang, H., Hirano, M., Rampersaud, E., Jansen, G. H., Donkervoort, S., Bigio, E. H., Brooks, B. R., Ajroud, K., Sufit, R. L., ... Siddique, T. (2011). Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia [Publisher: Nature Publishing Group]. *Nature*, *477*(7363), 211–215. https://doi.org/10.1038/nature10353

Depienne, C., & Mandel, J.-L. (2021). 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics*, *108*(5), 764–785. https://doi.org/10.1016/j.ajhg.2021.03.011

Dhanwani, R., Pham, J., Premlal, A. L. R., Frazier, A., Kumar, A., Pero, M. E., Bartolini, F., Dutra, J. R., Marder, K. S., Peters, B., Sulzer, D., Sette, A., & Lindestam Arlehamn, C. S. (2020). T Cell Responses to Neural Autoantigens Are Similar in Alzheimer's Disease Patients and Age-Matched Healthy Controls. *Frontiers in Neuroscience*, *14*, 874. https://doi.org/10.3389/fnins.2020.00874

Diallo, A., Jacobi, H., Cook, A., Labrum, R., Durr, A., Brice, A., Charles, P., Marelli, C., Mariotti, C., Nanetti, L., Panzeri, M., Rakowicz, M., Sobanska, A., Sulek, A., Schmitz-Hübsch, T., Schöls, L., Hengel, H., Melegh, B., Filla, A., ... Tezenas Du Montcel, S. (2018). Survival in patients with spinocerebellar ataxia types 1, 2, 3, and 6 (EUROSCA): A longitudinal cohort study. *The Lancet Neurology*, *17*(4), 327–334. https://doi.org/10.1016/S1474-4422(18)30042-5

Diekstra, F. P., Van Deerlin, V. M., Van Swieten, J. C., Al-Chalabi, A., Ludolph, A. C., Weishaupt, J. H., Hardiman, O., Landers, J. E., Brown, R. H., Van Es, M. A., Pasterkamp, R. J., Koppers, M., Andersen, P. M., Estrada, K., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., Van Damme, P., Melki, J., ... Veldink, J. H. (2014). C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: A genome-wide meta-analysis. *Annals of Neurology*, *76*(1), 120–133. https://doi.org/10.1002/ana.24198

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., ... Eberle, M. A. (2019). ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions (I. Birol, Ed.). *Bioinformatics*, *35*(22), 4754–4756. https://doi.org/10.1093/bioinformatics/btz431

Dolzhenko, E., van Vugt, J. J., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury,

Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H., van Es, M. A., ... Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, *27*(11), 1895–1903. https://doi.org/10.1101/gr.225672.117

Dolzhenko, E., Weisburd, B., Ibañez, K., Rajan-Babu, I.-S., Anyansi, C., Bennett, M. F., Billingsley, K., Carroll, A., Clamons, S., Danzi, M. C., Deshpande, V., Ding, J., Fazal, S., Halman, A., Jadhav, B., Qiu, Y., Richmond, P. A., Saunders, C. T., Scheffler, K., ... Eberle, M. A. (2022). REViewer: Haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Medicine*, *14*(1), 84. https://doi.org/10.1186/s13073-022-01085-z

Eitan, C., Siany, A., Barkan, E., Olender, T., Van Eijk, K. R., Moisse, M., Farhan, S. M. K., Danino, Y. M., Yanowski, E., Marmor-Kollet, H., Rivkin, N., Yacovzada, N. S., Hung, S.-T., Cooper-Knock, J., Yu, C.-H., Louis, C., Masters, S. L., Kenna, K. P., Van Der Spek, R. A. A., ... Hornstein, E. (2022). Whole-genome sequencing reveals that variants in the Interleukin 18 Receptor Accessory Protein 3'UTR protect against ALS. *Nature Neuroscience*, *25*(4), 433–445. https://doi.org/10.1038/s41593-022-01040-6

Elden, A. C., Kim, H. J., Hart, M. P., Chen-Plotkin, A. S., Johnson, B. S., Fang, X., Armakola, M., Geser, F., Greene, R., Lu, M. M., Padmanabhan, A., Clay-Falcone, D., McCluskey, L., Elman, L., Juhr, D., Gruber, P. J., Rüb, U., Auburger, G., Trojanowski, J. Q., ... Gitler, A. D. (2010). Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, *466*(7310), 1069–1075. https://doi.org/10.1038/nature09320

Estades Ayuso, V., Pickles, S., Todd, T., Yue, M., Jansen-West, K., Song, Y., González Bejarano, J., Rawlinson, B., DeTure, M., Graff-Radford, N. R., Boeve, B. F., Knopman, D. S., Petersen, R. C., Dickson, D. W., Josephs, K. A., Petrucelli, L., & Prudencio, M. (2023). TDP-43-regulated cryptic RNAs accumulate in Alzheimer's disease brains. *Molecular Neurodegeneration*, *18*(1), 57. https://doi.org/10.1186/s13024-023-00646-z

Evangelista, T., Weihl, C. C., Kimonis, V., Lochmüller, H., Clemen, C., Deshaies, R., Evangelista, T., Eymard, B., Greensmith, L., Hilton-Jones, D., Kimonis, V., Kley, R., Lochmüller, H., Meyer, H., Mozaffar, T., Noguchi, S., Ralston, S., Ridha, B., Udd, B., ... Brumhard, S. (2016). 215th ENMC International Workshop VCP-related multi-system proteinopathy (IBMPFD) 13–15 November 2015, Heemskerk, The Netherlands. *Neuromuscular Disorders*, *26*(8), 535–547. https://doi.org/10.1016/j.nmd.2016.05.017

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fanin, M., Savarese, M., Nascimbeni, A. C., Di Fruscio, G., Pastorello, E., Tasca, E., Trevisan, C. P., Nigro, V., & Angelini, C. (2015). Dominant muscular dystrophy with a novel *SYNE1* gene mutation. *Muscle & Nerve*, *51*(1), 145–147. https://doi.org/10.1002/mus.24357

Fecto, F., Yan, J., Vemula, S. P., Liu, E., Yang, Y., Chen, W., Zheng, J. G., Shi, Y., Siddique, N., Arrat, H., Donkervoort, S., Ajroud-Driss, S., Sufit, R. L., Heller, S. L., Deng, H. X., & Siddique, T. (2011). SQSTM1 Mutations in Familial and Sporadic Amyotrophic Lateral Sclerosis. *Archives of Neurology*, *68*(11), 1440–1446. https://doi.org/10.3390/ijms18122709

Fernandopulle, M. S., Prestil, R., Grunseich, C., Wang, C., Gan, L., & Ward, M. E. (2018). Transcription Factor-Mediated Differentiation of Human iPSCs into Neurons: Rapid differentiation of iPSCs into neurons. *Current Protocols in Cell Biology*, *79*(1), e51. https://doi.org/10.1002/cpcb.51

Ferraro, P. M., Agosta, F., Riva, N., Copetti, M., Spinelli, E. G., Falzone, Y., Sorarù, G., Comi, G., Chiò, A., & Filippi, M. (2017). Multimodal structural MRI in the diagnosis of motor neuron diseases [Publisher: Elsevier]. *NeuroImage: Clinical*, *16*(May), 240–247. https://doi.org/10.1016/j.nicl.2017.08.002

Fiesel, F. C., Weber, S. S., Supper, J., Zell, A., & Kahle, P. J. (2012). TDP-43 regulates global translational yield by splicing of exon junction complex component SKAR. *Nucleic Acids Research*, *40*(6), 2668–2682. https://doi.org/10.1093/nar/gkr1082

Figlewicz, D. A., Krizus, A., Martinoli, M. G., Meininger, V., Dib, M., Rouleau, G. A., & Julien, J.-P. (1994). Variants of the heavy neurofilament subunit are associated with the development of amyotrophic lateral sclerosis. *Human Molecular Genetics*, *3*(10), 1757–1761. https://doi.org/10.1093/hmg/3.10.1757

Fisher, E., & Feng, J. (2022). RNA splicing regulators play critical roles in neurogenesis. *WIREs RNA*, *13*(6), e1728. https://doi.org/10.1002/wrna.1728

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., ... Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Fratta, P., Nirmalananthan, N., Masset, L., Skorupinska, I., Collins, T., Cortese, A., Pemble, S., Malaspina, A., Fisher, E. M. C., Greensmith, L., & Hanna, M. G. (2014). Correlation of clinical and molecular features in spinal bulbar muscular atrophy. *Neurology*, *82*(23), 2077–2084. https://doi.org/10.1212/WNL.0000000000000507

Fratta, P., Collins, T., Pemble, S., Nethisinghe, S., Devoy, A., Giunti, P., Sweeney, M. G., Hanna, M. G., & Fisher, E. M. (2014). Sequencing analysis of the spinal bulbar muscular atrophy CAG expansion reveals absence of repeat interruptions. *Neurobiology of Aging*, *35*(2), 443.e1–443.e3. https://doi.org/10.1016/j.neurobiolaging.2013.07.015

Freischmidt, A., Wieland, T., Richter, B., Ruf, W., Schaeffer, V., Müller, K., Marroquin, N., Nordin, F., Hübers, A., Weydt, P., Pinto, S., Press, R., Millecamps, S., Molko, N., Bernard, E., Desnuelle, C., Soriani, M. H., Dorst, J., Graf, E., ... Weishaupt, J. H. (2015). Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nature Neuroscience*, *18*(5), 631–636. https://doi.org/10.1038/nn.4000

Gaiani, A., Martinelli, I., Bello, L., Querin, G., Puthenparampil, M., Ruggero, S., Toffanin, E., Cagnin, A., Briani, C., Pegoraro, E., & Sorarù, G. (2017). Diagnostic and prognostic biomarkers in amyotrophic lateral sclerosis: Neurofilament light chain levels in definite subtypes of disease. *JAMA Neurology*, *74*(5), 525–532. https://doi.org/10.1001/jamaneurol.2016.5398

Gardiner, S. L., Boogaard, M. W., Trompet, S., de Mutsert, R., Rosendaal, F. R., Gussekloo, J., Jukema, J. W., Roos, R. A. C., & Aziz, N. A. (2019). Prevalence of Carriers of Intermediate and Pathological Polyglutamine Disease–Associated Alleles Among Large Population-Based Cohorts. *JAMA Neurology*, *76*(6), 650. https://doi.org/10.1001/jamaneurol.2019.0423

Gasset-Rosa, F., Lu, S., Yu, H., Chen, C., Melamed, Z., Guo, L., Shorter, J., Da Cruz, S., & Cleveland, D. W. (2019). Cytoplasmic TDP-43 De-mixing Independent of Stress Granules Drives Inhibition of Nuclear Import, Loss of Nuclear TDP-43, and Cell Death. *Neuron*, *102*(2), 339–357.e7. https://doi.org/10.1016/j.neuron.2019.02.038

Gate, D., Saligrama, N., Leventhal, O., Yang, A. C., Unger, M. S., Middeldorp, J., Chen, K., Lehallier, B., Channappa, D., De Los Santos, M. B., McBride, A., Pluvinage, J., Elahi, F., Tam, G. K.-Y., Kim, Y., Greicius, M., Wagner, A. D., Aigner, L., Galasko, D. R., ... Wyss-Coray, T. (2020). Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. *Nature*, *577*(7790), 399–404. https://doi.org/10.1038/s41586-019-1895-7

Ghasemi, M., & Brown, R. H. (2018). Genetics of Amyotrophic Lateral Sclerosis. *Cold Spring Harbor Perspectives in Medicine*, *8*(5), a024125. https://doi.org/10.1101/cshperspect.a024125

Gitcho, M. A., Baloh, R. H., Chakraverty, S., Mayo, K., Norton, J. B., Levitch, D., Hatanpaa, K. J., White, C. L., Bigio, E. H., Caselli, R., Baker, M. C., Al-Lozi, M. T., Morris, J. C., Pestronk, A., Rademakers, R., Goate, A. M., & Cairns, N. J. (2008). TDP-43 A315T mutation in familial motor neuron disease. *Annals of Neurology*, *63*(4), 535–538. https://doi.org/10.1002/ana.21344

Glasmacher, S. A., Wong, C., Pearson, I. E., & Pal, S. (2020). Survival and Prognostic Factors in *C9orf72* Repeat Expansion Carriers: A Systematic Review and Meta-analysis. *JAMA Neurology*, *77*(3), 367. https://doi.org/10.1001/jamaneurol.2019.3924

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., Ogar, J. M., Rohrer, J. D., Black, S., Boeve, B. F., Manes, F., Dronkers, N. F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B. L., Knopman, D. S., Hodges, J. R., Mesulam, M., & Grossman, M. (2011). Classification of primary progressive aphasia and its variants. *Neurology*, *76*(11), 1006–1014. https://doi.org/10.1212/WNL.0b013e31821103e6

Gossye, H., Engelborghs, S., Van Broeckhoven, C., & van der Zee, J. (1999, February). C9orf72 Frontotemporal Dementia and/or Amyotrophic Lateral Sclerosis. In *GeneReviews® (Adam MP, Feldman J, Mirzaa GM, et al., editors.)* https://www.ncbi.nlm.nih.gov/books/NBK268647/

Greaves, C. V., & Rohrer, J. D. (2019). An update on genetic frontotemporal dementia. *Journal of Neurology*, *266*(8), 2075–2086. https://doi.org/10.1007/s00415-019-09363-4

Greenberg, S. A. (2019). Inclusion body myositis: Clinical features and pathogenesis. *Nature Reviews Rheumatology*, *15*(5), 257–272. https://doi.org/10.1038/s41584-019-0186-x

Greenberg, S. A., Pinkus, J. L., Kong, S. W., Baecher-Allan, C., Amato, A. A., & Dorfman, D. M. (2019). Highly differentiated cytotoxic T cells in inclusion body myositis. *Brain*, *142*(9), 2590–2604. https://doi.org/10.1093/brain/awz207

Greenway, M. J., Andersen, P. M., Russ, G., Ennis, S., Cashman, S., Donaghy, C., Patterson, V., Swingler, R., Kieran, D., Prehn, J., Morrison, K. E., Green, A., Acharya, K. R., Brown, R. H., & Hardiman, O. (2006). ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nature Genetics*, *38*(4), 411–413. https://doi.org/10.1038/ng1742

Griggs, R. C., Askanas, V., DiMauro, S., Engel, A., Karpati, G., Mendell, J. R., & Rowland, L. P. (1995). Inclusion body myositis and myopathies. *Annals of Neurology*, *38*(5), 705–713. https://doi.org/10.1002/ana.410380504

Gros-Louis, F., Lariviere, R., Gowing, G., Laurent, S., Camu, W., Bouchard, J. P., Meininger, V., Rouleau, G. A., & Julien, J. P. (2004). A frameshift deletion in peripherin gene associated with amyotrophic lateral sclerosis. *Journal of Biological Chemistry*, *279*(44), 45951–45956. https://doi.org/10.1074/jbc.M408139200

Gros-Louis, F., Dupré, N., Dion, P., Fox, M. A., Laurent, S., Verreault, S., Sanes, J. R., Bouchard, J.-P., & Rouleau, G. A. (2007). Mutations in SYNE1 lead

to a newly discovered form of autosomal recessive cerebellar ataxia. *Nature Genetics*, *39*(1), 80–85. https://doi.org/10.1038/ng1927

Guidetti, D., Sabadini, R., Ferlini, A., & Torrente, I. (2001). Epidemiological survey of X-linked bulbar and spinal muscular atrophy, or Kennedy disease, in the province of Reggio Emilia, Italy. *European Journal of Epidemiology*, *17*(6), 587–91. https://doi.org/10.1023/a:1014580219761

Hadano, S., Hand, C. K., Osuga, H., Yanagisawa, Y., Otomo, A., Devon, R. S., Miyamoto, N., Showguchi-Miyata, J., Okada, Y., Singaraja, R., Figlewicz, D. A., Kwiatkowski, T. J., Hosler, B. A., Sagie, T., Skaug, J., Nasir, J., Brown, R. H., Scherer, S. W., Rouleau, G. A., ... Ikeda, J. E. (2001). A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nature Genetics*, *29*(2), 166–173. https://doi.org/10.1038/ng1001-166

Haeusler, A. R., Donnelly, C. J., Periz, G., Simko, E. A., Shaw, P. G., Kim, M. S., Maragakis, N. J., Troncoso, J. C., Pandey, A., Sattler, R., Rothstein, J. D., & Wang, J. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease [Publisher: Nature Publishing Group]. *Nature*, *507*(7491), 195–200. https://doi.org/10.1038/nature13124

Haeusler, A. R., Donnelly, C. J., & Rothstein, J. D. (2016). The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease [Publisher: Nature Publishing Group]. *Nature Reviews Neuroscience*, *17*(6), 383–395. https://doi.org/10.1038/nrn.2016.38

Hallegger, M., Chakrabarti, A. M., Lee, F. C., Lee, B. L., Amalietti, A. G., Odeh, H. M., Copley, K. E., Rubien, J. D., Portz, B., Kuret, K., Huppertz, I., Rau, F., Patani, R., Fawzi, N. L., Shorter, J., Luscombe, N. M., & Ule, J. (2021). TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. *Cell*, *184*(18), 4680–4696.e22. https://doi.org/10.1016/j.cell.2021.07.018

Hardiman, O., Al-chalabi, A., Chiò, A., Corr, E. M., Robberecht, W., Shaw, P. J., & Simmons, Z. (2017). Amyotrophic lateral sclerosis. *Primer*, *3*(17071). https://doi.org/10.1038/nrdp.2017.71

Henden, L., Fearnley, L. G., Grima, N., McCann, E. P., Dobson-Stone, C., Fitzpatrick, L., Friend, K., Hobson, L., Chan Moi Fat, S., Rowe, D. B., D'Silva, S., Kwok, J. B., Halliday, G. M., Kiernan, M. C., Mazumder, S., Timmins, H. C., Zoing, M., Pamphlett, R., Adams, L., ... Williams, K. L. (2023). Short tandem repeat expansions in sporadic amyotrophic lateral sclerosis and frontotemporal dementia. *Science Advances*, *9*(18), eade2044. https://doi.org/10.1126/sciadv.ade2044

Hernandez Lain, A., Millecamps, S., Dubourg, O., Salachas, F., Bruneteau, G., Lacomblez, L., LeGuern, E., Seilhean, D., Duyckaerts, C., Meininger, V., Mallet, J., & Pradat, P.-F. (2011). Abnormal TDP-43 and FUS proteins in muscles of sporadic IBM: Similarities in a TARDBP-linked ALS patient.

Journal of Neurology, Neurosurgery & Psychiatry, *82*(12), 1414–1416. https://doi.org/10.1136/jnnp.2010.208868

Hobson, E. V., & McDermott, C. J. (2016). Supportive and symptomatic management of amyotrophic lateral sclerosis. *Nature Reviews Neurology*, *12*(9), 526–538. https://doi.org/10.1038/nrneurol.2016.111

Hogan, D. B., Jetté, N., Fiest, K. M., Roberts, J. I., Pearson, D., Smith, E. E., Roach, P., Kirk, A., Pringsheim, T., & Maxwell, C. J. (2016). The Prevalence and Incidence of Frontotemporal Dementia: A Systematic Review. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, *43*(S1), S96–S109. https://doi.org/10.1017/cjn.2016.25

Hu, D., Xia, W., & Weiner, H. L. (2022). CD8+ T cells in neurodegeneration: Friend or foe? *Molecular Neurodegeneration*, *17*(1), 59. https://doi.org/10.1186/s13024-022-00563-7

Humphrey, J., Birsa, N., Milioto, C., McLaughlin, M., Ule, A. M., Robaldo, D., Eberle, A. B., Kräuchi, R., Bentham, M., Brown, A.-L., Jarvis, S., Bodo, C., Garone, M. G., Devoy, A., Soraru, G., Rosa, A., Bozzoni, I., Fisher, E. M. C., Mühlemann, O., … Fratta, P. (2020). FUS ALS-causative mutations impair FUS autoregulation and splicing factor networks through intron retention. *Nucleic Acids Research*, *48*(12), 6889–6905. https://doi.org/10.1093/nar/gkaa410

Humphrey, J., Emmett, W., Fratta, P., Isaacs, A. M., & Plagnol, V. (2017). Quantitative analysis of cryptic splicing associated with TDP-43 depletion. *BMC Medical Genomics*, *10*(1), 38. https://doi.org/10.1186/s12920-017-0274-1

Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A., Hackett, J., Adamson, J., Lincoln, S., Dickson, D., Davies, P., Petersen, R. C., Stevens, M., De Graaff, E., Wauters, E., … Heutink, P. (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, *393*(6686), 702–705. https://doi.org/10.1038/31508

Ibañez, K., Jadhav, B., Zanovello, M., Gagliardi, D., Clarkson, C., Facchini, S., Garg, P., Martin-Trujillo, A., Gies, S. J., Galassi Deforie, V., Dalmia, A., Hensman Moss, D. J., Vandrovcova, J., Rocca, C., Moutsianas, L., Marini-Bettolo, C., Walker, H., Turner, C., Shoai, M., … Tucci, A. (2024). Increased frequency of repeat expansion mutations across different populations. *Nature Medicine*. https://doi.org/10.1038/s41591-024-03190-5

Ibañez, K., Polke, J., Hagelstrom, R. T., Dolzhenko, E., Pasko, D., Thomas, E. R. A., Daugherty, L. C., Kasperaviciute, D., Smith, K. R., Deans, Z. C., Hill, S., Fowler, T., Scott, R. H., Hardy, J., Chinnery, P. F., Houlden, H., Rendon, A., Caulfield, M. J., Eberle, M. A., … Zarowiecki, M. (2022). Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: A retrospective diagnostic accuracy and prospective clinical validation

study. *The Lancet Neurology*, *21*(3), 234–245. https://doi.org/10.1016/S1474-4422(21)00462-2

Irwin, K. E., Jasin, P., Braunstein, K. E., Sinha, I. R., Garret, M. A., Bowden, K. D., Chang, K., Troncoso, J. C., Moghekar, A., Oh, E. S., Raitcheva, D., Bartlett, D., Miller, T., Berry, J. D., Traynor, B. J., Ling, J. P., & Wong, P. C. (2024). A fluid biomarker reveals loss of TDP-43 splicing repression in presymptomatic ALS–FTD. *Nature Medicine*, *30*(2), 382–393. https://doi.org/10.1038/s41591-023-02788-5

Izumi, Y., Miyamoto, R., Morino, H., Yoshizawa, A., Nishinaka, K., Udaka, F., Kameyama, M., Maruyama, H., & Kawakami, H. (2013). Cerebellar ataxia with *SYNE1* mutation accompanying motor neuron disease. *Neurology*, *80*(6), 600–601. https://doi.org/10.1212/WNL.0b013e3182815529

Jeong, Y. H., Ling, J. P., Lin, S. Z., Donde, A. N., Braunstein, K. E., Majounie, E., Traynor, B. J., LaClair, K. D., Lloyd, T. E., & Wong, P. C. (2017). Tdp-43 cryptic exons are highly variable between cell types. *Molecular Neurodegeneration*, *12*(1), 13. https://doi.org/10.1186/s13024-016-0144-x

Jiang, N., Malone, M., & Chizari, S. (2023). Antigen-specific and cross-reactive T cells in protection and disease. *Immunological Reviews*, *316*(1), 120–135. https://doi.org/10.1111/imr.13217

Johnson, J. O., Butterfield, R. J., Mayne, K., Newcomb, T., Imburgia, C., Dunn, D., Duval, B., Feldkamp, M. L., & Weiss, R. B. (2021). Population-Based Prevalence of Myotonic Dystrophy Type 1 Using Genetic Analysis of Statewide Blood Screening Program. *Neurology*, *96*(7). https://doi.org/10.1212/WNL.0000000000011425

Johnson, J. O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V. M., Trojanowski, J. Q., Gibbs, J. R., Brunetti, M., Gronka, S., Wuu, J., Ding, J., McCluskey, L., Martinez-Lage, M., Falcone, D., Hernandez, D. G., Arepalli, S., Chong, S., Schymick, J. C., Rothstein, J. D., … Traynor, B. J. (2010). Exome Sequencing Reveals VCP Mutations as a Cause of Familial ALS. *Neuron*, *68*(5), 857–864. https://doi.org/10.1016/j.neuron.2010.11.036

Johnson, J. O., Pioro, E. P., Boehringer, A., Chia, R., Feit, H., Renton, A. E., Pliner, H. A., Abramzon, Y., Marangi, G., Winborn, B. J., Gibbs, J. R., Nalls, M. A., Morgan, S., Shoai, M., Hardy, J., Pittman, A., Orrell, R. W., Malaspina, A., Sidle, K., … Pirisi, A. (2014). Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis [Publisher: Nature Publishing Group]. *Nature Neuroscience*, *17*(5), 664–666. https://doi.org/10.1038/nn.3688

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein

structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M., & Kang, H. M. (2012). Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics*, *91*(5), 839–848. https://doi.org/10.1016/j.ajhg.2012.09.004

Kabashi, E., Valdmanis, P. N., Dion, P. A., Spiegelman, D., McConkey, B. J., Velde, C. V., Bouchard, J. P., Lacomblez, L., Pochigaeva, K., Salachas, F., Pradat, P. F., Camu, W., Meininger, V., Dupre, N., & Rouleau, G. A. (2008). TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nature Genetics*, *40*(5), 572–574. https://doi.org/10.1038/ng.132

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kay, C., Collins, J. A., Miedzybrodzka, Z., Madore, S. J., Gordon, E. S., Gerry, N., Davidson, M., Slama, R. A., & Hayden, M. R. (2016). Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology*, *87*(3), 282–288. https://doi.org/10.1212/WNL.0000000000002858

Kay, C., Collins, J. A., Wright, G. E., Baine, F., Miedzybrodzka, Z., Aminkeng, F., Semaka, A. J., McDonald, C., Davidson, M., Madore, S. J., Gordon, E. S., Gerry, N. P., Cornejo-Olivas, M., Squitieri, F., Tishkoff, S., Greenberg, J. L., Krause, A., & Hayden, M. R. (2018). The molecular epidemiology of Huntington disease is related to intermediate allele frequency and haplotype in the general population. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *177*(3), 346–357. https://doi.org/10.1002/ajmg.b.32618

Keller, C. W., Schmidt, J., & Lünemann, J. D. (2017). Immune and myodegenerative pathomechanisms in inclusion body myositis. *Annals of Clinical and Translational Neurology*, *4*(6), 422–445. https://doi.org/10.1002/acn3.419

Kenna, K. P., Van Doormaal, P. T., Dekker, A. M., Ticozzi, N., Kenna, B. J., Diekstra, F. P., Van Rheenen, W., Van Eijk, K. R., Jones, A. R., Keagle, P., Shatunov, A., Sproviero, W., Smith, B. N., Van Es, M. A., Topp, S., Kenna, A., Miller, J. W., Fallini, C., Tiloca, C., … Castellotti, B. (2016). NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nature Genetics*, *48*(9), 1037–1042. https://doi.org/10.1038/ng.3626

Keuss, M. J., Harly, P., Ryadnov, E., Jackson, R. E., Zanovello, M., Wilkins, O. G., Barattucci, S., Mehta, P. R., Oliveira, M. G., Parkes, J. E., Sinha, A., Correa-Sánchez, A. F., Oliver, P. L., Fisher, E. M., Schiavo, G., Shah, M., Burrone, J., & Fratta, P. (2024, June). Loss of TDP-43 induces synaptic dysfunction that is rescued by UNC13A splice-switching ASOs. https://doi.org/10.1101/2024.06.20.599684

Kim, H. J., Kim, N. C., Wang, Y. D., Scarborough, E. A., Moore, J., Diaz, Z., MacLea, K. S., Freibaum, B. D., Li, S., Molliex, A., Kanagaraj, A. P., Carter, R., Boylan, K. B., Wojtas, A. M., Rademakers, R., Pinkus, J. L., Greenberg, S. A., Trojanowski, J. Q., Traynor, B. J., ... Taylor, J. P. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS [Publisher: Nature Publishing Group]. *Nature*, *495*(7442), 467–473. https://doi.org/10.1038/nature11922

Kim, J. J., Vitale, D., Otani, D. V., Lian, M. M., Heilbron, K., the 23andMe Research Team, Aslibekyan, S., Auton, A., Babalola, E., Bell, R. K., Bielenberg, J., Bryc, K., Bullis, E., Cannon, P., Coker, D., Partida, G. C., Dhamija, D., Das, S., Elson, S. L., ... Mata, I. (2024). Multi-ancestry genome-wide association meta-analysis of Parkinson's disease. *Nature Genetics*, *56*(1), 27–36. https://doi.org/10.1038/s41588-023-01584-8

Klim, J. R., Williams, L. A., Limone, F., Guerra San Juan, I., Davis-Dusenbery, B. N., Mordes, D. A., Burberry, A., Steinbaugh, M. J., Gamage, K. K., Kirchner, R., Moccia, R., Cassel, S. H., Chen, K., Wainger, B. J., Woolf, C. J., & Eggan, K. (2019). ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. *Nature Neuroscience*, *22*(2), 167–179. https://doi.org/10.1038/s41593-018-0300-4

Korobeynikov, V. A., Lyashchenko, A. K., Blanco-Redondo, B., Jafar-Nejad, P., & Shneider, N. A. (2022). Antisense oligonucleotide silencing of FUS expression as a therapeutic approach in amyotrophic lateral sclerosis. *Nature Medicine*, *28*(1), 104–116. https://doi.org/10.1038/s41591-021-01615-z

Kumar, R., & Dhanda, S. K. (2020). Views on Global Genome Catalogues. *Bioinformation*, *16*(4), 297–300. https://doi.org/10.6026/97320630016297

Küsters, B., van Hoeve, B. J. A., Schelhaas, H. J., ter Laak, H., van Engelen, B. G. M., & Lammens, M. (2009). TDP-43 accumulation is common in myopathies with rimmed vacuoles. *Acta Neuropathologica*, *117*(2), 209–211. https://doi.org/10.1007/s00401-008-0471-2

Kwiatkowski, T. J., Bosco, D. A., LeClerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E. J., Munsat, T. L., Valdmanis, P. N., Rouleau, G. A., Hosler, B. A., Cortelli, P., De Jong, P. J., Yoshinaga, Y., Haines, J. L., Pericak-Vance, M. A., Yan, J., ... Brown, R. H. (2009). Mutations in the FUS/TLS gene on chromosome 16 cause

familial amyotrophic lateral sclerosis. *Science*, *323*(5918), 1205–1208. https://doi.org/10.1126/science.1166066

La Spada, A. R. (2022, December). Spinal and Bulbar Muscular Atrophy. In *GeneReviews® (Adam MP, Ardinger HH, Pagon RA, et al., editors)*. https://www.ncbi.nlm.nih.gov/books/NBK1333/

Langbehn, D., Brinkman, R., Falush, D., Paulsen, J., Hayden, M., & on behalf of an International Huntington's Disease Collaborative Group. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics*, *65*(4), 267–277. https://doi.org/10.1111/j.1399-0004.2004.00241.x

Laskaratos, A., Breza, M., Karadima, G., & Koutsis, G. (2021). Wide range of reduced penetrance alleles in spinal and bulbar muscular atrophy: A model-based approach. *Journal of Medical Genetics*, *58*(6), 385–391. https://doi.org/10.1136/jmedgenet-2020-106963

Laumont, C. M., & Perreault, C. (2018). Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cellular and Molecular Life Sciences*, *75*(4), 607–621. https://doi.org/10.1007/s00018-017-2628-4

Lee, E. B., Lee, V. M.-Y., & Trojanowski, J. Q. (2012). Gains or losses: Molecular mechanisms of TDP43-mediated neurodegeneration. *Nature Reviews Neuroscience*, *13*(1), 38–50. https://doi.org/10.1038/nrn3121

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, *50*(1), 151–158. https://doi.org/10.1038/s41588-017-0004-9

Ling, J. P., Chhabra, R., Merran, J. D., Schaughency, P. M., Wheelan, S. J., Corden, J. L., & Wong, P. C. (2016). PTBP1 and PTBP2 Repress Nonconserved Cryptic Exons. *Cell Reports*, *17*(1), 104–113. https://doi.org/10.1016/j.celrep.2016.08.071

Ling, J. P., Pletnikova, O., Troncoso, J. C., & Wong, P. C. (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*, *349*(6248), 650–655. https://doi.org/10.1126/science.aab0983

Ling, J. P., Wilks, C., Charles, R., Leavey, P. J., Ghosh, D., Jiang, L., Santiago, C. P., Pang, B., Venkataraman, A., Clark, B. S., Nellore, A., Langmead, B., & Blackshaw, S. (2020). ASCOT identifies key regulators of neuronal

subtype-specific splicing. *Nature Communications*, *11*(1), 137. https://doi.org/10.1038/s41467-019-14020-5

Liu, E. Y., Russ, J., Cali, C. P., Phan, J. M., Amlie-Wolf, A., & Lee, E. B. (2019). Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. *Cell Reports*, *27*(5), 1409–1421.e6. https://doi.org/10.1016/j.celrep.2019.04.003

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-{Δ}{Δ}CT Method. *Methods*, *25*(4), 402–408. https://doi.org/10.1006/meth.2001.1262

Logroscino, G., Piccininni, M., Graff, C., Hardiman, O., Ludolph, A. C., Moreno, F., Otto, M., Remes, A. M., Rowe, J. B., Seelaar, H., Solje, E., Stefanova, E., Traykov, L., Jelic, V., Rydell, M. T., Pender, N., Anderl-Straub, S., Barandiaran, M., Gabilondo, A., … Zulaica, M. (2023). Incidence of Syndromes Associated With Frontotemporal Lobar Degeneration in 9 European Countries. *JAMA Neurology*, *80*(3), 279. https://doi.org/10.1001/jamaneurol.2022.5128

Lois, C., Hong, E. J., Pease, S., Brown, E. J., & Baltimore, D. (2002). Germline Transmission and Tissue-Specific Expression of Transgenes Delivered by Lentiviral Vectors. *Science*, *295*(5556), 868–872. https://doi.org/10.1126/science.1067081

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lukavsky, P. J., Daujotyte, D., Tollervey, J. R., Ule, J., Stuani, C., Buratti, E., Baralle, F. E., Damberger, F. F., & Allain, F. H.-T. (2013). Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. *Nature Structural & Molecular Biology*, *20*(12), 1443–1449. https://doi.org/10.1038/nsmb.2698

Luty, A. A., Kwok, J. B., Dobson-Stone, C., Loy, C. T., Coupland, K. G., Karlström, H., Sobow, T., Tchorzewska, J., Maruszak, A., Barcikowska, M., Panegyres, P. K., Zekanowski, C., Brooks, W. S., Williams, K. L., Blair, I. P., Mather, K. A., Sachdev, P. S., Halliday, G. M., & Schofield, P. R. (2010). Sigma nonopioid intracellular receptor 1 mutations cause frontotemporal lobar degeneration-motor neuron disease. *Annals of Neurology*, *68*(5), 639–649. https://doi.org/10.1002/ana.22274

Ma, K.-Y., Schonnesen, A. A., He, C., Xia, A. Y., Sun, E., Chen, E., Sebastian, K. R., Guo, Y.-W., Balderas, R., Kulkarni-Date, M., & Jiang, N. (2021). High-throughput and high-dimensional single-cell analysis of antigen-specific CD8+ T cells. *Nature Immunology*, *22*(12), 1590–1598. https://doi.org/10.1038/s41590-021-01073-2

Ma, X. R., Prudencio, M., Koike, Y., Vatsavayai, S. C., Kim, G., Harbinski, F., Briner, A., Rodriguez, C. M., Guo, C., Akiyama, T., Schmidt, H. B., Cummings,

B. B., Wyatt, D. W., Kurylo, K., Miller, G., Mekhoubad, S., Sallee, N., Mekonnen, G., Ganser, L., ... Gitler, A. D. (2022). TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. *Nature*, *603*(7899), 124–130. https://doi.org/10.1038/s41586-022-04424-7

Manzano, R., Sorarú, G., Grunseich, C., Fratta, P., Zuccaro, E., Pennuto, M., & Rinaldi, C. (2018). Beyond motor neurons: Expanding the clinical spectrum in Kennedy's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *89*(8), 808–812. https://doi.org/10.1136/jnnp-2017-316961

Martinez-Val, A., Bekker-Jensen, D. B., Hogrebe, A., & Olsen, J. V. (2021). Data Processing and Analysis for DIA-Based Phosphoproteomics Using Spectronaut [Series Title: Methods in Molecular Biology]. In D. Cecconi (Ed.), *Proteomics Data Analysis* (pp. 95–107, Vol. 2361). Springer US. https://doi.org/10.1007/978-1-0716-1641-3_6

Maruyama, H., Morino, H., Ito, H., Izumi, Y., Kato, H., Watanabe, Y., Kinoshita, Y., Kamada, M., Nodera, H., Suzuki, H., Komure, O., Matsuura, S., Kobatake, K., Morimoto, N., Abe, K., Suzuki, N., Aoki, M., Kawata, A., Hirai, T., ... Kawakami, H. (2010). Mutations of optineurin in amyotrophic lateral sclerosis. *Nature*, *465*(7295), 223–226. https://doi.org/10.1038/nature08971

Mathieu, J., Allard, P., Potvin, L., Prévost, C., & Bégin, P. (1999). A 10-year study of mortality in a cohort of patients with myotonic dystrophy. *Neurology*, *52*(8), 1658–1658. https://doi.org/10.1212/WNL.52.8.1658

McCauley, M. E., & Baloh, R. H. (2019). Inflammation in ALS/FTD pathogenesis. *Acta Neuropathologica*, *137*(5), 715–730. https://doi.org/10.1007/s00401-018-1933-9

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Version Number: 3]. https://doi.org/10.48550/ARXIV.1802.03426

McKee, A. C., Stein, T. D., Kiernan, P. T., & Alvarez, V. E. (2015). The neuropathology of chronic traumatic encephalopathy. *Brain Pathology*, *25*(3), 350–364. https://doi.org/10.1111/bpa.12248

Melamed, Z., López-Erauskin, J., Baughn, M. W., Zhang, O., Drenner, K., Sun, Y., Freyermuth, F., McMahon, M. A., Beccari, M. S., Artates, J. W., Ohkubo, T., Rodriguez, M., Lin, N., Wu, D., Bennett, C. F., Rigo, F., Da Cruz, S., Ravits, J., Lagier-Tourenne, C., & Cleveland, D. W. (2019). Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. *Nature Neuroscience*, *22*(2), 180–190. https://doi.org/10.1038/s41593-018-0293-z

Mercado, P. A. (2005). Depletion of TDP 43 overrides the need for exonic and intronic splicing enhancers in the human apoA-II gene. *Nucleic Acids Research*, *33*(18), 6000–6010. https://doi.org/10.1093/nar/gki897

Mesulam, M. M., Wieneke, C., Thompson, C., Rogalski, E., & Weintraub, S. (2012). Quantitative classification of primary progressive aphasia at early and mild impairment stages. *Brain*, *135*(5), 1537–1553. https://doi.org/10.1093/brain/aws080

Meyerholz, D. K., & Beck, A. P. (2018). Principles and approaches for reproducible scoring of tissue stains in research. *Laboratory Investigation*, *98*(7), 844–855. https://doi.org/10.1038/s41374-018-0057-0

Miller, T. M., Cudkowicz, M. E., Genge, A., Shaw, P. J., Sobue, G., Bucelli, R. C., Chiò, A., Van Damme, P., Ludolph, A. C., Glass, J. D., Andrews, J. A., Babu, S., Benatar, M., McDermott, C. J., Cochrane, T., Chary, S., Chew, S., Zhu, H., Wu, F., … Fradette, S. (2022). Trial of Antisense Oligonucleotide Tofersen for *SOD1* ALS. *New England Journal of Medicine*, *387*(12), 1099–1110. https://doi.org/10.1056/NEJMoa2204705

Mitchell, J., Paul, P., Chen, H. J., Morris, A., Payling, M., Falchi, M., Habgood, J., Panoutsou, S., Winkler, S., Tisato, V., Hajitou, A., Smith, B. N., Vance, C., Shaw, C. E., Mazarakis, N. D., & De Belleroche, J. (2010). Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7556–7561. https://doi.org/10.1073/pnas.0914128107

Miyatake, S., Koshimizu, E., Fujita, A., Doi, H., Okubo, M., Wada, T., Hamanaka, K., Ueda, N., Kishida, H., Minase, G., Matsuno, A., Kodaira, M., Ogata, K., Kato, R., Sugiyama, A., Sasaki, A., Miyama, T., Satoh, M., Uchiyama, Y., … Matsumoto, N. (2022). Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *npj Genomic Medicine*, *7*(1), 62. https://doi.org/10.1038/s41525-022-00331-y

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.1

Moore, K. M., Nicholas, J., Grossman, M., McMillan, C. T., Irwin, D. J., Massimo, L., Van Deerlin, V. M., Warren, J. D., Fox, N. C., Rossor, M. N., Mead, S., Bocchetta, M., Boeve, B. F., Knopman, D. S., Graff-Radford, N. R., Forsberg, L. K., Rademakers, R., Wszolek, Z. K., Van Swieten, J. C., … Geschwind, D. (2020). Age at symptom onset and death and disease duration in genetic frontotemporal dementia: An international retrospective cohort study. *The Lancet Neurology*, *19*(2), 145–156. https://doi.org/10.1016/S1474-4422(19)30394-1

Murphy, N. A., Arthur, K. C., Tienari, P. J., Houlden, H., Chiò, A., & Traynor, B. J. (2017). Age-related penetrance of the C9orf72 repeat expansion. *Scientific Reports*, *7*(1), 2116. https://doi.org/10.1038/s41598-017-02364-1

Nalbandian, A., Donkervoort, S., Dec, E., Badadani, M., Katheria, V., Rana, P., Nguyen, C., Mukherjee, J., Caiozzo, V., Martin, B., Watts, G. D., Vesa, J., Smith, C., & Kimonis, V. E. (2011). The Multiple Faces of Valosin-Containing Protein-Associated Diseases: Inclusion Body Myopathy with Paget's Disease of Bone, Frontotemporal Dementia, and Amyotrophic Lateral Sclerosis. *Journal of Molecular Neuroscience*, *45*(3), 522–531. https://doi.org/10.1007/s12031-011-9627-y

Needham, M., Corbett, A., Day, T., Christiansen, F., Fabian, V., & Mastaglia, F. L. (2008). Prevalence of sporadic inclusion body myositis and factors contributing to delayed diagnosis. *Journal of Clinical Neuroscience*, *15*(12), 1350–1353. https://doi.org/10.1016/j.jocn.2008.01.011

Nelson, P. T., Dickson, D. W., Trojanowski, J. Q., Jack, C. R., Boyle, P. A., Arfanakis, K., Rademakers, R., Alafuzoff, I., Attems, J., Brayne, C., Coyle-Gilchrist, I. T. S., Chui, H. C., Fardo, D. W., Flanagan, M. E., Halliday, G., Hokkanen, S. R. K., Hunter, S., Jicha, G. A., Katsumata, Y., … Schneider, J. A. (2019). Limbic-predominant age-related TDP-43 encephalopathy (LATE): Consensus working group report. *Brain*, *142*(6), 1503–1527. https://doi.org/10.1093/brain/awz099

Neumann, M., Sampathu, D. M., Kwong, L. K., Truax, A. C., Micsenyi, M. C., Chou, T. T., Bruce, J., Schuck, T., Grossman, M., Clark, C. M., McCluskey, L. F., Miller, B. L., Masliah, E., Mackenzie, I. R., Feldman, H., Feiden, W., Kretzschmar, H. A., Trojanowski, J. Q., & Lee, V. M.-Y. (2006). Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science*, *314*(5796), 130–133. https://doi.org/10.1126/science.1134108

Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J. A., Kenna, B. J., Nalls, M. A., Keagle, P., Rivera, A. M., van Rheenen, W., Murphy, N. A., van Vugt, J. J., Geiger, J. T., Van der Spek, R. A., Pliner, H. A., Shankaracharya, Smith, B. N., … Twine, N. A. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron*, *97*(6), 1268–1283.e6. https://doi.org/10.1016/j.neuron.2018.02.027

Nishimura, A. L., Mitne-Neto, M., Silva, H. C., Richieri-Costa, A., Middleton, S., Cascio, D., Kok, F., Oliveira, J. R., Gillingwater, T., Webb, J., Skehel, P., & Zatz, M. (2004). A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. *American Journal of Human Genetics*, *75*(5), 822–831. https://doi.org/10.1086/425287

Noreau, A., Bourassa, C. V., Szuto, A., Levert, A., Dobrzeniecka, S., Gauthier, J., Forlani, S., Durr, A., Anheim, M., Stevanin, G., Brice, A., Bouchard, J.-P., Dion, P. A., Dupré, N., & Rouleau, G. A. (2013). *SYNE1* Mutations in Autosomal Recessive Cerebellar Ataxia. *JAMA Neurology*. https://doi.org/10.1001/jamaneurol.2013.3268

Olivé, M., Janué, A., Moreno, D., Gámez, J., Torrejón-Escribano, B., & Ferrer, I. (2009). TAR DNA-Binding Protein 43 Accumulation in Protein Aggregate Myopathies. *Journal of Neuropathology & Experimental Neurology*, *68*(3), 262–273. https://doi.org/10.1097/NEN.0b013e3181996d8f

Olney, N. T., Spina, S., & Miller, B. L. (2017). Frontotemporal Dementia. *Neurologic Clinics*, *35*(2), 339–374. https://doi.org/10.1016/j.ncl.2017.01.008

Opal, P., & Ashizawa, T. (2023, February). Spinocerebellar Ataxia Type 1. In *GeneReviews® (Adam MP, Feldman J, Mirzaa GM, et al., editors.)* https://www.ncbi.nlm.nih.gov/books/NBK1184/

Orlacchio, A., Babalini, C., Borreca, A., Patrono, C., Massa, R., Basaran, S., Munhoz, R. P., Rogaeva, E., St George-Hyslop, P. H., Bernardi, G., & Kawarai, T. (2010). SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain*, *133*(2), 591–598. https://doi.org/10.1093/brain/awp325

Ou, S. H., Wu, F., Harrich, D., García-Martínez, L. F., & Gaynor, R. B. (1995). Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. *Journal of Virology*, *69*(6), 3584–3596. https://doi.org/10.1128/jvi.69.6.3584-3596.1995

Pagani, M., Chiò, A., Valentini, M. C., Öberg, J., Nobili, F., Calvo, A., Moglia, C., Bertuzzo, D., Morbelli, S., De Carli, F., Fania, P., & Cistaro, A. (2014). Functional pattern of brain FDG-PET in amyotrophic lateral sclerosis. *Neurology*, *83*(12), 1067–1074. https://doi.org/10.1212/WNL.0000000000000792

Parkinson, N., Ince, P. G., Smith, M. O., Highley, R., Skibinski, G., Andersen, P. M., Morrison, K. E., Pall, H., Hardiman, O., Collinge, J., Shaw, P. J., & Fisher, E. M. (2006). ALS phenotypes with mutations in CHMP2B (charged multivesicular body protein 2B). *Neurology*, *67*(6), 1074–1077. https://doi.org/10.1212/01.wnl.0000231510.89311.8b

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Paulson, H. (2018). Repeat expansion diseases. In *Handbook of Clinical Neurology* (pp. 105–123, Vol. 147). Elsevier. https://doi.org/10.1016/B978-0-444-63233-3.00009-9

Pereverzev, A. P., Gurskaya, N. G., Ermakova, G. V., Kudryavtseva, E. I., Markina, N. M., Kotlobay, A. A., Lukyanov, S. A., Zaraisky, A. G., & Lukyanov, K. A. (2015). Method for quantitative analysis of nonsense-mediated mRNA decay at the single cell level. *Scientific Reports*, *5*(1), 7729. https://doi.org/10.1038/srep07729

Peters, S., Visser, A. E., D'Ovidio, F., Beghi, E., Chiò, A., Logroscino, G., Hardiman, O., Kromhout, H., Huss, A., Veldink, J. H., Vermeulen, R. C., & Van Den Berg, L. H. (2019). Associations of Electric Shock and Extremely Low-Frequency Magnetic Field Exposure with the Risk of Amyotrophic Lateral Sclerosis. *American Journal of Epidemiology*, *188*(4), 796–805. https://doi.org/10.1093/aje/kwy287

Philips, T., Bento-Abreu, A., Nonneman, A., Haeck, W., Staats, K. A., Geelen, V., Hersmus, N., Küsters, B., Van Den Bosch, L., Van Damme, P., Richardson, W. D., & Robberecht, W. (2013). Oligodendrocyte dysfunction in the pathogenesis of amyotrophic lateral sclerosis. *Brain*, *136*(2), 471–482. https://doi.org/10.1093/brain/aws339

Phukan, J., Elamin, M., Bede, P., Jordan, N., Gallagher, L., Byrne, S., Lynch, C., Pender, N., & Hardiman, O. (2012). The syndrome of cognitive impairment in amyotrophic lateral sclerosis: A population-based study. *Journal of Neurology, Neurosurgery and Psychiatry*, *83*(1), 102–108. https://doi.org/10.1136/jnnp-2011-300188

Polymenidou, M., & Cleveland, D. W. (2011). The seeds of neurodegeneration: Prion-like spreading in ALS [Publisher: Elsevier Inc.]. *Cell*, *147*(3), 498–508. https://doi.org/10.1016/j.cell.2011.10.011

Polymenidou, M., Lagier-Tourenne, C., Hutt, K. R., Huelga, S. C., Moran, J., Liang, T. Y., Ling, S.-C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J. P., Shiue, L., Bennett, C. F., Yeo, G. W., & Cleveland, D. W. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nature Neuroscience*, *14*(4), 459–468. https://doi.org/10.1038/nn.2779

Population estimates for the UK, England and Wales, Scotland and Northern Ireland: Mid-2019. (2020). https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019

Pottier, C., Bieniek, K. F., Finch, N. A., van de Vorst, M., Baker, M. C., Perkersen, R., Brown, P., Ravenscroft, T., van Blitterswijk, M., Nicholson, A. M., DeTure, M., Knopman, D. S., Josephs, K. A., Parisi, J. E., Petersen, R. C., Boylan, K. B., Boeve, B. F., Graff-Radford, N. R., Veltman, J. A., ... Rademakers, R. (2015). Whole-genome sequencing reveals important role for TBK1 and OPTN mutations in frontotemporal lobar degeneration without motor neuron disease [Publisher: Springer Berlin Heidelberg]. *Acta Neuropathologica*, *130*(1), 77–92. https://doi.org/10.1007/s00401-015-1436-x

Pottier, C., Ren, Y., Perkerson, R. B., Baker, M., Jenkins, G. D., Van Blitterswijk, M., DeJesus-Hernandez, M., Van Rooij, J. G. J., Murray, M. E., Christopher, E., McDonnell, S. K., Fogarty, Z., Batzler, A., Tian, S., Vicente, C. T., Matchett, B., Karydas, A. M., Hsiung, G.-Y. R., Seelaar, H., ... Rademakers, R. (2019).

Genome-wide analyses as part of the international FTLD-TDP whole-genome sequencing consortium reveals novel disease risk factors and increases support for immune dysfunction in FTLD. *Acta Neuropathologica*, *137*(6), 879–899. https://doi.org/10.1007/s00401-019-01962-9

Project MinE ALS Sequencing Consortium. (2018). Project MinE: Study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics*, *26*(10), 1537–1546. https://doi.org/10.1038/s41431-018-0177-4

Prpar Mihevc, S., Baralle, M., Buratti, E., & Rogelj, B. (2016). TDP-43 aggregation mirrors TDP-43 knockdown, affecting the expression levels of a common set of proteins. *Scientific Reports*, *6*(1), 33996. https://doi.org/10.1038/srep33996

Prudencio, M., Humphrey, J., Pickles, S., Brown, A.-L., Hill, S. E., Kachergus, J. M., Shi, J., Heckman, M. G., Spiegel, M. R., Cook, C., Song, Y., Yue, M., Daughrity, L. M., Carlomagno, Y., Jansen-West, K., de Castro, C. F., DeTure, M., Koga, S., Wang, Y.-C., ... Petrucelli, L. (2020). Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. *Journal of Clinical Investigation*, *130*(11), 6080–6092. https://doi.org/10.1172/JCI139741

Puls, I., Jonnakuty, C., LaMonte, B. H., Holzbaur, E. L., Tokito, M., Mann, E., Floeter, M. K., Bidus, K., Drayna, D., Oh, S. J., Brown, R. H., Ludlow, C. L., & Fischbeck, K. H. (2003). Mutant dynactin in motor neuron disease. *Nature Genetics*, *33*(4), 455–456. https://doi.org/10.1038/ng1123

Pulst, S. M. (2019, February). Spinocerebellar Ataxia Type 2. In *GeneReviews® (Adam MP, Feldman J, Mirzaa GM, et al., editors.)* https://www.ncbi.nlm.nih.gov/books/NBK1275/

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raaphorst, J., Beeldman, E., De Visser, M., De Haan, R. J., & Schmand, B. (2012). A systematic review of behavioural changes in motor neuron disease. *Amyotrophic Lateral Sclerosis*, *13*(6), 493–501. https://doi.org/10.3109/17482968.2012.656652

Ramachandran, S., Grozdanov, V., Leins, B., Kandler, K., Witzel, S., Mulaw, M., Ludolph, A. C., Weishaupt, J. H., & Danzer, K. M. (2023). Low T-cell reactivity to TDP-43 peptides in ALS. *Frontiers in Immunology*, *14*, 1193507. https://doi.org/10.3389/fimmu.2023.1193507

Rascovsky, K., Hodges, J. R., Knopman, D. S., Mendez, M. F., Kramer, J. H., Neuhaus, J., Van Swieten, J. C., Seelaar, H., Dopper, E. G., Onyike, C. U., Hillis, A. E., Josephs, K. A., Boeve, B. F., Kertesz, A., Seeley, W. W., Rankin, K. P., Johnson, J. K., Gorno-Tempini, M. L., Rosen, H., ... Miller, B. L. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of

frontotemporal dementia. *Brain*, *134*(9), 2456–2477. https://doi.org/10.1093/brain/awr179

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., Laaksovirta, H., van Swieten, J. C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A. M., Kaganovich, A., Scholz, S. W., Duckworth, J., Ding, J., Harmer, D. W., ... Traynor, B. J. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, *72*(2), 257–268. https://doi.org/10.1016/j.neuron.2011.09.010

Robberecht, W., & Philips, T. (2013). The changing scene of amyotrophic lateral sclerosis. *Nature Reviews Neuroscience*, *14*(4), 248–264. https://doi.org/10.1038/nrn3430

Roczniak-Ferguson, A., & Ferguson, S. M. (2019). Pleiotropic requirements for human TDP-43 in the regulation of cell and organelle homeostasis. *Life Science Alliance*, *2*(5), e201900358. https://doi.org/10.26508/lsa.201900358

Rose, M. (2013). 188th ENMC International Workshop: Inclusion Body Myositis, 2–4 December 2011, Naarden, The Netherlands. *Neuromuscular Disorders*, *23*(12), 1044–1055. https://doi.org/10.1016/j.nmd.2013.08.007

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P. C., Hentati, A., Donaldson, D., Goto, J., O'Regan, J. P., Deng, H. X., Rahmani, Z., Krizus, A., McKenna-Yasek, D., Cayabyab, A., Gaston, S. M., Berger, R., Tanzi, R. E., Halperin, J. J., Herzfeldt, B., ... Brown, R. H. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, *362*(6415), 59–62. https://doi.org/10.1038/362059a0

Rothwell, S., Cooper, R. G., Lundberg, I. E., Gregersen, P. K., Hanna, M. G., Machado, P. M., Herbert, M. K., Pruijn, G. J. M., Lilleker, J. B., Roberts, M., Bowes, J., Seldin, M. F., Vencovsky, J., Danko, K., Limaye, V., Selva-O'Callaghan, A., Platt, H., Molberg, Ø., Benveniste, O., ... for the Myositis Genetics Consortium. (2017). Immune-Array Analysis in Sporadic Inclusion Body Myositis Reveals HLA–DRB1 Amino Acid Heterogeneity Across the Myositis Spectrum. *Arthritis & Rheumatology*, *69*(5), 1090–1099. https://doi.org/10.1002/art.40045

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V., Wang, C., & Carter, A. B. (2018). Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines. *The Journal of Molecular Diagnostics*, *20*(1), 4–27. https://doi.org/10.1016/j.jmoldx.2017.11.003

Rüb, U., Schöls, L., Paulson, H., Auburger, G., Kermer, P., Jen, J. C., Seidel, K., Korf, H.-W., & Deller, T. (2013). Clinical features, neurogenetics and neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3,

6 and 7. *Progress in Neurobiology*, *104*, 38–66. https://doi.org/10.1016/j.pneurobio.2013.01.001

Sandell, S., Huovinen, S., Palmio, J., Raheem, O., Lindfors, M., Zhao, F., Haapasalo, H., & Udd, B. (2016). Diagnostically important muscle pathology in DNAJB6 mutated LGMD1D. *Acta Neuropathologica Communications*, *4*(1), 9. https://doi.org/10.1186/s40478-016-0276-9

Schöls, L., Bauer, P., Schmidt, T., Schulte, T., & Riess, O. (2004). Autosomal dominant cerebellar ataxias: Clinical features, genetics, and pathogenesis. *The Lancet Neurology*, *3*(5), 291–304. https://doi.org/10.1016/S1474-4422(04)00737-9

Schwab, C., Arai, T., Hasegawa, M., Yu, S., & McGeer, P. L. (2008). Colocalization of Transactivation-Responsive DNA-Binding Protein 43 and Huntingtin in Inclusions of Huntington Disease. *Journal of Neuropathology & Experimental Neurology*, *67*(12), 1159–1165. https://doi.org/10.1097/NEN.0b013e31818e8951

Seddighi, S., Qi, Y. A., Brown, A.-L., Wilkins, O. G., Bereda, C., Belair, C., Zhang, Y.-J., Prudencio, M., Keuss, M. J., Khandeshi, A., Pickles, S., Kargbo-Hill, S. E., Hawrot, J., Ramos, D. M., Yuan, H., Roberts, J., Sacramento, E. K., Shah, S. I., Nalls, M. A., … Ward, M. E. (2024). Mis-spliced transcripts generate de novo proteins in TDP-43-related ALS/FTD. *Science Translational Medicine*, eadg7162. https://doi.org/10.1126/scitranslmed.adg7162

Sephton, C. F., Cenik, C., Kucukural, A., Dammer, E. B., Cenik, B., Han, Y., Dewey, C. M., Roth, F. P., Herz, J., Peng, J., Moore, M. J., & Yu, G. (2011). Identification of Neuronal RNA Targets of TDP-43-containing Ribonucleoprotein Complexes. *Journal of Biological Chemistry*, *286*(2), 1204–1215. https://doi.org/10.1074/jbc.M110.190884

Sequeiros, J., Martins, S., & Silveira, I. (2012). Epidemiology and population genetics of degenerative ataxias. In *Handbook of Clinical Neurology* (pp. 227–251, Vol. 103). Elsevier. https://doi.org/10.1016/B978-0-444-51892-7.00014-0

Shiga, A., Ishihara, T., Miyashita, A., Kuwabara, M., Kato, T., Watanabe, N., Yamahira, A., Kondo, C., Yokoseki, A., Takahashi, M., Kuwano, R., Kakita, A., Nishizawa, M., Takahashi, H., & Onodera, O. (2012). Alteration of POLDIP3 Splicing Associated with Loss of Function of TDP-43 in Tissues Affected with ALS (Y. Nagai, Ed.). *PLoS ONE*, *7*(8), e43120. https://doi.org/10.1371/journal.pone.0043120

Sinnreich, M., Sorenson, E. J., & Klein, C. J. (2004). Neurologic Course, Endocrine Dysfunction and Triplet Repeat Size in Spinal Bulbar Muscular Atrophy. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, *31*(3), 378–382. https://doi.org/10.1017/S0317167100003486

Smith, B. N., Ticozzi, N., Fallini, C., Gkazi, A. S., Topp, S., Kenna, K. P., Scotter, E. L., Kost, J. E., Keagle, P., Miller, J. W., Calini, D., Vance, C., Daniel-

son, E. W., Troakes, C., Tiloca, C., Al-Sarraj, S., Lewis, E. A., King, A., Colombrita, C., ... Bertolin, C. (2014). Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron*, *84*(2), 324–331. https://doi.org/10.1016/j.neuron.2014.09.027

Smith, B. N., Topp, S., Fallini, C., Shibata, H., Chen, H. J., Troakes, C., King, A., Ticozzi, N., Kenna, K. P., Soragia-Gkazi, A., Miller, J. W., Sato, A., Dias, D. M., Jeon, M., Vance, C., Wong, C. H., De Majo, M., Kattuah, W., Mitchell, J. C., ... Shaw, C. E. (2017). Mutations in the vesicular trafficking protein Annexin A11 are associated with amyotrophic lateral sclerosis. *Science Translational Medicine*, *9*(388), 1–16. https://doi.org/10.1126/scitranslmed.aad9157

Snedden, A. M., Kellett, K. A., Lilleker, J. B., Hooper, N. M., & Chinoy, H. (2022). The role of protein aggregation in the pathogenesis of inclusion body myositis. *Clinical and Experimental Rheumatology*, *40*(2), 414–424. https://doi.org/10.55563/clinexprheumatol/pp0oso

Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., Kohno, Y., Takiyama, Y., Maeda, K., Doi, H., Koyano, S., Takeuchi, H., Kawamoto, M., Kohara, N., Ando, T., ... Sobue, G. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nature Genetics*, *51*(8), 1215–1221. https://doi.org/10.1038/s41588-019-0459-y

Sreedharan, J., Blair, I. P., Tripathi, V. B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J. C., Williams, K. L., Buratti, E., Baralle, F., De Belleroche, J., Mitchell, D., Leigh, P. N., Al-Chalabi, A., Miller, C. C., Nicholson, G. A., & Shaw, C. E. (2008). TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science*, *319*(5870), 1668–1672. https://doi.org/10.1126/science.1154584

Šušnjar, U., Škrabar, N., Brown, A.-L., Abbassi, Y., Phatnani, H., NYGC ALS Consortium, Phatnani, H., Fratta, P., Kwan, J., Sareen, D., Broach, J. R., Simmons, Z., Arcila-Londono, X., Lee, E. B., Van Deerlin, V. M., Shneider, N. A., Fraenkel, E., Ostrow, L. W., Baas, F., ... Buratti, E. (2022). Cell environment shapes TDP-43 function with implications in neuronal and muscle disease. *Communications Biology*, *5*(1), 314. https://doi.org/10.1038/s42003-022-03253-8

Synofzik, M., Smets, K., Mallaret, M., Di Bella, D., Gallenmüller, C., Baets, J., Schulze, M., Magri, S., Sarto, E., Mustafa, M., Deconinck, T., Haack, T., Züchner, S., Gonzalez, M., Timmann, D., Stendel, C., Klopstock, T., Durr, A., Tranchant, C., ... Bauer, P. (2016). SYNE1 ataxia is a common recessive ataxia with major non-cerebellar features: A large multi-centre study. *Brain*, *139*(5), 1378–1393. https://doi.org/10.1093/brain/aww079

Takano, H., Cancel, G., Ikeuchi, T., Lorenzetti, D., Mawad, R., Stevanin, G., Didierjean, O., Dürr, A., Oyake, M., Shimohata, T., Sasaki, R., Koide, R., Igarashi, S., Hayashi, S., Takiyama, Y., Nishizawa, M., Tanaka, H., Zoghbi, H., Brice, A., & Tsuji, S. (1998). Close Associations between Prevalences of Dominantly Inherited Spinocerebellar Ataxias with CAG-Repeat Expansions and Frequencies of Large Normal CAG Alleles in Japanese and Caucasian Populations. *The American Journal of Human Genetics*, *63*(4), 1060–1066. https://doi.org/10.1086/302067

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-b., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Tam, O. H., Rozhkov, N. V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., Propp, N., Fagegaltier, D., Harris, B. T., Ostrow, L. W., Phatnani, H., Ravits, J., Dubnau, J., Gale Hammell, M., Phatnani, H., Kwan, J., Sareen, D., Broach, J. R., Simmons, Z., … Harris, B. T. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Reports*, *29*(5), 1164–1177.e5. https://doi.org/10.1016/j.celrep.2019.09.066

Tanudisastro, H. A., Deveson, I. W., Dashnow, H., & MacArthur, D. G. (2024). Sequencing and characterizing short tandem repeats in the human genome. *Nature Reviews Genetics*, *25*(7), 460–475. https://doi.org/10.1038/s41576-024-00692-3

Teive, H. A. G., Meira, A. T., Camargo, C. H. F., & Munhoz, R. P. (2019). The Geographic Diversity of Spinocerebellar Ataxias (SCAs) in the Americas: A Systematic Review. *Movement Disorders Clinical Practice*, *6*(7), 531–540. https://doi.org/10.1002/mdc3.12822

Temiz, P., Weihl, C. C., & Pestronk, A. (2009). Inflammatory myopathies with mitochondrial pathology and protein aggregates. *Journal of the Neurological Sciences*, *278*(1-2), 25–29. https://doi.org/10.1016/j.jns.2008.11.010

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The 100,000 Genomes Project Pilot Investigators, Smedley, D., Smith, K. R., Martin, A., Thomas, E. A., McDonagh, E. M., Cipriani, V., Ellingford, J. M., Arno, G., Tucci, A., Vandrovcova, J., Chan, G., Williams, H. J., Ratnaike, T., Wei, W., Stirrups, K., Ibanez, K., Moutsianas, L., Wielscher, M., … Caulfield, M. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *New England Journal of Medicine*, *385*(20), 1868–1880. https://doi.org/10.1056/NEJMoa2035790

Tian, R., Gachechiladze, M. A., Ludwig, C. H., Laurie, M. T., Hong, J. Y., Nathaniel, D., Prabhu, A. V., Fernandopulle, M. S., Patel, R., Abshari, M., Ward, M. E., & Kampmann, M. (2019). CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron*, *104*(2), 239–255.e12. https://doi.org/10.1016/j.neuron.2019.07.014

Ticozzi, N., Vance, C., LeClerc, A. L., Keagle, P., Glass, J. D., McKenna-Yasek, D., Sapp, P. C., Silani, V., Bosco, D. A., Shaw, C. E., Brown, R. H., & Landers, J. E. (2011). Mutational analysis reveals the FUS homolog TAF15 as a candidate gene for familial amyotrophic lateral sclerosis. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, *156*(3), 285–290. https://doi.org/10.1002/ajmg.b.31158

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Župunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., & Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience*, *14*(4), 452–458. https://doi.org/10.1038/nn.2778

Udd, B., Juvonen, V., Hakamies, L., Nieminen, A., Wallgren-Pettersson, C., Cederquist, K., & Savontaus, M.-L. (1998). High prevalence of Kennedy's disease in Western Finland - is the syndrome underdiagnosed? *Acta Neurologica Scandinavica*, *98*(2), 128–133. https://doi.org/10.1111/j.1600-0404.1998.tb01732.x

Van Deerlin, V. M., Leverenz, J. B., Bekris, L. M., Bird, T. D., Yuan, W., Elman, L., Clay-falcone, D., Wood, E. M. C., Chen-Plotkin, A. S., Martinez-Lage, M., Steinbart, E., McCluskey, L., Grossman, M., Neumann, M., Wu, I. L., Yang, W. S., Kalb, R. G., Galasko, D. R., Montine, T. J., ... Yu, C. E. (2008). TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: A genetic and histopathological analysis. *The Lancet Neurology*, *7*(5), 409–416. https://doi.org/10.1016/S1474-4422(08)70071-1

Van Der Ende, E. L., Jackson, J. L., White, A., Seelaar, H., Van Blitterswijk, M., & Van Swieten, J. C. (2021). Unravelling the clinical spectrum and the role of repeat length in *C9ORF72* repeat expansions. *Journal of Neurology, Neurosurgery & Psychiatry*, *92*(5), 502–509. https://doi.org/10.1136/jnnp-2020-325377

Van Der Sanden, B. P., Corominas, J., De Groot, M., Pennings, M., Meijer, R. P., Verbeek, N., Van De Warrenburg, B., Schouten, M., Yntema, H. G., Vissers, L. E. L., Kamsteeg, E.-J., & Gilissen, C. (2021). Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. *Genetics in Medicine*, *23*(8), 1569–1573. https://doi.org/10.1038/s41436-021-01174-1

Van Es, M. A., Hardiman, O., Chio, A., Al-Chalabi, A., Pasterkamp, R. J., Veldink, J. H., & Van Den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *The Lancet*, *390*(10107), 2084–2098. https://doi.org/10.1016/S0140-6736(17)31287-4

Van Es, M. A., Veldink, J. H., Saris, C. G. J., Blauw, H. M., Van Vught, P. W. J., Birve, A., Lemmens, R., Schelhaas, H. J., Groen, E. J. N., Huisman, M. H. B., Van Der Kooi, A. J., De Visser, M., Dahlberg, C., Estrada, K., Rivadeneira, F., Hofman, A., Zwarts, M. J., Van Doormaal, P. T. C., Rujescu, D., ... Van Den Berg, L. H. (2009). Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nature Genetics*, *41*(10), 1083–1087. https://doi.org/10.1038/ng.442

Van Mossevelde, S., Engelborghs, S., Van Der Zee, J., & Van Broeckhoven, C. (2018). Genotype–phenotype links in frontotemporal lobar degeneration. *Nature Reviews Neurology*, *14*(6), 363–378. https://doi.org/10.1038/s41582-018-0009-8

Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., Van Der Spek, R. A., Võsa, U., De Jong, S., Robinson, M. R., Yang, J., Fogh, I., Van Doormaal, P. T., Tazelaar, G. H., Koppers, M., Blokhuis, A. M., Sproviero, W., Jones, A. R., Kenna, K. P., ... Veldink, J. H. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, *48*(9), 1043–1048. https://doi.org/10.1038/ng.3622

Vance, C., Rogelj, B., Hortobagyi, T., De Vos, K. J., Nishimura, A. L., Sreedharan, J., Hu, X., Smith, B. N., Ruddy, D., Wright, P., Ganesalingam, J., Williams, K. L., Tripathi, V. B., Al-Sarraj, S., Al-Chalabi, A., Leigh, P. N., Blair, I. P., Nicholson, G. A., De Belleroche, J., ... Shaw, C. E. (2009). Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science*, *323*(5918), 1208–1211. https://doi.org/10.1126/science.1165942

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., González-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., & Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, *5*, e11752. https://doi.org/10.7554/eLife.11752

Vega, M. V., Nigro, A., Luti, S., Capitini, C., Fani, G., Gonnelli, L., Boscaro, F., & Chiti, F. (2019). Isolation and characterization of soluble human full-length TDP-43 associated with neurodegeneration. *The FASEB Journal*, *33*(10), 10780–10793. https://doi.org/10.1096/fj.201900474R

Vegezzi, E., Ishiura, H., Bragg, D. C., Pellerin, D., Magrinelli, F., Currò, R., Facchini, S., Tucci, A., Hardy, J., Sharma, N., Danzi, M. C., Zuchner, S., Brais, B., Reilly, M. M., Tsuji, S., Houlden, H., & Cortese, A. (2024). Neurological disorders caused by novel non-coding repeat expansions: Clinical features and differential diagnosis. *The Lancet Neurology*, *23*(7), 725–739. https://doi.org/10.1016/S1474-4422(24)00167-4

Vogler, T. O., Wheeler, J. R., Nguyen, E. D., Hughes, M. P., Britson, K. A., Lester, E., Rao, B., Betta, N. D., Whitney, O. N., Ewachiw, T. E., Gomes, E., Shorter, J., Lloyd, T. E., Eisenberg, D. S., Taylor, J. P., Johnson, A. M., Olwin, B. B., & Parker, R. (2018). TDP-43 and RNA form amyloid-like

myo-granules in regenerating muscle. *Nature*, *563*(7732), 508–513. https://doi.org/10.1038/s41586-018-0665-2

Wahbi, K., Porcher, R., Laforêt, P., Fayssoil, A., Bécane, H. M., Lazarus, A., Sochala, M., Stojkovic, T., Béhin, A., Leonard-Louis, S., Arnaud, P., Furling, D., Probst, V., Babuty, D., Pellieux, S., Clementy, N., Bassez, G., Péréon, Y., Eymard, B., & Duboc, D. (2018). Development and Validation of a New Scoring System to Predict Survival in Patients With Myotonic Dystrophy Type 1. *JAMA Neurology*, *75*(5), 573. https://doi.org/10.1001/jamaneurol.2017.4778

Wallace, S. E., & Bean, L. J. (2022, October). Resources for Genetics Professionals — Genetic Disorders Caused by Nucleotide Repeat Expansions and Contractions. https://www.ncbi.nlm.nih.gov/books/NBK535148/

Wang, C., Ward, M. E., Chen, R., Liu, K., Tracy, T. E., Chen, X., Xie, M., Sohn, P. D., Ludwig, C., Meyer-Franke, A., Karch, C. M., Ding, S., & Gan, L. (2017). Scalable Production of iPSC-Derived Human Neurons to Identify Tau-Lowering Compounds by High-Content Screening. *Stem Cell Reports*, *9*(4), 1221–1233. https://doi.org/10.1016/j.stemcr.2017.08.019

Wanschitz, J. V., Dubourg, O., Lacene, E., Fischer, M. B., Höftberger, R., Budka, H., Romero, N. B., Eymard, B., Herson, S., Butler-Browne, G. S., Voit, T., & Benveniste, O. (2013). Expression of myogenic regulatory factors and myo-endothelial remodeling in sporadic inclusion body myositis. *Neuromuscular Disorders*, *23*(1), 75–83. https://doi.org/10.1016/j.nmd.2012.09.003

Watts, G. D., Thomasova, D., Ramdeen, S. K., Fulchiero, E. C., Mehta, S. G., Drachman, D. A., Weihl, C. C., Jamrozik, Z., Kwiecinski, H., Kaminska, A., & Kimonis, V. E. (2007). Novel VCP mutations in inclusion body myopathy associated with Paget disease of bone and frontotemporal dementia. *Clinical Genetics*, *72*(5), 420–426. https://doi.org/10.1111/j.1399-0004.2007.00887.x

Weihl, C. C., & Mammen, A. L. (2017). Sporadic inclusion body myositis – a myodegenerative disease or an inflammatory myopathy. *Neuropathology and Applied Neurobiology*, *43*(1), 82–91. https://doi.org/10.1111/nan.12384

Weihl, C. C., Temiz, P., Miller, S. E., Watts, G., Smith, C., Forman, M., Hanson, P. I., Kimonis, V., & Pestronk, A. (2008). TDP-43 accumulation in inclusion body myopathy muscle suggests a common pathogenic mechanism with frontotemporal dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*(10), 1186–1189. https://doi.org/10.1136/jnnp.2007.131334

Wilson, D. M., Cookson, M. R., Van Den Bosch, L., Zetterberg, H., Holtzman, D. M., & Dewachter, I. (2023). Hallmarks of neurodegenerative diseases. *Cell*, *186*(4), 693–714. https://doi.org/10.1016/j.cell.2022.12.032

Wu, C. H., Fallini, C., Ticozzi, N., Keagle, P., Sapp, P. C., Piotrowska, K., Lowe, P., Koppers, M., McKenna-Yasek, D., Baron, D. M., Kost, J. E., Gonzalez-Perez, P., Fox, A. D., Adams, J., Taroni, F., Tiloca, C., LeClerc, A. L., Chafe, S. C.,

Mangroo, D., ... Landers, J. E. (2012). Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis [Publisher: Nature Publishing Group]. *Nature*, *488*(7412), 499–503. https://doi.org/10.1038/nature11280

Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C. C., Soon, W. W. J., Bertin, N., Lin, C. B., Khor, C. C., DeGiorgio, M., Cheng, S., Bao, L., Karnani, N., Hwang, W. Y. K., Davila, S., Tan, P., Shabbir, A., Moh, A., ... Zhao, Y. (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell*, *179*(3), 736–749.e15. https://doi.org/10.1016/j.cell.2019.09.019

Xie, F., Wang, J., & Zhang, B. (2023). RefFinder: A web-based tool for comprehensively analyzing and identifying reference genes. *Functional & Integrative Genomics*, *23*(2), 125. https://doi.org/10.1007/s10142-023-01055-7

Yamashita, S., Kimura, E., Zhang, Z., Tawara, N., Hara, K., Yoshimura, A., Takashima, H., & Ando, Y. (2019). Muscle pathology of hereditary motor and sensory neuropathy with proximal dominant involvement with *TFG* mutation. *Muscle & Nerve*, *60*(6), 739–744. https://doi.org/10.1002/mus.26683

Yang, Y., Hentati, A., Deng, H. X., Dabbagh, O., Sasaki, T., Hirano, M., Hung, W. Y., Ouahchi, K., Yan, J., Azim, A. C., Cole, N., Gascon, G., Yagmour, A., Ben-Hamida, M., Pericak-Vance, M. A., Hentati, F., & Siddique, T. (2001). The gene encoding alsin, a protein with three guanine-nucleotide exchange factor domains, is mutated in a form of recessive amyotrophic lateral sclerosis. *Nature Genetics*, *29*(2), 160–165. https://doi.org/10.1038/ng1001-160

Yin, R., Ribeiro-Filho, H. V., Lin, V., Gowthaman, R., Cheung, M., & Pierce, B. G. (2023). TCRmodel2: High-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Research*, *51*(W1), W569–W576. https://doi.org/10.1093/nar/gkad356

Zampatti, S., Peconi, C., Campopiano, R., Gambardella, S., Caltagirone, C., & Giardina, E. (2022). C9orf72-Related Neurodegenerative Diseases: From Clinical Diagnosis to Therapeutic Strategies. *Frontiers in Aging Neuroscience*, *14*, 907122. https://doi.org/10.3389/fnagi.2022.907122

Zanovello, M., Ibáñez, K., Brown, A.-L., Sivakumar, P., Bombaci, A., Santos, L., Van Vugt, J. J. F. A., Narzisi, G., Karra, R., Scholz, S. W., Ding, J., Gibbs, J. R., Chiò, A., Dalgard, C., Weisburd, B., The American Genome Center (TAGC) consortium, Genomics England Research Consortium, Project MinE ALS Sequencing Consortium, The NYGC ALS Consortium, Ambrose, J. C., Arumugam, P., Bevers, R., ... Tucci, A. (2023). Unexpected frequency of the pathogenic *AR* CAG repeat expansion in the general population. *Brain*, *146*(7), 2723–2729. https://doi.org/10.1093/brain/awad050

Zanovello, M., Sorarù, G., Campi, C., Anglani, M., Spimpolo, A., Berti, S., Bussè, C., Mozzetta, S., Cagnin, A., & Cecchin, D. (2021). Brainstem glucose hypermetabolism in ALS/FTD and shorten survival: A^{18} F-FDG PET/MR study.

Journal of Nuclear Medicine, jnumed.121.262232. https://doi.org/10.2967/jnumed.121.262232

Zelinkova, H., Kolejakova, K. L., Spalek, P., Chandoga, J., Konkolova, J., & Bohmer, D. (2016). Molecular diagnosis of spinal and bulbar muscular atrophy in Slovakia. *Bratislava Medical Journal*, *116*(03), 137–141. https://doi.org/10.4149/BLL_2016_026

Zeng, Y., Lovchykova, A., Akiyama, T., Liu, C., Guo, C., Jawahar, V. M., Sianto, O., Calliari, A., Prudencio, M., Dickson, D. W., Petrucelli, L., & Gitler, A. D. (2024, January). TDP-43 nuclear loss in FTD/ALS causes widespread alternative polyadenylation changes. https://doi.org/10.1101/2024.01.22.575730

Zhang, Q., Bethmann, C., Worth, N. F., Davies, J. D., Wasner, C., Feuer, A., Ragnauth, C. D., Yi, Q., Mellad, J. A., Warren, D. T., Wheeler, M. A., Ellis, J. A., Skepper, J. N., Vorgerd, M., Schlotter-Weigel, B., Weissberg, P. L., Roberts, R. G., Wehnert, M., & Shanahan, C. M. (2007). Nesprin-1 and -2 are involved in the pathogenesis of Emery–Dreifuss muscular dystrophy and are critical for nuclear envelope integrity. *Human Molecular Genetics*, *16*(23), 2816–2833. https://doi.org/10.1093/hmg/ddm238

Zhang, S.-Q., Ma, K.-Y., Schonnesen, A. A., Zhang, M., He, C., Sun, E., Williams, C. M., Jia, W., & Jiang, N. (2018). High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nature Biotechnology*, *36*(12), 1156–1159. https://doi.org/10.1038/nbt.4282

Zhang, Y., Lee, Y., & Han, K. (2019). Neuronal function and dysfunction of CYFIP2: From actin dynamics to early infantile epileptic encephalopathy. *BMB Reports*, *52*(5), 304–311. https://doi.org/10.5483/BMBRep.2019.52.5.097

Zhao, W., Zhang, S., Zhu, Y., Xi, X., Bao, P., Ma, Z., Kapral, T. H., Chen, S., Zagrovic, B., Yang, Y. T., & Lu, Z. J. (2022). POSTAR3: An updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Research*, *50*(D1), D287–D294. https://doi.org/10.1093/nar/gkab702

Zimmermann, M., Mengel, D., Raupach, K., Haack, T., Neumann, M., & Synofzik, M. (2025). Frequency and neuropathology of HTT repeat expansions in FTD/ALS: Co-existence rather than causation. *Journal of Neurology*, *272*(1), 58. https://doi.org/10.1007/s00415-024-12822-2

Zweier, M., Begemann, A., McWalter, K., Cho, M. T., Abela, L., Banka, S., Behring, B., Berger, A., Brown, C. W., Carneiro, M., Chen, J., Cooper, G. M., Finnila, C. R., Guillen Sacoto, M. J., Henderson, A., Hüffmeier, U., Joset, P., Kerr, B., Lesca, G., … Rauch, A. (2019). Spatially clustering de novo variants in CYFIP2, encoding the cytoplasmic FMRP interacting protein 2, cause intellectual disability and seizures. *European Journal of Human Genetics*, *27*(5), 747–759. https://doi.org/10.1038/s41431-018-0331-z

Appendix

Appendix Tables

Appendix Table 1. Demographics for the *AR* work.
Appendix Table 2. WGS data for the *AR* work.
Appendix Table 3. WGS data for the RED work.
Appendix Table 4. Demographics for the RED work.
Appendix Table 5. Primers for the CE targeted RNA-seq panel.
Appendix Table 6. Proteomics normalised counts used to generate Figure 45.

Appendix Figures

Appendix Figure 1. Haplotype analysis for the region around the CAG repeat in the *AR* gene.

Appendix Figure 2. Distribution of normalised counts from the proteomics experiment.

Appendix Table 1. Total number of ExpansionHunter calls before and after visual QC in each cohort assessed, with threshold at 38 and 37 respectively, plus demographics

Cohort	ExpansionHunter version	Gender	Phenotype	Total participants	Total X chromosomes	Total EH calls ≥38 before visual QC	Total EH calls ≥38 after visual QC	X chromosome frequency ≥38 (95% C.I.)	Total EH calls ≥37 before visual QC	Total EH calls ≥37 after visual QC	X chromosome frequency ≥37 (95% C.I.)	Median age (1st-3rd Q)	Ethnicity
100k GP	EHv2.5	M	Non-neuro	13,072	13,072	4	2	1/6,536 (1,793-23,833)	7	7	1/1,867 (905-3,855)	47 (17-65)	AFR 1589, AFR-AMR 257, AFR-ASI 5, AFR-EUR 4, AMR 476, AMR-ASI 166, AMR-EAS 17, AMR-EUR 1566, ASI 3385, EAS 345, EUR 35828, EUR-EAS 6
		F	All	20,400	40,800	17	11	1/3,709 (2,071-6,642)	24	17	1/2,400 (1,499-3,844)		
		Combined		33,472	53,872	21	13	1/4,144 (2,422-7,090)	31	24	1/2,245 (1,509-3,340)		
GNOMAD	EHv3.2	M	All	14,947	14,947	8	5	1/2,989 (1,277-6,998)	8	7	1/2,135 (1,035-4,408)	NA	nfe: 12766, afr: 9261, oth: 395, amr: 2064, asj: 694, fin: 3188, eas: 313, sas: 203, ami: 179
		F	All	14,116	28,232	16	11	1/2,567 (1,433-4,596)	22	16	1/1,765 (1,086-2,866)		
		Combined		29,063	43,179	24	16	1/2,699 (1,661-4,384)	30	23	1/1,877 (1,251-2,817)		
NIH	EHv3.0	M	Ctrl	1,529	1,529	7	1	1/1,529 (271-8,661)	10	2	1/765 (210-2,787)	74 (63-83)	Europe: n=2286, North America: n=3233, Not known: n=1186
		F	All	5,176	10,352	4	2	1/5,176 (1,420-18,874)	10	3	1/3,451 (1,174-10,146)		
		Combined		6,705	11,881	11	3	1/3,960 (1,347-11,645)	20	5	1/2,376 (1,015-5,563)		
MINE	EHv3.1	M	Ctrl	1,272	1,272	2	2	1/636 (175-2,319)	2	2	1/636 (175-2,319)	63 (56-70)	Europe: n=2435, South Asian: n=7, West Asian: n=4, Mixed American: n=1, Not known: n=2631
		F	All	3,765	7,530	6	3	1/2,510 (854-7,380)	10	8	1/941 (477-1,857)		
		Combined		5,037	8,802	8	5	1/1,760 (752-4,121)	12	10	1/880 (478-1,620)		
SUMMARY		M		30,820	30,820	21	10	1/3,082 (1,674-5,674)	27	18	1/1,712 (1,083-2,707)		
		F		43,457	86,914	43	27	1/3,219 (2,213-4,683)	66	44	1/1,975 (1,472-2,651)		
		Combined		**74,277**	**117,734**	64	**37**	**1/3,182 (2,309-4,386)**	93	62	1/1,899 (1,482-2,434)		

Appendix Table 2. Sequencing data

Cohort	Chemistry	Instrument model	Read length	Coverage (average)	Genome build	Aligner	EH version	Total samples
100,000 Genomes Project	Truseq PCR-free	HiSeq X	2x150bp	36x	38	Illumina Isaac	2.5.5	56595
	Truseq PCR-free	HiSeq X	2x150bp	36x	37	Illumina Isaac	2.5.5	9840
	Truseq PCR-free	HiSeq X	2x125bp	35x	37	Illumina Isaac	2.5.5	8600
Project NIH	Truseq PCR-free	HiSeq X	2x150bp	35x	38	BWA mem	3.0.1	6705
Project MinE	PCR free	HiSeq2000	100bp - paired end	43x	37	Illumina Isaac	3.1.2	1177
	PCR free	HiSeqX	100bp - paired end	37x	37	Illumina Isaac	3.1.2	3902
GnomAD	670 PCR-plus, 10 PCR-free, 4 Unknown	BWA mem	2x 100bp (2%)	24x	38	BWA mem	3.2.2	668
	2063 PCR-plus, 7554 PCR-free, 18762 Unknown	BWA mem	2x 150bp	24x	38	BWA mem	3.2.2	28395

Appendix Table 3. Summary of the sequencing data used in this study.

Cohort	Chemistry	Instrument model	Read length	Genome build	EH version	Total samples
100K GP	Truseq PCR-free	HiSeq X	2x150bp	38	EHv3.2.2	34 190
TOPMed	PCR-Free	HiSeq X Ten	2x150bp	38	EHv3.2.2	47 986
1K GP3	PCR-Free	Illumina NovaSeq 6000	2x150bp	38	EHv3.2.2	2 504

Appendix Table 4. Demographics among WGS datasets.

Cohort	Total number unrelated genomes (neuro excluded)	Age			Sex Female:male	Super-populations				
		Mean	Median	Q1-Q3		AFR	AMR	EAS	EUR	SAS
100K GP	34 190	50 9	50	39-65	53.5%:46.5%	1211	636	294	29497	2552
TOPMed	47 986	60 88	62	53-70	63.5%:36.5%	11575	5038	972	30071	330
joint_100KGP_TOPMed	82 176	55 9	61	49-70	58.5%:41.5%	12786	5674	1266	59568	2882
1K GP3	2 504	> 18 years			50.8%:49.2%	661	347	504	503	489

Appendix Table 5

targets	primer_category	primer_sequence
AGRN	forward	CACGACGCTCTTCCGATCTGATGCTCAACTCCAGCCTCATG
AGRN	reverse_CE	CAGACGTGTGCTCTTCCGATCTTGCTGCCACCACCTTCAG
AGRN	reverse_annotated	CAGACGTGTGCTCTTCCGATCTAAGCCGCACAGCATTCCC
IGSF21	forward	CACGACGCTCTTCCGATCTAGTGTAACTTCAAGACAGATGGGC
IGSF21	reverse_CE	ACGTGTGCTCTTCCGATCTTCGTTCATTTCTTCATTCATAAACACC
IGSF21	reverse_annotated	ACGTGTGCTCTTCCGATCTTGTGAGTAGTTGGTGGAGAACATG
RAP1GAP	forward	TACACGACGCTCTTCCGATCTGTCCAGTGCCAGCAGCTTC
RAP1GAP	reverse_CE	ACGTGTGCTCTTCCGATCTCCACTGACATAAATCAGGCATGG
RAP1GAP	reverse_annotated	ACGTGTGCTCTTCCGATCTCACGTGCTATAGATGAAGGAGTCC
DLGAP3	forward	CCTACACGACGCTCTTCCGATCTGCCGGCACCATGTAACC
DLGAP3	reverse_CE	ACGTGTGCTCTTCCGATCTCACAGACCCACACAGTGACAC
DLGAP3	reverse_annotated	ACGTGTGCTCTTCCGATCTTCCCATCCTTGATGTCAGGATC
DLGAP3	reverse_CE	GACGTGTGCTCTTCCGATCTAGAGGGAGAGGGAGAGGAAG
TFAP2E	forward	CTACACGACGCTCTTCCGATCTCCTCCCGAGTGCCTCAAC
TFAP2E	reverse_CE	ACGTGTGCTCTTCCGATCTCACCATCCTCAGACTCATCCAAG
TFAP2E	reverse_annotated	ACGTGTGCTCTTCCGATCTAGGTTGAGCCCAATCTTCTCTAAC
GPSM2	forward	CACGACGCTCTTCCGATCTCAGCTTATAATATGACTCGATGGAGG
GPSM2	reverse_CE	ACGTGTGCTCTTCCGATCTTCATACACACACAACACACACTCAC
GPSM2	reverse_annotated	ACGTGTGCTCTTCCGATCTCTCCTGATTTACATAGACGTTCCC
RNF115	forward	CACGACGCTCTTCCGATCTACTACCGGAATATATATGTCCCAG
RNF115	reverse_CE	ACGTGTGCTCTTCCGATCTAAGCTCATGCACTCACTCTCTCTC
RNF115	reverse_annotated	ACGTGTGCTCTTCCGATCTGCTCTGCAAAATGTGTTGTTGTGG
AKT3	forward	CACGACGCTCTTCCGATCTAGACACAGTTTCTTCTCTGGAGTAAAC
AKT3	reverse_CE	ACGTGTGCTCTTCCGATCTGTACTTCCTGTTAGCCTGGTAAAC
AKT3	reverse_annotated	ACGTGTGCTCTTCCGATCTTCTAGTATCTGTCTCAGATGTTAC
AKT3	reverse_CE	GACGTGTGCTCTTCCGATCTCACTCCAACCTGGGCATTAC
KIF26B	forward	CACGACGCTCTTCCGATCTCGAGACTTCCACAGGCACATC
KIF26B	reverse_CE	ACGTGTGCTCTTCCGATCTACACACATTACCTAGGTCTGGG
KIF26B	reverse_annotated	ACGTGTGCTCTTCCGATCTAAGAAGTGGAAGGCCGATGTTTC
ADARB2	forward	CACGACGCTCTTCCGATCTAACCTTCCTCGCTCCTTTCAAG
ADARB2	reverse_CE	ACGTGTGCTCTTCCGATCTGAAGGGAGGGCTGATTCATTTG
ADARB2	reverse_annotated	CAGACGTGTGCTCTTCCGATCTTTGCCCACGTTGCGGTTC
PFKP	forward	CACGACGCTCTTCCGATCTGAGGCTCAAACATCGCAGAGG
PFKP	reverse_CE	AGACGTGTGCTCTTCCGATCTCCCATGAGCGTCTTCAGCG
PFKP	reverse_annotated	AGACGTGTGCTCTTCCGATCTTTGCAAGCAGCCTTCAGGC
PFKP	forward	CACGACGCTCTTCCGATCTGGAGGCTCAAACATCGCAGAG
PFKP	reverse_CE	ACGTGTGCTCTTCCGATCTAAATTCCCTCCACCTCTCTTCTG
GOLGA7B	forward	CACGACGCTCTTCCGATCTAGCAGAATGAGAAGATCTTTGCACC
GOLGA7B	reverse_CE	ACGTGTGCTCTTCCGATCTAAGAAGTCACAGACCCTCAGGTAG
GOLGA7B	reverse_CE	ACGTGTGCTCTTCCGATCTGCAGAGAAAGCATCACAGAGCTTC

PSD	forward	CACGACGCTCTTCCGATCTGCCTCTCCTCTGTCTCTCCTC
PSD	reverse_CE	ACGTGTGCTCTTCCGATCTGCCAGGGACATTCTTTTCTTCATAC
PSD	reverse_annotated	CAGACGTGTGCTCTTCCGATCTCAGAGCCCTGCCTTGAGC
PSD	reverse_annotated	TCAGACGTGTGCTCTTCCGATCTGCCAGGCGGGGAGTAAG
NSMCE4A	forward	CACGACGCTCTTCCGATCTACTGAATGCCGGTGACAAATTAAC
NSMCE4A	reverse_CE	ACGTGTGCTCTTCCGATCTCTTTGCCACCTCCATTGTGTTAC
NSMCE4A	reverse_annotated	ACGTGTGCTCTTCCGATCTTTGCTTTCTCTTTGCCCAAATCTG
SYT7	forward	CACGACGCTCTTCCGATCTACGCTACAAGAATTCCTTGGAGAC
SYT7	reverse_CE	ACGTGTGCTCTTCCGATCTAACTCTATCCCACAGGACAACG
SYT7	reverse_annotated	CAGACGTGTGCTCTTCCGATCTTTCCCGCTCCACCGCCAG
SYT7	reverse_CE	GACGTGTGCTCTTCCGATCTACACAGAAAGGAGGCAAGCG
RSF1	forward	ACACGACGCTCTTCCGATCTCTGAGTTGCCGTTCCCTGAG
RSF1	reverse_CE	ACGTGTGCTCTTCCGATCTCGAACAGTAGTGAGTAAGGAGCAG
RSF1	reverse_annotated	ACGTGTGCTCTTCCGATCTTCTGTCTGCAGTAACAGATTTGCC
RSF1	reverse_CE	ACGTGTGCTCTTCCGATCTCAGTAGTGAGTAAGGAGCAGGAG
UBASH3B	forward	ACACGACGCTCTTCCGATCTACCATCAAGCATGGATCGGC
UBASH3B	reverse_annotated	ACGTGTGCTCTTCCGATCTCATGGGAGAATAACCAGTCACATG
ARHGAP32	forward	CACGACGCTCTTCCGATCTCTTGAAAGAGAGGGTGTTTGGTTG
ARHGAP32	reverse_CE	ACGTGTGCTCTTCCGATCTACAGACGAAAAAGCTGAGTTCCAG
ARHGAP32	reverse_annotated	ACGTGTGCTCTTCCGATCTTCCACGATGCCATATCTCTCAATG
CEP290	forward	CACGACGCTCTTCCGATCTATGAACGTACAATCAGCAGTCTTG
CEP290	reverse_CE	ACGTGTGCTCTTCCGATCTAGTCTATCCATCCATCCATCCATC
CEP290	reverse_annotated	ACGTGTGCTCTTCCGATCTTCCAGGTCAACTTCTCTTTGATCC
CEP290	reverse_CE	ACGTGTGCTCTTCCGATCTTCCATCCGTTCAGCTATACGAAAG
MED13L	forward	CACGACGCTCTTCCGATCTAACTTCGTTAGGATTGGGAAATGG
MED13L	reverse_CE	ACGTGTGCTCTTCCGATCTTCCCACACATCATGTCTTTCTCTTG
MED13L	reverse_annotated	ACGTGTGCTCTTCCGATCTTGTGCATACATTACTTTCTCCATGC
ATP8A2	forward	CACGACGCTCTTCCGATCTTTGGTGACTCTTGAGGTTGTGAAG
ATP8A2	reverse_CE	ACGTGTGCTCTTCCGATCTAAGATCACACCACTTTACTCCAGC
ATP8A2	reverse_annotated	ACGTGTGCTCTTCCGATCTTTCATTAAGGTTTGATGTCCTGGC
G2E3	forward	CACGACGCTCTTCCGATCTGGAAAAGTGGCACTGAGGCTC
G2E3	reverse_CE	ACGTGTGCTCTTCCGATCTAGTGAGTCAAGATCCCACCATTG
G2E3	reverse_annotated	ACGTGTGCTCTTCCGATCTAACACAAGCAAGGTTCTGTGAGTC
MNAT1	forward	CACGACGCTCTTCCGATCTAAAGAAGAACAACTGCAGCAGATTC
MNAT1	reverse_annotated	ACGTGTGCTCTTCCGATCTGGTAGATCTATCTTTATGCTGAGCC
SEMA6D	reverse_CE	ACGTGTGCTCTTCCGATCTCATTCCAAGGCAGGAAATCCG
SEMA6D	reverse_annotated	CAGACGTGTGCTCTTCCGATCTGCGCGCTCTGCCAGTAAG
ONECUT1	forward	ACACGACGCTCTTCCGATCTTTCCGGAGGATGTGGAAGTG
ONECUT1	reverse_CE	ACGTGTGCTCTTCCGATCTCAAGTCCTACTCATCCTCCAGATC
ONECUT1	reverse_annotated	ACGTGTGCTCTTCCGATCTCTTTTTGGGTGTGTTGCCTCTATC
LINGO1	reverse_CE	ACGTGTGCTCTTCCGATCTGGCAGGTCAGACATCCAGAAG
LINGO1	reverse_annotated	CAGACGTGTGCTCTTCCGATCTTCCTCCTCCGACACCTCC

LINGO1	reverse_CE	ACGTGTGCTCTTCCGATCTGCAGGTCAGACATCCAGAAGG
ZNF423	forward	CACGACGCTCTTCCGATCTATCACTGTCAGCAGGACTTCGAG
ZNF423	reverse_CE	ACGTGTGCTCTTCCGATCTTACAAGGCTCTCCACACAAGC
ZNF423	reverse_annotated	GACGTGTGCTCTTCCGATCTAACATCCTTGCTGGAGGGAG
AARS	forward	CACGACGCTCTTCCGATCTTGGCTTAGAAGCAGATCTGGAATG
AARS	reverse_CE	ACGTGTGCTCTTCCGATCTATCACTTGAGCCCAGGAAGTTG
AARS	reverse_annotated	ACGTGTGCTCTTCCGATCTCCCAGAAGTTATCCTTCATGTTGC
NECAB2	forward	CACGACGCTCTTCCGATCTAATCAGATCCAGTCGCTGCTGAG
NECAB2	reverse_CE	ACGTGTGCTCTTCCGATCTCTCAATGAATGACCAAATCCTCCC
NECAB2	reverse_annotated	GACGTGTGCTCTTCCGATCTGCTTCCACACCAGGTCTGTG
KIAA0753	forward	CACGACGCTCTTCCGATCTATCATATCTCATGAGGCTGTAGGC
KIAA0753	reverse_CE	ACGTGTGCTCTTCCGATCTCTTCATTTTCTGAACGGCGCAAG
KIAA0753	reverse_annotated	ACGTGTGCTCTTCCGATCTTTCGCACATATCCTGAAGTTCAGC
MYO18A	forward	CACGACGCTCTTCCGATCTATTCCAGAGCTCAGCGAGCTC
MYO18A	reverse_CE	ACGTGTGCTCTTCCGATCTCCAGGCTTTGGCAGAGTTTTG
MYO18A	reverse_annotated	ACGTGTGCTCTTCCGATCTGTCTCATTCCAGGCCTCTTCTG
TSPOAP1	forward	ACACGACGCTCTTCCGATCTCTGCTGGGAGACAGCAACTG
TSPOAP1	reverse_CE	GACGTGTGCTCTTCCGATCTTTGTGTGGGCTGTGGAGTAC
TSPOAP1	reverse_annotated	ACGTGTGCTCTTCCGATCTGAGCTTTCAGGTCAGCAGAGG
TSPOAP1	reverse_CE	ACGTGTGCTCTTCCGATCTGTCTTACCCACCTTTCTTGGAG
TSPOAP1	reverse_annotated	ACGTGTGCTCTTCCGATCTGGAGCTTTCAGGTCAGCAGAG
MGAT5B	forward	ACACGACGCTCTTCCGATCTAGAAGCGGCTCATCAAAGGC
MGAT5B	reverse_CE	ACGTGTGCTCTTCCGATCTCCTCTGATGGAAGCTCCATATTTG
MGAT5B	reverse_annotated	ACGTGTGCTCTTCCGATCTGTGCCATGGATCTCCATGTATTTG
DLGAP1	forward	CACGACGCTCTTCCGATCTAAGCCTGCAATTTGGATCTTTTCC
DLGAP1	reverse_CE	ACGTGTGCTCTTCCGATCTCCATCCATCCATCCTTTATCCATC
DLGAP1	reverse_annotated	ACGTGTGCTCTTCCGATCTCAATCAGGAATTCTCTCGAGCAGAC
PRELID3A	forward	CACGACGCTCTTCCGATCTATTGAACACTCTGAAAGCGCTGTG
PRELID3A	reverse_CE	ACGTGTGCTCTTCCGATCTCTCCTTACAGGCACTGAACTCAG
PRELID3A	reverse_annotated	GACGTGTGCTCTTCCGATCTTCCTTGCTCGTCCTCTCCTC
ZNF521	forward	CACGACGCTCTTCCGATCTATCAATGCATCAAGTGTCAGATGG
ZNF521	reverse_CE	ACGTGTGCTCTTCCGATCTAGTTCGCGCATATGTGTTCTCTAG
ZNF521	reverse_annotated	ACGTGTGCTCTTCCGATCTTTTGGCAGGAGAGTCAAAGGTC
CELF5	forward	CACGACGCTCTTCCGATCTTACGAGCTCACGGTGCTCAAAG
CELF5	reverse_CE	ACGTGTGCTCTTCCGATCTTTCCTTATGGCTTCCGTGATCTC
CELF5	reverse_annotated	ACGTGTGCTCTTCCGATCTTCTGAGCTTTGATGGCGGAATC
CELF5	forward	CACGACGCTCTTCCGATCTTTTCGTGAAGTTCTCCTCCCAC
CELF5	reverse_CE	GACGTGTGCTCTTCCGATCTGACTGCGCCTCTTTCTCCAC
CELF5	reverse_annotated	AGACGTGTGCTCTTCCGATCTCTTGTCCGTGTCGGCGAAC
CELF5	reverse_CE	GACGTGTGCTCTTCCGATCTACGTGTATACCCACTCCAGG
HDGFL2	forward	CACGACGCTCTTCCGATCTACTTCACACCTGAGAAGAAAGCAG
HDGFL2	reverse_CE	ACGTGTGCTCTTCCGATCTTGATGCCAGCATCCCAGAATG

HDGFL2	reverse_annotated	CAGACGTGTGCTCTTCCGATCTTCCGAATCGGCCTTGGAG
HDGFL2	reverse_CE	GACGTGTGCTCTTCCGATCTTGCTTCCCTCCCTTCTGATG
INSR	forward	CACGACGCTCTTCCGATCTAAGAAGTTTCAGGAACCAAGGG
INSR	reverse_CE	ACGTGTGCTCTTCCGATCTCATCCATTCATTCAGTGCTTGTATGAC
INSR	reverse_annotated	ACGTGTGCTCTTCCGATCTCTCCCATCTCAGCAAGATCTTGTC
FBN3	forward	CACGACGCTCTTCCGATCTCTCTGCCTACAGGAAGCTGTG
FBN3	reverse_CE	ACGTGTGCTCTTCCGATCTTCACTCACAAAGAACACACACTGG
FBN3	reverse_annotated	ACGTGTGCTCTTCCGATCTAAGGAGCCAAGGCTGTTGATG
FBN3	reverse_CE	ACGTGTGCTCTTCCGATCTTCACAAAGAACACACACTGGGC
FBN3	forward	CACGACGCTCTTCCGATCTAACACGGATGGGAGCTACAAG
FBN3	reverse_CE	ACGTGTGCTCTTCCGATCTGCTGCAGTGAGCTGTGATTGTG
FBN3	reverse_annotated	ACGTGTGCTCTTCCGATCTATGACATTGACACACTGGCCATG
ELAVL3	forward	CACGACGCTCTTCCGATCTCAATGATGCAGACAAAGCCATCAAC
ELAVL3	reverse_CE	GACGTGTGCTCTTCCGATCTATCAGGTCACAGCCGGAGTC
ELAVL3	reverse_annotated	ACGTGTGCTCTTCCGATCTCGCTGACGTACAGGTTAGCATC
MAST1	forward	ACACGACGCTCTTCCGATCTGAAACTTCTCCCCCAACACC
MAST1	reverse_CE	ACGTGTGCTCTTCCGATCTTGACTTGCTCCTCTCACATATTCAC
MAST1	reverse_annotated	ACGTGTGCTCTTCCGATCTTTGGTGCCATAGCCAGATGAAG
ADGRL1	forward	CACGACGCTCTTCCGATCTTTTCCAGATGGAGAATGTGCAGTG
ADGRL1	reverse_CE	ACGTGTGCTCTTCCGATCTATTGCTTGAGGCCAGGAGTTTAAG
ADGRL1	reverse_annotated	AGACGTGTGCTCTTCCGATCTTTGTAGGTCCCAGGACAGG
UNC13A	forward	CACGACGCTCTTCCGATCTTCCACATCAGTGTGGAGATCAAAG
UNC13A	reverse_CE	ACGTGTGCTCTTCCGATCTTATTCAACAAACTAGTTCCTGGGG
UNC13A	reverse_annotated	ACGTGTGCTCTTCCGATCTATCTTCACGACCCCATTGTTCTG
UNC13A	reverse_CE	ACGTGTGCTCTTCCGATCTAAGAGACATACCCAGACACAAACG
CRTC1	forward	TACACGACGCTCTTCCGATCTACGGTACCGTCCTCTCTCC
CRTC1	reverse_CE	CAGACGTGTGCTCTTCCGATCTTTGGTGATGGTCGCCACC
CRTC1	reverse_annotated	AGACGTGTGCTCTTCCGATCTATTGGAGACTGGCGAGGTG
ZNF431	forward	CACGACGCTCTTCCGATCTTGTATCCTCTCAAGGAAGCAAGTG
ZNF431	reverse_CE	ACGTGTGCTCTTCCGATCTAATTCTGTTCTCTATGCCACTGGG
ZNF431	reverse_annotated	ACGTGTGCTCTTCCGATCTCTCCTCCAGAGAGAATTCTATGGC
GRAMD1A	forward	CCTACACGACGCTCTTCCGATCTAGCCCAGCCCTGCCCTG
GRAMD1A	reverse_CE	ACGTGTGCTCTTCCGATCTACTTGCAAGACAATTCACAGTCGG
GRAMD1A	reverse_annotated	AGACGTGTGCTCTTCCGATCTCAGGAGCTGCAGCCGTTTC
ZNF529	reverse_CE	TCAGACGTGTGCTCTTCCGATCTTCACCGCGAGCTCCTCC
ZNF529	reverse_annotated	ACGTGTGCTCTTCCGATCTAAACTTGAGTTGGCCATTAGTACC
ZNF527	forward	ACACGACGCTCTTCCGATCTGGATTCTGGGAGGTGACGTC
ZNF527	reverse_annotated	ACGTGTGCTCTTCCGATCTACAGCCATTCCTCCTTCTTCAG
POU2F2	forward	CCTACACGACGCTCTTCCGATCTGCGGGGCGGGCAGCATG
POU2F2	reverse_CE	ACGTGTGCTCTTCCGATCTTGTATTAATTCATGTCCAGGCAGG
POU2F2	reverse_annotated	ACGTGTGCTCTTCCGATCTTTTCGGTGTCTGTGTGCTCTGATG
POLD1	forward	CCTACACGACGCTCTTCCGATCTCGGGAGTCAGGGGTCAC

POLD1	reverse_CE	ACGTGTGCTCTTCCGATCTCCAAACTGAATGATCATTGGCTCC
POLD1	reverse_annotated	TCAGACGTGTGCTCTTCCGATCTCCCACGGGCCCGCTTTG
ZNF583	forward	CCTACACGACGCTCTTCCGATCTCAAGGACCGCGCCGAAG
ZNF583	reverse_CE	ACGTGTGCTCTTCCGATCTTCTTTCCACCCGATCATGATGTTG
ZNF583	reverse_annotated	ACGTGTGCTCTTCCGATCTGTTCTTCTGTCCTACTCCTTCAGC
PXDN	forward	CACGACGCTCTTCCGATCTTCATTCAGGAGCATGTACAGCATG
PXDN	reverse_CE	ACGTGTGCTCTTCCGATCTTAAAGTTTCTCCAGCCCATTTCTG
PXDN	reverse_annotated	ACGTGTGCTCTTCCGATCTTTTGCGATGAGGTTCAGGTACTG
TRAPPC12	forward	CACGACGCTCTTCCGATCTCAATTTGGAGCAAGGCTTAGCAG
TRAPPC12	reverse_CE	ACGTGTGCTCTTCCGATCTGAAAGGAGTGAGTTGCAAATGCAC
TRAPPC12	reverse_annotated	ACGTGTGCTCTTCCGATCTAGACAGTTTGCCATGGAGTACATC
TRAPPC12	reverse_CE	ACGTGTGCTCTTCCGATCTACATCGAGTAGGACAGAAGCTGTG
TRAPPC12	reverse_CE	ACGTGTGCTCTTCCGATCTCTCACACACGTTCAGCTCACAC
WDR35	forward	CACGACGCTCTTCCGATCTCTGACGGACAGAAGATCTGCATTG
WDR35	reverse_CE	ACGTGTGCTCTTCCGATCTTGATAGCTGTTGCTTTATAGGCTC
WDR35	reverse_annotated	ACGTGTGCTCTTCCGATCTTGTTACATGGGATAGCTGTATACC
WDR35	reverse_CE	ACGTGTGCTCTTCCGATCTCTCCTGATCTAGGTCTGGTCATTG
SLC5A7	forward	TACACGACGCTCTTCCGATCTTTTCGGGTGCCTGGTGATG
SLC5A7	reverse_CE	ACGTGTGCTCTTCCGATCTACAACTGGAATGGAGAGTGTACAC
SLC5A7	reverse_annotated	ACGTGTGCTCTTCCGATCTCTCTTCTGTAGTCTTGGGATCTGG
SLC5A7	reverse_CE	ACGTGTGCTCTTCCGATCTCCTTCGTCAGTCATCTTTACATTTTTG
SLC5A7	reverse_annotated	ACGTGTGCTCTTCCGATCTCCTCTTCTGTAGTCTTGGGATCTG
SPEG	forward	CACGACGCTCTTCCGATCTCTTGCAAAGCGGTCAATGAGTATG
SPEG	reverse_CE	ACGTGTGCTCTTCCGATCTAAGCACCAAGACATGGGTTAATAC
SPEG	reverse_annotated	CAGACGTGTGCTCTTCCGATCTCACGTCCACGTCCTGCAG
SPEG	reverse_CE	ACGTGTGCTCTTCCGATCTAACCAGGCATTCATGGGATCAG
ATG4B	forward	ACACGACGCTCTTCCGATCTACGAGAGCTTCCACTGCCAG
ATG4B	reverse_CE	ACGTGTGCTCTTCCGATCTCTCTTATACCCTTCAGATTCCCAAC
ATG4B	reverse_annotated	ACGTGTGCTCTTCCGATCTCTTGCTGGCACCAATCATTGAAG
ATG4B	reverse_CE	ACGTGTGCTCTTCCGATCTTCACATCCAGACACACGCTCTC
C20orf194	forward	CACGACGCTCTTCCGATCTTTTGACACAGAAATTCCTTTCCTGC
C20orf194	reverse_CE	ACGTGTGCTCTTCCGATCTTAAATTTAATCTGCAGCTCTGATG
C20orf194	reverse_annotated	ACGTGTGCTCTTCCGATCTTAGGGCTGTTTTCAGTTATATGGC
SLC24A3	forward	CACGACGCTCTTCCGATCTTGGCCAACAATGCTGAAATTGATG
SLC24A3	reverse_CE	ACGTGTGCTCTTCCGATCTTATCCATCCATCCATCCACTCTTC
SLC24A3	reverse_annotated	ACGTGTGCTCTTCCGATCTCAGCAGCTCGTCTACCATGATC
RALGAPA2	forward	CACGACGCTCTTCCGATCTACCCAATGAAGAATCACATGTTCTTC
RALGAPA2	reverse_CE	ACGTGTGCTCTTCCGATCTGGGATAGTCAGGACAACTTAAGG
RALGAPA2	reverse_annotated	ACGTGTGCTCTTCCGATCTCAGCAGCTTCCCACTCACTATG
CDH4	forward	CACGACGCTCTTCCGATCTATGATTACACGGCATTAATCTCCC
CDH4	reverse_CE	ACGTGTGCTCTTCCGATCTAAAGCCTCCGGTTCTCTTGAC
CDH4	reverse_annotated	ACGTGTGCTCTTCCGATCTCAACTTTGAAGTCCATGCTGTTGG

DIP2A	forward	CACGACGCTCTTCCGATCTATACTTGGACTTCAGCGTGTCAAC
DIP2A	reverse_CE	GACGTGTGCTCTTCCGATCTTCCCGAAGTGCTGGGATTAC
DIP2A	reverse_annotated	GACGTGTGCTCTTCCGATCTCACAGGGACACGTTGCTCTC
DIP2A	reverse_annotated	ACGTGTGCTCTTCCGATCTTACAGCTCACACTGCAGCTTTATG
SETD5	forward	CACGACGCTCTTCCGATCTACCAGTCACATCTCTTACTACTGC
SETD5	reverse_CE	ACGTGTGCTCTTCCGATCTATAACTACCTCGGGTGACAGG
SETD5	reverse_annotated	ACGTGTGCTCTTCCGATCTACAGCTGTTAGTGTTGGAGTCAAG
CADPS	forward	CACGACGCTCTTCCGATCTTCCTGCGTGATAAGGTCAATGAG
CADPS	reverse_CE	ACGTGTGCTCTTCCGATCTGGTAAGAAGTCTGCGTCAGGTC
CADPS	reverse_annotated	ACGTGTGCTCTTCCGATCTAAATATGAAGCTGTAAGTCCATCCG
SHQ1	forward	CACGACGCTCTTCCGATCTACCCAGCGTACATACTGAATGATC
SHQ1	reverse_CE	ACGTGTGCTCTTCCGATCTCCGGCCTGTTTCTACTTTTAATTTTC
SHQ1	reverse_annotated	ACGTGTGCTCTTCCGATCTTTTGTAAGGGAGACTTCCTTTAAGG
KALRN	forward	CACGACGCTCTTCCGATCTGCTGCTGTCTTTTTTCCGTATGTC
KALRN	reverse_CE	ACGTGTGCTCTTCCGATCTCTTCAAGGTTGAAGAGCAAAAGGG
KALRN	reverse_CE	ACGTGTGCTCTTCCGATCTAATCAGACACCACTCCTCCTGAG
KALRN	reverse_CE	GACGTGTGCTCTTCCGATCTTTCCCTCAGGTTTCCCAGAG
KALRN	reverse_CE	ACGTGTGCTCTTCCGATCTGGGACCAAGGTTCACAAATAGC
KALRN	reverse_CE	ACGTGTGCTCTTCCGATCTCACACACGCACACAGAGCAAG
IFT122	forward	CACGACGCTCTTCCGATCTAAGATCTTCTGCCTCCATGTCTTC
IFT122	reverse_CE	ACGTGTGCTCTTCCGATCTAAACACACGCTGCTGACAACAC
IFT122	reverse_annotated	ACGTGTGCTCTTCCGATCTGGCTTCCTTGAACAGTTTCCTATC
ETV5	forward	CACGACGCTCTTCCGATCTTAGCTGAAGCACAAGTTCCTGATG
ETV5	reverse_CE	ACGTGTGCTCTTCCGATCTTTGGATGAAGCCAAAGTTAACTGG
ETV5	reverse_annotated	ACGTGTGCTCTTCCGATCTTACAAGACGACAGCTCAGAGGAG
CDK7	forward	CACGACGCTCTTCCGATCTAGATTTTGGCCTGGCCAAATC
CDK7	reverse_CE	ACGTGTGCTCTTCCGATCTCCAGCAGCATTTGATTTACAGACAG
CDK7	reverse_annotated	ACGTGTGCTCTTCCGATCTATGTCCACACCTACACCATACATC
CDK7	reverse_CE	ACGTGTGCTCTTCCGATCTAGCTTATCAGTCATGTCCACACTG
EPB41L4A	forward	CACGACGCTCTTCCGATCTTGGTATTAAATTCTATGCTGAAGATCC
EPB41L4A	reverse_CE	ACGTGTGCTCTTCCGATCTTTCAGCCTGTCCTTTTGTGACTTC
EPB41L4A	reverse_annotated	GACGTGTGCTCTTCCGATCTCAGCAGTGTTGACGGGACAG
ATXN1	forward	CACGACGCTCTTCCGATCTAGAGAAAGAGTGGATTTCAGCCTG
ATXN1	reverse_CE	ACGTGTGCTCTTCCGATCTCACATACACACACGTTTCATACTGC
ATXN1	reverse_annotated	ACGTGTGCTCTTCCGATCTATTTTGCTTGAGTAGAAAGGGTGG
BCKDHB	forward	CACGACGCTCTTCCGATCTGTCTGTAACAAGTGCCTTGGATAAC
BCKDHB	reverse_CE	ACGTGTGCTCTTCCGATCTTTCCTGAGACACAGCATACAGTAAG
BCKDHB	reverse_annotated	ACGTGTGCTCTTCCGATCTAGCCAACAGTGCATCTAAAGACTC
UBE3D	forward	CACGACGCTCTTCCGATCTGGAACCAAAGGCAAATACCAAAG
UBE3D	reverse_CE	ACGTGTGCTCTTCCGATCTGATGTTGCAGTGAGCCGAGATC
UBE3D	reverse_annotated	ACGTGTGCTCTTCCGATCTACCTTGGTATGATAGGAAAACTCC
SYNE1	forward	CACGACGCTCTTCCGATCTGGCTGACAACAAATACATCATTCTGC

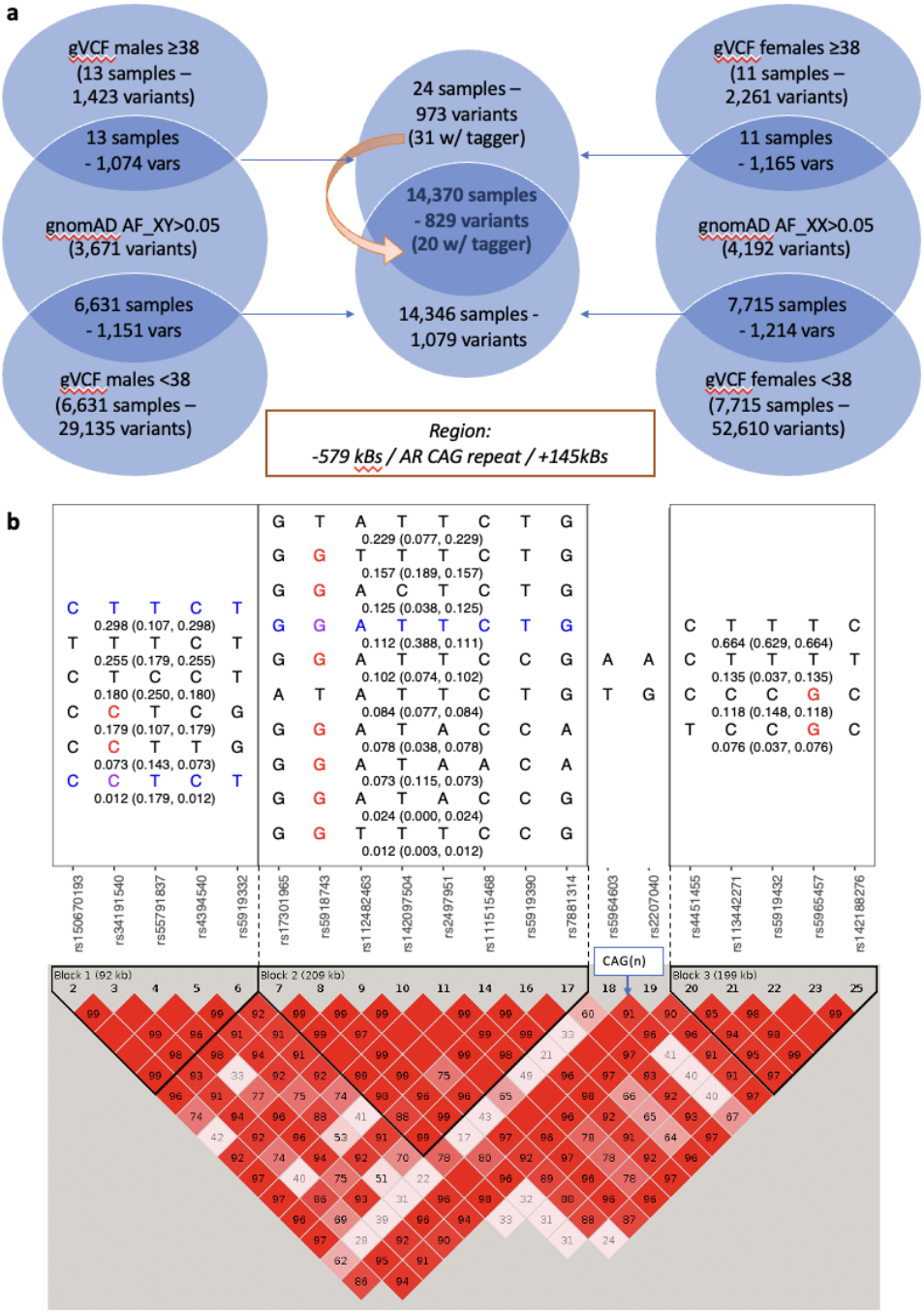
SYNE1	reverse_CE	ACGTGTGCTCTTCCGATCTTGTCAGATTATTCTCTATCGTCTATCC
SYNE1	reverse_annotated	ACGTGTGCTCTTCCGATCTCTTGTCACCACGCCTCTTTAATTC
SYNJ2	forward	CACGACGCTCTTCCGATCTTCATTAAAGGACAGTATGGCAAGC
SYNJ2	reverse_CE	ACGTGTGCTCTTCCGATCTGGTCACACAGTTGTAGCTACACTC
SYNJ2	reverse_annotated	ACGTGTGCTCTTCCGATCTGATTTCAGCATCTGGAATTCTGCC
IQCE	forward	CCTACACGACGCTCTTCCGATCTCCACCATGTTCCTGGGC
IQCE	reverse_CE	ACGTGTGCTCTTCCGATCTTGGAAGTCCAGAATCAAGGTGTTG
IQCE	reverse_annotated	ACGTGTGCTCTTCCGATCTTTTGCTTTCGTCTCCACATCAGAG
CAMK2B	forward	CCTACACGACGCTCTTCCGATCTCCCTGCCCATCTCCGAC
CAMK2B	reverse_CE	ACGTGTGCTCTTCCGATCTTTCTAATGTACAGAGACAGCTGCC
CAMK2B	reverse_annotated	ACGTGTGCTCTTCCGATCTTCAAAGTCACCGTTGTTGACGG
CAMK2B	reverse_CE	AGACGTGTGCTCTTCCGATCTTGCGCTCTGTGACCCTCAG
CAMK2B	forward	CACGACGCTCTTCCGATCTTCTGAAGCATTCCAACATCGTGC
CAMK2B	reverse_CE	ACGTGTGCTCTTCCGATCTAATGCAGCTGCTGACTGACTATTC
CAMK2B	reverse_annotated	ACGTGTGCTCTTCCGATCTGTAGTACTCTCTCGCCACAATGTC
ADCY1	forward	CACGACGCTCTTCCGATCTGAGCTGGTGAAACTCCTCAATGAG
ADCY1	reverse_CE	ACGTGTGCTCTTCCGATCTTAAAAGTATGTAAGCCAGCATCCC
ADCY1	reverse_annotated	ACGTGTGCTCTTCCGATCTGACACGCAGTAGTAGCAGTCC
TRRAP	forward	CACGACGCTCTTCCGATCTTGGTTGACCAGACCACTTTGATG
TRRAP	reverse_CE	ACGTGTGCTCTTCCGATCTGGGCTCCAGCCAGTATTACTC
TRRAP	reverse_annotated	ACGTGTGCTCTTCCGATCTACTGAGGAGATGACGTGACATTCTC
ACTL6B	forward	CACGACGCTCTTCCGATCTTACAGCAAACACGTCAAGTCTGAG
ACTL6B	reverse_CE	ACGTGTGCTCTTCCGATCTAGGCAGGAGGATTGCTTGAAC
ACTL6B	reverse_annotated	ACGTGTGCTCTTCCGATCTATGTTGTACTGCTCGAACATCAGC
PTPRN2	forward	CACGACGCTCTTCCGATCTAACAAAGACAAACTGGAGGAAACC
PTPRN2	reverse_CE	CAGACGTGTGCTCTTCCGATCTTCCGGACGGTCCCGGTAG
PTPRN2	reverse_annotated	ACGTGTGCTCTTCCGATCTATGAACTTGGTGGAGTCTTCTTGC
PTPRN2	reverse_CE	TCAGACGTGTGCTCTTCCGATCTCGGGCTAGGCACGCTGG
PTPRN2	forward	CACGACGCTCTTCCGATCTACCAGCAGTGACCTTCAAAGTG
PTPRN2	reverse_CE	ACGTGTGCTCTTCCGATCTCTTACCCATAGAAAAGCACGGAG
PTPRN2	reverse_annotated	ACGTGTGCTCTTCCGATCTCCGACTCCGGTTTGAAGAATTTTC
PTPRN2	reverse_CE	GACGTGTGCTCTTCCGATCTCCATAGAAAAGCACGGAGAG
PTPRN2	forward	CACGACGCTCTTCCGATCTACTTTTACCGCTACGAGGTGTC
PTPRN2	reverse_CE	ACGTGTGCTCTTCCGATCTTGTGAGTTGGATTGGCTGATGAG
PTPRN2	reverse_annotated	ACGTGTGCTCTTCCGATCTAAGTTCCTGGTCCATCACATACTG
STMN2	forward	CACGACGCTCTTCCGATCTACTGCTCAGCGTCTGCACATC
STMN2	reverse_CE	ACGTGTGCTCTTCCGATCTATGCTCACACAGAGAGCCAAATTC
STMN2	reverse_annotated	ACGTGTGCTCTTCCGATCTCAAGAGCAGATCAGTGACAGCATG
INTS8	forward	CACGACGCTCTTCCGATCTGGCATGGTAAATGGAGAAACTGAG
INTS8	reverse_CE	ACGTGTGCTCTTCCGATCTTACAGTTATATATAATTCCCCTTTC
INTS8	reverse_annotated	ACGTGTGCTCTTCCGATCTTCATAGACAGCTGAATTTGTGGAGC
ZFAT	forward	CACGACGCTCTTCCGATCTAGATTATTATTCCCCTTAGGCCTC

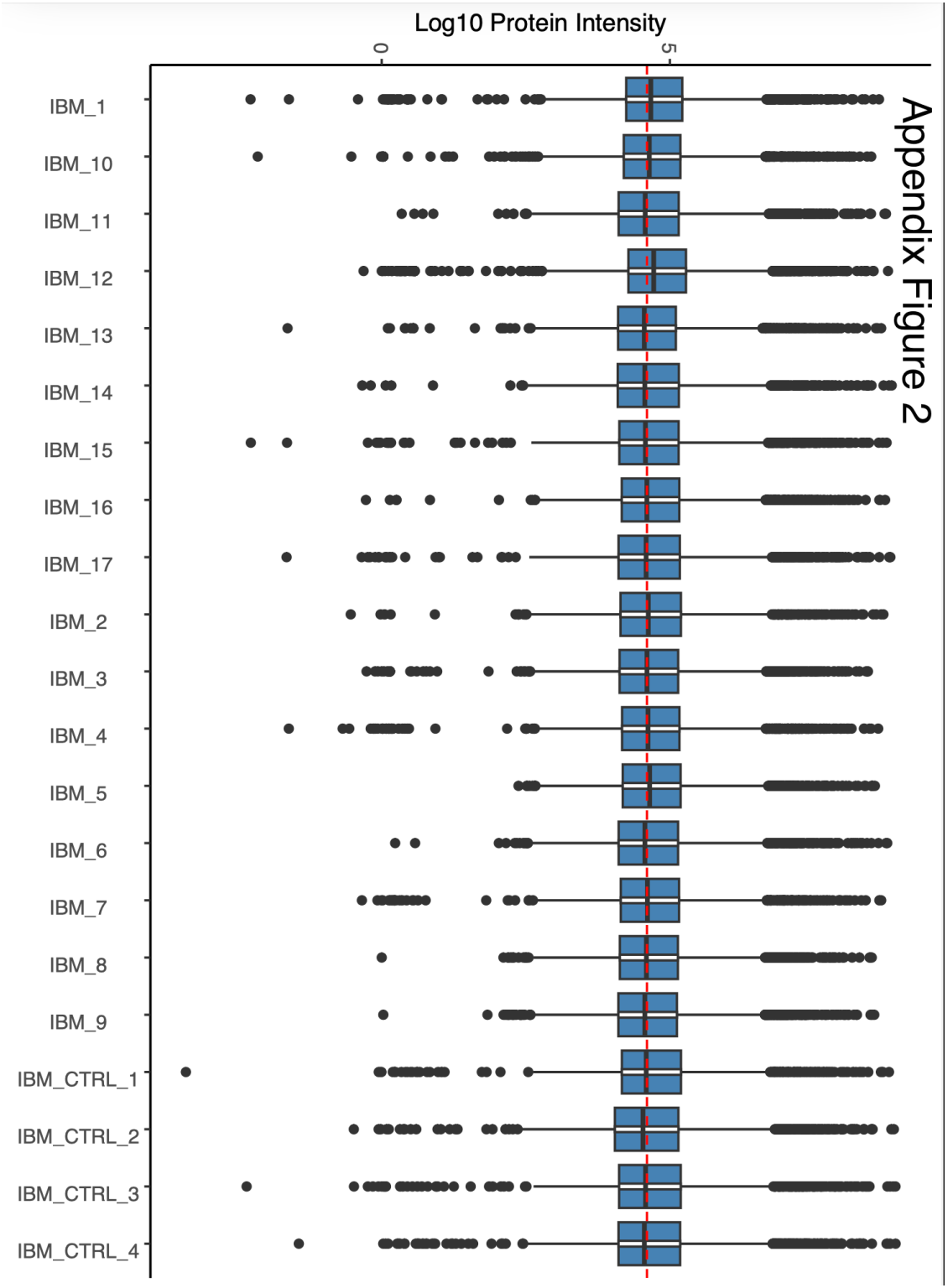
ZFAT	reverse_CE	ACGTGTGCTCTTCCGATCTGATCACACACACGAGAATGAGCTTG
ZFAT	reverse_annotated	ACGTGTGCTCTTCCGATCTCTTCTGTGCTGGACTTCTTCGTG
SCRT1	forward	TACACGACGCTCTTCCGATCTGTTCTCTTCGGCCGACCTG
SCRT1	reverse_CE	ACGTGTGCTCTTCCGATCTGACACACTCTCAAGTCACATGGTG
SCRT1	reverse_annotated	AGACGTGTGCTCTTCCGATCTCTTTGAGCAGCGCAGCCTC
PTPRD	forward	CACGACGCTCTTCCGATCTCTGTGGTTTTGAACTTCTTGAGGC
PTPRD	reverse_CE	ACGTGTGCTCTTCCGATCTCATGTGCACTTCTTTGGGATGTATG
PTPRD	reverse_annotated	ACGTGTGCTCTTCCGATCTTTTCAGCTGGAACACTTTCAGAGC
TUT7	forward	CACGACGCTCTTCCGATCTTTTATTTGTGGAAGAGAAGGGCAC
TUT7	reverse_CE	ACGTGTGCTCTTCCGATCTTGACTGGGGACTCAAAAATATTGTTG
TUT7	reverse_annotated	ACGTGTGCTCTTCCGATCTTTCCCTCATGATTCCTGCTGG
DAPK1	forward	CACGACGCTCTTCCGATCTCAGTTTATCATGACCGTGTTCAGG
DAPK1	reverse_CE	ACGTGTGCTCTTCCGATCTAGATGATAGGAAATGCATCCTCGC
DAPK1	reverse_annotated	ACGTGTGCTCTTCCGATCTTTCTTGATGAATTTGGCGGCATAC
PHF2	forward	ACACGACGCTCTTCCGATCTCAACCTGGACCTGCTCGAAG
PHF2	reverse_CE	ACGTGTGCTCTTCCGATCTCATTCATCCAGCCATTTAGCCATC
PHF2	reverse_annotated	ACGTGTGCTCTTCCGATCTGCATGTGAACCACATCTTTGTTGG
DNM1	forward	CCTACACGACGCTCTTCCGATCTTGACCCTTTCGGCCCTC
DNM1	reverse_CE	ACGTGTGCTCTTCCGATCTCACGAAATCAACATGGCAGTTCAC
DNM1	reverse_annotated	ACGTGTGCTCTTCCGATCTGAGGGATCTGTTTAGAGGTCGAAG
SPACA9	forward	ACACGACGCTCTTCCGATCTCAGATACCCGATCCTCGGTC
SPACA9	reverse_CE	ACGTGTGCTCTTCCGATCTCCAACCACTCCTCATTGTTCTAATC
SPACA9	reverse_annotated	ACGTGTGCTCTTCCGATCTTCACCTCATTCATTGTCTTCCTCTG
RAI2	forward	CACGACGCTCTTCCGATCTAGAGTTCCAGCACACTCGACC
RAI2	reverse_CE	ACGTGTGCTCTTCCGATCTCATCATTCCAGAAACATCCCTCATG
RAI2	reverse_annotated	ACGTGTGCTCTTCCGATCTAGAGAAGTCAAGACAGGCCTTATG
ABCD1	forward	CACGACGCTCTTCCGATCTTACGGTGGTGTGCTCTACAAG
ABCD1	reverse_CE	ACGTGTGCTCTTCCGATCTTCTCCTGAGGTGCTGACTCTTC
ABCD1	reverse_annotated	ACGTGTGCTCTTCCGATCTCTTTGCATGTCCTCCACTGAGTC
SPRY3	forward	CACGACGCTCTTCCGATCTGAGAAGAGATCCTTTGAGCCAGATG
SPRY3	reverse_CE	ACGTGTGCTCTTCCGATCTTAGGCAAATGACCTCCGTTCTG
SPRY3	reverse_annotated	ACGTGTGCTCTTCCGATCTGGTGGCACTTATGGTGATAAAGAG

Appendix Table 6

Genes	IBM_2	IBM_3	IBM_5	IBM_7	IBM_8	IBM_9	IBM_10	IBM_13	IBM_16	IBM_17	IBM_CTRL_1	IBM_CTRL_2	IBM_CTRL_3	IBM_CTRL_4
TRB	0.00	4679.06	0.00	24213.02	53695.91	68223.75	1066.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HLA-B	862077.06	3116475.00	865222.69	2484462.00	3111859.25	4198535.50	8673574.00	5461304.00	1474618.50	1277215.00	314991.41	142960.16	105858.72	142216.08
HLA-A	183473.41	957102.44	118309.15	873133.38	615638.38	745823.88	2252510.75	391002.66	357878.78	386216.19	72006.77	49155.58	22074.54	75176.51
CD3E	0.00	8329.99	0.00	7339.56	9422.17	4459.02	6817.65	2241.84	0.00	0.00	0.00	0.00	0.00	0.00
HLA-C	836546.19	1388416.50	167160.06	1427488.38	497653.47	546410.00	662145.31	1053735.38	754103.31	652907.25	20435.81	18140.26	79875.39	32769.24
CALR	622263.44	1515138.50	798429.19	990220.50	1012917.81	1029013.75	1463356.75	1157201.13	727150.38	604215.38	504780.81	385484.88	450560.94	406875.31
B2M	213312.83	358053.91	74585.80	351273.34	397972.81	469056.59	938201.13	474448.94	262968.53	141609.88	46107.59	36995.62	48144.86	29896.29
TAP1	103311.60	243147.52	129989.52	380490.88	170778.84	251224.09	391148.75	319413.41	247890.25	199929.28	30841.05	14853.02	8791.58	13040.71
TAP2	35267.82	75551.88	48676.25	151639.92	63707.83	96338.51	133473.92	128871.29	77948.96	75831.16	7448.05	5364.05	7414.11	3779.85
TARDBP	118483.30	224617.48	158173.11	92569.52	195615.91	370094.25	222329.59	165647.56	88307.74	83801.09	76571.06	81408.95	88001.05	80868.84
ERAP2	16327.07	6321.60	47329.65	51265.89	52771.18	62424.33	103057.59	32523.45	17517.59	0.00	10798.17	0.00	4994.78	5280.43
ERAP1	182389.61	247425.59	113092.48	329297.91	235649.19	304454.84	323862.66	327925.66	295053.41	191252.63	150724.78	73035.55	67939.99	60929.91

Appendix Figure 1: Haplotype analysis. **(a)** Diagram showing variant shortlisting across the 100K GP samples from European ancestry. Brown square indicates the region of interest. **(b) (top)** Table displaying the resulting haplotypes with their occurrence in our cohort (in brackets frequencies in cases and controls); red and blue indicate significant associations for variants and haplotypes, respectively. **(bottom)** LD plot displaying the blocks resulting from the 20 variants shortlisted in (a). The region surrounding the CAG repeat is not included in any block.





Appendix Figure 2