Elucidating the Genetic Landscape of the Frontotemporal Dementias using Next-Generation Sequencing and In-Silico Analyses

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Anupriya Dalmia
aus Mumbai, India

2021

Tag der mündlichen Prüfung: 17.12.2021

Dekan der Math.-Nat. Fakultät: Prof. Dr. Thilo Stehle
Dekan der Medizinischen Fakultät: Prof. Dr. Bernd Pichler

1. Berichterstatter: Prof. Dr. Peter Heutink
2. Berichterstatter: Prof. Dr. Thomas Gasser

Prüfungskommission:

Prof. Dr. Peter Heutink
Prof. Dr. Thomas Gasser
Prof. Dr. Stefan Bonn
Prof. Dr. Matthis Synofzik

Erklärung / Declaration:

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

„Elucidating the Genetic Landscape of the Frontotemporal Dementias using Next-Generation Sequencing and In-Silico Analyses“

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled “„Elucidating the Genetic Landscape of the Frontotemporal Dementias using Next-Generation Sequencing and In-Silico Analyses“ ”, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen,
Date: 25.2.2022
Signature: Anupriya Dalmia

Abstract

Anupriya Dalmia

Elucidating the Genetic Landscape of the Frontotemporal Dementias using Next-Generation Sequencing and In-Silico Analyses

Frontotemporal Dementia and Amyotrophic Lateral Sclerosis comprise a spectrum of heterogenous disorders that lie on the "FTD/ALS" spectrum, characterized by similar pathology and genetics but highly variable clinical symptoms that can impact behaviour, cognition or/and motor skills. These diseases have a late age at onset, rapid progression and debilitating symptoms that have devastated tens of thousands of families over the last few decades. To-date, treatment includes only symptom management.

Rapid advances in genomic technologies over the last decade have enabled scientists to investigate the complexities that underlie disease progression and identify some at-risk populations, but much of the genetic variability is still undiscovered. In this dissertation, we apply an in-depth and systematic approach to study the genetic landscape of FTD/ALS in protein coding genes as well as in non-coding genetic elements called long non-coding RNAs (lncRNAs).

In chapters 2 and 3, we apply a step-wise genetic screen of FTD/ALS patients to study the frequencies of both pathogenic and potentially pathogenic mutations in known neurodegenerative disease (NDD) genes. We discover an overlap of pathways previously thought to be associated with other NDDs such as Alzheimer's Disease and type B Kufs disease. As a result of our findings, we propose the inclusion of two genes, CTSF and SERPINA1, in future genetic screens for FTD/ALS. Through rare-variant association tests, we also find an excessive burden of rare, damaging

variants in human autophagy associated genes in FTD/ALS cases versus controls.

In chapters 4 and 5, we perform a two-phase study to investigate the role of lncRNAs in NDDs and healthy ageing. In phase 1, we perform antisense oligonucleotide based knockdowns in highly expressed lncRNAs in the brain. Additionally, we test these lncRNAs for evidence of cis-regulation of proximal genes. In phase 2, we design a genomewide CRISPRi experiment, including a novel sgRNA library targeting $\sim$4000 lncRNAs and 360 negative controls. This, to the best of our knowledge, is the first sgRNA library targeting a genomewide set of lncRNAs expressed in neuronal cell lines. Finally, we perform a series of in-silico analyses using both in-house and public data to gather functional evidence of lncRNAs inageing, cognitive impairment, antisense regulation of NDD genes, eQTL associatedgene regulation as well as those that were differentially expressed in FTD cases versus controls. As a result, we curate a list of 119 lncRNAs with evidence of function in human NDDs and healthy ageing. This is one of the first ever large-scale studies investigating the role of lncRNAs in neurodegeneration.

Acknowledgements

The amount of learning the last 3.5 years has given me is insurmountable. I'd like to thank my supervisors and mentors, Peter and Stefan, for guiding me while always ensuring I had enough freedom to explore my scientific curiosities. Peter also made it possible for me to hone my skills through the attendance of a multitude of courses and conferences during the start of my work, helping me build on important skill sets and critically think about the work that was to follow. It is also important for me to mention the hard work of my colleagues who performed the wet lab experiments that were paramount to this research. I'm thankful to Patrizia for the long hours we've spent analysing my results, for her expert opinions and her trust in mine. I'm also grateful to Thomas for offering invaluable insight at my yearly progress meetings.

I'd like to thank all my colleagues who have made sure my time as a PhD student was enriching, enlightening and exciting. I will miss all our lunchtime discussions, social excursions and being able to share our grief when an analysis didn't do what we expected it to. I'd also like to thank our colleagues who are part of the FANTOM and DESCRIBE-FTD projects for the invaluable work they are doing. Additionally, I am grateful for all the different projects I was given the opportunity to contribute to during my time at the DZNE.

My greatest gratitude extends to my parents and to my sister, Aditi, for all the doors they have opened for me throughout my life, for being constant cheerleaders and pillars of strength. My dog, Rio, who inspires me everyday to be a better version of myself and whom I have missed so dearly all these years of being away from home.

I am thankful to Kartik, who has offered me an ear to vent at and a laugh whenever it was needed, which is, always! To my friends who have been incredible study buddies - Aakriti, Ankita, Saloni, Avani, Anasuya and Tanmai - this wouldn't be

possible without you.

Completing the last leg my PhD in a pandemic when we had to move to doing almost everything virtually was a time of isolation and despair, specially considering how badly it hit India. There are four people who helped me get by, to whom I will always be grateful - Oskar, Joey, Alena and Danny.

Statement of Contributions

All chapters in this monography are written entirely by me.

CHAPTER 2 titled **"Exploring the genetic landscape of FTD in a German Cohort"**:
All Bioinformatics analyses performed in this chapter are by me. The wet lab experiments were conducted by Patrizia Rizzu and Noemia-Rita Alves-Fernandes. Datacollection of DNA samples under the DESCRIBE-FTD and DANCER-FTD studies isorganised and headed by Anja Schneider.

CHAPTER 3 titled **"Genetic Landscape of FTD/ALS in a broader Western European Population"**:
All Bioinformatics and statistical analyses performed in this chapter are by me. The wet lab experiments were conducted by Patrizia Rizzu and Noemia-Rita Alves-Fernandes. Data collection from the different cohorts is described in section 3.2 and 3.3.1.

CHAPTER 4 titled **"Investigating lncRNA function in neurons using ASO-based knockdowns"**:
All Bioinformatics and statistical analyses performed in this chapter are by me. The wet lab experiments were conducted by Salvador Rodriguez-Nieto, Ashutosh Dhingra and Noemia-Rita Alves-Fernandes under the supervision of Patrizia Rizzu. Selection of target lncRNAs for the pilot study was done with the assistance of Tenzin Nyima.

CHAPTER 5 titled **"Global exploration of lncRNA function using a pooled CRISPRi screen and in-silico experiments"**:
All Bioinformatics and statistical analyses performed in this chapter are by me. The wet lab experiments were conducted by Salvador Rodriguez-Nieto, Ashutosh Dhingra and Noemia-Rita Alves-Fernandes under the supervision of Patrizia Rizzu. The

sgRNA library was constructed with the assistance of Tyler Weirick and Chung Chau Hon from the RIKEN institute (Japan).

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer's Disease
AGD	Argyrophilic Grain Disease
ALS-BI	Amyotrophic Lateral Sclerosis with Behavioural Impairment
ALS-CI	Amyotrophic Lateral Sclerosis with Cognitive Impairment
ALS	Amyotrophic Lateral Sclerosis
AMP-AD	Accelerating Medicines Partnership Alzheimer's Disease
AS	Antisense
ASO	Antisense Oligonucleotides
BAM	Binary Alignment Map
BCL	Binary Base Call
BP	Base Pair
BV-FTD	Frontotemporal Dementia with Behavioural Impairment
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CAGE-SEQ	Cap Analysis Gene Expression - Sequencing
CBS	Corticobasal Syndrome
CI	Cognitive Impairment
CNV	Copy Number Variant
CPM	Counts Per Million
CRISPRI	Clustered Regularly Interspaced Palindromic Repeats Interference
CROP-SEQ	CRISPR Droplet Sequencing
DANCER	Degeneration Controls and Relatives
DEG	Differentially Expressed Genes
DGE	Differential Gene Expression
DNA	Deoxyribonucleic Acid
EQTL	Expression Quantitative Trait Loci
FANTOM	Functional Annotation of the Mammalian Genome
FTD-MND	Frontotemporal Dementia with Motor Neuron Disease

FTD	Frontotemporal Dementia
FTLD	Frontotemporal Lobar Degeneration
GATK	Genome Analysis Tool Kit
GLM	Generalized Linear Model
GO	Gene Ontology
GVCF	Genomic Variant Call Format
GWAS	Genome-wide Association Study
HD	Huntington Disease
HRE	Hexanucleotide Repeat Expansion
IBD	Identity by Descent
IBS	Indentity by State
IGV	Integrative Genomics Viewer
IMS	Integrative Medical Sciences
INDEL	Insertion or Deletion
KD	Knockdown
LNA	Locked Nucleic Acid
LNCRNA	Long non-coding RNA
LPA	Lopogenic-variant primary Progressive Aphasia
LQ-SS-CAGE	Low Quantity Single Strand Cap Analysis Gene Expression
MAF	Minor Allele Frequency
MCI	Mild Cognitive Impairment
MDS	Multidimensional Scaling
MLPA	Multiplex Ligation-dependent Probe Amplification
MRNA	Messenger Ribonucleic Acid
NB	Negative Binomial
NCI	No Cognitive Impairment
NDD	Neurodegenerative Diseases
NCRNA	Non-coding RNA
NFV-PPA	Nonfluent Variant Primary Progressive Aphasia
NGS	Next Generation Sequencing
NPC	Neuronal Precursor Cell
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD	Parkinson's Disease

PE	Paired-End
PNFA	Progressive Non-Fluent Aphasia
PPA	Primary Progressive Aphasia
PSPS	Progressive Supranuclear Palsy Syndrome
Q-PCR	Quantitative Polymerase Chain Reaction
QC	Quality Control
QL	Quasi-Likelihood
RIMOD-FTD	Risk and Modifying factors for Frontotemporal Dementia
RIN	RNA Integrity Number
RNA-SEQ	RNA-Sequencing
ROS-MAP	Religious Orders Study and Memory and Aging Project
RP-PCR	Repeat Primed-Polymerase Chain Reaction
RSX1	Rotterdam Study Exome Sequencing Database
SCRNA-SEQ	Single-Cell RNA Sequencing
SE	Single-End
SEMD	Semantic Dementia
SG-RNA	Single Guide RNA
SKAT-O	Optimized SNP-Set Kernel Association Test
SNP	Single Nucleotide Polymorphism
SV-PPA	Semantic Variant Primary Progressive Aphasia
TDP-43	TAR DNA-Binding Protein 43
TPM	Trasncripts Per Million
TSS	Transcription Start Sites
UBA	Ubiquitin-Associated
UCSC	University of California, Santa Cruz
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

For Shiva,

Chapter 1

Introduction

1.1 A Brief Background in Neurodegenerative Diseases, Genetics and the Motivation for this Dissertation

Neurodegenerative diseases (NDD) are progressive, debilitating disorders that are often hallmarked by neuronal loss. The most common neurodegenerative diseases are Alzheimer Disease (AD), followed by Parkinson disease (PD), Lewy Body Dementia, Frontotemporal Dementia (FTD), Amyotrophic Lateral Sclerosis (ALS), Huntington Disease (HD), and Prion diseases. Most of these diseases carry a significant genetic component to them. For many years, the treatment of neurodegenerative diseases has been limited to alleviating symptoms which on its own is a challenge in older patients who are largely impacted by NDDs. Via uncovering the genetic variability that leads to propensity to disease, we open avenues for identification of at-risk populations and to a better understanding of the underlying mechanisms of these complex diseases. In this dissertation, we will be exploring, in detail, the genetic landscape of FTD, often accompanied by motor symptoms consistent with ALS. Although FTD and ALS are two clinically distinct diseases, they share similarities in disease pathology (eg.,FUS, TDP-43 aggregation) and genetics (eg., C9Orf72, VCP).

1.2 Frontotemporal Dementia (FTD)

FTD is an umbrella term that encompasses a heterogeneous spectrum of disorders, the core of which include behavioral variant FTD (bvFTD), nonfluent/agrammatic variant primary progressive aphasia (nfvPPA), and semantic variant PPA (svPPA).

It is the third most common type of dementia, and is a leading type of early onset dementia (R. T. Vieira et al. 2013). FTD can present itself in the form of cognitive, behaviour, language, executive control and, sometimes, motor impairments. As the name suggests, the majorly affected areas in the brain of FTD patients are the frontal and temporal lobes.

Over time, the description of FTD has changed in terms of both categorisation and nomenclature. FTD (known then as 'Pick's Disease') was first described by Pick, a Czech neurologist, in 1892 (Pick, Girling, and Berrios 1994). Pick's patient suffered from left temporal lobe atrophy, causing language impairment, and which would presently be described as svPPA.

1.2.1 FTD Spectrum Disorders

1.2.1.1 Behavioural Variant Frontotemporal Dementia

Behavioural-variant frontotemporal dementia (bvFTD) is characterised by early changes in social behaviours such as personality and emotional response, as well as loss in executive control and pain response (Rascovsky et al. 2011). Patients are often unaware of these changes in their own behaviour. The new diagnostic consensus criteria for bvFTD require that for the diagnosis of possible bvFTD, three of the following behavioral/cognitive symptoms must be persistent or recurrent within the three first years of disease: behavioral disinhibition; apathy or inertia; loss of sympathy or empathy; perseverative, stereotyped or compulsive/ritualistic behavior; hyperorality and dietary changes; and, neuropsychological findings that include executive/generation deficits with relative sparing of memory and visuospatial functions.This is the most prevalent form of FTD, in the spectrum. About 12-15% of patients with bvFTD also develop motor neuron disease (Burrell et al. 2011).

1.2.1.2 Primary Progressive Aphasia

Primary progressive aphasia (PPA) is characterized by a progressive decline in linguistic skills of the patients. These language deficits are apparent during speech and language assessments.

1.2.1.2.1 Semantic Variant Primary Progressive Aphasia Semantic-variant primary progressive aphasia (svPPA) is categorized as: (i) left svPPA and (ii) right

svPPA, based on the affected temporal lobe. Patients with left svPPA have linguistic deficits, primarily loss of semantic knowledge and memory. Whereas, in right svPPA, behavioural symptoms predominate, including inappropriate social behaviour, change in personality, insomnia, loss of appetite and libido.

1.2.1.2.2 Nonfluent Variant Primary Progressive Aphasia Non-fluent variant primary progressive aphasia (nfv-PPA), also known as progressive non-fluent aphasia (PNFA), is characterized by speech impairment in the form of laboured speech and agrammatism. In addition, patients may also suffer from inability to comprehend complex sentences.

1.2.2 Related FTD Disorders

Other disorders that fall in the FTD spectrum include frontotemporal dementia with motor neuron disease (FTD-MND), progressive supranuclear palsy syndrome (PSP-S) and corticobasal syndrome (CBS).

1.2.3 Neuropathology of FTD

Frontotemporal dementia is caused by "Frontotemporal lobar degeneration" (FTLD), which is a neurodegenerative process involving selective neuronal loss and gliosis of the frontal and temporal lobes of the brain (Mackenzie et al. 2009). The different subtypes of FTD are associated with characteristic patterns of protein deposition. The three proteins involved in the majority of FTLD cases are: (i) the microtubule-associated protein tau (MAPT), (ii) the TAR DNA-binding protein with molecular weight 43 kDa (TDP-43), or the (iii) fused-in-sarcoma (FUS) protein. These are then categorized as FTLD-Tau, FTLD-TDP and FTLD-FUS, respectively. In 2011, abnormal expansions of a GGGGCC hexanucleotide repeat in a non-coding region of the C9orf72 gene was identified as the most common genetic cause of familial and sporadic forms of both FTD and ALS, and the basis of most families in which both conditions occur (DeJesus-Hernandez et al. 2011; Renton et al. 2011). These add to the heterogeneity of the neuropathology of FTD through RNA and/or protein toxicity.

1.2.3.1 FTLD-TDP

The most common class of FTLD is associated with TDP-43 proteinopathy, first described in 2006 (Neumann et al. 2006). TDP-43 is an RNA and DNA binding protein

with regulatory roles in numerous cellular processes: transcription, splicing, cell cycle regulation, apoptosis, microRNA biogenesis, mRNA transport to and local translation at the synapse and scaffolding for nuclear bodies (Buratti and Baralle 2008). In pathological conditions, TDP-43 is displaced from the nucleus to the cytoplasm, hyperphosphorylated, ubiquitinated and cleaved to produce C-terminal fragments (Bigio 2011). TDP-43 proteinopathy is also associated with other neurodegenerative diseases (NDD) such as MND with or without dementia, and Perry syndrome (Neumann et al. 2006). Occasionally, TDP-43 inclusions are also found in Alzheimer's disease, Parkinson dementia complex of Guam, and Lewy body disease.

1.2.3.2 FTLD-Tau

In about 45% cases of FTLD, the intraneural accumulation of filamentous, hyperphosphorylated microtubule-associated tau protein is observed (Boxer et al. 2013). Tau regulates axonal transport by maintaining microtubule stability. In pathological conditions, tau is hyperphosphorylated and assembled into insoluble filaments called **neurofibrillary tangles** that accumulate in neurons and/or glia (Michel Goedert and Spillantini 2011). Disorders related to FTD in which tauopathies are observed include Pick's disease (PiD), CBS, PSP-S, argyrophilic grain disease (AGD) (Josephs et al. 2011). FTLD-Tau is the most common neuropathological finding in patients with nfv-PPA/PNFA (Deramecourt et al. 2010).

1.2.3.3 FTLD-FUS

FUS-associated proteinopathies are seen in 5% of FTLD cases (Neumann et al. 2009). Like TDP-43, FUS is a DNA and RNA binding protein with regulatory roles in gene expression, transcription, RNA splicing, transport and translation (Lashley et al. 2011). While FUS is mainly expressed in the nucleus, it can shuttle between the cytoplasm and nucleus. In pathological conditions, FUS immunoreactive inclusions are seen in neurons and glial cells.

1.2.4 Genetics of FTD

FTD has a significant genetic component, with an estimated 43% of patients carrying a positive family history [at least one affected first-degree family member with dementia, ALS, or Parkinson's disease (PD)] and between 10.2% and 27% of FTD patients have an autosomal dominant presentation of the disease (Pottier et al. 2016).

In the 1990s, the definition of FTD was ambiguous and knowledge on the genetics of FTD was scarce. Several families diagnosed with FTD or related disorders, including Parkinsonism, were rapidly linked to chromosome 17q (Wijker et al. 1996) and in 1996, an International Consensus meeting identified 13 kindreds with evidence of linkage to 17q and renamed the disorder as frontotemporal dementia with parkinsonism linked to chromosome 17 (FTDP-17) (Foster et al. 1997).

Over the past decade, the following protein-coding genes have been consistently associated with FTD:

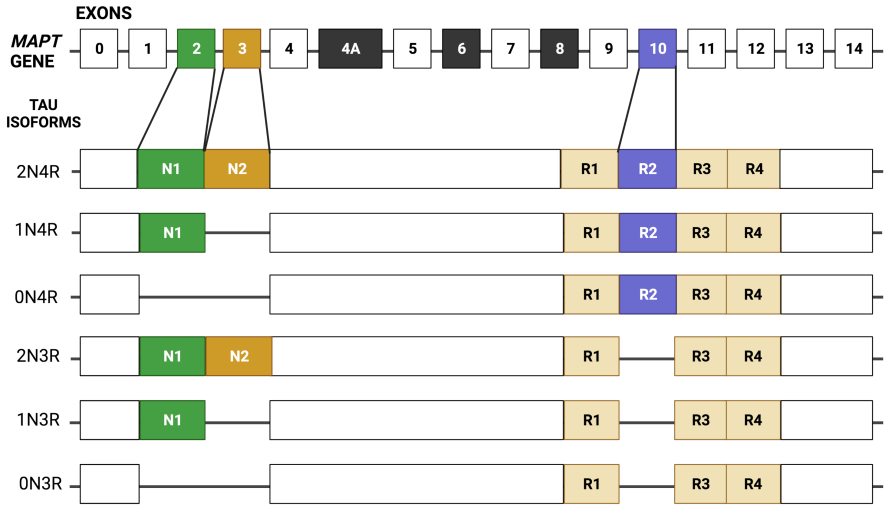
The Microtubule-Associated Protein Tau (MAPT), OMIM: 157140

As stated above, microtubule associated protein tau dysfunction is responsible for a large proportion of FTD cases, and the gene that encodes the tau protein, MAPT, was the first gene discovered to be associated with FTD. In 1998, the first mutations were reported in exons 9, 10 and 13, as well as in the splice site of intron 10 (Poorkaj et al. 1998; Hutton et al. 1998; Spillantini et al. 1998).

Tau is abundant in the brain, and within neurons it is primarily found in axons. Here it stabilizes microtubules and regulates neurite outgrowth. The interaction of tau with microtubules occurs primarily through the repeat 'microtubule-binding domains' in the C-terminus of tau. Tau also interacts with components of the plasma membrane through its amino terminal projection domain (Brandt, Léger, and Lee 1995; Gauthier-Kemper et al. 2011; Pooler et al. 2012). The MAPT gene is located on chromosome 17q21 (Neve et al. 1986) has 16 exons, of which exons 2, 3, 4A, 6, 8 and 10 can be alternatively spliced (M. Goedert et al. 1989). Exons 4A, 6 and 8 are not transcribed in the brain, and in total 6 isoforms of MAPT are present in the brain as a result of differential splicing of exons 2, 3 and/or 10. Tau proteins can either be 3R or 4R tau i.e., with 3 or 4 binding repeats respectively, depending on the alternative splicing of exon 10, which encodes the second repeat domain, R2 (M. Goedert and Jakes 1990). These are the binding units for microtubules and are essential for maintenance of their stability and dynamics (Dixit et al. 2008). In the mature human brain, the ratio of 3R:4R is 1:1, with the relative amounts varying between cell types and brain regions (M. Goedert et al. 1989).

Mutations in MAPT are responsible for 10-20% of familial FTD cases and almost 3%

FIGURE 1.1: The 6 isoforms of MAPT expressed in the human brain.



The exons marked in white are constitutive exons and those marked in dark grey are not expressed in the human brain.The aspects of the N-terminal projection domain, N1 (green) and N2 (mustard), are produced from exons 2 and 3, respectively. Exon 10 encodes the second aspect of the microtubule-binding repeat domain, R2 (purple). (Created with BioRender.com)

of sporadic FTD cases (Benussi, Padovani, and Borroni 2015). The mean age at onset for patients with pathogenic MAPT mutations is 55 years old which is lower than that for other FTD genes (Seelaar et al. 2011). Pathogenic mutations in MAPT are usually either missense mutations or deletions in exons 1, 9–13, or splice mutations in the intron that follows exon 10. In general, mutations that affect the alternate splicing of exon 10 causes a relative increase in 4R tau, are associated with neuronal and glial pathology that resembles sporadic PSP-S and CBS, and are often associated with prominent parkinsonism. In contrast, mutations in exons 9, 11, 12, and 13 lead to a predominance of neuronal inclusions (either Pick bodies composed of 3R tau or AD-like NFT composed of both 3R and 4R tau isoforms) and more often cause dementia. (I. R. A. Mackenzie and Neumann 2016).

Most polymorphisms in the MAPT gene are in complete linkage disequilibrium with each other and are inherited as two separate haplotypes, H1 and H2. The predominant haplotype is H1, which has been linked to sporadic tauopathies, PSP-S and CBS (M. Baker et al. 1999) whereas H2 has been linked to early age at onset in familial FTD (Ghidoni et al. 2006).

Granulin (GRN), OMIM: 138945

Several families with linkage to chromosome 17q21 but without pathogenic MAPT mutations and thus, without characteristic tau inclusions, were found in the late 1990s (Lendon et al. 1998). These patients carried ubiquitin-positive neuronal intranuclear inclusions which were later characterized as TDP-43 inclusions (Sun and Eriksen 2011). Mutations in the GRN gene, which lies 2 centimorgans from the MAPT gene, were identified as the cause for these TDP-43 positive FTD patients (Cruts et al. 2006; Matt Baker et al. 2006). GRN encodes a secreted growth factor with a role in inflammation, tissue development and tumorigenesis and has increased expression in microglia in patients with neurodegenerative diseases (Bateman and Bennett 2009). Over 70 GRN mutations are known to be causal for FTD, causing haploinsufficiency, and sometimes, non-functional or unstable proteins (Eriksen and Mackenzie 2008).

Age at onset for FTD in GRN mutation carriers is highly variable, with an average of 65 years old. The penetrance is high but incomplete: 50% by age 60 and 90% by age 70 (Seelaar et al. 2011). There is significant variation in the associated clinical presentation, even among individuals within the same family. Most present as bvFTD or nfvPPA and some degree of parkinsonism is common but ALS is exceptionally rare (I. R. A. Mackenzie 2007). Mutations in GRN are all heterozygous and are responsible for 5-20% of familial and 1-5% of sporadic FTD cases (Rademakers, Neumann, and Mackenzie 2012).

Over 70 GRN mutations are known to be causal for FTD, including frameshift, nonsense, missense, and splice mutations, but also with rare partial deletions and a complete deletion of GRN (Gijselinck, Van Broeckhoven, and Cruts 2008). All pathogenic mutations uniformly lead to a 50% loss in GRN protein levels leading to disease through haploinsufficiency. These include missense mutations that introduce premature stop codons or those that alter the initiation codon or mutations causing intron retention. These lead to degradation of the mutant GRN mRNA by nonsense-mediated decay (Baker et al., 2006).

Despite the clinical variability, most if not all GRN mutation carriers present with

TDP-43 pathology at autopsy (I. R. A. Mackenzie et al. 2006). The exact link between GRN haploinsufficiency and TDP-43 pathology is not known and remains a topic of active research (Kleinberger et al. 2013).

The Open Reading Frame 72 of Chromosome 9 (C9orf72), OMIM: 614260

The most common genetic cause of both familial and sporadic FTD and ALS are expansions of an intronic hexanucleotide repeat (GGGGCC) in the C9Orf72 gene (DeJesus-Hernandez et al. 2011; Renton et al. 2011). These repeat expansions are responsible for 21% of familial FTD cases, and almost 6% of sporadic FTD cases (DeJesus-Hernandez et al. 2011; Majounie et al. 2012; Rademakers, Neumann, and Mackenzie 2012). In healthy individuals, there are usually 2-24 non-coding hexanucleotide GGGGCC repeats but in diseased individuals, these expansions can occur hundreds to thousands of times (Seelaar et al. 2011). As with GRN mutation families, there may be tremendous clinical heterogeneity among members of a family with the C9orf72 mutation. The neuropathology is a combination of FTLD-TDP and typical ALS with TDP-43 inclusions in a wide range of neuroanatomical regions ((Irwin et al. 2015; I. R. A. Mackenzie and Neumann 2016). The most common clinical presentation is bvFTD, ALS, or the combination of both (Pottier et al. 2016).

The exact mechanism of how these repeats cause the disease is unknown but there is evidence of haploinsufficiency through loss of gene expression or/and gain of function with secondary RNA toxicity (Benussi, Padovani, and Borroni 2015). In addition to TDP-43 pathology, all C9orf72 mutation carriers present with neuronal inclusions in the cerebellar granule cell layer, hippocampal pyramidal neurons, and other neuroanatomical sites that stain positively for proteins of the ubiquitin proteasome system (such as ubiquitin and p62) but are negative for TDP-43. These inclusions were recently found to be composed of dipeptide repeat proteins (DPRs), translated from the GGGGCC repeat through unconventional repeat-associated non-ATG translation (Ash et al. 2013; Mori, Weng, et al. 2013). Three DPRs (poly-GP, poly-GA, and poly-GR) are generated from the sense strand and abundantly detected in the cerebellum and hippocampus of expansion carriers. DPRs from the antisense strand (poly-PA, poly-PR, and again poly-GP) are also generated (Gendron et al. 2013; Mori, Arzberger, et al. 2013). The anatomical distribution of DPR pathology is highly consistent among cases, regardless of the clinical features and shows no

correlation with the regional pattern of neurodegeneration or TDP-43 pathology (I. R. Mackenzie et al. 2013).

Adding to this complexity is another consistent feature that is seen in C9Orf72 hexanucleotide expansion carriers: intracellular aggregates of RNA, composed of the mutant sense and antisense transcripts (DeJesus-Hernandez et al. 2011; Gendron et al. 2013; Mizielinska et al. 2013). These RNA foci are present in up to 50% of neuronal nuclei in key anatomical regions and less frequently seen in neuronal cytoplasm and in glial cells.

The TARDNA Binding Protein 43 Encoding Gene (TARDBP), OMIM: 605078
As stated previously, TDP-43 is an RNA-binding protein that forms heterogeneous nuclear ribonucleoprotein complexes (hnRNP) encoded by TARDBP on chromosome 1. It has a role in transcription, RNA splicing and microRNA processing (Sieben et al. 2012). Mutations in the TARDBP gene are typically associated with ALS and are rare in FTD.

Fused in Sarcoma Gene (FUS), OMIM: 137070
FUS gene is located on the chromosome 16q11.22 and is also a member of the hnRNP family (Sieben et al. 2012). Mutations in FUS are very commonly causative for ALS (Sieben et al. 2012; Kwiatkowski et al. 2009; Vance et al. 2009), and have also been observed in FTD patients, specially those with FTD-MND (Van Langenhove et al. 2010; Yan et al. 2010). Pathological FUS inclusions are present in most cases without TDP-43 and tau inclusions, accounting for almost 10% of FTLD cases, described as FTLD-FUS (Mackenzie et al. 2010). FUS-containing aggregates are seen both in the nucleus and the cytoplasm of neurons and glial cells (Mackenzie et al. 2010; Belzil et al. 2011; Chiò et al. 2011; DeJesus-Hernandez et al. 2010). FUS is involved in a variety of cellular processes such as transcription, splicing, RNA localization and degradation and DNA damage (Lagier-Tourenne, Polymenidou, and Cleveland 2010).

The Valosin-Containing Protein (VCP), OMIM: 601023

Valosin-containing protein is encoded by the VCP gene which is located on chromosome 9p13.3 and is involved in protein degradation, membrane fusion, transcriptional activation and apoptosis (Seelaar et al. 2011). There are 19 known pathogenic mutations in VCP and 80% of the carriers have a positive family history. Mutations in VCP have been described in FTD patients in the FTLD-TDP molecular subgroup (Mackenzie and Neumann 2016). These patients are usually diagnosed with bvFTD.

The Chromatin-Modifying 2B (CHMP2B), OMIM: 609512

The CHMP2B gene is located on chromosome 3p11.2 and carries pathological mutations for familial FTD, most commonly presenting as bvFTD (Stokholm et al. 2013). CHMP2B encodes a component of the heterometric ESCRT-III complex (Endosomal Sorting Complex Required for Transport III) that plays a role in the recycling or degradation of cell surface receptors, and is expressed in neurons of all major brain regions. (Han et al. 2012)

The TANK-Binding Kinase (TBK1), OMIM: 604834

TBK1 has recently been linked to ALS and FTD (Freischmidt et al. 2015) and over 100 variants including loss of function variants, in-frame deletions, and missense variants have been reported in ALS, FTD, or ALS-FTD patients, thus making TBK1 the third or fourth most frequent genetic cause of FTD after C9orf72, GRN and MAPT. (Freischmidt et al. 2017).

The Sequestome 1 (SQSTM1), OMIM: 601530

The SQSTM1 gene is located on chromosome 5q35 and encodes the p62 protein that is a stress-responsive ubiquitin binding protein commonly found in neuronal cytoplasmic inclusions. P62 plays a role in protein degradation via the proteasome, in protein aggregation and in autophagy (Bjørkøy et al. 2006; Seibenhener et al. 2004). Increased p62 immunoreactivity has been observed in patients with neurological disorders such as AD, dementia with Lewy bodies, FTLD, Parkinson disease (PD) and Huntington disease (HD) (Kuusisto et al. 2002; Zatloukal et al. 2002; Nakaso et al. 2004). Additionally, patients with FTLD or ALS carrying the C9orf72 repeat expansion present abundant neuronal p62-positive inclusions (Murray et al. 2011; Al-Sarraj et al. 2011).

1.2.5 Challenges with the clinical distinction of FTD and Alzheimer's Disease

Patients with FTD may have superimposed amyloid-beta pathology, which is a hallmark of Alzheimer's disease (Rohrer et al. 2011). Recently, amyloid-beta plaques have been observed to facilitate tau aggregation in AD models (He et al. 2018). Additionally, in FTD patients, amyloid pathology has been associated with a worse performance in several cognitive tests (He et al. 2018; Naasan et al. 2014) and CSF amyloid has been associated with increased volumetric loss(He et al. 2018; Naasan et al. 2014; Ljubenkov et al. 2018). A recent study examining the association of CSF amyloid-beta with mortality rate in FTD patients reported that patients who died earlier had a significantly lowered CSF amyloid-beta than those who did not. (D. Vieira et al. 2019).

Owing to these overlapping neuropathologies and clinical symptoms, it is often difficult to elucidate the diagnosis and AD patients are often misdiagnosed as bvFTD patients especially if they have an early age at onset. While definitive diagnoses can only be arrived at by studying neuropathology, genetic testing of patients has proven to be a helpful tool in clinically diagnosing patients and avoiding false positives.

1.2.6 The Epidemiology of Frontotemporal Dementia

FTD prevalence was estimated between 0.01-4.61 per 1000 persons and the incidence between 0.01-2.5 per 1000 person/year (Hogan et al. 2016). The same study noted that the behavioural variant of FTD was almost four times as common as the primary progressive aphasias. In recent dementia cohorts, FTD cases have been found to account for 1.6-7% of dementia cases, making it the second leading cause of adult-onset dementia after AD (Religa et al. 2015)(van der Flier and Scheltens 2018). Due to several factors, these numbers are likely underestimated:

- FTD is underdiagnosed, several neuropathological studies confirm that as much as 9% of the elderly population irrespective of cognitive impairment has FTD pathology at the time of death (Beach et al. 2015).

- Due to the overlapping symptoms of FTD with other diseases, the time taken to diagnose FTD correctly is usually longer than other forms of dementia. There is no single test that can conclusively diagnose FTD, and clinicians often have to investigate family history, conduct cognitive and behavioural examinations, rule out other disorders that can cause similar symptoms (for eg., sleep apnea) and run blood tests and brain imaging to rule out metabolic deficiencies (for eg., deficiency of Vitamin B12 can cause neurological symptoms) and cardiovascular illness (for eg., checking for tumors, subdural hematomas, hydrocephalus, etc.). Due to these reasons, arriving at an early diagnosis of FTD is challenging and as the patient ages and the disease progresses, usually other health issues associated with old age arise, making a definitive diagnosis harder. An estimated 30% cases are, thus, misdiagnosed and can only be confirmed post mortem.

- Older publications and study exclude several of the new syndromes that lie in the FTD/ALS disease spectrum.

- Non-referrals: In cases of psychiatric, amnestic and/or late-onset presentations of FTD, as well as in cases of an overlap of behavioural, cognitive and motor presentations, the diagnosis is often overlooked and underestimated.

Despite FTD being a devastating and prevalent disease, it's complexity makes it difficult to arrive at a timely diagnosis. Advances in neuropsychology, neuroimaging and cerebrospinal fluid (CSF) biomarkers and genetics have improved FTD diagnosis making the need for accurate FTD/ALS biomarkers imperative. Currently, the two most widely used biomarkers to distinguish FTD vs AD are P-Tau$_{181}$ and A$\beta_{1\text{-}42}$/A$\beta_{1\text{-}40}$ ratio. We will explore in this thesis how genetics can be an important aid to arriving at a clinical diagnosis by studying the frequencies of damaging variants that lie in known NDD genes and by studying new genes that are potential risk genes and should be included in genetic screens for FTD/ALS, along with exploring pathways that may suffer insults in patients at risk for FTD/ALS.

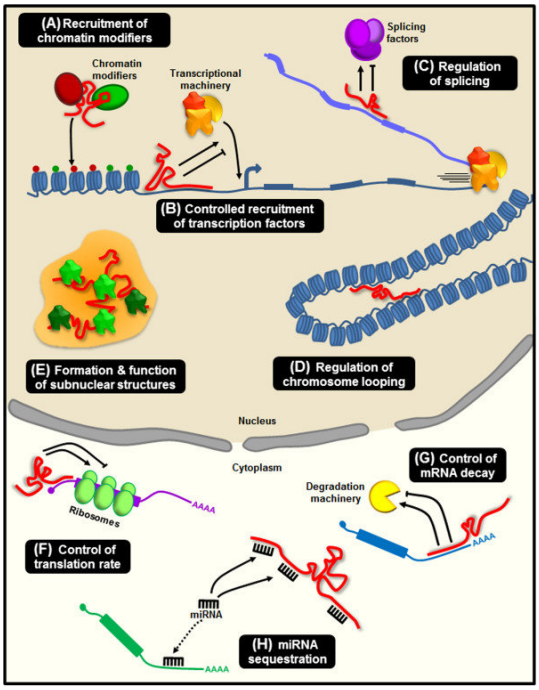
1.2.6.1 Non-coding genetic elements

Most of the human genome is transcribed at some stage - embryonic development, growth or disease progression but only 2% of it encodes proteins. Due to the low

expression levels of these non-coding RNAs (ncRNAs) compared to canonical mRNAs encoding proteins, for several decades they were referred to as "junk DNA" and their functional relevance was left uninvestigated. Evidence of important regulatory roles played by ncRNAs, especially long non-coding RNAs, has emerged rapidly over the last decade (Beermann et al. 2016). LncRNAs are broadly defined as non-coding RNA molecules longer than 200 nucleotides. Efforts to discover, annotate and characterize lncRNAs has revealed a massive atlas of > 27,000 human lncRNA genes (Hon et al. 2017). Despite not being translated into proteins, lncRNAs are molecules with a myriad of regulatory functions ranging from chromatin modification, splicing, mRNA decay, protein translation, protein stability and maintenance of the transcriptional machinery. The deregulation of lncRNAs has been associated with several human diseases (Wapinski and Chang 2011) disease, yet the function of a majority of these lncRNAs eludes scientists till today. Some questions that remain pertinent to the field are whether it is all or just a fraction of lncRNAs that carry important regulatory functions and whether it is the RNA product or the act of transcription that is functionally relevant. Another important facet to consider while studying lncRNAs is that the functionality of lncRNA loci is often revealed by assessing the selective constraints (Davydov et al. 2010) and genetic variations (Farh et al. 2015) within their regulatory regions than their transcript sequences. Below, we highlight some known mechanisms by which lncRNAs regulate and moderate transcription, translation, splicing and degradation machinery.

LncRNAs are involved in regulating histone modifications at the chromatin level via interactions with histone-associated acetylase and methylase and recruiting chromatin modification complexes at the chromatin level by acting as molecular scaffolds. By acting as co-factors or inhibitors to transcription factors, lncRNAs can regulate transcription in both directions i.e., activation and interference. Alternative splicing is an important mechanism for spatial and temporal regulation of gene expression and for proteomic diversity. By interacting with the splicing machinery, lncRNAs can regulate both alternative splicing of mRNAs and lncRNAs. Micrornas (miRNAs) are a class of ncRNAs that between 18-25 nucleotides in length, do not have an open reading frame and are widely expressed in eukaryotes. They play an important role in targeting mRNA for cleavage or directing translational inhibition to negatively regulate mRNA expression. Many lncRNAs act as “miRNA

FIGURE 1.2: Mechanisms for long non-coding RNA (lncRNA) function (Publication: Neguembor, Jothi, and Gabellini, 2014. Licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/)



A) LncRNAs (in red) are able to recruit chromatin modifiers mediating the deposition of activatory (green dots) or repressive (red dots) histone marks. (B) LncRNAs control the recruitment of transcription factors and core components of the transcriptional machinery. (C) LncRNAs can directly bind mRNAs and modulate splicing events. (D-E) LncRNAs participate in the higher order organization of the nucleus by mediating chromatin looping (D) and as structural components for the formation and function of nuclear bodies (E). (F) LncRNAs control translation rates favoring or inhibiting polysome loading to mRNAs. (G) LncRNAs modulate mRNA decay protecting mRNA from degradation or, alternatively, mediating the recruitment of degradation machinery. (H) LncRNAs can act as miRNA sponges, thus favoring the expression of the mRNAs targeted by the sequestered miRNA.

sponges" by acting as endogenous target mimics and sequestering miRNAs. Recent studies have also suggested roles of lncRNAs in regulating the stability of mRNA post-transcriptionally (Zhang et al. 2019).

1.3 Sequencing Technologies

High throughput sequencing technologies have paved the way for fast and relatively cheap large scale studies involving the human genome and transcriptome.

These technologies allow for sequencing of DNA and RNA much more quickly than the previously used Sanger Sequencing, which is a technique which was developed

in 1977 to determine nucleotide sequences using the "chain-termination method" (Sanger, Nicklen, and Coulson 1977), and as such revolutionised the study of genomics and molecular biology. Illumina and Agilent Technologies are USA-based Biotechnology companies that provide most of the platforms for these sequencing technologies.

In the above sections, we have reviewed the merits of studying the genome of patients with a diagnosis of a complex, polygenic disease like FTS/ALS. Along with the genome, studying the transcriptome of patients and comparing it with that of a healthy population can offer useful hints on disease progression. The transcriptome comprises all the RNA molecules transcribed from the DNA. It not only forms the basis of all proteins, but also non-coding RNAs i.e., miRNAs, lncRNAs, etc. Studying insults to the transcriptome offers a valuable window into studying disease mechanisms.

In this thesis, we utilise a number of next-generation sequencing (NGS) approaches using short-read sequences, the basis of which are outlined below:

1.3.1 Whole Exome/Genome Sequencing Analysis

There are up to 1 billion listed SNPs in the dbSNP database for homo sapiens (Sherry 2001). Rare single nucleotide variants (SNVs), small INDELs and CNVs have been demonstrated to underlie many disorders, but remain difficult to study due to their low minor allele frequency (MAF). WES/WGS are revolutionary tools in studying rare variation through high throughput, high quality and depth data from large cohorts of patients in a scalable fashion. Although WES offers some obvious attractions of lower costs, simplification of variant analysis and data storage, there are several merits to WGS over WES even when studying protein coding genes. Despite evidence of incremental improvements in exome capture technology over time, WGS has greater uniformity of sequence read coverage and reduced biases in the detection of non-reference alleles than WES. Exome-seq achieves 95% SNP detection sensitivity at a mean on-target depth of 40 reads, whereas WGS only requires a mean of 14 reads. Some reasons that cause a lower sensitivity in SNP detection in WES include PCR amplification, which tends towards lower coverage in GC-rich regions

due to annealing during amplification, and the preferential capture of reference sequence alleles, which biases the allele distribution away from alternate alleles at heterozygous SNP sites. WES produces a relatively heterogeneous profile of read coverage over target regions when compared to the more homogeneous WGS. Since disease-causing mutations are not biased towards easy or hard to sequence areas of the genome for either WES or WGS, there are arguments to be made to invest in the more expensive WGS when studying rare diseases (Meynert et al. 2014).

Each human being carries 4-5 million, primarily benign, SNPs in their genome. Out of these variations, only a minority are unambiguously deleterious and introduce premature stop codons or impact normal mRNA splicing. The most frequent class of genetic variation that occurs is a missense mutation which introduces a different amino acid by altering a single codon. An estimated 2% people carry a missense mutation in any given gene (Andrews, Sjollema, and Goodnow 2013). A major challenge, thus, arises in predicting whether these mutations are damaging and alter the function of the corresponding protein, especially when a patient carries a mutation in a gene of interest to their phenotype.

Several damage prediction algorithms like Polyphen-2 (Adzhubei et al. 2010), SIFT (Kumar, Henikoff, and Ng 2009), as well as scoring algorithms like CADD (Kircher et al. 2014) help infer the deleterious effects of rare variation. We compare these prediction methods in Table 1.1. The widely used PolyPhen2 and CADD tools integrate a number of different information sources, including sequence and structure-based features (and in the case of CADD, the results of other tools such as VEP (McLaren et al. 2016), data from the ENCODE project (Consortium and The ENCODE Project Consortium 2004) and information from the UCSC browser tracks (Kent 2002)), and use a machine learning approach to categorize variants as benign or deleterious. SIFT predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids. The score is the normalized probability that the amino acid change is tolerated so scores nearer zero are more likely to be deleterious. The qualitative prediction is derived from this score such that substitutions with a score < 0.05 are called 'deleterious' and all others are called 'tolerated'. However, to ascertain a variant as causal, family history is required to confirm co-segregation of the variant with

TABLE 1.1: Comparison of damage prediction algorithms for genetic variation

NAME	CATEGORY	SCORE USED	INFORMATION USED
SIFT (Kumar, Henikoff, and Ng 2009)	Function prediction	1 - Score	Protein sequence conservation among homologs, physico-chemical similarity between alternate amino acids
Polyphen-2 (Adzhubei et al. 2010)	Function prediction	Score	Eight protein sequence features, three protein structure features
CADD (Kircher et al. 2014)	Ensemble score	Score	diverse genomic features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements and functional predictions.

disease.

Large scale exome sequencing studies such as the ones conducted by the Genome Aggregation Database (gnomAD) (Karczewski et al. 2020) are powerful tools in deciphering the MAF of a rare variant in a number of different populations. The gnomAD database describes genetic variation from 125,748 exomes and 15,708 genomes, comprising over 270 million variants making it the largest catalogue of human variant data (https://gnomad.broadinstitute.org/).

1.3.2 RNA Sequencing

Prior to RNA-Sequencing (RNA-Seq), the preferred technology used to quantify RNA was microarrays with a predefined set of oligonucleotides. The development of RNA-Seq has enabled the sequencing of the entire transcriptome at low costs. For RNA-Seq, RNA is extracted from samples of interest and reverse transcribed to cDNA. For gene expression measurements, a typical DNA fragment is between 50 - 200 bp in length. The cDNA libraries are then ligated to sequencing adapters on a flow cell and amplified via PCR. Then, these amplified cNA fragments are sequenced. After each sequencing event, the nucleotide is determined via a fluorescent signal. In the case of paired-end sequencing, the cDNA fragment is also sequenced in the reverse direction from the opposite end, yielding a forward and reverse set of reads. These reads enable higher quality sequence alignment to the reference genome. As stated above, studying insults to the transcriptome offer valuable clues

in studying disease progression, when compared against a control dataset.

Some limitations that accompany RNA-Seq are the inability to identify novel reads using short-reads provided by Illumina technology. Short reads rarely span across several splice junctions and thus make it difficult to directly infer all full-length transcripts. In addition, it is difficult to identify transcription start and end sites accurately using RNA-Seq. Also, conventional RNA-Seq methods do not capture the transcriptomic composition of individual cells. Since the transcriptome of each individual cell is extremely dynamic and reflects its functionality, there have been advances in single-cell sequencing technologies to circumvent biases that come with bulk sequencing.

1.3.3 CAGE Sequencing

Cap Analysis of Gene Expression Sequencing (CAGE-Seq) is a technology designed to capture the 5'-end of the mRNA usually in short 27 nucleotide long fragments (Shiraki et al. 2003). The beginning of the 5' end of the mRNA corresponds with the transcription start site (TSS), which CAGE-Seq is able to capture accurately. Quantification of TSSs of all the genes in a transcriptome enables identification and characterization of gene promoters as well as enhancers and helps study promoter/enhancer activity, addressing a major limitation of conventional RNA-Seq. Hence, CAGE-Seq is a fairly low-cost sequencing technique that has helped in studying a multitude of mechanisms such as promoter switching, transcriptional activation/inactivation, differential promoter usage, etc.

Although RNA and CAGE-seq offer two completely different functionalities in quantifying random RNA fragments and TSSs, respectively, the combination of these two sequencing approaches can be extremely powerful in studying the transcriptome.

1.3.4 Single-cell RNA Sequencing

So far, we have discussed technologies that involve bulk-sequencing, which involves studying the gene expression of a tissue sample that consists of a heterogenous mixture of cells.

Single-cell RNA-seq (scRNA-seq) is one of the newest and most active fields of RNA-seq with a unique set of opportunities and challenges, one of which is its high cost. The first step, and most important, step in the scRNA-seq protocol is to isolate viable, single cells from the tissue of interest. Next, isolated individual cells are lysed to allow capture of as many RNA molecules as possible. In order to specifically analyse polyadenylated mRNA molecules, and to avoid capturing ribosomal RNAs, poly[T]-primers are commonly used. Next, poly[T]-primed mRNA is converted to complementary DNA (cDNA) by a reverse transcriptase. Depending on the scRNA-seq protocol, the reverse-transcription primers will also have other nucleotide sequences added to them, such as adaptor sequences for detection on NGS platforms, unique molecular identifiers to mark unequivocally a single mRNA molecule, as well as sequences to preserve information on cellular origin. The minute amounts of cDNA are then amplified either by PCR or, in some instances, by in vitro transcription followed by another round of reverse transcription—some protocols opt for nucleotide barcode-tagging at this stage to preserve information on cellular origin. Then, amplified and tagged cDNA from every cell is pooled and sequenced by NGS, using library preparation techniques, sequencing platforms and genomic-alignment tools similar to those used for bulk samples (Haque et al. 2017).

1.4 Genomewide Association Studies

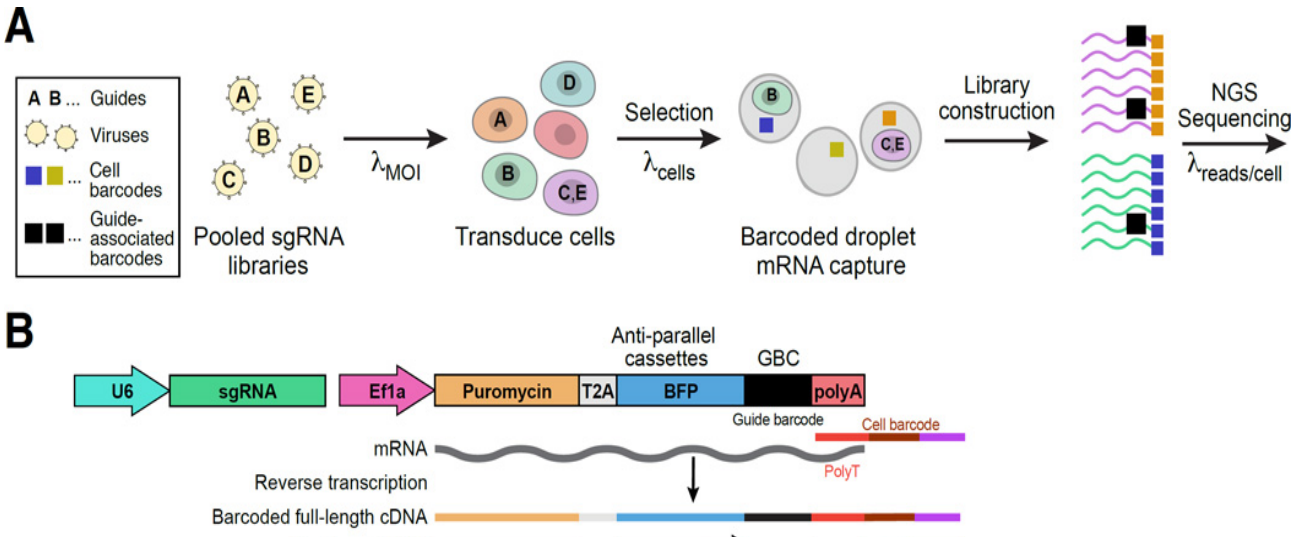
Genomewide association studies (GWAS) are powered to detect polygenic effects where allele frequencies are compared between cases and controls or associated with continuous traits. The pivotal technological advancements that enabled GWA studies were the development of microarray technology combined with growing catalogues of common human SNPs and the human reference genome. By cataloguing the common SNP pool through large scale SNP discovery and haplotype mapping, researchers did not need to assay every SNP to capture genome wide information (which contains redundant information due to linkage disequilibrium), but rather could scale down to a subset of SNPs that tagged each LD block and would fit on a single microarray, typically consisting of 200 thousand to 2 million probes. Subsequently, after genotyping a participant, the individual's patterns of genetic variation could be matched against a database of more complete haplotypes and the missing, unmeasured genetic variation could be accurately recovered through statistical

imputation [239, 45, 220]. Of course, such methods only work for common genetic variation with a minor allele frequency (MAF) > 1%; nonetheless, for common variants, these innovations enable high throughput and cost effective genotyping across thousands of participants, leading to successful mapping of polygenic traits. Despite the routine success of GWA studies, only a small fraction of the 10,000 independent variant-trait associations from GWA studies have led to the identification of specific genes or molecular mechanisms underlying complex diseases and traits. Increasing our knowledge of the effect of trait-associated genetic variation on specific genes and molecular mechanisms would enable targeted development of efficacious treatments and interventions. The knowledge gap is due to the fact that the vast majority of the GWAS loci for complex traits lie in non-coding portions of the genome.. Furthermore, sets of variants are commonly inherited in tandem, due to LD, thereby obscuring the actual causal variant (or series of causal variants) identified in the region of a GWAS association. The most recent GWAS conducted for FTD was in 2014 comparing 3526 patients with FTD and 9402 healthy controls (Ferrari et al. 2014).

1.5 CRISPRi

CRISPR interference (CRISPRi) is an RNA based method for targeted silencing of transcription in cell models. The CRISPRi system is derived from the Streptococcus pyogenes CRISPR (clustered regularly interspaced palindromic repeats) pathway, requiring only the coexpression of a catalytically inactive Cas9 protein and a customizable single guide RNA (sgRNA). The Cas9-sgRNA complex binds to DNA elements complementary to the sgRNA and causes a steric block that halts transcript elongation by RNA polymerase, resulting in the repression of the target gene. (Larson et al. 2013). Pooled CRISPR screens that couple genetic perturbations with single-cell transcriptomics (Perturb-Seq, also referred to as CROP-seq) have emerged as powerful tools to characterize the consequences of genetic perturbations. In these experiments, each cell stochastically receives one guide RNA out of a guide RNA library, enabling high numbers of perturbations to be assayed in a single experiment (A. Dixit et al. 2016). Single-cell RNA-seq is then used to retrieve the identity of the gRNA in each cell along with its effect on the transcriptome, including changes in the expression of single genes, as well as large transcriptomic rearrangements (A. Dixit et al. 2016; Datlinger et al. 2017).

FIGURE 1.3: Perturb-seq: pooled screening of transcriptional profiles of perturbations (A) Overview. (B) Perturb-seq vector. (Publication: A. Dixit et al. 2016, with permission from Elsevier)



Chapter 2

Exploring the genetic landscape of FTD in a German Cohort

2.1 ABSTRACT

The aim of this study is to investigate the frequencies of genetic variation in neurodegenerative disease genes in individual FTD/ALS cases from different parts of Germany. In addition to studying FTD/ALS genes, we aim to look at genetic variation in neurodegenerative diseases such as AD, and genes involved in pathways such as the endo-lysosomal pathway, cholesterol homeostasis pathway and autophagy.

We use an in-depth, systematic approach in studying 463 German patients to 1) identify the percentage of carriers of the pathogenic C9Orf72 HRE in the population using RP-PCR, 2) identify large INDELs in GRN and MAPT genes using MLPA, 3) identify all known pathogenic mutations in a preconceived list of 22 genes using exome-sequencing, 4) identify potentially pathogenic mutations using a step-wise genetic screening strategy and 5) study the burden of rare damaging variants human autophagy associated genes in FTD/ALS cases versus controls.

While our findings in some of the most common FTD genes - C9Orf72, GRN, MAPT, TBK1 - remain consistent with the literature, we were able to expand the genetic landscape of FTD/ALS via some unusual findings. We found pathogenic and potentially pathogenic mutations in APP, PSEN1, PSEN2 genes which are associated with AD and in the CTSF gene which is associated with Type B Kufs Disease. Finally, we were able to identify 4 human autophagy genes that carried an excessive

burden of rare damaging variants in FTD/ALS patients, with the top candidate being the SERPINA1 gene.

2.2 INTRODUCTION

The frontal and temporal lobes of the brain are the major affected areas in patients with FTD. FTD can also result from different underlying pathologies eg., tau (the protein product of the MAPT gene), TDP-43 (the protein product of the TARDBP gene), or amyloid pathology, indicating a highly complex and converging clinical and genetic landscape. There are very few studies exploring the diverse spectrum of genetic risk variants in NDD genes and the relative proportions in which they contribute to the diverse genetic architecture of FTD/ALS. There are even fewer studies comprising a German population. A study published in 2017 used exome-sequencing data from 121 unrelated FTD subjects from South Germany to study the genetic landscape of FTD (Blauwendraat et al., 2017). Here, we present the largest genetic study, to the best of our knowledge, on a German population of 463 patients from 9 different parts of Germany that lie on the FTD/ALS spectrum of disorders. These patients are all part of the DESCRIBE-FTD study which began in 2016 headed by Anja Schneider in an effort to "describe the course of FTD in its various clinical manifestations in detail, to gain a better understanding of the underlying pathology and to identify parameters that enable diagnosis and prediction of the course of the disease" (https://www.dzne.de/en/research/studies/clinical-studies/describe/describe-ftd/).

FTD has a significant genetic component, with an estimated 43% of patients carrying a positive family history [at least one affected first-degree family member with dementia, ALS, or Parkinson's disease (PD)] and between 10.2% and 27% of FTD patients have an autosomal dominant presentation of the disease (Pottier et al. 2016). Rare variants with minor allele frequencies (MAF) less than 0.01 play an important role in the etiology of complex and polygenic diseases such as FTD and ALS.

One of the most common ways to genetically test patients with a clinical FTD/ALS diagnosis is to perform a repeat primed PCR experiment to measure the length of

the C9Orf72 hexanucleotide repeat expansion. Additionally, some studies include genetic testing for GRN and MAPT using Sanger-sequencing. These, however, only explain 10% of the heritability of genetic FTD, leaving a lot of questions unanswered about the missing heritability of FTD/ALS. We use whole exome-sequencing in this study to investigate less common FTD and ALS genes such as commonly CHMP2B, TREM2, SQSTM1, FUS, TARDBP, SIGMAR1 and VCP. Additionally, we investigate new genes involved in cholestrol homeostastis, lipofuscinosis and autophagy in an attempt to shed more light into the missing heritability of FTD and ALS.

In addition to the need for a widening of the genetic landscape of FTD/ALS to account for this missing heritability, there is a need to study the manifold pathways that have been implicated in FTD to provide a more complete picture of its molecular pathogenesis.

Using methods such as repeat primed Sanger sequencing (to check for C9Orf72 repeat expansions), Multiplex ligation-dependent probe amplification (MLPA) for detecting large deletions and duplications in MAPT and GRN genes which are often undetected in exome sequencing studies, whole exome sequencing (WES) and whole genome sequencing (WGS), we provide a systematic and thorough analysis of the frequencies of mutations in NDD genes in 463 individual FTD cases from a German cohort. Lastly, we perform a proof-of-concept association study to investigate the burden of damaging variants in autophagy-associated genes in FTD/ALS cases.

2.3 METHODS

2.3.1 Subjects

A total of 463 subjects with German ancestry were recruited for this study from the Describe-FTD (https://www.dzne.de/en/research/studies/clinical-studies/describe/describe-ftd/) and DANCER-FTD (https://www.dzne.de/forschung/studien/klinische-studien/dancer/) cohorts. All of the patients are individual cases, some with positive family history. We have a total of 435 whole exome sequencing samples and 24 whole genome sequencing samples from this cohort. Three patients carrying the C9Orf72 hexanucleotide repeat expansion and one with a pathogenic MAPT mutation were

TABLE 2.1: Clinical Diagnoses of the Subjects included in this Study

Clinical Diagnosis	
bvFTD	143
bvFTD + ALS	20
bvFTD + CBS	1
IPS	1
LPA	28
PNFA	62
PPA	53
SemD (+bvFTD)	29
DANCER	80
FTD + ALS	2
ALS	42
FTD with exact diagnosis not known	3
TOTAL	**463**

DANCER-FTD, Degeneration Controls and Relatives of FTD patients; ALS, amyotrophic lateral sclerosis; bvFTD, behavioral variant; lvPPA, logopenic variant primary progressive aphasia; PNFA, progressive non-fluent aphasia; svPPA, semantic variant primary progressive dementia; SemD, semantic dementia.

not exome or genome sequenced.

The aims of the DESCRIBE-FTD study are to describe the course of FTD-ALS in detail, in their various characteristic clinical forms, to improve our understanding of the underlying pathology and to identify effective parameters for diagnosis and forecasting of the diseases' progression. This work is expected to illuminate the causes of the diseases, and to provide a basis for better therapies that can be applied at earlier stages of the diseases. For this reason, in addition to genetic data, the collection of biomaterials such as CSF, saliva, urine and blood is also being conducted. In some cases, MRI and PET scans from patients are also documented.

2.3.1.1 Clinical Characteristics

The distributions of clinical diagnoses of the subjects in the study cohort are highlighted in Table 2.1 and the sex distribution is highlighted in Table 2.2. Each individual's patient ID, sex and clinical diagnosis is presented in Supplementary Table A.1.

TABLE 2.2: Sex of the Subjects included in this study

Sex	
Male	255 (55%)
Female	208 (45%)

2.3.2 Kinship Identification Analysis

To identify potential sample swaps or kinship between the samples, the joint VCF file was converted to PLINK binary file format. Following that, the software KING (Manichaikul et al. 2010) was used to calculate pairwise kinship coefficients. Close relatives can be inferred fairly reliably based on the estimated kinship coefficient. Range >0.354, [0.177, 0.354], [0.0884, 0.177] and [0.0442, 0.0884] corresponds to duplicate/MZ twin, 1st-degree, 2nd-degree, and 3rd-degree relationships respectively. For association tests, one member of each pair with kinship coefficient >0.0884 were removed.

2.3.3 Genetic Screening Strategy

All patients were tested for pathogenic C9Orf72 hexanucleotide repeat expansion, the most common genetic cause of FTD in central Europe. Following that, multiplex ligation-dependent probe amplification (MLPA) was performed to detect large insertions and deletions in MAPT and GRN genes. Finally, protein-coding variants in FTD and other NDD genes were analysed using NGS data.

2.3.3.1 Detection of the C9Orf72 HRE

To detect the C9orf72 hexanucleotide repeat expansion, we used the AmplideX PCR/CE C9Orf72 kit designed specifically to detect the GGGGCC repeats in the C9Orf72 gene by Asuragen. The approach is based on a Repeat-Primed PCR (RP-PCR) design to profile repeat sequences in the C9Orf72 gene. A cut-off value of 30 repeats was used to define expanded repeats.

2.3.3.2 Detection of genetic deletions and duplications in GRN and MAPT genes

Subjects were screened for GRN and MAPT gene deletions and duplications using the multiplex-ligation probe amplification (MLPA) method using the SALSA MLPA probemix kit developed at MRC-Holland b.v.

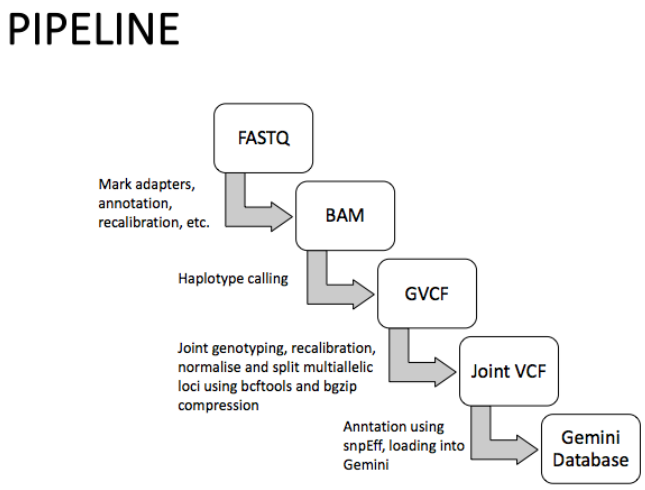


FIGURE 2.1: Steps for analysis of NGS data

2.3.3.3 Whole Exome Sequencing and Data Processing

Library preparation and sequencing were performed using Agilent SureSelect Human All Exon V7 and Illumina NovaSeq 6000, respectively. Using Picard's FastqToSam tool, raw fastq files were converted to unaligned bam files. Adapter sequences were marked using Picard's MarkIlluminaAdapters software. Ubam files were converted once again to fastq files using Picard's SamToFastq tool.

Following the GATK guidelines, raw reads were aligned to the hg19 human reference genome (ucsc.hg19.fasta) using bwa mem (v-0.7.17). BAM files were merged and sorted using Picard's MergeBamAlignment and SortSam tools, respectively. PCR duplicates were marked using Picard's MarkDuplicates and indexing was done using BuildBamIndex. Base quality scores were recalibrated using GATK's (v-4.0.8.1) ApplyBQSR function. Finally, gVCF files for each sample were generated using GATK HaplotypeCaller.

Joint genotyping was performed on all gVCF files using GATK's GenotypeGVCF function, followed by site-level filtering using GATK VariantRecalibrator and GATK ApplyRecalibration. The SNV VQSR model was trained using SNP sites from HapMap 3.3 [International HapMap, C. et al. A second generation human haplotype map of

over 3.1 million SNPs. Nature 449, 851–861 (2007).], 1000 Genomes Project (1000GP) sites found to be polymorphic on Illumina Omni 2.5M SNP arrays59, 1000GP Phase 1 high-confidence SNPs60, and dbSNP61 (v138) with a Ti/Tv ratio set to 2.8. The indel VQSR model was trained using high-confidence indel sites from 1000GP and dbSNP (v138) with VQSLOD = 99.9.

Variant normalisation was performed using Bcftools norm function to normalise indels and split multiallelic loci. Annotation to CHROM:POS:REF:ALT was performed using BCFTools annotate function.

An alternative step of manual hard filtering was performed but discarded as it led to several false negative results and is not required as per GATK Best Practices.

VCFTools was applied to include only those reads with PASS in the filter field to generate a filtered VCF file.

SNP annotation was performed using snpEFF, followed by bgzip compression. GEMINI (GEnome MINIng) software (Paila et al. 2013) was used to visualise and identify and prioritize high confidence likely pathogenic variants in known FTD genes as well as potentially novel FTD genes.

Through exome sequencing, all coding variants were analysed in the following genes:

gene == 'APP' or gene =='C9orf72 ' or gene =='CHCHD10' or gene =='CHMP2B' or gene =='FUS' or gene =='GRN' or gene =='HNRNPA1' or gene =='HNRNPA2B1' or gene =='MAPT' or gene =='OPTN' or gene =='PRKAR1B' or gene =='PSEN1' or gene =='PSEN2' or gene =='SIGMAR1' or gene =='SQSTM1' or gene =='TARDBP' or gene =='TBK1' or gene =='TREM2' or gene =='UBQLN2' or gene =='VCP' or gene =='TIA1'

Patients with 'confirmed pathogenic mutations' - confirmed using the ClinVar pathogenicity status and functional evidence in the literature - in the candidate FTD genes above were then considered "solved" for their disease phenotype but were not yet removed from further analysis.

2.3.3.4 Discovering "potentially" pathogenic mutations in FTD genes

Variants were then filtered for being non-synonymous coding variants in the list of genes described above, carrying a gnomAD (v2.1.1) minor allele frequency < 0.0005 or missing, CADD (Phred score >= 20) and predicted as damaging/deleterious by at least one of the following in-silico damage prediction algorithms: (i) SIFT, (ii) Polyphen-2. In addition, existing literature on these variants was studied to further corroborate our findings.

2.3.4 Sanger Sequencing to confirm WES findings

For every variant detected using the exome and genome sequencing data, we validated its presence using Sanger sequencing. Using the Primer3 software (Untergasser et al. 2012), primers were designed to amplify a region 250-400 base pairs in length, including the mutation being validated. PCR products were then purified with EXO-SAP and sequenced using the BigDye terminators v3 on an AI3500 system (Thermo Fisher Scientific). These sequencing data are then analysed using SeqScape software by Thermo Fisher Scientific or the CLC Workbench suite (https://digitalinsights.qiagen.com/).

2.3.5 Optimized Sequence Kernel Association Test

2.3.5.1 Pre-Processing

As we do not have enough power to check for global over-representation of genes belonging to particular pathways, we used evidence from literature to select pathways hypothesized to be associated with FTD. Since several recent studies point to an association of autophagy with FTD, we performed a proof of concept study to test the same. To test for an excessive burden of deleterious variants genes involved in autophagy in FTD, all non-synonymous variants within the autophagy gene set from the Human Autophagy Database (http://autophagy.lu/) were extracted.

All non-synonymous, protein-coding variants with MAF < 0.01 from 214 autophagy-related genes were selected for further analysis. Sex was used as a covariate in all statistical tests.

A joint VCF file with cases and controls was generated using the JointGenotyping, Recalibration and Variant and Genotype Quality-Control steps mentioned previously in section 2.3.3.3. This joint VCF file was used to perform the association tests.

2.3.5.2 Burden Tests vs Kernel-based Tests

The power to detect the association of a variant to a trait decreases as the MAF decreases. One way to overcome the limitations with statistical power for detecting effects of rare variants is by testing cumulative effects of genetic variants in genetic regions or SNP sets, such as genes. Burden tests collapse rare variants in a genetic region into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the region. Because all burden tests implicitly assume that all the rare variants in a region are causal and affect the phenotype in the same direction with similar magnitudes, they suffer from a substantial loss of power when these assumptions are violated. (Neale et al. 2011; Basu and Pan 2011).

On the other hand, kernel-based test methods, such as the sequence kernel association test (SKAT) (Lee et al. 2012), are non-burden tests. Instead of aggregating variants, SKAT aggregates individual variant-score test statistics with weights when SNP effects are modeled linearly. More generally, SKAT aggregates the associations between variants and the phenotype through a kernel matrix and can allow for SNP-SNP interactions, i.e., epistatic effects. SKAT is especially powerful when a genetic region has both protective and deleterious variants or many noncausal variants. Although SKAT provides attractive power and makes few assumptions about rare-variant effects, it has several limitations. It can be less powerful than burden tests if a large proportion of the rare variants in a region are truly causal and influence the phenotype in the same direction (Basu and Pan 2011). In addition, large-sample-based p value calculations, which SKAT uses, can produce conservative type I errors for small-sample case-control sequencing association studies, which could lead to power loss. This is especially an issue in cohorts with small sample sizes, which is common in exome-sequencing studies.

SKAT-O is an optimal unified test which automatically behaves like the burden test when the burden test is more powerful than SKAT, and behaves like SKAT when the SKAT is more powerful than the burden test. (Lee et al. 2012)

2.3.5.3 SKAT-O Analysis

Since DANCER-FTD individuals do not have a confirmed diagnosis for FTD/ALS, we removed all DANCER individuals from the study. In addition, we removed one member of each pair of related individuals with degree of relatedness > 2. In total, we were left with 442 affected cases. We ran the SKAT-O analysis using the EPACTS (Efficient and Parallelizable Association Container Toolbox) software to test which rare variants, in aggregate, were associated with the disease phenotype. A total of 1732 neurologically healthy subjects from the Rotterdam Study Exome Sequencing Database I (RSX1) were used as controls. All selected rare variants were with MAF < 0.01 and were non-synonymous coding variants i.e., missense, frameshift, stop-loss or stop-gain. Sex of the subjects was included as a covariate in all of the association tests. A conservative Bonferroni corrected p-value of 2 x 10-4 was chosen to adjust for the 214 consecutive tests being performed.

2.4 RESULTS

2.4.1 Kinship Analysis

Kinship coefficients for all pairwise comparisons between the 459 (WES and WGS) subjects were calculated. The cohort consisted of 20 pairs with degree of relatedness >2, of which 3 were duplicates or monozygotic (MZ) twins and the rest were first or second degree relatives (Table 2.3). One of each pair of duplicates/MZ twins were removed from further analysis. From the first and second degree relatives, one of each pair was removed from the Gene-wise rare variant association tests performed later in this study.

2.4.2 A brief overview of the identified pathogenic pathogenic variants

A total 463 individual patients from the Describe cohort were included in this study. Our approach included an initial screening for C9Orf72 repeat expansions using a

TABLE 2.3: Kinship Analysis Results for all individuals in the DESCRIBE-FTD and DANCER-FTD cohorts

ID1	ID2	N_SNP	HetHet	IBS0	Kinship	Degree of relatedness
1110293648	1108072661	296373	0.1921	0.0001	0.4929	duplicate/MZ twin
1108070361	1108072576	421915	0.1515	0.0001	0.4921	duplicate/MZ twin
143802329_BN	1094820054	263126	0.1774	0.0003	0.4907	duplicate/MZ twin
1108103927	1108103462	359743	0.1271	0.005	0.2897	1st-degree
1108061609	1110278837	587353	0.0716	0.0036	0.2782	1st-degree
1108062215	DNA28000A	253019	0.1099	0.0075	0.2739	1st-degree
1108070339	1108072555	306216	0.1153	0.0004	0.2655	1st-degree
1108061572	1108062553	590567	0.0627	0.0003	0.2646	1st-degree
1110275692	1108062078	604447	0.0633	0.0003	0.2645	1st-degree
1108103927	1108090096	300564	0.1066	0.0004	0.2633	1st-degree
1108061658	1108062553	592714	0.0624	0.0003	0.2626	1st-degree
1110270536	1110279932	357742	0.1196	0.0077	0.2622	1st-degree
1110278837	1110306374	330237	0.1047	0.0004	0.2612	1st-degree
1108090096	1108103462	306373	0.1126	0.0005	0.2608	1st-degree
1108061585	1094818187	340162	0.1048	0.0003	0.2599	1st-degree
1108061609	1110306374	349990	0.1043	0.0003	0.2588	1st-degree
1110271041	1110306350	364388	0.1039	0.0004	0.258	1st-degree
1108070712	1108090161	492284	0.0956	0.0067	0.2534	1st-degree
1108061572	1108061658	606956	0.0683	0.0043	0.2524	1st-degree
1110279939	1110308063	357429	0.1048	0.0005	0.2399	1st-degree

repeat primed PCR, which were identified in 21 individuals.

Following this, MLPA and exome sequencing revealed a total of 27 additional pathogenic variants (Table 2.4 and 2.5). These were also verified using Sanger sequencing.

Thus, in total we identified 50 confirmed pathogenic variants from the 21 selected NDD genes (See Methods: Genetic Screening Strategy) i.e., $\sim$ 11% of the patients carried known pathogenic mutations.

Out of these 50 subjects who carried pathogenic mutations, almost 50% were C9Orf72 repeat expansions (n = 21) and the rest were distributed across the following FTD genes: GRN (n=12), MAPT (n=6), APP (n=1), CHCHD10 (n=1), FUS (n=2), PSEN1 (n=1), SQSTM1 (n=1), TARDBP (n=1), TBK1 (n=2), CTSF (n=1) and VCP (n=2). No known pathogenic mutations were found in, CHMP2B, HNRNPA1, HNRNPA2B1, OPTN, PRKAR1B, PSEN2, SIGMAR1, TIA1, TREM2 or UBQLN2. In addition, we found a pathogenic mutation in the CTSF genes associated with neuronal ceroid lipofuscinosis, making it a total 51 subjects who were genetically solved for their disease.

It is important to note that three of the subjects carrying pathogenic mutations (two in GRN and one in MAPT) were confirmed as identical or MZ twins through our kinship analysis. One pair (1108070361 and 1108072576 carrying GRN mutation NM_002087.3:c.882T>C;p.Tyr294Ter) was confirmed as identical by the sampling lab at DZNE Bonn, while the other two many be identical repeated samples or MZ twins (1110293648 and 1108072661 carrying GRN mutation NM_002087.3;c. 675_676delCA; p.Ser226Trpfs; 143802329_BN and 1094820054 carrying MAPT mutation NM_005910.5:c.1090C>T p.Pro364Ser).

2.4.3 Potentially pathogenic variants identified in NDD genes

To further uncover the genetic landscape of FTD in the patients not carrying known pathogenic mutations, "potentially pathogenic" mutations with gnomAD MAF < 0.0005, CADD score >= 20 by at least one of the following in-silico predictive algorithms for pathogenicity: (i) SIFT, (ii) Polyphen-2 in these candidate FTD genes were

TABLE 2.4: Eight pathogenic Single Nucleotide Variations (SNVs) were identified in the GRN gene in 12 patients. The gnomAD minorallele frequencies (MAF) reported here are from exomes (v2.1.1).

VCF ID	IMPACT	REFGENE FEATURE	DIAGNOSIS	GNOMAD MAF
chr17:42428777:T:G	stop_gained	NM_002087.3:c.882T>C;p.Tyr294Ter	bvFTD, bvFTD	N
chr17:42428169:G:A	splice_donor_variant	NM_002087.3:c.708+1G>A	LPA, PNFA	0.000007985
chr17:42428134:CCA:C	frameshift_variant	NM_002087.3;c. 675_676delCA; p.Ser226Trpfs	bvFTD, bvFTD	N
chr17:42429455:C:T	stop_gained	NM_002087.4:c.1252C>T;p.Arg418Ter	PNFA, PPA	N
chr17:42426558:C:A	missense_variant	NM_002087.4 c.26 C>A p.Ala9Asp	PNFA	N
chr17:42427669:T:TA	frameshift_variant	NM_002087 c.424dupA p.Met142fs	bvFTD	N
chr17:42422705:A:G	intron_variant	NM_002087.3:c.-8+3A>G.	PPA	N
chr17:42428403:A:G	splice_acceptor_variant	NM_002087.4(GRN):c.709-2A>G	LPA	N

TABLE 2.5: Pathogenic SNVs identified in other FTD/NDD genes

GENE	VCF ID	IMPACT	REFGENE FEATURE	DIAGNOSIS	GNOMAD MAF
CHCHD10	chr22:24109646:G:A	missense_variant	NM_213720.3(CHCHD10): c.176C>T (p.Ser59Leu)	bvFTD	N
MAPT	chr17:44087784:C:T	intronic_variant affecting the splicing of exon 10	NM_005910.5(MAPT); c.915+16C>T	bvFTD	N
MAPT	chr17:44087755 C>T	missense_variant	NM_005910.5(MAPT): c.902C>T (p.Pro301Leu)	bvFTD, DANCER, bvFTD	0.0000053
MAPT	chr17:44096076 C>T	missense_variant	NM_005910.5(MAPT): c.1090C>T (p.Pro364Ser)	SemD, SemD	N
APP	chr21:27264165:C:T	missense_variant	NM_000484.4: c.2080G>A;p;Asp694Asn	bvFTD	N
FUS	chr16:31202752:C:T	missense_variant	NM_004960.3(FUS): c.1574C>T; p.Pro525Leu	ALSgen	0.000003977
FUS	chr16:31193959:ATTC:A	disruptive_inframe_deletion	NM_004960.3 c170-172 del p.Ser57Del	bvFTD	0.0001710
PSEN1	chr14:73637653:C:T	missense_variant	NM_000021.4(PSEN1): c.236C>T;p.Ala79Val	SemD	0.00001193
SQSTM1	chr5:179263439:CT:C	frameshift_variant	SQSTM1 p.Asp391fs 394ter	PPA	N
TARDBP	chr1:11082266:A:G	missense_variant	NM_007375.3(TARDBP): c.800A>G (p.Asn267Ser)	LPA	0.00007566
TBK1	chr12:64860701:C:T	stop_gained	NM_013254.4: c.379C>T; p.Arg127ter	bvFTD+ALS	N
TBK1	chr12:64891000:TGAA:T	disruptive_inframe_deletion	NM_013254.4: c.1922_1924AAG[2]; p.Glu643del	bvFTD	0.000008301
VCP	chr9:35067907:G:A	missense_variant	NM_007126.5(VCP): c.283C>T (p.Arg95Cys)	bvFTD	0.000003977
VCP	chr9:35065355:C:G	missense_variant	NM_007126.5: c.469G>C p.Gly157Arg	bvFTD+ALS	N

analysed. In addition, literature on these variants was studied to further corroborate our findings. Twenty-three patients carried 21 variants of interest in the following FTD/NDD genes: APP (n=2), CHMP2B (n=1), MAPT (n=3), FUS (n=1), SIGMAR1 (n=1), SQSTM1 (n=3), TREM2 (n=1), TIA1 (n=1), PSEN1 (n=2), PSEN2 (n=2), PRKAR1B (n=1), TBK1 (n=3). Details on each variant are described in Table 2.6. Three patients (one with bvFTD and two from the DANCER-FTD cohort) carried a missense variant in CHMP2B (rs149380040; NM_014043.4(CHMP2B):c.581C>T (p.Ser194Leu), predicted by SIFT as deleterious) first reported by (Ghanim et al. 2010) where is was also carried by an FTD patient. It has also been hypothesized to be pathogenic in other studies (Blue et al. 2018), although for AD and not FTD. Two patients with ALS (one with ALSbi and one with ALSci) carried a missense variant in TREM2 (rs142232675, NM_018965.3 (TREM2):c.259G>A (p.Asp87Asn), predicted by Polyphen-2 as damaging). This variant has multiple times been associated with Alzheimer's Disease (AD) (Guerreiro et al. 2013), (Jin et al. 2015). An enrichment of rare variants in TREM2 has been observed in both FTD and AD patients (Cuyvers et al. 2014). In addition to these and consistent with previous reports on genetic FTD, we found 6 potentially pathogenic mutations in MAPT and TBK1. We found three potentially pathogenic mutations in MAPT: three missense variants NM_016835.4(MAPT):c.664C>A(p.Arg222Ser), rs1463829855; 44061110G>C;p.(Glu314Gln) and rs763728305; NP_058519.3:p.(Pro494Leu) and one frame-shift variant rs953116486; NP_058519.3:p.(Gly144fs). In TBK1, we found one missense variant rs576726084; NP_037386.1:p.(Asn22His) and one splice acceptor variant NM_013254.3:c.(1644-5_1644-2del). We also found potentially pathogenic variants in genes less commonly associated with FTD: APP, FUS, PRKAR1B, SQSTM1, SIGMAR1, PSEN1, PSEN2, and TIA1. Interestingly, one patient with PNFA carried two potentially pathogenic missense variants in the SQSTM1 gene.

In addition, genes associated to lipofuscinosis (CTSF) and cholesterol homeostasis (CYP27A1) pathways have been recently associated to FTD (Blauwendraat et al., 2017). We checked for potentially damaging mutations in these genes and identified four missense variants (Table 2.7) of interest: NM_003793.4(CTSF):c.1133A>G (p.Asn378Ser), NM_003793.4(CTSF):c.160C>G (p.Arg54Gly), NM_003793.4(CTSF):c.692A>G (p.Tyr231Cys) and NM_000784.4(CYP27A1):c.491G>A (p.Arg164Gln). Out of these, the mutation NM_003793.4(CTSF):c.692A>G (p.Tyr231Cys) carried by a bvFTD patient in the DESCRIBE-FTD cohort is a confirmed pathogenic mutation identified in

TABLE 2.6: Potentially pathogenic variants found in FTD or ND-Dgenes in the DESCRIBE-FTD patient and DANCER cohort

GENE	VCF ID	Impact	REFGENE FEATURE	INDIVIDUAL ID	GNOMAD AF
APP	chr21:27394296:GCTT:G	disruptive_inframe_deletion	NC_000021.8:g.27394298_27394300TTC; p.Glu241del	1094813058	0.00008075
	chr21:27348341:C:T	missense_variant	NM_000484.4:c.1225G>C (p.Val409Leu)	1108061605	0.000003982
CHMP2B	chr3:87302911:C:T	missense_variant	NM_014043.4(CHMP2B):c.581C>T (p.Ser194Leu)	1108090096,1108103462, 1108103927	0.00005191
FUS	chr16:31201424:G:A	missense_variant	NC_000016.9:g.31201424G>A; p.Arg378Gln	1108060751	0.00003185
MAPT	chr17:44060592:G:GC	frameshift_variant	NC_000017.10:g.44060598dup; p.Gly144fs	1108070820	AC0
	chr17:44060834:C:A	missense_variant	NM_016835.4(MAPT):c.664C>A (p.Arg222Ser)	1108103438	0.0002801
	chr17:44068926:C:T	missense_variant	NC_000017.10:g.44068926C>T; p.Pro494Leu	1108061237	0.000003980
PRKAR1B	chr7:751042:C:T	missense_variant	NC_000007.13:g.751042C>A; p.Cys34Tyr	1110306398	0.00005567
PSEN1	chr14:73614806:C:G	missense_variant	p.Arg27Gly	1110306359	N
	chr14:73659482:A:C	missense_variant	p.Ile227Leu	1108061596	0.000007
PSEN2	chr1:227069693:C:T	missense_variant	NC_000001.10:g.227069693C>T; p.Arg62Cys	1110296425	0.00001991
	chr1:227073369:C:T	missense_variant	NM_000447.3(PSEN2):c.487C>T (p.Arg163Cys)	1108062212	0.000008086
SIGMAR1	chr9:34637322:A:G	missense_variant	NM_005866.4(SIGMAR1):c.247T>C (p.Phe83Leu)	1094818144	0.00001294
SQSTM1	chr5:179260077:G:A	missense_variant	NM_003900.5:c.800G>A; p.Arg267His	1110278743	0.00007159
	chr5:179248075:C:G	missense_variant	NM_003900.5(SQSTM1):c.139C>G (p.Leu47Val)	1110290332	0
	chr5:179248090:G:T	missense_variant	NM_003900.5(SQSTM1):c.154G>T (p.Ala52Ser)	1110290332	0
TBK1	chr12:64849714:A:G	missense_variant	NC_000012.11:g.64849714A>C; p.Asn22Asp	DNA27716A	0.000003985
	chr12:64889471:TTAAA:T	splice_acceptor_variant	NM_013254.3(TBK1):c.1644-5_1644-2del	DNA28066A	0.00009677
	chr12:64858199:A:G	missense_variant	NM_013254.4(TBK1):c.314A>G (p.Tyr105Cys)	1108062060	0.000004007
TIA1	chr2:70439871:C:T	missense_variant	NC_000002.11:g.70439871C>T	Proband_29	0.000003977
TREM2	chr6:41129133:C:T	missense_variant	NM_018965.3(TREM2):c.259G>A (p.Asp87Asn)	1108103532, 1108103500	0.0009702

AC0 in the GNOMAD MAF column indicates that allele count in gnomAD datasets is zero i.e., no high confidence genotype was found for this variant.

TABLE 2.7: Variants found in the CTSF and CYP27A1 genes belonging to the lipofuscinosis and cholestrol homeostatis pathways, respectively.

GENE	VCF ID	IMPACT	REFGENE FEATURE	DIAGNOSIS	GNOMAD MAF
CTSF	chr11:66332390:T:C	missense_variant	NM_003793.4(CTSF):c.1133A>G (p.Asn378Ser)	Proband_27	0.0001804
CTSF	chr11:66335798:G:C	missense_variant	NM_003793.4(CTSF):c.160C>G (p.Arg54Gly)	1108071469	0.00003166
CTSF	chr11:66333791:T:C	missense_variant	NM_003793.4(CTSF):c.692A>G (p.Tyr231Cys)	1110293601	0.00004773
CYP27A1	chr2:219676989:G:A	missense_variant	NM_000784.4(CYP27A1):c.491G>A (p.Arg164Gln)	1108070740	0.0002824

For the variant NM_003793.4(CTSF):c.160C>G (p.Arg54Gly) the gnomAD MAF for exomes is AC0 and hence here the MAF from the genomes is reported. This variant was only found in one female of Non-Finnish European descent.

cases of adult-onset neuronal ceroid lipofuscinosis (Smith et al. 2013).

2.4.4 Rare-variant Association Analysis

The Human Autophagy Database (HADb) consists of 214 human genes involved directly or indirectly in autophagy, as described in the literature. A SKAT-O test was performed using EPACTS software to test which rare variants, in aggregate, were associated with the disease phenotype. A total of 1732 neurologically healthy subjects from the Rotterdam Study Exomes I (RSX1) were used as controls. All selected rare variants were with MAF < 0.01 and were non-synonymous coding variants i.e., missense, frameshift, stop-loss or stop-gain. Sex of the subjects was included as a covariate in all of the association tests. A conservative Bonferroni corrected p-value of 2 x 10-4 was chosen to adjust for the 214 consecutive tests being performed. A

TABLE 2.8: SKAT-O Analysis to study the burden of deleterious variants in Human Autophagy genes in FTD/ALS patients

CHR	BEGIN	END	GENE	NS	P-VALUE	STATRHO
14	94844843	94849388	SERPINA1	2174	3.69E-12	0
5	78076288	78280974	ARSB	2174	1.46E-06	0
11	64662597	64684483	ATG2A	2174	3.91E-06	0.2
14	62187212	62213683	HIF1A	2174	0.000024387	0

NS: total number of subjects tests out of which 1732 were controls and 442 were FTD/ALS patients; STATRHO: The ratio rho of 1 corresponds to a pure Burden test, and a ratio rho of 0 corresponds to purely an (original) SKAT test.

significant association of rare and non-synonymous variants in SERPINA1, ARSB, ATG2A and HIF1A with FTD was identified (Table 2.8).

2.5 DISCUSSION

The main goals of this study were tripartite. Firstly, to genetically solve a cohort of 463 German patients for their FTD/ALS clinical diagnosis, or, characterize them as C9orf72 repeat-negative and negative for known pathogenic and potentially pathogenic variants in FTD and NDD genes. Secondly, to elucidate the wide genetic landscape of FTD/ALS in a Germany wide population. Finally, to study the association of damaging rare variants in human autophagy-related genes with FTD cases versus controls.

Our study started with an in-depth systematic genetic analysis of the FTD/ALS spectrum, which comprises a heterogenous group of neurodegenerative disorders. There are multiple proteinopathies associated with FTD-ALS with TDP-43 proteinopathies being the most common in genetic FTD. There are majorly a result of C9orf72 hexanucleotide repeat expansions (HREs). The first step, hence, to solve these patients for their disease phenotype was to screen for C9orf72 HREs with repeat length > 30 and pathogenic variants in MAPT and GRN were screened using MLPA.

51 out of 463 i.e., 11% of the subjects were genetically solved for their disease phenotype using our genetic screening strategy. The heritability of FTD varies between 10-40% depending on the clinical variant of FTD/ALS. As expected, the most commonly found cause for FTD/ALS in the cohort was the presence of C9orf72 HREs.

The second most common genetic causes of FTD are pathogenic mutations in MAPT and GRN genes. Consistent with this, we identified 12 patients with pathogenic GRN mutations and 5 with pathogenic MAPT mutations.

Three of the five MAPT mutations were the P301L mutation in exon 10, which is the most common disease-causing MAPT mutation. It has been shown that P301L mutation carriers typically present with the symptoms of bvFTD. A recent study (Clarke et al. 2021) implicated the early involvement of the anterior cingulate in presymptomatic P301L mutation carriers. Since the function of the anterior cingulate is to modulate attention and executive functions by influencing response selection, and lesions of the anterior cingulate have produced inattention and apathy (To et al. 2017). The anterior cingulate is also thought to play a critical role in social cognition via contextual integration and evaluating the behaviour of others (Apps, Rushworth, and Chang 2016). Apathy, executive dysfunction and social cognitive impairment are all core symptoms for the diagnosis of bvFTD. Consistent with this, two of the three P301L mutation carriers in the Describe cohort were diagnosed with bvFTD while the third one belongs to the DANCER group and is a first degree relative of one of the former, confirmed through our kinship analysis. A recent study showed that the variability in age at onset and at death in MAPT mutation carriers is highly correlated to family membership (Moore et al. 2020). Considering this and also the fact that pathogenic MAPT mutations are usually fully penetrant, it is likely that the asymptomatic P301L mutation carrier is the offspring of their symptomatic relative. Magnetic Resonance Imaging (MRI) of such a patient would help better understand the patterns of atrophy in asymptomatic P301L mutation carriers.

Interestingly, some patients also carried pathogenic and potentially pathogenic mutations in APP, PSEN1 and PSEN2 genes which are typical Alzheimer's disease genes. While PSEN1 and PSEN2 mutations have been linked to FTD in the past, these associations are rare (Raux et al. 2000) and likely a product of a misdiagnosis of AD.

Amongst other genes uncommonly associated with FTD that were tested were CTSF (Cathepsin F, a lysosomal protease) and CYP27A1 (cytochrome P450 oxidase, involved in cholesterol homeostasis). We found a pathogenic missense mutation in

CTSF, NM_003793.4(CTSF):c.692A>G (p.Tyr231Cys), which is an exon 5 substitution located within the I29 propeptide inhibitor domain.This mutation affects a highly conserved amino acid and is predicted to be damaging by SIFT and 'probably damaging' by PolyPhen-2. It has previously been seen in a patient with Kuf's disease which is an adult-onset neuronal ceroid lipofuscinosis (Smith et al. 2013).

Commonly, in a genetic screen for FTD/ALS patients, the genes examined are C9orf72, GRN, MAPT and TBK1. The less commonly FTD/ALS genes that may be considered in a more extensive screen are CHMP2B, TREM2, SQSTM1, FUS, TARDBP, SIGMAR1 and VCP. Here, we show that the genetic and molecular landscape of clinical FTD/ALS is much wider. We found pathogenic and potentially pathogenic mutations in genes involved in mitochondrial function (CHCHD10), lysosomal pathways (CTSF), cholesterol homeostasis (CYP27A1), apoptosis and stress granule dynamics (TIA1), amyloid pathology and other forms of dementia (APP, PSEN1, PSEN2, PRKAR1B). There are several recent and older studies showing that the susceptibility to the frontotemporal dementias can be the result of an insult to a number of key cellular pathways. In learning the complexity of the pathogenic and molecular mechanisms of FTD and ALS, we uncover more potential molecular therapeutic targets that can facilitate early intervention and aid in improving the quality of life of at-risk individuals.

In our cohort, we did not find any pathogenic or potentially pathogenic mutations in UBQLN2 which is a human autophagy gene linked to FTD/ALS (Deng et al. 2011). However, the mechanism through which UBQLN2 causes FTD/ALS remains unclear. A recent study suggested that mutations in UBQLN2 impede autophagy by reducing autophagosome acidification through loss of function (Wu et al. 2020). As a proof of concept study, we performed an association test using SKAT-O analysis to study the burden of rare damaging variants in genes involved in human autophagy in FTD cases. Autophagy has been linked to FTD in several recent studies, including the most recent Genome-wide Association Study for FTD (Ferrari et al. 2014).

It is important, in an association test, to account for population stratification and cryptic relatedness that can confound the study and produce false positive or negative results. As we do not know the family history of our patients, and they belong

to a highly homogenous cohort, we performed a kinship analysis. One member of each pair of related samples with degree of relatedness lesser than or equal to 2, was removed. In addition, all DANCER subjects were removed from the case cohort, despite some being carriers of pathogenic mutations, as they were clinically well at the time of the study. Sex was included as a covariate in the study to correct for confounding due to sex-specific genetic elements. Critically, the implementation of association tests for joint consideration of Autophagy genes significantly improves statistical power over single gene and variant tests (Maeda, Otomo, and Otomo 2019; Zuk et al. 2014). Using a highly conservative Bonferroni corrected p-value threshold, 4 autophagy-related genes were found to be significantly associated with FTD: SERPINA1, HIF1A, ATG2A and ARSB.

There have been several speculative studies concerning SerpinA1, which is a serine protease inhibitor, and its role in neuroinflammation and neurodegeneration. A study published in 2020 showed SerpinA1 upregulation and post-translational modifications may be a common feature for several neurodegenerative disorders (Abu-Rumeileh et al. 2020), whereas under normal conditions this gene is strictly downregulated throughout the body. SerpinA1 protein has also shown to interact with several ALS-associated molecules, including FUS (Ebbert et al. 2017). In a recent gene-based association study, SERPINA1 was shown to be associated with progressive non-fluent aphasia (PNFA) but not with bvFTD (Mishra et al. 2017). This was also seen in the FTD GWAS study from 2014 (Ferrari et al. 2014). It may be interesting to test if SERPINA1 mutations are specific to the PNFA variant of FTD. This makes SERPINA1 a strong candidate in future FTD genetic screens.

As a survival mechanism, hypoxic conditions can be a trigger for autophagy induction. Transcription factor HIF1A is capable of transcriptional activation of a number of genes involved in angiogenesis, erythropoiesis (eg., VEGF and erythropoietin) and autophagy. The HIF1A-target gene Bnip3 (BCL2/adenovirus E1B 19 kDa interacting protein 3) encodes a putative BH3-only (BCL2 homology domain 3-only) protein that is necessary and sufficient to induce autophagy by competitively binding BCL2 (B-cell CLL/lymphoma 2) and disrupting the BCL2–BECN1 interaction (Bellot et al. 2009). It would be fruitful also to investigate the roles of angiogenesis and erythropoiesis in FTD/ALS due to the strong signal from HIF1A in our SKAT-O

analysis.

Mutations in ARSB, which is a lysosomal enzyme N-acetylgalactosamine-4-sulfatase (arylsulfatase B, ARSB) are known to cause defects in the autophagic pathway, ubiquitination and mitochondrial function (Tessitore, Pirozzi, and Auricchio 2009). This gene was also a candidate in a functional prioritization and SKAT-O study done for Parkinson's disease((Robak et al. 2017; Jansen et al. 2017) once again indicating a convergence of cellular pathways for a host of neurodegenerative diseases.

Lastly, the ATG2A (Autophagy-related gene 2), is a lipid transfer protein that aids in the transfer of lipids between membranes (Maeda, Otomo, and Otomo 2019). Lipid metabolism and autophagy are closely linked mechanisms that are worthy candidates for further investigation as insults to these pathways could impact FTD/ALS risk.

Usage of biomaterials collected as a part of the DESCRIBE-FTD study, especially serum GRN levels (for mutation and non-mutation carriers), amyloid-beta levels as well as white and grey matter changes should be examined for a more composite picture of clinical FTD/ALS.

Thus, we present the largest to-date genetic analysis of Germany wide FTD/ALS patients. The frequency of C9Orf72 mutations (5.5%) was comparable to that is seen previously in the study by Blauwendraat et al. in 2017 which consisted of 121 German FTD patients. In a pan-European study published in 2013 consisting of both sporadic and familial cases of FTD/ALS (van der Zee et al. 2013), frequencies of pathogenic C9Orf72 HRE in different populations were examined and found to be an overall frequency ranging between 6.09% and 7.86% for Belgium, Italy and Portugal, which is close to the average overall European frequency of 8.38%. However, a marked enrichment was observed in the Spanish (25.49%) and Swedish (21.33%) patient cohorts. In contrast, in this study, the frequency of C9Orf72 pathogenic expansions in the German patients was only 3.52%.

For GRN, we saw a frequency of 2.8%, which is lower than previously reported German frequencies of 5.8% (Blauwendraat et al. 2017) and those from other cohorts:

UK 8.4% (Rohrer et al. 2009), Dutch 4% (Bronner et al. 2007), French 4.8% (Le Ber et al. 2007). In the Belgian FTD cohort, GRN mutations explain a significantly higher number of genetic FTD cases $\sim$ 10% (Cruts et al. 2006).

In MAPT, we saw a mutation frequency of 1.6% which is comparable to those seen in Sweden, US and France at 0 (Fabre et al. 2001), 1.2% (Huey et al. 2006) and 2.9% (Le Ber et al. 2007) respectively but significantly lower than that seen in the Dutch population population at 17.8% (Rizzu et al. 1999). In the German study by Blauwendraat et al. 2017, no MAPT mutations were found.

For TBK1, we saw a mutation frequency of 0.5%, which is comparable to pan-European frequencies (van der Zee et al. 2017) but lower than those seen in Belgian cohorts at 1.7% (Gijselinck et al. 2015). Similar to MAPT, in the German study by Blauwendraat et al. 2017, no TBK1 mutations were found.

It is evident that mutation frequencies vary greatly between populations, which could both be a result of ascertainment or sampling bias or true biological differences. Since very few high powered genetic studies exist, it is difficult to make conclusive remarks on population frequencies of even the most common FTD genes i.e., C9Orf72, MAPT and GRN, yet our study makes an effort to fill this gap in the literature for a Germany wide population, with patients recruited from 9 different centres in Germany, aiding in correcting the effects of a sampling bias. All mutations were confirmed again using Sanger-sequencing and those that cannot be detected easily via exome-sequencing such as larger insertions and deletions were confirmed using MLPA. These frequencies would be useful in planning future genetic screens as well as epidemiological studies for patients on the FTD/ALS spectrum.

Chapter 3

Genetic Landscape of FTD/ALS in a broader Western European Population

3.1 ABSTRACT

This study uses next generation sequencing to study the genetic landscape of FTD/ALS patients in a western european population. In addition to studying known FTD/ALS genes, we study other less common FTD/ALS genes as well as genes associated with amyloidosis, cholestrol homeostasis and lipufuscinosis. As a large portion of genetic FTD/ALS remains unexplained, we also perform a genomewide rare variant association study to study gene-wise burden of rare damaging variants in FTD/ALS patients versus controls. To this effect, we study the pathways these genes belong to. As a proof of concept, we also performed a rare variant association study for human autophagy genes.

Key takeaways from the genetic screen are that the leading causes of genetic FTD, after the C9Orf72 HRE, remain pathogenic mutations in the GRN, MAPT and TBK1 genes. We found an unexpectedly high number of potentially pathogenic mutations in the SQSTM1 gene, which is a gene integral to the human autophagosome. We also find both pathogenic and potentially pathogenic mutations in genes involved in amyloidosis, cholestrol homeostasis and lipufuscinosis. This points to a convergence of varied pathways and disease mechanisms that could drive frontotemporal cortex pathology in FTD/ALS patients, as has been hypothesized previously.

Finally, as a result of our findings from the rare variant association studies and NGS analysis, we propose SERPINA1 as a candidate for future genetic screens for FTD/ALS.

3.2 INTRODUCTION

So far, we have looked at the genetics of FTD/ALS in a purely German population, as a part of the DESCRIBE-FTD study. In this chapter, we explore FTD in a wider Western European population, albeit still very homogenous. Here, we include participants from (i) the Genetic Frontotemporal dementia Initiative (GENFI) https://www.genfi.org/, (ii) Risk and Modifying factors for Frontotemporal Dementia (RiMod-FTD) project https://www.rimod-ftd.org/, (iii) a strictly consecutive study conducted at the Department for Neurodegenerative Diseases, Center for Neurology, Tübingen, Germany, from 2009 to 2014 for unrelated patients as well as the (iv) DESCRIBE-FTD study.

The GENFI consortium currently consists of sites across the UK, Netherlands, Belgium, France, Spain, Portugal, Italy, Germany, Sweden, Finland and Canada. In our study, we include samples collected at the Barcelona (Spain) and Coimbra (Portugal) sites. The RiMod-FTD study aims to generate a multi-omics data resource for in-depth research on FTD and its molecular mechanisms. Here, we focus on analysing genetic data from patients with a clinical diagnosis of FTD/ALS from the sites in the UK, Spain, Italy, the Netherlands and France. There is no known cure for FTD. All of these studies and initiatives have a collective goal to develop markers which will help identify the disease at its earliest stage as well as markers that allow the progression of the disease to be tracked. This can be done by studying molecular mechanisms and pathways that may be insulted in FTD/ALS patients as well as studying the genetic landscape of FTD, as was explored in our previous chapter. Here, we increase the scope of our work with a large, more varied cohort in not only studying the genetic landscape, but also performing a genome-wide association analysis to uncover pathways associated with FTD/ALS.

3.3 METHODS

3.3.1 SUBJECTS

3.3.1.1 Cases

Since C9Orf72 HRE carriers were not exome sequenced for all of the different cohorts, hereon we will be excluding this mutation from our results to avoid a misrepresentation of the % of carriers. Our combined sequencing approach yielded a total of 27 confirmed pathogenic variants in the 23 analyzed FTD/ALS and other dementia genes, identified in 34 different individuals (Table 3.2). In addition, 31 different 'potentially pathogenic' mutations were found in 37 different patients (Table 3.3).
In addition to the 463 subjects from the DESCRIBE-FTD cohort described in chapter 1, we included an additional 371 clinical FTD/ALS patients from 8 other Western European cohorts (Table 3.1). The DZNE cohort consisted of 110 unrelated FTD subjects of Caucasian ancestry (over 90% from Southern Germany) who were recruited at the Department for Neurodegenerative Diseases, Center for Neurology, Tübingen, Germany, from 2009 to 2014. All subjects were clinically diagnosed with FTD according to international consensus criteria (Neary et al. 1998) The UK cohort consists of 27 clinical FTD subjects exome sequenced by the Institute of Prion Diseases, MRC Prion Unit at UCL in a study headed by Dr Simon Mead. The Clarimon cohort samples were collected in Spain at the Genetics of Neurodegenerative Disorders Unit at Hospital de la Santa Creu Sant Pau headed by Dr Jordi Clarimon and Dr Oriol Dols-Icardo(Dols-Icardo et al. 2018). The Netherlands/Dutch cohort consists of 57 samples with patients on the FTD/ALS spectrum of disorders, in a study supervised by Dr John Van Swieten. There are 37 samples in the French cohort, collected as part of a study headed by Dr Isabelle Le Ber. All 29 of the subjects in the Italian cohort consist of a positive family history for NDD, these were sequenced under a project headed by the German Center for Neurodegenerative Diseases (DZNE), Tuebingen. The UK, Dutch, French, Clarimon and Italian cohorts are all part of the multinational RiMod-FTD study (https://www.rimod-ftd.org/). The GENFI-Barcelona and GENFI-Coimbra projects were headed by Dr. Raquel Sanchez-Valle and Dr Isabel Santana, respectively.

TABLE 3.1: Clinical information for the 371 subjects included in this study in addition to the individuals from the DESCRIBE-FTD and DANCER-FTD cohorts

Cohort	DZNE	UK	The Netherlands	France	Italy	GENFI - Barcelona	GENFI - Coimbra	Clarimon
Total number of samples	**110**	**27**	**57**	**37**	**29**	**13**	**44**	**54**
Family history								
Familial	22	20	47	33	29	12	39	10
Sporadic	74	6	2	-	-	-	-	44
Unknown	14	1	8	4		1	5	-
Age at onset								
Mean	62.33	53.72	59.09	62.33	64	62.7	65.05	62.1
Median	62	52.5	59.9	64.5	66	63	67	-
SD	11.89	8.1	8.71	11.48	6	10.45	7.84	-
Initial diagnosis								
FTD	73	27	55	37	29	12	40	13
ALS	6	-	-	-	-	-	-	29
FTD-ALS	1	-	-	-	-	1	3	12
Other	-	-	2	-	-	-	1	-
Secondary diagnosis								
bvFTD	29	15	37	8	18	4	24	-
PFNA	23	2	8	-	5	-	-	-
PPA	2	-	-	-	1	8	-	-
SD	4	5	8	1	4	-	-	-
FTD-MND	10	1	4	11	1	1	-	-
FTD-PSP	1	-	-	17	-	-	-	-
Parkinsonism	7	-	8	17	-	-	3	-
FTD-ataxia	4	-	-	-	-	-	4	-
FTD-AD	2	-	-	-	-	-	3	-
FTD-CBS	-	4	-	-	-	-	3	-

3.3.1.2 Controls

The control dataset consisted of 1732 exomes from the Rotterdam Study Exome Sequencing Database (RSX1), which is a prospective cohort study ongoing since 1990 in the city of Rotterdam in the Netherlands (Hofman et al. 2013).

3.3.2 Data Pre-Processing and Quality Control

Sequence alignment and variant calling: Sequence reads were aligned to the human reference genome (hg19) using Burrows Wheeler Aligner (BWA) for short read sequencing (Li and Durbin 2009). The data was sorted, index and PCR duplicates marked using Picard tools from Broad Institute. Variant calling and SNP/INDEL recalibration was performed using the GATK Best Practices pipeline as described in Chapter 1.

A combined variant call file (VCF) was generated for all exomes including the FTD/ALS cases, DANCER cohort and RSX1 (n=2562) using GATK's GenotypeGVCFs function, followed by by site-level filtering using GATK VariantRecalibrator and GATK ApplyRecalibration. The SNV VQSR model was trained using SNP sites from HapMap 3.3 [International HapMap, C. et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861 (2007).], 1000 Genomes Project (1000GP) sites found to be polymorphic on Illumina Omni 2.5M SNP arrays59, 1000GP Phase 1 high-confidence SNPs60, and dbSNP61 (v138) with a Ti/Tv ratio set to 2.8. The indel VQSR model was trained using high-confidence indel sites from 1000GP and dbSNP (v138) with VQSLOD = 99.9.

Variant normalisation: Variant normalisation was performed using Bcftools norm function to normalise indels and split multiallelic loci. Annotation to CHROM:POS:REF:ALT was performed using Bcftools annotate function.

Quality control: Finally, GenotypeQC was performed for the VCF file to remove all sites with MAC (minor allele count) < 1, DP <10 and GQ <20 using VCFTools (Danecek et al. 2011).

3.3.3 SNP/INDEL Annotation and Detection

Variant Annotation: SNP/INDEL annotation was performed using snpEFF, followed by bgzip compression.

Known gene analysis: Following annotation, a SQLite database was generated for the variants using GEMINI (GEnome MINIng) (Paila et al. 2013). This database was used to visualise and identify and prioritize variants of interest as described in the Genetic Screening Strategy for confirmed and potentially pathogenic variants in NDD genes. These NDD genes were 'APP, C9orf72, CHCHD10, CHMP2B, CTSF, CYP27A1, FUS, GRN, HNRNPA1, HNRNPA2B1, MAPT, OPTN, PRKAR1B, PSEN1, PSEN2, SIGMAR1, SQSTM1, TARDBP, TBK1, TREM2, UBQLN2, VCP AND TIA1.

3.3.4 Gene-wise Association Analyses

Data pre-processing: For the rare-variant association study, 17 cryptically related samples identified by our kinship analysis were removed. In addition, 80 subjects from the DANCER cohort were also removed, despite some being carriers of pathogenic FTD/ALS mutations to ensure a dataset with clear case and control status. In total, we had 745 cases and 1732 controls for the rare-variant association tests.

The joint VCF file was annotated using the EPACTS 'anno' function. This adds the "ANNO = [function]:[genename]" entry into the INFO field based on the gencode V7 database. A marker group file containing the list of markers per group using the EPACTS 'make-group' function. The variants were filtered for being non-synonymous, more specifically stop loss/gain, frameshift, missense and splice variants. The group file consisted of a total 20205 genes. A ped file was generated for the 2477 case and control samples.

Optimal Sequence Kernel Association Test (SKAT-O): A genomewide SKAT-O test was performed for rare variants (MAF < 0.01) using EPACTS. Sex was included as a covariate. Principal component analysis (PCA) was performed using PLINK (Purcell et al. 2007) to check for population stratification. All genes that passed the significance threshold of a Bonferroni corrected p-value of 2 x 10-6 (0.05/20205) were

reported.

Studying Autophagy genes and their association with FTD/ALS: A list of 214 human autophagy-related genes was obtained from the Human Autophagy Database (http://autophagy.lu/) which is developed at the Laboratory of Experimental Cancer Research headed by Dr Guy Berchem. Using this list, a Autophagy specific marker group file was generated from the main genome wide marker file for non-synonymous mutations: stop loss/gain, frameshift, missense and splice variants. SKAT-O using EPACTS was run on these 214 autophagy-related genes, as a proof-of-concept study. All genes that passed the significance threshold of a Bonferroni corrected p-value of 2 x 10-3 (0.05/214) were reported.

3.3.5 Replication Dataset

The replication cohort consisted of 2451 subjects with a clinical FTD/ALS diagnosis and 4029 controls. Patients with FTD were diagnosed according to the Neary criteria (Neary et al. 1998) or the Movement Disorders Society criteria (Höglinger et al. 2017) for PSP. Patients with ALS were diagnosed according to the El Escorial criteria (Ido et al. 2021). All participants included in the aged, healthy control cohort were free of neurological disease based on a history and neurological examination (mean age = 77.0 years of age at collection, interquartile range = 69.0 - 86.0). All participants were of a European ancestry. The genomic DNA was extracted from whole blood or cerebellar brain tissue and PCR-free, paired-end, non-indexed libraries were conducted. Whole genome sequencing was performed using Illumina HiSeq X Ten sequencer using 150 base pair (bp) paired-end cycles. These data were entirely produced by the National Institute of Health (NIH), Bethesda, MD.

The data was processed using GATK (2016) Best Practices, implemented in the workflow description language (WDL). SNPs and INDELs were called from the processed data using GATK Best Practices workflow for joint discovery and Variant Quality Score Recalibration (VQSR). The average sequencing read-depth after filtering by alignment quality was 35x and the mean coverage per genome was 36.3.

Rare variant association tests were performed for variants with MAF < 0.01 and MAC > 3 using SKAT-O for the four following categories:

- Missense (i.e. only exonic variants). Number of genes tested = 17595
- Loss-of-function (LOF) (i.e. stop, frameshift, splice variants). Number of genes tested = 14228
- Missense and LOF (i.e. variants from group 1 + group 2). Number of genes tested = 18693
- CADD score > 12 (all non-synonymous variants with CADD score > 12). Number of genes tested = 3152

Covariates adjusted for were sex, consensus_age, PC1, PC2, PC3, PC5, PC8.

3.3.6 Using a Genomewide Association Study For FTD as Validation For Our Findings

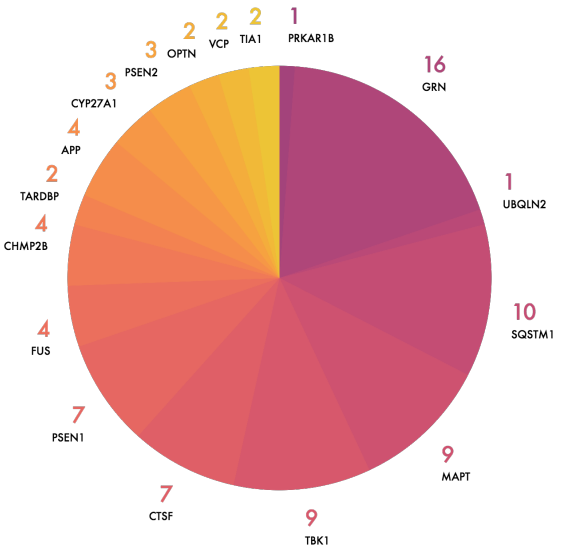
The most recent GWAS for FTD was performed in 2014 which analysed samples from 3562 patients with FTD and 9402 neurologically healthy controls (Ferrari et al. 2014). The study discovered loci linked to immune function, lysosomal biology and autophagy associated significantly or of suggestive significance with FTD. In addition to a combined analysis, the study conducted individual analyses for subtypes of FTD: bvFTD (1377 patient samples versus 2754 control samples), SemD (308 cases versus 616 controls), PNFA (269 cases versus 538 controls) and FTD-MND (200 cases versus 400 controls). It is important to consider here that the number of samples are much too small for a highly powered study and to get genome wide significant results. For this reason, the study also reports several SNPs that showed suggestive p-values (between 1x10-6 and 1x10-7).

3.4 RESULTS

3.4.1 Pathogenic and Potentially Pathogenic Variants identified in our subset of FTD and NDD genes

Presented here are all of the mutations that passed the criteria mentioned in the 'Genetic Screening Strategy' of Chapter 2 for the identification of pathogenic as well as 'potentially pathogenic' variants.

FIGURE 3.1: Pathogenic and potentially pathogenic variants in NDD genes in 831 European individuals.



Relative frequencies of mutations in neurodegenerative disease (NDD) genes in a cohort of 831 subjects on the clinical FTD/ALS spectrum, including 80 subjects from the DANCER cohort of DZNE. C9orf72 expansions are excluded as in some of the cohorts, those with a C9orf72 HRE were not sequenced, so to include them partially would present an inaccurate estimation of the frequency. As expected, the second most common cause of genetic FTD-ALS are mutations in GRN. Pathogenic and potentially pathogenic mutations were found in 9% of the patients, excluding C9orf72 HRE which accounted for almost 50% of the mutations in the DESCRIBE-FTD cohort. Thus, the findings remain consistent.

Since C9Orf72 HRE carriers were not exome sequenced for all of the different cohorts, hereon we will be excluding this mutation from our results to avoid a misrepresentation of the % of carriers. Our combined sequencing approach yielded a total of 29 confirmed pathogenic variants in the 23 analyzed FTD/ALS and other dementia genes, identified in 35 different individuals (Table 3.2). In addition, 47 different 'potentially pathogenic' mutations were found in 54 different individuals (Table 3.2).

As expected, the highest number of pathogenic mutations observed in the FTD/ALS cohorts was in the GRN (n=11) and MAPT (n=5) genes, followed by TBK1 (n=3). Carriers of pathogenic variants in less common FTD/ALS genes like CHMP2B, CHCHD10, SQSTM1 and VCP was also noted. Consistent with what was observed in the DESCRIBE-FTD cohort, we also observed a number of pathogenic mutations in genes linked to other forms of dementia: APP, PSEN1, PSEN2, TARDBP. Lastly, we found pathogenic mutations in cholestrol homeostasis and lipofuscinosis genes: CYP27A1, CTSF.

We observed an unexpectedly high number of variants in the SQSTM1 gene, which is integral to the autophagosome. Two patients carry confirmed pathogenic SQSTM1 mutations whereas 7 others carry potentially pathogenic variants. Out of these, one indibividual carries two potentially pathogenic missense mutations in the SQSTM1 gene (subject ID: 1110290332, clinical diagnosis: PNFA, variants: rs779786150 and rs74855562).

Interestingly, a female patient affected with ALS-FUS (ID: 1108071469) as confirmed by the clinical diagnosis and our genetic screen, carries an additional potentially pathogenic variant in Cathepsin F (CTSF) which is associated with Neuronal Ceroid Lipofuscinosis. She carries the P525L pathogenic variant in the FUS gene which has been reported previously in multiple individuals affected with amyotrophic lateral sclerosis (Kwiatkowski et al. 2009; Conte et al. 2012). In addition, she carries The p.R54G variant (also known as c.160C>G), located in coding exon 1 of the CTSF gene. This variant results from a C to G substitution at nucleotide position 160. The arginine at codon 54 is replaced by glycine, an amino acid with dissimilar properties. This amino acid position is well conserved in available vertebrate species.

Similarly, another female patient affected with PPA (ID:1108071455) carries a pathogenic

TABLE 3.2: Confirmed Pathogenic Mutations across 831 clinical FTD/ALS patients

GENE	EFFECT	VCF ID	PATIENT ID	COHORT	AA CHANGE
APP (0.12%)	Missense	chr21:27264165:C:T	1108070846	DESCRIBE-FTD	p.Asp694Asn
CHMP2B (0.12%)	Missense	chr3:87294943:G:A	BRI0050	France	p.Arg69Gln
CYP27A1 (0.12%)	Missense	chr3:87294943:G:A	30477_TCCTGAGCCTCTCTAT_L006	Italy	p.Asn179Ser
CTSF (0.12%)	Missense	chr11:66333791:T:C	1110293601	DESCRIBE-FTD	p.Tyr231Cys
	Missense	chr16:31202752:C:T	1108071469	DESCRIBE-FTD	p.Pro526Leu
FUS (0.24%)	Disruptive inframe deletion	chr16:31193959:ATTC:A	1108062070	DESCRIBE-FTD	p.Ser57Del
	Missense	chr17:42426558:C:A	1108071455	DESCRIBE-FTD	p.Ala9Asp
	Frameshift	chr17:42428134:CCA:C	1094820054/143802329_BN	DESCRIBE-FTD	p.Ser226fs
	Stop gain	chr17:42428777:T:G	1108070361/1108072576	DESCRIBE-FTD	p.Tyr294*
	Splice donor	chr17:42428169:G:A	21170, 1108060722, 1110269093	DZNE & DESCRIBE-FTD	c.708+1G>A
	Splice acceptor	chr17:42428403:A:G	1108061724	DESCRIBE-FTD	c.709-2A>G
	Stop gain	chr17:42429455:C:T	1110260291, 1108082284	DESCRIBE-FTD	p.Arg418Ter
	Frameshift	chr17:42427669:T:TA	1110279576	DESCRIBE-FTD	p.Met142fs
GRN (1.3%)	Intron	chr17:42422705:A:G	1108070740	DESCRIBE-FTD	c.-8+3A>G
	Intronic_variant affecting the splicing of exon 10	chr17:44087784:C:T	1108072673	DESCRIBE-FTD	c.915+16C>T
	Missense	chr17:44087755 C>T	1110275692, 1108062078, 1110272198	DESCRIBE-FTD	p.Pro301Leu
MAPT (0.6%)	Missense	Chr17:44096076 C>T	1110293648/1108072661	DESCRIBE-FTD	p.Pro364Ser
	Missense	chr14:73664760:C:T	SM011013	Dutch	p.Pro264Leu
PSEN1 (0.24%)	Missense	chr14:73637653:C:T	1110304305	DESCRIBE-FTD	p.Ala79Val
	Missense	chr5:179260777:C:T	Sample_6	France	p.Pro387Leu
SQSTM1 (0.24%)	Frameshift	chr5:179263439:CT:C	1110262312	DESCRIBE-FTD	p.Asp391fs 394ter
	Missense	chr1:11082610:G:A	22049	DZNE	p.Ala382Thr
TARDBP (0.24%)	Missense	chr1:11082266:A:G	1108098137	DESCRIBE-FTD	p.Asn267Ser
	Disruptive inframe deletion	chr12:64891000:TGAA:T	Proband_29, EGAR00001567159_AB1850	DESCRIBE-FTD & Spain/Clarimon	p.Glu643del
TBK1 (0.36%)	Stop gain	chr12:64860701:C:T	Proband_28	DESCRIBE-FTD	p.Arg127*
	Missense	chr9:35067907:G:A	1108062094	DESCRIBE-FTD	Arg95Cys
VCP (0.24%)	Missense	chr9:35065355:C:G	1110273372	DESCRIBE-FTD	p.Gly157Arg
CHCHD10 (0.24%)	Missense	chr22:24109646:G:A	BRI0013, 1110270593	France & DESCRIBE-FTD	p.Ser59Leu

GRN mutation (rs63751243, p.Ala9Asp) as well as a potentially pathogenic variant in CYP27A1 (rs148417330, p.Arg164Gln). The CYP27A1 variant is predicted as possibly damaging and deleterious by Polyphen2 and SIFT, respectively, and has been speculated to be associated with cholesterol storage disease but the clinical significance of this variant remains uncertain.

3.4.2 Rare Variant Association Studies

For the genomewide SKAT-O analysis, we tested a total set of 20205 genes and all rare (MAF < 1%) non-synonymous variants. The test yielded a list of 35 genes that passed Bonferroni corrected significance threshold [0.05/20205 = 2.5 x 10-6] (Table 3.4). Out of these 35 genes, 4 genes (MUC16, MUC5B, MUC4 and MUC20) belong to the MUC family of genes which we excluded as potential candidates as these are frequently reported as hitters in WES datasets (Fuentes Fajardo et al. 2012).

A thorough literature search on each of the 31 candidate genes (MUC genes excluded) lead to the highlighting of a set of functions and pathways that these genes were integral to (Table 3.6). Consistent with the results from the DESCRIBE-FTD

TABLE 3.3: Potentially Pathogenic Mutations across 831 clinical FTD/ALS patients

GENE	EFFECT	VCF ID	PATIENT ID	COHORT	AA CHANGE	SNP ID
	Missense	chr21:27425601:C:G	30300_CGTACTAGCTCTCTAT_L006	Italy	p.Arg140Thr	rs772020679
	Missense	chr21:27348341:C:T	1108061605	DESCRIBE-FTD	p.Val409Leu	rs1487805466
APP (0.36%)	In-frame deletion	chr21:27394296:GCTT:G	1094813058	DESCRIBE-FTD	p.Glu241del	rs754150568
CHMP2B (0.36%)	Missense	chr3:87302911:C:T	1108103462, 1108090096, 1108103927	DESCRIBE-FTD	p.Ser194Leu	rs149380040
	Missense	chr11:66335798:G:C	1108071469	[HTML]FFFDESCRIBE-FTD	p.Arg54Gly	rs776443007
	Missense	chr11:66332390:T:C	S08D4446, SM008079, 24708, Proband_27	Dutch, DZNE, DESCRIBE-FTD	p.Asn378Ser	rs148080813
CTSF (0.72%)	Missense	chr11:66331465:A:C	18272	DZNE	p.Leu465Trp	None
CYP27A1 (0.24%)	Missense	chr2:219676989:G:A	1108070740, SOL878_18669	DESCRIBE-FTD & UK	p.Arg164Gln	rs148417330
	Missense	chr16:31201424:G:A	1108060751	DESCRIBE-FTD	p.Arg378Gln	rs1269972112
FUS (0.24%)	Missense	chr16:31195226:G:A	21368_TCCTGAGCCTCTCTAT_L002	Italy	p.Gly80Ser	rs776474571
	Missense	chr17:42427095:G:A	BRI0013	France	p.Gly109Arg	rs766292113
	Missense	chr17:42429101:C:T	21284, 22935	DZNE	p.Pro373Ser	rs912111761
	Missense	chr17:42429743:C:T	BRI0003	France	p.Pro483Leu	rs774128685
	Missense	chr17:42429576:C:T	1104378616	Germany	p.Pro458Leu	rs63750537
GRN (0.6%)	Missense	chr17:42429500:C:T	EGAR00001567164_AB1857	Spanish_Clarimon	p.Arg433Trp	rs63750412
	Missense	chr17:44060834:C:A	1108103438	DESCRIBE-FTD	p.Arg222Ser	rs150983093
MAPT (0.48%)	Missense	chr17:44068926:C:T	1108061237	DESCRIBE-FTD	Pro494Leu	rs763728305
	Missense	chr10:13154564:G:A	EX387-BAR	GENFI_Barcelona	p.Val161Met	rs776058639
OPTN (0.24%)	Missense	chr10:13164416:C:T	15944	DZNE	p.Arg271Cys	rs540943401
PRKAR1B (0.12%)	Missense	chr7:751042:C:T	1110306398	DESCRIBE-FTD	p.Cys34Tyr	rs530392908
	Splice Acceptor	chr14:73673092:A:G	17920	DZNE		None
	Missense	chr14:73659462:G:A	Sample_4	France	p.Arg220Gln	rs763831389
	Missense	chr14:73614806:C:T	17654	DZNE	p.Arg27Cys	None
	Missense	chr14:73659482:A:C	1108061596	DESCRIBE-FTD	p.Ile227Leu	rs199842082
PSEN1 (0.6%)	Missense	chr14:73614806:C:G	1110306359	DESCRIBE-FTD	p.Arg27Gly	None
	Missense	chr1:227069693:C:T	1110296425	DESCRIBE-FTD	p.Arg62Cys	rs142892469
	Missense	chr1:227076676:T:C	16724	DZNE	p.Leu271Pro	rs1211631545
	Missense	chr1:227076547:A:G	EGAR00001567126_AB1792	Spanish_Clarimon	p.Tyr228Cys	rs200410369
PSEN2 (0.36%)	Missense	chr1:227073369:C:T	1108062212	DESCRIBE-FTD	p.Arg163Cys	rs200931244
SIGMAR1 (0.12%)	Missense	chr9:34637322:A:G	1094818144	DESCRIBE-FTD	p.Phe83Leu	rs773344340
	Missense	chr5:179260077:G:A	1110278743	DESCRIBE-FTD	p.Arg267His	rs149424705
	Missense	chr5:179252155:C:T	12621	DZNE	p.Pro228Leu	rs151191977
	Missense	chr5:179260687:A:T	21200	DZNE	p.Gln357Leu	rs1415449512
	Disruptive Inframe Deletion	chr5:179252182:TGAA:T	BRI0029	France	p.Lys238del	rs767056938,rs796052214
	Missense	chr5:179263444:C:G	22135	DZNE	p.Pro392Ala	None
	Missense	chr5:179248075:C:G	1110290332	DESCRIBE-FTD	p.Leu47Val	rs779786150
SQSTM1 (0.84%)	Missense	chr5:179248090:G:T	1110290332	DESCRIBE-FTD	p.Ala52Ser	rs748555662
	Missense	chr12:64849687:G:T	SM009510	Dutch	p.Asp13Tyr	None
	Disruptive Inframe Deletion	chr12:64858113:GACA:G	EGAR00001567146_AB1836	Spanish_Clarimon	p.Thr79del	rs748007618
	Splice Acceptor	chr12:64889471:TTAAA:T	DNA28066A	DESCRIBE-FTD	:c.1644-5_1644-2del	rs755646937
	Missense	chr12:64858199:A:G	1108062060	DESCRIBE-FTD	p.Tyr105Cys	rs1366668789
TBK1 (0.72%)	Missense	chr12:64849714:A:G	DNA27716A	DESCRIBE-FTD	p.Asn22Asp	rs576726084
	Missense	chr2:70443421:T:A	FPD016-006_2	France	p.Gln228Leu	rs763253859
TIA1 (0.24%)	Missense	chr2:70439871:C:T	Proband_29	DESCRIBE-FTD	p.Ala281Thr	rs768554955
TREM2 (0.24%)	Missense	chr6:41129133:C:T	1108103532, 1108103500	DESCRIBE-FTD	p.Asp87Asn	rs142232675
UBQLN2 (0.12%)	Missense	chrX:56590707:C:T	29291_TCCTGAGCCTCTCTAT_L004	Italy	p.Thr134Ile	rs764837088

cohort SKAT-O analysis for autophagy associated genes, we picked up three human autophagy-associated genes that passed genomewide signifcance in our SKAT-O analysis across the Western European population: SERPINA1, ATG2A, ZNF418. In addition, TRIM64B and TRIM43 were also genomewide significant. While there is not much known about the TRIM64B and TRIM43 genes, the TRIM family of genes has roles in immunity and autophagy (Hatakeyama 2017).

Additionally, SKAT-O analyses for 214 human autophagy genes performed using 745 western European cases (DANCER-FTD removed and closely related relatives removed) and 1732 controls produced 3 genes that were above the Bonferroni corrected significance threshold of 0.0002: SERPINA1, ATG2A and ARSB. This is consistent with what we see in the DESCRIBE-FTD cohort.

3.4.2.1 Evaluation of variants in candidate autophagy genes

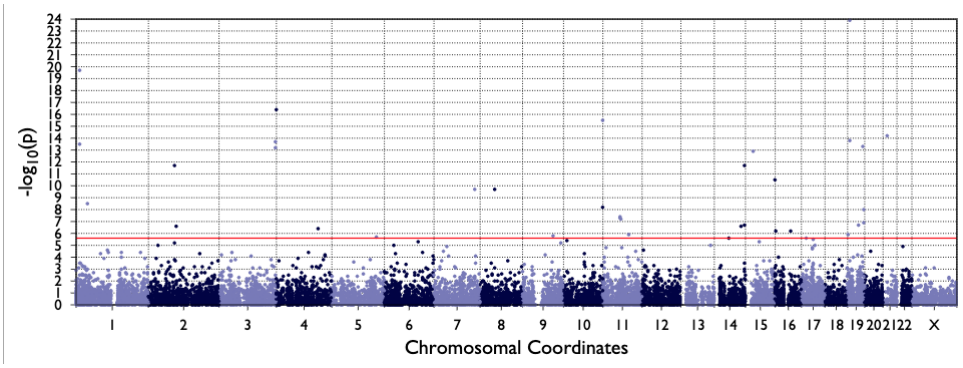
In SERPINA1, we found 6 potentially pathogenic variants in patients with FTD/ALS (Table 3.7) and one confirmed pathogenic variant (NM_001127701.1(SERPINA1):c.194T>C (p.Leu65Pro); rs28931569, patient ID: DNA28576A, cohort: DESCRIBE-FTD) in a female patient with PPA.
In ATG2A, we found 8 potentially pathogenic missense variants and 1 damaging splice donor variant (Table 3.8). Lastly, in ARSB we found 5 potentially pathogenic missense variants (Table 3.8).

3.4.3 Replication Cohort: Rare Variant Association Analysis

We found that none of the candidate genes that were found in our SKAT-O analyses were significantly associated with FTD/ALS in the replication cohort of 2451 subjects with clinical FTD/ALS and 4029 controls. However, some genes did show suggestive p-values of lesser than 0.05 in the association analyses.

In the analysis consisting of rare, loss-of-function (stop, frameshift and splice) variants only, the total number of genes tested were 14228. Here, rare LOF variants in HIF1A were associated with the cases status with a p-value of 0.001.

FIGURE 3.2: Manhattan Plot for Genomewide SKAT-O Analysis



Chromosomal coordinates on the x-axis and test p-values on the y-axis. This plot illustrates gene-based collapsing of non-synonymous rare variants (MAF<1%) using SKAT-O. The red line illustrates a threshold for genome wide significance.

FIGURE 3.3: Q-Q plot of test statistics from Genomewide SKAT-O for gene-based collapsing of rare (MAF<1%) non-synonymous variants in 745 FTD/ALS patients versus 1732 controls.

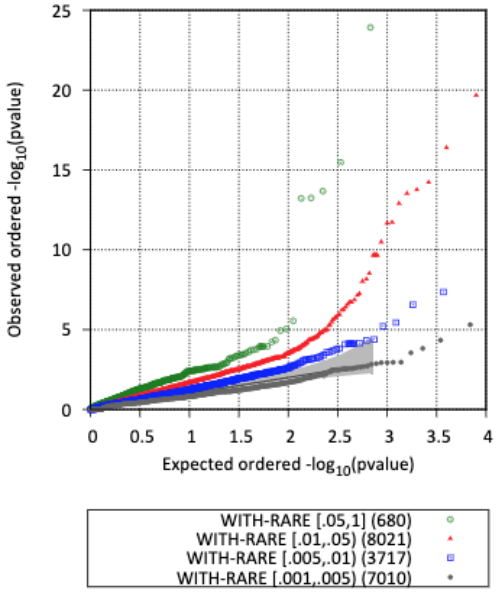


TABLE 3.4: Results from Genomewide SKAT-O for gene-based collapsing of rare (MAF<1%) non-synonymous variants in 745 FTD/ALS patients versus 1732 controls.

CHROM	BEGIN	END	GENE	NS	FRACTION WITH RARE	NUM. ALL VARS	NUM. PASS VARS	NUM. SING VARS	PVALUE
19	8961981	9091811	MUC16	2477	0.31449	419	287	149	1.17E-24
1	12835153	12837669	PRAMEF12	2477	0.040775	18	11	5	2.09E-20
4	1388337	1389276	CRIPAK	2477	0.018975	17	12	7	4.07E-17
11	1244408	1282813	MUC5B	2477	0.21679	235	165	67	3.38E-16
21	15588514	15599586	RBM11	2477	0.030682	9	6	2	5.91E-15
19	9005235	9011448	AC008734.1	2477	0.036334	30	25	13	1.74E-14
3	195474159	195538675	MUC4	2477	0.21235	307	185	74	2.11E-14
1	12853390	12856111	PRAMEF1	2477	0.029471	28	13	3	2.92E-14
19	55174498	55179377	LILRB4	2477	0.051272	37	28	6	5.57E-14
3	195343638	195460073	MUC20	2477	0.056116	37	16	4	5.96E-14
15	28630450	28632820	GOLGA8F	2477	0.027453	9	7	0	1.24E-13
14	106204131	106209368	IGHG1	2477	0.014534	8	7	2	1.88E-12
2	90248938	90249273	IGKV1D-43	2477	0.026241	6	4	0	2.14E-12
16	1306295	1308333	TPSD1	2477	0.020993	22	16	9	3.29E-11
8	52733050	52811580	PCMTD1	2477	0.029067	12	8	2	2.15E-10
7	142457341	142460870	PRSS1	2477	0.01413	10	7	4	2.15E-10
1	40945001	40961562	ZNF642	2477	0.023819	14	13	7	2.97E-09
10	135094807	135116328	TUBGCP2	2477	0.031893	25	24	13	6.80E-09
19	58352751	58371456	ZNF587	2477	0.033912	15	9	4	9.71E-09
11	61015892	61018692	PGA4	2477	0.0080743	8	8	5	4.46E-08
11	61015862	61018692	PGA5	2477	0.0109	10	10	6	5.62E-08
11	64662597	64684483	ATG2A	2477	0.038757	35	32	13	7.07E-08
19	58437573	58438315	ZNF418	2477	0.015745	12	12	8	1.35E-07
14	106109658	106111126	IGHG2	2477	0.020186	21	19	12	1.97E-07
19	37850550	37854580	HKR1	2477	0.029067	15	12	5	1.97E-07
14	94844843	94849388	SERPINA1	2477	0.034316	17	10	2	2.61E-07
2	96259774	96261953	TRIM43	2477	0.0068631	6	6	5	2.82E-07
4	144435225	144474296	SMARCA5	2477	0.026645	14	13	8	3.83E-07
16	5139186	5147676	FAM86A	2477	0.012919	10	7	5	5.67E-07
16	55844441	55866952	CES1	2477	0.034719	26	21	7	5.91E-07
19	1987562	1997470	BTBD2	2477	0.012919	7	5	3	1.14E-06
11	89603920	89609112	TRIM64B	2477	0.019378	9	8	2	1.28E-06
9	107546653	107646756	ABCA1	2477	0.044812	37	31	15	1.61E-06
5	154320711	154346305	MRPL22	2477	0.021801	10	9	4	2.01E-06
17	19641709	19646691	ALDH3A1	2477	0.03149	12	10	4	2.37E-06

NS: Number of phenotype samples with non-missing genotypes; FRACTION WITH RARE: Fraction of individuals carrying care variants below MAF 1%; NUM. ALL VARS: Total number of non-synonymous variants in the group; NUM. PASS VARS: Number of variants passing the MAF threshold; NUM. SING VARS: Number of singletons among variants that passed the MAF threshold; PVALUE: Raw p-value of SKAT-O test

TABLE 3.5: Functional Annotation for Genomewide Significant Candidate Genes obtained from the SKAT-O analysis

PATHWAY	GENES
IMMUNITY	LILRB4, IGHG1, IGKV1D-43, TRIM64B, TRIM43, IGHG2
AUTOPHAGY	SERPINA1, ATG2A, ZNF418, TRIM64B, TRIM43
LIPID METABOLISM	ATG2A, NPC1L1, ABCA1
ENERGY METABOLISM/MITOCHONDRIA	MRPL22, SMARCA5
MENTAL ILLNESS	ZNF642, ZNF587
ION TRANSPORT	SLC12A1
DIGESTION	PGA4, PGA5

TABLE 3.6: Human Autophagy Genes associated with FTD/ALS: Results from SKAT-O for gene-based collapsing of rare (MAF<1%) non-synonymous variants 745 FTD/ALS patients versus 1732 controls.

CHR	BEGIN	END	GENE	NS	FRACTION WITH RARE	NUM. ALL VARS	NUM. PASS VARS	NUM. SING VARS	PVALUE
14	94844843	94849388	SERPINA1	2477	0.034316	17	10	2	2,61E-03
11	64662597	64684483	ATG2A	2477	0.038757	35	32	13	7.07e-08
5	78076288	78280974	ARSB	2477	0.0060557	13	9	6	0.00025895

NS: Number of phenotype samples with non-missing genotypes; FRACTION WITH RARE: Fraction of individuals carrying care variants below MAF 1%; NUM. ALL VARS: Total number of non-synonymous variants in the group; NUM. PASS VARS: Number of variants passing the MAF threshold; NUM. SING VARS: Number of singletons among variants that passed the MAF threshold; PVALUE: Raw p-value of SKAT-O test

TABLE 3.7: Rare potentially pathogenic SERPINA1 variants in the 831 clinical FTD/ALS patients

EFFECT	VCF ID	PATIENT ID	COHORT	AA CHANGE	SNP ID
Missense	chr14:94845805:G:A	SM009517	Dutch	S354F	rs201788603
Missense	chr14:94847290:G:T	18925	DZNE	P279T	rs759736224
Missense	chr14:94849022:C:T	BRI0014	French	V185M	rs147247134
Missense	chr14:94849061:C:A	S03D3303	Dutch	G172W	rs112030253
Missense	chr14:94849325:C:T	24568	DZNE	A84T	rs111850950
Missense	chr14:94849364:T:G	EGAR00001567163_AB1856	Spanish_Clarimon	S71R	rs11575873

TABLE 3.8: Rare potentially pathogenic ATG2A and ARSB variants in the 831 clinical FTD/ALS patients

GENE	EFFECT	VCF ID	PATIENT ID	COHORT	AA CHANGE	SNP ID
ATG2A	Missense	chr11:64665815:T:A	DNA-049-BAR	GENFI_Barcelona	p.His1566Leu	rs144122454
	Missense	chr11:64668386:C:T	105395_C	GENFI_Coimbra	p.Gly1435Asp	None
	Missense	chr11:64669781:G:T	SD95_3294	Dutch	p.Asp1287Glu	None
	Missense	chr11:64673302:C:T	1094820054/143802329_BN	DESCRIBE-FTD	p.Val1095Met	rs201916479
	Missense	chr11:64673837:C:T	1110308381, DNA28539A	DESCRIBE-FTD	p.Arg1051His	None
	Missense	chr11:64676485:C:G	102385_C	GENFI_Coimbra	p.Arg781Thr	rs762585246
	Missense	chr11:64678146:G:A	1108061994	DESCRIBE-FTD	p.Ala550Val	None
	Missense	chr11:64681615:T:C	2167-C	GENFI_Coimbra	p.Glu142Gly	None
	Splice Donor	chr11:64681809:C:A	14520	DZNE		None
ARSB	Missense	chr5:78076288:C:T	1110272175	DESCRIBE-FTD	p.Val512Met	rs201928777
	Missense	chr5:78181468:C:A	S09D11646	Dutch	p.Ala361Ser	rs752599167
	Missense	chr5:78181489:G:C	BRI0055	French	p.Leu354Val	None
	Missense	chr5:78264927:T:G	102330-C	GENFI_Coimbra	p.Lys134Thr	None
	Missense	chr5:78260418:C:T	12977_CAGAGAGGTAGATCGC_L007	Italy	p.Gly171Ser	None

TABLE 3.9: GWAS analysis results for 269 PNFA cases versus 538 controls (Ferrari et al., 2014)

GENE	SNP ID	CHR	BP	REF	ALT	P-VALUE	CASES	CONTROLS
SERPINA1	rs11628917	14	94843719	C	T	2,58E-06	269	538
SERPINA1	rs17751614	14	94841542	C	T	2,79E-06	269	538
SERPINA1	rs1243160	14	94854877	G	A	5,04E-06	269	538

In the analysis consisting of rare missence (i.e., only exonic variants), the number of genes tested were 17595. Here, rare missense variants SERPINA1 and TPSD1 had p-values = 0.04 each, in the SKAT-O analysis.

Finally, in the analysis that included both missense and LOF rare variants, SERPINA1 and TPSD1 had p-values 0.04 and 0.03, respectively.

3.4.4 Validation using the Genomewide Association Study for FTD

As described in the Methods section 3.3.6, the subjects included in this study include 3562 FTD cases and 9402 neurologically healthy controls. The study discovered loci linked to immune function, lysosomal biology and autophagy associated significantly or of suggestive significance with FTD.

In the comparison of 269 PNFA cases versus 538 controls, SERPINA1 showed suggestive p-values of 10-6 (Table 3.9). P-values did not reach genomewide significance probably due to small sample size.

3.5 DISCUSSION

In this study, we examine individual cases of FTD/ALS in a wider western european population of 831 subjects. In addition to investigating pathogenic and potentially pathogenic variants in known NDD genes, we performed a genomewide rare variant association study and another association study to check for the burden of rare damaging variants in human autophagy associated genes.

Excluding the C9orf72 HRE pathogenic variants, which are the most common cause of genetic FTD, we find that 35 patients carry confirmed pathogenic mutations, explaining greater than 4% of the cases. The reason to exclude C9orf72 HRE variants is that while in the DESCRIBE-FTD and DANCER-FTD cohorts, we systematically perform RP-PCR to detect the C9Orf72 HRE lengths, in the other Wester European cohorts, in several cases, the C9Orf72 carriers were exluded prior to exome sequencing. This would lead to a confounding of true ratios when accounting for the C9Orf72 HRE variants in some cohorts and not the others. Out of the 35 confirmed pathogenic mutations, the highest numbers are in GRN (1.3%), MAPT (0.6%) and TBK1 (0.4%), as is expected.

Interestingly, we found pathogenic mutations in less common FTD/ALS genes, and those that are commonly associated with other form of dementias/Amyloidosis: APP, PSEN1, PSEN2, TARDBP. We also found pathogenic mutations in CHCHD10 (mitochondrial), CTSF (lipofuscinosis) and CYP27A1 (cholestrol homeostasis) genes. This provides more evidence towards the theory that the pathology of the frontotemporal cortex of patients on the FTD/ALS spectrum of diseases is a result of multiple disease mechanisms (Ferrari et al. 2016). It is also possible that the patients were incorrectly diagnosed as clinical syndromes of bvFTD and AD are often clinically indisguishable. Similarly, the patient carrying a known pathogenic variant of CTSF may indeed be affected by Kufs disease.

Among the potentially pathogenic variants found, we found an unexpectedly high number of variants (n=7) in the SQSTM1 gene. The SQSTM1 gene encodes the Sequeestome-1 protein which is integral to the human autophagosome. This lends further proof into the role of autophagy as a central pathway in FTD/ALS disease prognosis and pathology.

We also found several examples of patients that carried more than one damaging mutation. In some cases, these mutations were in genes that share a pathway. For example, patient 'Proband-29' from the DESCRIBE-FTD cohort carries a pathogenic TBK1 mutation and a potentially pathogenic TIA1 mutation. Both these genes have roles in stress granule associated pathways. Whereas, patient '1094820064' carries a pathogenic GRN mutation and a potentially pathogenic ATG2A mutation. There

can be several ways in which dual mutation carriers are impacted with relation to progression of disease: 1) it could lead to an earlier age at onset 2) it could lead to more aggressive and diverse symtoms and 3) the effects of one mutation may drive the disease pathology, masking the effects of the other. These hypothesis, however, are yet to be functionally confirmed.

Genes belonging to several pathways that are commonly linked with FTD/ALS were found to be significantly associated with the 745 FTD/ALS cases that we tested. These include immune function, autophagy, lipid metabolism, mitochondria/energy metabolism. In addition, we found a gene involved in ion transport (SLC12A1) and two involved in digestion (PGA4, PGA5) to be significantly associated with FTD/ALS. Several studies in the last decade have pointed towards a link between the gastrointestinal tract and neurodegenerative diseases (Landqvist Waldö et al. 2014). Most complaints regarding digestive issues and somatic discomfort in FTD patients have gone unexplained. PGA4 and PGA5 encode protein precursors of the digestive enzyme Pepsin and are abnormamilities in these proteins have been associated with Gastritis.

Two genes that were consistently significantly associated with the FTD/ALS case cohorts are ATG2A and SERPINA1 both at the genomewide and the pathway-specific (autophagy) levels. To validate this, we found that three SNPs in the SERPINA1 gene were hits in the FTD GWAS for PNFA cases. In addition, we found one PPA patient in the DESCRIBE-FTD cohort who carries a confirmed pathogenic SERPINA1 mutation. In addition, we found 6 damaging SERPINA1 variants in the patient cohort, that may be pathogenic. Ours is not the first study to implicate SERPINA1 in ALS pathology, a 2017 study implicates SERPINA1 in both sporadic ALS and C9Orf72 associated-ALS (Ebbert et al. 2017). We propose that SERPINA1 is a strong candidate to be included in future FTD/ALS genetic screens.

We propose that the variants found in ATG2A and SERPINA1 be studied further to understand the functional mechanism in which they might be contributing to disease pathology. In addition, validating the variants found in ATG2A and SERPINA1 in our patient population using Sanger sequencing would be an important next step.

One of the major limitations in our efforts to uncover the missing heritability of FTD/ALS and, more widely, neurodegenerative disease is the absence of knowledge of the roles of non-coding genetic elements in disease and development. So far, in our whole exome sequencing based approaches, we were only able to study coding variants which make up only 1% of the human genome. The GWA study that we used to corroborate our findings includes a vast majority of SNPs in non-coding regions of the genome, which we were unable to decipher with our existing NGS data. Additionally, when hypothesizing complex disease mechanisms including a convergence of pathways, it is important to acknowledge the role of non-coding elements such as lncRNAs, enhancers, etc., to better understand disease pathology and progression.

Chapter 4

Investigating lncRNA function in neurons using ASO-based knockdowns

4.1 ABSTRACT

Non-coding genetic elements are often overlooked in studying the mechanisms that underlie complex polygenic neurodegenerative disorders and traits. Long non-coding RNAs (lncRNAs) constitute the majority of transcripts in the mammalian genomes, yet the functions of a majority of them remain unclear. Here, we aim to design a study to functionally annotate lncRNAs that are highly expressed in the human brain and nervous system. In this pilot study, we conduct a systematic knock down of the expression of 16 lncRNAs using Antisense Oligonucleotides (ASOs) in human-induced pluripotent stem cell-derived neurons (hIPSC-derived neurons), followed by CAGE-Sequencing to study changes in the transcriptome. The results of our perturbation screen exhibit the need for the development of a more robust, scalable and cost-effective methodology to functionally annotate lncRNAs and their role in human neurodegeneration.

4.2 INTRODUCTION

For several decades, most of the non-coding RNA (ncRNAs) species were dismissed as products of spurious transcription and treated as "junk DNA". Research in the field of neurodegeneration has largely been focused on the 20,000 protein coding genes, which make up about 2% of the genome, leaving a large gap in the literature

with relation to ncRNAs. Recently, however, it has been shown that the expression of ncRNAs in the brain is dynamically regulated in an activity-dependent and spatiotemporally controlled manner, suggesting precise regulatory roles of ncRNAs in brain development and function. (Salta and De Strooper 2017).

Long noncoding RNAs (lncRNAs) are defined as those RNAs with at least one transcript of length > 200 nucleotides that is not translated into a protein. An estimated 40% of the genes for lncRNAs are specifically expressed in the brain tissue. Most lncRNAs have low abundance and lack typical signatures of selective constraints. Additionally, a substantial fraction of lncRNAs are unstable (Andersson, Refsing Andersen, et al. 2014) and originate from regulatory regions of other functional units such as promoter upstream transcripts (PROMPTs) (Andersson, Refsing Andersen, et al. 2014; Preker et al. 2008; Andersson, Gebhard, et al. 2014)) and enhancer RNAs (Andersson, Refsing Andersen, et al. 2014; Preker et al. 2008; Andersson, Gebhard, et al. 2014). For these reasons, despite a few well characterized examples of lncRNAs, the functional relevance of a majority of them remains unclear.

It is also important to note that in the case of lncRNAs, it is often the act of transcription and the location of the transcript that is of functional relevance rather than the sequence of the transcript itself. Due to this reason, it is imperative to have accurate 5′ sequences for lncRNAs. The FANTOM5 (Hon et al. 2017) study helped fill a tremendous gap via CAGE-Sequencing as most transcriptomes up till then were built via RNA-Seq and had inaccurate 5′ sequences. The FANTOM5 project generated a comprehensive atlas of 27,919 human lncRNA genes with high-confidence 5′ ends and expression profiles across 1,829 samples from the major human primary cell types. In addition, this study characterized lncRNAs that overlap expression quantitative trait loci (eQTL)-associated single nucleotide polymorphisms (SNPs) of mRNAs and are co-expressed with the corresponding mRNAs. This is suggestive of potential roles in transcriptional regulation. The data generated under the FANTOM5 study offers a host of opportunities in studying the functional relevance of lncRNAs in neurodegeneration and in healthy ageing.

The genome assembly under the FANTOM project is collectively referred to as the

FANTOM CAT genome and the lncRNA gene classes under FANTOM CAT are defined as follows:

Divergent lncRNAs: genes with its strongest CAGE cluster within $\pm$ 2kb on the opposite strand of any CAGE clusters of GENCODEv19 protein coding genes or pseudogenes.
Sense intronic lncRNA: lncRNA genes 1) initiating within the intron of another FANTOM CAT gene, 2) with at least 50% of their genic region overlapping with the genic region of any other genes, 3) with its strongest CAGE cluster not overlapping exons of other genes, and 4) containing 10 CAGE reads, or otherwise defined as 'other sense overlap RNA'.
Antisense lncRNAs: genes with $\geq$ 50% of their genic region overlapping with the genic region of GENCODEv19 protein coding genes or pseudogenes on the opposite strand.
Intergenic lncRNAs: the remaining lncRNA genes that could not be assigned to any of the above categories.

In this chapter, we use the data generated under the FANTOM5 study to design a pilot study to investigate the functions of 20 highly expressed lncRNAs in the brain and central nervous system.

4.3 METHODS

4.3.1 Experiment Design

Two control human induced pluripotent stem cell (hiPSC) lines GM23280 and ND41865 were transduced with a lentiviral vector to induce Neurogenin-2 (NGN2) expression to produce hiPSC derived cortical neurons using a published protocol (Dhingra et al. 2020). These cell lines were transfected with a maximum of 5 ASOs for each lncRNA target and RNA was extracted and collected at days 0, 3 and 8 after transfection. ASOs with a minimum qPCR knockdown efficiency of 50% were selected. The top 3 (highest qPCR KD efficiency and minimum total RNA yield of 200 nanograms) ASOs for each target were extracted and the RNA products for days 3 and 8 were sent for CAGE-Sequencing after library preparation. In addition, a positive control (MALAT-1), an untreated sample and a scramble control was included for each time

point.

4.3.2 Next Generation Sequencing

CAGE-Seq is a powerful method to identify transcription start sites (TSSs) of capped RNAs while simultaneously measuring transcript levels. For the pilot experiments using ASO-based perturbation, we used a new experimental protocol for sequencing called Low Quantity (LQ) single strand (ss) CAGE "LQ-ssCAGE" developed and performed at the RIKEN Center for Integrative Medical Sciences (IMS), Yokohama, Japan. Two libraries were prepared with the same data and sequenced with 100-bp paired-end and 50-bp paired-end, respectively.

Pre-processing. The fastq files were aligned to the FANTOM CAGE-Associated Transcriptome (CAT) reference genome (Hon et al. 2017) which is based on the human genome build 38 (hg38) using STAR (v 2.5.3). First, expression for CAGE promoters was estimated by counting the numbers of mapped CAGE tags falling under the 379,952 promoter regions of FANTOM 6 CAT gene models (Ramilowski et al. 2020). Next, the expression of the corresponding 124,047 genes was estimated by summing up the expression values of all promoters assigned to a given gene.

Analysis. The clustering of the data was visualised using MDS plotting in R using the 'plotMDS' function. The gene-wise expression counts were used to conduct differential gene expression (DGE) analysis at the two time points - day 3 and day 8 - in R using 'DeSeq2' (Love, Huber, and Anders 2014). Pairwise comparisons for each lncRNA knockdown versus the untreated control on each time point were conducted using the different ASOs as technical replicates. A log fold change cut-off of 0.5 was implemented. The differentially expressed genes (DEG) lists were used to conduct gene ontology and pathway enrichment analyses using G:Profiler (Raudvere et al. 2019). In addition, gene expression patterns were studied between time points, i.e., those that were consistently up/downregulated upon KD of their corresponding lncRNA or those that changed direction of expression between days 3 and 8. Additionally, likelihood ratio tests (LRT) with a reduced model was used to conduct a time-series analysis on 'DeSeq2' using the 'day' as a factor to fit all the DEGs in one model and test for any differences over the different time points. Each

cell line was analysed separately.

4.3.3 Target Selection

For the pilot phase of the study, 20 lncRNA targets (Supplementary Table A.2) were selected using in-house CAGE-Sequencing data from (i) human frontal and temporal brain regions of neurologically healthy controls obtained under a Material Transfer Agreement from the Netherlands Brain Bank of Neurological Disorders and MRC, Kings College London, (ii) hiPSC controls on days 0 and 10 following NGN2 transduction, (iii) hiPSCs with a pathogenic C9orf72 point mutation (P310Q and R140Q) on days 0 and 10 following NGN2 transduction and (iv) hiPSCs with pathogenic C9orf72 hexanucleotide repeat expansion (HRE) on days 0 and 10 following NGN2 transduction. The selection was based on high expression levels of lncRNAs in the above 8 categories as well as a thorough literature search that aided in the hypothesis that these 20 lncRNAs play a role in human neurodegeneration.

4.3.3.1 Feature Map Construction

For each of the 20 lncRNAs, we constructed a genomic map of a 1 Mb region around the lncRNA (500 Kb on either side) to study the flanking genes of these lncRNAs, using the UCSC genome browser (Kent 2002) tracks, we added custom tracks from the NONCODE ((Kent 2002; Liu et al. 2005), LNCipedia (Kent 2002; Liu et al. 2005; Volders et al. 2013) and FANTOM5 databases to study the enhancer regions, promoter regions and other relevant information of cis-genes to these target lncRNAs being knocked down. In addition to screenshots of these feature maps, we generated easily editable browser views on the UCSC genome browser for any further modifications that may be required.

4.3.4 ASO Design

ASOs were designed as RNase H-recruiting locked nucleic acid (LNA) phosphorothioate gapmers with a central DNA gap flanked by 2-4 LNA nucleotides at the 5′ and 3′ ends of the ASOs. For each lncRNA target, the unspliced transcript sequence from FANTOM CAT was used as a template for designing a minimum of 5 ASOs

per lncRNA. For more details on ASO design, refer to the supplemental methods in the FANTOM6 PILOT paper (Ramilowski et al. 2020). Finally, we shortlisted the ASOs designed based on minimal off-target effects and spanning the transcript on multiple unique locations.

4.3.5 Cis-regulation In-Silico Analysis

4.3.5.1 Co-expression analysis

Three human neuronal precursor cell (NPC) lines GM23280, ND41865 and ST12761-CL49 were transduced with a lentiviral vector to induce Neurogenin-2 (NGN2) expression to produce hiPSC derived cortical neurons using a published protocol (Dhingra et al. 2020). RNA libraries were prepared on days 0, 8 and 16 following transfection using the Illumina RNA Stranded Gold kit. RNA-Sequencing was performed using the Illumina NextSeq 550 to produce 150 bp paired-end reads.

As primary sequencing output, BCL basecall files were produced. These BCL files were converted to fastQ files using 'bcl2fastq' software by Illumina. In addition to BCL to fastQ conversion, this tool also demultiplexes the samples in the same step. Adapter sequences were trimmed using 'cutadapt' (Martin 2011). FastQC (Andrews S., 2010) and MultiQC (Ewels et al. 2016) softwares were used to visualise the quality of the sequencing data prior to conducting the co-expression analyses. These sequencing data were quantified using 'salmon' (Patro et al. 2017), using the FANTOM CAT transcript reference FASTA which is based on the hg38 genome build. Transcript abundance files were imported from the 'salmon' output using 'tximport' tool (Soneson, Love, and Robinson 2015) in R. A 'DGEList' object was created using the 'edgeR' software (Robinson, McCarthy, and Smyth 2010) in R and the data was normalised using 'calcNormFactors' in 'edgeR'. A matrix of CPMs (counts per million) was generated using 'edgeR'. The data was annotated using FANTOM CAT annotation.

It was discussed previously that the function of a lncRNA is often in the location of it's transcript and the act of transcription rather than it's transcript itself. A common mechanism for lncRNAs to regulate gene function is via cis-regulation of nearby

genes through their promoter and enhancer regions as well as by recruiting transcriptional machinery. To test this, we studied the co-expression of the 20 target lncRNAs, with genes within 500kb up and downstream of their transcript. Pearson correlation coefficients (r) were calculated for each lncRNA gene with it's cis genes using the 'cor' function in R. In addition to individual pearson correlation coefficients with each of it's cis genes, the average coefficient to check for an overall effect was also calculated using the 'colMeans' and 'abs' functions in R.

4.3.5.2 Hi-C visualization

To visualise chromosomal interactions between the 20 selected lncRNAs and the genes in-cis, we used the 3D Genome Browser (Wang et al. 2018) that has extensive data on Hi-C interactions for different cell and tissue types. Under the Hi-C heatmaps, the UCSC genome browser (Navarro Gonzalez et al. 2021) for that genomic region can be imbibed to visualise chromatin interactions and other 'omics' data simultaneously.

4.4 RESULTS

4.4.1 Targets Selected

For the pilot phase of the long non-coding RNAs investigative study, we selected 20 lncRNAs for perturbation (Table 4.1). The criteria for selection is highlighted in the Methods section 4.3.3.

4.4.2 ASO based knockdown experiments

The experimental design for the ASO-based KD screens is described in the Methods section 4.3.1.

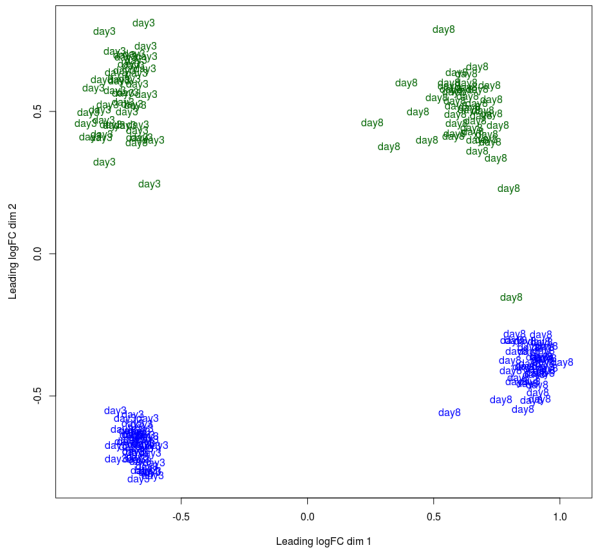
During the experimental phase, four of the selected lncRNAs - LINC00599, DHRS4-AS1, AC013394.2 and RP11-1094M14.1 - showed inadequate knockdown (KD) efficiency or extremely low yield for all the tested ASOs and hence were removed from further analysis. Thus, a total of 16 lncRNAs were knocked down during the pilot phase.

TABLE 4.1: Target lncRNAs selected for ASO based perturbations as Phase 1 of the study

Gene ID	Gene Name	Gene Type	Chr	Start	End	Strand	ASOs
ENSG00000253230	LINC00599	intergenic	8	9886104	9905802	-	5
ENSG00000196810	CTBP1-AS2	antisense	4	1249468	1251187	+	5
ENSG00000249673	NOP14-AS1	antisense	4	2934915	2937841	+	5
ENSG00000244879	GABPB1-AS1	antisense	15	50355484	50356358	+	5
ENSG00000267321	RP11-1094M14.11	intergenic	17	35568099	35570884	+	5
ENSG00000227252	AC105760.2	antisense	2	236959770	237085774	-	5
ENSG00000215256	DHRS4-AS1	divergent	14	23934047	23954171	-	5
ENSG00000227354	RBM26-AS1	antisense	13	79406290	79407590	+	5
ENSG00000225377	RP5-1103G7.4	antisense	20	311124	325268	-	5
ENSG00000176840	MIR7-3HG	intergenic	19	4769132	4770184	+	5
ENSG00000245937	CTC-228N24.3	divergent	5	127940425	128083072	-	5
ENSG00000215447	BX322557.10	divergent	21	45288081	45290578	+	5
ENSG00000254635	WAC-AS1	antisense	10	28512561	28532626	-	5
ENSG00000231365	RP11-418J17.1	antisense	1	119140416	119142200	+	5
ENSG00000270066	SCARNA2	intergenic	1	109100198	109100612	+	5
ENSG00000247240	UBL7-AS1	antisense	15	74461264	74481302	+	5
ENSG00000272888	AC013394.2	sense intronic	15	92882722	92883861	+	4
ENSG00000214401	KANSL1-AS1	antisense	17	46193575	46196721	+	5
ENSG00000260448	RP11-449H11.1	intergenic	16	25067126	25107097	-	5
ENSG00000249456	RP11-298J20.4	sense overlapping	10	124916919	124917057	+	2

These annotations are based on human genome build 38. The lncRNA 'Gene Type' is characterized by annotations in the FANTOM 'CAT gene category'.

FIGURE 4.1: MDS plot showing 4 clear clusters from CAGE-Sequencing expression data from ASO based lncRNA perturbations based on day of differentiation and cell line. In green, is the cell line ND41865 and in blue is the cell line GM23280.



4.4.3 Transcriptomics Analysis

Using the 182 CAGE-Sequencing samples from the ASO based perturbations, as well as controls, we performed preliminary quality control measures to check for data quality. The data clustered in 4 clear clusters based on the cell line (ND31865 and GM23280) as well as day of differentiation (day 3 and day 8) [Figure 4.1], as is expected. However, we noted that the depth of sequencing of this data is extremely low with a large number of 0 expressed genes i.e., no reads mapped to the given gene. In several cases, the lncRNAs to be perturbed were 0 expressed in the untreated as well as negative control (CA) samples which would limit the meaningfulness of the DGE analysis.

4.4.4 Differential Gene Expression (DGE) Analysis

In addition to the above mentioned limitations, due to a lack of biological replicates, statistically significant results could not be achieved in a DGE analysis. The DEGs consistently showed an enrichment of gene ontology terms 'axon guidance' and 'neuron development'. Results from the DGE analysis are described in detail in the Supplementary Text C.1.

4.4.5 Co-Expression Analysis for Cis-Genes

It is well known that lncRNAs often regulate the expression levels of protein coding genes in their proximity, especially antisense lncRNAs. Using RNA-Sequencing data from the same cell lines - ND41685 and GM23280 - we performed in-silico analyses to check for correlation of expression levels of the target lncRNAs with their cis-genes (+/- 500 Kb) as described in the Methods section 4.3.5 of this chapter.

Supplementary table A.6 highlights the pearson correlation coefficients expression between our preliminary target set of 16 lncRNAs and their "cis-genes". We found that the expression of the AC105760.2 lncRNA gene is highly positively correlated with the AC105760.3 (r=0.97), ACKR3 (r=0.73) and COPS8 (r=0.69) genes. The AC105760.2 gene is antisense to the ACKR3 and COPS8 genes, indicative of antisense regulation of transcription. ACKR3 is an important regulator of axon guidance in the oculomotor system (Whitman et al. 2019) and the COPS8 gene which is highly expressed in the brain encodes a subunit of the COPS9 signalosome which is a highly conserved protein complex with functions as an important regulator in multiple signalling pathways (Seeger et al. 1998).

The expression of the GAPBP1-AS1 gene which is head-to-head antisense to its coding counterpart GAPBP1 gene is highly negatively correlated to it (r = -0.69). In contrast to AC105760.2, this could be an example of lncRNA mediated transcriptional repression by disrupting the binding of the transcriptional machinery at the GABPB1 promoter region.

In contrast, the expression of the KANSL1-AS1 gene is highly positively correlated with the KANSL1 gene (r=0.93). In addition, Hi-C maps from the 3-D genome browser [explained in the Methods section 4.3.1] also show interaction of the KANSL1-AS1 gene with the KANSL1 gene (Supplementary Fig B.3).

For the NOP14-AS1 gene, Hi-C interactions showed interaction with the FAM193A (r = 86) and HTT (r = 0.6) genes. The RP11-298J20.4 gene expression is highly positively correlated with the gene encoding the Zinc finger RANBP2-type containing 1 protein (r = 0.9) which has important roles in deubiquitinization (Zhang et al. 2018).

The RP11-298J20.4 gene completely overlaps the ZRANB1 gene. The RP11-449H11.1 lncRNA, also called the LCMT1-AS1 gene, is highly positively correlated with the LCMT1 gene (r = 0.95). Hi-C data also confirmed an interaction between the LCMT1-AS1 gene and the LCMT1 gene locus (Supplementary Fig B.4).

The expression of the UBL7-AS1 gene was highly correlated with several of its cis-genes. It was negatively correlated with STOML1 (r = -0.87), CSK (r = -0.68), SEMA7A (r = -0.96), CYP1A1 (r = -0.85), ULK3 (r = -0.96), FAM219B (r = -0.94) and CTD-3254N5.1 (r = -0.88). It was positively correlated with SCAMP2 (r = 0.94), RPP25 (r = 0.65), RP11-60L3.1 (r = 0.65) and RP11-10O17.3 (r = 0.95). Hi-C data showed an interaction between the UBL7-AS1 and the CYP1A1 gene.

Lastly, the WAC-AS1 gene showed high negative correlation of expression with the WAC gene (r = -0.91). These two genes are also head-to-head antisense to each other.

4.5 DISCUSSION

Despite growing evidence on the function of lncRNAs, the work done on functionally characterizing them is lagging, specially in the field of human neuroscience. In this pilot study, we designed an experiment to study the effect of knocking down the expression of 16 lncRNAs in human iPSC-derived cortical neurons. Since KD efficiencies obtained with ASOs were not correlated with the lncRNA expression levels in our cell lines, their subcellular localisation or their genomic annotation, we were able to apply the same KD technology to various sub-classes of lncRNAs. We saw several limitations in our study - i) that it was not scalable due to the time consuming KD experiments and it would be impossible to study a large number of lncRNAs using this approach, ii) iPSCs often undergo spontaneous differentiation and are not a highly stable cell model to work with, iii) due to the lack of scalability and the costs of the ASOs, the experiments were also not cost effective for a larger study, iv) due to the low depth of sequencing obtained using LQ-ssCAGE meaningful downstream analyses could not be performed with only very highly expressed and variable genes being flagged by DGE.

The co-expression analyses performed using the RNA-sequencing data from neuronal precursor cells (NPC) from the same cell lines as that of the HiPSCs lent support towards the hypothesis of antisense regulation by several of the antisense lncRNAs - KANLS1-AS1, NOP14-AS1, GABPB1-AS1, WAC-AS1, LCMT1-AS1 and UBL7-AS1. We noted that the NPC lines were much more stable to work with than the iPSC counterparts and the data produced was highly reproducible.

To overcome the shortcomings of our ASO-based pilot phase, we propose a CRISPRi-pooled screening to perform a genomewide perturbation of lncRNAs expressed in NPCs using a custom sgRNA library. We propose the use of NPCs over iPSCs due to their increased stability and scalability aiding in a more robust and reproducible study. In the next chapter, we further develop the pooled CRISPRi screen for lncRNAs, as well as conduct several in-silico analyses to shortlist candindate lncRNAs with plausible roles in neurodegeneration and ageing.

Chapter 5

Global exploration of lncRNA function using a pooled CRISPRi screen and in-silico experiments

5.1 ABSTRACT

In our pilot study described in the previous chapter, we performed ASO-based perturbations of expression of lncRNAs that are highly expressed in the human brain. However, due to the lack of scalability of the study, we propose a genomewide CRISPR-interference (CRISPRi) model of perturbing lncRNA gene expression. This would help address, at a much wider scale, the gap in literature regarding lncRNA function.

Here, we design a CRISPRi based perturbation study for 3804 lncRNAs and first-ever novel short guide RNA (sgRNA) library with 30002 sgRNAs targeting lncRNAs expressed in cortical neurons derived from human neuronal precursor cell lines.

The rationale behind this study centers around the need to perform large-scale studies when studying lncRNAs due to a massive majority of them remaining uncharacterized. For that, reason, in addition to designing a genomewide CRISPRi study, we performed a series of in-silico experiments using in-house and public data to curate a list of 58 candidate lncRNAs that carry strong evidence for function in human neurodegenerative diseases and healthy ageing.

5.2 INTRODUCTION

Global exploration of lncRNA function is essential to broadening the scope with which we study them. We have spoken previously of the mechanisms with which lncRNAs can regulate, enhance and suppress important biological processes such as gene transcription, protein translation, post translational modifications and RNA degradation, yet the work done to investigate their function remains heavily limited to cancer biology and other common diseases.

We stress on the importance of investigating lncRNA function in an unbiased genomewide study to increase the number of known functional lncRNA loci. One of the only large-scale CRISPR-based study for lncRNAs was conducted in 2017 by Liu et al. This study comprised a screen to assess the function of 16,401 lncRNAs in seven different human cell lines. A considerable number (500) of the tested lncRNAs influenced cell growth and in almost all cases the function was highly cell type—specific, often limited to just one cell type. The cell types used in this study were either cancer cell lines or undifferentiated iPSCs. They found the effects of lncRNA KD particularly enhances in iPSCs in the form of reduced cell growth which could be attributed to the increased instability of iPSCs also recorded by us.

A short-guide RNA "guides" the Cas9 nuclease to the genomic region of interest and is central to designing a successful CRISPR experiment. In designing sgRNA libraries, there are several caveats to consider. The first being the kind of CRISPR experiment being performed - knockouts, inactivation or activation of target genes. CRISPRi/a use a catalytically inactive Cas9 system that enables regulation of gene expression. Secondly, a pre-selection of genes whose TSS are to be targeted and ensuring high "on-target" efficiency. Thirdly, minimizing off-target effects which would markedly skew interpretations of results. For the prediction of cleavage efficiency and specificity, numerous computational approaches have been developed for scoring sgRNA activity. Previously, standard alignment methods were used to determine sgRNA specificity by aligning the sgRNAs to a reference genome and off-target sequences/loci returned. Off-late, more sophisticated scoring learning-based methods using training datasets to score sgRNA off-target effects have been implemented. A training model would consider the different featured affecting specificty

such as GC content, DNA methylation, chromatin structure, RNA secondary structure, etc.

We propose a pooled CRISPRi screen for genomewide lncRNA loci in NPCs with the primary goal to shortlist lncRNAs that impact cell growth phenotype. In designing ~ 7 sgRNAs per target lncRNA, we only label those lncRNAs as "essential" that cause a consistent repression of cell growth across all sgRNAs. Once a sub-group of lncRNAs essential to neuronal cell growth has been established, scalable experiments for CROP-Seq would be the next step in studying specific lncRNA mediated gene regulation and cell fate.

In this chapter, we also elaborate on several in-silico analyses applied on large public datasets as well as in-house datasets to produce a list of candidate lncRNAs with evidence of function in cognitive impairment, healthy ageing, and neurodegeneration. Such systematic and unbiased approaches to uncover functional lncRNA loci pave the way to a significantly better understanding of a largely ignored section of genetics in human neurodegeneration and healthy neuronal development.

5.3 METHODS

5.3.1 Selection of Candidate lncRNAs using In-Silico Analyses

Following the pilot phase of the study where lncRNA targets were selected on a primarily technical basis, we sought to perform in-silico functional analyses on public and in-house data to study the functional role of lncRNAs in human neurodegeneration and development.

In silico-analyses led to short-selection of lncRNAs of interest and helped generate hypotheses for their potential roles. The following analyses were formed:

5.3.1.1 Using CAGE-Sequencing data from frontal and temporal brain regions from neurologically healthy controls, as well as pathogenic FTD mutation carriers:

These data were a part of the RiMod-FTD project and post-mortem brain tissues were obtained under a Material Transfer Agreement from the Netherlands Brain Bank. DGE Analysis between controls and FTD mutation carriers (C9orf72-HRE, pathogenic MAPT and GRN mutations) was conducted. Pairwise comparisons were performed between controls and mutation carriers of each gene using 'edgeR' in R. A log-fold change cut-off of 2 was used and a FDR cut-off of 0.05. Frontal and temporal samples were analysed separately.

5.3.1.2 Using the Illumina TruSeq Neurodegeneration Panel (Supplementary Table A.3):

The Illumina TruSeq Neurodegeneration Panel was a result of a collaborative effort by researchers in the Neuroscience community. It consists of 118 risk genes associated with common neurodegenerative disease including FTD, ALS, Alzheimer's Disease, Parkinson's Disease, Dementia with Lewy Bodies, Dystonia and others.

Anti-sense regulation is a common mechanism of lncRNA mediated transcriptional regulation of protein coding genes. To test this, we identified lncRNAs head-to-head antisense to these genes involved in human neurodegeneration. We used the FANTOM CAT genome browser and ZENBU visualization tool [https://fantom.gsc.riken.jp/zenbu/] to study the genomic position of these lncRNAs.

5.3.1.3 Using FANTOM5 data (Hon et al. 2017) to assess eQTL-mRNA correlation of expression for eQTL associated SNPs at lncRNA loci that overlap GWAS hits for neurodegenerative traits using GWAS Catalogue (Buniello et al. 2019).

The FANTOM5 study evaluated the expression correlation of lncRNA-mRNA pairs linked by eQTL-associated SNPs and produced 5264 pairs involving 3166 lncRNAs that were significantly co-expressed. The extent of co expression was measured by absolute Spearman's rho and an eQTL-linked lncRNA-mRNA pair was defined as

'implicated in eQTL' when (1) the distance between the pair was 101.5kb and (2) the pair were significantly more co-expressed than the 75th percentile of the matched background correlation (one-tailed binomial test, P<0.05).

To this effect, we used CAGE-Sequencing data produced from frontal lobe tissue of 119 neurologically healthy individuals (Blauwendraat et al. 2016). Combining CAGE-Sequencing, genotype and exome data, this study identified 2410 eQTLSs and showed that non-coding transcripts are more likely to contain an eQTL than coding transcripts, in particular antisense transcripts. The study also uses data from the GWAS Catalogue to explore possible biological consequences of candidate GWAS variants in the associated region by correlating transcript expression levels with corresponding eQTLs. We annotated the CAGE clusters produced using FANTOM CAT transcripts and detected lncRNAs containing eQTLs that are GWAS hits for neurological traits and disorders associated with the brain such as ALS, bipolar disorder, cognitive performance, autism, multiple sclerosis, migraine, intelligence, hippocampal atrophy, PD, Schizophrenia, white matter hyperintensity burden and alcohol dependence. In addition, using FANTOM5 data, we checked if these lncRNAs are also significantly co-expressed with their corresponding mRNA.

5.3.1.4 Using the RNA-Seq data from the dorsolateral prefrontal cortex of autopsied individuals enrolled in the Religious Orders Study (ROS) or the Rush Memory and Aging Project (MAP), which are jointly designed prospective studies of aging and dementia with detailed, longitudinal cognitive phenotyping during life and a quantitative, structured neuropathologic examination after death (Bennett et al. 2012).

The data consists of a total number of 639 RNA-Sequencing files, belonging to a non-Hispanic white Caucasian population, with 63.9% females and 36.1% males.

Samples were extracted using Qiagen's miRNeasy mini kit and the RNase free DNase Set, and quantified by Nanodrop and quality was evaluated by Agilent Bioanalyzer. Sequencing was performed on the Illumina HiSeq with 101bp paired-end reads and achieved coverage of 150M reads of the first 12 samples. These 12 samples served as a deep coverage reference and included 2 males and 2 females of non-impaired, mild

cognitive impaired, and Alzheimer's cases. This is batch "0". The remaining samples were sequenced with coverage of 50M reads. The libraries were constructed and pooled according to the RIN scores such that similar RIN scores would be pooled together. Varying RIN scores results in a larger spread of insert sizes during library construction and leads to uneven coverage distribution throughout the pool. An additional 57 samples were submitted at a later date to the platform and run on an updated protocol requiring only 250ng of input RNA. This protocol is a modification of Illumina's TruSeq protocol to include long insert sizes and also be strand specific. These late samples were sequenced in batch 2 on plates 7 and 8.

Batch 2 (dorsolateral prefrontal cortex, posterior cingulate cortex, head of caudate nucleus): Sequencing was done on Illumina NovaSeq6000 sequence using 2 x 100bp cycles targeting 30 million reads per sample.

Batch 3 and 4 (dorsolateral prefrontal cortex, posterior cingulate cortex): Libraries were normalized for molarity and sequenced on a NovaSeq 6000 (Illumina) at 40-50M reads, 2 x 150 bp paired-end.

Pre-processing. Raw data in the form of BAM files was downloaded from the AMP-AD Knowledge Portal [https://www.synapse.org/#!Synapse:syn3219045]. These files were converted to paired-end FastQ files using bedtools 'SamToFastq' function. Quality control (QC) checks on the FastQ files were performed using 'fastqc' and QC reports generated. Base quality trimming and adapter clipping was performed using Trimmomatic (v.0.36) (Bolger, Lohse, and Usadel 2014). Using the FANTOM CAT transcript reference FASTA file, a genome index was generated for alignment-free quantification of the RNASeq data using Salmon (v-0.8.2) (Patro et al. 2017). Transcript abundance files generated by Salmon were loaded using the 'tximport' package in R. Gene-level summarization of counts was performed as the current biological question did not require looking at transcripts individually. All zero/lowly expressed genes were filtered and a 'DGEList' with normalised counts generated using the 'calcNormFactors' function of edgeR was prepared for further processing.

A DGEList object is an easy to manipulate data object in R that contains count information for samples, genes and additional information on genomic features. The

'sample' data-frame contains information on library sizes, sequencing depth for each sample as well as the 'group' each sample belongs to, in our case, the grouping is done by the cognitive diagnosis of the individual.

MDS plotting to visualise the data and clustering by various covariates such as sex, Braak Stage, batch, cognitive diagnosis (cogdx), RIN scores, APOE genotype and measurement of neuritic plaque (ceradsc) using the 'PlotMDS' function in R.

From the RNASeq data, one batch (batch 7) clustered away from all of the other data and was filtered out from further analyses to prevent confounding due to batch effects. In addition, all patients that had other forms of dementia (unrelated to Alzheimer's disease) were removed from downstream analysis.

Finally, 189 subjects with no cognitive impairment (NCI), 142 subjects with mild cognitive impairment (MCI) and 201 subjects with Alzheimer's disease and no other form of dementia were retained for downstream analyses.

Differential Gene Expression Analysis.

Generalised Linear Models:
Generalised linear models (GLMs) are an extension of classical linear models to non-normally distributed data. GLMs specify probability distributions according to their mean-variance relationship, for example the quadratic mean-variance relationship specified above for read counts. Assuming that an estimate is available for g, so the variance can be evaluated for any value of μgi, GLM theory can be used to fit a log-linear model

$$log\mu_{gi} = x_i^T \beta_g + logN_i$$

for each gene. Here x_i is a vector of covariates that specifies the treatment conditions applied to RNA sample i, and $_g$ is a vector of regression coefficients by which the covariate effects are mediated for gene g. The quadratic variance function specifies the negative binomial GLM distributional family. The use of the negative binomial distribution is equivalent to treating the π_{gi} as gamma distributed.

Quasi negative binomial:

The negative binomial (NB) model can be extended with quasi-likelihood (QL) methods to account for gene specific variability from both biological and technical sources (Lund et al. 2012). Under the QL framework, the variance of the count ygi is a quadratic function of the mean,

$$var(y_{gi}) = \sigma_g^2(\mu_{gi} + \phi\mu_{gi}^2),$$

where ϕ is the NB dispersion parameter and ${\sigma^2}_{\mathrm{g}}$ is the QL dispersion parameter.

Any increase in the observed variance of ygi will be modelled by an increase in the estimates for ϕ and/or ${\sigma^2}_{\mathrm{g}}$. In this model, the NB dispersion ϕ is a global parameter whereas the QL is gene-specific, so the two dispersion parameters have different roles. The NB dispersion describes the overall biological variability across all genes. It represents the observed variation that is attributable to inherent variability in the biological system, in contrast to the Poisson variation from sequencing. The QL dispersion picks up any gene-specific variability above and below the overall level. In 'edgeR', the QL dispersion estimation and hypothesis testing is done by using the functions 'glmQLFit' and 'glmQLFTest', respectively.

The data was grouped by the 'cogdx' score: 1 for NCI, 2 for MCI and 4 for AD. The design matrix consisted of 3 variables: 'group', 'sex' and 'batch' and was generated using the 'model.matrix function'.

Pairwise comparisons were performed for (i) NCI vs. AD subjects, (ii) MCI vs. AD subjects and (iii) NCI vs. MCI subjects using edgeR's 'makeContrasts' function in R. An FDR threshold of 0.05 was considered.

The DEGs from each pairwise comparison were combined into a vector of 625 gene names that were differentially expressed in at least one comparison. A count table with normalised log counts for these 625 DEGs for the 532 ROSMAP subjects was generated using the 'logCPM' function in edgeR.

K-Means Clustering:

Kmeans algorithm is an iterative algorithm that tries to partition the dataset to 'k' number of pre-defined distinct non-overlapping sub-groups (clusters) where each data point belongs to only one group. A good clustering algorithm would minimize intra-cluster differences between data points of one cluster and maximise inter-cluster differences.

To study the patterns of expressions of DEGs between NCI, MCI and AD, we performed kmeans clustering using the 'ConsensusClusterPlus' library in R (Wilkerson and Hayes 2010). Optimal 'k' was calculated using calculations of the proportion of ambiguous clustering (PAC) for each value of k (values tested: 1-10) tested via consensus clustering.

Means of normalised counts from each group (NCI, MCI and AD) were taken per gene for each cluster. These mean counts were plotted on a scale of 0-1 to study the trend of change in expression using the 'GGPlot2' package in R (Wickham 2009).

Finally, lncRNAs from these k-means clusters that followed a consistent trend of increasing or decreasing in gene expression with the increase in cognitive impairment (NCI $\rightarrow$ MCI $\rightarrow$ AD) were selected as targets for downstream functional analysis.

5.3.1.5 Using CAGE-Sequencing data from the frontal lobe tissue of neurologically healthy individuals who died of causes unrelated to neurodegeneration ranging from the age of 2-95 years (Blauwendraat et al. 2016).

CAGE-Sequencing was performed on frozen human frontal lobe material that was collected from 119 neurologically healthy individuals. Sample information is provided in supplementary table A.4. CAGE cluster count data was downloaded from the original study (Blauwendraat et al. 2016). The counts were re-annotated using a lifted over bed file for human genome build 38 and annotated using the FANTOM CAT GTF file. R packages 'ChIPseeker' (Yu, Wang, and He 2015)and 'stringi' (Gagolewski, 2020) were used for the visualisation of CAGE peaks and annotation. Exploratory analyses to study the effect of covariates such as brain bank and sex were performed using the 'plotMDS' function of edgeR in R. As expected, clustering

was seen for brain banks and thus, to remove confounding batch effects, only samples from University of Maryland were considered as they were highest in number (n=74).

A design matrix using the 'model.matrix' function in R was generated and modeled on the variable 'Age'. First, a linear modelling approach was taken using the 'limma' (Ritchie et al. 2015) package's 'lmFit' function in R. Our code fitted the linear model, smoothened the standard errors with the Empirical Bayes method using the 'eBayes' function and displayed the top 20 ranked genes from the linear model fit using 'topTable' function in R. A summary of the results and differentially expressed genes with age was generated using the 'decideTests' function in 'limma'.

As this approach does not account for non-linear trends, we decided to use a spline regression approach to capture a non-linear relationship of gene expression with age.

Restricted cubic splines:
A restricted cubic split is essentially a piecewise cubic polynomial, and the number of these "pieces" is dictated by the number of windows used (Gauthier, Wu, and Gooley 2020)polynomial, and these windows are defined by "knots". Restrictions are imposed so that the splines are continuous and the curve is smooth so as to leave no gaps between these knots. In addition, in a restricted cubic spline model, the curve is linear before the first knot and after the last one.

To apply this, we used the 'splines' package in R with 5 degrees of freedom (supplied to the model with the argument 'df'). In doing this, we fit a spline to the expression of each gene, using age of each sample as a covariate. Following this, we obtained a fit object using 'lmFit' and 'eBayes' from limma testing for any response of expression to age.

Finally, we used the results from both the simple model and the spline model, using the log-fold change from the simple model and the p-values from the spline model. To this effect, we obtained linear directions of expression change with age per gene and p-values assuming a non-linear dependence of gene expression with age.

Binned age groups model. Additionally, we applied a binned model approach where we formed three bins of ages: 2-25 years, 26 - 45 years, and 46-72 years old representing development, adolescent and ageing phases respectively. Using this age-range variable as the contrasting factor in the contrast matrix, we generated another linear model. Using a FDR threshold of 0.05 and below, we extracted all the genes in common between the spline regression model and the binned age groups model and used those genes as our trustworthy set of differentially expressed genes with healthy ageing. Next, we performed k-means clustering on this set of DEGs and plotted the fitted expression values of the genes per cluster against age. lncRNAs that followed significant patterns of increase or decrease in expression at important stages of ageing were selected for further investigation.

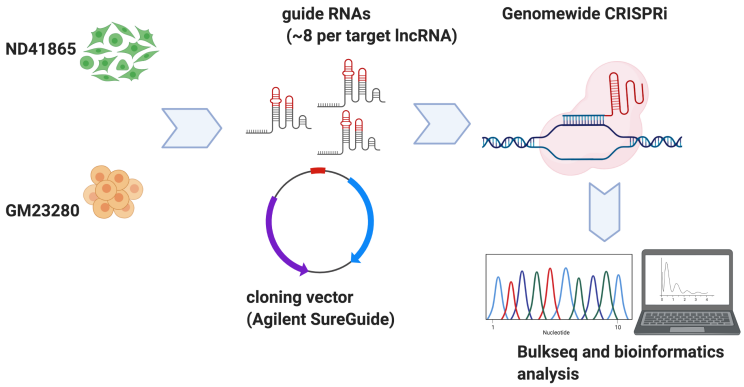
5.3.2 GENOME-WIDE CRISPRI OF LNCRNAS

5.3.2.1 Experimental Design

Pooled CRISPR interference (CRISPRi) screens are powerful tools in studying the functional role of genes being perturbed. Here, we designed a custom library of 30,000 single-guide RNAs (sgRNAs) targeting a total of 3587 lncRNAs. We made use of the inactivated version of the most widely used CRISPR-Cas9 system, the Streptococcus pyogenes dCas9 protein, which can complex with a 110-nucleotide sgRNA containing a 20-nt sequence that complementary binds to the target DNA region and causes a steric block that halts transcriptional elongation by RNA polymerase. This results in the repression of the target gene, in our case, the target lncRNA encoding gene.

We used two human fibroblast derived neuronal precursor cell (NPC) lines from neurologically healthy controls, GM23280 and ND41865 as cellular models in the experimental setup. Both cell lines were transduced with a lentiviral vector to induce Neurogenin-2 (NGN2) expression to produce NPC derived cortical neurons using a published protocol (Dhingra et al. 2020). A commercially available lentiviral vector from Agilent (SureGuide) was selected for our custom sgRNA library, this vector construct includes a U6 promoter, puromycin selection marker and a GFP reporter. To prevent multiplicity of infection (MOI) from exceeding 1, an MOI threshold of 0.3

FIGURE 5.1: Pooled CRISPRi screen. A primary experiment with genomewide CRISPRi for 3600 lncRNA targets. The initial screen will only check for survival as a phenotype and will be used to shortlist "essential" lncRNAs for a perturb-Seq experiment. [Figure designed using BioRender]



was selected. Once the full pooled library of vectors containing the sgRNAs is introduced to the NPCs at day 0 of differentiation, cells would be harvested on days 8 and 16 for PCR amplification. The PCR product would be RNA and CAGE sequenced to study those lncRNAs "essential" to the differentiation and growth of neurons, i.e., those that consistently cause a death phenotype in the cells. These shortlisted lncRNAs would then be part of a more specific CROP-Sequencing study, which entails single cell RNA sequencing of pooled genetic perturbation screens.

The basic experimental design is highlighted in figure 5.1. As a positive control, sgRNAs targeting MALAT1 lncRNA were designed and used. In addition to sgRNAs targeting the lncRNAs, we also designed 360 sgRNAs acting as negative controls, both mapping and non-mapping. To ensure that the pooled sgRNA library is equally represented, we conducted several quality control steps which will be highlighted below.

5.3.2.2 lncRNA target selection for genomewide CRISPRi

In order to ensure that the lncRNAs being targeted in our genome wide CRISPRi screen are expressed in our cellular models, we used RNA-Sequencing data from

days 0, 8 and 16 of differentiation to cortical neurons for our model cell lines: GM23280 and ND41865. As highlighted above, the differentiation protocol used involves transduction with a lentiviral vector to induce Neurogenin-2 (NGN2) expression to produce NPC derived cortical neurons as described in a published protocol (Dhingra et al. 2020).

RNA libraries were prepared on days 0, 8 and 16 following transfection using the Illumina RNA Stranded Gold kit. RNA-Sequencing was performed using the Illumina NextSeq 550 to produce 150 bp paired-end reads.

As primary sequencing output, BCL files were produced. These BCL files were converted to fastQ files using 'bcl2fastq' software by Illumina. In addition to BCL to fastQ conversion, this tool also demultiplexes the samples in the same step. Adapter sequences were trimmed using 'cutadapt' (Martin 2011). FastQC (Andrews S., 2010) and MultiQC (Ewels et al. 2016) softwares were used to visualise the quality of the sequencing data prior to conducting the co-expression analyses. These sequencing data were quantified using 'salmon' (Patro et al. 2017), using the FANTOM CAT transcript reference FASTA which is based on the hg38 genome build. Transcript abundance files were imported from the 'salmon' output using 'tximport' tool (Soneson, Love, and Robinson 2015) in R. A 'DGEList' object was created using the 'edgeR' software (Robinson, McCarthy, and Smyth 2010) in R and the data was normalised using 'calcNormFactors' in 'edgeR'. A matrix of CPMs (counts per million) was generated using 'edgeR'. The data was annotated using FANTOM CAT annotation. For selection, we used a gene expression threshold and selected all lncRNAs that were expressed in each cell line and at each time point at this minimum threshold. For each selected lncRNA, we collated gene information data from the FANTOM CAT server including biotype, to study their genomic positions with respect to their cis-protein coding genes.

5.3.2.3 sgRNA library design

A commonly considered limitation of CRISPR/Cas9 systems is the unexpected cleavage of unintended sites in the genome, i.e., off-target effects. As discussed in the introduction to this chapter, deep-learning based methods help to markedly improve

the ability to predict sgRNA off-target propensity versus previously used alignment-based methods. We applied a published pipeline (Horlbeck et al. 2016) for sgRNA library design for CRISPRi that utilises deep-learning based methods to optimize the design of sgRNAs targeting novel and non-coding genes in humans. This pipeline was trained on data collected from 30 CRISPRi and 9 CRISPRa screens and used the FANTOM consortium's annotation to define the TSSs. This algorithm has previously been used to design genomescale CRISPRi libraries for humans with highly precise sgRNA activity (Horlbeck et al. 2016).

The steps in the pipeline for sgRNA library curation are three-fold:

1. *Learning sgRNA predictors from empirical data.* This step involves loading up empirical data and generating TSS annotations using the FANTOM CAT dataset. Parameters are calculated and fitted in this step for empirical sgRNAs.

2. *Applying a machine learning model to predict sgRNA activity.* This step includes finding all sgRNAs in our genomic regions of interest and prediction of their activity.

3. *Constructing sgRNA libraries.* This step involves the scoring of sgRNAs for their off-target potential and picking the top sgRNAs based on their predicted activity scores and off-target filtering.

In addition to the sgRNAs for the target lncRNAs, we also designed 360 sgRNAs as negative controls that match the base composition of the library by calculating the base frequency at each position of the sgRNA and then generating random sequences weighed by this frequency. 180 of these negative controls are mapping controls which means they map to a random part of the genome far from a variety of genomic features such as genes, enhancers, etc., whereas the remaining 180 are non-mapping which means they do not map anywhere in the genome.

Since the pipeline designs sgRNAs based on the TSSs associated to a given gene and some lncRNAs can have up to 28 TSSs, this can drastically increase the number of sgRNAs needed in a library. To reduce the impact of lncRNAs with many TSSs, the pipeline filters out low abundance TSSs as well as merges TSSs that are located

very near to one another, here we filtered 30% of the CAGE tags associated with a merged on the same strand within 1kb (as long as the resulting loci was less than 1kb). The primary TSS from FANTOM5 was used as a heuristic as the top TSS stays at the top, irrespective. The CAGE-Sequencing samples used for predicting sgRNAs were hiPSC derived cortical neurons from the same cell lines as our model cell lines: GM23280 and ND41865. A minimum percentage of reads of 30% per gene was used as a cutoff for the CAGE clusters. We initially predicted a total of 20 sgRNAs per loci, most of which are filtered out leaving an average of 8 sgRNAs per loci. We dropped any lncRNAs for which at least 7 unique sgRNAs without off-target effects could not be predicted. In addition, if a gene has multiple promoters and one of the promoters did not pass, then the entire gene was excluded.

5.3.2.4 Quality control for sgRNA library representation

To check for library completeness, we implemented two different sequencing strategies:

1. PLASMID SEQUENCING: The plasmid library was directly sequenced.

2. AMPLICON SEQUENCING: The region containing the protospacer within the plasmid was amplified via PCR and then sequenced. Three replicates of the amplicon library were produced using different concentrations of the RNA product: 10ng, 12ng and 20ng.

The steps highlighted in Figure 5.2 were followed in the analysis of the sequencing data and production of count tables. A custom FASTA file using the sgRNA sequences and the flanking regions was generated for alignment of the reads.
Following the count table generation, exploratory analysis to check for the library representation was performed. We used Lorenz curves to visualise the distribution of the counts, which was originally developed by Max O. Lorenz in 1905 for representing inequality of the wealth distribution. In a Lorenz curve, complete equality would be a straight diagonal line with a slope of 1 (the area between this curve and itself is 0, so the Gini coefficient is 0). The Gini Coefficient or Gini Index measures the inequality among the values of a variable. Higher the value of an index, more dispersed is the data. Alternatively, the Gini coefficient can also be calculated as the half of the relative mean absolute difference. A well distributed dataset would have

FIGURE 5.2: Steps for the analysis of the sequencing data and production of count tables for the sgRNA library

the curve as close to the diagonal line of equality as possible and a low Gini coefficient of < 0.2.

5.4 RESULTS

5.4.1 Candidate lncRNA Selection using Public and In-house Datasets

In this section, results from in-silico analysis using public as well as in-house datasets to curate a list of candidate lncRNAs with evidence of function in neurodegeration, cognitive impairment, or healthy ageing.

5.4.1.1 LncRNAs DE in FTD cases versus controls

Using CAGE-Sequencing data from frontal and temporal brain regions from neurologically healthy controls as well as pathogenic mutation carriers for FTD in the GRN, MAPT or C9Orf72 genes, we found a total of 25 lncRNAs that were significantly differentially expressed either in one or both of the tissues (Table 5.1).

5.4.1.2 LncRNAs anti-sense to protein-coding NDD genes

Using the Illumina TruSeq Neurodegeneration Panel, we shortlisted 5 lncRNAs that were antisense to protein-coding genes involved in neurodegeneration (Table 5.2).

Out of these, only 1 lncRNA, EPHA-AS1, was expressed in our cell lines.

5.4.1.3 LncRNAs overlapping eQTLs that are GWAS hits for neurogeneration or developmental traits

Using CAGE-Sequencing data from the frontal lobe tissue of neurologically healthy controls and data from the GWAS catalog, we found 9 lncRNAs that overlap eQTLs that are GWAS hits for neurodegenerative/ developmental traits or diseases (Table 5.3).

5.4.1.4 LncRNAs DE with increasing/decreasing cognitive impairment

Out of the 8 batches in the ROSMAP data, batch 7, clustered separately (Supplementary Fig. B.1). To remove confounding due to this batch effect, all the samples from batch 7 were removed from further analysis. Once filtered for batch 7, the data showed clustering based on sex, as is expected. In addition, all individuals with other causes of cognitive impairment were filtered. For differential expression analyses, sex and batch were included as covariates in all design matrices. We found 625 genes that were significantly ($p<0.05$) in at least one of the pairwise comparisons (AD vs NCI, MCI vs NCI or AD vs MCI). Using k-means clustering with k=5, we divided the DEGs into 5 clusters of genes that follow a specific pattern of expression. Of these 5 clusters, 2 of them showed consistent increase and decrease with increasing cognitive impairment (CI) (Fig. 5.3). In addition to extracting the lncRNA lists from each of these clusters, we also performed extensive literature search on each lncRNA and checked for expression levels in our cell lines. In addition, we checked for expression in the brain tissue and neurons using the FANTOM5 datasets. Finally, we selected 7 lncRNAs that significantly decreased in expression with increasing CI and 4 lncRNAs that significantly increased with increasing CI (Table 5.4).

5.4.1.5 LncRNAs following specific patterns of gene expression with increasing age in neurologically healthy subjects

For this analysis, only samples (n=74) from a single brain bank from the University of Maryland were considered to avoid confounding due to brain bank. Using the

TABLE 5.1: lncRNAs differentially expressed in FTD-causing mutation carriers versus controls using CAGE-Sequencing data from the frontal and temporal lobe tissues.

Gene ID	Gene Name	CAT Gene Class	Temporal Lobe	Frontal Lobe
ENSG00000241956	CTC-340A15.2	lncRNA_intergenic	GRN	GRN
ENSG00000232784	AC067961.1	lncRNA_intergenic	NO	GRN
CATG00000090157	CATG00000090157.1	lncRNA_intergenic	NO	GRN, MAPT
ENSG00000249937	RP11-454P21.1	lncRNA_intergenic	GRN	GRN
ENSG00000214548	MEG3	lncRNA_intergenic	NO	GRN
ENSG00000142396	ERVK3-1	lncRNA_sense_intronic	NO	GRN
ENSG00000271327	RP11-1109F11.3	lncRNA_divergent	GRN	GRN
ENSG00000226281	RP1-80N2.2	lncRNA_intergenic	NO	MAPT
CATG00000017188	CATG00000017188.1	lncRNA_divergent	NO	GRN
ENSG00000228794	RP11-206L10.11	lncRNA_intergenic	NO	GRN
ENSG00000235070	AC068138.1	lncRNA_intergenic	NO	GRN, MAPT
ENSG00000228400	AC079154.1	lncRNA_divergent	GRN	GRN
ENSG00000227053	RP11-395B7.4	lncRNA_divergent	GRN	GRN
CATG00000107731	CATG00000107731.1	lncRNA_intergenic	MAPT, GRN, ALL	GRN, MAPT, ALL
CATG00000099144	CATG00000099144.1	lncRNA_sense_intronic	NO	MAPT
ENSG00000228988	RP4-677H15.4	lncRNA_divergent	NO	GRN
ENSG00000230658	KLHL7-AS1	lncRNA_divergent	NO	GRN
CATG00000041744	CATG00000041744.1	lncRNA_intergenic	GRN	GRN, ALL
ENSG00000221857	CTD-2527I21.4	lncRNA_divergent	GRN	GRN
ENSG00000248810	RP11-362F19.1	lncRNA_divergent	MAPT	NO
ENSG00000237517	DGCR5	lncRNA_intergenic	GRN	NO
ENSG00000245384	AC004053.1	lncRNA_divergent	GRN	NO
CATG00000071146	CATG00000071146.1	lncRNA_intergenic	GRN	NO
ENSG00000188825	LINC00910	lncRNA_intergenic	GRN	NO
ENSG00000231721	MKLN1-AS1	lncRNA_divergent	C9orf72	NO

methods described in section 5.3.1.5, we shortlisted a total of 34 lncRNAs from cluster 1 (Fig 5.4.(a)) and 35 lncRNAs from cluster 2 (Fig 5.4.(b)). Out of these, 8 lncRNAs passed the minimum expression level thresholds in our cell line data (Table 5.5).

5.4.2 Genomewide CRISPRi Study

In this section, we elaborate on the results obtained from our novel sgRNA library design as well as the selection of target lncRNAs for the genomwide CRISPRi screen, the experimental set up for which is described in the Methods section 5.3.2.

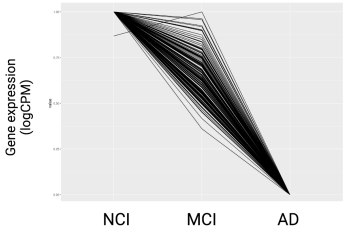
TABLE 5.2: lncRNAs that are antisense to genes on the Illumina Neurodegeneration TruSeq panel with 118 protein-coding genes involved in human neurodegeneration

GENE ID	GENE NAME	CAT Gene Class	ANTISENSE PROTEIN CODING GENE
ENSG00000229153.1	EPHA1-AS1	lncRNA, divergent	EPHA1
ENSG00000248309.1	MEF2C-AS1	p-ncRNA, divergent	MEF2C
ENSG00000117242.7	PINK1-AS	lncRNA, antisense	PINK1
ENSG00000237737.1	DCTN1-AS1	e-lncRNA, divergent	DCTN1
ENSG00000264589.1	MAPT-AS1	p-lncRNA,divergent	MAPT

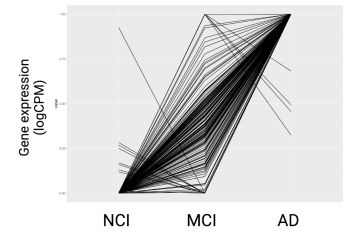
TABLE 5.3: lncRNAs that overlap eQTLs that are GWAS hits for developmental or neurodegenerative traits/disorders

GENE ID	GENE NAME	CAT GENE CLASS	DISEASE TRAIT	SNPs	eQTL-mRNA Coexpression
ENSG00000214401	KANSL1-AS1	lncRNA_divergent	Parkinson's disease	rs11012, rs183211, rs199515, rs199533, rs415430	ARHGAP27
ENSG00000233797	UFL1-AS1	lncRNA_divergent	Migraine	rs11757063, rs11759769	FUT9
CATG00000042617	CATG00000042617.1	lncRNA_divergent	Intelligence (childhood)	rs13387221	None
ENSG00000247728	RP11-932O9.7	lncRNA_divergent	Epilepsy	rs143536437	FAN1
CATG00000009038	CATG00000009038.1	lncRNA_divergent	Migraine without aura	rs1485395	None
CATG00000088147	CATG00000088147.1	lncRNA_divergent	Alcoholism (alcohol use disorder factor score)	rs2140418	NA (CAT Permissive gene)
CATG00000080172	CATG00000080172.1	lncRNA_intergenic	Attention deficit hyperactivity disorder	rs2199161	None
ENSG00000224086	LL22NC03-86G7.1	lncRNA_divergent	Multiple sclerosis	rs2283792	MAPK1, PPIL2
CATG00000004877	CATG00000004877.1	lncRNA_divergent	Schizophrenia	rs4757144	None

FIGURE 5.3: Scaled plots showing trends of change in expression with increasing CI using the ROSMAP datasets.



(a) Cluster 1: A total of 176 genes that consistently decrease in expression with increasing cognitive impairment. Out of these 176 genes, 41 are lncRNAs.

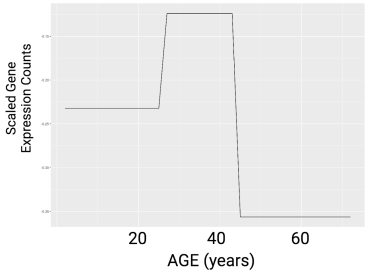


(a) Cluster 3: A total of 177 genes that consistently increase in expression with increasing cognitive impairment. Out of these 177 genes, 22 are lncRNAs.

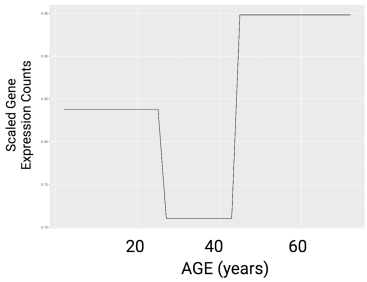
TABLE 5.4: lncRNAs that follow a consistent trend of increase or decrease in expression with increasing cognitive impairment

GENE ID	GENE NAME	CAT GENE CLASS
lncRNAs that follow a trend of downregulation with increasing CI:		
CATG00000039993	CATG00000039993.1	lncRNA_antisense
ENSG00000231312	AC007246.3	lncRNA_divergent
ENSG00000271614	LINC00936	lncRNA_divergent
CATG00000111167	CATG00000111167.1	lncRNA_intergenic
ENSG00000270607	RP11-359E10.1	lncRNA_divergent
ENSG00000132832	RP11-445H22.3	lncRNA_intergenic
CATG00000109338	CATG00000109338.1	lncRNA_divergent
lncRNAs that follow a trend of upregulation with increasing CI:		
ENSG00000235823	LINC00263	lncRNA_intergenic
ENSG00000238230	LINC00391	lncRNA_intergenic
ENSG00000235423	RP11-282O18.3	lncRNA_divergent
ENSG00000182310	LINC00085	lncRNA_intergenic

FIGURE 5.4: Plots of normalised and scaled average gene expression counts for genes that follow specific trends of expression between three key phases of ageing: development/adolescence (2-25 years), adulthood (26-45 years), ageing (46-72 years).



(a) A smoothened curve showing average expression counts in cluster 1 for genes that increase in expression during the developmental/adoleselent phase (2-25 years) and decrease in expression during healthy ageing (46-72 years).



(a) A smoothened curve showing average expression counts in cluster 2 for genes that decrease in expression during the developmental/adoleselent phase (age 2-25 years) and increase in expression during healthy ageing (46-72 years).

TABLE 5.5: lncRNAs that follow specific trends of increase and decrease in expression during the developmental, adolescent and ageing phases of life in neurologically healthy individuals

GENE ID	GENE NAME	CAT GENE CLASS	CLUSTER
CATG00000034323	CATG00000034323.1	lncRNA_divergent	2
CATG00000070525	CATG00000070525.1	lncRNA_divergent	1
CATG00000103964	CATG00000103964.1	lncRNA_divergent	2
ENSG00000242759	LINC00882	lncRNA_intergenic	1
ENSG00000242759	LINC00882	lncRNA_intergenic	1
ENSG00000254226	CTB-12O2.1	lncRNA_intergenic	1
ENSG00000259372	CTD-2240J17.1	lncRNA_intergenic	2
ENSG00000268751	SCGB1B2P	lncRNA_intergenic	2

Here, clusters 1 and 2 refer to lncRNAs belonging to Fig 5.4 (a) and Fig 5.4 (b) respectively.

5.4.2.1 Target lncRNA selection

The target lncRNA list for the genome wide CRISPRi experiment included 3804 lncRNAs that are expressed with a minimum 1 CPM threshold in the RNA-Sequencing data from NPC cell lines from GM23280 and ND41865 on days 0, 8 and 16 of differentiation into cortical neurons. In addition to the highly expressed lncRNAs, our target lncRNA list included all of the lncRNAs selected via the in-silico analyses described in sections 5.3.1. We have collated a comprehensive list of physical features, gene biotypes and FANTOM CAT gene classes using FANTOM5 datasets for each selected lncRNA.

5.4.2.2 sgRNA library

During the sgRNA library design, any lncRNA for which less than 7 sgRNAs could be designed was dropped. In addition, if a gene has multiple promoters and one of those did not pass the selection criteria, then the lncRNA was excluded from the target set. As a result of this, 204 lncRNAs were dropped out of the selected 3804 lncRNAs, resulting in a target set of 3587 lncRNAs. For each of these, a minimum of 7 sgRNAs were designed.

The sgRNA library designed was compatible with the Agilent vector SureGuide CRISPR Library Solutions with a U6 promoter region with 60 nucleotides, variable region with 20 nucleotides and a scaffold region with 60 nucleotides making the

length of each sgRNA 140-nt long. Our SureGuide Custom Amplified pooled library consists of 30,0002 sgRNAs.

In addition to the target lncRNAs, we designed sgRNAs for 360 negative controls, 180 of which are mapping and 180 are non-mapping controls:

Mapping negative controls: Map to regions of the genome far from a variety of genomic features such as genes, enhancers, etc.

Non-mapping negative controls: Do not map to any known location on the genome.

5.4.2.3 Quality Control for Representation of sgRNAs in the Pooled Library

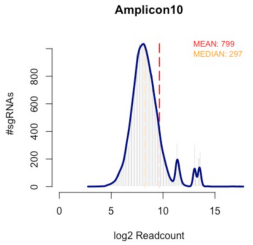
To check for adequate representation of sgRNAs in our pooled library, we plotted their distribution using a smoothed histogram (Fig 5.5 (a)) and a Lorenz curve (Fig. 5.5 (b)). In addition, we checked for representation of sgRNAs/target lncRNA (Fig 5.5 (c)).

As expected, we obtained a bell shaped distribution curve, with a few under and over-represented sgRNAs in either end of the curve. The Lorenz curve (Fig 5.5 (b)) confirmed a near equal distribution of sgRNAs across the pool. We also checked for the representation of sgRNAs per target lncRNA by adding up the counts of reads that mapped to each sgRNA targeting a single lncRNA. In doing this, we saw an even better representation of sgRNA distribution (Gini coefficient = 0.05 and AUC = 0.475) (Fig 5.5(c)).

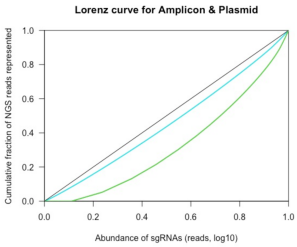
5.5 DISCUSSION

In an effort to effectively study the role of lncRNA biology in neurodegeneration and healthy neuronal development, we curated a genomewide CRISPRi study with a novel and customized sgRNA library targeting all highly expressed lncRNAs in neuronal precursor cell lines. A pooled screen would allow the pre-selection of "essential" lncRNAs based on their impact of their repression on the cell growth phenotype. While several previous studies have investigated the role of lncRNAs using

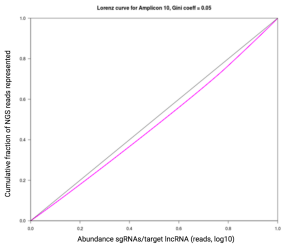
FIGURE 5.5: Representation of sgRNAs in our pooled library of 30002 guides targeting 3857 lncRNAs and 360 negative controls.



(a) Smoothed histogram showing sgRNA expression levels in Amplicon (10ng). The left and right tails of the histogram represent under and over-represented sgRNAs, respectively.



(b) Lorenz Curve with the blue curve showing the distribution of sgRNAs in the sequenced amplicon library and green curve that for the plasmid library with Gini coefficients 0.09 and 0.35, repectively.



(c) Lorenz Curve depicting the distribution of sgRNAs per lncRNA in the amplicon library with Gini coefficients 0.05.

cancer cell lines, this is the first genomewide study, to our knowledge, that examines their role in neuronal cell lines. A pooled screen is a cost-effective means to study their genomewide impact with a proposed next step being CROP-Seq with all filtered "essential" lncRNAs.

We use a deep-learning based model to predict sgRNA activity and control for off-target effects as opposed to traditional alignment based methods to increase precision. Our study involves several layers of quality control measures to ensure a successful study design - all lncRNAs that did not have at least 7 unique sgRNAs without off-target effects as predicted by the deep-learning algorithm were dropped, low abundance TSSs were filtered to reduce the impact of lncRNAs that have a large number of TSSs and the sgRNA library representation to ensure library completeness was performed in steps detailed in section 5.3.2.4. Through quality control steps involving the sequencing of the pooled sgRNA library and studying the counts of the reads that map to each guide, we were able to conclude that the custom sgRNA library is well represented and the expression levels of the sgRNAs is largely uniform both at sgRNA and at lncRNA level. This is important to ensure that drop-out sgRNAs do not result in false positive results. In chapter 4, we noted several limitations when working with iPSC derived neurons such as their scope for spontaneous differentiation and the cost and time involved. Thus, here, we developed a study

using NPCs which do not pose the risks of spontaneous differentiation and are, in turn, more stable, time and cost effective.

Another major strength of this study is the usage of an unbiased approach in the pre-selection of lncRNAs wherein we included all known annotated lncRNAs from the permissive and robust FANTOM CAT gene lists during the target selection process. In doing so, we account for the confounding effects of selection biases from known literature, which is important in the case of lncRNAs as the functions of the majority of lncRNAs remain unclear to-date. Additionally, the in-silico analyses produced a total of 119 lncRNAs with evidence of possible function in neurodegeneration and development.

To conclude, this study offers a comprehensive approach towards the global exploration of the role of lncRNAs in neurodegeneration and development. Results from this study would aid in building a public database of CRISPRi results for meta-analyses of cell-type specific effects of lncRNAs. While such databases exist for the coding genome, it is largely absent for lncRNAs, especially in the field of neurodegenerative diseases.

Chapter 6

Conclusion

Due to their late onset, rapid progression and debilitating symptoms, neurodegenerative diseases are possibly the most devastating set of illnesses. Primary clinical presentations of the disease include dementia, parkinsonism and motor dysfunction. Once these symptoms start developing, symptomatic treatment and management of the disease is the only solution. We now know that the cause or effect of neurodegenerative diseases is often associated to the aggregation of proteins. These proteins are predominantly tau, α-synuclein, TDP-43 or amyloid. These proteinopathies are often overlapped between different neurodegenerative disorders, much like the clinical symptoms, making an accurate diagnosis of disease difficult and affecting the statistical significance of most studies. In the course of the work done under this dissertation, we witnessed this first hand. Over the span of just 3 years, the diagnoses of patients were often changed based on their evolving clinical symptoms. All of these factors make it extremely difficult to diagnose, study and intercept at-risk individuals in time. The advancement of genomic technologies over the last decade has enabled researchers to study disease progression, associated pathways and genes more closely.

The work presented in this dissertation is, in part, an effort to study the genetic variability of FTD-ALS, which comprises a heterogeneous set of disorders with varying clinical and morphological phenotypes. The second chapter titled "Exploring the genetic landscape of FTD in a German Cohort" aims to study a largely homogenous population of German patients and assess the distribution of genetic risk variants and pathogenic variants in this population. We use an in-depth, systematic and streamlined approach in studying 463 patients in the DESCRIBE-FTD study to 1) identify the percentage of carriers of the pathogenic C9Orf72 HRE in the population

using RP-PCR, 2) identify large INDELs in GRN and MAPT genes using MLPA, 3) identify all known pathogenic mutations in a preconceived list of 22 genes, 4) identify potentially pathogenic mutations using a genetic screening strategy described in section 2.4.3.4 and 5) study the burden of rare damaging variants human autophagy-associated genes in FTD/ALS cases versus controls. In order to ensure accurate representation of frequencies, we also performed a kinship analysis and removed all cryptically related pairs with degree of relatedness = 2 or less. While our findings in some of the most common FTD genes - C9Orf72, GRN, MAPT, TBK1 - remain consistent with the literature, we were able to expand the genetic landscape of FTD-ALS via some unusual findings. We found pathogenic and potentially pathogenic mutations in APP, PSEN1, PSEN2 genes which are associated with AD and the CTSF gene which is associated with Type B Kufs Disease, which is an adult-onset neuronal ceroid lipofuscinosis, associated with a severe, early-onset neuropsychiatric phenotype with early epileptic seizures (Smith et al. 2013). We, thus, show here that adult-onset FTD and ALS can be caused by mutations in amyloid genes as well as in the Cathepsin F gene. We concur with a previous study (Blauwendraat et al. 2018) that proposes the inclusion of the CTSF gene in future genetic screens for FTD. Finally, we were able to identify 4 human autophagy genes that carried an excessive burden of rare damaging variants in FTD/ALS patients, with the top candidate being the SERPINA1 gene. It is important to also comment on the fact that we find a rare pathogenic mutation in the UBA domain of the SQSTM1 gene in one of the patients, which is also a gene integral to the human autophagosome. This study had the major benefit of coming from an extremely homogeneous population and of having similar technologies being used to perform the sequencing, making our data considerably less noisy. All of our findings were verified using Sanger sequencing and visualised on the IGV browser.

In Chapter 3 titled "Genetic Landscape of FTD/ALS in a broader Western European Population", we include a geographically wider group of affected individuals to study changes in frequencies of genes that carry pathogenic and potentially pathogenic mutations as well as to conduct a genomewide rare variant association test to identify potential candidate genes involved in disease progression. Using a systematic approach similar to that in chapter 2, we discovered the distribution of pathogenic variants and potentially pathogenic variants in the previously selected

22 NDD genes, CTSF and CYP27A1. CYP27A1 a cholesterol homeostasis gene that has previously been linked with FTD ((Blauwendraat et al. 2018). Interestingly, we found a pathogenic CYP27A1 mutation in an FTD patient from the Italian cohort, once again, broadening the genetic landscape of genetic FTD/ALS. Additionally, we found 31 potential candidate genes involved in a multitude of pathways - immunity, autophagy, lipid metabolism, ion transport and digestion - that tested significantly in our rare variant association test. All of these pathways have been previously hypothesized to be associated with FTD/ALS, and our findings could lead a window into investigation of how these pathways converge in disease progression. Interestingly, SERPINA1 also crossed genome wide significance in our rare-variant association tests using 731 cases from western Europe. As validation, we found that the genomewide association study conducted by Ferrari et al., in 2014 also showed suggestive evidence of SERPINA1 gene variants being associated with PNFA. Our findings here strongly suggest the inclusion of SERPINA1 as a candidate gene in future FTD screens. Consistent with the findings of Chapter 2, we find a high burden of rare damaging variants in autophagy genes in cases versus controls. We also find an additional pathogenic variant as well as 7 potentially pathogenic variants in the SQSTM1 gene, adding to our belief that the human autophagosome is a key pathway in FTD/ALS disease progression.

One major limitation of our genetic study is the absence of family data. Although we see several variants with strong evidence of being damaging, we are unable to ascertain their pathogenicity without observing co-segregation within a family, which forms a foundational part of studying rare variation in human genetics. If a variant can be observed to co-segregate with a phenotype within a family, the evidence for its association with the disease is greatly strengthened. Integrating genomic findings with family data provides excellent opportunities to find highly penetrant rare variants, and thus discover important disease mechanisms. However, due to the costs involved, as well as due to patient privacy concerns, large scale studies seldom comprise a thorough family history and clinically healthy relatives are seldom analysed. With the decreasing costs of sequencing, technological advancements as well as large integrated public databases of genetic variation, there is increased scope for integrative studies involving family data. This would be a massive step in uncovering the missing heritability of FTD/ALS and what I hope lies in the immediate

future of FTD genetics.

Additionally, with these aforementioned decreased costs and technological advancements, the potential to conduct WGS experiments also rises. With exome-seq, we capture a very small percentage of the genome, ignoring important regulatory genetic elements. In addition, WGS provides a much more uniform coverage of the genome, including exomes, as discussed in chapter 2. As we now know, the often ignored and highly prevalent class of ncRNAs - lncRNAs - have several mechanisms by which they regulate gene expression. Being able to study whole genomes of patients, including their families, would be a second major step in studying disease modifying/causing genetic variation in FTD/ALS.

In an attempt to uncover the missing heritability and often ignored aspect of studying the genetics of a rare and polygenic disease, we dedicated our efforts in Chapters 4 and 5 to develop experimental designs and a systematic approach to investigate the roles of long non-coding RNAs in NDDs. We present a pilot study investigating 20 highly expressed lncRNAs in the brain using ASO based KD and transcriptomics analysis of days 3 and 8 after knockdown. Due to several limitations in this study pertaining to a low depth of sequencing, high costs and the time involved in ASO based KD experiments leading to a lack of scalability, we progressed to the second phase of our study involving a genomewide CRISPRi for which we have pre-selected $\approx$3600 lncRNAs that are expressed in our neuronal precursor cell lines and generated a novel sgRNA library containing 30002 sgRNAs targeting each of the target lncRNAs as well as 360 negative controls. Several quality control steps were performed to ensure the integrity of the sgRNA library and ensure each sgRNA is well represented.

In addition to selecting for high expression, we performed a series of in-silico analyses using both in-house and public data to gather functional evidence of lncRNAs in ageing, cognitive impairment, antisense regulation of NDD genes, eQTL associated gene regulation as well as those that were differentially expressed in FTD cases versus controls. One of the reasons, we believe, for lncRNAs to go uninvestigated in the field of neuroscience is because the task of investigating 4000 lncRNAs that are expressed in the human brain without a hypothesis seems monumental and daunting.

Our unbiased approach in selecting potentially functional lncRNAs has resulted in the curation of a list of 119 lncRNAs with evidence for roles in neurodegeneration and ageing. Integrating these results into our novel sgRNA library which targets all highly expressed lncRNAs in NPCs offers a unique opportunity for the global exploration of lncRNA function in the field of neuroscience. This is, to the best of our knowledge, the first large scale perturbation study of lncRNAs in neurons.

Our efforts in the course of the work done under Chapters 2, 3 4 and 5 have been to uncover as much of the genetic variability that underlies the FTD/ALS spectrum as possible in the scope of the project, and, later, to also investigate non-coding genetic elements which have consistently proven to play important regulatory roles. We propose the inclusion of the CTSF and SERPINA1 genes in future FTD/ALS genetic screens, and further investigation of the human autophagosome with relation to FTD/ALS. We identified several potentially pathogenic variants which would require functional validation before the carriers of these variants can be considered "solved" for their diagnosis. Finally, our work with the lncRNA study paves the way for future scientists to have a starting point in investigating the roles played by lncRNAs in both neurodegeneration and healthy ageing.

In conclusion, massive strides have been made in the past decade towards the development of sophisticated, low-cost and high coverage sequencing technologies, which we have used to our advantage in investigating new genes and variants that play potential role in FTD/ALS disease mechanisms. Several studies, including ours, lack family data which significantly lowers their potential to conclusively annotate pathogenic variation. A strategic approach to accelerate the understanding of the yet undiscovered genetics of NDDs would, in my opinion, be through the inclusion of family data, investment into whole genome sequencing and increased investigation of non-coding genetic elements, along with a collaborative approach that involves publicly accessible integration of genetic findings.

Bibliography

[1] Abu-Rumeileh, Samir, Steffen Halbgebauer, Petra Steinacker, Sarah Anderl-Straub, Barbara Polischi, Albert C. Ludolph, Sabina Capellari, Piero Parchi, and Markus Otto. 2020. "CSF SerpinA1 in Creutzfeldt-Jakob Disease and Frontotemporal Lobar Degeneration." Annals of Clinical and Translational Neurology 7 (2): 191–99.

[2] Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. "A Method and Server for Predicting Damaging Missense Mutations." Nature Methods 7 (4): 248–49.

[3] Al-Sarraj, Safa, Andrew King, Claire Troakes, Bradley Smith, Satomi Maekawa, Istvan Bodi, Boris Rogelj, Ammar Al-Chalabi, Tibor Hortobágyi, and Christopher E. Shaw. 2011. "p62 Positive, TDP-43 Negative, Neuronal Cytoplasmic and Intranuclear Inclusions in the Cerebellum and Hippocampus Define the Pathology of C9orf72-Linked FTLD and MND/ALS." Acta Neuropathologica 122 (6): 691–702.

[4] Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. "An Atlas of Active Enhancers across Human Cell Types and Tissues." Nature 507 (7493): 455–61.

[5] Andersson, Robin, Peter Refsing Andersen, Eivind Valen, Leighton J. Core, Jette Bornholdt, Mette Boyd, Torben Heick Jensen, and Albin Sandelin. 2014. "Nuclear Stability and Transcriptional Directionality Separate Functionally Distinct RNA Species." Nature Communications 5 (November): 5336.

[6] Andrews, T. D., G. Sjollema, and C. C. Goodnow. 2013. "Understanding the Immunological Impact of the Human Mutation Explosion." Trends in Immunology 34 (3): 99–106.

[7] Apps, Matthew A. J., Matthew F. S. Rushworth, and Steve W. C. Chang. 2016. "The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others." Neuron 90 (4): 692–707.

[8] Ash, Peter E. A., Kevin F. Bieniek, Tania F. Gendron, Thomas Caulfield, Wen-Lang Lin, Mariely Dejesus-Hernandez, Marka M. van Blitterswijk, et al. 2013. "Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS." Neuron 77 (4): 639–46.

[9] Baker, M., I. Litvan, H. Houlden, J. Adamson, D. Dickson, J. Perez-Tur, J. Hardy, T. Lynch, E. Bigio, and M. Hutton. 1999. "Association of an Extended Haplotype in the Tau Gene with Progressive Supranuclear Palsy." Human Molecular Genetics 8 (4): 711–15.

[10] Baker, M., I. Litvan, H. Houlden, J. Adamson, D. Dickson, J. Perez-Tur, J. Hardy, T. Lynch, E. Bigio, and M. Hutton. 1999. "Association of an Extended Haplotype in the Tau Gene with Progressive Supranuclear Palsy." Human Molecular Genetics 8 (4): 711–15.

[11] Basu, Saonli, and Wei Pan. 2011. "Comparison of Statistical Tests for Disease Association with Rare Variants." Genetic Epidemiology 35 (7): 606–19.

[12] Bateman, Andrew, and Hugh P. J. Bennett. 2009. "The Granulin Gene Family: From Cancer to Dementia." BioEssays. https://doi.org/10.1002/bies.200900086.

[13] Beach, Thomas G., Charles H. Adler, Lucia I. Sue, Geidy Serrano, Holly A. Shill, Douglas G. Walker, Lihfen Lue, et al. 2015. "Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program." Neuropathology: Official Journal of the Japanese Society of Neuropathology 35 (4): 354–89.

[14] Beermann, Julia, Maria-Teresa Piccoli, Janika Viereck, and Thomas Thum. 2016. "Non-Coding RNAs in Development and Disease: Background, Mechanisms, and Therapeutic Approaches." Physiological Reviews 96 (4): 1297–1325.

[15] Bellot, Grégory, Raquel Garcia-Medina, Pierre Gounon, Johanna Chiche, Danièle Roux, Jacques Pouysségur, and Nathalie M. Mazure. 2009. "Hypoxia-Induced Autophagy Is Mediated through Hypoxia-Inducible Factor Induction of BNIP3 and BNIP3L via Their BH3 Domains." Molecular and Cellular Biology 29 (10): 2570–81.

[16] Belzil, Veronique V., Hussein Daoud, Judith St-Onge, Anne Desjarlais, Jean-Pierre Bouchard, Nicolas Dupre, Lucette Lacomblez, et al. 2011. "Identification of Novel FUS Mutations in Sporadic Cases of Amyotrophic Lateral Sclerosis." Amyotrophic Lateral Sclerosis: Official Publication of the World Federation of Neurology Research Group on Motor Neuron Diseases 12 (2): 113–17.

[17] Bennett, David A., Julie A. Schneider, Zoe Arvanitakis, and Robert S. Wilson. 2012. "Overview and Findings from the Religious Orders Study." Current Alzheimer Research 9 (6): 628–45.

[18] Benussi, Alberto, Alessandro Padovani, and Barbara Borroni. 2015. "Phenotypic Heterogeneity of Monogenic Frontotemporal Dementia." Frontiers in Aging Neuroscience 7 (September): 171.

[19] Bigio, Eileen H. 2011. "TDP-43 Variants of Frontotemporal Lobar Degeneration." Journal of Molecular Neuroscience. https://doi.org/10.1007/s12031-011-9545-z.

[20] Bjørkøy, Geir, Trond Lamark, and Terje Johansen. 2006. "p62/SQSTM1: A Missing Link between Protein Aggregates and the Autophagy Machinery." Autophagy 2 (2): 138–39.

[21] Blauwendraat, Cornelis, Carlo Wilke, Javier Simón-Sánchez, Iris E. Jansen, Anika Reifschneider, Anja Capell, Christian Haass, et al. 2018. "The Wide Genetic Landscape of Clinical Frontotemporal Dementia: Systematic Combined Sequencing of 121 Consecutive Subjects." Genetics in Medicine: Official Journal of the American College of Medical Genetics 20 (2): 240–49.

[22] Blauwendraat, Cornelis, Margherita Francescatto, J. Raphael Gibbs, Iris E. Jansen, Javier Simón-Sánchez, Dena G. Hernandez, Allissa A. Dillman, et al. 2016. "Comprehensive Promoter Level Expression Quantitative Trait Loci Analysis of the Human Frontal Lobe." Genome Medicine 8 (1): 65.

[23] Blue, Elizabeth E., Joshua C. Bis, Michael O. Dorschner, Debby W. Tsuang, Sandra M. Barral, Gary Beecham, Jennifer E. Below, et al. 2018. "Genetic Variation in Genes Underlying Diverse Dementias May Explain a Small Proportion of Cases in the Alzheimer's Disease Sequencing Project." Dementia and Geriatric Cognitive Disorders 45 (1-2): 1–17.

[24] Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." Bioinformatics 30 (15): 2114–20.

[25] Boxer, Adam L., Michael Gold, Edward Huey, Fen-Biao Gao, Edward A. Burton, Tiffany Chow, Aimee Kao, et al. 2013. "Frontotemporal Degeneration, the next Therapeutic Frontier: Molecules and Animal Models for Frontotemporal Degeneration Drug Development." Alzheimer's & Dementia: The Journal of the Alzheimer's Association 9 (2): 176–88.

[26] Brandt, R., J. Léger, and G. Lee. 1995. "Interaction of Tau with the Neural Plasma Membrane Mediated by Tau's Amino-Terminal Projection Domain." Journal of Cell Biology. https://doi.org/10.1083/jcb.131.5.1327.

[27] Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." Nucleic Acids Research 47 (D1): D1005–12.

[28] Buratti, Emanuele, and Francisco E. Baralle. 2008. "Multiple Roles of TDP-43 in Gene Expression, Splicing Regulation, and Human Disease." Frontiers in Bioscience: A Journal and Virtual Library 13 (January): 867–78.

[29] Burrell, James R., Matthew C. Kiernan, Steve Vucic, and John R. Hodges. 2011. "Motor Neuron Dysfunction in Frontotemporal Dementia." Brain. https://doi.org/10.1093/brain/awr195.

[30] Chiò, Adriano, Andrea Calvo, Cristina Moglia, Irene Ossola, Maura Brunetti, Luca Sbaiz, Shiao-Lin Lai, Yevgeniya Abramzon, Bryan J. Traynor, and Gabriella Restagno. 2011. "A de Novo Missense Mutation of the FUS Gene in a 'true' Sporadic ALS Case." Neurobiology of Aging. https://doi.org/10.1016/j.neurobiolaging.2010.05.016.

[31] Clarke, Mica T. M., Frédéric St-Onge, Jean-Mathieu Beauregard, Martina Bocchetta, Emily Todd, David M. Cash, Jonathan D. Rohrer, and Robert Laforce Jr. 2021. "Early Anterior Cingulate Involvement Is Seen in Presymptomatic MAPT P301L Mutation Carriers." Alzheimer's Research & Therapy 13 (1): 42.

[32] Consortium, The Encode Project, and The ENCODE Project Consortium. 2004. "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science. https://doi.org/10.1126/science.1105136.

[33] Conte, Amelia, Serena Lattante, Marcella Zollino, Giuseppe Marangi, Marco Luigetti, Alessandra Del Grande, Serenella Servidei, Federica Trombetta, and Mario Sabatelli. 2012. "P525L FUS Mutation Is Consistently Associated with a Severe Form of Juvenile Amyotrophic Lateral Sclerosis." Neuromuscular Disorders: NMD 22 (1): 73–75.

[34] Cruts, Marc, Ilse Gijselinck, Julie van der Zee, Sebastiaan Engelborghs, Hans Wils, Daniel Pirici, Rosa Rademakers, et al. 2006. "Null Mutations in Progranulin Cause Ubiquitin-Positive Frontotemporal Dementia Linked to Chromosome 17q21." Nature 442 (7105): 920–24.

[35] Cuyvers, Elise, Karolien Bettens, Stéphanie Philtjens, Tim Van Langenhove, Ilse Gijselinck, Julie van der Zee, Sebastiaan Engelborghs, et al. 2014. "Investigating the Role of Rare Heterozygous TREM2 Variants in Alzheimer's Disease and Frontotemporal Dementia." Neurobiology of Aging 35 (3): 726.e11–19.

[36] Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." Bioinformatics 27 (15): 2156–58.

[37] Datlinger, Paul, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. 2017. "Pooled CRISPR Screening with Single-Cell Transcriptome Readout." Nature Methods 14 (3): 297–301.

[38] Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." PLoS Computational Biology 6 (12): e1001025.

[39] DeJesus-Hernandez, Mariely, Ian R. Mackenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, et al. 2011. "Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS." Neuron 72 (2): 245–56.

[40] DeJesus-Hernandez, Mariely, Jannet Kocerha, Nicole Finch, Richard Crook, Matt Baker, Pamela Desaro, Amelia Johnston, et al. 2010. "De Novo Truncating FUS Gene Mutation as a Cause of Sporadic Amyotrophic Lateral Sclerosis." Human Mutation 31 (5): E1377–89.

[41] Deng, Han-Xiang, Wenjie Chen, Seong-Tshool Hong, Kym M. Boycott, George H. Gorrie, Nailah Siddique, Yi Yang, et al. 2011. "Mutations in UBQLN2 Cause Dominant X-Linked Juvenile and Adult-Onset ALS and ALS/dementia." Nature 477 (7363): 211–15.

[42] Deramecourt, V., F. Lebert, B. Debachy, M. A. Mackowiak-Cordoliani, S. Bombois, O. Kerdraon, L. Buée, C-A Maurage, and F. Pasquier. 2010. "Prediction of Pathology in Primary Progressive Language and Speech Disorders." Neurology 74 (1): 42–49.

[43] Dhingra, Ashutosh, Joachim Täger, Elisangela Bressan, Salvador Rodriguez-Nieto, Manmeet-Sakshi Bedi, Stefanie Bröer, Eldem Sadikoglou, et al. 2020. "Automated Production of Human Induced Pluripotent Stem Cell-Derived Cortical and Dopaminergic Neurons with Integrated Live-Cell Monitoring." Journal of Visualized Experiments: JoVE, no. 162 (August). https://doi.org/10.3791/61525.

[44] Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, et al. 2016. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens." Cell 167 (7): 1853–66.e17.

[45] Dixit, Ram, Jennifer L. Ross, Yale E. Goldman, and Erika L. F. Holzbaur. 2008. "Differential Regulation of Dynein and Kinesin Motor Proteins by Tau." Science 319 (5866): 1086–89.

[46] Dols-Icardo, Oriol, Alberto García-Redondo, Ricardo Rojas-García, Daniel Borrego-Hernández, Ignacio Illán-Gala, José Luís Muñoz-Blanco, Alberto

Rábano, et al. 2018. "Analysis of Known Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Genes Reveals a Substantial Genetic Burden in Patients Manifesting Both Diseases Not Carrying the Expansion Mutation." Journal of Neurology, Neurosurgery, and Psychiatry 89 (2): 162–68.

[47] Ebbert, Mark T. W., Christian A. Ross, Luc J. Pregent, Rebecca J. Lank, Cheng Zhang, Rebecca B. Katzman, Karen Jansen-West, et al. 2017. "Conserved DNA Methylation Combined with Differential Frontal Cortex and Cerebellar Expression Distinguishes C9orf72-Associated and Sporadic ALS, and Implicates SERPINA1 in Disease." Acta Neuropathologica 134 (5): 715–28.

[48] Eriksen, Jason L., and Ian R. A. Mackenzie. 2008. "Progranulin: Normal Function and Role in Neurodegeneration." Journal of Neurochemistry 104 (2): 287–97.

[49] Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." Bioinformatics 32 (19): 3047–48.

[50] Fabre, S. F., C. Forsell, M. Viitanen, M. Sjögren, A. Wallin, K. Blennow, M. Blomberg, C. Andersen, L. O. Wahlund, and L. Lannfelt. 2001. "Clinic-Based Cases with Frontotemporal Dementia Show Increased Cerebrospinal Fluid Tau and High Apolipoprotein E epsilon4 Frequency, but No Tau Gene Mutations." Experimental Neurology 168 (2): 413–18.

[51] Farh, Kyle Kai-How, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, et al. 2015. "Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants." Nature. https://doi.org/10.1038/nature13835.

[52] Ferrari, Raffaele, Dena G. Hernandez, Michael A. Nalls, Jonathan D. Rohrer, Adaikalavan Ramasamy, John B. J. Kwok, Carol Dobson-Stone, et al. 2014. "Frontotemporal Dementia and Its Subtypes: A Genome-Wide Association Study." Lancet Neurology 13 (7): 686–99.

[53] Ferrari, Raffaele, Paola Forabosco, Jana Vandrovcova, Juan A. Botía, Sebastian Guelfi, Jason D. Warren, UK Brain Expression Consortium (UKBEC), et

al. 2016. “Frontotemporal Dementia: Insights into the Biological Underpinnings of Disease through Gene Co-Expression Network Analysis.” Molecular Neurodegeneration 11 (February): 21.

[54] Flier, Wiesje M. van der, and Philip Scheltens. 2018. “Amsterdam Dementia Cohort: Performing Research to Optimize Care.” Journal of Alzheimer’s Disease: JAD 62 (3): 1091–1111.

[55] Foster, Norman L., Kirk Wilhelmsen, Anders A. F. Sima, Margaret Z. Jones, Constance J. D’Amato, Sid Gilman, and Conference Participants. 1997. “Frontotemporal Dementia and Parkinsonism Linked to Chromosome 17: A Consensus Conference.” Annals of Neurology. https://doi.org/10.1002/ana.410410606.

[56] Freischmidt, Axel, Kathrin Müller, Albert C. Ludolph, Jochen H. Weishaupt, and Peter M. Andersen. 2017. “Association of Mutations in TBK1 With Sporadic and Familial Amyotrophic Lateral Sclerosis and Frontotemporal Dementia.” JAMA Neurology 74 (1): 110–13.

[57] Freischmidt, Axel, Thomas Wieland, Benjamin Richter, Wolfgang Ruf, Veronique Schaeffer, Kathrin Müller, Nicolai Marroquin, et al. 2015. “Haploinsufficiency of TBK1 Causes Familial ALS and Fronto-Temporal Dementia.” Nature Neuroscience 18 (5): 631–36.

[58] Fuentes Fajardo, Karin V., David Adams, NISC Comparative Sequencing Program, Christopher E. Mason, Murat Sincan, Cynthia Tifft, Camilo Toro, Cornelius F. Boerkoel, William Gahl, and Thomas Markello. 2012. “Detecting False-Positive Signals in Exome Sequencing.” Human Mutation 33 (4): 609–13.

[59] Gauthier-Kemper, Anne, Carina Weissmann, Nataliya Golovyashkina, Zsofia Sebö-Lemke, Gerard Drewes, Volker Gerke, Jürgen J. Heinisch, and Roland Brandt. 2011. “The Frontotemporal Dementia Mutation R406W Blocks Tau’s Interaction with the Membrane in an Annexin A2–dependent Manner.” Journal of Cell Biology. https://doi.org/10.1083/jcb.201007161.

[60] Gauthier, J., Q. V. Wu, and T. A. Gooley. 2020. “Cubic Splines to Model Relationships between Continuous Variables and Outcomes: A Guide for Clinicians.” Bone Marrow Transplantation 55 (4): 675–80.

[61] Gendron, Tania F., Kevin F. Bieniek, Yong-Jie Zhang, Karen Jansen-West, Peter E. A. Ash, Thomas Caulfield, Lillian Daughrity, et al. 2013. "Antisense Transcripts of the Expanded C9ORF72 Hexanucleotide Repeat Form Nuclear RNA Foci and Undergo Repeat-Associated Non-ATG Translation in c9FTD/ALS." Acta Neuropathologica 126 (6): 829–44.

[62] Ghanim, Mustapha, Léna Guillot-Noel, Florence Pasquier, Ludmila Jornea, Vincent Deramecourt, Bruno Dubois, Isabelle Le Ber, Alexis Brice, and French Research Network on FTD and FTD/MND. 2010. "CHMP2B Mutations Are Rare in French Families with Frontotemporal Lobar Degeneration." Journal of Neurology 257 (12): 2032–36.

[63] Ghidoni, Roberta, Simona Signorini, Laura Barbiero, Elena Sina, Paola Cominelli, Aldo Villa, Luisa Benussi, and Giuliano Binetti. 2006. "The H2 MAPT Haplotype Is Associated with Familial Frontotemporal Dementia." Neurobiology of Disease 22 (2): 357–62.

[64] Gijselinck, I., C. Van Broeckhoven, and M. Cruts. 2008. "Granulin Mutations Associated with Frontotemporal Lobar Degeneration and Related Disorders: An Update." Human Mutation 29 (12): 1373–86.

[65] Gijselinck, Ilse, Sara Van Mossevelde, Julie van der Zee, Anne Sieben, Stéphanie Philtjens, Bavo Heeman, Sebastiaan Engelborghs, et al. 2015. "Loss ofTBK1is a Frequent Cause of Frontotemporal Dementia in a Belgian Cohort." Neurology.

[66] Goedert, M., and R. Jakes. 1990. "Expression of Separate Isoforms of Human Tau Protein: Correlation with the Tau Pattern in Brain and Effects on Tubulin Polymerization." The EMBO Journal 9 (13): 4225–30.

[67] Goedert, M., M. G. Spillantini, R. Jakes, D. Rutherford, and R. A. Crowther. 1989. "Multiple Isoforms of Human Microtubule-Associated Protein Tau: Sequences and Localization in Neurofibrillary Tangles of Alzheimer's Disease." Neuron 3 (4): 519–26.

[68] Goedert, Michel, and Maria Grazia Spillantini. 2011. "Pathogenesis of the Tauopathies." Journal of Molecular Neuroscience. https://doi.org/10.1007/s12031-011-9593-4.

[69] Guerreiro, Rita, Aleksandra Wojtas, Jose Bras, Minerva Carrasquillo, Ekaterina Rogaeva, Elisa Majounie, Carlos Cruchaga, et al. 2013. "TREM2 Variants in Alzheimer's Disease." The New England Journal of Medicine 368 (2): 117–27.

[70] Haque, Ashraful, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. 2017. "A Practical Guide to Single-Cell RNA-Sequencing for Biomedical Research and Clinical Applications." Genome Medicine 9 (1): 75.

[71] Han, Jeong-Ho, Hyun-Hee Ryu, Mi-Hee Jun, Deok-Jin Jang, and Jin-A Lee. 2012. "The Functional Analysis of the CHMP2B Missense Mutation Associated with Neurodegenerative Diseases in the Endo-Lysosomal Pathway." Biochemical and Biophysical Research Communications. https://doi.org/10.1016/j.bbrc.2012.04.041.

[72] Hatakeyama, Shigetsugu. 2017. "TRIM Family Proteins: Roles in Autophagy, Immunity, and Carcinogenesis." Trends in Biochemical Sciences. https://doi.org/10.1016/j.tibs.2017.01.002.

[73] He, Zhuohao, Jing L. Guo, Jennifer D. McBride, Sneha Narasimhan, Hyesung Kim, Lakshmi Changolkar, Bin Zhang, et al. 2018. "Amyloid- Plaques Enhance Alzheimer's Brain Tau-Seeded Pathologies by Facilitating Neuritic Plaque Tau Aggregation." Nature Medicine 24 (1): 29–38.

[74] Hofman, Albert, Sarwa Darwish Murad, Cornelia M. van Duijn, Oscar H. Franco, André Goedegebure, M. Arfan Ikram, Caroline C. W. Klaver, et al. 2013. "The Rotterdam Study: 2014 Objectives and Design Update." European Journal of Epidemiology 28 (11): 889–926.

[75] Hogan, David B., Nathalie Jetté, Kirsten M. Fiest, Jodie I. Roberts, Dawn Pearson, Eric E. Smith, Pamela Roach, Andrew Kirk, Tamara Pringsheim, and Colleen J. Maxwell. 2016. "The Prevalence and Incidence of Frontotemporal Dementia: A Systematic Review." The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques 43 Suppl 1 (April): S96–109.

[76] Höglinger, Günter U., Gesine Respondek, Maria Stamelou, Carolin Kurz, Keith A. Josephs, Anthony E. Lang, Brit Mollenhauer, et al. 2017. "Clinical Diagnosis of Progressive Supranuclear Palsy: The Movement Disorder Society

Criteria." Movement Disorders: Official Journal of the Movement Disorder Society 32 (6): 853–64.

[77] Hon, Chung-Chau, Jordan A. Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen J. L. Rackham, Julian Gough, Elena Denisenko, et al. 2017. "An Atlas of Human Long Non-Coding RNAs with Accurate 5' Ends." Nature 543 (7644): 199–204.

[78] Horlbeck, Max A., Luke A. Gilbert, Jacqueline E. Villalta, Britt Adamson, Ryan A. Pak, Yuwen Chen, Alexander P. Fields, et al. 2016. "Compact and Highly Active next-Generation Libraries for CRISPR-Mediated Gene Repression and Activation." eLife 5 (September). https://doi.org/10.7554/eLife.19760.

[79] Huey, Edward D., Jordan Grafman, Eric M. Wassermann, Pietro Pietrini, Michael C. Tierney, Bernardino Ghetti, Salvatore Spina, et al. 2006. "Characteristics of Frontotemporal Dementia Patients with a Progranulin Mutation." Annals of Neurology 60 (3): 374–80.

[80] Hutton, M., C. L. Lendon, P. Rizzu, M. Baker, S. Froelich, H. Houlden, S. Pickering-Brown, et al. 1998. "Association of Missense and 5'-Splice-Site Mutations in Tau with the Inherited Dementia FTDP-17." Nature 393 (6686): 702–5.

[81] Ido, Bademain Jean Fabrice, Imen Kacem, Mahamadi Ouedraogo, Amina Nasri, Saloua Mrabet, Amina Gargouri, Mouna Ben Djebara, Bawindsongré Jean Kabore, and Riadh Gouider. 2021. "Sensitivity of Awaji Criteria and Revised El Escorial Criteria in the Diagnosis of Amyotrophic Lateral Sclerosis (ALS) at First Visit in a Tunisian Cohort." Neurology Research International 2021 (January): 8841281.

[82] Irwin, David J., Nigel J. Cairns, Murray Grossman, Corey T. McMillan, Edward B. Lee, Vivianna M. Van Deerlin, Virginia M-Y Lee, and John Q. Trojanowski. 2015. "Frontotemporal Lobar Degeneration: Defining Phenotypic Diversity through Personalized Medicine." Acta Neuropathologica 129 (4): 469–91.

[83] Jansen, Iris E., Hui Ye, Sasja Heetveld, Marie C. Lechler, Helen Michels, Renée I. Seinstra, Steven J. Lubbe, et al. 2017. "Discovery and Functional Prioritization of Parkinson's Disease Candidate Genes from Large-Scale Whole Exome Sequencing." Genome Biology 18 (1): 22.

[84] Jin, Sheng Chih, Minerva M. Carrasquillo, Bruno A. Benitez, Tara Skorupa, David Carrell, Dwani Patel, Sarah Lincoln, et al. 2015. "TREM2 Is Associated with Increased Risk for Alzheimer's Disease in African Americans." Molecular Neurodegeneration 10 (April): 19.

[85] Josephs, Keith A., John R. Hodges, Julie S. Snowden, Ian R. Mackenzie, Manuela Neumann, David M. Mann, and Dennis W. Dickson. 2011. "Neuropathological Background of Phenotypical Variability in Frontotemporal Dementia." Acta Neuropathologica 122 (2): 137–53.

[86] Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." Nature 581 (7809): 434–43.

[87] Kent, W. J. 2002. "The Human Genome Browser at UCSC." Genome Research. https://doi.org/10.1101/gr.229102.

[88] Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." Nature Genetics 46 (3): 310–15.

[89] Kleinberger, Gernot, Anja Capell, Christian Haass, and Christine Van Broeckhoven. 2013. "Mechanisms of Granulin Deficiency: Lessons from Cellular and Animal Models." Molecular Neurobiology 47 (1): 337–60.

[90] Kumar, Prateek, Steven Henikoff, and Pauline C. Ng. 2009. "Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm." Nature Protocols 4 (7): 1073–81.

[91] Kuusisto, E., A. Salminen, and I. Alafuzoff. 2002. "Early Accumulation of p62 in Neurofibrillary Tangles in Alzheimer's Disease: Possible Role in Tangle Formation." Neuropathology and Applied Neurobiology. https://doi.org/10.1046/j.1365-2990.2002.00394.x.

[92] Kwiatkowski, T. J., Jr, D. A. Bosco, A. L. Leclerc, E. Tamrazian, C. R. Vanderburg, C. Russ, A. Davis, et al. 2009. "Mutations in the FUS/TLS Gene on

Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis." Science 323 (5918): 1205–8.

[93] Lagier-Tourenne, Clotilde, Magdalini Polymenidou, and Don W. Cleveland. 2010. "TDP-43 and FUS/TLS: Emerging Roles in RNA Processing and Neurodegeneration." Human Molecular Genetics 19 (R1): R46–64.

[94] Landqvist Waldö, Maria, Alexander Frizell Santillo, Lars Gustafson, Elisabet Englund, and Ulla Passant. 2014. "Somatic Complaints in Frontotemporal Dementia." American Journal of Neurodegenerative Disease 3 (2): 84–92.

[95] Larson, Matthew H., Luke A. Gilbert, Xiaowo Wang, Wendell A. Lim, Jonathan S. Weissman, and Lei S. Qi. 2013. "CRISPR Interference (CRISPRi) for Sequence-Specific Control of Gene Expression." Nature Protocols 8 (11): 2180–96.

[96] Lashley, Tammaryn, Jonathan D. Rohrer, Rina Bandopadhyay, Charles Fry, Zeshan Ahmed, Adrian M. Isaacs, Jack H. Brelstaff, et al. 2011. "A Comparative Clinical, Pathological, Biochemical and Genetic Study of Fused in Sarcoma Proteinopathies." Brain: A Journal of Neurology 134 (Pt 9): 2548–64.

[97] Le Ber, Isabelle, Julie van der Zee, Didier Hannequin, Ilse Gijselinck, Dominique Campion, Michèle Puel, Annie Laquerrière, et al. 2007. "Progranulin Null Mutations in Both Sporadic and Familial Frontotemporal Dementia." Human Mutation 28 (9): 846–55.

[98] Lee, Seunggeun, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, David C. Christiani, Mark M. Wurfel, and Xihong Lin. 2012. "Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies." American Journal of Human Genetics 91 (2): 224–37.

[99] Lendon, C. L., T. Lynch, J. Norton, D. W. McKeel Jr, F. Busfield, N. Craddock, S. Chakraverty, et al. 1998. "Hereditary Dysphasic Disinhibition Dementia: A Frontotemporal Dementia Linked to 17q21-22." Neurology 50 (6): 1546–55.

[100] Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." Bioinformatics 25 (14): 1754–60.

[101] Liu, Changning, Baoyan Bai, Geir Skogerbø, Lun Cai, Wei Deng, Yong Zhang, Dongbo Bu, Yi Zhao, and Runsheng Chen. 2005. "NONCODE: An Integrated Knowledge Database of Non-Coding RNAs." Nucleic Acids Research 33 (Database issue): D112–15.

[102] Ljubenkov, Peter A., Adam M. Staffaroni, Julio C. Rojas, Isabel E. Allen, Ping Wang, Hilary Heuer, Anna Karydas, et al. 2018. "Cerebrospinal Fluid Biomarkers Predict Frontotemporal Dementia Trajectory." Annals of Clinical and Translational Neurology 5 (10): 1250–63.

[103] Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." Genome Biology 15 (12): 550.

[104] Lund, Steven P., Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth. 2012. "Detecting Differential Expression in RNA-Sequence Data Using Quasi-Likelihood with Shrunken Dispersion Estimates." Statistical Applications in Genetics and Molecular Biology 11 (5). https://doi.org/10.1515/1544-6115.1826.

[105] Mackenzie, Ian R. A., Matt Baker, Stuart Pickering-Brown, Ging-Yuek R. Hsiung, Caroline Lindholm, Emily Dwosh, Jennifer Gass, et al. 2006. "The Neuropathology of Frontotemporal Lobar Degeneration Caused by Mutations in the Progranulin Gene." Brain: A Journal of Neurology 129 (Pt 11): 3081–90.

[106] Mackenzie, Ian R. A. 2007. "The Neuropathology and Clinical Phenotype of FTD with Progranulin Mutations." Acta Neuropathologica 114 (1): 49–54.

[107] Mackenzie, Ian R. A., Manuela Neumann, Eileen H. Bigio, Nigel J. Cairns, Irina Alafuzoff, Jillian Kril, Gabor G. Kovacs, et al. 2009. "Nomenclature for Neuropathologic Subtypes of Frontotemporal Lobar Degeneration: Consensus Recommendations." Acta Neuropathologica 117 (1): 15–18.

[108] Mackenzie, Ian R. A., Manuela Neumann, Eileen H. Bigio, Nigel J. Cairns, Irina Alafuzoff, Jillian Kril, Gabor G. Kovacs, et al. 2010. "Nomenclature and Nosology for Neuropathologic Subtypes of Frontotemporal Lobar Degeneration: An Update." Acta Neuropathologica 119 (1): 1–4.

[109] Mackenzie, Ian R., Thomas Arzberger, Elisabeth Kremmer, Dirk Troost, Stefan Lorenzl, Kohji Mori, Shih-Ming Weng, et al. 2013. "Dipeptide Repeat Protein Pathology in C9ORF72 Mutation Cases: Clinico-Pathological Correlations." Acta Neuropathologica 126 (6): 859–79.

[110] Mackenzie, Ian R. A., and Manuela Neumann. 2016. "Molecular Neuropathology of Frontotemporal Dementia: Insights into Disease Mechanisms from Postmortem Studies." Journal of Neurochemistry 138 Suppl 1 (August): 54–70.

[111] Maeda, Shintaro, Chinatsu Otomo, and Takanori Otomo. 2019. "The Autophagic Membrane Tether ATG2A Transfers Lipids between Membranes." eLife 8 (July). https://doi.org/10.7554/eLife.45777.

[112] Majounie, Elisa, Alan E. Renton, Kin Mok, Elise G. P. Dopper, Adrian Waite, Sara Rollinson, Adriano Chiò, et al. 2012. "Frequency of the C9orf72 Hexanucleotide Repeat Expansion in Patients with Amyotrophic Lateral Sclerosis and Frontotemporal Dementia: A Cross-Sectional Study." Lancet Neurology 11 (4): 323–30.

[113] Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." Bioinformatics 26 (22): 2867–73.

[114] Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." EMBnet.journal 17 (1): 10.

[115] McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." Genome Biology 17 (1): 122.

[116] Meynert, Alison M., Morad Ansari, David R. FitzPatrick, and Martin S. Taylor. 2014. "Variant Detection Sensitivity and Biases in Whole Genome and Exome Sequencing." BMC Bioinformatics 15 (July): 247.

[117] Mishra, Aniket, Raffaele Ferrari, Peter Heutink, John Hardy, Yolande Pijnenburg, Danielle Posthuma, and International FTD-Genomics Consortium. 2017. "Gene-Based Association Studies Report Genetic Links for Clinical Subtypes of Frontotemporal Dementia." Brain: A Journal of Neurology 140 (5): 1437–46.

[118] Mizielinska, Sarah, Tammaryn Lashley, Frances E. Norona, Emma L. Clayton, Charlotte E. Ridler, Pietro Fratta, and Adrian M. Isaacs. 2013. "C9orf72 Frontotemporal Lobar Degeneration Is Characterised by Frequent Neuronal Sense and Antisense RNA Foci." Acta Neuropathologica 126 (6): 845–57.

[119] Moore, Katrina M., Jennifer Nicholas, Murray Grossman, Corey T. McMillan, David J. Irwin, Lauren Massimo, Vivianna M. Van Deerlin, et al. 2020. "Age at Symptom Onset and Death and Disease Duration in Genetic Frontotemporal Dementia: An International Retrospective Cohort Study." Lancet Neurology 19 (2): 145–56.

[120] Mori, Kohji, Shih-Ming Weng, Thomas Arzberger, Stephanie May, Kristin Rentzsch, Elisabeth Kremmer, Bettina Schmid, et al. 2013. "The C9orf72 GGGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS." Science 339 (6125): 1335–38.

[121] Murray, Melissa E., Mariely DeJesus-Hernandez, Nicola J. Rutherford, Matt Baker, Ranjan Duara, Neill R. Graff-Radford, Zbigniew K. Wszolek, et al. 2011. "Clinical and Neuropathologic Heterogeneity of c9FTD/ALS Associated with Hexanucleotide Repeat Expansion in C9ORF72." Acta Neuropathologica 122 (6): 673–90.

[122] Naasan, Georges, Howard J. Rosen, Gil Dan Rabinovici, Bruce L. Miller, Jon Elofson, Giovanni Coppola, Anna Karydas, Jamie Fong, William Seeley, and William Jagust. 2014. "O4-01-04: AMYLOID IN DEMENTIA ASSOCIATED WITH FAMILIAL FTLD: NOT AN INNOCENT BYSTANDER." Alzheimer's & Dementia. https://doi.org/10.1016/j.jalz.2014.04.387.

[123] Nakaso, Kazuhiro, Yuko Yoshimoto, Toshiya Nakano, Takao Takeshima, Yoko Fukuhara, Kenichi Yasui, Shigeru Araga, Toru Yanagawa, Tetsuro Ishii, and Kenji Nakashima. 2004. "Transcriptional Activation of p62/A170/ZIP during the Formation of the Aggregates: Possible Mechanisms and the Role in Lewy Body Formation in Parkinson's Disease." Brain Research 1012 (1-2): 42–51.

[124] Navarro Gonzalez, Jairo, Ann S. Zweig, Matthew L. Speir, Daniel Schmelter, Kate R. Rosenbloom, Brian J. Raney, Conner C. Powell, et al. 2021. "The UCSC Genome Browser Database: 2021 Update." Nucleic Acids Research 49 (D1): D1046–57.

[125] Neale, Benjamin M., Manuel A. Rivas, Benjamin F. Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M. Purcell, Kathryn Roeder, and Mark J. Daly. 2011. "Testing for an Unusual Distribution of Rare Variants." PLoS Genetics 7 (3): e1001322.

[126] Neary, D., J. S. Snowden, L. Gustafson, U. Passant, D. Stuss, S. Black, M. Freedman, et al. 1998. "Frontotemporal Lobar Degeneration: A Consensus on Clinical Diagnostic Criteria." Neurology 51 (6): 1546–54.

[127] Neguembor, Maria Victoria, Mathivanan Jothi, and Davide Gabellini. 2014. "Long Noncoding RNAs, Emerging Players in Muscle Differentiation and Disease." Skeletal Muscle 4 (1): 8.

[128] Neumann, Manuela, Deepak M. Sampathu, Linda K. Kwong, Adam C. Truax, Matthew C. Micsenyi, Thomas T. Chou, Jennifer Bruce, et al. 2006. "Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis." Science 314 (5796): 130–33.

[129] Neumann, Manuela, Sigrun Roeber, Hans A. Kretzschmar, Rosa Rademakers, Matt Baker, and Ian R. A. Mackenzie. 2009. "Abundant FUS-Immunoreactive Pathology in Neuronal Intermediate Filament Inclusion Disease." Acta Neuropathologica 118 (5): 605–16.

[130] Neve, R. L., P. Harris, K. S. Kosik, D. M. Kurnit, and T. A. Donlon. 1986. "Identification of cDNA Clones for the Human Microtubule-Associated Protein Tau and Chromosomal Localization of the Genes for Tau and Microtubule-Associated Protein 2." Brain Research 387 (3): 271–80.

[131] Paila, Umadevi, Brad A. Chapman, Rory Kirchner, and Aaron R. Quinlan. 2013. "GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations." PLoS Computational Biology 9 (7): e1003153.

[132] Park, Sun Ah, Sang Il Ahn, and Jean-Marc Gallo. 2016. "Tau Mis-Splicing in the Pathogenesis of Neurodegenerative Disorders." BMB Reports 49 (8): 405–13.

[133] Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." Nature Methods 14 (4): 417–19.

[134] Pick, Arnold, D. M. Girling, and G. E. Berrios. 1994. "On the Relationship between Senile Cerebral Atrophy and Aphasia." History of Psychiatry. https://doi.org/10.1177/0957154x9400502007.

[135] Pooler, Amy M., Alessia Usardi, Catherine J. Evans, Karen L. Philpott, Wendy Noble, and Diane P. Hanger. 2012. "Dynamic Association of Tau with Neuronal Membranes Is Regulated by Phosphorylation." Neurobiology of Aging. https://doi.org/10.1016/j.neurobiolaging.2011.01.005.

[136] Poorkaj, P., T. D. Bird, E. Wijsman, E. Nemens, R. M. Garruto, L. Anderson, A. Andreadis, W. C. Wiederholt, M. Raskind, and G. D. Schellenberg. 1998. "Tau Is a Candidate Gene for Chromosome 17 Frontotemporal Dementia." Annals of Neurology 43 (6): 815–25.

[137] Pottier, Cyril, Thomas A. Ravenscroft, Monica Sanchez-Contreras, and Rosa Rademakers. 2016. "Genetics of FTLD: Overview and What Else We Can Expect from Genetic Studies." Journal of Neurochemistry 138 Suppl 1 (August): 32–53.

[138] Preker, Pascal, Jesper Nielsen, Susanne Kammler, Søren Lykke-Andersen, Marianne S. Christensen, Christophe K. Mapendano, Mikkel H. Schierup, and Torben Heick Jensen. 2008. "RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters." Science 322 (5909): 1851–54.

[139] Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." American Journal of Human Genetics 81 (3): 559–75.

[140] Rademakers, Rosa, Manuela Neumann, and Ian R. Mackenzie. 2012. "Advances in Understanding the Molecular Basis of Frontotemporal Dementia." Nature Reviews. Neurology 8 (8): 423–34.

[141] Ramilowski, Jordan A., Chi Wai Yip, Saumya Agrawal, Jen-Chien Chang, Yari Ciani, Ivan V. Kulakovskiy, Mickaël Mendez, et al. 2020. "Functional Annotation of Human Long Noncoding RNAs via Molecular Phenotyping." Genome Research 30 (7): 1060–72.

[142] Rascovsky, Katya, John R. Hodges, David Knopman, Mario F. Mendez, Joel H. Kramer, John Neuhaus, John C. van Swieten, et al. 2011. "Sensitivity of Revised Diagnostic Criteria for the Behavioural Variant of Frontotemporal Dementia." Brain: A Journal of Neurology 134 (Pt 9): 2456–77.

[143] Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. "g:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update)." Nucleic Acids Research 47 (W1): W191–98.

[144] Raux, G., R. Gantier, C. Thomas-Anterion, J. Boulliat, P. Verpillat, D. Hannequin, A. Brice, T. Frebourg, and D. Campion. 2000. "Dementia with Prominent Frontotemporal Features Associated with L113P Presenilin 1 Mutation." Neurology 55 (10): 1577–78.

[145] Religa, Dorota, Seyed-Mohammad Fereshtehnejad, Pavla Cermakova, Ann-Katrin Edlund, Sara Garcia-Ptacek, Nicklas Granqvist, Anne Hallbäck, et al. 2015. "SveDem, the Swedish Dementia Registry - a Tool for Improving the Quality of Diagnostics, Treatment and Care of Dementia Patients in Clinical Practice." PloS One 10 (2): e0116538.

[146] Renton, Alan E., Elisa Majounie, Adrian Waite, Javier Simón-Sánchez, Sara Rollinson, J. Raphael Gibbs, Jennifer C. Schymick, et al. 2011. "A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD." Neuron 72 (2): 257–68.

[147] Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." Nucleic Acids Research 43 (7): e47.

[148] Rizzu, Patrizia, John C. Van Swieten, Marijke Joosse, Masato Hasegawa, Martijn Stevens, Aad Tibben, Martinus F. Niermeijer, et al. 1999. "High Prevalence of Mutations in the Microtubule-Associated Protein Tau in a Population Study of Frontotemporal Dementia in the Netherlands." The American Journal of Human Genetics.

[149] Robak, Laurie A., Iris E. Jansen, Jeroen van Rooij, André G. Uitterlinden, Robert Kraaij, Joseph Jankovic, International Parkinson's Disease Genomics

Consortium (IPDGC), Peter Heutink, and Joshua M. Shulman. 2017. "Excessive Burden of Lysosomal Storage Disorder Gene Variants in Parkinson's Disease." Brain: A Journal of Neurology 140 (12): 3191–3203.

[150] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." Bioinformatics 26 (1): 139–40.

[151] Rohrer, J. D., R. Guerreiro, J. Vandrovcova, J. Uphill, D. Reiman, J. Beck, A. M. Isaacs, et al. 2009. "The Heritability and Genetics of Frontotemporal Lobar Degeneration." Neurology 73 (18): 1451–56.

[152] Rohrer, Jonathan D., Tammaryn Lashley, Jonathan M. Schott, Jane E. Warren, Simon Mead, Adrian M. Isaacs, Jonathan Beck, et al. 2011. "Clinical and Neuroanatomical Signatures of Tissue Pathology in Frontotemporal Lobar Degeneration." Brain: A Journal of Neurology 134 (Pt 9): 2565–81.

[153] Rosso, Sonia M., Laura Donker Kaat, Timo Baks, Marijke Joosse, Inge de Koning, Yolande Pijnenburg, Daniëlle de Jong, et al. 2003. "Frontotemporal Dementia in The Netherlands: Patient Characteristics and Prevalence Estimates from a Population-Based Study." Brain: A Journal of Neurology 126 (Pt 9): 2016–22.

[154] Salta, Evgenia, and Bart De Strooper. 2017. "Noncoding RNAs in Neurodegeneration." Nature Reviews. Neuroscience 18 (10): 627–40.

[155] Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." Proceedings of the National Academy of Sciences of the United States of America 74 (12): 5463–67.

[156] Seeger, M., R. Kraft, K. Ferrell, D. Bech-Otschir, R. Dumdey, R. Schade, C. Gordon, M. Naumann, and W. Dubiel. 1998. "A Novel Protein Complex Involved in Signal Transduction Possessing Similarities to 26S Proteasome Subunits." FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology 12 (6): 469–78.

[157] Seelaar, H., J. D. Rohrer, Y. A. L. Pijnenburg, N. C. Fox, and J. C. van Swieten. 2011. "Clinical, Genetic and Pathological Heterogeneity of Frontotemporal Dementia: A Review." Journal of Neurology, Neurosurgery & Psychiatry. https://doi.org/10.1136/jnnp.2010.212225.

[158] Seibenhener, M. Lamar, M. Lamar Seibenhener, Jeganathan Ramesh Babu, Thangiah Geetha, Hing C. Wong, N. Rama Krishna, and Marie W. Wooten. 2004. "Sequestosome 1/p62 Is a Polyubiquitin Chain Binding Protein Involved in Ubiquitin Proteasome Degradation." Molecular and Cellular Biology. https://doi.org/10.1128/mcb.24.18.8055-8068.2004.

[159] Sherry, S. T. 2001. "dbSNP: The NCBI Database of Genetic Variation." Nucleic Acids Research. https://doi.org/10.1093/nar/29.1.308.

[160] Shiraki, Toshiyuki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, et al. 2003. "Cap Analysis Gene Expression for High-Throughput Analysis of Transcriptional Starting Point and Identification of Promoter Usage." Proceedings of the National Academy of Sciences of the United States of America 100 (26): 15776–81.

[161] Sieben, Anne, Tim Van Langenhove, Sebastiaan Engelborghs, Jean-Jacques Martin, Paul Boon, Patrick Cras, Peter-Paul De Deyn, Patrick Santens, Christine Van Broeckhoven, and Marc Cruts. 2012. "The Genetics and Neuropathology of Frontotemporal Lobar Degeneration." Acta Neuropathologica 124 (3): 353–72.

[162] Smith, Katherine R., Hans-Henrik M. Dahl, Laura Canafoglia, Eva Andermann, John Damiano, Michela Morbin, Amalia C. Bruni, et al. 2013. "Cathepsin F Mutations Cause Type B Kufs Disease, an Adult-Onset Neuronal Ceroid Lipofuscinosis." Human Molecular Genetics 22 (7): 1417–23.

[163] Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences." F1000Research 4 (December): 1521.

[164] Spillantini, M. G., J. R. Murrell, M. Goedert, M. R. Farlow, A. Klug, and B. Ghetti. 1998. "Mutation in the Tau Gene in Familial Multiple System Tauopathy with Presenile Dementia." Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.95.13.7737.

[165] Stokholm, Jette, Thomas W. Teasdale, Peter Johannsen, Jorgen E. Nielsen, Troels Tolstrup Nielsen, Adrian Isaacs, Jerry M. Brown, Anders Gade, and Frontotemporal dementia Research in Jutland Association (FReJA) consortium. 2013. "Cognitive Impairment in the Preclinical Stage of Dementia in FTD-3 CHMP2B Mutation Carriers: A Longitudinal Prospective Study." Journal of Neurology, Neurosurgery, and Psychiatry 84 (2): 170–76.

[166] Sun, Li, and Jason L. Eriksen. 2011. "Recent Insights into the Involvement of Progranulin in Frontotemporal Dementia." Current Neuropharmacology. https://doi.org/10.2174/157015911798376361.

[167] Tessitore, Alessandra, Marinella Pirozzi, and Alberto Auricchio. 2009. "Abnormal Autophagy, Ubiquitination, Inflammation and Apoptosis Are Dependent upon Lysosomal Storage and Are Useful Biomarkers of Mucopolysaccharidosis VI." PathoGenetics 2 (1): 4.

[168] Thul, Peter J., and Cecilia Lindskog. 2018. "The Human Protein Atlas: A Spatial Map of the Human Proteome." Protein Science. https://doi.org/10.1002/pro.3307.

[169] To, Wing Ting, Dirk De Ridder, Tomas Menovsky, John Hart, and Sven Vanneste. 2017. "The Role of the Dorsal Anterior Cingulate Cortex (dACC) in a Cognitive and Emotional Counting Stroop Task: Two Cases." Restorative Neurology and Neuroscience 35 (3): 333–45.

[170] Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. 2012. "Primer3—new Capabilities and Interfaces." Nucleic Acids Research. https://doi.org/10.1093/nar/gks596.

[171] Van Langenhove, T., J. van der Zee, K. Sleegers, S. Engelborghs, R. Vandenberghe, I. Gijselinck, M. Van den Broeck, et al. 2010. "Genetic Contribution of FUS to Frontotemporal Lobar Degeneration." Neurology 74 (5): 366–71.

[172] Vance, Caroline, Boris Rogelj, Tibor Hortobágyi, Kurt J. De Vos, Agnes Lumi Nishimura, Jemeen Sreedharan, Xun Hu, et al. 2009. "Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6." Science 323 (5918): 1208–11.

[173] Vieira, Daniela, João Durães, Inês Baldeiras, Beatriz Santiago, Diana Duro, Marisa Lima, Maria João Leitão, Miguel Tábuas-Pereira, and Isabel Santana. 2019. "Lower CSF Amyloid-Beta1–42 Predicts a Higher Mortality Rate in Frontotemporal Dementia." Diagnostics. https://doi.org/10.3390/diagnostics9040162.

[174] Vieira, Renata Teles, Leonardo Caixeta, Sergio Machado, Adriana Cardoso Silva, Antonio Egidio Nardi, Oscar Arias-Carrión, and Mauro Giovanni Carta. 2013. "Epidemiology of Early-Onset Dementia: A Review of the Literature." Clinical Practice and Epidemiology in Mental Health: CP & EMH 9 (June): 88–95.

[175] Volders, Pieter-Jan, Kenny Helsens, Xiaowei Wang, Björn Menten, Lennart Martens, Kris Gevaert, Jo Vandesompele, and Pieter Mestdagh. 2013. "LNCipedia: A Database for Annotated Human lncRNA Transcript Sequences and Structures." Nucleic Acids Research 41 (Database issue): D246–51.

[176] Wang, Yanli, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, et al. 2018. "The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions." Genome Biology 19 (1): 151.

[177] Wapinski, Orly, and Howard Y. Chang. 2011. "Long Noncoding RNAs and Human Disease." Trends in Cell Biology. https://doi.org/10.1016/j.tcb.2011.04.001.

[178] Whitman, Mary C., Noriko Miyake, Elaine H. Nguyen, Jessica L. Bell, Paola M. Matos Ruiz, Wai-Man Chan, Silvio Alessandro Di Gioia, et al. 2019. "Decreased ACKR3 (CXCR7) Function Causes Oculomotor Synkinesis in Mice and Humans." Human Molecular Genetics 28 (18): 3113–25.

[179] Wickham, Hadley. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media.

[180] Wijker, M., Z. K. Wszolek, E. C. Wolters, M. A. Rooimans, G. Pals, R. F. Pfeiffer, T. Lynch, R. L. Rodnitzky, K. C. Wilhelmsen, and F. Arwert. 1996. "Localization of the Gene for Rapidly Progressive Autosomal Dominant Parkinsonism and Dementia with Pallido-Ponto-Nigral Degeneration to Chromosome 17q21." Human Molecular Genetics 5 (1): 151–54.

[181] Wilkerson, Matthew D., and D. Neil Hayes. 2010. "ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking." Bioinformatics 26 (12): 1572–73.

[182] Wu, Josephine J., Ashley Cai, Jessie E. Greenslade, Nicole R. Higgins, Cong Fan, Nhat T. T. Le, Micaela Tatman, et al. 2020. "ALS/FTD Mutations in UBQLN2 Impede Autophagy by Reducing Autophagosome Acidification through Loss of Function." Proceedings of the National Academy of Sciences of the United States of America 117 (26): 15230–41.

[183] Yan, J., H-X Deng, N. Siddique, F. Fecto, W. Chen, Y. Yang, E. Liu, et al. 2010. "Frameshift and Novel Mutations in FUS in Familial Amyotrophic Lateral Sclerosis and ALS/dementia." Neurology 75 (9): 807–14.

[184] Yu, Guangchuang, Li-Gen Wang, and Qing-Yu He. 2015. "ChIPseeker: An R/Bioconductor Package for ChIP Peak Annotation, Comparison and Visualization." Bioinformatics 31 (14): 2382–83.

[185] Zatloukal, Kurt, Cornelia Stumptner, Andrea Fuchsbichler, Hans Heid, Martina Schnoelzer, Lukas Kenner, Reinhold Kleinert, Marco Prinz, Adriano Aguzzi, and Helmut Denk. 2002. "p62 Is a Common Component of Cytoplasmic Inclusions in Protein Aggregation Diseases." The American Journal of Pathology 160 (1): 255–63.

[186] Zee, Julie van der, Ilse Gijselinck, Lubina Dillen, Tim Van Langenhove, Jessie Theuns, Sebastiaan Engelborghs, Stéphanie Philtjens, et al. 2013. "A Pan-European Study of the C9orf72 Repeat Associated with FTLD: Geographic Prevalence, Genomic Instability, and Intermediate Repeats." Human Mutation 34 (2): 363–73.

[187] Zee, Julie van der, Ilse Gijselinck, Sara Van Mossevelde, Federica Perrone, Lubina Dillen, Bavo Heeman, Veerle Bäumer, et al. 2017. "TBK1 Mutation Spectrum in an Extended European Patient Cohort with Frontotemporal Dementia and Amyotrophic Lateral Sclerosis." Human Mutation 38 (3): 297–309.

[188] Zhang, Peijing, Zhenna Xiao, Shouyu Wang, Mutian Zhang, Yongkun Wei, Qinglei Hang, Jongchan Kim, et al. 2018. "ZRANB1 Is an EZH2 Deubiquitinase and a Potential Therapeutic Target in Breast Cancer." Cell Reports. https://doi.org/10.1016/j.celrep.2018.03.078

[189] Zhang, Xiaopei, Wei Wang, Weidong Zhu, Jie Dong, Yingying Cheng, Zujun Yin, and Fafu Shen. 2019. "Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels." International Journal of Molecular Sciences 20 (22).

[190] Zuk, Or, Stephen F. Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J. Daly, Benjamin M. Neale, Shamil R. Sunyaev, and Eric S. Lander. 2014. "Searching for Missing Heritability: Designing Rare Variant Association Studies." Proceedings of the National Academy of Sciences of the United States of America 111 (4): E455–64. https://doi.org/10.3390/ijms20225573

Appendix A

Supplementary Tables

TABLE A.1: Sex and Clinical Diagnosis for each individual in the DESCRIBE-FTD study

ID	SEX	CLINICAL DIAGNOSIS
1108061245	M	bvFTD
1108070583	F	SemD
143801405	M	ALSci
143801406	M	ALSci
143802250	M	ALSci
1108061214	M	bvFTD
1108061219	M	PNFA
1108061223	M	PNFA
1108061224	M	bvFTD
1108061242	M	PNFA
1108061264	F	bvFTD
1108061269	M	bvFTD
1108061277	F	ALSci
1108061865	F	bvFTD
1108061934	M	PNFA
1108061968	M	PNFA
1108061994	M	PPA

1108070338	F	bvFTD
1108070339	F	LPA
1108070355	M	ALSci
1108070361	M	bvFTD
1108070362	F	bvFTD
1108070789	F	bvFTD + ALS
1108070791	F	LPA
1108070820	F	bvFTD
1108070846	F	bvFTD
1108070853	F	SemD
1108070854	M	bvFTD
1108070861	M	SemD
1108070870	M	PNFA
1108070872	M	PNFA
1108103532	M	ALSbi
1108103578	F	bvFTD
1110272175	F	PPA
1110272198	F	bvFTD
1110273342	M	bvFTD
1110273372	M	bvFTD + ALS
1110273389	F	bvFTD + ALS
1104377680	M	PPA
1104378542	M	bvFTD
1104378567	F	PNFA
1104378580	M	PNFA
1104378592	M	bvFTD

1104378597	M	PNFA
1104378602	M	bvFTD
1104378616	M	bvFTD
1104378623	F	SemD
1104378632	F	SemD
1104378633	M	bvFTD
1104386304	M	LPA
1108060751	M	LPA
1108070693	M	ALSbi
1108070697	F	ALSbi
1108070698	M	ALSbi
1108070728	M	bvFTD
1108070777	F	PNFA
1108071463	M	PPA
1108071483	F	PNFA
1108071534	M	bvFTD
1108071537	M	bvFTD
1108071538	F	LPA
1108071539	F	PNFA
1108071542	F	bvFTD + ALS
1108072517	F	PPA
1108072533	M	bvFTD
1108072534	F	bvFTD
1108072546	M	LPA
1108072576	M	bvFTD
1108072597	M	SemD

1108072615	M	PPA
1108072618	F	bvFTD
1108072620	F	PNFA
1108072626	F	ASLci
1108072653	F	PPA
1108072655	F	ASLci
1108072661	F	SemD
1108072671	M	SemD
1108090096	F	bvFTD
1108090111	M	SemD
1108090114	M	bvFTD
1108096909	F	SemD
1108103414	M	SemD
1108103417	M	bvFTD
1108103438	F	PPA
1108103443	F	bvFTD
1108103450	M	bvFTD
1108103492	F	PPA
1108103494	F	bvFTD
1108103500	F	ALSci
Proband_26	F	bvFTD
Proband_27	F	PNFA
Proband_28	M	bvFTD + ALS
Proband_29	M	bvFTD
Proband_30	M	bvFTD
Proband-31	M	LPA

1094813058	M	bvFTD + ALS
1094818106	F	ALSbi
1094818108	F	bvFTD
1094818121	M	PPA
1094818127	M	PNFA
1094818149	M	ALSci
1094818155	M	LPA
1094818182	M	PNFA
1094818184	F	SemD
1094818187	M	PPA
1094818189	M	PNFA
1094818191	F	bvFTD
1094820054	F	bvFTD
1104378544	F	DANCER
1104378563	M	DANCER
1104378578	F	PNFA
1104378589	M	bvFTD
1104378599	M	DANCER
1104378627	M	ALS
1108060780	M	bvFTD
1108070517	M	bvFTD
1108070689	M	DANCER
1108070712	F	DANCER
1108070716	F	PNFA
1108070727	F	bvFTD
1108070740	F	PPA

1108070775	M	ALS
1108071455	F	PNFA
1108071469	F	ALSgen
1108072555	F	DANCER
1108072640	F	PNFA
1108072678	M	ALSni
1108077178	M	IPS
1108077191	M	ALSbi
1108090161	F	DANCER
1108103433	M	ALS
1108103451	M	DANCER
1108103462	F	DANCER
1108103497	M	DANCER
1110274699	F	PNFA
1110274700	F	bvFTD
1110274706	M	bvFTD
1110274723	F	PNFA
1110274724	F	SemD
1110274730	M	SemD
1110278743	M	PPA
1110279920	M	SemD
1110279922	M	ALSci
1110279932	M	bvFTD
1110279933	F	SemD
1110279934	M	bvFTD
1110279943	F	bvFTD

1110279944	M	bvFTD
1110279947	M	bvFTD
1110279958	F	SemD
1110279969	M	PPA
1110279971	F	PNFA
1110279972	F	bvFTD
1110296372	M	LPA
1110296382	F	bvFTD + ALS
1110296390	F	PNFA
1110296402	M	bvFTD
1110296404	M	bvFTD
1110296409	F	bvFTD
1110296414	M	bvFTD + ALS
1110296425	M	PPA + ALS
1110300661	M	bvFTD
1110304262	F	bvFTD + ALS
1110304281	F	bvFTD
1110304286	M	PNFA
1110304305	M	SemD
1110306330	M	SemD
1110306333	M	PNFA
1110306350	F	bvFTD
1110306353	F	ALSgen
1110306359	M	PPA
1110306368	F	LPA
1110306374	F	PNFA

1110306390	M	bvFTD
1110306398	F	bvFTD
1110306408	M	PNFA
1110306410	F	bvFTD
1110306417	M	bvFTD
1110308063	F	bvFTD
1110308088	F	PPA
1110308099	F	ALSbi
1110308101	F	LPA
1110308102	M	PPA
1110308104	M	bvFTD
1110308119	F	bvFTD
1110308121	M	PPA
1110308128	M	bvFTD
1110308137	M	PPA
1110308143	M	bvFTD
1110308381	F	bvFTD
143802321	F	ALScbi
1094818110	M	DANCER
1094818111	F	DANCER
1094818119	F	DANCER
1094818134	F	DANCER
1094818143	M	DANCER
1094818144	M	DANCER
1094818157	M	DANCER
1094818158	M	DANCER

1094818165	F	DANCER
1094818167	M	DANCER
1094818168	F	DANCER
1094819461	F	ALSci
1094819474	F	SemD
1104377711	F	DANCER
1104378566	F	DANCER
1108061572	M	DANCER
1108061574	F	DANCER
1108061580	M	DANCER
1108061581	F	DANCER
1108061582	M	DANCER
1108061583	M	DANCER
1108061585	F	DANCER
1108061596	F	DANCER
1108061600	F	DANCER
1108061605	M	DANCER
1108061609	F	DANCER
1108061618	F	DANCER
1108061629	M	DANCER
1108061630	M	DANCER
1108061645	M	DANCER
1108061649	M	DANCER
1108061652	F	DANCER
1108061656	M	DANCER
1108061658	M	DANCER

1108061701	M	PNFA
1108061724	F	LPA
1108062049	M	ALSci
1108062051	M	bvFTD+ALS
1108062056	F	PNFA
1108062065	M	bvFTD
1108062070	F	bvFTD
1108062078	F	bvFTD
1108062080	F	PPA
1108062089	M	bvFTD
1108062094	M	bvFTD
1108062099	M	PPA
1108062106	F	PNFA
1108062120	F	PNFA
1108062140	M	ALSci
1108062141	M	PPA
1108062146	M	bvFTD
1108062148	F	PNFA
1108062149	F	LPA
1108062155	F	DANCER
1108062158	F	PPA
1108062160	F	PPA
1108062169	F	PNFA
1108062170	F	bvFTD
1108062172	M	PNFA
1108062173	F	bvFTD

1108062179	M	bvFTD
1108062188	M	bvFTD
1108062191	M	bvFTD
1108062192	F	PNFA
1108062193	M	PNFA
1108062194	M	PNFA
1108062197	M	bvFTD
1108062200	F	bvFTD
1108062205	F	ALSci
1108062212	M	bvFTD
1108062215	F	SemD
1108062216	F	bvFTD
1108062220	M	bvFTD+ALS
1108062231	M	PPA
1108062233	M	PNFA
1108062238	M	ALSci
1108070779	F	DANCER
1108072652	F	DANCER
1108094924	M	DANCER
1108094941	F	DANCER
1108097705	M	bvFTD
1108097750	M	bvFTD
1108098137	F	LPA
1108098141	F	SemD
1108098151	M	bvFTD
1108098156	F	PNFA

1108098166	F	bvFTD
1108098178	M	bvFTD+ALS
1108098193	F	LPA
1108098204	M	LPA
1108098223	M	bvFTD
1108098226	F	bvFTD
1108098228	F	bvFTD
1110259399	M	DANCER
1110259418	F	DANCER
1110259424	M	DANCER
1110259474	F	DANCER
1110260543	M	ALSci
1110262264	F	bvFTD
1110262273	M	PPA
1110262283	M	bvFTD
1110262293	M	ALSci
1110262312	F	PPA
1110262320	F	PPA
1110262336	M	bvFTD
1110262346	M	bvFTD+ALS
1110262352	M	ALSci
1110262353	M	PNFA
1110268039	F	DANCER
1110270521	M	bvFTD
1110270523	F	bvFTD
1110270531	M	bvFTD

1110270533	M	bvFTD
1110270536	F	bvFTD
1110270561	M	PPA
1110270562	M	bvFTD+ALS
1110270570	M	SemD
1110270579	M	bvFTD
1110270592	F	bvFTD
1110270593	M	bvFTD
1110270601	M	PPA
1110270602	M	bvFTD
1110270603	M	bvFTD
1110270605	F	bvFTD
1110271410	M	bvFTD
1110278837	F	DANCER
1110279939	F	DANCER
1110279963	M	DANCER
1110288778	M	PPA
1110289024	M	LPA
1110289031	F	PPA
1110289044	F	bvFTD
1110289055	M	LPA
1110289057	M	LPA
1110289059	F	bvFTD
1110289068	M	bvFTD
1110289071	F	bvFTD
1110289078	F	PPA

1110289079	M	PNFA
1110289082	M	bvFTD
1110289103	M	LPA
1110290332	F	PNFA
1110293566	F	PNFA
1110293590	M	bvFTD
1110293592	M	bvFTD+ALS
1110293625	F	PPA
1110293648	F	SemD
1110294320	M	PPA
1110306338	M	DANCER
1110306358	F	DANCER
1110306361	M	DANCER
1110306382	M	DANCER
1110306385	F	DANCER
1110306393	F	DANCER
1110306403	F	DANCER
1110306418	F	DANCER
1110308297	M	SemD
1110308320	M	bvFTD
1094804234	F	DANCER
1094804267	F	DANCER
1094819535	F	DANCER
1108062060	F	LPA
1108062075	M	bvFTD
1108062553	F	bvFTD

1108098209	M	DANCER
1108103927	M	DANCER
1108103957	M	DANCER
1110260249	M	PNFA
1110260270	M	LPA
1110260273	F	PPA
1110260276	F	PPA
1110260289	F	ALSci
1110260291	M	PPA
1110260294	F	PPA
1110260302	F	ALSci
1110260305	M	ALSci
1110260308	M	PPA
1110260310	M	bvFTD+ALS
1110260312	M	bvFTD
1110260323	F	LPA
1110260332	F	PNFA
1110260336	M	ALSci
1110260337	M	ALScbi
1110260339	M	bvFTD
1110260340	F	ALScbi
1110260537	M	ALSci
1110260610	M	bvFTD
1110268226	M	DANCER
1110268241	F	DANCER
1110268242	F	DANCER

1110268253	M	DANCER
1110268255	F	DANCER
1110268266	F	DANCER
1110268522	F	DANCER
1110269088	M	LPA
1110269093	F	PNFA
1110269105	M	ALSci
1110269118	M	bvFTD
1110269120	F	PNFA
1110269127	M	bvFTD
1110269142	M	PPA
1110269161	F	ALSci
1110271019	M	PNFA
1110271041	M	DANCER
1110271060	M	DANCER
1110275659	F	PPA
1110275660	F	SemD
1110275692	M	DANCER
1110293601	M	bvFTD
1110293639	F	DANCER
1108061210	F	bvFTD
1108061237	M	bvFTD
1108062007	M	bvFTD
1108062043	M	PPA
1108082216	M	bvFTD
1108082221	F	bvFTD

1108082284	F	PNFA
1108060722	F	LPA
1108072516	M	SemD
1108072526	M	PPA
1108072574	M	PNFA
1108072582	M	LPA
1108072609	F	PNFA
1108072611	M	PPA
1108072612	M	PNFA
1108072621	F	SemD
1108072625	M	bvFTD
1108072635	M	PNFA
1108072683	M	bvFTD
1108072698	M	bvFTD+ALS
1108090099	M	PNFA
1108090106	M	bvFTD+ALS
1108090139	M	SemD
1108090155	F	bvFTD
DNA28066	M	bvFTD
DNA27830	M	bvFTD
DNA28549	F	PPA
DNA28337	F	FTD-ALS
DNA28035	F	bvFTD
DNA28539	M	bvFTD-CBS
DNA28556	F	SD-bvFTD
DNA28299	M	PNFA

DNA27716	F	FTD
DNA28214	M	bvFTD
DNA28315	F	bvFTD
DNA26624	F	bvFTD
DNA27039	M	Not Known
DNA28560	M	PNFA
DNA28562	M	PPA
DNA28576	F	PPA
DNA23480	F	bvFTD
DNA26585	M	svPPA
DNA27068	M	bvFTD
DNA27131	M	bvFTD
DNA27166	M	bvFTD
DNA27548	M	PNFA
DNA27556	F	PNFA
DNA28000	M	PNFA
DNA28596	F	PPA
DNA22915	M	lvPPA
DNA28360	M	FTD-ALS
DNA28392	M	SD
143802332	F	bvFTD+ALS
143802333	F	PPA
143802330	M	bvFTD
143802328	F	LPA
143802327	F	MCI
143802331	F	bvFTD

143802316	M	ALSbi
143802334	M	bvFTD+ALS
1108061251	F	bvFTD
143802329	F	bvFTD

TABLE A.2: lncRNA targets for ASO based perturbations in phase 1 of the non-coding RNAs study

transcript ID	gene ID	gene Name	gene Type	chr	chrom Start	chrom End	strand
ENCT00000432381.C1	ENSG00000253230	LINC00599	lincRNA	chr8	9886104	9905802	-
ENST00000514984.C1	ENSG00000196810	CTBP1-AS2	antisense	chr4	1249468	1251187	+
FTMT21500014085.C1	ENSG00000249673	NOP14-AS1	antisense	chr4	2934915	2937841	+
FTMT25900027609.C1	ENSG00000244879	GABPB1-AS1	processed_transcript	chr15	50355484	50356358	+
FTMT26800001703.C1	ENSG00000267321	RP11-1094M14.11	lincRNA	chr17	35568099	35570884	+
MICT00000210402.C1	ENSG00000227252	AC105760.2	antisense	chr2	236959770	237085774	-
ENCT00000131568.C1	ENSG00000215256	DHRS4-AS1	processed_transcript	chr14	23934047	23954171	-
HBMT00000385880.C1	ENSG00000227354	RBM26-AS1	antisense	chr13	79406290	79407590	+
MICT00000212167.C1	ENSG00000225377	RP5-1103G7.4	antisense	chr20	311124	325268	-
ENST00000540211.C1	ENSG00000176840	MIR7-3HG	lincRNA	chr19	4769132	4770184	+
ENST00000499346.C2	ENSG00000245937	CTC-228N24.3	lincRNA	chr5	127940425	128083072	-
ENST00000454115.C2	ENSG00000215447	BX322557.10	processed_transcript	chr21	45288081	45290578	+
HBMT00000161886.C1	ENSG00000254635	WAC-AS1	antisense	chr10	28512561	28532626	-
ENST00000457043.C1	ENSG00000231365	RP11-418J17.1	antisense	chr1	119140416	119142200	+
ENST00000458748.C1	ENSG00000270066	SCARNA2	lincRNA	chr1	109100198	109100612	+
ENST00000499217.T0	ENSG00000247240	UBL7-AS1	antisense	chr15	74461264	74481302	+
ENST00000553829.C1	ENSG00000272888	AC013394.2	processed_transcript	chr15	92882722	92883861	+
ENST00000398275.T0	ENSG00000214401	KANSL1-AS1	antisense	chr17	46193575	46196721	+
MICT00000129125.C1	ENSG00000260448	RP11-449H11.1	lincRNA	chr16	25067126	25107097	-
FTMT24000007081.C1	ENSG00000249456	RP11-298J20.4	sense_overlapping	chr10	124916919	124917057	+

TABLE A.3: Genes in the Illumina TruSeq Neurodegeneration Panel

APP	CSF1R	GIGYF2	ALS2
PSEN1	TRIP4	TBP	MAPT
PSEN2	TP53INP1	HTRA2	GRN
APOE	VPS35	SOD1	TMEM106B
CLU	SNCA	TARDBP	RAB38
PICALM	LRRK2	OPTN	CTSC
CR1	PRKRA	VCP	BTNL2
BIN1	GBA	FUS	TOMM40
CD33	RAB39B	PFN1	CLCN6
MS4A4A	TMEM230	SQSTM1	MARK2
MS4A6E	RAB7L1	UBQLN2	MARK4
CD2AP	GCH1	CHMP2B	EP300
EPHA1	VPS13C	ANG	AKT1
ABCA7	PARK2	NEFH	SGTA
CASS4	PINK1	TBK1	ELAVL1
CELF1	PARK7	NEK1	TOR1A
FERMT2	ATP13A2	CHCHD10	THAP1
INPP5D	PLA2G6	TUBA4A	APTX
MEF2C	FBXO7	UNC13A	ATM
NME8	SYNJ1	SARM1	PRRT2
PTK2B	DNAJC6	C21orf2	ANO3
SLC24A4	SCARB2	EPHA4	TH
RIN3	CHCHD2	LMNB1	ATP1A3
SORL1	PANK2	SPAST	DNMT1
ZCWPW1	POLG	DCTN1	ITM2B
TREM2	TAF1	FIG4	NOTCH3

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
BLSA-1556	US Caucasian	BLSA (Juan Troncosco)	male	86	control	6.8	8
BLSA-1672	US Caucasian	BLSA (Juan Troncosco)	male	84	control	7.4	17
BLSA-1805	US Caucasian	BLSA (Juan Troncosco)	male	82	control	8.1	5.5
BLSA-1838	US Caucasian	BLSA (Juan Troncosco)	female	95	control	5.9	10
BLSA-1839	US Caucasian	BLSA (Juan Troncosco)	male	88	control	5.4	14
BLSA-1889	US Caucasian	BLSA (Juan Troncosco)	male	92	control	8.7	4
BLSA-1924	US Caucasian	BLSA (Juan Troncosco)	male	83	control	7.3	6
BLSA-1961	US Caucasian	BLSA (Juan Troncosco)	female	90	control	6.1	14
BLSA-2069	US Caucasian	BLSA (Juan Troncosco)	female	92	control	7	18
JHU-705	US Caucasian	Hopkins (Juan Troncosco)	male	73	control	6.2	9
JHU-719	US Caucasian	Hopkins (Juan Troncosco)	male	66	control	8.1	10
MIAMI-2112	US Caucasian	Miami	male	42	n/a	8.7	2
SH-01-14	US Caucasian	Sun Health	female	78	lung cancer	6.3	3.33
SH-01-31	US Caucasian	Sun Health	male	81	respiratory arrest	6.6	2.75
SH-02-08	US Caucasian	Sun Health	male	95	complications of metastatic melanoma	7.1	3.5
SH-02-12	US Caucasian	Sun Health	male	92	congestive heart failure, copd	6.7	3.83

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
SH-03-15	US Caucasian	Sun Health	male	80	congestive heart failure	7.3	3.25
SH-03-17	US Caucasian	Sun Health	male	90	congestive heart failure	6.6	2.83
SH-03-50	US Caucasian	Sun Health	male	91	congestive heart failure; aspiration pneumonia tremor disorder	8.7	3.33
SH-04-08	US Caucasian	Sun Health	male	73	acute myocardial infarction stroke and chronic subdural hematoma	7.2	2.25
SH-06-05	US Caucasian	Sun Health	female	88	complications of hip fracture due to fall	7.8	4.5
SH-07-63	US Caucasian	Sun Health	female	87	pneumonia, metastatic cancer	8.7	2.5
SH-07-73	US Caucasian	Sun Health	female	76	copd epilepsy	9	4
SH-08-23	US Caucasian	Sun Health	male	85	cardiac arrythmia, chf, coronary heart disease, multiple myeloma	7.5	12.25
SH-08-44	US Caucasian	Sun Health	female	91	copd	8	2.5
SH-92-05	US Caucasian	Sun Health	male	82	lung cancer	8.6	2
SH-94-35	US Caucasian	Sun Health	female	78	lung cancer	7.7	1.25
SH-95-02	US Caucasian	Sun Health	male	70	cardiac and/or respiratory failure with intestinal bleeding	8.1	3

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
SH-95-34	US Caucasian	Sun Health	male	82	end-stage lung cancer	7.3	1.66
SH-96-08	US Caucasian	Sun Health	male	78	pancreatic cancer with intra-abdominal metastases	8.6	1.66
SH-96-13	US Caucasian	Sun Health	female	85	pancreatic cancer essential tremor, spasmodic dysphonia	6.7	2.75
SH-96-22	US Caucasian	Sun Health	male	94	pneumonia	8.1	2
SH-96-30	US Caucasian	Sun Health	male	90	cardiac and/or respiratory failure	7.3	2.16
SH-96-32	US Caucasian	Sun Health	female	85	abdominal lymphoma	8.3	1
SH-96-35	US Caucasian	Sun Health	male	84	cardiac and/or respiratory failure	8.2	2.66
SH-96-44	US Caucasian	Sun Health	female	82	cardiac and/or respiratory failure non-diagnostic alzheimer changes	7.4	1.5
SH-97-17	US Caucasian	Sun Health	male	78	cardiac and/or respiratory failure	6.5	2.66
SH-97-37	US Caucasian	Sun Health	male	83	cardiac and/or respiratory failure	8.1	3.16
SH-97-53	US Caucasian	Sun Health	male	91	cardiac and/or respiratory failure	8.2	2.66
SH-98-23	US Caucasian	Sun Health	male	68	cardiac and/or respiratory failure	7.8	2
SH-98-27	US Caucasian	Sun Health	male	63	acute intracerebral hemorrhage	7.7	1.5

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
SH-98-32	US Caucasian	Sun Health	male	83	cardiac and/or respiratory failure muscular dystrophy	8.2	3
SH-99-14	US Caucasian	Sun Health	male	86	chf, ischemic cardiomyopathy	8.4	2.5
SH-99-29	US Caucasian	Sun Health	male	81	cardiac and/or respiratory failure, emphysema, pneumothorax	8.8	3.75
SH-99-44	US Caucasian	Sun Health	male	69	hepatocellular cancer	7.5	2.16
UMARY-1027	US Caucasian	U maryland Brain Bank	male	22	multiple injury	8.1	9
UMARY-1028	US Caucasian	U maryland Brain Bank	male	39	compressional asphyxia and chest injuries	7.8	14
UMARY-1037	US Caucasian	U maryland Brain Bank	male	19	narcotic intoxication	8.7	11
UMARY-1064	US Caucasian	U maryland Brain Bank	female	40	toxic/metabolic-acute narcotic intoxication	7.5	19
UMARY-1076	US Caucasian	U maryland Brain Bank	male	17	accident, ruptured aneurysm	8.1	19
UMARY-1078	US Caucasian	U maryland Brain Bank	female	17	multiple injuries	8.4	12
UMARY-1079	US Caucasian	U maryland Brain Bank	female	19	toxic/metabolic (i.e. drug related)	8.3	16
UMARY-1104	US Caucasian	U maryland Brain Bank	male	35	multiple injuries	8.6	12
UMARY-1113	US Caucasian	U maryland Brain Bank	male	56	has cvd	7.1	17
UMARY-1158	US Caucasian	U maryland Brain Bank	male	16	cardiomegaly	8	15

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
UMARY-1170	US Caucasian	U maryland Brain Bank	male	58	cardiac arrhythmia/endocarditis	7.5	24
UMARY-1185	US Caucasian	U maryland Brain Bank	male	4	drowning	6.3	17
UMARY-1209	US Caucasian	U maryland Brain Bank	female	39	chest and abdominal injuries	6.4	17
UMARY-1226	US Caucasian	U maryland Brain Bank	male	23	drowning	8.2	21
UMARY-1260	US Caucasian	U maryland Brain Bank	male	42	multiple injuries	8.5	8
UMARY-1326	US Caucasian	U maryland Brain Bank	male	37	cardiovascular disease	7.9	12
UMARY-1406	US Caucasian	U maryland Brain Bank	female	38	cad	6.3	22
UMARY-142	US Caucasian	U maryland Brain Bank	male	16	accident, head injuries	8.5	7
UMARY-1486	US Caucasian	U maryland Brain Bank	female	22	multiple injuries	8	10
UMARY-1496	US Caucasian	U maryland Brain Bank	female	53	cardiomyopathy	7.1	19
UMARY-1535	US Caucasian	U maryland Brain Bank	male	34	abdominal injuries	6.4	16
UMARY-1540	US Caucasian	U maryland Brain Bank	male	28	multiple injuries	7.3	7
UMARY-1544	US Caucasian	U maryland Brain Bank	male	32	multiple injuries	8.2	12
UMARY-1570	US Caucasian	U maryland Brain Bank	male	48	cardiovascular disease	8	14
UMARY-1571	US Caucasian	U maryland Brain Bank	female	18	multiple injuries	9.1	8
UMARY-1573	US Caucasian	U maryland Brain Bank	female	32	multiple injury	7.1	12

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
UMARY-1578	US Caucasian	U maryland Brain Bank	male	53	cardiovascular disease	7.7	17
UMARY-1584	US Caucasian	U maryland Brain Bank	female	18	multiple injuries	8	15
UMARY-1609	US Caucasian	U maryland Brain Bank	female	33	cad	8	24
UMARY-1613	US Caucasian	U maryland Brain Bank	female	41	multiple drug intoxication	8.7	8
UMARY-164	US Caucasian	U maryland Brain Bank	male	17	head injury	8.5	16
UMARY-1652	US Caucasian	U maryland Brain Bank	male	19	narcotic intoxication	5.6	21
UMARY-1668	US Caucasian	U maryland Brain Bank	male	19	narcotic and cocaine intoxication	6.2	24
UMARY-1713	US Caucasian	U maryland Brain Bank	male	23	head and neck injuries	6.9	8
UMARY-177	US Caucasian	U maryland Brain Bank	male	16	head injury	7.1	19
UMARY-1794	US Caucasian	U maryland Brain Bank	male	21	multiple injury	7.1	17
UMARY-1795	US Caucasian	U maryland Brain Bank	female	49	cardiovascular disease	7	23
UMARY-1796	US Caucasian	U maryland Brain Bank	male	16	multiple injury	8.3	16
UMARY-1797	US Caucasian	U maryland Brain Bank	male	43	multiple injuries	7.7	18
UMARY-1825	US Caucasian	U maryland Brain Bank	male	48	cardiovascular disease	7.6	20
UMARY-1846	US Caucasian	U maryland Brain Bank	female	20	multiple injuries	7.8	9

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
UMARY-1864	US Caucasian	U maryland Brain Bank	female	2	bronchiolitis	7.3	8
UMARY-1865	US Caucasian	U maryland Brain Bank	male	55	cardiovascular disease	6.2	16
UMARY-1909	US Caucasian	U maryland Brain Bank	male	40	cardiovascular disease	7.9	20
UMARY-1936	US Caucasian	U maryland Brain Bank	male	46	cardiovascular disease	8.5	13
UMARY-288	US Caucasian	U maryland Brain Bank	male	42	multiple injury	7.2	18
UMARY-4263	US Caucasian	U maryland Brain Bank	male	61	cardiac arrest	8.5	6
UMARY-4540	US Caucasian	U maryland Brain Bank	male	25	multiple injuries	8.2	23
UMARY-4598	US Caucasian	U maryland Brain Bank	male	45	dilated cardiomyopathy	7.4	6
UMARY-4638	US Caucasian	U maryland Brain Bank	female	15	chest injury	8.8	5
UMARY-4640	US Caucasian	U maryland Brain Bank	female	47	pneumonia	7.2	5
UMARY-4724	US Caucasian	U maryland Brain Bank	female	16	multiple injury	7.3	15
UMARY-4726	US Caucasian	U maryland Brain Bank	male	28	multiple injuries	8.1	6
UMARY-4729	US Caucasian	U maryland Brain Bank	male	24	multiple injuries	7.4	10
UMARY-4781	US Caucasian	U maryland Brain Bank	male	45	as cvd	8.2	17
UMARY-4782	US Caucasian	U maryland Brain Bank	male	18	head and chest injuries	7	17
UMARY-4789	US Caucasian	U maryland Brain Bank	female	72	accident, exsanguination	6.9	19

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
UMARY-4841	US Caucasian	U maryland Brain Bank	female	42	accident, multiple injuries	8.6	17
UMARY-4842	US Caucasian	U maryland Brain Bank	male	47	airway obstruction of food bolus	7.6	12
UMARY-4848	US Caucasian	U maryland Brain Bank	male	16	drowning	8	15
UMARY-4903	US Caucasian	U maryland Brain Bank	female	31	quetiapine/alcohol intoxication	8.2	5
UMARY-4915	US Caucasian	U maryland Brain Bank	male	49	cardiovascular disease	8.2	5
UMARY-5028	US Caucasian	U maryland Brain Bank	male	67	multiple injuries	7.8	18
UMARY-5078	US Caucasian	U maryland Brain Bank	male	48	neck injuries/alcohol use	8.3	23
UMARY-5079	US Caucasian	U maryland Brain Bank	male	33	drowning complicated by alcohol intoxication	8.1	16
UMARY-5081	US Caucasian	U maryland Brain Bank	male	48	pulmonary embolism	8.1	20
UMARY-602	US Caucasian	U maryland Brain Bank	male	27	accident, multiple injuries	6.4	15
UMARY-604	US Caucasian	U maryland Brain Bank	male	43	cardiovascular disease	7.6	15
UMARY-818	US Caucasian	U maryland Brain Bank	male	27	multiple injury	7.8	10
UMARY-819	US Caucasian	U maryland Brain Bank	male	18	chest injury	7.5	28
UMARY-871	US Caucasian	U maryland Brain Bank	male	42	toxic/metabolic (i.e. drug related)	7.9	19

TABLE A.4: Clinical characteristics of 119 neurologically healthy individuals including ethnicity, sex, cause of death, age at death, RNA integrity number, post-mortem interval and brain bank origin.

NIH-ID	Ethnicity	Institution	Sex	Age	Diagnosis	RIN	PMI
UMARY-879	US Caucasian	U maryland Brain Bank	male	21	multiple injuries	7.8	13
UMARY-880	US Caucasian	U maryland Brain Bank	female	48	cardiovascular disease	8	12
UMARY-933	US Caucasian	U maryland Brain Bank	male	20	lightning strike	7.5	12

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
AC105760.2	MLPH	-0.259162302
	AC105760.3	0.976248665
	ACKR3	0.733528511
	COL6A3	-0.495729674
	COPS8	0.694589046
	AC112721.1	-0.107759587
	AC112721.2	0.069244192
	AC112715.2	0.632123428
	AC105760.2	1
	AC011286.1	-0.366453195
CTBP1-AS2	TACC3	-0.97677351
	MAEA	0.82571035
	IDUA	0.77820788
	FGFRL1	-0.93208586
	TMEM175	0.94343933
	DGKQ	0.93783160
	SLC26A1	0.76223366
	SPON2	0.43867903
	CTBP1	0.88872396
	UVSSA	0.30942468
	SLBP	-0.96075757
	CPLX1	0.95044029
	FAM53A	-0.77907227

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	RNF212	0.66450322
	GAK	0.54413812
	CRIPAK	0.62357935
	CTBP1-AS2	1
	Y_RNA	0.05080485
	AC016773.1	-0.89000181
	NKX1-1	0.09359980
	RP11-1398P2.1	-0.82960552
	RP11-440L14.1	-0.18395445
	RP11-460I19.2	-0.40139976
	RP11-20I20.2	-0.65274337
	RP11-572O17.1	-0.79807969
	RP11-20I20.4	-0.18710818
CTC-228N24.3	SLC12A2	-0.10080643
	PRRC1	0.21498084
	CTXN3	-0.13729584
	CTC-228N24.1	0.27064149
	CTC-228N24.3	1
	KDELC1P1	-0.14017047
	HNRNPKP1	0.87492285
GABPB1-AS1	TRPM7	-0.546249920
	GABPB1	-0.694097039

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	USP8	-0.445177281
	SPPL2A	-0.472254883
	SLC27A2	-0.362387502
	HDC	-0.522962755
	USP50	0.028408457
	RN7SL494P	-0.187566301
	GABPB1-AS1	1
	RNA5SP395	0.291729333
	RP11-120K9.2	0.291828535
	RP11-507J18.2	0.633254541
	MIR4712	0.676245358
KANSL1-AS1	KANSL1	0.93001545
	SPPL2C	-0.29246136
	ARL17A	-0.08593673
	MAPT	-0.40552355
	KANSL1-AS1	1
	Y_RNA	-0.11594160
	RP11-259G18.1	0.35727720
	RP11-669E14.6	0.20685734
	RP11-259G18.2	0.56213911
	RP11-259G18.3	0.86435548
	MAPT-AS1	-0.35697465
	RP11-293E1.1	-0.57374141

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	RP11-995C19.2	0.56781560
BX322557.10	ITGB2	-0.60488529
	FAM207A	-0.43181227
	SLC19A1	-0.43409850
	LINC00334	0.25670138
	COL18A1	-0.79800141
	C21orf67	0.30350393
	PTTG1IP	-0.79312115
	COL18A1-AS1	-0.16227332
	LINC00315	0.73953910
	SUMO3	-0.50956694
	POFUT2	0.10737579
	ADARB1	0.96396706
	BX322557.10	1
	LINC00205	0.91564365
	COL18A1-AS2	-0.69033642
	LINC00162	0.17374937
	ITGB2-AS1	-0.24061099
	AL133493.2	-0.14518831
	LINC00163	0.48311429
	SSR4P1	0.40090242
	AL773604.8	-0.86414331
	LINC00316	0.17062443

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	PRED57	0.75454634
	AP001579.1	-0.18286411
	LL21NC02-1C16.2	-0.43679911
MIR7-3HG	MPND	0.822900500
	UHRF1	-0.678646953
	C19orf10	-0.582033339
	SHD	0.940306547
	FSD1	0.574735638
	PLIN3	-0.674710967
	KDM4B	0.771010335
	TICAM1	-0.008665925
	FEM1A	0.654320984
	SH3GL1	-0.539040741
	DPP9	-0.345063053
	TMIGD2	-0.260115170
	CHAF1A	-0.680092154
	UBXN6	0.651992981
	HDGFRP2	-0.490549935
	PLIN4	-0.541003616
	SEMA6B	0.814507047
	LRG1	0.070240761
	MIR7-3HG	1
	STAP2	-0.618004084

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	TNFAIP8L1	-0.611366719
	ARRDC5	-0.428015322
	AC005594.3	-0.624472173
	PLIN5	-0.398852434
	RN7SL121P	0.271402803
	MIR4746	-0.570215789
	CTB-50L17.16	0.800643345
	CTB-50L17.7	-0.649260675
	CTB-50L17.5	-0.716653092
	CTB-50L17.14	0.529841231
	CTB-50L17.2	0.532734338
	CTC-518P12.6	0.597386467
	CTC-482H14.5	0.464586106
	CTB-50L17.9	0.093847903
	AC007292.6	0.015173373
	AC005523.3	-0.347859331
	AC007292.7	0.762933053
NOP14-AS1	RNF4	-0.785177137
	SH3BP2	0.677639406
	NOP14	-0.741262294
	ADD1	0.864596247
	MFSD10	0.654727353
	HGFAC	0.678399241

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	FAM193A	0.857552501
	GRK4	0.772744861
	RGS12	0.661888450
	TNIP2	-0.842290176
	MSANTD1	0.563846303
	HTT	0.604181029
	RNU6-204P	0.849475577
	NOP14-AS1	1
	HTT-AS	0.748285930
RBM26-AS1	NDFIP2	-0.19179700
	RBM26	0.60665191
	RBM26-AS1	1
	LINC01068	0.62740210
	LINC00382	-0.29305918
	CCT5P2	0.18277709
	NDFIP2-AS1	-0.41824845
	NIPA2P5	0.18194962
	RNA5SP33	-0.11119175
RP11-298J20.4	ZRANB1	0.905133306
	LHPP	0.548831551
	FAM175B	0.006253551
	CTBP2	-0.150617862

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	FAM53B	-0.110807600
	METTL10	-0.737092584
	MRPS21P6	0.635922363
	NPM1P31	-0.629486134
	RPS10P18	0.307748802
	NKX1-2	0.334737819
	RP11-464O2.2	-0.255088086
	RP11-298J20.4	1
	RP11-12J10.3	0.008867627
	MIR4296	0.617877437
	RP13-238F13.5	0.192614321
RP11-418J17.1	TBX15	0.027012972
	PHGDH	0.861691442
	WARS2	0.758896298
	HAO2	0.643692341
	ZNF697	-0.796422725
	HSD3BP4	0.107293578
	HSD3B1	-0.884411835
	HSD3B2	0.250614594
	RBMX2P3	0.122394608
	WARS2-IT1	0.587933712
	GAPDHP32	-0.511722360
	RP11-418J17.2	0.422131778

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	RP11-418J17.3	0.672782327
	GAPDHP58	-0.432750882
	RP11-418J17.1	1
	RP5-834N19.1	0.746114278
	RP5-871G17.5	-0.846753430
	RPS3AP12	0.597247655
	RP4-712E4.2	-0.296025748
	GAPDHP33	0.007699265
RP11-449H11.1	TNRC6A	0.617632276
	AQP8	0.588593257
	ARHGAP17	-0.184172343
	ZKSCAN2	0.890674193
	SLC5A11	-0.007255253
	LCMT1	0.947180487
	AC012317.1	0.851108297
	AC008731.1	0.097265129
	CTD-2540M10.1	0.258727019
	RP11-266L9.2	0.531161525
	RP11-449H11.1	1
	RP11-266L9.1	0.518468439
	RP11-266L9.5	0.866253244
	RP11-266L9.4	0.944214914
	RP11-266L9.3	0.816711721

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
RP5-1103G7.4	TRIB3	0.862904195
	CSNK2A1	-0.632975457
	RBCK1	-0.715385693
	NRSN2	0.718540526
	TBC1D20	0.574165755
	TCF15	0.268999506
	DEFB129	0.176244926
	SOX12	0.755152288
	ZCCHC3	-0.851299762
	DEFB128	0.170709179
	C20orf96	-0.021005752
	RP5-1103G7.4	1
	RP5-1103G7.10	0.454070254
SCARNA2	SARS	-0.359686333
	WDR47	-0.681299841
	STXBP3	0.623226628
	KIAA1324	-0.587573980
	CLCC1	0.794317972
	GPSM2	0.804089205
	PRPF38B	0.542137389
	PSRC1	0.716770933
	SORT1	-0.627421671

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	SYPL2	-0.712106317
	PSMA5	0.901305252
	FNDC7	0.190317509
	CELSR2	-0.134052559
	GPR61	-0.705871504
	HENMT1	-0.361946307
	AKNAD1	0.516978538
	ATXN7L2	-0.607183236
	CYB561D1	-0.242718702
	C1orf194	0.912766847
	AMIGO1	-0.710222647
	TAF13	-0.304836810
	SPATA42	0.663108053
	RNU6V	-0.296964989
	TMEM167B	-0.073843455
	MYBPHL	-0.660270585
	RP5-1160K1.3	-0.114239101
	RANP5	0.203214966
	RP11-20O24.4	0.749199090
	SCARNA2	1
	RP5-1065J22.8	0.522957713
UBL7-AS1	STOML1	-0.86758350
	CSK	-0.68716247

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	ISLR	-0.08641940
	STRA6	0.53902040
	SEMA7A	-0.95721007
	UBL7	-0.63088147
	CYP11A1	-0.31309370
	PML	-0.38663937
	CYP1A1	-0.84561346
	ULK3	-0.96020919
	CCDC33	-0.28294778
	SCAMP2	0.94051358
	CYP1A2	0.15251514
	LMAN1L	-0.35482357
	ISLR2	-0.40170752
	RPP25	0.64773543
	COX5A	-0.50501277
	FAM219B	-0.94106326
	MPI	0.13616452
	EDC3	0.55735767
	CLK3	0.59258610
	ARID3B	-0.66779010
	CPLX3	-0.54330067
	UBL7-AS1	1
	RP11-247C2.2	-0.31932202
	RP11-60L3.1	0.64590751

TABLE A.6: Correlation of gene expression of lncRNAs selected in Phase 1 with their cis-genes.

TARGET lncRNA	GENE NAME	PEARSON CORRELATION (r)
	RP11-10O17.1	-0.06269609
	CTD-2311M21.2	-0.05321472
	CTD-2235H24.2	0.56471408
	CTD-3154N5.1	-0.88307722
	RP11-665J16.1	-0.13481345
	RP11-414J4.2	0.02760883
	RP11-10O17.3	0.95179780
	CTD-2311M21.3	0.30206826
	MIR4513	-0.02956941
WAC-AS1	BAMBI	0.4877750562
	WAC	-0.9054852004
	MPP7	0.7163033389
	C10orf126	-0.3582700848
	RP11-351M16.1	-0.3926570675
	RP11-492M23.2	-0.3722339956
	LINC00837	0.6390271109
	RP11-478H13.1	-0.3619376593
	WAC-AS1	1

TABLE A.5: KEGG pathways enriched using DEGs from CTBP1-AS2 knocked down samples versus untreated controls on day 8 of differentiation into cortical neurons

KEGG ID	Pathway	N	DE	P.DE
path:hsa03010	Ribosome	153	76	4,01E-05
path:hsa04714	Thermogenesis	231	92	4,92E-01
path:hsa05016	Huntington disease	193	78	2,53E+01
path:hsa00190	Oxidative phosphorylation	133	58	7,74E+02
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	149	61	5,67E+03
path:hsa05012	Parkinson disease	142	57	6,02E+04
path:hsa05010	Alzheimer disease	171	63	4,54E+05
path:hsa05168	Herpes simplex virus 1 infection	491	126	1,41E+09
path:hsa01100	Metabolic pathways	1487	327	2,25E+09
path:hsa00510	N-Glycan biosynthesis	50	21	7,65E+09
path:hsa04110	Cell cycle	124	39	0.000233089108581167
path:hsa03030	DNA replication	36	16	0.00024787641721035
path:hsa04260	Cardiac muscle contraction	86	29	0.000397619201743428
path:hsa03420	Nucleotide excision repair	47	18	0.000933640580601728
path:hsa03018	RNA degradation	79	26	0.00117334134723666
path:hsa04723	Retrograde endocannabinoid signaling	148	42	0.00146521952745896
path:hsa03460	Fanconi anemia pathway	54	19	0.00220206990987292
path:hsa00513	Various types of N-glycan biosynthesis	39	15	0.00232547031954342
path:hsa04210	Apoptosis	136	38	0.00323723037976521
path:hsa05223	Non-small cell lung cancer	66	21	0.00526410504879415

Appendix B

Supplementary Figures

FIGURE B.1: MDS Plot showing clustering of 639 ROSMAP based on the batch

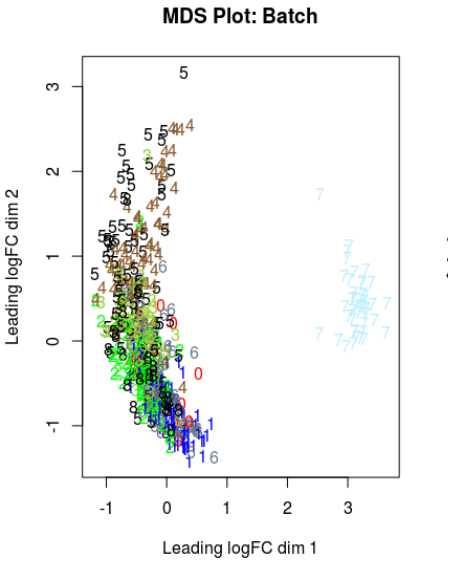


FIGURE B.2: MDS Plot showing clustering of filtered 532 samples based on sex (0 = Females; 1 = Males)

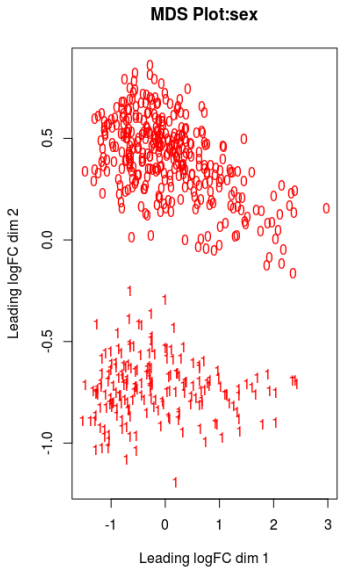


FIGURE B.3: KANSL1-AS1 Hi-C map showing chromatin interaction between the KANSL1-AS1 gene and the KANSL1 gene obtained from the 3-D Genome Browser (Wang et al. 2018).

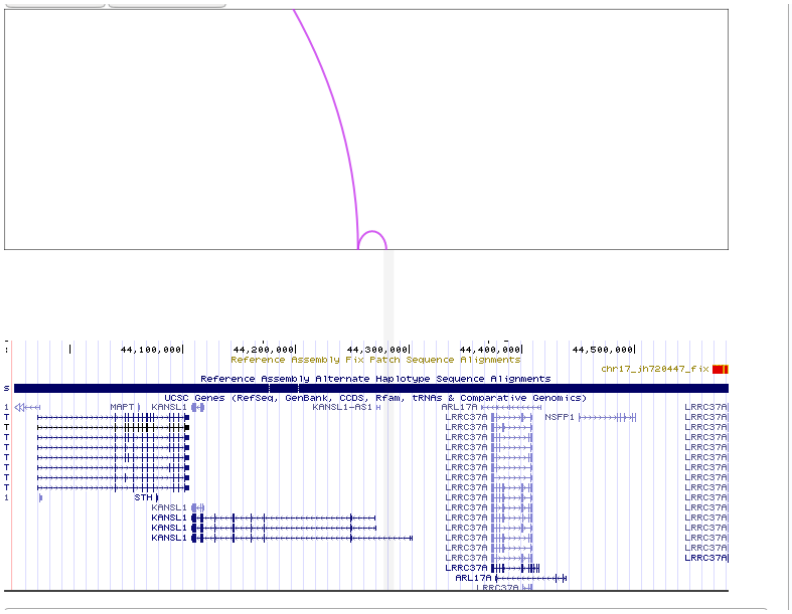
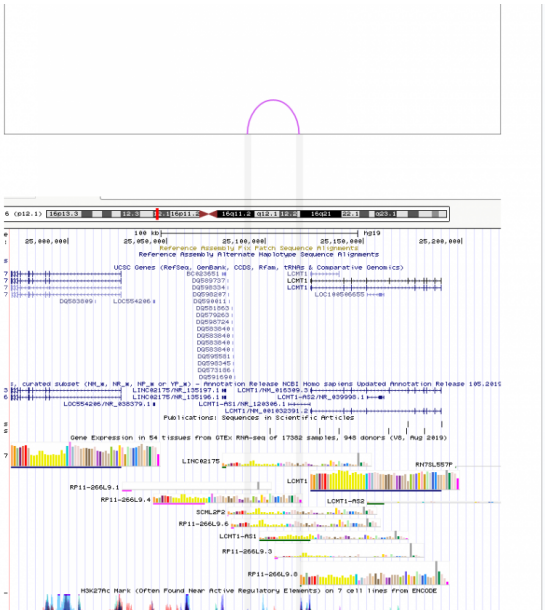


FIGURE B.4: LCMT1-AS1 Hi-C map showing chromatin interaction between the LCMT1-AS1 gene and the LCMT1 gene obtained from the 3-D Genome Browser (Wang et al. 2018).



Appendix C

Supplementary Text

C.1 Differential Gene Expression Analysis for the Pilot ASO-based lncRNA expression perturbation study

During the experimental phase, four of the selected lncRNAs - LINC00599, DHRS4-AS1, AC013394.2 and RP11-1094M14.1 - showed inadequate knockdown (KD) efficiency or extremely low yield for all the tested ASOs and hence were removed from further analysis. Here, we present individually the results of each of the remaining 16 lncRNAs.

1. CTBP1-AS2: This lncRNA had 3 ASOs that achieved above threshold KD efficiency and was thus selected for sequencing and further analysis. Although, from the CAGE-Sequencing it was observed that the lncRNA is 0 expressed in all scramble controls as well as untreated samples except for the untreated sample from day 8 of the ND41865 cell line. Interestingly, the KEGG pathways that were enriched from using the DEGs from day 8 between KDs and the untreated controls included Huntington disease, oxidative phosphorylation, Parkinson disease, Alzheimer's disease as well as apoptosis and nucleotide excision repair with p-value < 0.005 [Supplementary Table A.5]. Interestingly, nucleotide excision repair (KEGG PATH:HSA03420) was also enriched in the DEGs from day 3 of differentiation with p-value = 1.9 x 10-9.

2. NOP14-AS1: For the ND41865 cell line, NOP14-AS1 ASO based knockdowns showed high KD efficiency and were seen as highly expressed in the scramble and untreated control cell lines, as well. Gene ontology analysis performed on DEGs from days 3 and 8 of differentiation yielded some common biological processes with p-value < 0.005: nervous system development (GO:0007399),

neurogenesis (GO:0022008), generation of neurons (GO:0048699) and central nervous system development (GO:0007417). Additionally, the Human Protein Atlas (Thul and Lindskog 2018) shows that the NOP14-AS1 gene is highly expressed in the brain as compared to other tissues.

3. GABPB1-AS1: For GABPB1-AS1, expression of the gene was seen in scramble and untreated controls for both cell lines on days 3 and 8 of differentiation. Two ASOs were selected in ND41865 and three in GM23280 based on KD efficiency. Since out of the two ASOs selected in ND41865, one showed 0% KD efficiency on day 8 of differentiation in ND41865, we present results only from the CAGE-Sequencing data from the GM23280 cell line. The top GO term enriched in DEGs from day 3 and day 8 of differentiation was nervous system development (GO:0007399). In addition, the DEGs were also enriched for neurogenesis (GO:0022008), neuron development (GO: 0048666), synaptic signalling (GO:0099536), brain development (GO:0007420), and neuronal projection development/morphogenesis(GO:0031175/0048812).

4. AC105760.2: This lncRNA gene was only knocked down in the ND41865 cell line and two ASOs passed the required KD efficiency threshold. On checking the expression of the AC105760.2 gene in the scramble and untreated controls in the ND41865 cell line, 0 expression was seen on both days 3 and 8 of differentiation due to the low depth of sequencing. DEGs from days 3 and 8 of differentiation were enriched for apoptosis (KEGG PATH:HSA04210) with p-values 0.006 and 0.0004 respectively.

5. RBM26-AS1: For this lncRNA, ASO based knockdowns were performed using 3 ASOs in the ND41865 cell line whereas for the GM23280 cell line, ASOs were toxic on day 8 of differentiation and were removed from further analysis. Presented here are the results from the CAGE-Sequencing data from the ND41865 cell line. The CAGE-Sequencing data for the scramble control sample for day 3 of differentiation as well as all knockdown samples for all ASOs on both days 3 and 8 of differentiation showed 0 expression of the RBM26-AS1 gene. DEGs from days 3 and 8 of differentiation were enriched for axon guidance (KEGG PATH:HSA04360) with p-values 0.001 and 3.4 x 10-8 respectively.

6. RP5-1103G7.4: ASO based knockdown experiments were performed successfully in both ND41865 and GM2320 cell lines for this lncRNA gene. We present

results from both these cell lines - analysed separately - here. One GO term that was enriched in DEGs from both cell lines was cellular response to DNA damage stimulus (GO: 0006974). In addition, in the ND41865 cell line, the DEGs from both days 3 and 8 were also enriched for signal transduction by p53 class mediator (GO:0072331). The p53 signalling pathway (KEGG PATH:HSA04115) was also enriched in DEGs from both cell lines.

7. MIR7-3HG: Knockdowns for this gene were only performed in the ND41865 cell line as the KD efficiency in GM23230 was extremely low. DEGs from both days 3 and 8 of differentiation in the ND41865 cell line were enriched for the KEGG pathway term apoptosis (KEGG PATH:HSA04210). No additional remarkable findings were seen. The CAGE-Sequencing data showed 0 expression of the MIR7-3HG gene in the scramble control A (CA) on day 3 of differentiation.

8. CTC-228N24.3: For both cell lines, adequate KD efficiency was obtained for 3 ASOs on both days of differentiation, 3 and 8. Two commonly significantly enriched KEGG pathways in DEGs from both cell lines and time points are axon guidance (KEGG PATH:HSA04360) and p53 signalling pathway (KEGG PATH:HSA04115).

9. BX322557.10: For both cell lines, adequate KD efficiency was obtained for 3 ASOs on both days of differentiation, 3 and 8. Interestingly, in the DEGs obtained from the GM23280 cell line, the most highly enriched (p-value < 10-5) GO terms were nervous system development (GO:0007299), central nervous system development (GO:0007417), neuron differentiation (GO:0030182), neuron development (GO:0048666), neurogenesis (GO:0022008), head development (GO:0060322), brain development (GO:0007420), generation of neurons (GO:0048699), neuron projection morphogenesis (GO:0048812), neuron projection development (GO:0031175), synapse assembly (GO:0007416), axon development (GO:0061564), forebrain development (GO:0030900), regulation of synapse assembly (GO:0051963), telencephalon development (GO:0021537), axonogenesis (GO:0007409), synaptic signalling (GO:0099536), regulation of neuron differentiation (GO:0045664) and developmental process (GO:0032502). However, these findings were not replicated in the KD data from the ND41865 cell line. An interesting KEGG pathway that was enriched in the DEGs from

the ND41865 KD data was mismatch repair (KEGG PATH:HSA03430). However, this finding was not replicated in the GM23280 cell line.

10. WAC-AS1: For this lncRNA gene, only 1 ASO was selected as the others resulted in extremely poor KD efficiency or were toxic. DGE analysis yielded a list of 2891 and 4545 DEGs for days 3 and 8 respectively in the ND41865 cell line, but without technical and biological replicates, it is difficult to infer the significance of these data.

11. RP11-418J17.1: For both cell lines, adequate KD efficiency was obtained for 3 ASOs on both days of differentiation, 3 and 8. DEGs in both cell lines and for both days of differentiation 3, and 8 had one common pathway term enriched, axon guidance (KEGG PATH:HSA04360) and one GO term, nervous system development (GP:0007399). In addition, several other neuron and synaptic junction related terms were also enriched for amongst the DEG as was seen in the DGE results from the BX322557.10 lncRNA gene.

12. SCARNA2: For both cell lines, adequate KD efficiency was obtained for 3 ASOs on both days of differentiation, 3 and 8. Similar to the results seen in KD data from RP11-418J17.1 and BX322557.10, DEGs from both cell lines as well as days of differentiation 3 and 8 were enriched for several neuronal development GO terms and pathways: axon guidance (KEGG PATH:HSA04360), nervous system development (GP:0007399), synapse (GO:0045202), presynapse (GO:0098793), axon development (GO:0061564), neurogenesis (GO:0022008), and other related terms.

13. UBL7-AS1: For both cell lines, adequate KD efficiency and RNA yield was obtained for 3 ASOs on each day of differentiation, days 3 and 8. The DEGs from both cell lines and time points were enriched for the KEGG pathway term axon guidance (KEGG PATH:HSA04360). DEGs from day 3 of differentiation in the GM23280 cell line were also enriched for nervous system development (GP:0007399), neuron development (GO:0048666), neuron differentiation (GO:0030182), neurogenesis (GO:0022008), trans-synaptic signalling (GO:0099537), generation of neurons (GO:0048699), neuron projection morphogenesis (GO:0048858), axon development (GO:0061564), nerve development (GO:0021675) and central nervous development (GO:0007417). These findings however are not replicated in the ND41865 cell line, where the DEGs

are majorly enriched for ncRNA processing (GO:0034470) and rRNA processing (GO:00106072). The CAGE-Sequencing data showed 0 expression for the UBL7-AS1 gene in scramble and untreated controls on day 8 of differentiation in the GM23280 cell line.

14. KANSL1-AS1: Two ASOs were selected for knockdowns in both cell lines for the KANSL1-AS1 gene, one of these ASOs gave only 30% KD efficiency on day 3 of differentiation in the ND41865 cell line but the KD sufficiency was adequate on day 8 as well as on both days of differentiation in the GM23280 cell line. From the DEG lists obtained from the ND41865 cell lines, we found enrichment of several neurodegeneration and related pathways: Oxidative phosphorylation (KEGG PATH:HSA00190), Parkinson disease (KEGG PATH:HSA05012), Ubiquitin mediated proteolysis (KEGG PATH:HSA04120), Huntington Disease (KEGG PATH:HSA05016), Autophagy (KEGG PATH:HSA04140) and Amyotrophic Lateral Sclerosis (KEGG PATH:HSA05014). These findings however were not replicated in the DEGs obtained from the GM23280 cell line data.

15. RP11-449H11.1: Only one ASO was selected for this target lncRNA gene as the others were either toxic or resulted in extremely low KD efficiency. The CAGE-Sequencing data showed 0 expression for scramble as well as untreated controls for all time points in the GM23280 cell line. The DEGs obtained between KD samples and scramble controls showed an enrichment for axon guidance (KEGG PATH:HSA04360) and calcium signalling pathway (KEGG PATH:HSA04020) but without technical and biological replication, it is difficult to ascertain the significance of these results.

16. RP11-298J20.4: This lncRNA gene was only knocked down in the ND41865 cell line using 2 ASOs, as in the GM23280 cell line low KD efficiency and toxicity was obtained. DEGs from day 3 and 8 were enriched for three common KEGG pathways: base excision repair (KEGG PATH:HSA03410), nucleotide excision repair (KEGG PATH:HSA03420) and axon guidance (KEGG PATH:HSA04360).