Computational methods and analyses to dissect the pathogenesis of Frontotemporal Dementia

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät

und

der Medizinischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Kevin Menden

aus Schwäbisch-Gmünd, Deutschland

November 2020

Tag der mündlichen Prüfung: 08.04.2021

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. Stefan Bonn
Prüfungskomission:	Prof. Dr. Peter Heutink
	Prof. Dr. Stefan Bonn
	Prof. Dr. Kay Nieselt
	Prof. Dr. Thomas Gasser

Erklärung / Declaration

Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel: „Computational methods and analyses to dissect the pathogenesis of Frontotemporal Dementia“ 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 „Computational methods and analyses to dissect the pathogenesis of Frontotemporal Dementia”, 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, den Unterschrift/Signature

Contributions

Chapter 2

The project described in chapter 2 was initiated by me together with Peter Heutink and Stefan Bonn. I had the initial idea for the project, designed and implemented the algorithm, performed testing of the algorithm and processed and analyzed all public datasets necessary. Together with Stefan Bonn I wrote the majority of the manuscript. Stefan Bonn furthermore helped in algorithm and general project design. Mohamed Marouf and Sergio Oller helped with model selection and creating the python package. Anupriya Dalmia helped with processing of the pancreas datasets. Katrin Kloiber and Peter Heutink helped with manuscript writing. Daniel Sumner Magruder and Sergio Oller designed and implemented the web application.

Chapter 3

The RiMod-FTD project, which is the basis for chapter 3, was initiated by Peter Heutink. This chapter builds on various datasets generated in experiments that were designed and performed by Ashutosh Dhingra, Melissa Castillo Lizardo, Noémia Rita Fernandes, Eldem Sadikoglou, Salvador Rodriguez Nieto and Patrizia Rizzu. Part of the small RNA-seq experiments was performed by the group of André Fischer at the DZNE Göttingen, by Cemil Kerimoglu and Lalit Kaurani. The experiments generating the iPSC-derived microglia were performed by Deborah Kronenberg-Versteeg and Ashutosh Dhingra. Tenzin Nyima and Margheritta Francescatto furthermore helped with processing of the CAGE-seq data. Pathology scoring was done by Manuela Neumann. All computational analyses that led to the generation of the results presented in this chapter were generated by me. Peter Heutink and Patrizia Rizzu helped with manuscript writing.

Chapter 4

This chapter builds on partly the same data as used in chapter 3 (see contributions Chapter 3). This chapter was initiated by me and I furthermore performed all data analyses, analysis interpretations and wrote the manuscript/chapter.

Für Benita
Für meine Eltern

Abstract

Frontotemporal Dementia (FTD) is a devastating neurodegenerative disorder that typically manifests before the age of 65 and is characterized by progressive degeneration of frontal and temporal lobes as well as behavioural changes and problems with speech. Although great advancements in our understanding of FTD have been made during the last decades, there still does not exist a treatment that halts the progression of this disease. It is therefore necessary to further advance our understanding of the molecular mechanisms that cause FTD and that drive disease progression forward. In this thesis, I have contributed to the field of FTD research through the development of computational methods and the analysis of multi-omics datasets in order the develop new hypotheses for disease mechanisms in FTD.

Studying complex tissues such as the human brain requires to carefully consider the contributions of diverse components of such systems. A major factor in transcriptomic experiments is cell type composition, as every cell has a unique transcriptional profile. In chapter 2, we have developed a deep learning-based cell type deconvolution algorithm that outperforms other methods and, importantly, also works well on post-mortem human brain tissue. The algorithm was rigorously tested and made available to the community as an open source, accessible python package and as web application.

In chapter 3, we have analysed multi-omics datasets from post-mortem human brain tissue of FTD patients with mutations in the genes GRN, MAPT and C9orf72. Using an integrative data analysis approach, we could identify common and distinct affected pathways in these three genetic FTD subtypes. We leveraged the rich multi-omics datasets to identify new aspects of the disease, such as vulnerable neurons and increasing blood vessel percentages. In-depth analysis could highlight several regulator molecules, such as micro RNAs and transcription factors, that are likely to play important roles in FTD and therefore depict promising subjects for future studies.

In chapter 4, we have performed co-expression module analysis of transcriptional data from seven different brain regions of patients with genetic subtypes of FTD. Using this comprehensive dataset, we have highlighted regions that are transcriptionally affected in different FTD subtypes and regions that seem not to suffer from the disease. We furthermore highlighted region- and disease-specific co-expression modules and pinpointed hub genes of potentially important function for these modules. Our analysis is the first that evaluates transcriptional deregulation at such diversity in FTD, and therefore provides

valuable novel insights for the field of FTD.

Kurzfassung

Die Frontotemporale Demenz (FTD) ist eine verheerende neurodegenerative Erkrankung, die sich typischerweise vor dem 65. Lebensjahr manifestiert und durch eine fortschreitende Degeneration der Stirn- und Schläfenlappen sowie durch Verhaltensänderungen und Sprachprobleme gekennzeichnet ist. Obwohl in den letzten Jahrzehnten große Fortschritte in unserem Verständnis von FTD gemacht wurden, gibt es immer noch keine Behandlung, die das Fortschreiten dieser Krankheit aufhält. Es ist daher notwendig, unser Verständnis der molekularen Mechanismen, die FTD verursachen und das Fortschreiten der Krankheit vorantreiben, weiter zu vertiefen. In dieser Arbeit habe ich durch Methodenentwicklung and Datenanalyse neue Hypothesen für potenzielle Krankheismechanismen bei FTD aufgestellt, und so zur FTD Forschung beigetragen.

Die Untersuchung komplexer Gewebe wie des menschlichen Gehirns erfordert eine sorgfältige Abwägung der Beiträge der verschiedenen Komponenten solcher Systeme. Ein wichtiger Faktor bei transkriptomischen Experimenten ist die Zusammensetzung des Gewebes aus verschiedenen Zelltypen, da jeder Zelltyp ein einzigartiges Transkriptionsprofil aufweist. In Kapitel 2 haben wir einen auf maschinellem Lernen basierenden Algorithmus entwickelt, welcher die Zelltypenzusammensetzung eines Gewebes anhand transkriptomischer Daten vorhersagt. Unser Algorithmus übertrifft andere, moderne Algorithmen und liefert insbesondere auch bei postmortalem menschlichem Hirngewebe gute Ergebnisse. Der Algorithmus wurde rigoros getestet und als open-source Software anderen Forschern zur Verfügung gestellt.

In Kapitel 3 haben wir multimodale Datensätze aus postmortalem Humanen Hirngewebe von FTD-Patienten mit Mutationen in den Genen GRN, MAPT und C9orf72 analysiert. Mit einem integrativen Ansatz der Datenanalyse konnten wir Gemeinsamkeiten und Unterschiede der verschiedenen genetischen FTD-subtypen identifizieren. So konnten wir zum Beispiel besonders anfällige Neuronenarten identifizieren und einen generellen Anstieg der Blutgefäßdichte ausmachen. Außerdem konnten Gene und Moleküle identifiziert werden, welche mit hoher Wahrscheinlichkeit eine regulatorische Rolle im Verlauf der Krankheit spielen und daher von großem Interesse für zukünftige FTD Studien sein könnten.

Im vierten Kapitel haben wir mithilfe von Genexpressionsdaten aus sieben verschiedenen Hinregiongen von Patienten mit genetischen FTD Subtypen eine Ko-expressionsanalyse durchgeführt und konnten so neue Erkenntnisse über die Anfälligkeit verschiedener Ge-

hirnregionen gewinnen. So konnten wir Gehirnregionen identifizieren, welche bei bestimmten Subgruppen besonders betroffen sind, sowie Regionen, welche scheinbar von FTD verschont bleiben. Durch tiefergehende Analysen konnten wir Gen-Module und potentiell wichtige Gene mit zentralen, regulatorischen Funktionen identifizieren. Unsere Studie ist die erste Analyse, welche FTD in solch großer Diversität, mit sieben Gehirnregionen und drei verschiedenen FTD-Subtypen, untersucht und liefert daher wichtige neue Erkenntnisse für das Feld der FTD Forschung.

Acknowledgments

First I would like to thank Peter Heutink for supervising my PhD, allowing me a lot of scientific freedom and for just being a great boss in general. Peter's guidance has helped me a lot to grow as a scientist and as a person. A big thanks also goes to Stefan Bonn, who supervised the bioinformatic side of my PhD and whose help was absolutely invaluable during this time. I also want to thank the third member of my advisory board, Thomas Gasser, for taking the time to supervise me and give helpful comments during our meetings.
Next, I want to thank Patrizia Rizzu, without whom we bioinformaticians could not exist because there would be no data. I also want to thank all the other members of the wet lab that have helped generating all this nice data! Of course I want to thank my fellow bioinformatics PhD students for all the lunch breaks and scientific as well as non-scientific discussions.
I am most grateful, however, to my parents, who have supported me in innumerable ways throughout my studies and life and still continue to do so, and to Benita, who listened to me during times of frustration and always managed to cheer me up. Thank you!

Contents

Chapter 1

Introduction

In this introduction, I will give a detailed presentations of several topics that are of relevance to this thesis. In section 1.1 I will give an introduction to frontotemporal dementia and its clinical and pathological symptoms, and furthermore discuss our current knowledge about the underlying genetics and molecular disease mechanisms. In section 1.2 I will review the biology of the transcriptome and cover experimental and computational methods for the analysis of it. Lastly, I will cover the topic of machine learning and deep neural networks, as these are central topics for chapter 2. I will conclude this introduction by outlining the aims of the thesis.

1.1 Frontotemporal Dementia

Frontotemporal Dementia (FTD) is a neurodegenerative disorder that is characterized by progressive deterioration of the frontal and temporal lobes [19]. It is a pre-senile dementia, meaning it typically manifests before the age of 65. Among pre-senile dementias, FTD is the second most common form after Alzheimer's disease (AD). Arnold Pick was the first to describe a patient with FTD in 1892, leading to "Pick's disease" as the initial term for the FTD disease spectrum. Both clinically and pathologically, FTD is a heterogeneous disorder.

1.1.1 Clinical Features

Depending on the clinical features, FTD can be divided into three subtypes. The most common form is behavioural variant FTD (bvFTD). Patients with bvFTD are mainly characterized by changes in their behaviour. Striking features are social disinhibition with inappropriate and compulsive behaviours [154]. Apathy and diminished empathy are additional symptoms and patients often exhibit hyperorality behaviours such as binge-eating or drinking. The memory is relatively unaffected in early bvFTD, but deficits in executive functions are often present. The other two clinical variants of FTD are primary progressive aphasias (PPAs), which are characterized by language impairments. Non-fluent variant PPA (nfvPPA) or agrammatic variant PPA leads to impairment in speech generation. Patients have grammatic errors in their language and problems in

the comprehension of long sentences. The speech is slow and halting with sound errors [67]. Finally, patients with semantic variant PPA (svPPA) have usually no problems with speech production and grammar, but they lose the semantic meaning and comprehension of words. Impaired object knowledge, dyslexia and dysgraphia are additional features that characterize svPPA [67]. Additional to the three main clinical manifestations of FTD, a significant number of patients develop symptoms of motor neuron disease (MND). In fact, over recent years, compelling evidence has accumulated that led to the classification of FTD and MNDs such as amyotrophic lateral sclerosis (ALS) as two extreme ends of a disease spectrum [36]. At the one extreme, patients present only with cognitive symptoms (FTD) and at the other end only motoric deficits are observed (MND/ALS). Motor syndromes with FTD are categorized as progressive supranuclear palsy (PSP), corticobasal syndrome (CBS) or ALS [143].

1.1.2 Neuropathology

Like the clinical symptoms, the neuropathology of FTD is highly heterogeneous. The major pathological hallmark is frontotemporal lobar degeneration (FTLD), which is defined by neuronal loss, atrophy and gliosis [88]. FTLD is the common underlying cause of all variants of FTD. On the cellular level, FTLD is characterized by pathologic inclusions of protein accumulates. Today, three specific subtypes of FTLD have been identified which are defined by the primarily accumulating proteins: microtubule-associated protein tau (MAPT), TAR DNA binding protein 43 (TDP-43) and fused in sarcoma (FUS). The corresponding pathologies are termed FTLD-tau, FTLD-TDP and FTLD-FET, respectively (Fig.1.1). FTLD-tau and FTLD-TDP make up most cases (90-95%) [57], with FTLD-tau accounting for approximately 40% of cases and FTLD-TDP for 50%.

Abnormal accumulation of the tau protein is not an exclusive feature of FTLD-tau, but also occurs in several other neurodegenerative diseases, most prominently in AD [206]. The gene microtubule associated protein tau (MAPT) consists of multiple isoforms that give rise to proteins with different chemical properties. Most relevant for FTD is the alternative splicing of exon 10, which leads to two tau variants with either three or four microtubule-binding repeat domains, referred to as 3R and 4R tau. Pathologically, FTLD-tau can be further divided into four molecular subtypes: Pick's disease (PiD), PSP, corticobasal degeneration (CBD), globular glial tauopathy (GGT) and argyrophilic grain disease (AGD) [122]. PiD is characterized by neuronal loss and swollen neurons, as well as neuronal cytoplasmic inclusions called Pick bodies. PiD typically presents without MND, in contrast to PSP which is primarily a movement disorder. In PSP, neurodegeneration affects several subcortical brain regions apart from frontal and temporal lobes. Both PSP and CBD manifest with glial tau inclusions, which is less frequently observed in PiD. GGT, as the name suggests, manifests with globular, widespread tau pathology with inclusions in astrocytes and oligodendrocytes [5]. Finally, AGD is a

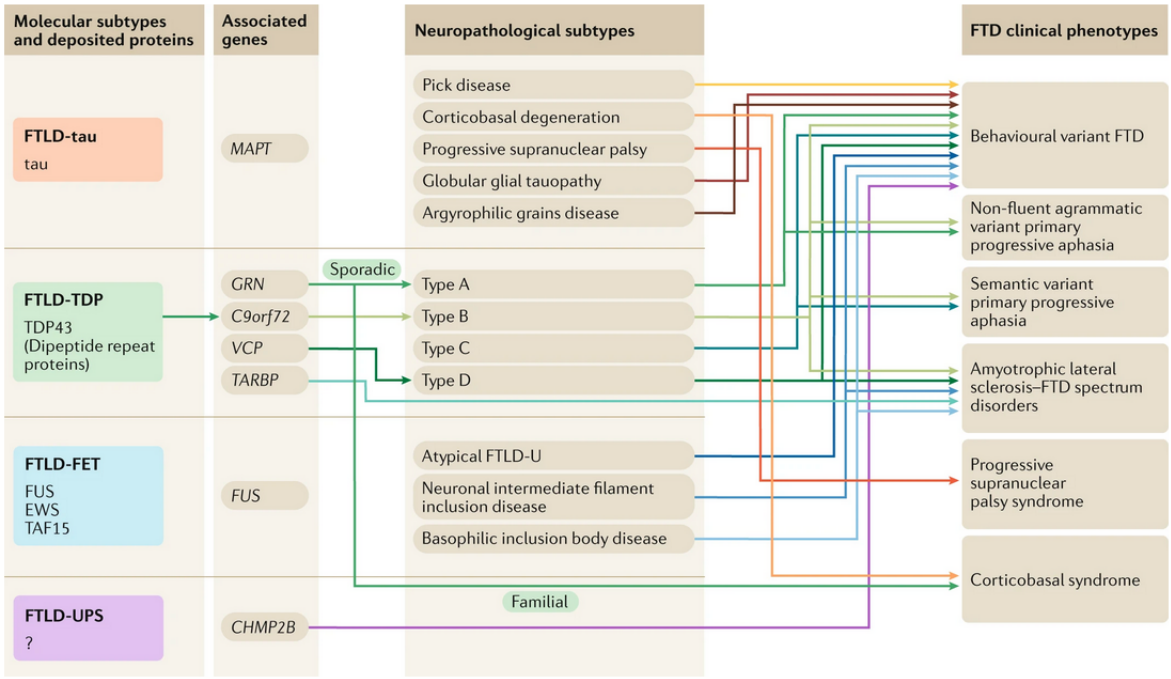

Figure 1.1: Subtypes of FTLD and FTD. The table presents an overview of pathological subtypes and associated genes as well as clinical FTD variants. From Panza et al. [143] with permission from Springer Nature

late onset tauopathy that is characterized by argyrophilic grains, oligodendrocytic coiled bodies and neuronal pretangles [161]. All the above discussed pathologies contain tau inclusions that are composed mainly of 4R tau.

FTLD-TDP is characterized by pathological inclusions of the TAR DNA-binding protein 43 (TDP-43), which has numerous molecular functions in RNA processing [121]. As with tau aggregates, TDP-43 inclusions are not an FTD-exclusive pathology, but are also the major pathology in ALS [138, 10]. Hyperphosphorylation, ubiquitination and N-terminal truncation are pathological features of accumulating TDP-43 protein [137]. FTLD-TDP is characterized in subtypes A, B, C and D, depending on the precise neuropathological features such as the distribution of cellular inclusions.

Finally, FTLD-FET is characterized by pathological inclusions of the protein encoded by the gene fused-in-sarcoma (FUS), which co-aggregates with the two FET-family member proteins Ewings's sarcoma (EWS) and TATA-binding protein-associated factor 15 (TAF15), hence the terminology [122, 137]. FTLD-FET is further characterized in the pathological subtypes atypical FTLD-U, neuronal intermediate filament inclusion disease (NIFID) and basophilic inclusion body disease (BIBD).

1.1.3 Genetics of FTD

FTD has a large genetic component, with estimates of up to 43% of cases having a positive family history and 10-17% being caused by an autosomal dominant mutation [151]. During the past decades, mutations in several genes have been identified as causes for FTD. In 1998, Hutton and colleagues identified missense and 5'-splice-site mutations in the MAPT gene as causes for FTD [84]. The tau protein encoded by MAPT interacts with microtubules and is important for their stabilization and organization [206]. Many pathogenic MAPT mutations lead to a decreased ability of tau to interact with microtubules, which leads to their destabilization and an excess of unbound tau protein in the cell, possibly leading to an increased accumulation potential. FTD caused by MAPT mutations is characterized by neuronal and microglial inclusions of tau (FTLD-tau). Intronic MAPT mutations lead to a ratio change of isoforms with either three or four microtubule-binding repeats (3R and 4R, respectively). Mutations in coding regions are also mostly found in these repeat regions and often lead to a reduced microtubule-binding ability of the resulting protein [185].

Loss of function (LOF) mutations in the granulin (GRN) gene were identified as another cause for FTD in 2006 [18, 45]. Mutations in GRN lead to a 50% reduction in GRN protein product and thus cause FTD by haploinsufficiency. FTD caused by mutations in GRN is characterized by ubiquitin-positive inclusions which are mainly composed of the TDP-43 protein (FTLD-TDP). GRN is expressed in neurons and microglia, where it is involved in the inflammatory response and many other cellular processes [97]. The exact mechanisms that lead from GRN haploinsufficiency to TDP-43 aggregates and neurodegeneration are not well understood.

In 2011, two independent research groups identified an expansion of the hexanucleotide repeat of sequence GGGGCC in the first intron of the C9orf72 gene as a cause for both FTD and ALS [158, 48]. Now, this mutation is the most common known monogenetic cause for FTD and ALS [196]. How the same mutation leads to two clinically quite different diseases is not entirely clear, but it demonstrates that ALS and FTD are part of a heterogeneous disease spectrum. Repeat expansions in C9orf72 lead to pathological inclusions of TDP-43, which is also the most prominent pathology in ALS [115]. Like GRN mutations, it is not known how the C9orf72 mutation leads to TDP-43 accumulation. Another pathological feature of the hexanucleotide repeat expansions (HRE) is the unconventional non-ATG translation of the repeat, which leads to dipeptide repeats (DPRs) that form insoluble inclusions [13]. The non-translated repeat RNA molecules can furthermore form toxic RNA foci, which represent another potentially important pathogenic mechanism of the C9orf72 mutation [200]. Finally, the repeat expansion leads to a reduced expression of the C9orf72 protein product, which can at least partly be attributed to hypermethylation at the C9orf72 locus [23].

While mutations in MAPT, GRN and C9orf72 account for most familial cases, other causal mutations have been identified that are observed much less frequently. A mutation in the Charged multivesicular body protein 2B (CHMP2B), which is an important component of the endosomal ESCRT-III complex, was identified as causal for FTD in 2005 [180]. Other rare causes of FTD are mutations in the genes vasolin-containing protein (VCP), sequestome 1 (SQSTM1), TANK binding kinase 1 (TBK1), optineurin (OPTN), TDP-43, FUS and CHCHD10 [151].

1.1.4 Disease Mechanisms

Ultimately, to develop treatments for FTD, it is important to have a precise understanding of the molecular mechanisms that are underlying disease development and progression. The identification of several causal mutations for FTD depicted important starting points for the determination of causal mechanisms, leading to great advances in our understanding of the molecular biology of FTD. Nevertheless, it is still unclear how precisely mutations in very different genes lead to a seemingly similar disease and what the underlying pathways of this process are. In the following, I will give an introduction into what is currently known about molecular pathways involved in the disease.

The tau protein, encoded by the gene MAPT, is involved in many neurodegenerative diseases, termed tauopathies. In physiological conditions, it binds to microtubules, where it has a stabilizing role. As mentioned earlier, some FTD-causing MAPT mutations lead to different isoform ratios of 4R and 3R tau. These two tau isoforms have different chemical properties, such as microtubule binding affinity. FTLD-tau subtypes often have accumulations of only a specific tau isoform, or sometimes both [51]. It has been shown that overexpression of one tau isoform can lead to a dysfunction in axonal transport, by reducing the localization of microtubules to axons and impairing their transport mechanisms [184]. Another suggested mechanism is the binding of 3R and 4R isoforms to specific sites at microtubule. Overexpression of one isoform could then lead to increased amounts of unbound tau protein due to saturation of binding sites [185]. Changes in ratio of 3R and 4R tau are usually caused by intronic or splice site mutations. However, missense mutations in coding regions have also been shown to decrease the microtubule binding activity of tau, consequently leading to diminished microtubule assembly [47]. Furthermore, multiple studies have shown that many pathogenic MAPT mutations lead to an increased seeding and aggregation capability of tau, making it more likely to form pathological inclusions [169]. Hyperphosphorylation of the tau protein is another mechanism observed in tauopathies like AD and FTLD-tau that can lead to impaired binding of tau to microtubules and the formation of neurofibrillary tangles (NFTs) [29] (Fig. 1.2). It is unclear which mechanism - impaired microtubule function or increased tau aggregation - is the main cause of neurodegeneration. Most likely, both play a role in pathogenesis [185].

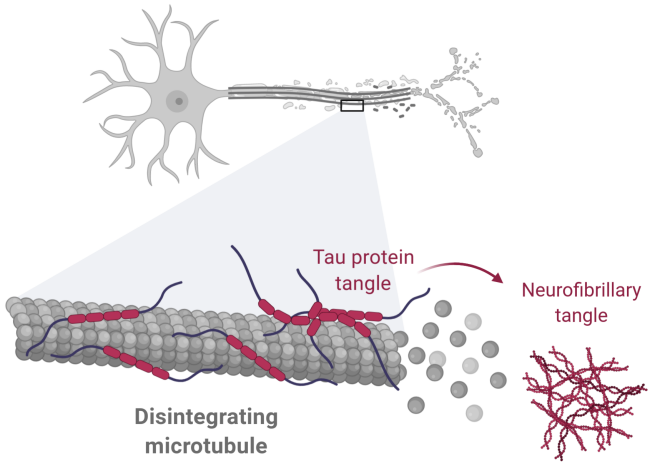

Figure 1.2: Destabilization of microtubules. Impaired binding of tau to microtubules leads to microtubule destabilization. Unbound, hyperphosphorylated tau builds neurofibrillary tangles. (Created with BioRender.com)

Because mutations in MAPT lead to aggregation of the protein encoded by the gene itself, the causal reason for tau aggregation can be assigned to the initial mutation. However, for FTLD-TDP pathology, it is much more difficult to determine how mutations in GRN or C9orf72, two genes seemingly unrelated to TDP-43, lead to aggregation of the latter. As most pathogenic GRN mutations lead to haploinsufficiency, loss-of-function is the most likely disease mechanism for GRN. The suggested roles of GRN in the brain are diverse, ranging from neurite outgrowth, neuron survival and differentiation to anti-inflammatory effects [43, 197]. In the brain, GRN is mainly expressed in microglia, and studies have shown that GRN deficiency can lead to stronger activation of microglia after injury, suggesting that GRN keeps the inflammatory response under control [189]. Another study in mice has shown that GRN-deficiency leads to age-dependent up-regulation of lysosomal and innate immunity genes in microglia, as well as enhanced synaptic pruning [120]. Therefore, the authors of the study suggest, immune system activation and synaptic pruning are likely drivers rather than consequences of neurodegeneration in FTD caused by GRN mutations. Interestingly, homozygous GRN mutations cause neuronal ceroid lipofuscinosis (NCL) instead of FTD. NCL is a lysosomal storage disorder, supporting the involvement of GRN in lysosome biology [181]. Indeed, several studies have associated abnormalities in lysosomes with GRN deficiency, such as enlarged lysosomes in GRN-deficient mice or in induced pluripotent stem cell (iPSC) neurons from FTD patients with GRN mutations [148].

As described in the previous section, pathological effects from HREs in the C9orf72 gene are caused by diverse mechanisms: down-regulation of C9orf72, DPRs and RNA foci. Possible toxicity has been shown for all of these effects, however it remains to be determined to which extent these effects are involved in neurodegeneration [190]. Studies in model systems and mice suggest that reduced expression of the C9orf72 gene is not sufficient to induce severe neurodegeneration [35]. However, recent studies suggest that reduced levels of C9orf72 can work synergistically with other gain-of-function effects of the mutation, thereby exacerbating their pathogenicity [228]. The protein encoded by C9orf72 has been shown to be structurally similar to Rab guanine nucleotide exchange factors (RabGEFs), which are proteins that are involved in cellular trafficking and autophagy pathways [113]. Building on this finding, multiple studies have provided evidence that C9orf72 plays important roles in autophagy pathways. Interestingly, Sellier and colleagues could show in neurons that loss of C9orf72 impairs autophagy and leads to protein aggregates of TDP-43 and the p62 protein, given a modifying mutation in the Ataxin-2 gene is present as well [173]. Apart from autophagy, it was also confirmed that C9orf72 is involved in cellular trafficking pathways [9].
It is debated to which extent DPRs play a role as a gain-of-function mechanism in FTD pathogenesis. In post-mortem human brain tissue, DPRs have been found in various brain regions, but their localization does not correlate with FTLD-TDP pathology [71, 66]. However, a recent study found that the poly-GR subtype of DPRs does correlate with neurodegeneration and co-localizes with TDP-43 in dendrites [166]. Nevertheless, more research is needed to more deeply investigate the role of DPRs in FTD caused by C9orf72 mutations.
Finally, the bidirectional transcription of the C9orf72 HRE leads to RNA molecules that form aggregates, or RNA foci, with potentially pathological functions [48]. These foci have been observed in the nuclei of multiple cell types in post-mortem brains from patients with FTD as well as in iPSC-derived neurons [71]. Nevertheless, it remains unclear if and how RNA foci have pathological effects. For instance, Mizielinska and colleagues found that RNA foci are abundant in FTD patients with C9orf72, being found in 51% of frontal neurons, although they did not correlate with TDP-43 or p62 inclusions. While this indicates that they are not necessary for the major pathological hallmark of FTD to evolve, it does not rule out pathological functions [129].

Above, I have described the initial insults caused by mutations in MAPT, GRN and C9orf72 that ultimately lead to FTD. As the disease develops and progresses, many more molecular pathways are affected, that are to some extent common to patients with mutations in all three genes and also to other neurodegenerative disorders (see Fig. 1.3). These pathways lay downstream of the causal mutations and it might be their dysfunction that ultimately leads to neurodegeneration.

Cellular trafficking pathways are affected in several neurodegenerative disorders [65], such as endo-lysosomal trafficking pathways [203], which are important for functional

autophagy. As discussed earlier, both GRN and C9orf72 likely play vital roles in lysosomal and autophagy related biology, similar to other genes with causal FTD mutations such as CHMP2B. Furthermore, microtubules are important for the formation of autophagosomes, thereby providing a potential link between MAPT and autophagy as well [100]. Autophagy and lysosomes are important mechanisms that enable cells to dispose of dysfunctional organelles or proteins. If non-functional, proteins or organelles that would ideally be removed can accumulate in the cell and subsequently build aggregates. For instance, a recent study indicates that lysosome-deficiency or defective vesicular transport might be the mechanisms that lead to a failure in clearance of tau proteins, while dysfunctional lysosomes might be the cause of aggregating TDP-43 in GRN mutation carriers [17].

Another common feature of genetic FTD subtypes are stress granules (SGs). SGs are membraneless organelles that are formed through liquid-liquid phase separation by ribonucleoproteins (RNPs) and can contain several RNA and protein species. They dynamically form during stress and quickly disassemble afterwards [211]. It has been proposed by Wolozin and Ivanov that SGs could serve as the initial location for protein aggregation, as they harbor proteins in greatly increased concentrations and TDP-43 has been shown to colocalize to SGs [211]. According to their hypothesis, SGs could form as a response to stresses that are mediated by the initial insults from disease-causing mutations, e.g. down-regulation of GRN or hyperphosphorylation of tau. Dysfunctions in trafficking and autophagy pathways, as discussed above, could impair the cellular capabilites of removing excess SGs, leading to their accumulation and thereby providing optimal seeding environments for pathological protein inclusions. While the different steps in this hypothesis need to be experimentally validated, SGs very likely play important roles in FTD pathogenesis.

During recent years, researchers of neurodegenerative diseases have increasingly focused on neuroinflammatory processes, which are now believed to play key roles in many diseases, including FTD [65]. Increased microglial activation in frontotemporal regions in FTD has already been identified in 2004 [37]. Recent studies found inflammation-associated co-expression modules of genes up-regulated in FTD and partly correlating with TDP-43 pathology [193, 186]. Increased levels of tumor necrosis factor (TNF), a major driver of inflammation, have furthermore been found in patients with FTLD-TDP caused by GRN mutations [127]. The gene GRN itself has important functions in the immune system, and has been shown to regulate microglia and the immune response in several studies [32]. Recently, a study of C9orf72 knockout mice has shown that C9orf72 loss-of-function leads to an increased interferon response [124]. The authors found that impaired autophagy leads to decreased degradation of the stimulator of interferon genes (STING) protein, thus provoking an increased interferon response. This goes in line with previous findings which have shown that C9orf72 is required for microglial function [141]. As with other pathological mechanisms in FTD, their precise role needs to be

further examined. However, there is compelling evidence showing that neuroinflammation is an important component of at least some FTD subtypes.

As a last example of a major molecular pathway that is involved in many neurodegenerative diseases and that might also play important roles in FTD, I want to discuss the dysfunction of mitochondria. Mitochondria generate the energy necessary for cellular functions, and are thus crucial for cell survival. Neurons are large cells with high energy demands that are dependent on functional mitochondria and their transport to all parts of the cell, making them especially vulnerable to mitochondrial dysfunction [65]. In a properly functioning cell, malfunctioning mitochondria are removed through a process called mitophagy, which is a specialized form of autophagy. Hence, impaired autophagy can lead to suboptimal mitochondrial quality control, ultimately leading to an impaired energy production. Interestingly, missense mutations in the gene CHCHD10 have been identified to be a likely cause for FTD-ALS [20]. While the exact function of CHCHD10 is still unknown, it is a mitochondrial protein, thus suggesting a mitochondrial origin of FTD. Furthermore, Choi and colleagues recently generated mouse models for C9orf72 FTD/ALS with inducible DPRs. They found that DPRs led to mitochondrial dysfunction and proposed that this might be an early event in the disease [44].

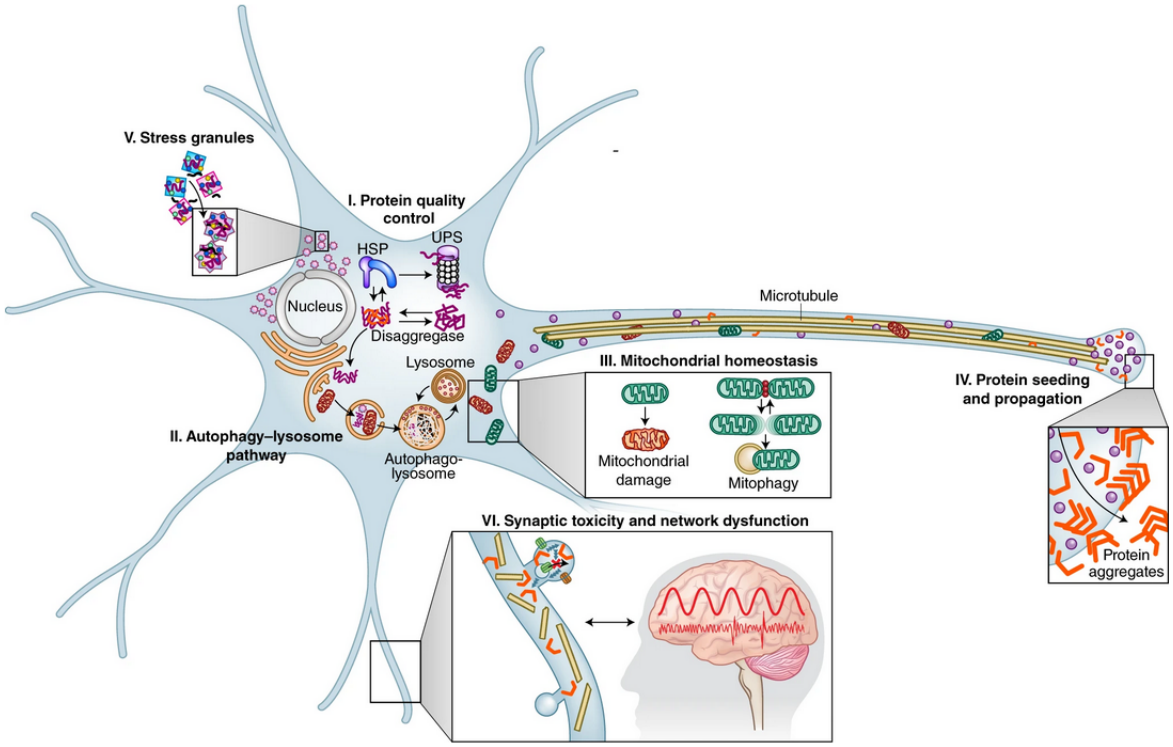

Figure 1.3: Commonly affected pathways in neurodegenerative diseases include protein quality control (I), autophagy and lysosomal pathways (II), mitochondrial homeostasis (III), protein seeding and aggregation (IV), the formation of stress granules (V) and mechanisms leading to synaptic toxicity (VI). Figure taken from Gan et al. [65] with permission from Springer Nature.

With the advancement of our understanding of the molecular disease mechanisms underlying FTD, potential treatment strategies can emerge which can be tested and might ultimately lead to a drug that halts the progression of FTD. Treatment strategies that are currently investigated focus mainly on inhibition and clearance of tau and TDP-43 aggregates [143]. Further strategies include microtubule stabilization, inhibition of tau phosphorylation, modulating GRN and C9orf72 expression and modulation of autophagy [143]. A more detailed understanding of FTD molecular biology is necessary to determine which strategies are the most promising ones without leading to undesired side-effects.

1.2 Transcriptomics

The transcriptome comprises all RNA molecules in a cell that are transcribed from the DNA. It is central to all cellular functions, as it is the basis for all proteins as well as

non-coding RNA molecules such as micro RNAs (miRNAs) or long non-coding RNAs (lncRNAs) in the cell. Given its key role for cellular biology, it is not surprising that the transcriptome is subject to pathological changes in many diseases, leading to dysfunction in sometimes essential molecular pathways. It is because of this vital importance that transcriptomic profiling has become one of the most important methods in molecular cell biology that helped scientists to unravel the mechanisms behind many complex diseases. Here, I want to outline the technologies and computational algorithms used for transcriptomic profiling as well as the different molecules and regulatory mechanisms that are part of or affect the transcriptome.

1.2.1 Molecular Biology

The transcriptome encompasses a diverse set of molecules, not all of which are fully understood in their functions. Messenger RNA (mRNA) is the most important RNA molecule, as it is translated into proteins, which are the central molecular machines of every cell. An eukaryotic protein coding gene consists of non-coding (introns) and coding (exons) regions. After transcription, intronic regions are spliced out of the immature mRNA molecule to yield a mature mRNA which only consists of exons. Only this mature mRNA is then translated into a functional protein. Through different combinations of exons or the usage of alternative splice sites that define different intron-exon boundaries, many genes can be transcribed into multiple variants, or isoforms, with unique biological functions. This process is called alternative splicing, and is an important mechanism in development and other molecular pathways [21, 144].

MiRNAs are small RNA molecules of 21-22 nucleotides length that play important roles in transcriptional and translational regulation. They bind to the 3'-untranslated regions (3'-UTR) of mRNAs through a complex mechanism that is facilitated mainly through a short, 7-8 basepair (bp) complementary sequence in their 5'-region, called the seed sequence [74]. Because of this short complementary sequence, every miRNA has the potential to target several genes, sometimes hundreds or thousands. Binding to the 3'-end of a target gene leads to translational repression and increased mRNA degradation. The latter is caused by destabilization of the mRNA, and evidence suggests that this is the major mechanisms of action for miRNAs [82]. Regulation by miRNAs is an important mechanisms, whose dysfunction has been implicated in a variety of diseases [147].

In eukaryotes, transcription of RNA from DNA is carried out by RNA polymerase enzymes. There are three specific polymerases: Polymerase (Pol) I, Pol II and Pol III. Each of these polymerases is responsible for transcribing a specific set of RNA molecules. Protein-coding mRNA is transcribed by Pol II, for example [167]. Polymerases bind to accessible gene promoters to initiate transcription. This binding to promoters is facilitated by transcription factors (TFs), which depict the most important regulatory molecules in the transcriptome (Fig. 1.4). TFs bind to specific DNA elements that con-

sist of a sequence motif for which the respective TF has a much higher binding affinity compared to other DNA sequences [107]. DNA elements that are bound by TFs are also referred to as transcription factor binding sites (TFBS). The selective binding of TFs to certain DNA elements enables them to initiate the transcription of a specific set of genes. Apart from TFBSs in promoter regions, TFs can also bind to other regulatory elements such as enhancers [183]. Binding to enhancer elements activates gene expression of genes that can be located either upstream or downstream of the enhancer at a significant distance from the gene promoter. The prominent role of TFs in transcriptomic regulation makes them interesting targets to study and for therapeutic intervention.

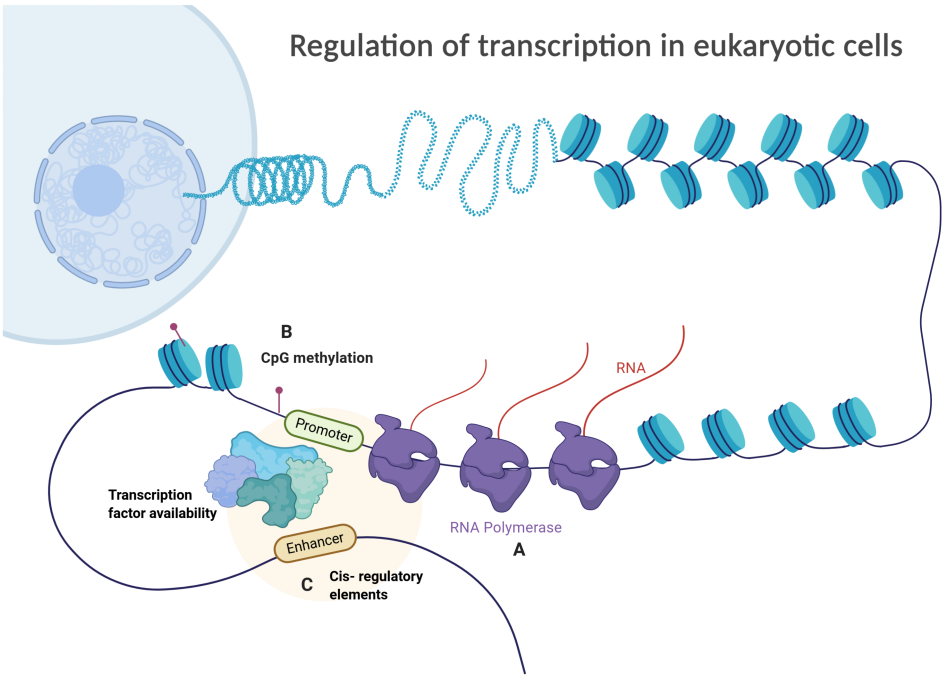

Figure 1.4: (Adapted from Regulation of Transcription in Eukaryotic Cells, by B. Atanasovska, BioRender.com (2020), Retrieved from https://app.biorender.com/biorender-templates)

Apart from regulation by TFs, epigenetics, or the heritable modification of DNA, is another important regulatory mechanism of gene expression. A prominent example for DNA modifications is the addition of methyl groups. The most generic form of this modification is the methylation of cytosine residues, which usually occurs when preceding a guanine residue (CpG dinucleotide) [68]. CpG dinucleotides are largely depleted from the genome, except for so-called CpG islands which usually occur in promoter regions [85] (Fig. 1.4). Methylation of cytosine residues in CpG island of promoters can result in decreased promoter activity and hence gene expression by impairing the

possibility of TFs to bind the promoter [68]. CpG methylation can also activate gene expression through different mechanisms. The possibilities of influence on gene expression by DNA methylation are diverse and complex and remain a field of active research. It is, however, well established that DNA methylation plays a crucial role during development and in many diseases. Several technologies have been developed for the study of DNA methylation. Some are based on DNA sequencing technology, while others are based on microarrays. For the latter, the most common platform is the Infinium assay from Illumina. Based on this assay, the MethylationEPIC BeadChip can be used to examine 850,000 CpG sites for methylation. All methylation data used in this thesis was generated using the MethylationEPIC BeadChip.

1.2.2 Sequencing Technology

Technological advancements have paved the way for large-scale studies by making transcriptome profiling relatively cheap. The first technology that allowed for high-throughput transcriptional profiling were microarrays – small chips with attached and predefined oligonucleotides. To measure RNA quantities, the RNA is extracted from a sample, reverse-transcribed into complementary DNA (cDNA), labelled with a fluorescent dye and then hybridized to the microarray chip. Gene expression can then be measured through the strength of the fluorescent signal of each probe. Nowadays, the most commonly used technique to measure gene expression is RNA-sequencing (RNA-seq), whose success is based on the rapid development and decrease in cost of DNA sequencing technology. The most common platform for sequencing technology in general comes from Illumina. For RNA-seq, RNA from each sample is extracted from a sample and reverse-transcribed to cDNA. For gene expression measurements, DNA fragment lengths between 50 and 200 bp are typical. The cDNA libraries are then ligated to sequencing adapters on a flowcell and amplified via polymerase chain reaction (PCR). Finally, the amplified cDNA fragments are sequenced. In the case of Illumina sequencing, this is done by a so-called sequencing-by-synthesis process (Fig. 1.5). First, the complementary reverse strands are removed from the cDNA on the flow cell and a primer is attached to the now single-stranded sequence. Fluorescently labelled nucleotides are then incorporated by a DNA polymerase one at a time. After each sequencing round, the incorporated nucleotide is determined using the fluorescent signal, which is different for each of the four nucleotides. Thus, the sequence of the DNA fragment can be determined with high precision. In the case of paired-end sequencing, the DNA fragment is additionally sequenced in the reverse direction from the opposing end, yielding two reads (forward and reverse reads). The availability of forward and reverse reads has advantages for data analysis steps like sequence alignment or genome assembly.

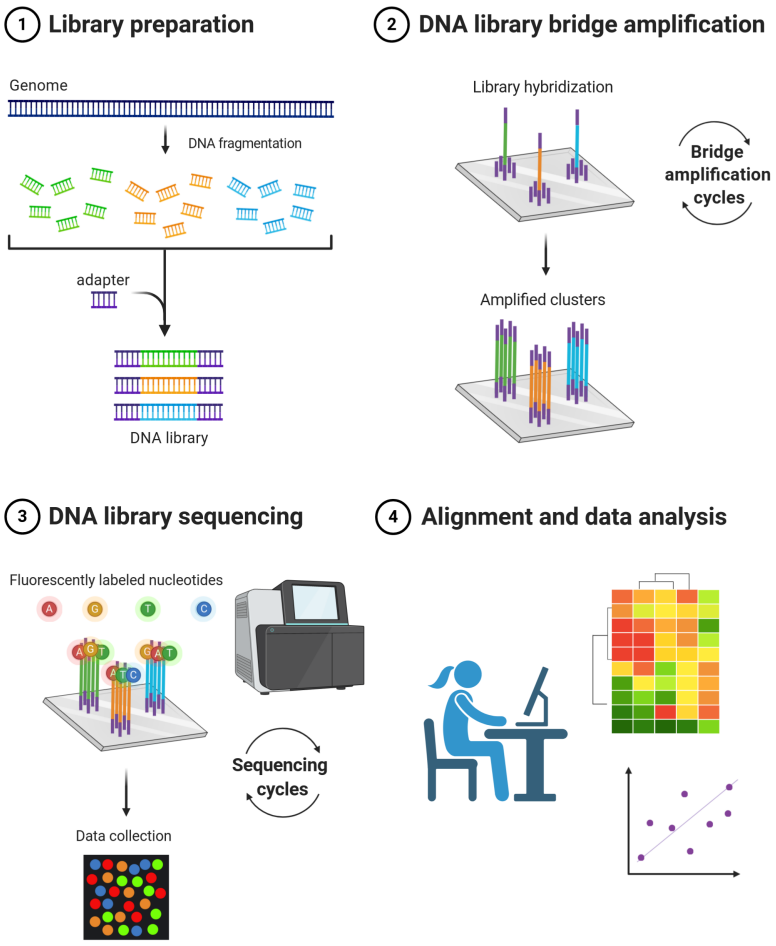

Figure 1.5: Illustration of Illumina Next Generation Sequencing. **1** DNA is cut into smaller fragments and ligated with sequencing adapters to yield a sequencing library for each sample. **2** The libraries are hybridized to the flow cell and amplified in clusters using bridge amplification. **3** Fluorescently labeled nucleotides are added to the flow cell to synthesize the complementary DNA strands. The fluorescent signal is captured during every cycle. **4** Sequencing data can be aligned to a reference genome, for example, in the final data analysis step. (Adapted from Next Generation Sequencing (Illumina), by BioRender.com (2020), Retrieved from https://app.biorender.com/biorender-templates)

Cap analysis of gene expression sequencing (CAGE-seq) is another sequencing-based technology that uses the 5'-cap of RNA molecules to extract their 5'-end sequence [177]. The very beginning of the 5'-end corresponds to the transcription start site (TSS), hence CAGE-seq can be used to accurately determine and quantify TSSs of all genes in the transcriptome. The position and distribution of CAGE-seq TSS tags enables the identi-

fication and characterization of gene promoters. Quantification of tags adds information about the promoter activity.

Technologies like RNA-seq and CAGE-seq are typically used to study the gene expression of a tissue sample that consists of a mixture of different cell types or of a homogeneous sample of cells that consists of thousands of cells. In recent years, however, it has become quite common to study gene expression in individual, single cells using single cell RNA-seq (scRNA-seq) technology. While scRNA-seq is still relatively expensive compared to bulk RNA-seq, rapid technological advancements have helped to decrease the cost, which has led to a strong increase in the number of studies that use scRNA-seq and to a vast improvement of our understanding of single cell biology. The first step in a scRNA-seq protocol is to isolate single cells and to bring them in solution. Once in solution, one of a variety of different scRNA-seq technologies can be used. One of the most common technologies comes from 10X Genomics. For example, the 10X Genomics Chromium system mixes the cell solution with enzymes, partitioning oil and gel beads, which contains barcodes and unique molecular identifier (UMI) adapters. The Chromium system creates pairs of single cells and beads that are encapsulated in an emulsion, so called Gel Bead-in-emulsions (GEMs). In each GEM, the RNA is reverse transcribed into DNA and labelled with the bead-specific barcode and UMIs. Finally, sequencing adapters are attached, resulting in a sequencing library that can be sequenced using standard short-read sequencing technology as described above. Because every GEM contains a unique barcode, it is possible to assign reads to their cells of origin, enabling the study of gene expression at single cell resolution. A disadvantage of scRNA-seq is that because of the low amount of RNA that can be obtained from a single cell, the number of genes that can be quantified is much lower compared to bulk RNA-seq. These low expression measurements lead to stochastically missing genes (zero counts), so-called dropouts.

1.2.3 Algorithms for transcriptome analysis

RNA-seq experiments can generate large amounts of data in the form of millions of sequencing reads that require further computational processing in order to extract meaningful biological insights. A typical workflow for the analysis of RNA-seq data starts by assessing the quality of the reads from the experiment, demultiplexing pooled samples and, when necessary, the trimming of remaining adapters or low-quality bases. The cleaned reads then need to be aligned to a reference genome to determine their genes of origin. The first efficient algorithm for the alignment of two biological sequences was proposed by Needleman and Wunsch in 1970 [136]. The Needleman-Wunsch algorithm is based on dynamic programming and can generate a global sequence alignment in $O(n^2)$ time, where n is the length of both sequences (if they are of the same length). In 1981, the first algorithm for local sequence alignment was proposed as a variation of the Needleman-Wunsch algorithm by Smith and Waterman, which has the same time complexity as the Needleman-Wunsch algorithm [182]. These algorithms represented hallmarks of bioin-

formatics but are not useful in their original form for the large amounts of data generated to date. Thus, in the last decades, numerous algorithms for the efficient and precise alignment of sequences have been proposed that can tackle the challenge to accurately align millions of reads against large reference genomes in reasonable time, using limited computational resources. Modern alignment algorithms need to be extremely efficient, and thus usually use a seed-matching scheme to avoid unnecessary alignment – a concept initially introduced by Altschul et al. with the BLAST algorithm [7]. By efficiently searching for exact seed matches between a read and the reference genome, the number of sequence alignments to be calculated can be dramatically reduced. Apart from being efficient, RNA-seq alignment algorithms should be able to deal with unknown splice junctions, which was first addressed by the TopHat algorithm in 2009 [192]. Nowadays many efficient and precise algorithms for sequence alignment or genome mapping exist, two popular examples are STAR [52] and HISAT2 [94].

Once the genomic origins of sequencing reads have been determined by sequence alignment, the next step is to count reads that stem from a specific gene to get gene expression values. Numerous specialized tools have been developed for this task as well, with HTSeq and featureCounts as two popular examples [8, 114]. Alignment-free algorithms like kallisto [31] or Salmon [146] allow to directly quantify gene counts without the alignment step, which is extremely efficient and enables users to quantify transcript abundances on a normal desktop computer in a matter of minutes.

Finally, in order to compare RNA abundance estimates across different samples, the gene counts need to be normalized for library size (number of sequencing reads of a library, i.e. sample) and, if desired, for gene length. This step is usually included in the algorithms used for gene abundance estimation or in the tools used for differential gene expression (DGE) analysis, which is the most common downstream analysis for RNA-seq data. DGE analysis determines whether the difference in gene expression between two groups is statistically significant. Popular tools that can be used for DGE analysis and that have been specifically developed to consider the intricacies of RNA-seq data are DESeq2 and edgeR [119, 41].

1.2.4 Cell Type Deconvolution

While it is nowadays possible to study gene expression on the single-cell level using scRNA-seq technology, the vast majority of data that has been generated and is still being generated stems from so called ‘bulk’ RNA-seq, which effectively measures the gene expression of a tissue, comprised of a mixture of different cell types. In recent years, countless scRNA-seq studies have shown that gene expression profiles diverge significantly between different cell types. It is thus abundantly clear that bulk RNA-seq is measuring a mixture of different expression profiles, which is greatly dependent on the cell type composition of the sample under investigation. It can therefore be problematic

to compare gene expression between two different sample groups with systematic differences in cell type composition. For instance, one cell type could be depleted in one group, leading to seemingly lower expression of genes specific to this cell type. While this issue has been overlooked in many RNA-seq studies, scientists have developed many methods to identify or address this problem.

The most prominent group of algorithms for this problem have been termed as cell type deconvolution algorithms. The goal of cell type deconvolution, from here on simply referred to as deconvolution, is to estimate the fractions of different cell types in a tissue sample from its gene expression profile. Deconvolution is based on the assumption that a bulk RNA-seq expression profile is the weighted sum of its constituting cell type-specific expression profiles. It can thus be defined as in equation (1.1):

$$M = G \times F \tag{1.1}$$

Where M is a $g \times n$ expression matrix, with g number of genes and n samples. The matrix G represents a $g \times c$ expression matrix, where c defines the number of different cell types in the mixture. The matrix F is a $c \times n$ matrix that defines the relative proportion of the different cell types in the samples.

Provided that the cell type specific gene expression profile matrix (GEP) is known and the number of samples exceeds the number of cell types, it is possible to estimate F by formulating the problem as a minimization problem with some objective loss function that measures the distance between F and $G \times F$. In theory, this problem can be solved by simple linear regression (see section 1.3.1). In practice, however, certain constraints must be taken into consideration. The first is a non-negativity constraint, as cell type fractions cannot be negative. The second constraint must enforce that cell type fractions from all cell types sum up to 1, as they together represent 100% of the tissue.

One of the first methods for cell type deconvolution was proposed by Abbas et al. in 2009 [4]. The authors used simple least squares regression to estimate the cell composition of immune cells in microarray blood data from patients with Systemic Lupus Erythematosus (SLE). The non-negativity constraint was addressed with an iterative scheme, in which the authors set negative coefficients to zero and repeated the procedure until all coefficients are positive. Numerous other methods have subsequently been proposed which use different optimization procedures and regression algorithms and add different regularization methods to the objective function.

What all deconvolution algorithms of the above described form have in common is that they are strongly dependent on a suitable GEP matrix. Only when the GEP matrix G resembles the true cell type expression profiles in the tissue under investigation, can the deconvolution yield good results. It is therefore crucial that the data used to generate

the GEP matrix is of good quality. The data used typically stems from expression profiling (either microarray or RNA-seq) of manually filtered cells or, more recently, from scRNA-seq studies. Furthermore, not all genes convey information that is useful for cell deconvolution, as many essential genes adhere to similar expression patterns across most cell types. Many algorithms thus also try to subset the GEP matrix to the set of genes that is most suitable for deconvolution. For instance, the popular deconvolution algorithm CIBERSORT infers genes that are differentially expressed in each cell type compared to the remaining cell types and then uses a specific number of genes with the highest fold-changes from each cell type for the GEP matrix. The authors chose the number of genes per cell type that yields the GEP matrix with the lowest condition number [139].

As scRNA-seq datasets for many tissues have become increasingly abundant, new deconvolution algorithms have been developed that are specifically designed to use scRNA-seq data for GEP construction. For instance, the Bseq-SC algorithm introduced a procedure to generate GEP matrices from scRNA-seq data that can be used with CIBERSORT [22]. Briefly, Bseq-SC rescales expression values to account for the variability in total RNA content of different cell types and then uses the average of normalized expression values across cells of the same type. Marker genes are selected based on high expression, high variation and cell type restriction. Another method that incorporates a framework both for GEP creation and deconvolution is MuSiC (Multi-Subject Single Cell Deconvolution) [205]. Instead of pre-selecting marker genes, MuSiC assigns weights to each gene which are based on the concept of marker gene consistency. To assure that markers are not specific to a certain subject, but consistent across subjects, MuSiC down-weighs genes with high cross-subject variance and up-weighs genes with low cross-subject variance. Additionally, MuSiC includes gene- and subject-specific intercepts in formulation of the optimization problem to adjust for protocol biases between bulk- and scRNA-seq data. In another recent work, Frishberg and colleagues proposed the method of Cell Population Mapping (CPM), which goes beyond cell types and uses the distinct cellular states of cells within a cell type to infer cell composition [62]. Instead of creating a single GEP matrix, CPM repeatedly samples cells from the input scRNA-seq dataset as GEP matrix and performs deconvolution via support vector regression (SVR). This procedure is repeated a defined number of steps and the estimated cell composition outputs are then averaged. CPM then uses dimensionality reduction algorithms to infer the abundance of cells in specific cellular states. As this is not part of the classical deconvolution algorithm, I will not further discuss this here. In 2019, the authors of the CIBERSORT algorithm introduced a new version, termed CIBERSORTx, which builds upon the original CIBERSORT algorithm and was specifically designed to use scRNA-seq data as reference [140]. For GEP creation, CIBERSORTx generates five reference samples by sampling 50% of cells from a specific cell type and then averaging the expression profiles. These samples are then used to generate a GEP matrix using the same procedure as CIBERSORT. Additional features of CIBERSORTx are two batch correction methods

that were developed to account for systematic differences between different platforms, such as bulk RNA-seq and scRNA-seq. Moreover, Newman and colleagues introduced another feature with CIBERSORTx, that allows to not only estimate cell compositions, but group-wise cellular expression profiles. For this, first the matrix of cellular fractions, F, is estimated, and then a new optimization problem is formulated as

$$H_i \times F = M_i \tag{1.2}$$

Where H_i is the i-th row of an expression matrix of $g \times c$ dimensions and M_i is the i-th row of the mixture matrix M, corresponding to the i-th gene. As F has been previously estimated using G, it can now be used to solve for H_i, which represents the cell type level expressions of the gene i across all samples. CIBERSORTx solves this problem via non-negative least squares (NNLS). When subsetting M and F for sufficiently large sample groups, group- and cell type-specific expression profiles can be estimated.

The field of cell type deconvolution is very active and numerous other algorithms have been suggested over the recent years. It is has now also become more common to apply deconvolution algorithms to bulk RNA-seq datasets in order to adjust for biases and to uncover new insights.

1.3 Machine Learning and Deep Neural Networks

The field of machine learning (ML) is concerned with the development of mathematical models that learn to solve defined tasks from data. It is often divided into the categories of supervised and unsupervised learning. Supervised learning algorithms learn from datasets that consist of input training data and corresponding output target values. Their purpose is to learn a function that maps input samples to their corresponding target values. Such a function can then be used to predict unknown target values from new input data. Unsupervised learning algorithms only require input data without target values. They are developed for various purposes, such as to identify meaningful patterns in the data that allow it to be represented in fewer dimensions. In this section, I will give an introduction to supervised machine learning algorithms in general and deep neural networks (DNNs) in particular, as these are relevant for chapter 2 of this thesis.

1.3.1 Linear Regression

Consider a dataset of N data points or observations x_n which are associated with corresponding target values y_n, where $n = 1, ..., N$. The goal of a supervised learning algorithm is to predict y given x for unseen values of x. Here, as an introductory example for a supervised learning algorithm, I will briefly discuss linear regression, which is one of the simplest ML algorithms. A linear regression model has the following form:

$$\hat{y}_i = w_0 + w_1 x_i^1 + ... + w_D x_i^D \tag{1.3}$$

Where D is the dimension of the input feature vector x_i and x_i^d is the d-th element of x_i. The target variable $\hat{y}_i$ is the predicted output of the linear model, and w is the vector of weights or parameters of size $D+1$. The goal of linear regression is to learn a vector w that yields a predicted target value $\hat{y}_i$ that is as close as possible to the true target value y_i. As can be seen from (1.3), the output of the model is a linear combination of the input features, weighted according to the vector w. The additional element w_0 of the weight vector is also referred to as bias or intercept. Because x and w are vectors, equation (1.3) can also be written as:

$$\hat{y}_i = x_i^T w + \varepsilon = y(x_i, w) + \varepsilon \tag{1.4}$$

Where w is the vector $(w_0, ..., w_D)$ and x_i^T is the transpose of the vector $(x_i^0, ..., x_i^D)$. To make the simpler vector form possible, we have added the additional element $x_i^0 = 1$ to x_i. Notice further that equation (1.4) now contains the error term ε, which is a Gaussian distributed variable that models the noise.

In order to find the optimal weight vector we have to minimize a cost or loss function L that defines how well our linear model $y(x_i, w)$ predicts the target value y_i. In the case of linear regression, a common and simple loss function is the squared error:

$$L(\hat{y}_i, y_i) = \frac{1}{N} \sum_{n=1}^{N} (y_i - \hat{y}_i)^2 \tag{1.5}$$

An analytical solution of w that minimizes $L(\hat{y}_i, y_i)$ can be found using the method of ordinary least squares (OLS). This can be seen when formulating the whole training set as a matrix X of dimensions $N \times D$, hence every row contains an observation and the different columns correspond to different features of the training data. When using y to denote the target values of all observations, we can formulate the linear model as

$$Xw = y \tag{1.6}$$

Using OLS, the optimal weight vector can now be calculated using the following equation:

$$w = (X^T X)^{-1} X^T y \tag{1.7}$$

Here, I will not go into the details on how to arrive at the solution (1.7). Instead, in the next section, I will introduce an alternative way of arriving at an optimal weight vector, which is more relevant to the topic of deep neural networks.

1.3.2 Gradient Descent

Using OLS, we can calculate the optimal solution to suitable optimization problems analytically. However, for large and complex datasets, OLS can be inefficient because it involves the inversion of the matrix $X^T X$. As X becomes large, this can imply a significant computational burden. OLS is also not applicable to all ML algorithms and problems. For these reasons, many modern machine learning algorithms are usually trained using some variation of the *gradient descent* algorithm, which was first proposed by Cauchy in 1847.

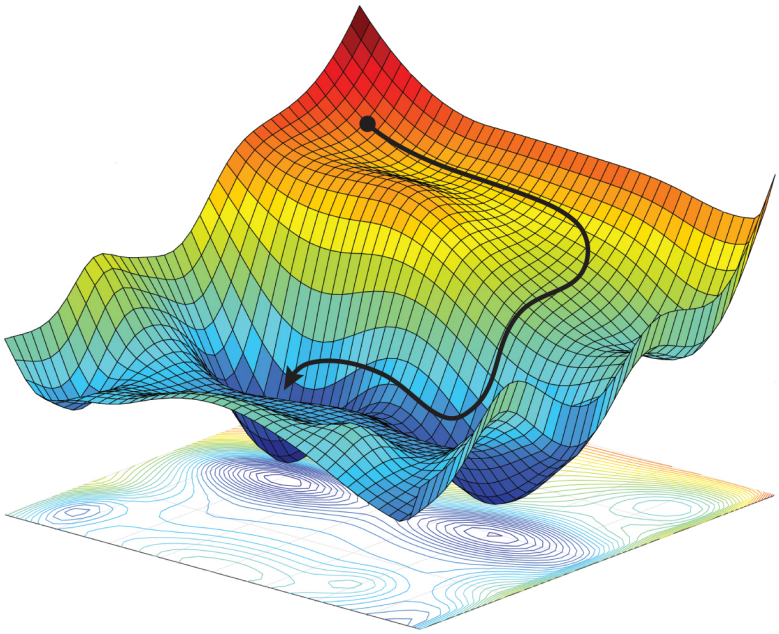

Figure 1.6: Illustration of gradient descent. The image shows a loss function where the two horizontal axes represent parameters of a ML model and the vertical axis as well as the color, represents the loss. Gradient descent seeks to reach the points of minimal loss using the curvature of the loss function. Source: A.Amini, D.Rus, Massachusets Institute of Technology, adapted by M.Atarod/Science [83]

The idea behind gradient descent is that it is possible to iteratively minimize an objective function by calculating the partial derivatives of the loss function with respect to every parameter of the model. The partial derivative, or gradient, defines the slope of the loss function with respect to a parameter. Hence, to decrease the loss function, we have to move into the direction of the negative gradient. Consider the loss function as a function of the model parameters $L(w)$, then one step of gradient descent is defined as:

$$w_{i+1} = w_i - \alpha \nabla L(w_i) \tag{1.8}$$

Where w_i is the weights vector at step i, $\nabla L(w_i)$ is the gradient of $L(w)$ at the point w_i and α is an empirically chosen step size or learning rate. If α is sufficiently small, a gradient descent step guarantees that $F(w)$ decreases with each step.

$$F(w_i) \geq F(w_{i+1}) \tag{1.9}$$

If the learning rate is chosen too large, the parameter update can even lead to an increase of the loss function because the local or global minimum is "overstepped". Gradient descent is a powerful method that allows to minimize (or maximize) any defined and differentiable loss function. At each step, gradient descent can be applied to the complete dataset, or to a subset of samples. The latter is referred to as stochastic gradient descent (SGD). Compared to OLS, gradient descent is not guaranteed to find the optimum, but rather finds a statistical approximation. In practise, this is sufficient. During the last decade, several optimized versions of gradient descent (GD) have been developed, that improve certain aspects of the original algorithm. Examples are AdaGrad [54], RMSProb (by Tieleman Hinton) and Adam [96], which builds on the other two algorithms. Briefly, improvements on GD are mainly based on an adaptive learning rate and momentum. Intuition for the latter can be gained by considering the meaning of momentum in physics. Here, the direction of the gradient update is not only dependant on the current step, but also takes into account the directions of previous steps, which can lead to better optimization performance. Similarly, Adam uses adaptive learning rates for each parameter, instead of a fixed, single learning rate for all. These modifications have led to significant performance improvements and make Adam one of the most popular optimization algorithms today.

1.3.3 Deep Neural Networks

Deep neural networks (DNNs) are driving today's most advanced machine learning systems [110]. Because of their success, research on DNNs and their applications has massively increased during the last decade. This led to major breakthroughs in complex problems like natural language processing and computer vision. However, DNNs are not a recent invention. The foundations for DNNs were laid over 60 years ago by Frank Rosenblatt, who described the perceptron algorithm, which is a binary classification algorithm. A perceptron calculates a weighted sum of an input vector and then applies an activation function to it, which either outputs 0 or 1. This algorithm can, however, only learn to classify linearly separable problems. In the following decade, multilayer pereceptrons (MLPs) were developed, which are constructed from multiple perceptrons and essentially represent feed forward neural networks (FFNNs), which are the most basic version of DNNs. The terms deep learning and artificial neural networks were introduced later, but essentially refer to MLPs as well. The reason why the field of deep learning has only now begun to produce such impressive results is usually attributed to the vast increase in computational power and the increasing availability of large and high quality

datasets.

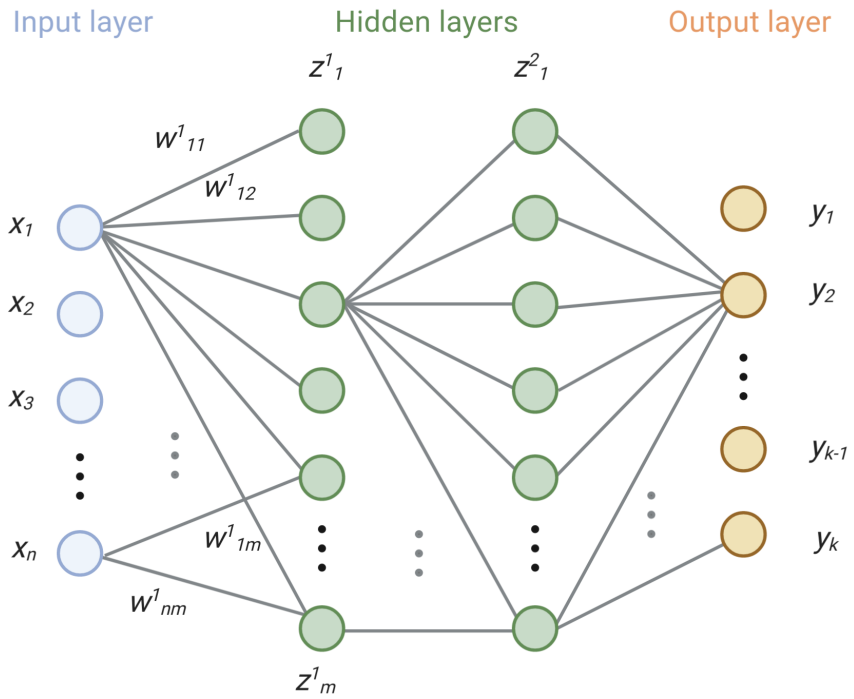

Figure 1.7: Graphical illustration of a deep neural network with two hidden layers. The input layer (blue) consists of n features $x_1 \ldots x_n$. Each node is connected to all nodes in the following layer via the weights w. Here, w^1_{nm} signifies the weight between the node x_n of the input layer and the node $m-th$ node of the first hidden layer, z^1_m. The nodes in the hidden layers (green) are connected in the same way. Finally, the last hidden layer, z^2, is connected to the output layer y (orange). Not all nodes and weights are shown, but indicated dots. Adapted from [27] (Created with BioRender.com)

A DNN consists of several units or neurons which are organized in multiple layers. An example is shown in Fig.1.7. Each neuron takes as input the weighted outputs of all neurons from the previous layer. The neuron then calculates the sum of these (weighted) inputs and applies an activation function to produce its output. The input data to a DNN is called the input layer and the last layer is called the output layer. For instance, the value z^1_j of the j-th neuron of the first hidden layer can be calculated as in (1.10).

$$z^1_j = \sum_i w_{ij} x_i \tag{1.10}$$

Where the variable w_{ij} depicts the value of the weight between the input feature x_i and

the hidden unit z_j^1. The value of z_j^1 is then calculated by an activation function $f(.)$:

$$z_j^1 = f(z_j^1) \tag{1.11}$$

This process is then repeated across all the DNN layers up to the output layer. One important feature of DNNs is that the activation function $f(.)$ must be a non-linear function, for instance a sigmoid function like the logistic function:

$$f(z) = \frac{1}{1+\exp^{-z}} \tag{1.12}$$

Today, the most popular activation function is the rectified linear unit (ReLU) [135]:

$$f(z) = max(0,z) \tag{1.13}$$

Depending on the optimization problem, a different or no activation function might be used for the output layer. For instance, a sigmoid function can be used in the case of a binary classification problem. For regression problems, usually no activation function is used for the output layer. For classification problems with more than two classes, the softmax function can be used, which enforces that all output values add up to 1, thus representing a normalized probability distribution:

$$f(z_i) = \frac{\exp^{z_i}}{\sum_{j=1}^{C} \exp^{z_j}} \tag{1.14}$$

Where C is the number of classes and z_i depicts the value of the i-th unit of the output layer.

1.3.4 Backpropagation

Through their complex structure of interconnected neurons, DNNs can approximate complex functions. In fact, a sufficiently large DNN can approximate any function [110]. Similar to linear regression, we have to adjust the weights of a DNN in order for it to learn a classification or regression problem such that it minimizes a loss function. To train a DNN using GD or variations of it, we have to calculate all partial derivatives of the loss function with respect to every weight. These derivatives can be efficiently calculated using the backpropagation algorithm, wich was proposed independently by several reasearch groups in the 1970ies. The algorithm and its potential for neural network training was popularized by a paper 1986 paper from Rumelhart, Hinton and Williams [165]. Backpropagation works by first calculating the partial derivatives of the loss function with respect to each node in the output layer. Then, the chain rule of derivatives is applied to backpropagate this error through the network and to calculate the partial

derivatives of the loss function with respect to every weight in the network. Because of its similarity to the forward pass, which calculates the activations of the neurons, backpropagation is also referred to as the backward pass. A drawback of DNN optimization using GD and backpropagation is that the optimization problem is non-convex, hence it contains multiple local minima in which the learning algorithm could potentially get stuck. However, because of the high-dimensionality of DNNs, most local minima are so-called saddle points of the loss function, which can be escaped. In practise, GD and backpropagation yield very good optimization results for DNNs and are the standard method for optimization today.

1.3.5 Regularization

One of the most important characteristics that any ML algorithm should have, is the ability to generalize to unseen data. When training a supervised learning algorithm, it is usually the goal to apply the trained algorithm to new data with unknown target values. For instance, one could train a classification algorithm on a dataset to predict whether a patient has a disease or whether they are healthy. The data from new patients might be generated slightly differently, e.g. at a different hospital or with a different machine. In this case, the distribution of the training data could be different from the distribution of the target data, leading to decreased prediction performance because the weights of the algorithm are trimmed to the training data distribution. Even when training and prediction data are from exactly the same distribution, the prediction performance can decrease on the prediction or test data. This effect is called overfitting, and methods that try to alleviate it are referred to as regularization.

According to the general understanding, strong overfitting occurs when the capacity of a ML model is too high for the task or training data. Here, the concept of capacity defines the ability of the model to approximate complex functions. A model with high capacity can learn many and very complex functions, while low-capacity models might only be able to learn simple functions. When trained on a simple task, a high capacity model can be overfitting by learning features that are specific to samples in the training data. This is visualized in Fig. 1.8.

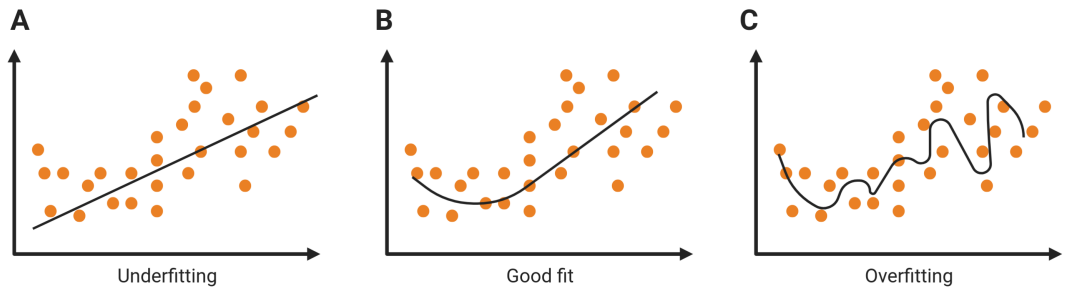

Figure 1.8: Exemplary illustration of model overfitting. Shown is a simple task of predicting a target value (y, vertical axis) given only a single feature (x, horizontal axis) **A** A simple linear model cannot fit the data because it there is a non-linear dependency between x and y. The model is underfitting. **B** An ideal model captures the non-linear structure and is able to nicely fit the data. This model would perform best on new data. **C** A model with too high capacity can overfit the data. This model will have better performance on the training data, but worse performance on new, unseen data. (Created with BioRender.com)

Whereas too complex models can be overfitting, models that are too simple can also underfit, meaning they do not achieve optimal performance even on the training data. This can be seen in Fig. 1.8A. For a given task, a ML practitioner therefore needs to find the optimal model, which neither under- nor overfits. The balance between underfitting and overfitting is also often called the bias-variance-trade-off. Underfitting models have too much bias and overfitting models have too much variance. Trying to reduce one of them usually means increasing the other, hence the trade-off.

Because of its centrality to ML, numerous methods for regularization have been developed. One way of regularizing a ML model is to add certain regularization terms to the loss function. Consider a squared error loss function of the form

$$L(y,x) = \sum_{i=n}^{N} (y_i - h_w(x_i))^2 \tag{1.15}$$

Here I have used $h_w(.)$ to denote the ML model with parameters w, for example a linear regression model. Given a set of input features x_i, this function tries to approximate the true output values y_i. N is the number of samples in the training dataset. One way of regularizing such a model is to penalize very large weights by adjusting the loss in (1.15):

$$L(y,x) = \sum_{i=n}^{N} (y_i - h_w(x_i))^2 + \lambda \sum_{n=1}^{N} |w_i| \tag{1.16}$$

Where λ is a regularization parameter that controls the strength of the regularization.

For any $\lambda > 0$, large values for w will increase the loss function and are thus less favorable when the objective is to minimize the loss function. This is called L1 regularization. Linear regression with L1 regularization is also called LASSO regression. The intuition behind L1 regularization is that complex models have larger weights, and therefore penalizing large weights will lead to more simple models with better generalization capabilities. A similar regularization method is L2 regularization, for which the loss function is modified as follows:

$$L(y,x) = \sum_{i=n}^{N} (y_i - h_w(x_i))^2 + \lambda \sum_{n=1}^{N} w_i^2 \tag{1.17}$$

The only difference to L1 regularization is that the squared weight vector is used in the added regularization term.

While L1 and L2 regularization work great for many models, other techniques are more common for the regularization of DNNs. One of the most important and most frequently used methods is dropout regularization [79]. Dropout regularization works by randomly removing a specified fraction of neurons from each layer during every training step. Usually, input and output layers are held fixed. The intuition behind dropout is to prevent the DNN from relying strongly on specific neurons, because they might be unavailable during training when dropped out. It also prevents co-adaption of neurons, which is the strong dependence of neurons in a layer on specific neurons in a previous layer. By forcing the DNN not to rely on few single neurons, it learns internal feature representations that are generally useful to solving the problem. Dropout is nowadays widely used in many large deep learning models, which often have thousands of neurons in each layer.

Another form of regularization, that is especially used in computer vision models, is data augmentation. The term basically refers to manipulating the training dataset in such a way as to introduce slight variations, thereby artificially increasing the training data size. This works because it is generally known that large datasets improve generalization capabilities as they make it harder for a model to overfit. To augment data, one can inject, for instance, random gaussian noise to the input data, or change the orientation of input images. However, data augmentation must be done with care, as it is not always obvious which data manipulations help with generalization while not changing the input data too much.

Finally, when using some variant of SGD to train a large DNN, one can stop the training process before converging on the minimal loss value. By stopping the training process early, a sub-optimal performance on the training dataset is achieve. However, this can effectively prevent overfitting. Intuitively, in the early training process, a DNN learns general features that allow the loss function to decrease rapidly. Later in the training process, weight adjustments only lead to smaller decreases of the loss and these adjustments

can often be data- or even sample-specific. Early stopping is commonly used nowadays. To determine the optimal point of stopping, typically the data is split in training, validation and test dataset. The validation dataset is not used for training, but for evaluation of model generalization throughout the training process. Once the loss only decreases for the training dataset but not for the validation dataset, training should be stopped. The test dataset is only used once for the final method .

1.3.6 Deep Learning in Biology

Driven by the successes of deep learning-based systems in computer vision and natural language processing and by the rapidly growing amount of data in biology, many researchers have applied deep learning to biological problems [58]. For example, Kelley and colleagues trained a deep learning system on large amounts of sequence data to predict chromatin accessibility which surpassed the state-of-the-art (SOTA) [91]. Similar models have been used to predict enhancers [116] or to predict the effect of genetic variants on expression and disease risk [226]. Even the effects of variants in noncoding sequences can be predicted with such models [227]. Poplin and colleagues have introduced an algorithm for single nucleotide polymorphism (SNP) calling that is based on DNNs [150]. In their study, the authors show that their method exceeds the performance of other SOTA tools. Another area where deep learning methods have been used is the prediction of miRNA targets based on sequence and other informative features [42]. Raw data from Oxford Nanopore sequencing machines - a specialized sequencing technology for long reads - has been transformed to DNA sequence using Hidden Markov Models (HMMs) for some time. Nowadays, DNN-based methods have greatly surpassed the accuracy of HMMs, thereby improving the quality of the sequencing data [209]. Finally, numerous algorithms have been developed for scRNA-seq data, which is ideally suited for ML algorithms because of the high sample numbers generated for single experiments. Marouf and colleagues have used DNN-based generative models to generate realistic data when trained on scRNA-seq datasets [123] and Lotfollahi and colleagues have used generative deep learning models to predict the results of single-cell perturbations [118].
This list is by no means extensive and countless other methods have been developed which I am not mentioning here. With the mentioned examples I merely want to demonstrate that deep learning, besides leading to groundbreaking advances in computer vision and natural language processing, has also started to transform and improve algorithms in computational biology.

1.4 Aims of the thesis

As I have described in some detail in this chapter, great advances have been made in our understanding of FTD pathology, genetics and disease mechanisms during recent decades. Nevertheless, there does not exist a viable treatment option to this date. In order to develop remedies, it is required to further advance and deepen our knowledge of diesase mechanisms and pathways, and to identify new potential drug targets that can be tested. The goal of this thesis was to integrate and analyse multi-omics FTD-related datasets and contribute to FTD research.

The majority of the data analysed in this thesis was generated from post-mortem human brain tissue, which depicts a complex mixture of cell types. Differences in cell type compositions are therefore a major potential bias in the data that has to be corrected for. In chapter 2, I have addressed this issue by developing a novel, deep learning-based algorithm for cell type deconvolution which I have tested on post-mortem human brain tissue data. In chapters 3 and 4, I have used this algorithm and various other computational methods to extract meaningful insights from diverse genomics datasets related to FTD.

Chapter 2

Deep learning-based cell composition analysis from tissue expression profiles

Disclaimer

This chapter was published in similar form in *Science Advances* at 22.07.2020 under the title "Deep learning-based cell composition analysis from tissue expression profiles" [126]. I initiated the project together with Peter Heutink and Stefan Bonn, who also helped in algorithm design and general project design as well as manuscript writing. Mohamed Marouf and Sergio Oller helped with model selection and creating the python package. Anupriya Dalmia helped with data processing. Katrin Kloiber and Peter Heutink helped with manuscript writing. Daniel Sumner Magruder and Sergio Oller designed the web application.

2.1 Abstract

Cell type deconvolution of tissue expression profiles is an important tool of modern transcriptomics. However, current deconvolution algorithms are not designed to take full advantage of the diverse range of scRNA-seq datasets available today. To address this problem, we have developed a single cell-assisted deconvolutional neuronal network (Scaden), that uses gene expression information from multiple sources to infer the cellular composition of tissues. Scaden is trained on scRNA-seq data to engineer discriminative features that confer robustness to bias and noise, making complex data preprocessing and feature selection unnecessary. We show that Scaden outperforms current state-of-the-art algorithms on multiple datasets, stemming from various tissues and experimental methods. Scaden leverages training data from multiple datasets to build robustness against dataset-specific bias. Scaden is published as a fully open source software package and as a web application, making it accessible to scientists from diverse backgrounds.

2.2 Introduction

The analysis of tissue-specific gene expression using next-generation sequencing technologies like RNA-seq is a centerpiece of the molecular characterization of biological and medical processes [80]. A well-known limitation of tissue-based RNA-seq is that it typically measures average gene expression across many molecularly diverse cell types that can have distinct cellular states [55]. A change in gene expression between two conditions can therefore be attributed to a change in the cellular composition of the tissue or a change in gene expression in a specific cell population, or a mixture of the two. To deconvolve the cell type composition from a change in gene expression is especially important in systems with cellular proliferation (e.g., cancer) or cellular death (e.g., neuronal loss in neurodegenerative diseases) due to systematic cell population differences between experimental groups [105].

To account for this problem, several computational cell deconvolution methods have been proposed during the last years [15, 130]. These algorithms use gene expression profiles (GEPs) of cell type–specifically expressed genes to estimate cellular fractions using linear regression to detect, interpret, and possibly correct for systematic differences in cellular abundance between samples [15]. While the best-performing linear regression algorithms for deconvolution seem to be variations of support vector regression [139, 132, 140, 205, 62], the selection of an optimal GEP is a field of active research [132]. It has been recently shown that the design of the GEP is the most important factor in most deconvolution methods, as results from different algorithms strongly correlate given the same GEP [195].

In theory, an optimal GEP should contain a set of genes that are predominantly expressed within each cell population of a complex sample [202]. They should be stably expressed across experimental conditions, for example, across health and disease, and resilient to experimental noise and bias. However, bias is typically inherent to biomedical data and is imparted, for instance, by intersubject variability, variations across species, different data acquisition methods, different experimenters, or different data types. The negative impact of bias on deconvolution performance can be partly improved by using large, heterogeneous GEP matrices [195]. It is therefore not surprising that recent advancements in cell deconvolution relied almost exclusively on sophisticated algorithms to normalize the data and engineer optimal GEPs [132].

While GEP-based approaches lay the foundational basis of modern cell deconvolution algorithms, we hypothesize that deep neural networks (DNNs) could create optimal features for cell deconvolution, without relying on the complex generation of GEPs. DNNs such as multilayer perceptrons are universal function approximators that achieve state-of-the-art performance on classification and regression tasks. Whereas this feature is of little importance for strictly linear input data, it makes DNNs superior to linear regres-

sion algorithms as soon as data deviate from ideal linearity. This means, for instance, that as soon as data are noisy or biased and classical linear regression algorithms may falter, the hidden layer nodes of the DNN learn to represent higher-order latent representations of cell types that do not depend on input noise and bias. We theorize, therefore, that by using gene expression information as network input, hidden layer nodes of the DNN would represent higher-order latent representations of cell types that are robust to input noise and technical bias.

An obvious limitation of DNNs is the requirement for large training data to avoid overfitting of the machine learning model. While ground-truth information on tissue RNA-seq cell composition is scarce, one can use single-cell RNA-seq (scRNA-seq) data to obtain large numbers of in silico tissue datasets of predefined cell composition [176, 92]. We do this by subsampling and subsequently merging cells from scRNA-seq datasets, this approach being limited only by the availability of tissue-specific scRNA-seq data. It is to be noted that scRNA-seq data suffer from biases, such as dropout, to which RNA-seq data are not subject to [78]. While this complicates the use of scRNA-seq data for GEP design [205], we surmise that latent network nodes could represent features that are robust to these biases.

On the basis of these assumptions, we developed a single cell–assisted deconvolutional DNN (Scaden) that uses simulated bulk RNA-seq samples for training and predicts cell type proportions for input expression samples of cell mixtures. Scaden is available as downloadable software package and web application (https://scaden.ims.bio). Scaden is trained on publicly available scRNA-seq and RNA-seq data, does not rely on specific GEP matrices, and automatically infers informative features. Last, we show that Scaden deconvolves expression data into cell types with higher precision and robustness than existing methods that rely on GEP matrices.

2.3 Results

2.3.1 Scaden overview, model selection and training

In this part, we focus on the design and optimization of Scaden by training, validation, and testing on in silico data. Note that the generation of in silico data is a strictly linear mathematical operation. Our aim in this context, to corroborate Scaden's basic functionality, is to show that Scaden's performance compares with (but not necessarily exceeds) that of state-of-the-art algorithms.

The basic architecture of Scaden is a DNN that takes gene counts of RNA-seq data as input and outputs predicted cell fractions (Fig. 2.1). To optimize the performance of the DNN, it is trained on data that contain both the gene expression and the real cell type fraction information (Fig. 2.1a). The network then adjusts its weights to minimize

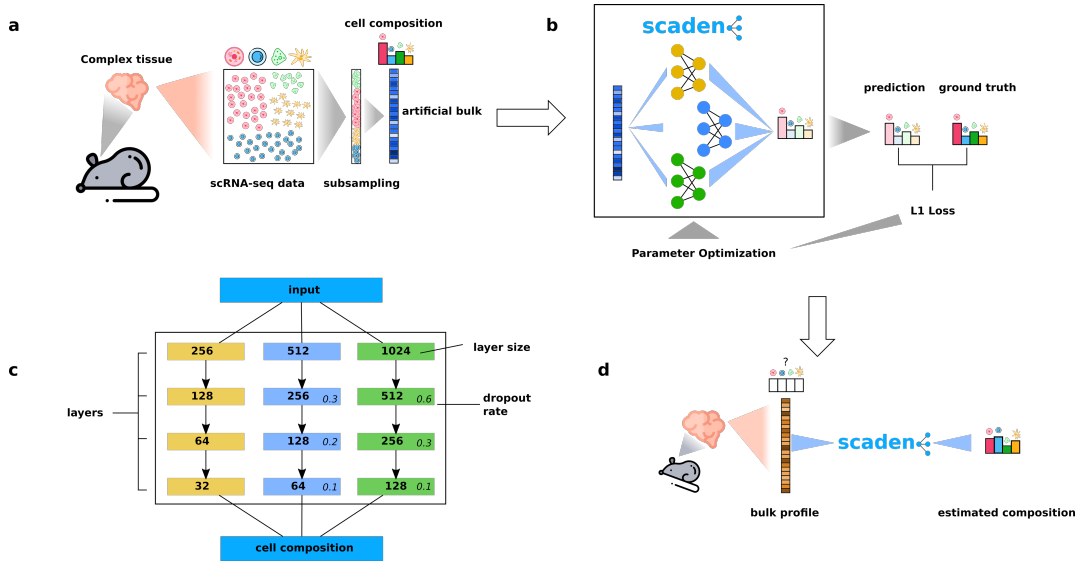

Figure 2.1: Overview of Scaden algorithm. **a** Data from scRNA-seq experiments is analyzed, normalized and labeled per cell type. Expression profiles of cells are then subsampled to yield artificial bulk gene expression samples of known cell composition. **b** A Scaden ensemble model is trained using the generated labeled training data. **c** Model architecture of Scaden. Scaden consists of three separate DNNs with four hidden layers each, of varying layer sizes and dropout rates. The final predictions are the average of all three models. **d** A trained Scaden model can then be used to estimate the unknown cellular composition of a bulk tissue expression sample.

the error between the predicted cell fractions and the real cell fractions (Fig.2.1b-d). We restricted feature selection to the removal of "uninformative" genes that have either zero expression or an expression variance below 0.1, leaving 10,000 genes for training. In our hands, this feature selection step decreases training time and memory usage.

For the model selection and training, we made use of the large numbers of artificial bulk RNA-seq datasets with defined composition that can be generated in silico from published scRNA-seq and RNA-seq datasets (simulated tissues; Fig. 2.1A and tables 2.1 and 2.2). The only constraint is that the scRNA-seq and RNA-seq data must come from the same tissue as the bulk data subject to deconvolution.

To find the optimal DNN architecture for cell deconvolution, we generated bulk peripheral blood mononuclear cell (PBMC) RNA-seq data from four publicly available scRNA-seq datasets (tables 2.1 and 2.3). We performed leave-one-dataset-out cross-validation, training Scaden on mixtures of synthetic datasets from three scRNA-seq datasets and evaluating the performance on simulated tissue from a fourth scRNA-seq dataset.

We used the root mean square error (RMSE), Pearson's correlation coefficient (r), the slope and intercept of the regression fitted for ground-truth and predicted cell fractions, and Lin's concordance correlation coefficient (CCC) (17) to assess algorithmic performance. The CCC is a measure sensitive not only to scatter but also to deviations from linearity (slope and intercept). Within the main text, we report on CCC and RMSE values only; other metrics can be found in the Supplementary Materials.

The final Scaden model is an ensemble of the three best-performing models (table 2.4), and the final cell type composition estimates are the averaged predictions of all three ensemble models (Fig. 2.1). Using an ensemble of models increased the deconvolution performance as compared to single best models (table 2.6). Details of the model and hyperparameters are given in table 2.5. We also evaluated the effect of the size of the training dataset on Scaden deconvolution performance, repeating leave-one-dataset-out cross-validation on PBMC data with training dataset sizes from 150 up to 15,000 samples (Fig. 2.2B). The increase in CCC value starts to level off from about 1500 simulated samples for this dataset but continues to increase slowly with sample size. We specifically addressed the question to what degree the DNN, trained on simulated samples, tends to overfit, failing to generalize to real bulk RNA-seq data. To understand after how many steps a model trained on in silico data overfits on real RNA-seq data, we trained Scaden on simulated data from an ascites scRNA-seq dataset (table 2.1; 6000 samples) and evaluated the loss function on a corresponding annotated RNA-seq dataset (18) (table 2.2; three samples) as a function of the number of steps (Fig. 2.2). All models converged after approximately 5000 steps and slightly overfit when trained for longer. On the basis of this result, we opted for an early-stop approach after 5000 steps for evaluation on real bulk RNA-seq data.

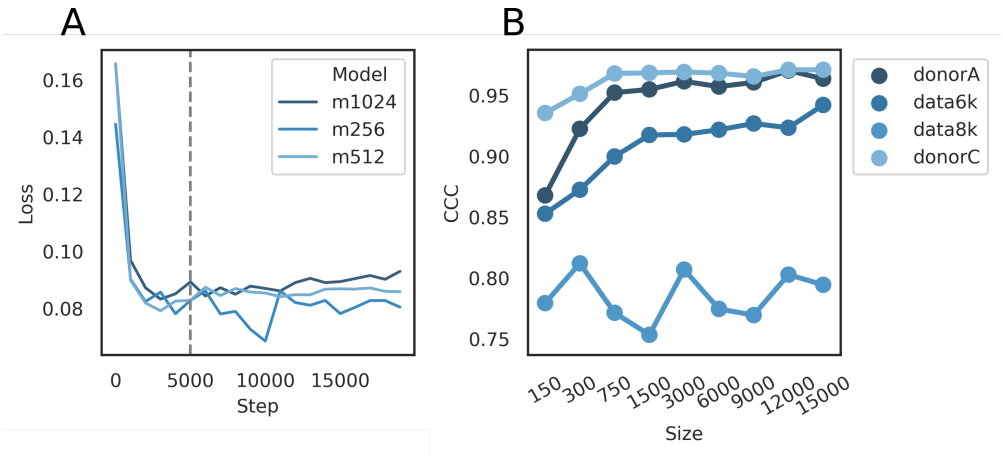

Figure 2.2: Effect of long trainnig and dataset size on Scaden performance. **A** Loss (mean absolute error, y-axis) of different Scaden models with respect to training steps (x-axis) for a mini-batch size of 32. **B** Performance of Scaden ensemble (CCC, y-axis) for increasing training dataset sizes (x-axis). Performance was measured using leave-one-dataset-out cross validation: Scaden was trained on three datasets and tested on the left-out dataset.

2.3.2 Comparison of deconvolution algorithms on simulated data

We then compared Scaden to four state-of-the-art GEP-based cell deconvolution algorithms, CIBERSORT (CS) (6), CIBERSORTx (CSx) (7), Multi-subject Single Cell deconvolution (MuSiC) (8), and Cell Population Mapping (CPM) (9). While CS relies on hand-curated GEP matrices, CSx, MuSiC, and CPM can generate GEPs using scRNA-seq data as input.

To get an initial estimate of Scaden's deconvolution fidelity, we trained the model on 24,000 simulated PBMC RNA-seq samples from three datasets and tested its performance in comparison to CS, CSx, MuSiC, and CPM on a fourth dataset of 500 samples each (e.g., training on data6k, data8k, and donorA and evaluation on donorC). We used corresponding scRNA-seq datasets for the construction of GEPs as input for CSx and MuSiC, and CPM. For CS, we used the PBMC-optimized LM22 GEP matrix (6), which was developed by the CS authors for the deconvolution of human PBMC data.

For two of four test datasets (donorA and donorC), Scaden obtained the highest CCC and lowest RMSE, followed by CSx, MuSiC, CS, and CPM (fig. 2.3 and table 2.7). CSx and MuSiC obtained the highest CCC values for the data8k and data6k datasets, respectively. Scaden obtained the highest average CCC and lowest RMSE (0.88 and 0.08, respectively), followed by MuSiC (0.85 and 0.10), CSx (0.83 and 0.11), CS (0.63 and 0.15), and CPM (0 and 0.20). As expected, all algorithms that use scRNA-seq data

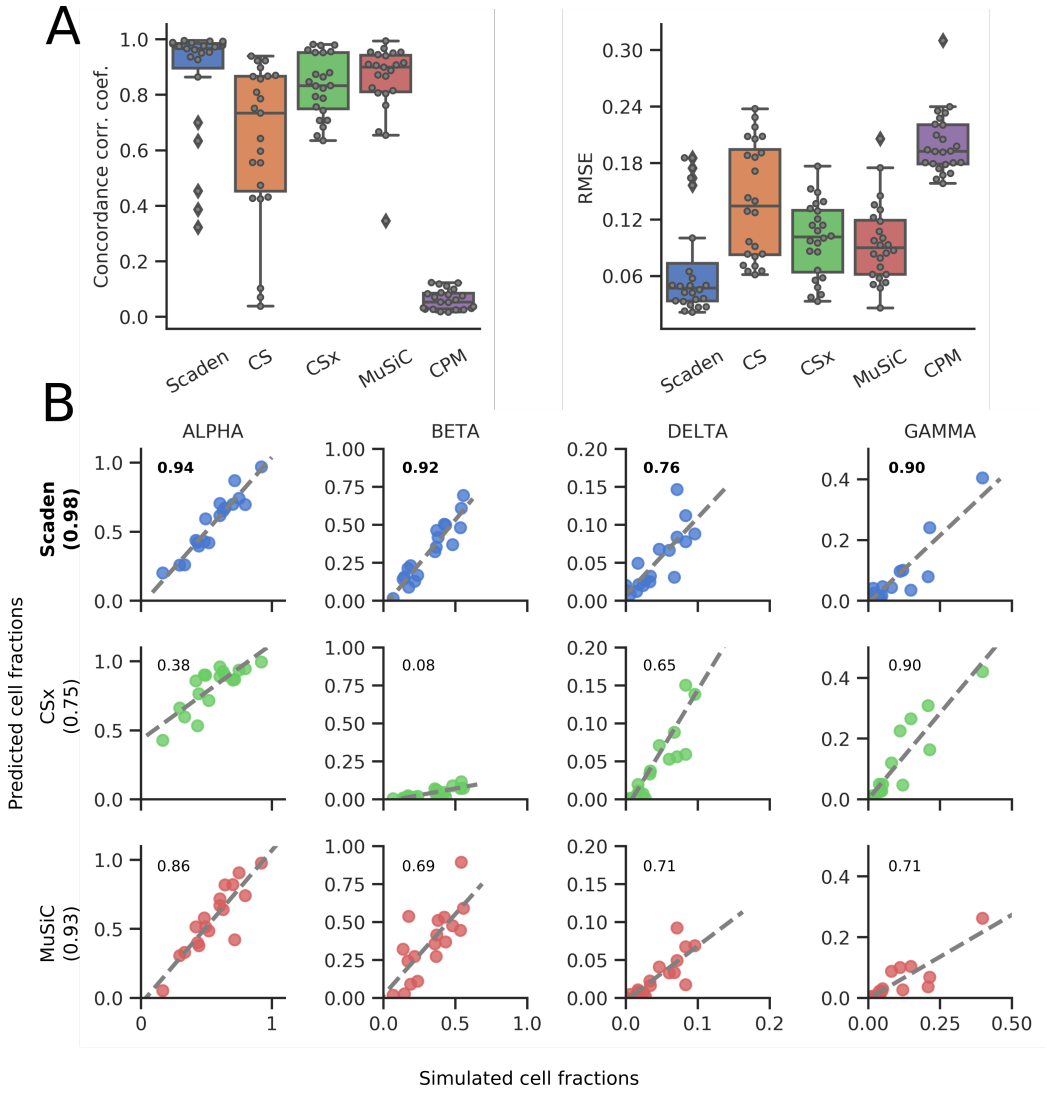

Figure 2.3: **A** Boxplots of the cell type prediction CCC and RMSE for four simulated PBMC datasets. Tables S14 and S16 contain information on the five (six for CS) cell types used. **B** Scatterplots for four pancreas cell types of ground-truth (x axis) and predicted values (y axis) for Scaden, CSx, and MuSiC on artificial pancreas data (20). Numbers inside the plotting area and in parenthesis signify CCC values.

as reference performed well, with the notable exception of CPM. We want to mention that CPM focuses on the reconstruction of continuous spectra of cellular states, while it incorporates cell deconvolution as an additional feature. We therefore report CPM's deconvolution performance in the Supplementary Materials from here on. On average, Scaden also obtained the highest correlation and the best intercept and slope values on simulated PBMC data (table 2.7). A closer inspection on a per–cell type basis (Fig. 2.3A) revealed that Scaden yields consistently higher CCC values and lower RMSEs when compared to the other algorithms.

A specific feature of the MuSiC algorithm is that it preferentially weighs genes according to low intersubject and intracell cluster variability for its GEP, which increases deconvolution robustness when high-expression heterogeneity is observed between human participants, for example (8). To understand whether Scaden can use multisubject information to increase its deconvolution performance, we trained Scaden, CSx, and MuSiC on scRNA-seq pancreas data from several participants (19) and assessed the performance on a separate simulated pancreas RNA-seq dataset (20). To allow for direct comparison, we chose the same pancreas training and test datasets that were used in the original MuSiC publication (table 2.1). To enable Scaden to leverage the heterogeneity of multisubject data, training data were generated separately for every participant in the dataset (see Methods). CSx cannot profit from multisubject data but performed well on the artificial PBMC datasets and was therefore included in the comparison. The best average performance (across cell types) is achieved by Scaden (CCC = 0.98), closely followed by MuSiC (CCC = 0.93), while CSx does not perform as well (CCC = 0.75; Fig. 2.3B and table 2.8). On a per–cell type basis, Scaden's predictions are clearly superior to the other two algorithms for all cell types. This provides strong evidence that Scaden, by separating training data generation for each participant, can learn intersubject heterogeneity and outperform specialized multisubject algorithms such as MuSiC on the cell type deconvolution task.

In addition, we wanted to test how the best-performing deconvolution algorithms Scaden, MuSiC, and CSx behave when unknown cell content is part of the mixture. To test this, all cells falling into the "Unknown" category were removed from the training or reference PBMC datasets but added to the simulated mixture samples at fixed percentages (5, 10, 20, and 30%; see Methods). Scaden obtains the highest CCC for all tested percentages of unknown cell content (fig. 2.5 and table 2.9). The general deconvolution performance declines linearly with increasing percentage of unknown content for all tested algorithms, indicating that Scaden, MuSiC, and CSx have a similar robustness against unknown mixture content.

We next compared the runtime and memory footprint of Scaden and MuSiC on an Intel Xeon six-core central processing unit (CPU) to the runtime of the CSx web application. Scaden is the only algorithm that requires the generation of in silico training data, which

takes 13 min for 2000 samples with a peak memory usage of 8 GB. Similar values were obtained for the human brain data. Next, we used the PBMC data to benchmark the runtime and memory consumption of the deconvolution task. For Scaden, model training took 11 min and cell fraction prediction 8 s for 500 samples, using less than 1-GB memory. We used the web application of CSx with batch correction to deconvolve the 500 PBMC samples in 35 min. MuSiC took only 2 min and 15 s to deconvolve all 500 samples, with the memory usage peaking at 4.5 GB. As Scaden can take advantage of a graphics processing unit (GPU), we additionally compared training duration on an AMD Ryzen 5 2600 CPU and GeForce RTX 2600 GPU on the same machine. Training on the CPU took 9 min and 39 s, while it took only 3 min and 2 s on the GPU, corresponding to a roughly three times shorter runtime for Scaden if a GPU is available.

2.3.3 Robust deconvolution of bulk expression data

The true use case of cell deconvolution algorithms is the cell fraction estimation of tissue RNA-seq data. In particular for noisy and biased bulk RNA-seq data, we hypothesize that Scaden's latent feature representations might help it to more robustly predict cell fractions as compared to GEP-based algorithms.

We therefore assessed the performance of Scaden, CS, CSx, and MuSiC to deconvolve two publicly available human PBMC bulk RNA-seq datasets, for which curated GEP matrices and RNA-seq data with associated ground-truth cell type compositions from flow cytometry are available (see the "Data availability" section). We will refer to these datasets that consists of 12 samples each as PBMC1 (21) and PBMC2 (10) (table 2.2). Both datasets have similar cell type compositions across samples, with CD4 and CD8 T cells making up the biggest fractions. Deconvolution for all methods was performed as described in the previous section, with the difference that data from all four PBMC scRNA-seq datasets were now deployed for Scaden training. Results are given in Fig. 2.4 (A to C) and tables 2.10 and 2.11.

On the PBMC1 dataset and using all cell types, Scaden obtained the highest CCC and lowest RMSE (0.56 and 0.13), while CSx (0.55 and 0.16) and CS (0.43 and 0.15) performed well yet notably worse than Scaden (Fig. 2.4A and tables 2.10 and 2.11). CPM (0 and 0.18) and MuSiC (0.19 and 0.32) both failed to deconvolve the cell fractions of the PBMC1 data. Scaden also obtained the best CCC and RMSE (0.68 and 0.08) on the PBMC2 dataset, while CS (0.58 and 0.10) and CSx (0.42 and 0.13) obtained good deconvolution results. Similar to the PBMC1 data deconvolution results, CPM (0.16 and 0.11) and MuSiC (0.13 and 0.30) did not perform well on the PBMC2 deconvolution task. In addition to CCC and RMSE metrics, Scaden achieves the best correlation, intercept, and slope on both PBMC datasets.

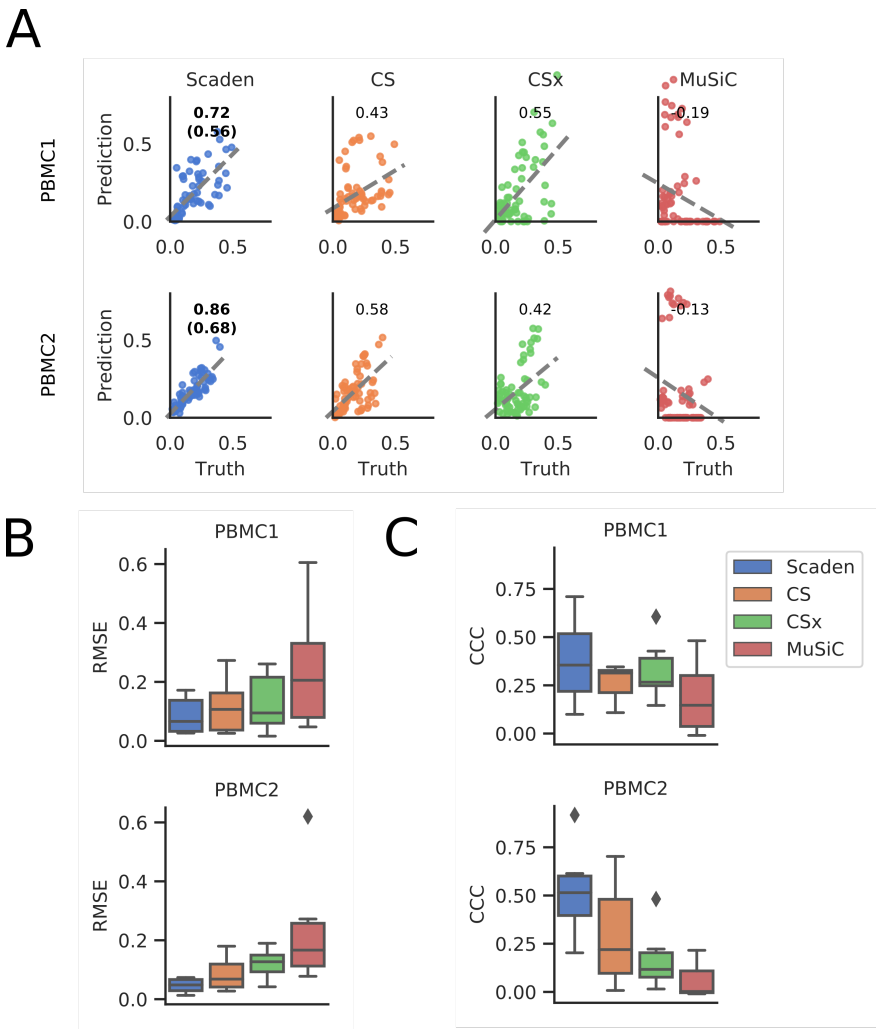

Figure 2.4: **A** Per–cell type scatterplots of ground-truth (x axis) and predicted values (y axis) for Scaden, CS, CSx, and MuSiC on real PBMC1 and PBMC2 cell fractions. Numbers inside the plotting area signify CCC values. For Scaden, the CCC using only scRNA-seq training data is shown in parenthesis, and the CCC using mixed scRNA-seq and RNA-seq training data is shown without parentheses. **B** Boxplots of RMSE values for real PBMC1 and PBMC2 data. **C** CCC values for real PBMC1 and PBMC2 data.

In particular, Scaden outperforms classical algorithms on a per–cell type basis (Fig. 2.4, B and C). These results show weaker correlations and a strong dependence on the cell type. A closer examination of the metrics in table 2.11 shows that the largest variations are found in the slope and intercept.

We further evaluated how good the Scaden ensemble performs compared to the best single DNN model (M512, 512 nodes input layer). While the M512 model shows good deconvolution performance on the PBMC1 (CCC, 0.57) and PBMC2 (CCC, 0.68) datasets, the ensemble model achieves the best average cross-validation performance (table 2.6). We therefore opted to use the ensemble method to reduce interdataset performance variation observed with M512 and other single models.

An additional algorithmic feature of Scaden is that it seamlessly integrates increasing amounts of training data, which can be of different types, such as a combination of simulated tissue and real tissue data with cell fraction information. In theory, even limited real tissue training data could make Scaden robust to data type bias and consequently improve Scaden's deconvolution performance on real tissue data. We therefore trained Scaden on a mix of simulated PBMC and real PBMC2 (12 samples) data and evaluated its performance on real PBMC1 data (Fig. 2.4, A and B and tables 2.10 and 2.11). While the training contained very little (2%) real data, Scaden's CCC increased from 0.56 to 0.72, and the RMSE decreased from 0.13 to 0.10. We observed similar performance increases when Scaden was trained on simulated PBMC and real PBMC1 data and evaluated on real PBMC2 data (Fig. 2.4, A and B, tables 2.10 and 2.11). Next, we wanted to investigate how a Scaden model trained on only few real samples compares to the models trained on simulated or simulated and real data. While a Scaden model trained on only bulk PBMC1 samples (n = 12) deconvolves PBMC2 data with a CCC of 0.62, it does not reach the CCC of models trained on simulated data (CCC of 0.68) or on simulated and bulk data (CCC of 0.86). We would also not advise training models on so few training samples, as these models are usually overfit.

This further validates that Scaden reliably deconvolves tissue RNA-seq data into the constituent cell fractions and that very accurate deconvolution results can be obtained if reference and target datasets are from the same experiment.

We next wanted to test how the algorithm performs on postmortem human brain tissue of a subsample from the Religious Orders Study and Memory and Aging Project (ROSMAP) study [3], for which ground-truth cell composition information was recently measured by immunohistochemistry (41 samples with all cell types given) [145]. The data provided by this study consist of bulk RNA-seq data from the dorsolateral prefrontal cortex and pose a special challenge due to the complexity of its cell type composition, which is further complicated by the fact that the data originate from brains of healthy individuals as well as patients with Alzheimer's disease (AD) at various stages of neuronal

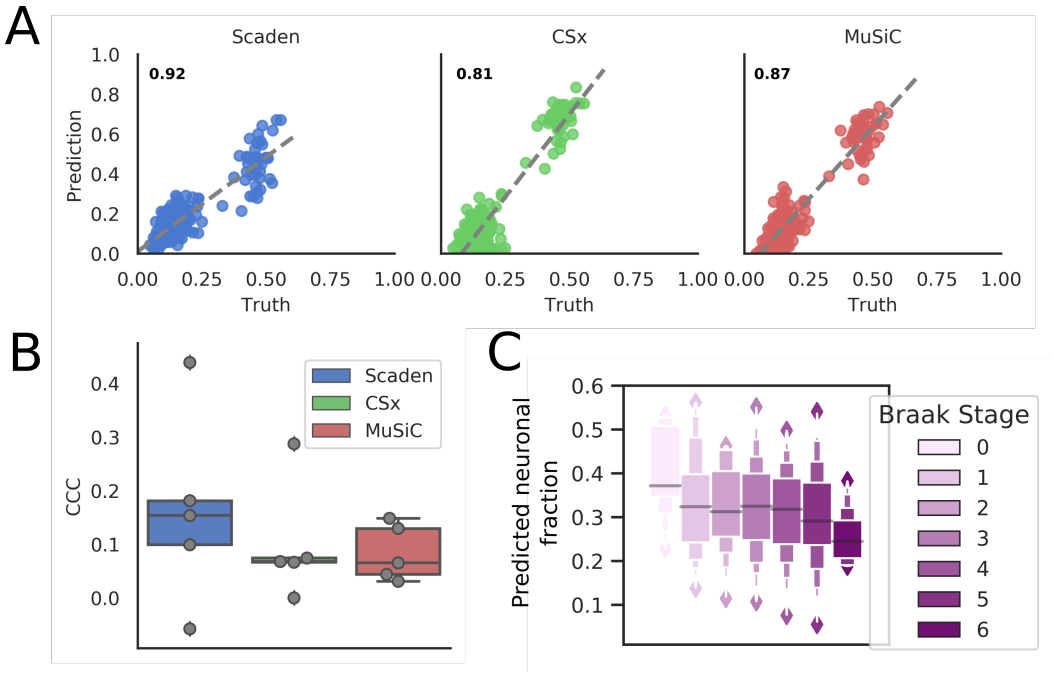

Figure 2.5: **A** Prediction of human brain cell fractions of the ROSMAP dataset using the Darmanis dataset as a reference: scatterplots of ground-truth (x axis) and predicted values (y axis) for Scaden, CSx, and MuSiC of data. CCC values are shown as inserts. **B** Per–cell type CCC values for ROSMAP using the Darmanis data as a reference. **C** Neuronal content determined by Scaden trained on mouse brain data and evaluated on the Braak stage of the ROSMAP study.

loss. As reference datasets, we used the scRNA-seq dataset provided by Darmanis et al. [46] from the anterior temporal lobe of living patients and the Lake dataset that isolates nuclei of neurons from two (visual and frontal) cortical regions from a postmortem brain and subjects them to RNA-seq [106]. From these, we generated 2000 training samples (Darmanis) and 4000 samples (two regions from the Lake dataset).

Figure 2.5 A shows the deconvolution results for all three algorithms with the Darmanis (scRNA-seq) reference dataset. Scaden achieves the highest CCC value (0.92) followed by MuSiC (0.87) and CSx (0.81; table 2.12). Compared to Scaden, MuSiC and CSx overestimate neural percentages, leading to higher RMSE values of 0.09 and 0.12, respectively (Scaden, 0.06). Notably, all methods showed a lower CCC on the per–cell type level (Fig. 2.4 B), demonstrating that some per–cell type correlations are poor, either in slope, intercept, variance, or a combination of them. This emphasizes the need for a cell type–specific inspection of results and highlights that, depending on the dataset, cell type–specific deconvolution results can be far from perfect.

In addition to comparing the predictive power of Scaden, CSx, and MuSiC on human

brain tissue with different reference datasets, we also tested how the choice of reference datasets affected Scaden's deconvolution results. Notably, all methods substantially drop in performance when the Lake single-nucleus RNA-seq dataset is used as a reference as we had presumed (fig. 2.6 A). We want to emphasize that Scaden, in contrast to CSx and MuSiC, has the possibility to simultaneously use both datasets as reference, whereas for CSx and MuSiC, the user has to choose one of the two, unaware of which will give the correct results.

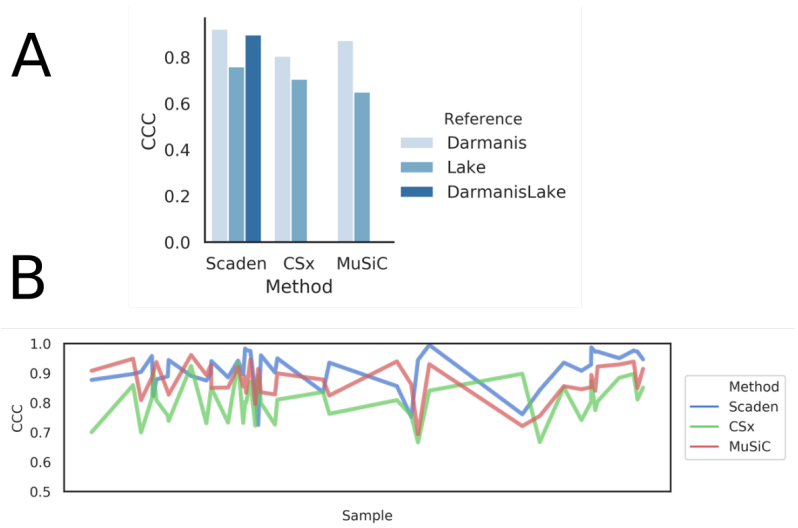

Figure 2.6: Deconvolution performance comparison on post-mortem human brain data from ROSMAP study. A)CCC valuesfor prediction of post-mortemhuman brain (ROSMAP) cell fractions for Scaden , CSx, and MuSiC on the reference data sets Darmanis (lightblue, leftmost bars) and Lake (medium-shaded blue), as well as both (for Scaden only, dark blue).Interestingly, addition of the Lake dataset into training data affected Scaden performance only slightly. B)Per sample results for prediction of postmorten human brain (ROSMAP) cell fractions on the Darmanis reference datasets using Scaden, CSx, and MuSiC

We found that the performance of Scaden was almost unaffected when the Lake dataset was added to the Darmanis training samples (CCC = 0.90, RMSE = 0.06; fig. 2.6A and table 2.12). These results show that cell deconvolution with Scaden is robust to training data bias (Darmanis single-cell versus Lake single-nucleus data). An added benefit of Scaden is that it allows for the inclusion and mixing of different scRNA-seq experiments in the training dataset, further increasing its robustness (fig. 2.6A). Last, when calculating the CCC values on a per-sample basis, Scaden achieves the best scores for most samples (fig. 2.6B).

In a next step, we wanted to assess whether Scaden's deconvolution performance was robust across species by trying to predict the cell fractions of the ROSMAP study [3] with a Scaden model trained on in silico data from five mouse brain scRNA-seq datasets (table 2.1). Intriguingly, Scaden was able to achieve a CCC value of 0.83 and an RMSE of 0.079, showing that Scaden can reliably deconvolve RNA-seq data across related species.

The ROSMAP study also contains information on the Braak stages [30] corresponding to 390 human postmortem prefrontal cortex samples, which correlate with the severity and progression stage of AD and the degree of neuronal loss. We used the Scaden model trained on artificial data generated from five mouse brain scRNA-seq datasets to predict neuronal cell fractions of this larger human dataset. Overall, Scaden's cell fraction predictions capture the increased neuronal loss with increasing Braak stage (Fig. 2.5 C). The largest drop in neural percentage is observed at stage 5, when the neurodegeneration typically reaches the prefrontal cortex of the brain.

Given the robustness with which Scaden predicts tissue RNA-seq cell fractions using scRNA-seq training data, even across species, we next wanted to investigate whether an scRNA-seq–trained Scaden model can also deconvolve other data types. To this end, we measured the deconvolution performance on a bulk PBMC microarray dataset (20 samples) [139] of a Scaden model trained on scRNA-seq and RNA-seq PBMC data (see above). We compared Scaden to CS using the microarray-derived LM22 matrix. CS achieved a slightly higher CCC and slightly lower total RMSE (0.72 and 0.11) than Scaden (0.71 and 0.13), while Scaden obtained the highest average CCC (0.50) compared to CS (0.39; table 2.13). Notably, in this scenario, Scaden was trained entirely on simulated scRNA-seq and RNA-seq data, while CS's LM22 GEP was optimized on PBMC microarray data.

Overall, we provide strong evidence that Scaden robustly deconvolves tissue data across tissues, species, and even data types.

2.4 Discussion

Scaden is a novel deep learning–based cell deconvolution algorithm that, in many instances, compares favorably in both prediction robustness and accuracy to existing deconvolution algorithms that rely on GEP design and linear regression. We believe that Scaden's performance relies to a large degree on the inherent feature engineering of the DNN. The network does not only select features (genes) for regression but also creates new features that are optimal for the regression task in the nodes of the hidden layers. These hidden features are nonlinear combinations of the input features (gene expression), which makes it notoriously difficult to explain how a DNN works [220]. It is important to highlight that this feature creation is fundamentally different from all other existing

cell deconvolution algorithms, which rely on heuristics that select a defined subset of genes as features for linear regression.

Another advantage of this inherent feature engineering is that Scaden can be trained to be robust to input noise and bias (e.g., batch effects). Noise and bias are all prevalent in experimental data, because of different sample quality, sample processing, experimenters, and instrumentation, for example. If the network is trained on different datasets of the same tissue, however, then it learns to create hidden features that are robust to noise and bias, such as batch effects. This robustness is pivotal in real-world cell deconvolution use cases, where the bulk RNA data for deconvolution and the training data (and therefore the network and GEP) contain different noise and biases. In this study, we tested Scaden with training data from scRNA-seq datasets generated with a variety of different protocols and could not identify a specific protocol that is not suitable. While especially recent cell deconvolution algorithms include batch correction heuristics before GEP construction, Scaden optimizes its hidden features automatically when trained on data from various batches. Potential protocol-specific biases can therefore be alleviated when employing training data from multiple protocols.

The robustness to noise and bias, which might be due to hidden feature generation, is especially evident in Scaden's ability to deconvolve across data types. A network trained on in silico bulk RNA-seq data can seamlessly deconvolve microarray data of the same tissue. This is quite noteworthy, as microarray data are known to have a reduced dynamic range and several hybridization-based biases compared to RNA-seq data. In other words, Scaden can deconvolve bulk data of types that it has never been trained on, even in the face of strong data type bias. This raises the possibility that Scaden trained on scRNA-seq data might reliably deconvolve other bulk omics data as well, such as proteomic and metabolomic data. This assumption is strengthened by the fact that Scaden, trained on scRNA-seq data, attains state-of-the-art performance on the deconvolution of bulk RNA-seq data, two data types with very distinct biases [78].

As highlighted in the introduction, a drawback for many DNNs is the large amount of training data required to obtain robust performance. Here, we used scRNA-seq data to create in silico bulk RNA-seq data of predefined type (target tissue) with known composition, across datasets. This immediately highlights Scaden's biggest limitation, the dependency on scRNA-seq data of the target tissue. In this study, we have shown that Scaden, trained solely on simulated data from scRNA-seq datasets, can outperform GEP-based deconvolution algorithms. We did observe, however, that the addition of labeled RNA-seq samples to the training data did substantially improve deconvolution performance in the case of PBMC data. We therefore believe that efforts to increase the similarity between simulated training data and the target bulk RNA-seq data could increase Scaden's performance further. Mixtures of in silico bulk RNA-seq data and publically available RNA-seq data, of purified cell types, for example, could further increase the

deconvolution performance of Scaden. Furthermore, domain adaptation methods can be used to improve performance of models that are trained on data (here, scRNA-seq data) that are similar to the target data (here, RNA-seq data) [14]. In future versions, Scaden's simple multilayer perceptron architecture could leverage domain adaptation to further stabilize and improve its cell deconvolution performance.

Scaden uses an ensemble approach by averaging the predictions of three different models to increase performance and improve generalization. Increasing the number of models per ensemble would allow for the estimation of the prediction uncertainty. While not implemented in this study, this could be an interesting extension to Scaden's ensemble architecture.

Recent cell deconvolution algorithms have used cell fraction estimates to infer cell type–specific gene expression from bulk RNA-seq data. It is straightforward to use Scaden's cell fraction estimates to infer per-group [105] and per-sample [140] cell type–specific gene expression using simple regression or non-negative matrix factorization, respectively. We would like to add a note of caution, however, as the error of cell fraction estimates, which can be quite large, is propagated into the gene expression calculations and will affect any downstream statistical analysis.

While Scaden achieves good performance on the samples and tissues used in this study, it is important to keep in mind that cell type similarity, sample heterogeneity, and complexity, as well as experimental noise and bias, can severely limit deconvolution accuracy. Furthermore, Scaden is currently not attempting to model cell size differences in its algorithm, which might be useful to consider for the interpretation of prediction results.

In summary, the deconvolution performance, robustness to noise and bias, and the flexibility to learn from large numbers of in silico datasets, across data types (scRNA-seq and RNA-seq mixtures) and potentially even tissues, make us believe that DNN-based architectures will become an algorithmic mainstay of cell type deconvolution.

2.5 Methods

2.5.1 Datasets and preprocessing

scRNA-seq datasets. The following human PBMC scRNA-seq datasets were downloaded from the 10X Genomics data download page: 6k PBMCs from a Healthy Donor, 8k PBMCs from a Healthy Donor, Frozen PBMCs (Donor A), and Frozen PBMCs (Donor C). Throughout this paper, these datasets are referred to with the handles data6k, data8k, donorA, and donorC, respectively. It was not intended to incorporate as many datasets as possible. Instead, these four datasets were chosen with the goal to dispose

of a set of samples with consistent cell types and gene expression. This limited our choice to datasets that displayed clearly identifiable cell types for the majority of cells. The Ascites scRNA-seq dataset was downloaded from https://figshare.com as provided by Schelker et al. [171]. Pancreas and mouse brain datasets were downloaded from the scRNA-seq dataset collection of the Hemberg laboratory [76]. The human brain datasets from Darmanis et al. [46] and Lake et al. [106] where downloaded from Gene Expression Omnibus (GEO) with accession numbers GSE67835 and GSE97930, respectively. A table listing all datasets including references to the original publications can be found in table S1.

scRNA-seq preprocessing and analysis. All datasets were processed using the Python package Scanpy (v. 1.2.2) [210] following the Scanpy's reimplementation of the popular Seurat's clustering workflow. First, the corresponding cell-gene matrices were filtered for cells with less than 500 detected genes and genes expressed in less than five cells. The resulting count matrix for each dataset was filtered for outliers with high or low numbers of counts. Gene expression was normalized to library size using the Scanpy function "normalize_per_cell." The normalized matrix of all filtered cells and genes was saved for the subsequent data generation step.

The following processing and analysis steps had the sole purpose of assigning cell type labels to every cell. All cells were clustered using the louvain clustering implementation of the Scanpy package. The louvain clustering resolution was chosen for each dataset, using the lowest possible resolution value (low-resolution values lead to less clusters) for which the calculated clusters appropriately separated the cell types. The top 1000 highly variable genes were used for clustering, which were calculated using Scanpy's "filter_genes_dispersion" function with parameters min_mean = 0.0125, max_mean = 3, and min_disp = 0.5. Principal components analysis was used for dimensionality reduction.

To identify cell types, marker genes were investigated for all cell types in question. For PBMC datasets, useful marker genes were adopted from public resources such as the Seurat tutorial for 2700 PBMCs [170]. Briefly, interleukin-7 receptor (IL7R) was taken as marker for CD4 T cells, LYZ for monocytes, MS4A1 for B cells, GNLY for natural killer cells, FCER1A for dendritic cells, and CD8A and CCL5 as markers for CD8 T cells. For all other scRNA-seq datasets, marker genes and expected cell types were inferred from the original publication of the dataset. For instance, to annotate cell types of the mouse brain dataset from Zeisel et al. [225], we used the same marker genes as Zeisel and colleagues. We did not use the same cell type labels from the original publications because a main objective was to assure that cell type labeling is consistent between all datasets of a certain tissue.

Cell type annotation was performed manually across all the clusters for each dataset,

such that all cells belonging to the same cluster were labeled with the same cell type. The cell type identity of each cluster was chosen by crossing the cluster's highly differentially expressed genes with the curated cell type's marker genes. Clusters that could not be clearly identified with a cell type were grouped into the "Unknown" category.

Tissue datasets for benchmarking. To assess the deconvolution performance on real tissue expression data, we used datasets for which the corresponding cell fractions were measured and published. The first dataset is the PBMC1 dataset, which was obtained from Zimmermann et al. [229]. The second dataset, PBMC2, was downloaded from GEO with accession code GSE107011 [132]. This dataset contains both RNA-seq profiles of immune cells (S4 cohort) and from bulk individuals (S13 cohort). As we were interested in the bulk profiles, we only used 12 samples from the S13 cohort from these data. Flow cytometry fractions were collected from the Monaco et al. publication [132].

In addition to the above mentioned two PBMC datasets, we used Ascites RNA-seq data. This dataset was provided by the authors, and cell type fractions for this dataset were taken from the supplementary materials of the publication [171].

For the evaluation on pancreas data, artificial bulk RNA-seq samples created from the scRNA-seq dataset of Xin et al. [215] were used. This dataset was downloaded from the resources of the MuSiC publication [205]. The artificial bulk RNA-seq samples used for evaluation were then created using the "bulk_construct" function of the MuSiC tool.

To assess how Scaden and the GEP algorithms deal with the presence of unknown cell types, we generated PBMC bulk RNA samples from the four scRNA-seq datasets (6000 each). The undefined amount of unknown cells that was generated by this approach was removed to be replaced by defined amounts of 5, 10, 20, and 30% of unknown cells, respectively. Cell fractions of all four samples were predicted with Scaden trained on the other three.

Performance on these samples was then assessed to test robustness against unseen cell types in the bulk mixture. Scaden was trained on samples from all datasets but the test dataset, while CSx and MuSiC used data8k as a reference.

The microarray dataset GSE65133 was downloaded from GEO, and cell type fractions were taken from the original CS publication [139].

Last, we wanted to get insights into neurodegenerative cell fraction changes in the brain. While it is known that neurodegenerative diseases like AD are accompanied by a gradual loss of brain neurons, stage-specific cell type shifts are still hard to come by. Here, we use the ROSMAP study cortical RNA-seq dataset along with the corresponding clinical metadata, to infer cell type composition over six clinically relevant stages of neurode-

generation [3]. Furthermore, to assess deconvolution accuracy on postmortem human brain tissue, we used 41 samples from the ROSMAP, for which cell composition information from immunohistochemistry [145] was recently released and for which fractions for all cell types were reported. The ROSMAP RNA-seq data were downloaded from www.synapse.org/. The cell composition values were provided by the authors of the study [145].

RNA-seq preprocessing and analysis. For the RNA-seq datasets analyzed in this study, we did not apply any additional processing steps but used the obtained count or expression tables directly as downloaded for all datasets except the ROSMAP dataset. For the latter, we generated count tables from raw FastQ files using Salmon (33) and the GRCh38 reference genome. FastQ files from the ROSMAP study were downloaded from Synapse (www.synapse.org).

2.5.2 Simulation of bulk RNA-seq samples from scRNA-seq data

Scaden's DNN requires large amounts of training RNA-seq samples with known cell fractions. This explains why the generation of artificial bulk RNA-seq data is one of the key elements of the Scaden workflow.

To generate the training data, preprocessed scRNA-seq datasets were used (see the "Datasets and preprocessing" section), comprising the gene expression matrix and the cell type labels. Artificial RNA-seq samples were simulated by subsampling cells from individual scRNA-seq datasets; cells from different datasets were not merged into samples to preserve within-subject relationships. Datasets generated from multiple participants were split according to participant, and each subsampling was constrained to cells from one participant to capture the cross-subject heterogeneity and keep subject-specific gene dependencies.

The exact subsampling procedure is described in the following. First, for every simulated sample, random fractions were created for all different cell types within each scRNA-seq dataset using the random module of the Python package NumPy. Briefly, a random number was chosen from a uniform distribution between 0 and 1 using the NumPy function "random.rand()" for each cell type, and then this number was divided by the sum of all random numbers created to ensure the constraint of all fractions adding up to 1

$$f_c = \frac{r_c}{\sum_{C_{all}} r_c} \tag{2.1}$$

where r_c is the random number created for cell type c and C_{all} is the set of all cell types. Here, f_c is the calculated random fraction for cell type c. Then, each fraction was multiplied with the total number of cells selected for each sample, yielding the number of cells to choose for a specific cell type

$$N_c = f_c * N_{total} \tag{2.2}$$

where N_c is the number of cells to select for the cell type c, and N_{total} is the total number of cells contributing to one simulated RNA-seq sample (500, in this study). Next, N_c cells were randomly sampled from the scRNA-seq gene expression matrix for each cell type c. Afterward, the randomly selected single-cell expression profiles for every cell type are then aggregated by summing their expression values, to yield the artificial bulk expression profile for this sample.

Using the above-described approach, cell compositions that are strongly biased toward a certain cell type or are missing specific cell types are rare among the generated training samples. To account for this and to simulate cell compositions with a heavy bias to and the absence of certain cell types, a variation of the subsampling procedure was used to generate samples with sparse compositions, which we refer to as sparse samples. Before generating the random fractions for all cell types, a random number of cell types was selected to be absent from the sample, with the requirement of at least one cell type constituting the sample. After these leave-out cell types were chosen, random fractions were created and samples generated as described above. The average cell type proportions of the training dataset generated as described above are equal for all cell types. This allows for unbiased deconvolution as the true cell composition of a given tissue is not known beforehand. Using different sampling distributions (e.g., Gaussian and Uniform) or excluding sparse samples did not change Scaden's deconvolution performance notably on the simulated PBMC datasets. This shows that Scaden is relatively robust to training data generated by different sampling procedures.

Using this procedure, we generated 32,000 samples for the human PBMC training dataset, 14,000 samples for the human pancreas training dataset, 6000 samples for human brain, and 30,000 samples for the mouse brain training dataset (table 2.3).

Artificial bulk RNA-seq datasets were stored in "h5ad" format using the Anndata package [210], which allows to store the samples together with their corresponding cell type ratios while also keeping information about the scRNA-seq dataset of origin for each sample. This allowed to access samples from specific datasets, which is useful for cross-validation.

2.5.3 Scaden overview

The following section contains an overview of the input data preprocessing, the Scaden model, model selection, and how Scaden predictions are generated.

Input data preprocessing. The data preprocessing step is aimed to make the input data

more suitable for machine learning algorithms. To achieve this, an optimal preprocessing procedure should transform any input data from the simulated samples or from the bulk RNA-seq to the same feature scale. Before any scaling procedure can be applied, it must be ensured that both the training data and the bulk RNA-seq data subject to prediction share the same features. Therefore, before scaling, both datasets are limited to contain features (genes) that are available in both datasets. In addition, uninformative genes that have either zero expression or an expression variance below 0.1 were removed, leaving 10,000 genes for model training and inference. The two-step processing procedure used for Scaden is described in the following:

First, to account for heteroscedasticity, a feature inherent to RNA-seq data, the data were transformed into logarithmic space by adding a pseudocount of 1 and then taking the Logarithm (base 2).

Second, every sample was scaled to the range [0,1] using the MinMaxScaler() class from the Sklearn preprocessing module. Per-sample scaling, unlike per-feature scaling that is more common in machine learning, assures that intergene relative expression patterns in every sample are preserved. This is important, as our hypothesis was that a neural network could learn the deconvolution from these intergene expression patterns

$$x_{scaled,i} = \frac{x_i - min(X_i)}{max(X_i) - min(X_i)} \tag{2.3}$$

where $x_{scaled,i}$ is the log2 expression value of gene x in sample i, X_i is the vector of log2 expression values for all genes of sample i, $min(X_i)$ is the minimum gene expression of vector X_i, and $max(X_i)$ is the maximum gene expression of vector X_i.

Note that all training datasets are stored as expression values and are only processed as described above. In the deployment use case, the simulated training data should contain the same features as in the bulk RNA-seq sample that shall be deconvolved.

Model selection. The goal of model selection was to find an architecture and hyperparameters that robustly deconvolve simulated tissue RNA-seq data and, more importantly, real bulk RNA-seq data. Because of the very limited availability of bulk RNA-seq datasets with known cell fractions, model selection was mainly optimized on the simulated PBMC datasets. To capture interexperimental variation, we used leave-one-dataset-out cross-validation for model optimization: A model was trained on simulated data from all but one dataset, and performance was tested on simulated samples from the left-out dataset. This allows to simulate batch effects between datasets and helps to test the generalizability of the model. In the process of model selection and (hyper-) parameter optimization, performed on PBMC and Ascites datasets, we found three models with different architectures and dropout rates but comparable performance. To address

overfitting in individual models, we decided to use a combination of models, expecting this to serve as another means of regularization. We did not test multiple combinations but rather used an informed choice with varying layer sizes and dropout regularization, with the goal to increase model diversity. We observed that the average of an ensemble of models generalized better to the test sets than individual models. Model training and prediction is done separately for each model, with the prediction averaging step combining all model predictions (tables 2.4 and 2.6). We provide a list of all tested parameters in the Supplementary Materials (table 2.5).

Final Scaden model. The Scaden model learns cell type deconvolution through supervised training on datasets of simulated bulk RNA-seq samples simulated with scRNA-seq data. To account for model biases and to improve performance, Scaden consists of an ensemble of three DNNs with varying architectures and degrees of dropout regularization. All models of the ensemble use four layers of varying sizes between 32 and 1024 nodes, with dropout regularization implemented in two of the three ensemble models. The exact layer sizes and dropout rates are listed in table 2.4. The rectified linear unit is used as activation function in every internal layer. We used a Softmax function to predict cell fractions, as we did not see any improvements in using a linear output function with consecutive non-negativity correction and sum-to-one scaling. Python (v. 3.6.6) and the TensorFlow library (v. 1.10.0) were used for implementation of Scaden. A complete list of all software used for the implementation of Scaden is provided in table S15.

Training and prediction. After the preprocessing of the data, a Scaden ensemble can be trained on simulated tissue RNA-seq data or mixtures of simulated and real tissue RNA-seq data. Parameters are optimized using Adam with a learning rate of 0.0001 and a batch size of 128. We used an L1 loss as optimization objective

$$L1(y_i, \hat{y_i}) = |y_i - \hat{y_i}| \tag{2.4}$$

where y_i is the vector of ground-truth fractions of sample i and $\hat{y_i}$ is the vector of predicted fractions of sample i. Each of the three ensemble models is trained independently for 5000 steps. This "early stopping" serves to avoid domain overfitting on the simulated tissue data, which would decrease the model performance on the real tissue RNA-seq data. We observed that training for more steps lead to an average performance decrease on real tissue RNA-seq data. To perform deconvolution with Scaden, a bulk RNA-seq sample is fed into a trained Scaden ensemble, and three independent predictions for the cell type fractions of this sample are generated by the trained DNNs. These three predictions are then averaged per cell type to yield the final cell type composition for the input bulk RNA-seq sample

$$\hat{y_c} = \frac{\hat{y_c^1} + \hat{y_c^2} + \hat{y_c^3}}{3} \tag{2.5}$$

where $\hat{y_c}$ is the final predicted fraction for cell type c and $\hat{y_c^i}$ is the predicted fraction for cell type c of model i.

Scaden requirements. Currently, a disadvantage of the Scaden algorithm is the necessity to train a new model for deconvolution if no perfect overlap in the feature space exists. This constraint limits the usefulness of pretrained models. Once trained, however, the prediction runtime scales linearly with sample numbers and is usually in the order of seconds, making Scaden a useful tool if deconvolution is to be performed on very large datasets. While the requirements are dataset dependent, the Scaden demo was profiled to require a peak of 3.2 GB of random-access memory (RAM) during the DNN training process, so a computer with 8 GB of RAM should be able to run it smoothly. In our tests with an Intel(R) Xeon(R) CPU E5-1630 workstation, the demo could run in 22 min, spending most of the CPU time in the DNN training process. The most prominent and obvious issue of Scaden is the difference between simulated scRNA-seq data used for training and the bulk RNA-seq data subject to inference. While Scaden is able to transfer the learned deconvolution between the two data types and achieves state-of-the-art performance, we hypothesize that efforts to improve this translatability could improve Scaden's prediction accuracy even further. Algorithmic improvements are therefore likely to address this issue and are planned for future releases.

2.5.4 Algorithm comparison

We used several performance measures to compare Scaden to four existing cell deconvolution algorithms, CS with LM22 GEP, CSx, MuSiC, and CPM. To compare the performance of the five deconvolution algorithms, we measured the RMSE, Lin's CCC, Pearson product moment correlation coefficient r, and R2 values, comparing real and predicted cell fractions estimates. In addition, to identify systematic prediction errors and biases, slope and intercept for the regression lines were calculated. These metrics are defined as follows

$$RMSE(y,\hat{y}) = avg(\sqrt{(y-\hat{y})^2} \tag{2.6}$$

$$r(y,\hat{y}) = \frac{cov(y,\hat{y})}{\sigma_y \sigma_{\hat{y}}} \tag{2.7}$$

$$R^2(y,\hat{y}) = r(y,\hat{y})^2 \tag{2.8}$$

$$slope(y,\hat{y}) = \frac{\Delta y}{\Delta \hat{y}} \tag{2.9}$$

$$CCC(y,\hat{y}) = \frac{2r\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})} \tag{2.10}$$

where y are the ground-truth fractions, $\hat{y}$ are the prediction fractions, σ_x is the SD of x, $cov(y,\hat{y})$ is the covariance of y and $\hat{y}$, and μ_y, $\mu_{\hat{y}}$ are the mean of the predicted and ground-truth fractions, respectively.

All metrics were calculated for all data points of a dataset and separately for all data points of a specific cell type. For the latter approach, we then averaged the resulting values to recover single values. While the metrics calculated on all data points might be sufficient, we deem that the cell type–specific deconvolution might, in many instances, be of even greater interest. It is noteworthy in this context that cell type–specific deconvolution performance can be quite weak, depending on the dataset. This is true for all tested deconvolution algorithms, while Scaden achieves best performance.

CIBERSORT. CS is a cell convolution algorithm based on specialized GEPs and support vector regression. Cell composition estimations were obtained using the CS web application (https://cibersort.stanford.edu/). For all deconvolutions with CS, we used the LM22 GEP, which was generated by the CS authors from 22 leukocyte subsets profiled on the HGU133A microarray platform.

Because the LM22 GEP matrix contains cell types at a finer granularity than what was used for this study, predicted fractions of subcell types were added together. For cell grouping, we used the mapping of subcell types to broader types given by figure 6 from Monaco et al. [132]. We provide a table with the exact mappings used here in the Supplementary Materials (table 2.13). The deconvolution was performed using 500 permutations with quantile normalization disabled for all datasets but GSE65133 (Microarray), as is recommended for RNA-seq data. We used default settings for all other CS parameters.

CIBERSORTx. CSx is a recent variant of CS that can generate GEP matrices from scRNA-seq data and use these for deconvolution. For additional deconvolution robustness, it applies batch normalization to the data. All signature matrices were created by uploading the labeled scRNA-seq expression matrices and using the default options. Quantile normalization was disabled. For deconvolution on simulated data, no batch normalization was used. For all bulk RNA-seq datasets, the S-Mode batch normalization was chosen. All PBMC datasets were deconvolved using a GEP matrix generated from the data6k dataset (for simulated samples from data6k, a donorA GEP matrix was chosen).

MuSiC. MuSiC is a deconvolution algorithm that uses multisubject scRNA-seq datasets

as GEP matrices in an attempt to include heterogeneity in the matrices to improve generalization. While MuSiC tries to address similar issues of previous deconvolution algorithms by using scRNA-seq data, the approach is very different. For deconvolution, MuSiC applies a sophisticated GEP-based deconvolution algorithm that uses weighted non-negative least-squares regression with an iterative estimation procedure that imposes more weight on informative genes and less weight on noninformative genes.

The MuSiC R package contains functionality to generate the necessary GEP matrix given an scRNA-seq dataset and cell type labels. To generate MuSiC deconvolution predictions on PBMC datasets, we used the data8k scRNA-seq dataset as reference data for MuSiC and follow the tutorial provided by the authors to perform the deconvolution. For deconvolution of artificial samples generated from the data8k dataset, we provided MuSiC with the data6k dataset as a reference instead.

MuSiC was developed with a focus on multisubject scRNA-seq datasets, in which the algorithm tries to take advantage from the added heterogeneity that these datasets contain, by calculating a measure of cross-subject consistency for marker genes. To assess how Scaden performs on multisubject datasets compared to MuSiC, we evaluated both methods on artificial bulk RNA-seq samples from human pancreas. We used the bulk_construct function from MuSiC to combine the cells from all 18 participants contained in the scRNA-seq dataset from Xin et al. [215] to generate artificial bulk samples for evaluation. Next, as a multisubject reference dataset, we used the pancreas scRNA-seq dataset from Segerstolpe et al. [172], which contains single-cell expression data from 10 different participants, 4 of which with type 2 diabetes. For Scaden, the Segerstolpe scRNA-seq dataset was split by participants, and training datasets were generated for each participant, yielding in total 10,000 samples. For MuSiC, a processed version of this dataset was downloaded from the resources provided by the MuSiC authors [205] and used as an input reference dataset for the MuSiC deconvolution. Deconvolution was then performed according to the MuSiC tutorial, and performance was compared according to the above-defined metrics.

Cell Population Mapping. CPM is a deconvolution algorithm that uses single-cell expression profiles to identify a so-called "cell population map" from bulk RNA-seq data [62]. In CPM, the cell population map is defined as composition of cells over a cell-state space, where a cell state is defined as a current phenotype of a single cell. Contrary to other deconvolution methods, CPM tries to estimate the abundance of all cell states and types for a given bulk mixture, instead of only deconvolving the cell types. As input, CPM requires an scRNA-seq dataset and a low-dimensional embedding of all cells in this dataset, which represents the cell-state map. As CPM estimates abundances of both cell states and types, it can be used for cell type deconvolution by summing up all estimated fractions for all cell states of a given cell type, a method that is implemented in the scBio R package, which contains the CPM method. To perform deconvolution with

CPM, we used the data6k PBMC scRNA-seq dataset as an input reference for all PBMC samples. For samples simulated from the data6k dataset, we used the data8k dataset as a reference. According to the CPM paper, a dimension reduction method can be used to obtain the cell-state space. We therefore used Uniform Manifold Approximation and Projection (UMAP), a dimension reduction method widely used for scRNA-seq data, to generate the cell-state space mapping for the input scRNA-seq data. Deconvolution was then performed using the CPM function of the scBio package with an scRNA-seq dataset and accompanying UMAP embedding as input.

2.6 Supplementary Data

Tissue	Name	cells	Subjects	Source
PBMC	data6k	5,419	1	10X Genomics
PBMC	data8k	8,381	1	10X Genomics
PBMC	donorA	2,900	1	10X Genomics
PBMC	donorC	9,519	1	10X Genomics
Mouse Brain	Tasic	1,679	1	[191]
Mouse Brain	Zeisel	3,005	1	[225]
Mouse Brain	Romanov	2,881	1	[162]
Mouse Brain	Campbell	21,086	1	[38]
Mouse Brain	Chen	14,437	1	[40]
Pancreas	Segerstolpe	3,514	10	[172]
Pancreas	Baron	8,569	4	[22]
Ascites	Ascites	3,114	3	[171]
Human Brain	Darmanis	465	1	[46]
Human Brain	Lake	27,416	1	[106]

Table 2.1: scRNA-seq datasets used for the generation of simulated tissues for Scaden training.

Tissue	Name	Samples	Reference
PBMC	PBMC1	12	[229]
PBMC	PBMC2	12	[132]
Pancreas	Xin	18	[215]
Human Brain	ROSMAP	390	[3]
Ascites	Ascites	3	[171]

Table 2.2: Bulk tissue RNA-seq datasets used for performance evaluation.

Tissue	Samples	Datasets	Size
PBMC	32,000	4	1.5 GB
Pancreas	14,000	2	0.6 GB
Human Brain	6,000	2	0.32 GB
Ascites	6,000	1	0.38 GB
Mouse Brain	30,000	5	1.5 GB

Table 2.3: Number of samples, datasets and size of the simulated training data.

Model	Layers	Layer sizes	Dropout rates
M256	4	256, 128, 64, 32	0, 0, 0, 0
M512	4	512, 256, 128, 64	0, 0.3, 0.2, 0.1
M1024	4	1024, 512, 256, 128	0, 0.6, 0.3, 0.1

Table 2.4: Architectures of deep neural network models used in Scaden ensemble. All models use an L1 as a loss function, ReLU activation for all layers but the last, and softmax activation for the last layer.

Parameter	Values tested
Batch size	32, 64, 128, 256, 512
Layers	2, 3, 4
Layer sizes	2048, 1024, 512, 256, 128, 64, 32, 16
Dropout rate 0 - 0.8heightLoss function	 L1, L2

Table 2.5: Hyperparameters used for model optimization

	Scaden Ensemble	M256	M512	M1024
CCC	0.914	0.898	0.909	0.907

Table 2.6: Comparison of Scaden models and the Scaden ensemble on four PBMC scRNA-seq datasets. Concordance correlation coefficient was calculated on all datasets separately and then averaged.

Method	DS	RMSE	Slope	Correlation	Intercept	CCC
CPM	data6k	0.192	0.03	0.082	0.162	0.053
CPM	data8k	0.185	0.048	0.263	0.159	0.093
CPM	donorA	0.239	-0.081	-0.259	0.18	-0.147
CPM	donorC	0.189	0.038	0.102	0.16	0.066
CS	data6k	0.163	0.508	0.57	0.082	0.566
CS	data8k	0.136	0.551	0.708	0.075	0.687
CS	donorA	0.137	0.605	0.767	0.066	0.746
CS	donorC	0.168	0.45	0.522	0.092	0.517
CSx	data6k	0.106	0.756	0.824	0.041	0.821
CSx	data8k	0.097	0.744	0.863	0.043	0.854
CSx	donorA	0.125	0.696	0.81	0.051	0.801
CSx	donorC	0.094	0.829	0.865	0.029	0.864
MuSiC	data6k	0.086	0.848	0.887	0.025	0.886
MuSiC	data8k	0.136	0.663	0.728	0.056	0.725
MuSiC	donorA	0.1	0.811	0.883	0.031	0.88
MuSiC	donorC	0.084	0.897	0.896	0.017	0.896
Scaden	data6k	0.104	0.747	0.83	0.042	0.825
Scaden	data8k	0.133	0.625	0.73	0.063	0.722
Scaden	donorA	0.035	0.92	0.988	0.013	0.985
Scaden	donorC	0.046	0.849	0.973	0.025	0.964

Table 2.7: Deconvolution evaluation on simulated PBMC data.

Method	Celltype	RMSE	Correlation	Slope	Intercept	CCC
CSx	ALPHA	0.282	0.816	0.691	0.431	0.375
CSx	Average	0.171	0.845	0.891	0.1	0.499
CSx	BETA	0.309	0.833	0.175	-0.017	0.078
CSx	DELTA	0.04	0.812	1.567	-0.013	0.647
CSx	GAMMA	0.052	0.921	1.131	0.0	0.897
CSx	Total	0.212	0.79	1.113	-0.028	0.746
MuSiC	ALPHA	0.11	0.887	1.108	-0.042	0.863
MuSiC	Average	0.087	0.835	0.861	-0.008	0.744
MuSiC	BETA	0.148	0.752	1.067	0.017	0.694
MuSiC	DELTA	0.023	0.817	0.716	-0.003	0.707
MuSiC	GAMMA	0.068	0.881	0.552	-0.003	0.711
MuSiC	Total	0.099	0.938	1.078	-0.019	0.929
Scaden	ALPHA	0.067	0.949	1.071	-0.034	0.942
Scaden	Average	0.051	0.902	1.031	-0.02	0.881
Scaden	BETA	0.07	0.936	1.152	-0.045	0.916
Scaden	DELTA	0.024	0.807	1.012	0.008	0.764
Scaden	GAMMA	0.045	0.914	0.89	-0.008	0.901
Scaden	Total	0.055	0.978	1.033	-0.008	0.976

Table 2.8: Deconvolution performance on simulated pancreas data from Xin et al. on a per cell-type level.

Method	Content	CCC	RMSE	Correlation	Intercept	Slope
CSx	0.02	0.731	0.097	0.738	0.032	0.841
CSx	0.05	0.751	0.092	0.754	0.035	0.823
CSx	0.1	0.715	0.092	0.719	0.041	0.797
CSx	0.2	0.694	0.091	0.703	0.039	0.807
CSx	0.3	0.632	0.099	0.637	0.057	0.714
MuSiC	0.02	0.793	0.084	0.803	0.016	0.921
MuSiC	0.05	0.799	0.083	0.805	0.018	0.908
MuSiC	0.1	0.739	0.089	0.745	0.032	0.841
MuSiC	0.2	0.669	0.095	0.679	0.04	0.8
MuSiC	0.3	0.665	0.101	0.687	0.028	0.861
Scaden	0.02	0.934	0.041	0.944	0.033	0.837
Scaden	0.05	0.925	0.044	0.936	0.035	0.825
Scaden	0.1	0.902	0.046	0.915	0.042	0.792
Scaden	0.2	0.846	0.054	0.859	0.051	0.747
Scaden	0.3	0.798	0.063	0.816	0.063	0.686

Table 2.9: Deconvolution performance on datasets with added unknown mixture contents.

Method	Dataset	RMSE	Correlation	Slope	Intercept	CCC
CPM	PBMC1	0.18	-0.003	-0.003	0.167	-0.003
CPM	PBMC2	0.114	-0.203	-0.094	0.182	-0.155
CS	PBMC1	0.147	0.437	0.491	0.085	0.434
CS	PBMC2	0.101	0.594	0.754	0.041	0.577
CSx	PBMC1	0.16	0.603	0.925	0.012	0.552
CSx	PBMC2	0.13	0.456	0.67	0.055	0.424
MuSiC	PBMC1	0.316	-0.235	-0.468	0.245	-0.189
MuSiC	PBMC2	0.299	-0.197	-0.542	0.257	-0.127
Scaden	PBMC1	0.104	0.722	0.805	0.032	0.717
Scaden	PBMC2	0.052	0.855	0.848	0.025	0.855

Table 2.10: Deconvolution performance on real PBMC RNA-seqdatasets PBMC1 and PBMC2. Scaden was trained on a mixture of in silico and real bulk RNA-seq data, the remaining tools used either scRNA-seq datasets as reference (CPM, MuSiC, CSx) or a in-built reference (CS).

Method	Dataset	Celltype	RMSE	Corr.	Slope	Intercept	CCC
Scaden_SC	PBMC1	Total	0.131	0.564	0.644	0.059	0.559
Scaden_SC	PBMC2	Total	0.077	0.684	0.689	0.052	0.684
Scaden_all	PBMC1	Total	0.104	0.722	0.805	0.032	0.717
Scaden_all	PBMC2	Total	0.052	0.855	0.848	0.025	0.855
Scaden_SC	PBMC1	Bcells	0.033	0.648	0.172	0.006	0.083
Scaden_SC	PBMC1	CD4Tcells	0.228	0.633	0.492	-0.055	0.149
Scaden_SC	PBMC1	CD8Tcells	0.101	0.603	0.761	0.108	0.562
Scaden_SC	PBMC1	Monocytes	0.178	0.556	0.885	0.173	0.186
Scaden_SC	PBMC1	NK	0.087	0.81	0.531	0.137	0.312
Scaden_SC	PBMC1	Unknown	0.029	0.577	0.361	0.009	0.287
Scaden_SC	PBMC2	Bcells	0.012	0.936	0.977	0.002	0.935
Scaden_SC	PBMC2	CD4Tcells	0.145	0.767	0.682	-0.057	0.119
Scaden_SC	PBMC2	CD8Tcells	0.049	0.67	0.403	0.129	0.587
Scaden_SC	PBMC2	Monocytes	0.078	0.865	0.994	0.071	0.558
Scaden_SC	PBMC2	NK	0.071	0.629	0.314	0.14	0.276
Scaden_SC	PBMC2	Unknown	0.025	0.247	0.217	0.044	0.209
Scaden_all	PBMC1	Bcells	0.031	0.668	0.188	0.007	0.1
Scaden_all	PBMC1	CD4Tcells	0.151	0.638	0.652	-0.017	0.345
Scaden_all	PBMC1	CD8Tcells	0.096	0.6	0.704	0.123	0.569
Scaden_all	PBMC1	Monocytes	0.172	0.518	0.777	0.184	0.177
Scaden_all	PBMC1	NK	0.036	0.804	0.488	0.058	0.71
Scaden_all	PBMC1	Unknown	0.026	0.64	0.41	0.01	0.365
Scaden_all	PBMC2	Bcells	0.013	0.936	0.94	0.0	0.917
Scaden_all	PBMC2	CD4Tcells	0.074	0.772	0.769	-0.005	0.373
Scaden_all	PBMC2	CD8Tcells	0.051	0.672	0.398	0.106	0.562
Scaden_all	PBMC2	Monocytes	0.072	0.895	1.058	0.049	0.614
Scaden_all	PBMC2	NK	0.045	0.69	0.301	0.103	0.467
Scaden_all	PBMC2	Unknown	0.023	0.241	0.178	0.043	0.203

Table 2.11: Deconvolution performance on real PBMC RNA-seq data for Scaden models trained only on scRNA-seq simulated tissues (Scaden_SC) or on a mix of simulated and real tissue (Scaden_all).

Method	Reference	CCC	RMSE	Correlation	Slope	Intercept
Scaden	Darmanis	0.922	0.056	0.924	0.992	0.028
Scaden	Lake	0.76	0.144	0.918	1.730	0.103
Scaden	Darmanis& Lake	0.898	0.064	0.899	0.948	0.029
MuSiC	Darmanis	0.873	0.089	0.951	1.445	0.075
MuSiC	Lake	0.65	0.119	0.652	0.696	0.078
CSx	Darmanis	0.805	0.123	0.94	1.663	0.094
CSx	Lake	0.706	0.183	0.95	2.130	0.115

Table 2.12: Deconvolution performance on bulk RNA-seq data from post-mortem human brain tissue (ROSMAP). Metrics are reported for deconvolution with different reference datasets (Darmanis and Lake) and additionally with both reference datasets for Scaden (Darmanisand Lake).

Method	Celltype	CCC	Correlation	Intercept	R2	RMSE	Slope
CIBERSORT	Average	0.391	0.612	0.087	0.396	0.103	0.516
CIBERSORT	Bcells	0.122	0.33	0.029	0.109	0.068	0.109
CIBERSORT	CD4Tcells	0.629	0.658	0.199	0.433	0.095	0.537
CIBERSORT	CD8Tcells	0.285	0.635	0.018	0.404	0.12	0.375
CIBERSORT	Monocytes	0.295	0.741	0.19	0.548	0.17	0.779
CIBERSORT	NK	0.623	0.698	-0.003	0.487	0.059	0.78
CIBERSORT	Total	0.717	0.728	0.026	0.53	0.11	0.869
Scaden	Average	0.498	0.726	-0.032	0.536	0.115	1.006
Scaden	Bcells	0.431	0.728	0.012	0.53	0.055	0.388
Scaden	CD4Tcells	0.64	0.778	-0.195	0.606	0.153	1.474
Scaden	CD8Tcells	0.474	0.543	0.02	0.294	0.104	0.635
Scaden	Monocytes	0.43	0.838	0.033	0.702	0.191	1.764
Scaden	NK	0.516	0.741	-0.029	0.549	0.074	0.77
Scaden	Total	0.705	0.749	-0.015	0.561	0.126	1.067

Table 2.13: Deconvolution performance comparison of CS (LM22) and Scaden on the GSE65133 Microarray dataset. Pleasenote that the LM22 GEP used for CS was created using PBMC microarray data, while Scaden was trained on simulated scRNA-seq PBMC datasets.

Chapter 3

Integrative analysis of Multi-Omics FTD Data

Disclaimer

A manuscript describing the content of this chapter is in preparation. All data analysis leading to the results presented here was carried out by me. Tenzin Nyima and Margheritta Francescatto helped with processing of the CAGE-seq data. Experiments for generation of the different datasets analysed in this chapter were designed and performed by Ashutosh Dhingra, Melissa Castillo Lizardo, Noémia Rita Fernandes, Eldem Sadikoglou, Salvador Rodriguez Nieto and Patrizia Rizzu. Pathology scoring was done by Manuela Neumann. Part of the small RNA-seq experiments was performed by the group of André Fischer at the DZNE Göttingen, by Cemil Kerimoglu and Lalit Kaurani. Peter Heutink designed, initiated and organized the RiMod-FTD project.

3.1 Abstract

To develop treatments for the neurodegenerative disorder FTD a better understanding of the underlying molecular disease mechanisms is needed. The goal of RiMod-FTD is to create a multi-omics and multi-model resource of datasets relevant for different disease subtypes. Here, we present an integrative analysis of multi-omics datasets from phase 1 of the RiMod-FTD project. Using RNA-seq, smRNA-seq, methylation and CAGE-seq data, we have identified a selective vulnerability of excitatory neurons in FTD and an enrichment of endothelial cells in all disease groups. We have furthermore detected potential disease mechanisms that drive neuroinflammation in FTD-GRN and identified several miRNAs that inhibit FTD-relevant cellular processes. This first integrative analysis of the RiMod-FTD resource provides valuable new insights and shows how this resource will be able to help advance the field of FTD research in the future.

3.2 Introduction

FTD is a devastating form of dementia that typically manifests before the age of 65. The underlying pathology behind FTD is frontotemporal lobar degeneration (FTLD), which is characterized by a progressive deterioration of frontal and temporal lobes. For a detailed introduction to current knowledge about genetics, pathology and disease mechanisms in FTD I would like to refer the reader to section 1.1. Briefly, FTD symptoms include behavioural changes and deterioration of language production and understanding. Depending on the clinical symptoms, FTD is divided into the behavioural variant FTD subtype (bvFTD) or the language subtypes primary progressive aphasia (PPA) and non-fluent variant FTD (nfvFTD).

With up to 43% of familial cases, FTD has a large genetic component. During the recent decades, mutations in several genes have been identified as causes for FTD. Approximately half of all familial cases are caused by mutations in one of the genes MAPT, GRN or C9orf72. While mutations in MAPT lead to pathological aggregations of the tau protein, mutations in GRN and C9orf72 lead to aggregations of TDP-43. C9orf72 mutations furthermore lead to RNA foci and dipeptide repeats (DPRs) as additional pathological features (see section 1.1.2). Significant advances have been made recently with regards to understanding the molecular mechanisms that underlie FTD (see section 1.1.4). However, as of today, no treatment that can halt or slow the progression of FTD exists. Further research is therefore necessary to increase our knowledge about causal mechanisms in order to develop remedies. Given that FTD is a heterogeneous disease with differing underlying pathology, caused by genes with very different functions, it is of crucial importance to expand our knowledge about commonalities and differences between the different disease subtypes. This will help us to determine whether treatments for FTD in general can be developed, or whether treatments for different subtypes are needed.

The Risk and Modifying factors in Frontotemporal Dementia (RiMod-FTD) project is funded by the EU Joint Programme - Neurodegenerative Disease (JPND) research to address the above mentioned challenges. The project is a consortium effort by 10 different groups throughout Europe. The goal of RiMod-FTD is to identify common and distinct disease mechanisms in different FTD subtypes. To achieve this, a multi-omics and multi-model approach is used. Post-mortem brain tissue from FTD patients with mutations in MAPT, GRN or C9orf72 and healthy controls were collected. The tissue samples were profiled using multiple 'omics' technologies: RNA-seq, CAGE-seq, small RNA-sequencing (smRNA-seq) and proteomics. The multi-omics approach allows to examine the molecular mechanisms in end-stage FTD with high precision, as various different regulatory mechanisms and disease aspects can be studied and compared between patient groups. Different brain regions are studied to identify potential regional differences. Furthermore, mouse models for the different mutations in MAPT, GRN and C9orf72 are generated which allow to study the temporal progression of the disease and

the effects of the mutations in a controlled environment. As post-mortem human brain data only represents the end stage of the disease, the longitudinal data that can be generated with mouse models is an important addition. Finally, induced pluripotent stem cell (iPSC) models of specific cell types were derived from patient cells. The cellular models depict valuable systems that can be used to test hypotheses derived from human and mouse data.

In this chapter, we have intensively analysed and integrated the multi-omics data from post-mortem human brain tissue. Specifically, RNA-seq, CAGE-seq, smRNA-seq and methylation data from the frontal lobe were analysed. We have identified pathways that are dysregulated in all investigated FTD genetic subtypes and pathways that are distinctly affected in FTD subtypes. Using the multi-omics data, we have examined regulatory mechanisms and identified several potential key regulators and pathways, which constitute new avenues to pursue for FTD research.

3.3 Results

3.3.1 Multi-omics Data Resource for Frontotemporal Dementia

As part of RiMod-FTD, data from seven different brain regions (frontal, temporal and occipital lobes, hippocampus, caudate, putamen, cerebellum) was collected and processed. In this chapter, I present the results from the analysis of the data from the frontal lobe, which is the most comprehensively studied brain region. The datasets consist of 47 to 49 samples for RNA-seq, CAGE-seq, smRNA-seq and Illumina InfiniumEPIC methylation profiling (see Table 3.2). Additionally, CAGE-seq data from mouse models of MAPT and GRN mutations and RNA-seq data from iPSC-derived neurons generated from FTD patients with mutations in GRN, MAPT or C9orf72, was analysed. Apart from unraveling new disease mechanism, an important objective of the RiMod-FTD consortium was to create a public data resource for FTD. The value of a high-quality, central resource for a specific research question, such as understanding a disease, has been proven by other projects such as ROSMAP [2] and promises to advance and accelerate the FTD research field as well.

Dataset	samples	FTD-GRN	FTD-MAPT	FTD-C9orf72	controls
RNA-seq	47	7	11	13	16
CAGE-seq	49	8	12	13	16
smRNA-seq	47	8	13	13	13
Methylation	48	7	13	14	14

Table 3.1: Post-mortem human brain datasets analysed in this chapter with repsective number of total samples and samples per group.

	Age	PMI	pH	RIN	M / F
FTD-MAPT	60.7 (7.9)	389.8 (120.9)	6.4 (0.2)	7.1 (0.9)	7 / 5
FTD-GRN	63.0 (8.3)	284.2 (54.3)	6.4 (0.2)	6.7 (0.7)	2 / 6
FTD-C9orf72	63.8 (8.2)	366.1 (131.9)	6.4 (0.3)	6.2 (0.8)	5 / 8
Control	79.7 (11.5)	370.2 (62.5)	6.6 (0.3)	7.5 (1.0)	6 / 10

Table 3.2: Average characteristics of samples from GFM. The average metrics for each sample group are shown with the standard deviation in bracketes. PMI corresponds to time after death until samples collection in minutes. The other metrics are age at death, pH-value, RNA integrity number (RIN) and the number of males (M) and females (F) in the respective group.

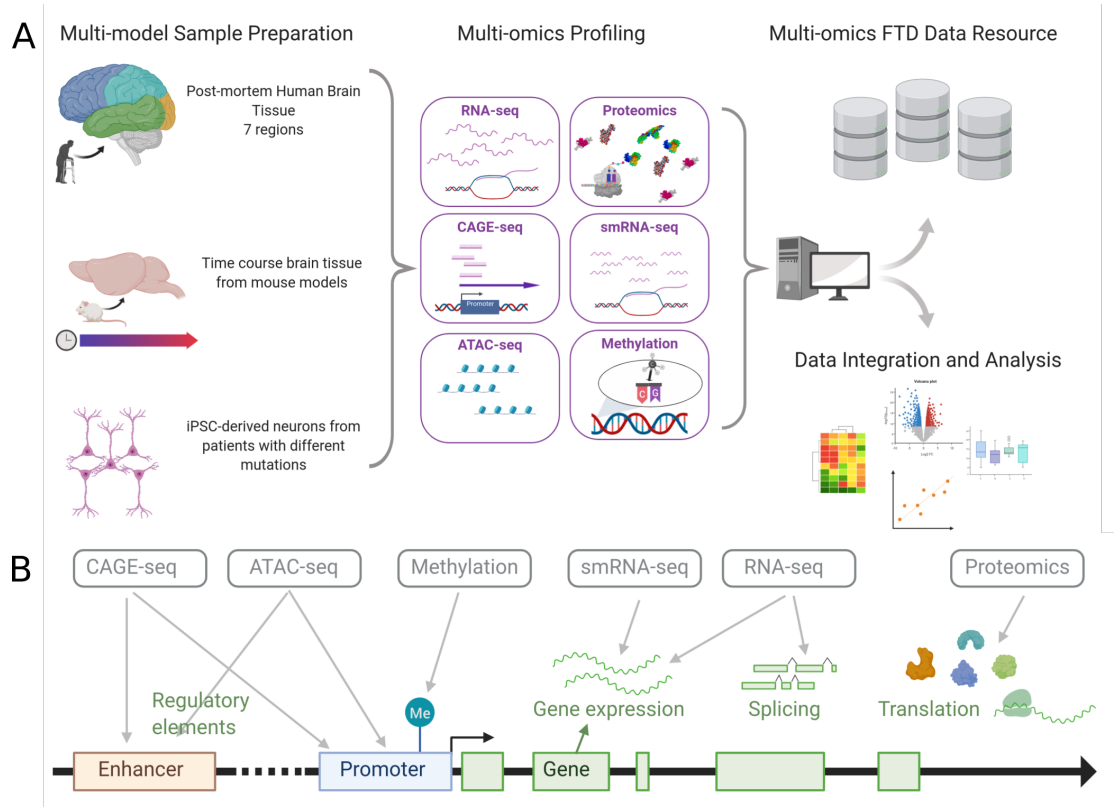

Figure 3.1: The RiMod-FTD project. **A** Data from three different sources is sampled: Human post-mortem brain tissue of FTD patients with mutations in GRN, MAPT or C9orf72 and healthy controls, brain tissue of mouse models for the genetic FTD subtypes and iPSC-derived neurons, generated from cells of FTD patients with known mutations. All samples are then profiled using a multi-omics approach. Finally, the data will be uploaded to yield a multi-omics FTD data resource and analyzed to generate new hypotheses for FTD disease mechanisms. **B** The multi-omics approachs allows to examine several mechanisms that are important for gene expression and function, such as regulation at the promoter and enhancer, DNA methylation or regulation by miRNAs.

3.3.2 Affected Genes, Pathways and Cell types in FTD

To identify generally dysregulated pathways in the frontal lobe of FTD brains, we analysed the RNA-seq and smRNA-seq data for differential expression and sample heterogeneity. Principal component analysis (PCA) revealed high sample heterogeneity (Fig. 3.2A) in the RNA-seq gene expression data. However, visual inspection of the first two principal components also indicates a difference between control and FTD samples. We observed the largest number of differentially expressed genes (DEGs) for FTD-GRN, followed by FTD-MAPT and FTD-C9orf72 (Fig. 3.2B, adj. P-value < 0.05). Because we detected only relatively few DEGs for FTD-C9orf72 (148), we have focused on the

comparison and analysis of FTD-GRN and FTD-MAPT. DE analysis of miRNAs using the smRNA-seq data yielded 88, 42 and 36 differentially expressed miRNAs in FTD-MAPT, FTD-GRN and FTD-C9orf72, respectively (adj. P-value < 0.05, absolute log fold-change > 0.6, Fig. 3.2C). We generated putative miRNA-target pairings by correlating the expression of DE miRNAs with their predicted targets (see Methods) using matching RNA-seq and smRNA-seq samples. We retained only predicted targets with a negative correlation of at least -0.4 with their respective miRNA.

Due to the neurodegenerative nature of FTD, it is likely that there exists a systematic difference in cell composition between cases and controls which can affect differential gene expression analysis due to differences in gene expression between cell types. To account for this common problem, we applied a conservative approach by removing all DEGs with a high likelihood of being false-positives from the analysis. Briefly, we performed cell type deconvolution analysis to estimate cell type fractions for each sample using our recently developed deconvolution method Scaden [126] (Fig. 3.10A). Then, we selected DEGs with high cell type specificity and correlated their expression with the fractions of the major expressing cell type (see Methods). False positive DEGs that are caused by systematic increase or decrease of a specific cell type will show high correlation with the cell type fractions and can thus be identified and removed from the analysis. All further analyses were based on the filtered set of DEGs, unless otherwise specified. Note that this method could only be applied to the total RNA-seq dataset because similar cell type specificity data was not available for other data types.

Next, we performed pathway enrichment analyses of up- and down-regulated DEGs using the go:Profiler tool [155]. Both for FTD-MAPT and FTD-GRN, mitochondrial pathways are the most significantly down-regulated pathways (Fig. 3.2D, Fig. 3.9). Other down-regulated pathways in FTD-GRN include cellular localization, protein ubiquitination and the synaptic vesicle cycle. Among up-regulated pathways, FTD-GRN and FTD-MAPT share extracellular matrix organization and circulatory system development. In FTD-GRN, cell surface receptor signaling pathways and immune system pathways are furthermore enriched. In FTD-MAPT, we detect various developmental pathways enriched among up-regulated genes.

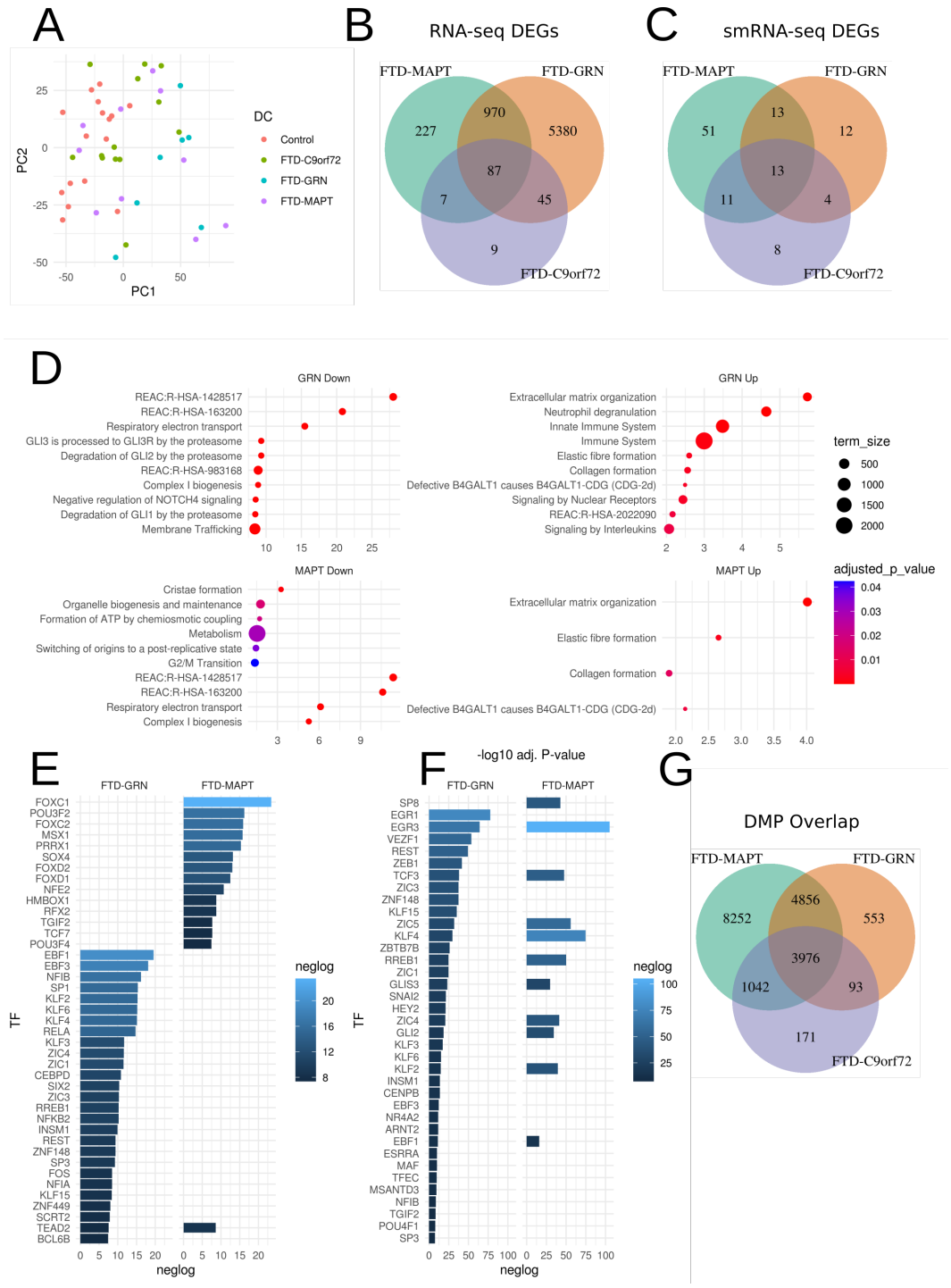

Figure 3.2: Gene- and Pathway-level transcriptional changes in FTD. **A** Principal component analysis of variance stabilized RNA-seq expression values, colored by group. **B** Overlap between RNA-seq DEGs from different disease groups. **C** Overlap between smRNA-seq DEGs from different disease groups. **D** Enriched Reactome pathways in RNA-seq up- and down-regulated DEGs. Shown are the then most significant pathways per group; the x-axis signifies the negative log10 P-value. Color corresponds to adjusted P-value and node size corresponds to the number of genes in a pathway. **E**, **F** Best candidates for active and inactiveTFs in FTD-GRN and FTD-MAPT, respectively. The x-axis signifies the negative log10 P-value. **G** Overlap of DMPs in different disease groups.

Next, to better understand relevant regulatory mechanisms, we generated a set of candidate driver transcription factors (TFs) using the frontal lobe CAGE-seq data. As CAGE-seq provides accurate maps of transcription start sites (TSSs) of genes, it can be used to locate promoter regions, which are usually in close proximity to the TSS. We expanded the regions around differentially expressed CAGE tag clusters and performed transcription factor binding site (TFBS) enrichment analysis in these expanded regions using the Homer software [75] and motifs from the JASPAR database [207](see Methods for details, similar to [11]). We selected all TFs with significant enrichment (P-value $\leq$ 0.001) for either up-regulated or down-regulated CAGE clusters as candidate regulators. We considered genes as potential targets of a TF if a TFBS could be found in their promoter region. While TF expression is not necessarily a good measure for activity, we selected only TFs with some evidence for differential expression in the RNA-seq data (adj. P-value < 0.05, not filtered for cell composition) to focus on the TFs with the most evidence of up- or down-regulated activity. TEAD2 is the only predicted active TF common to FTD-GRN and FTD-MAPT (Fig. 3.2E), while there is greater overlap among inactive TFs (here: inactive TF = has down-regulated targets, Fig. 3.2F).

DNA methylation is another important regulatory mechanism that can affect gene expression. We used the methylation data from the frontal lobe to examine epigenetic changes in FTD using the Illumina Infinium EPIC array, which measures over 850,000 CpG sites. We considered only the most variable CpG sites (28,173) and corrected for confounding variables using surrogate variable analysis to perform differential methylation analysis (see Methods). We detected 18,126, 9,478 and 5,282 significantly differentially methylated positions (DMPs) for FTD-MAPT, FTD-GRN, and FTD-C9orf72, respectively (Fig. 3.2G). The largest epigenetic aberrations appear to occur in FTD-MAPT. The C9orf72 repeat expansion is known to be associated with hypermethylation, [214] which we wanted to confirm in our data. Indeed, a CpG site located at the 5'-end of the C9orf72 gene, only 14 bp away from the repeat expansion, is hypermethylated, (Fig. 3.3A). Due to a low variance, this CpG site does not survive our variance filtering approach. Pathway enrichment analysis of genes in proximity to DMPs yielded enrichment of genes involved in nervous system development for hypermethylated CpG sites and genes involved in system development for hypomethylated CpG sites (Fig. 3.3C). As hypermethylation at CpG sites at promoter regions is associated with decreased expression, these results are in agreement with the findings from the RNA-seq pathway enrichment analysis. In part, the strong enrichment for nervous system development might also be caused by remnants of cell composition bias that could not be corrected entirely by surrogate variable analysis. We then performed biological age prediction with the methylation data to detect signs of accelerated aging in FTD using the Wenda algorithm [73]. Wenda underestimates the predicted age for all groups, albeit to a lesser extent for FTD groups (Fig. 3.3B). These results could therefore indeed be interpreted as signs for accelerated aging in FTD.

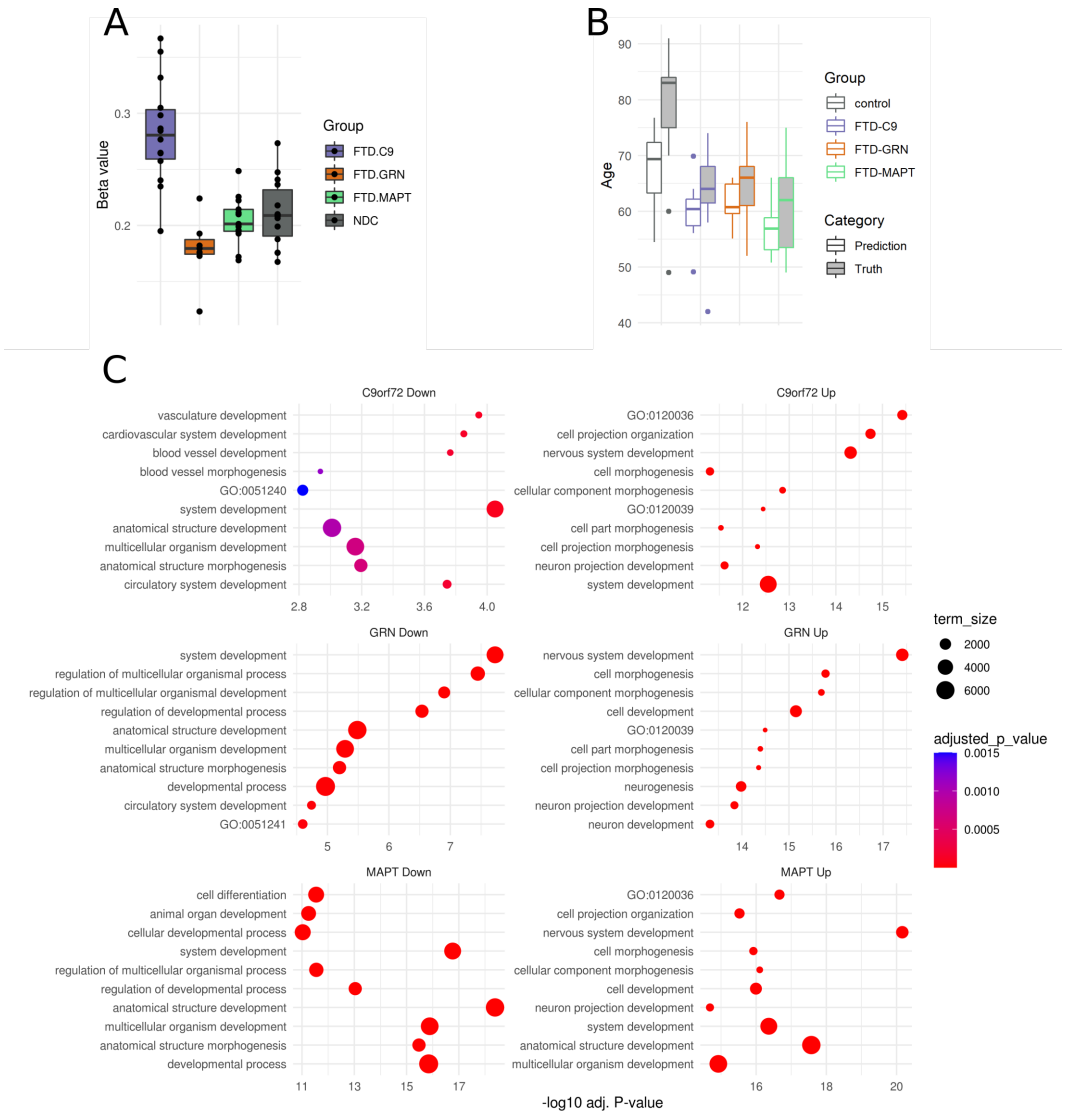

Figure 3.3: Differential methylation in FTD. **A** Methylation beta values of different disease groups at a CpG locus proximal to the 5'-end of the C9orf72 gene. **B** Actual and predicted age using the Wenda algorithm, colored by different disease groups. **C** Top 10 pathway enrichment results for gene in proximity to hyper- and hypo-methylated CpG sites in all three disease groups.

3.3.3 Loss of excitatory neurons and enrichment of endothelial cells in FTD

We then inspected cell composition changes in FTD genetic subtypes using the results from the RNA-seq deconvolution analysis (Methods). Owing to the prominent decrease of neuronal fractions (Fig. 3.10A), virtually all other cell types show increasing percentages. We therefore calculated the percentage difference of average fractions for all cell types between FTD groups and controls. This allows to compare the increase of fractions between different cell types. Strongest neuronal loss was observed for FTD-GRN, followed by FTD-MAPT and FTD-C9orf72 (Fig. 3.4A). This is in accordance with studies which have shown that the frontal lobe is most strongly affected in FTD-GRN [28, 222, 39]. According to our analysis, neuronal loss can be primarily attributed to loss of excitatory neurons, while fractions of inhibitory neurons stay relatively stable. This is in agreement with recent studies that found excitatory neurons to be especially vulnerable to tau pathology [64] and detected an important role of glutamatergic neurotransmission in FTD [142, 134]. Excitatory glutamatergic neurons might therefore be especially vulnerable in this disease. Closer examination of the KEGG pathway 'glutamatergic synapse' suggests that AMPA receptors are mainly affected, while we could not see signs of dysregulation for NMDA receptors (Fig. 3.12).

To validate the results from the computational deconvolution, we correlated fractions of excitatory neurons with manually determined pathology scores (Fig. 3.4B). Loss of excitatory neurons shows strong negative correlation with pathology scores (Pearson's correlation coefficient = -0.78, P-value = 2.8e-07). Deconvolution results therefore agree very well with experimental findings, as neuronal loss is expected to correlate with more severe pathology.

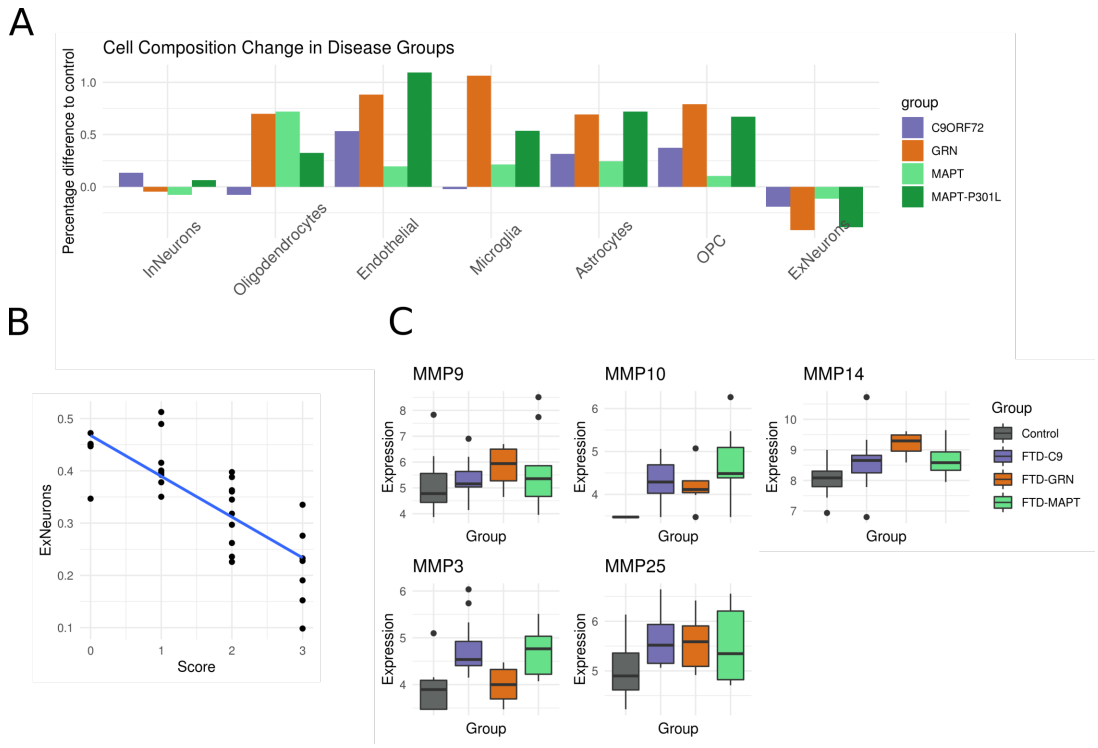

Figure 3.4: Cellular composition changes and matrix metalloproteinases in FTD. **A** Depicted is the difference between average fractions of a certain groups compared to the average of controls. The difference is calculated as percentages. Hence a value of 1.0 indicates a 100% increase compared to the controls. **B** Linear regression of excitatory neuron fractions against pathology scores given by a pathologist. **C** Normalized expression values for selected matrix metalloproteinase genes in different FTD diseas groups.

The strongest growth in percentage compared to the baseline is observed for endothelial cells in most disease groups, with the notable exception of FTD-GRN, where microglial cells show the strongest increase. When regressing neuronal fractions against other cell type fractions, endothelial cells have the worst fit (R2 = 0.42), hence their fractions are less well explained by neural loss. Circulatory system development is an enriched biological process in all three disease groups, hence suggesting a potential involvement of angiogenic pathways in the disease. The role of the vasculature in FTD is relatively unexplored, however, Bennet et al. recently found increased vasculature growth in mouse models of FTD-MAPT with P301L mutation [25]. The investigated mice showed increased vessel growth, albeit vessels were of smaller diameter and often blocked. Interestingly, endothelial enrichment in FTD-MAPT is particularly strong in patients with P301L mutation (Fig. 3.4B). Another recent study observed a particular microvascular structure with increased frequency in brains of patients with frontotemporal lobar degeneration (FTLD) [56]. The authors hypothesized that these structures might be caused by increased angiogenesis. Many genes involved in circulatory system devel-

opment are up-regulated in FTD-GRN and FTD-MAPT. The cell deconvolution therefore points towards a novel involvement of the vascular system in FTD, which should be further investigated.

The prominent increase of microglial fractions in FTD-GRN highlights the strong immune system component of this disease subtype. GRN is highly expressed in microglia and well-known for having important functions in the immune response [120]. The here observed strong enrichment of microglia points toward increased microglia activity in FTD-GRN, which is not observed in similar strength for any of the other FTD genetic subtypes.

3.3.4 Matrix Metalloproteinases are up-regulated in FTD

Next, we examined the DEGs with the largest fold-changes, as these depict good candidates for genes that drive pathologic mechanisms. In both FTD-GRN and FTD-MAPT, the genes with the largest absolute log fold changes are down-regulated. In FTD-MAPT, the gene UTP14C has a log-fold change of - 21.9. UTP14C is thought to be important for spermatogenesis - its contribution to FTD pathology is unclear. In FTD-GRN, the three genes with the largest absolute fold changes are TBC1D3 (-30.0), TBC1D3F (-29.9) and KRT17P7 (-29.7). The latter is a pseudogene, while the two former genes are GTPase activating proteins for the gene RAB5 and involved in vesicle-mediated transport, making them interesting candidates for further investigation given the known importance of transport in FTD.

Both for FTD-GRN and FTD-MAPT, matrix metalloproteinase enzymes (MMPs) are among the DEGs with the largest fold-changes (Fig. 3.4C, Fig. 3.10B). Specifically, MMP10 is among the top 20 DEGs for FTD-GRN and MMP9 for FTD-MAPT. MMP14 is differentially expressed in all groups, MMP9 and MMP25 are DE in FTD-GRN and FTD-MAPT (Fig. 3.4C). While not DE in every group, MMP10 has positive fold-changes in all groups and is barely detectable in control samples, which is consistent with current knowledge as MMPs are only expressed at basal levels in the adult brain [33]. We therefore looked for signs of MMP deregulation in our neural cell models and found MMP15 and tissue inhibitor of metalloproteinases 2 (TIMP2) to be significantly down-regulated in FTD-GRN neurons (Fig. 3.11). In FTD-MAPT neurons we find elevated levels of MMP2, albeit only in the group with IVS10+16 mutation, which, however, is different from the mutations in the post-mortem brain data.This confirms what was found by Biswas and colleagues for this mutation [28]. The authors could show that inhibition of MMP2 and MMP9 protected neurons from stress-induced cell death. Finally, the metalloproteinase Mmp12 is up-regulated in 21 month old GRN knockout mice, confirming that insufficient GRN levels can lead to increased MMP expression. Elevated RNA levels of MMP genes have been reported for many neurodegenerative diseases, indicating their importance in neurodegenerative mechanisms [164, 95].

This prominent enrichment of extracellular matrix associated pathways among up-regulated genes might be a cause or consequence of the here observed strong up-regulation of MMPs which can degrade ECM components. MMPs are furthermore tightly involved in the inflammatory response, and can activate the tumor necrosis factor (TNF) gene, for instance [199]. Several studies have shown promising effects of MMP inhibition in model systems [60].Inflammatory cytokines, reactive oxygen species (ROS) or hypoxia can lead to the activation of MMPs, among others [157, 33], which depicts a potential connection to the prominent down-regulation of mitochondrial pathways. Dysfunctional mitochondria are a primary source of ROS and might therefore lead to the activation of MMPs [90]. Moreover, MMPs can digest the extracellular matrix (ECM) and stimulate increased production of growth factors, thereby promoting growth of blood vessels [157]. Increased MMP expression might therefore be a cause of the observed blood vessel development and enrichment of endothelial cells.

The precise role of MMPs in FTD remains relatively unexplored but the frequent observation of MMP up-regulation in neurodegenerative disorders and their tight involvement in crucial processes such as neuroinflammation, highlights MMPs as interesting candidates for potential drug targets.

3.3.5 Co-expression Module Analysis

To better understand transcriptional changes and regulatory mechanisms, it is often helpful to cluster genes into modules with similar expression or function. Therefore, we performed tissue-specific functional module detection with HumanBase [104]. The HumanBase platform uses community detection algorithms to find gene clusters given a list of genes by using tissue-specific gene-gene associations that have been compiled using massive public datasets, which allows to link both genes with and without annotation to functional pathways. We divided DEGs into up- and down-regulated genes as we were looking for active and repressed modules in FTD. Resulting modules are tested for enrichment of biological process terms. The module enrichment results agreed with the DEG-based enrichment analysis (Tables 3.3, 3.4, 3.5, 3.6), as up- and down-regulated modules were enriched for similar pathways compared to up- and down-regulated DEGs. Then, to determine in which cell types the detected modules are of most importance, we performed expression weighted cell type enrichment (EWCE) analysis [179]. EWCE determines significant enrichment of cell type specific genes in a given geneset (see Methods). Both for FTD-MAPT and FTD-GRN, most modules show specificity for a few cell types (Fig. 3.5 C and D). Up-regulated modules in both groups are significantly enriched for endothelial genes (P-value < 0.1). Genes within these modules have been associated with blood vessel development (FTD-MAPT M6-up) and endothelial cell growth (FTD-GRN M4-up) by HumanBase (Fig. 3.5 A and B).

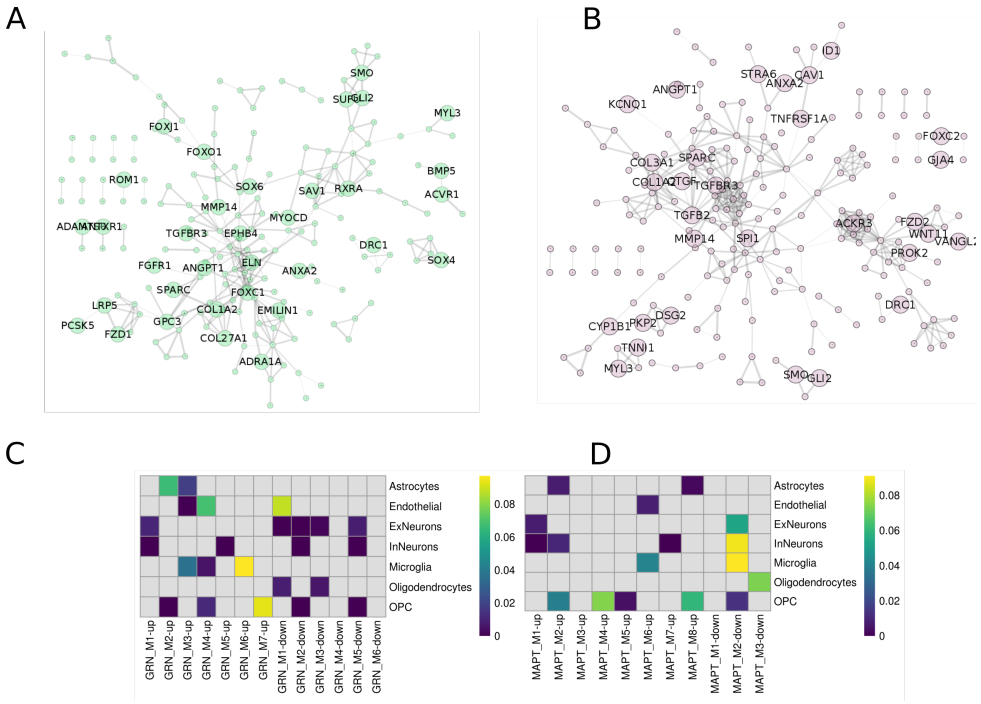

Figure 3.5: Cell type specificity of modules and circulatory system development related modules. **A**, **B** PPI networks of genes up-regulated in FTD-MAPT and FTD-GRN (log-fold-change ¿ 1), respectively. Genes involved in the biological process "circulatory system development" are highlighted. **C**, **D** EWCE analysis results for HumanBase modules of FTD-GRN and FTD-MAPT, respectively.

3.3.6 Increased inflammatory response in FTD-GRN

Given the prominent increase in microglial cell fractions in FTD-GRN and the strong enrichment of immune system pathways among up-regulated genes, we wanted to further investigate neuroinflammation in FTD-GRN. First, we examined FTD-GRN modules for enrichment of immune system-related terms. Indeed, several up-regulated modules are enriched for genes related to the immune system, whereas we could not find enrichment among down-regulated modules. The module FTD-GRN M1-up contains genes important for neutrophil migration and response to interleukines (Fig. 3.6 A, D). Both modules M3-up and M4-up contain genes relevant to NF-kappa-B (NFkB) signaling, as well as genes involved in tumor necrosis factor (TNF) production (Fig. 3.6 B and C, respectively). Finally, the module M6-up is enriched for genes involved in T cell activation. Modules M3-up, M4-up and M6-up are furthermore enriched for microglial-specific genes (Fig. 3.6 D). Interestingly, several necroptosis-related genes are up-regulated (M1-up: TLR3, TLR8, RIPK3; M4-up: RIPK2), suggesting this pathway as a potential driver

of neuronal death. The necroptosis cell death pathway is deregulated in several neurodegenerative disorders [224], and a recent study has shown that TBK1, a genetic cause of ALS and FTD (here down-regulated in FTD-GRN), is an endogenous inhibitor of RIPK1, an upstream regulator of RIPK3 [216]. The authors showed that embryonic lethality of TBK1-knockout mice is dependent on RIPK1 activity. RIPK1-dependent apoptosis therefore likely plays a pivotal role in FTD/ALS caused by TBK1 mutations. It will be important to further determine whether GRN mutations lead to alternative routes of necroptosis (e.g. via RIPK3). In a recent review, Molnár and colleagues have discussed several available drugs that could potentially regulate necroptosis [131], highlighting the potential of this pathway as a drug target.

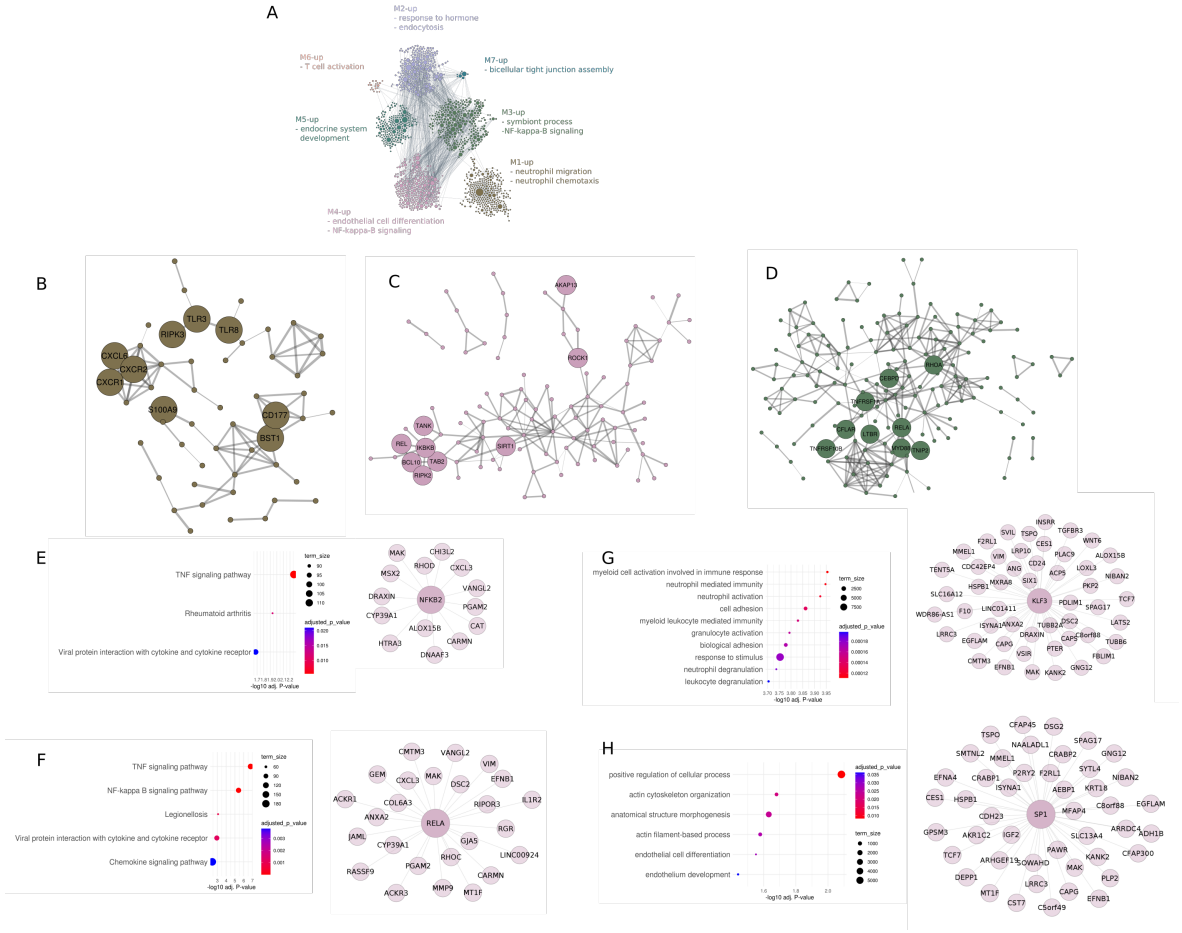

Figure 3.6: Neuroinflammation in FTD-GRN. **A** Up-regulated HumanBase modules in FTD-GRN with most significant terms. **B** Protein-protein interaction network (PPI, from STRING-DB) of FTD-GRN M1-up module. Genes involved in necroptosis, interleukin response and neutrophil migration are indicated. **C** PPI network of FTD-GRN M4-up module. Genes involved in NFkB signaling are indicated. **D** PPI network of FTD-GRN M3-up module. Genes involved in NFkB signaling and CEPBD are indicated. **E**, **F** KEGG pathway enrichment of predicted targets of TFs NFKB2 and RELA, respectively. **G**, **H** GO:BP pathway enrichment of predicted targets of TFs KLF3 and SP1, respectively.

To identify potential regulatory drivers behind the active immune response in FTD-GRN, we inspected predicted active TFs in FTD-GRN derived from the CAGE-seq data (Fig.3.2E). Interestingly, the TFs Nuclear Factor Kappa B Subunit 2 (NFKB2) and RELA (previously NFKB3) are predicted as active TFs in FTD-GRN. NFKB2 and RELA together form the NFkB transcription factor complex. RELA is a member of the M3-up module. Enrichment analysis of predicted NFKB2 and RELA targets in FTD-GRN us-

ing go:Profiler revealed TNF signaling and NFkB signaling as the most significantly enriched KEGG pathways (3.6E), suggesting NFKB2 and RELA as potential driver genes of increased TNF signaling. Furthermore, enrichment analysis indicated targets of the TFs SP1 and KLF3 as highly enriched among genes in the FTD-GRN M3-up module. Both TFs are also predicted as active in our analysis. Predicted KLF3 targets are enriched for genes involved in myeloid cell activation(3.6G) . SP1 target genes do not show a strong enrichment, but have roles in actin cytoskeleton organization and endothelial cell differentiation, among others (3.6H). The TF CCAAT Enhancer Binding Protein Delta (CEPBD) is another predicted active TF in FTD-GRN and member of the M3-up module. Enrichment analysis of CEPBD targets did not reveal any significant enrichment, however CEPBD is known to be an important regulator of inflammation, and has been implicated in neuroinflammation, for instance in Alzheimer's disease (AD)[98, 99]. Studies have shown that CEPBD can increase oxidative stress through reactive oxygen species (ROS) production [204]. Moreover, CEPBD can potentially activate MMPs [152].

To closer examine which parts of the NFkB and TNF signaling pathways are affected in FTD-GRN and in FTD in general, we inspected the fold-changes of genes of the corresponding KEGG pathways. Interestingly, the pro-inflammatory cytokine Interleukin 1 Beta (IL1B) is down-regulated in all disease groups, although only significantly in FTD-MAPT (Fig. 3.13A). Typically, IL1B expression is reported to be increased in neurodegenerative diseases [174]. Similarly, the inflammatory cytokine Interleukin 6 (IL6) has negative fold-changes in all disease groups. Downstream effector genes with positive fold-changes include multiple chemokines, Interleukin 18 Receptor 1 (IL18R1) and several metalloproteinases. Up-regulation of NFkB- and TNF-signaling genes is stronger in FTD-GRN (Fig. 3.13B). The TNF Receptor Associated Factor 3 (TRAF3) gene has negative fold-changes in all disease groups, although only reaching significance in FTD-GRN. The Dynamin-like protein 1 (DRP1), a downstream effector gene of the necroptosis pathway has negative fold-changes in all disease groups, suggesting necroptosis to either not be the primary cell death pathway, or necroptosis-induced cell death in a DRP1-independent manner [133].

We furthermore investigated predicted targets of down-regulated miRNAs for functions in immune system pathways. However, we could not detect any significant enrichment in relevant pathways, which indicates that neuroinflammation in FTD-GRN is not regulated by miRNAs. Similarly, genes proximal to hypomethylated CpG sites are not enriched for immune system related pathways, excluding epigenetic alterations as a major driver for neuroinflammation. While we did not detect prominent signals for neuroinflammation in FTD-MAPT, the FTD-MAPT module M3-up contains several genes involved in T cell and TNF signaling (EZR, RAB29, CARD8, HIPK1). However, the data suggests that neuroinflammation in FTD-MAPT and FTD-C9orf72 is much less prominent compared to FTD-GRN.

3.3.7 Energy Metabolism is impaired in FTD

Among the most significantly down-regulated pathways in FTD-GRN and FTD-MAPT are several pathways involved in energy metabolism and oxidative phosphorylation (Fig. 3.2D). We therefore hypothesize that dysfunctional mitochondria are common features of genetic subgroups in end-stage FTD. The modules FTD-GRN M1-down and FTD-MAPT M1-down are both most significantly associated with the term NADH dehydrogenase complex assembly (Tables 3.3, 3.4). Notably, the FTD-GRN M1-down module furthermore contains the mitochondrial gene CHCHD10 which is a genetic cause of FTD [20]. While not part of the corresponding FTD-MAPT module, CHCHD10 is also significantly down-regulated in FTD-MAPT (adj. P-value: 0.027, log fold-change: -0.56). This adds further evidence to the importance of CHCHD10 and moreover mitochondrial function in FTD pathogenesis.

Further inspection of the FTD-MAPT and FTD-GRN M1-up modules revealed that they contain several NADH:Ubiquinone Oxidoreductase Subunit genes (Fig. 3.7 B and C), which are necessary for functional oxidative phosphorylation and hence energy production. Interestingly, both down-regulated modules are significantly enriched for genes involved in intracellular transport, the FTD-GRN module is furthermore enriched for autophagy-related genes (Fig. 3.7). Impaired transport of mitochondria is a well-known feature of neurodegenerative diseases [61]. Especially large neurons are vulnerable to both mitochondrial dysfunction and impaired cellular transport, as they need to transport mitochondria to the ends of their lengthy axons [63]. The M1-down modules provide a potential link between mitochondrial dysfunction and impaired energy production.

Further evidence for the importance of the M1-down modules in FTD pathogenesis and neurodegeneration in general comes from several genes of the FTD-GRN M1-down module that are known to play important roles in neurodegeneration. Apart from CHCHD10, the module also harbours the genes Superoxide Dismutase 1 (SOD1), Dynactin Subunit 1 (DCTN1), PTEN Induced Kinase 1 (PINK1) and Huntingtin (HTT). All of these genes show lower expression values in all genetic subgroups, albeit they do not reach significant levels in every group (Fig. 3.7A). SOD1 encodes an enzyme that protects against superoxide radicals and mutations in SOD1 have long been known to cause Amyotrophic lateral sclerosis and TDP-43 inclusions [163]. Mutations in DCTN1 also lead to TDP-43 pathology and to a neurodegenerative disease called Perry syndrome (PS) [102]. DCTN1 is involved in axonal transport and can also cause FTD-like symptoms in some patients. PINK1 is a mitochondrial kinase that protects cells from dysfunctional mitochondria, and mutations in PINK1 are a well-known cause of Parkinson's disease (PD) [194]. Finally, poly-Q repeat expansions in the gene HTT cause Huntington's disease (HD), another neurodegenerative disorder. Evidence suggests that HTT is important for many mitochondrial functions, hence dysfunction of this gene leads to mitochondrial deficits [217].

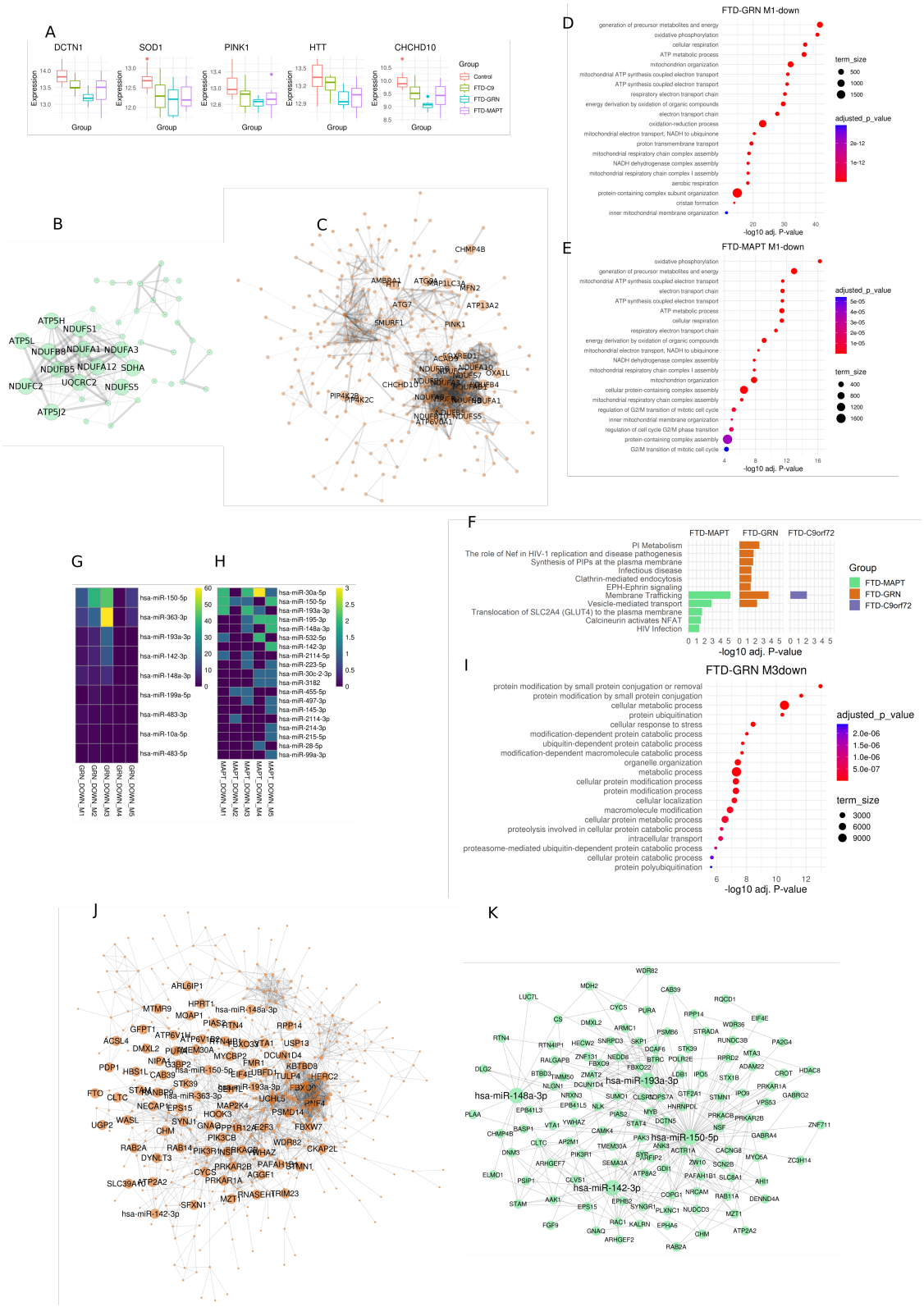

Figure 3.7: Impaired oxidative phosphorylation and cellular trafficking in FTD. **A** Expression levels (variance stabilized with DESeq2) of the genes CHCHD10, PINK1 and HTT in different groups. **B** STRING-DB PPI of FTD-MAPT M1-down module. Genes involved in oxidative phosphorylation are indicated. **C** PPI of FTD-GRN M1-down module. Genes involved in NADH dehydrogenase complex assembly and mitophagy are indicated, as well as CHCHD10.

3.3.8 Cellular trafficking pathways are inhibited by miRNAs

Cellular transport is thought to play a key role in FTD pathogenesis. Impaired trafficking can affect protein and mitochondria homeostasis. Here, we have shown that mitochondria function is strongly impaired in end-stage FTD and that transport pathways are tightly connected to this pathology. Next, we wanted to identify potential regulatory molecules that might be responsible for the observed changes. Based on literature research, we could not detect an obvious link between our set of predicted active TFs and cellular transport pathways, suggesting that TF regulation is not a primary driver of dysfunctional cellular transport. Pathway enrichment analysis of genes proximal to hypermethylated CpG-sites revealed strong enrichment for neuronal system related pathways (Fig. 3.3C) in FTD-MAPT and FTD-GRN, yet no enrichment for cellular trafficking pathways. We then performed pathway enrichment analysis with predicted targets of up-regulated miRNAs in all disease groups. Interestingly, for all groups, the most significantly enriched biological process (GO:BP) is cellular localization and the most significant Reactome pathway is membrane trafficking (Fig. 3.7F). Up-regulated miRNAs in FTD therefore seem to primarily target cellular transport pathways, and might play important roles in dysfunctional transportation.

To detect modules and genes predominantly targeted by up-regulated miRNAs, we calculated the intersection-over-union (IoU) of up-regulated miRNA targets with down-regulated modules for FTD-GRN and FTD-MAPT (Fig. 3.7 G and H). In both groups, the miRNA hsa-miR-150-5p has large IoU values with several modules, specifically with the FTD-GRN modules M2-down and M3-down, and the FTD-MAPT modules M1-down and M5-down. In FTD-GRN, the M3-down module is most strongly targeted by miRNAs, while in FTD-MAPT, several modules have large IoU values. We selected the FTD-GRN M3-down module for further inspection, as it is heavily targeted by miRNAs and contains genes involved in metabolic processes and cellular localization (Fig. 3.7I, Table 3.5). Five miRNAs have putative target genes in this module: hsa-miR-150-5p, hsa-miR-142-3p, hsa-miR-193a-3p, hsa-miR-148a-3p and hsa-miR-363-3p. All of these miRNAs are also significantly up-regulated in FTD-MAPT, although hsa-miR-363-3p does not survive the LFC cutoff we have used to define DE miRNAs (see Methods). We generated networks of the above mentioned candidate miRNAs combined with a PPI network of the FTD-GRN M3-down module (Fig. 3.7J) and a PPI network of all predicted targets in FTD-MAPT (Fig. 3.7K), as we could not detect a similar module in FTD-MAPT. In total, we observed 31 common putative miRNA targets in both networks. Among them are the genes CHM and RAB2A, which are involved in RAB GTPase signaling, an important pathway for cellular trafficking.

Next, we selected three miRNAs for further characterization in our cellular model systems: hsa-miR-193a-3p, hsa-miR-150-5p and hsa-miR-19b-3p (Fig. 3.8A). The former two miRNAs are DE in all three disease groups and have many targets among module

genes (Fig. 3.7 J K). The miRNA hsa-miR-19b-3p is up-regulated in all disease groups, although it does not reach significance after (FTD-MAPT and FTD-C9orf72) or before (FTD-GRN) multiple testing correction. Nevertheless, down-regulated genes were predicted to be enriched for targets of hsa-miR-19b-3p by g:Profiler, the miRNA is known to inhibit autophagy [230] and it is highly expressed in neurons. We performed RNA-seq experiments on iPSC-derived neurons and microglia (Methods) that were transfected with miRNA mimics and inhibitors for the three selected miRNAs. Inhibition of miR-150-5p in neurons had no detectable effect (2 DEGs). The miR-150-5p mimic resulted in 32 significantly down-regulated genes without relevant pathway enrichment (GO:BP 'involuntary skeletal muscle contraction, P-value 1.214×10-2). In microglia, however, the miR-150-5p mimic had strong effects, leading to 237 down-regulated and 236 up-regulated DEGs. Up-regulated DEGs are involved in cellular transport and myeloid activation (Fig. 3.8B). Up-regulation of miR-150-5p can thus lead to microglia activation. Down-regulated genes are enriched for morphogenesis and nervous system development pathways. Inhibition of miR-150-5p had even stronger effects (3221 DEGs), indicating an important function of this miRNA in microglia.

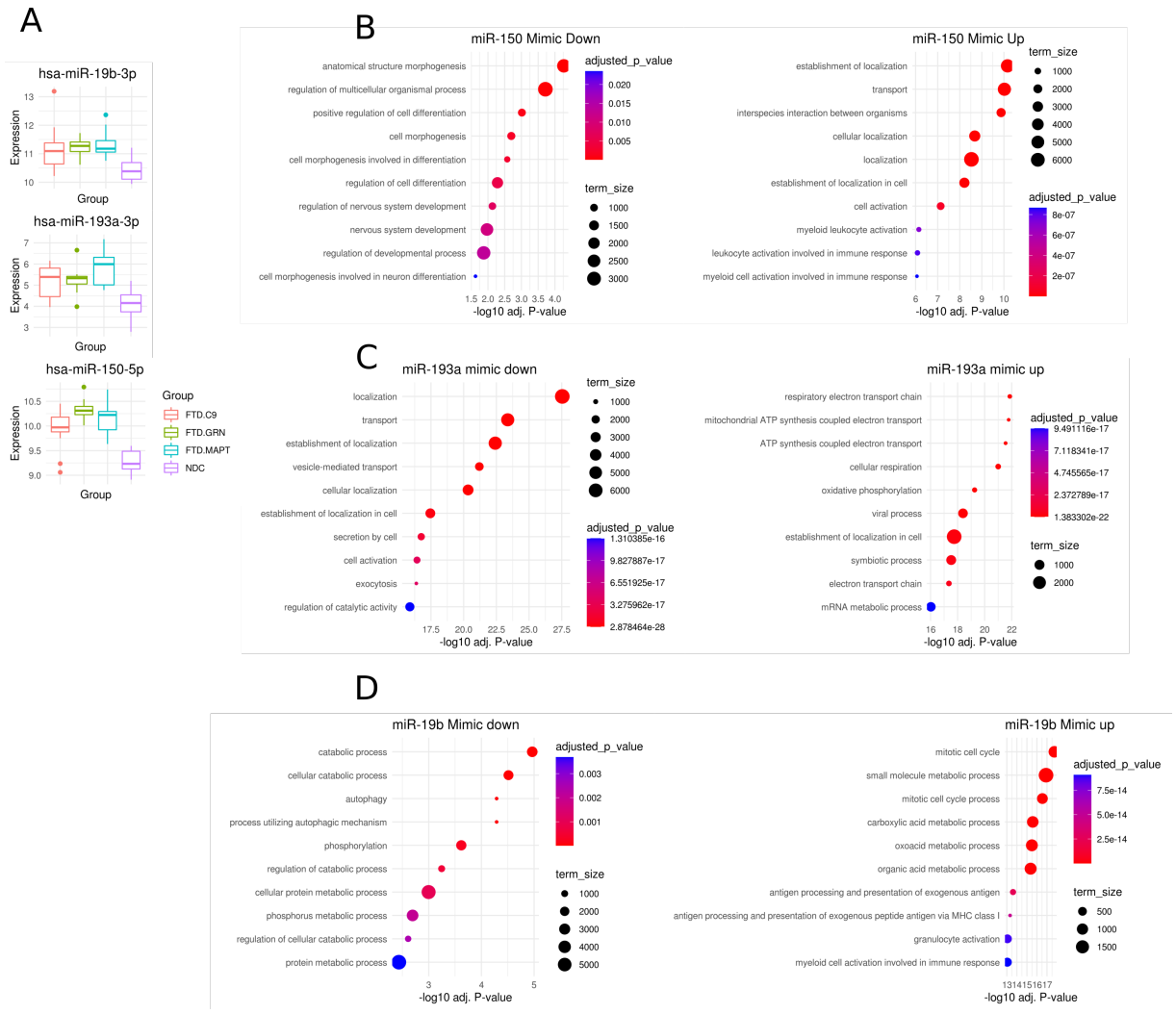

Figure 3.8: Effects of miRNA mimic and inhibitor experiments in iPSC-derived microglia. **A** Boxplots of normalized expression values for the selected miRNAs. **B**, **C**, **D** The top ten most significantly enriched biological processes of up- and down-regulated genes after transfection with mimics for miR-150-5p, miR-193a-3p and miR-19b-3p, respectively. Node size corresponds to the number of genes in the biological process term and node color corresponds to the P-value adjusted for multiple testing.

Transfection of miR-193a-3p mimic and inhibitor was only successful in microglia, where the mimic had strong effects with 1756 down-regulated and 1474 up-regulated genes. Up-regulated genes were enriched of mitochondrial functions like oxidative phosphorylation, while down-regulated genes were enriched for localization and vesicle-mediated transport pathways (Fig. 3.8C).

In neurons, the miR-19b-3p mimic resulted in 89 down- and 137 up-regulated DEGs

(inhibitor: 8 down-regulated, 31 up-regulated). Genes down-regulated by the mimic and up-regulated in the inhibitor experiment are involved in neuronal system pathways, enriched for miR-19b-3p targets according to g:Profiler and share 17 common genes, providing strong evidence for these genes to be regulated by miR-19b-3p. In microglia, stronger effects of the miR-19b-3p mimic compared to the inhibitor were observed (1518 compared to 608 DEGs), suggesting that this miRNA plays a minor role in microglia. Genes down-regulated by the miR-19b-3p mimic were enriched for catabolic processes, autophagy and also vesicle-mediated transport, up-regulated genes were enriched for cell cycle and immune system related genes (Fig. 3.8D). These results provide strong evidence that some of the up-regulated miRNAs in FTD brains target genes involved in cellular trafficking and autophagy pathways. Up-regulation of the microglial miRNA miR-150-5p can moreover potentially activate microglia and might therefore contribute to neuroinflammation.

3.4 Discussion

Here, we present the data from phase 1 of the RiMod-FTD project, a multi-omics, multi-model data resource for FTD research. Generated by the RiMod-FTD consortium during several years, the resource depicts a valuable tool for FTD researchers that will help to accelerate scientific progress towards a better understanding of relevant disease mechanisms in FTD. Additional multi-omics data from cell lines, mouse models and other brain regions will be added in future phases.

In this study, we have systematically analysed RNA-seq, CAGE-seq, smRNA-seq and methylation datasets from the GFM of patients with mutations in GRN, MAPT or C9orf72 and healthy controls. From all three genetic disease groups, we detected the largest transcriptional dysregulation for the FTD-GRN group, and only relatively small changes in FTD-C9orf72. In line with these findings, deconvolution analysis revealed that patients with GRN mutations are subject to the greatest neuronal loss from all groups. These results indicate either a potential selective vulnerability of frontal lobe cells to GRN mutations or more destructive neurodegeneration in FTD-GRN in general. Indeed, it has been previously observed that frontal lobe atrophy is more pronounced in FTD-GRN compared to FTD-MAPT and FTD-C9orf72 [208, 221].

Our deconvolution analysis furthermore indicates that excitatory neurons are affected the strongest in all genetic disease groups of FTD, which was confirmed in the RiMod-FTD proteomics data (Mediema et al., 2020) as well. Recently, evidence from multiple studies has accumulated pointing toward strong involvement of glutamatergic synapses in FTD [26]. While it has been previously reported that densities of both ionotropic glutamate receptors, AMPA and NMDA receptors, are reduced in post-mortem brain tissue of FTD patients [134], we see evidence for AMPA receptors to be particularly affected.

In brains of GRN mutation carriers, we have found increased microglial cell numbers and strong signals of neuroinflammation. These findings agree with recent studies which have shown that microglial burden is considerably stronger in FTD-GRN compared to the other genetic subtypes [168, 212]. Woolacott et al. furthermore observed that FTD-GRN cases had severe to very severe microglial dystrophy and an abundance of phagocytic and antigen presenting microglia. Microglial activation was also observed in FTD-C9orf72 and FTD-MAPT, albeit less prominent. Closer examination of neuroinflammation in FTD-GRN highlights the TFs NFKB2, RELA, KLF3 and SP1 as potential key inflammatory regulators, leading to activation of the NFkB- and TNF-signaling pathways. Our data further indicates that a potential downstream effect of the neuroinflammation is the activation of the necroptosis pathway. Negative fold-changes of the necroptosis-pathway gene DRP1 implies a DRP1-independent path to cell death. The necroptosis cell death pathway is deregulated in several neurodegenerative disorders [224], and a recent study has shown that TBK1, a genetic cause of ALS and FTD (here down-regulated in FTD-GRN), is an endogenous inhibitor of RIPK1, an upstream regulator of RIPK3 [216]. The authors showed that embryonic lethality of TBK1-knockout mice is dependent on RIPK1 activity, suggesting that the necroptosis pathway is indeed an important part of FTD pathogenesis. In a recent review, Molnár and colleagues have discussed several available drugs that could potentially regulate necroptosis [131], highlighting the potential of this pathway as a drug target. The particularly strong neuroinflammation in FTD-GRN compared to other genetic subgroups implies that anti-inflammatory treatments might be more effective and helpful in this subgroup. We have furthermore highlighted several genes and pathways that depict promising starting points for further investigation of neuroinflammation in FTD-GRN.

Pathway enrichment and deconvolution analyses have pointed toward increased blood vessel abundance and growth in FTD brains compared to controls. Enrichment for blood vessel associated proteins was also found in the RiMod-FTD proteomics data (Mediema et al., 2020). It is generally not known how and if the vasculature system is involved in FTD pathogenesis, although recent studies have observed abnormalities in a mouse model of tau pathology and post-mortem human brains [25, 56]. To our knowledge, angiogenesis as a pathological feature in several genetic FTD subtypes has not been reported before and therefore depicts an important subject to study for future FTD research.

In all three disease groups, we have observed prominent up-regulation of ECM pathways and MMP enzymes. While it has been increasingly recognized that MMPs are important regulators in many neurodegenerative diseases [33, 160], the role of MMPs in FTD pathogenesis has been of minor interest to most researchers in the field. Observing MMPs as strongly up-regulated in all genetic disease groups, however, suggests that they might be important regulators in end-stage FTD biology. In mouse models of ALS, inhibition of the MMPs MMP2 and MMP9 could indeed prolong survival and re-

duce symptoms [93, 201, 117]. Moreover, TIMP3, which is up-regulated in our data, was found to be partly responsible for neuronal apoptosis in a ALS model [111], which points towards TIMP3 as a potential apoptosis mechanism in FTD. Given their important biological functions and their involvement in all genetic FTD subgroups, we think it will be fruitful and important to further investigate how MMPs contribute to FTD and whether they can be exploited as drug targets.

We have observed strong enrichment for genes involved in oxidative phosphorylation and other mitochondrial pathways in FTD-MAPT and FTD-GRN among down-regulated genes. This implies that the energy metabolism is greatly affected at least in the end-stage of FTD. Given that mutations in genes important for mitochondria, such as SOD1 and CHCHD10, can lead to ALS or FTD, the involvement of energy metabolism pathways in the disease is highly likely.

Impaired cellular trafficking mechanisms is very likely a key feature of FTD pathogenesis and it has been shown multiple times that FTD-causal mutations lead to trafficking deficits [9, 213, 219]. However, it is not always clear which mechanisms continue to dysfunctional transport mechanisms. Here, using a multi-omics approach, we have found evidence that elevated expression of several miRNAs might contribute to the inhibition of genes important for cellular transport. For three selected miRNAs, we could experimentally validate that up-regulation leads to impaired cellular transport, autophagy or immune system pathways. Up-regulation of these miRNAs has thus clearly pathogenic potential and could contribute to disease progression. To our knowledge, we are the first to link up-regulated miRNAs with inhibited cellular transport pathways in FTD. Additional studies are necessary to further validate this hypothesis, which, if replicated, highlights the reported miRNAs as promising drug targets.

Using data from the multi-omics RiMod-FTD resource, we have identified several regulator candidates of potential importance in FTD pathogenesis. Data from future RiMod-FTD phases will help to further verify and refine those hypotheses and will thus help to advance the field of FTD research.

3.5 Methods

3.5.1 Donor samples employed in this study

Post mortem human brains

Tissues were obtained under a Material Transfer Agreement from the Netherlands Brain Bank, and additional samples were provided by the Queen Square Brain Bank of Neurological Disorders and MRC, King College London. Demographic details about human brain samples are summarized in Table 1. GFM and GTM tissue from each subject was

divided into three pieces for transcriptomic, proteomic and epigenetic experiments in a dry-ice bath using precooled scalpels and plasticware.

hIPS-derived NGN2 neurons and miRNA mimics and inhibitors transfection

smNPC were derived from hiPSc cells (Cell line id: GM23280 obtained from the Coriell Institute) using the protocol described in P Reinhardt et al. [156]. The differentiation protocol from smNPC to neurons involves over-expression of Neurogenin-2 (NGN2) using a modified version of the NGN2 lentiviral inducible vector system (single vector pLV_TRET_hNgn2_UBC_BSD_T2A_rtTA3). The detailed description about protocol, reagents and media composition is available in Dhingra et al. [50].

Briefly, stable NGN2 smNPC are grown for six days in expansion medium N2B27 supplemented with CHIR99021 (CHIR) 3 µM, Purmorphamine (PMA) 0.5 µM and L-ascorbic acid 2-phosphate magnesium (AA) 64 mg/l. For differentiation, cells are plated (80,000 cells/cm2) onto Poly L-orithine and laminin coated plates in N2B27 medium supplemented with doxycycline (dox) at 2.5 µg/mL, and 2 µM DAPT. On day 4 of differentiation, transfection was performed in n=3 replicate plates using lipofectamine RNAiMax (ThermoFisher Scientific) with a final concentration of miRNA mimic and inhibitors (miR-19b-3p and miR-1505p mimics and inhibitors from Qiagen and miR-193a-3p mimic and inhibitor from ThermoFisher Scientific) in the range of 5 to 10 nM as per the manufactures' guidelines along with their corresponding controls. Next day (day 5 of differentiation), the complete media was changed with N2B27 media supplemented with dox, 10 ng/mL brain-derived neurotrophic factor (BDNF), 10 ng/mL glial cell-derived neurotrophic factor (GDNF), 10 ng/mL neurotrophic factor 3 (NT-3), 1 µg/mL Laminin, and 10 µM DAPT. Thereafter, half media was changed on day 8 of differentiation. On day 11, cells were gently washed with PBS and processed for RNA isolation .

hIPS-derived microglia and miRNA mimics and inhibitors transfection

hiPSCs were differentiated as previously described (van Wilgenburg et al, 2013 [198]). In brief, 3 x 10$\hat{6}$ iPSCs were seeded into an Aggrewell 800 well (STEMCELL Technologies) to form embryoid bodies (EBs), in mTeSR1 and fed daily with medium plus 50ng/ml BMP4 (Miltenyi Biotec), 50ng/ml VEGF (Miltenyi Biotec), and 20ng/ml SCF (RD Systems). Four-day EBs were then differentiated in 6-well plates (15 EBs/well) in X-VIVO15 (Lonza) supplemented with 100ng/ml M-CSF (Miltenyi Biotec), 25ng/ml IL-3 (Miltenyi Biotec), 2mM Glutamax (Invitrogen Life Technologies), and 0.055mM beta-mercaptoethanol (Thermo Fisher Scientific), with fresh medium added weekly. Microglial precursors emerging in the supernatant after approximately 1 month were collected and isolated through a 40um cell strainer and plated in N2B27 media supple-

mented with 100 ng/ml M-CSF, 25 ng/ml interleukin 34 (IL-34) for differentiation. Thereafter, the media is reshred every 2 days supplemented with 100 ng/ml M-CSF, and 25 ng/ml IL-34. The cells were cultured for additional 6 days with media refresh every 2 days. On day 7 of maturation, transfection was performed in n=3 replicate plates using lipofectamine RNAiMax with a final concentration of miRNA mimics and inhibitors in the range of 5 to 10 nM as per the manufactures' guidelines along with their corresponding controls (miR-19b-3p and miR-1505p mimics and inhibitors from Qiagen and miR-193a-3p mimic and inhibitor from ThermoFisher Scientific). Next day complete media was refreshed. On day 11, cells were gently washed with PBS and processed for RNA isolation.

Tissue was homogenized in chilled hypotonic buffer (Tris- HCL 50mM pH 7.5, KCl 100mM, MgCl2 12mM, Nonidet-P40 1%, DTT 1mM (ThermoFischer), RNAse Out 200U (ThermoFischer), Protease Inhibitor (Roche) with a dounce tissue grinder in ice. After 5 minutes centrifugation at 1000g at 4°C supernatant was used for transcriptomic and proteomic experiments, while the pellet containing nuclei was reserved for epigenomics analysis.

3.5.2 Genetic analysis

Genomic DNA was isolated from 50 mg of GFM frozen brain tissue by using the Qiamp DNA mini kit (Qiagen) following the manufacturer protocol. DNA concentration and purity were assessed by nanodrop measurement. DNA integrity was evaluated by loading 100 nanogram per sample on a 0,8% agarose gel and comparing size distribution to a size standard. Presence of C9orf72-HRE in postmortem brain tissues and hIPS cells was confirmed by primed repeat PCR according to established protocols. Reported mutations for MAPT and GRN were verified by sanger sequencing.

3.5.3 Transcriptomic procedures

RNA isolation from human brain tissue

Total RNA for CAGE-seq and RNAseq was isolated from ±100mg of frozen brain tissue with TRIzol reagent (Thermo Fischer Scientific) according to the manufacturer recommendation, followed by purification with the RNeasy mini columns (Qiagen) after DNAse treatment.

Total RNA for smallRNA-seq was isolated from frozen tissue using the TRIzol reagent (ThermoFischer Scientific). After isopropanol precipitation and 80% ethanol rinsing RNA pellet was resuspended in RNAse free water and up to 10 micrograms of RNA was incubated with 2U of Ambion DNAse I (ThermoFischer) at 37°C for 20 minutes. DNA-free RNA samples were then further purified by phenol-chloroform-isoamyl-alchol extraction followed by ethanol precipitation.

RNA isolation from hIPS-derived NGN2 neurons

At day 11 of NGN2 driven differentiation hIPS-neurons were carefully rinsed with PBS and lysed in TRIzol LS reagent (ThermoFischer Scientific) according to the manufacturer recommendation. Further DNAse treatment and purification were carried out as described for the total RNA purification for smallRNAseq

RNA QC

For each RNA sample, RNA concentration (A260) and purity (A260/280 and A260/230) was determined by Nanodrop measurement and RNA integrity (RIN) was assessed on a Bioanalyser 2100 system and/or Tape station 41200 (Agilent Technologies Inc.)

CAGE-seq libraries

CAGE-seq libraries were prepared from 5 micrograms of RNA from frozen brain tissues according to a published protocol [188]. Libraries were sequenced on a HiSeq 2000 and/or HiSeq2500 on a 1x50 bp single read flow cell (Illumina) at an average of 20M reads/sample (I need to double check this value..)

RNAseq libraries

RNAseq libraries were prepared from 1 microgram of total RNA from frozen brain tissue and IPS-derived neurons using the TruSeq Stranded Total RNA with Ribo-Zero Gold kit (Illumina) according to the protocol specifications. RNAseq libraries from brain tissue were sequenced on a Hiseq2500 and HISeq4000 on a 2x100 bp paired end (PE) flow cell (Illumina) at an average of 100M PE/sample. RNA seq libraries from IPS-derived neurons were sequenced on a 300 cycles flow cell on a NextSeq550 sequencer (Illumina) at an average of 50M PE reads/sample

smallRNAseq libraries

Small RNA-sequencing was divided into two different parts for technical reasons, processed at the DZNE in Tübingen and Göttingen. For the Tübingen part, smallRNAseq libraries from frozen tissue and IPS-derived neurons were prepared starting from 2 micrograms of total RNA using the Nextflex Small RNA-seq kit v3 (Bioo Scientific). Libraries were sequenced on a NextSeq550 on a 75 cycles flow cell. The data generated in Göttingen was produced using the NEBNext Small RNA Library Prep Set from New England Biolabs Inc.

Methylation assay

To assess the methylation status of over 850000 CpG sites in promoter, gene body and enhancer regions we have used the MethylationEPIC bead chip arrays (Illumina). Bisulfite conversion of genomic DNA, genome amplification, hybridization to the beadchips, washing, staining and scanning procedure was performed by Atlas Biolabs (Atlas Biolabs, Berlin, Germany). Cases and controls DNAs were distributed randomly across each array.

HumanBase Module Analysis

Functional gene modules were generated using the HumanBase tool at: https://hb.flatironinstitute.org/. We split DEGs into up- and down-regulated genes and generated modules for the separated list with HumanBase. Modules were then downloaded for further analyses.

RNA-seq processing and analysis

Raw FastQ files were processed using the RNA-seq pipeline from nf-core (nf-core/rnaseq v1.3) [59], with trimming enabled. Gene quantification was subsequently done using Salmon (v0.14.1) [146] on the trimmed FastQ files. Alignment and mapping were performed against the human genome hg38. DESeq2 (v.1.26.0) [119] was used to perform differential expression analysis. We corrected for the covariates gender and PH-value. Genes were considered differentially expressed when having a Benjamini-Hochberg corrected P-value below 0.05.

Cell type deconvolution and filtering

We performed cell type deconvolution on the RNA-seq data using Scaden [125]. For training we used the human brain training dataset used in the Scaden publication. Each ensembl model was trained for 5000 steps. To filter differentially expressed genes for false positives caused by cell composition bias, we first calculated the correlation of gene expression with cell type fraction.Then, we calculated a cell type specificity score as defined in Skene et al. [178] for each gene available in the scRNA-seq dataset from Darmanis et al. [46]. We filtered out all genes that had a specificity score of at least 0.5 and a positive correlation of at least 0.4 with the cell type fractions of the most specific cell type. Relative changes in cell type composition were quantified by first calculating the average fractions of a cell type for all groups and then calculating the procentual change of cell fractions compared to the average control fractions. Functional module detection was done by dividing DEGs into up- and down-regulated genes and supplying these separately to HumanBase [69]. We selected ‘nervous system’ as tissue.

Cell type enrichment analysis

We performed cell type enrichment analysis for several genesets using the EWCE R package [179]. Cell type specificity of genes was calculated from the single-cell RNA-seq cortex dataset of Darmanis and colleagues [46]. EWCE analysis was done following instructions from https://github.com/NathanSkene/EWCE.

CAGE-seq processing and analysis

Sequencing adapters and barcodes in CAGE-seq FastQ files were trimmed using Skewer (v.0.1.126) [87]. Sequencing artefacts were removed using TagDust (v1.0) [109]. Processed reads were then aligned against the human genome hg38 using STAR (v.2.4.1) [52]. CAGE detected TSS (CTSS) files were created using CAGEr (v1.10.0) [70]. With CAGEr, we removed the first G nucleotide if it was a mismatch. CTSS were clustered using the 'distclu' method with a maximum distance of 20 bp. For exact commands used we refer to the reader to the scripts used in this pipeline: https://github.com/dznetubingen/cageseq-pipeline-mf.

Transcription factor activity analysis

To identify candidate regulatory transcription factors, we first performed differential expression analysis with all CAGE-seq clusters (see RNA-seq analysis). Then, we extracted the sequence 600 bp up-stream and 300 bp downstream around all detected clusters. We used Homer [75] to look for significant TFBS enrichment in the regions around up- and down-regulated clusters. When calculating enrichment, we considered all extracted regions that are not part of the set of interest as background. The complete pipeline can be found at https://github.com/KevinMenden/tf-activity.

smRNA-seq processing and analysis

After removing sequencing adapters, all FastQ files were uploaded to OASIS2 [153] for analysis. Subsequent differential expression analysis was performed on the counts yielded from OASIS2, using DESeq2 and correcting for gender and PH-value, as was done for the RNA-seq data. Additionally, we added a batch variable to the design matrix to correct for the two different batches of this dataset. For the target prediction analysis, we first downloaded all targets from mirBase [103]. Then, we correlated the expression of miRNAs with their predicted targets using matching samples from the RNA-seq data. We removed all predicted targets with a correlation above -0.4, thus only considering miRNA-target pairings with high negative correlation.

Methylation data processing and analysis

The Infinium MethylationEPIC BeadChip data was analyzed using the minfi R package [12]. We removed all sites with a detection P-value above 0.01, on sex chromosomes and with single nucleotide polymorphisms (SNPs). Data normalization was done using stratified quantile normalization. Sites with a standard deviation below 0.1 were considered uninformative and filtered out, to increase detection power. Surrogate variable analysis [112] was performed to determine confounding factors. Differential methylation analysis was done using the limma package [159] and controlling for the detected surrogate variables. Sites with a Benjamini-Hochberg [24] adjusted P-value below 0.05 were considered differentially methylated.

Age prediction

We predicted the biological age of donors using the methylation data and the Wenda algorithm [73]. Training data was kindly provided by the authors of Wenda. We subsetted the data for CpG sites found in our data (11,729) sites and performed the prediction as described at https://github.com/PfeiferLabTue/wenda.

Analysis of mRNA-seq data from cellular models

This section describes the analysis of mRNA-seq data generated for the miRNA mimic and inhibitor experiments and the iPSC-derived neurons from patients with mutations in MAPT, GRN or C9orf72. FastQ files were mapped and gene counts quantified using Salmon and differential expression analysis was performed with DESeq2 (see post-mortem brain RNA-seq analysis). DEGs were examined for pathway enrichment using go:Profiler. For iPSC-derived neurons from patients with mutations in GRN, DEGs were additionally submitted to HumanBase to identify deregulated functional modules.

Assessment of degeneration

For assessment of neurodegeneration, HE stained paraffin sections of the frontal and temporal cortex were graded as absent (0), mild (1), moderate (2) and severe (3) based on the presence of spongiosis, neuronal loss and gliosis.

3.6 Supplementary Material

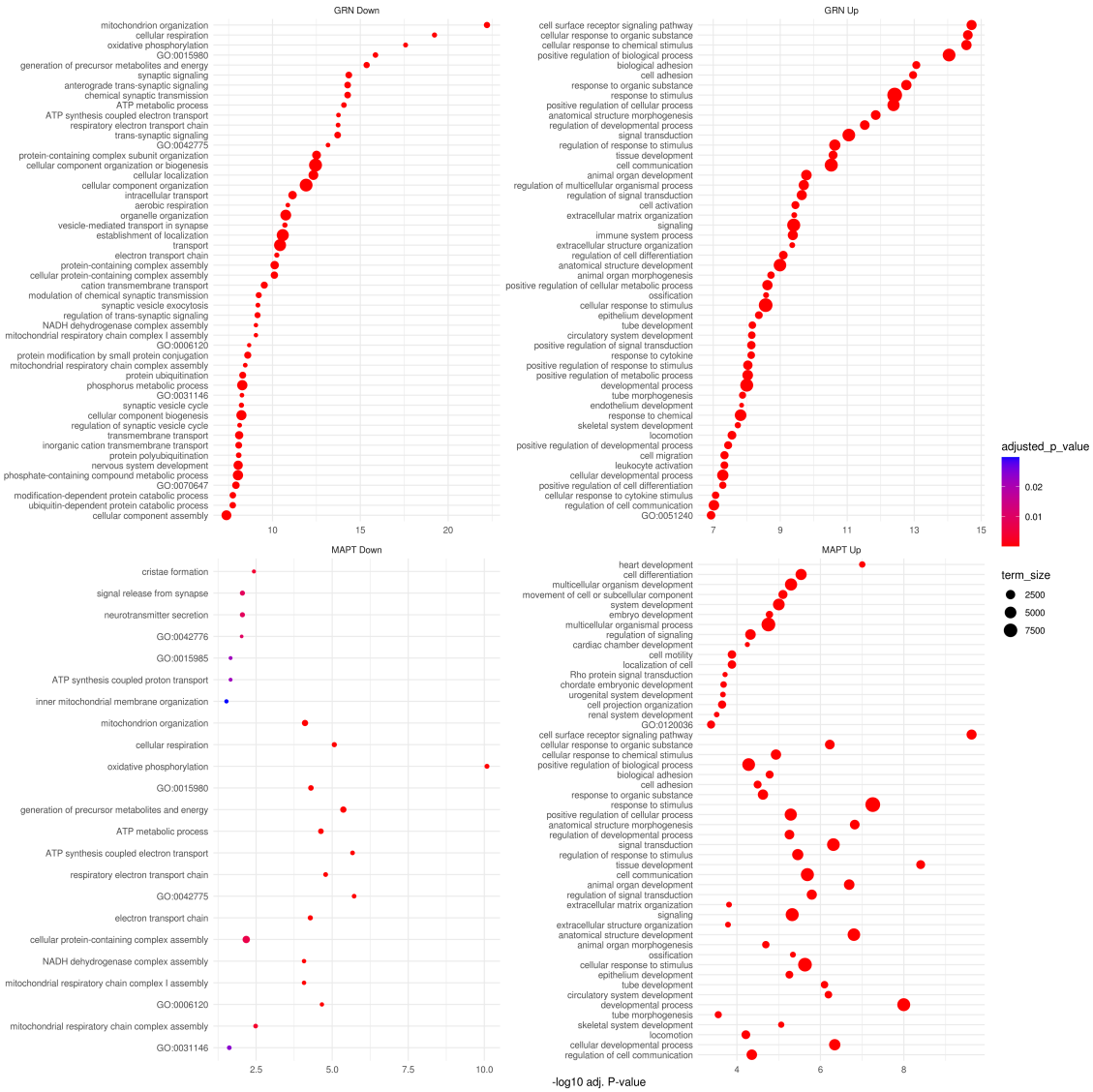

Figure 3.9: Pathway enrichment analysis of up- and down-regulated genes in FTD-MAPT and FTF-GRN. Shown are the 50 most significantly enriched gene ontology biological processes. Node size corresponds to genes in the respective pathway, node color corresponds to the multiple-testing adjusted P-value.

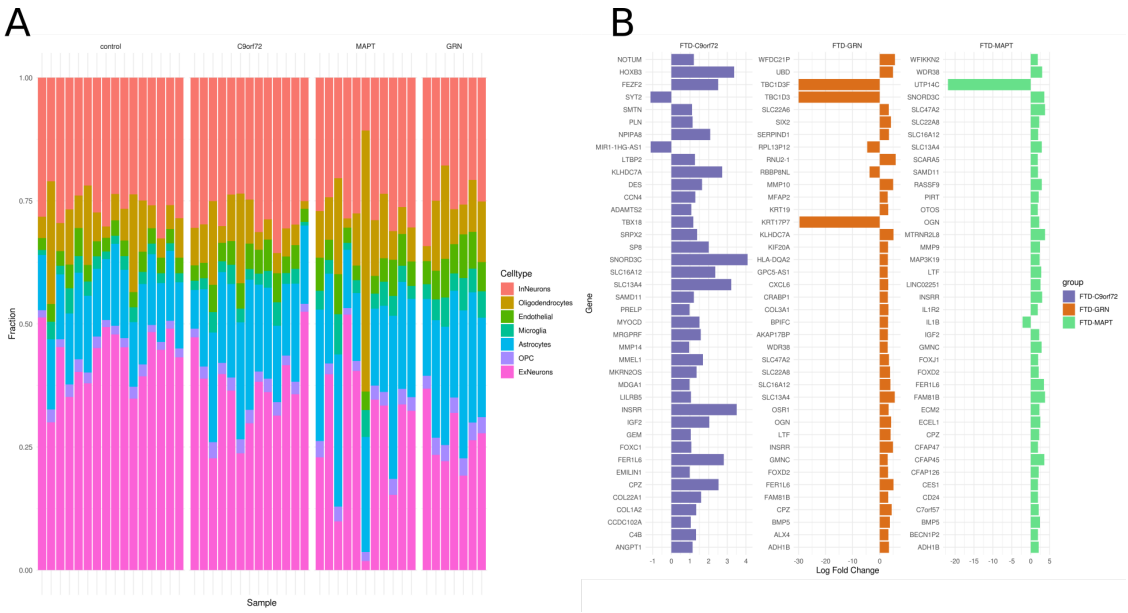

Figure 3.10: Cell composition and top differentially expressed genes in FTD. **A** Cell composition of individual samples as predicted by Scaden divided by disease group. Cell types are depcited in different colors. **B** The DEGs with the highest fold-changes are shown for FTD-C9orf72, FTD-GRN and FTD-MAPT (RNA-seq data).

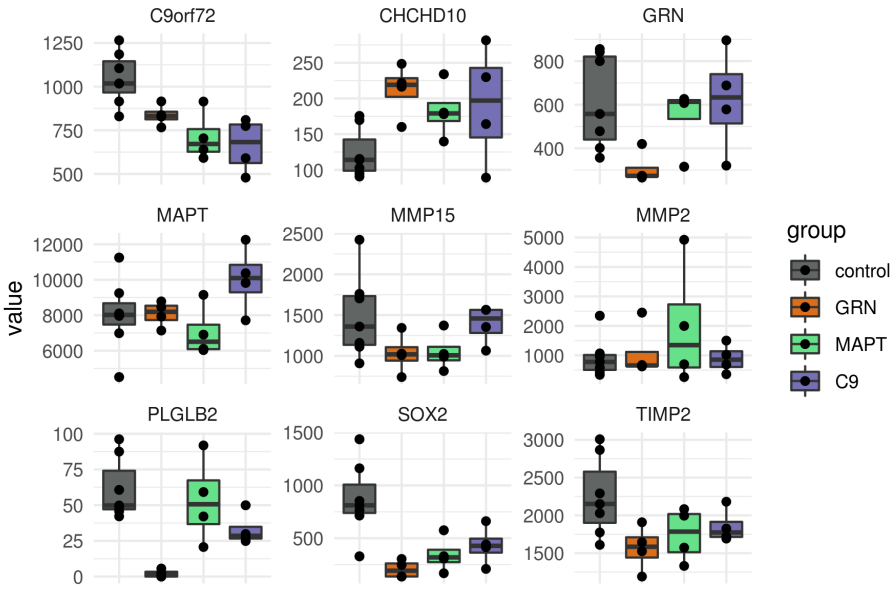

Figure 3.11: Expression values of selected genes in iPSC-derived neurons from FTD patientes carrying mutations in GRN, MAPT or C9orf72.

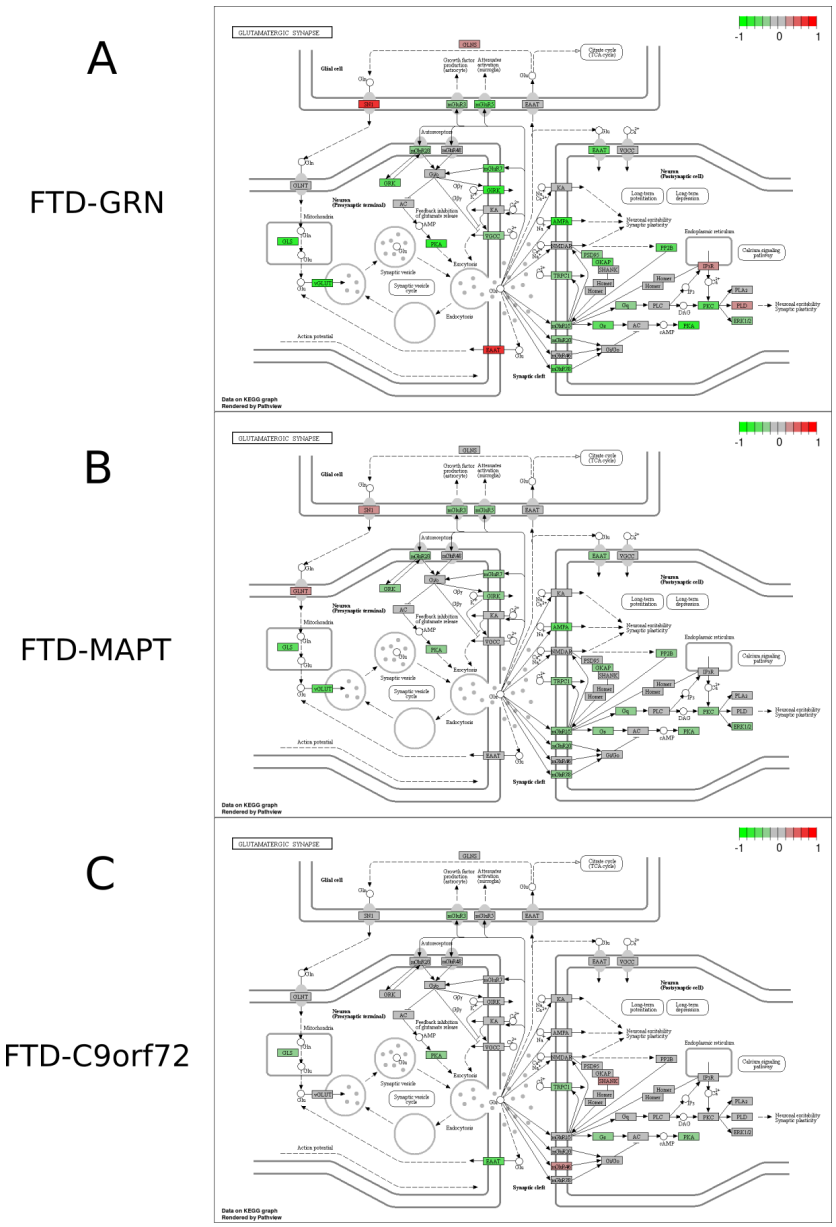

Figure 3.12: The glutamatergic synapse pathway in FTD (KEGG). **A**, **B**, **C** Shown are the average log fold-changes of genes corresponding to nodes in the KEGG pathway for FTD-GRN, FTD-MAPT and FTD-C9orf72, respectively. Genes were not filtered for significant differential expression.

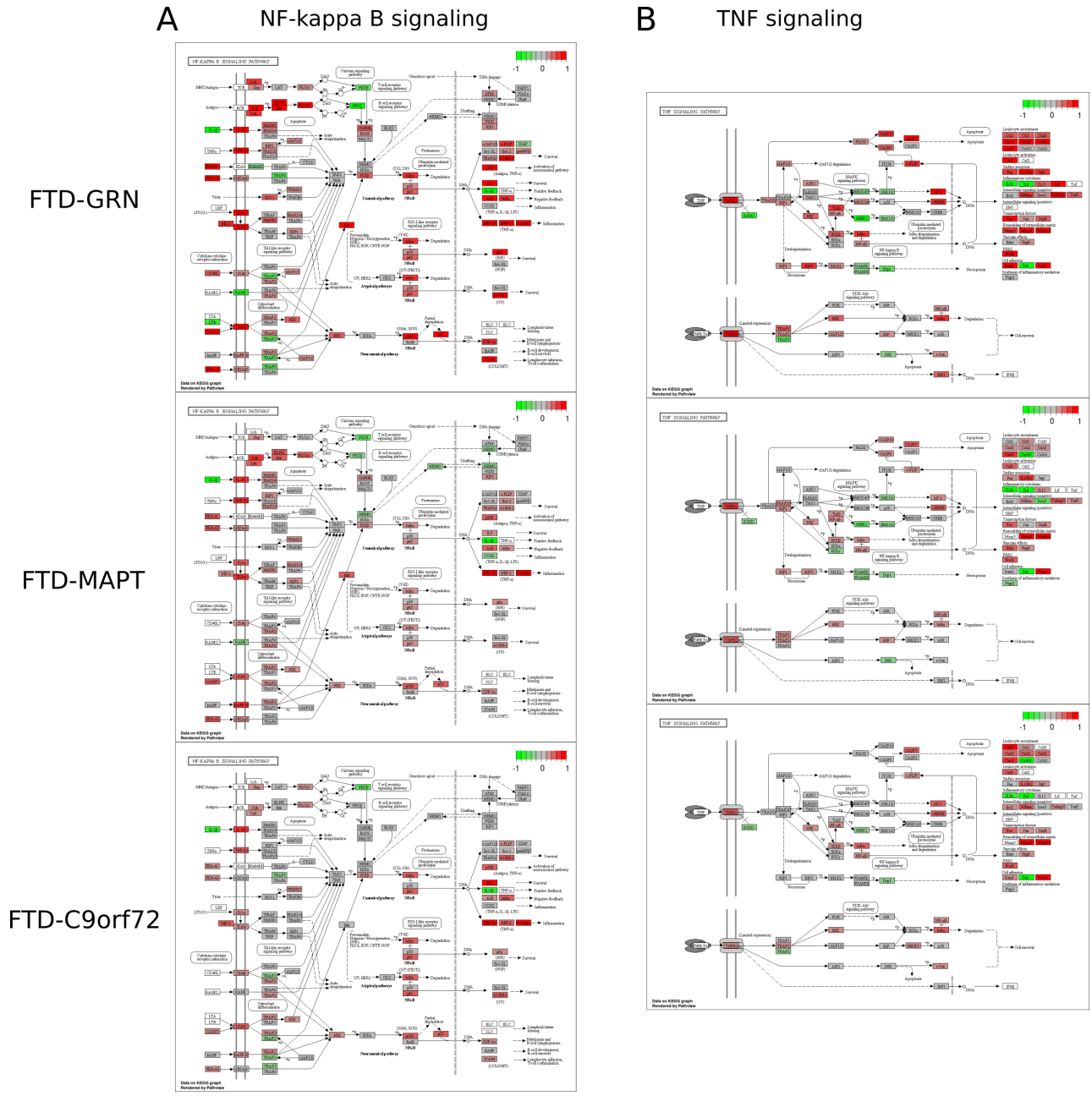

Figure 3.13: Fold-changes of genes in immune system pathways. **A** Fold-changes of genes of the NF-kappa B signaling pathways. **B** Fold-changes of genes of the TNF-signaling pathway.

Module	Top terms (max 5)	Genes	Terms
M1	mitochondrion organization, NADH dehydrogenase complex assembly, mitochondrial respiratory chain complex I assembly, mitochondrial respiratory chain complex assembly, generation of precursor metabolites and energy	538	201
M2	monovalent inorganic cation transport, potassium ion transmembrane transport, regulation of cation transmembrane transport, cellular potassium ion transport, potassium ion transport	577	85
M3	protein polyubiquitination, dephosphorylation, protein modification by small protein removal, peptidyl-serine modification, regulation of proteasomal protein catabolic process	791	151
M4	respiratory chain complex IV assembly, cytochrome complex assembly, cellular respiration, energy derivation by oxidation of organic compounds, mitochondrion organization	28	15
M5	epithelial cilium movement involved in determination of left/right asymmetry, epithelial cilium movement, striated muscle contraction, motile cilium assembly, inner dynein arm assembly	287	8
M6	epithelial cell migration, tissue migration, epithelium migration, ameboidal-type cell migration, epithelium development	3	5
M7	mitochondrion organization	21	1

Table 3.3: HumanBase FTD-GRN down-regulated modules and associated biological processes.

Module	Top terms (max 5)	Genes	Terms
M1	NADH dehydrogenase complex assembly, mitochondrial respiratory chain complex I assembly, mitochondrial respiratory chain complex assembly, cellular respiration, mitochondrion organization	49	49
M2	protein stabilization, regulation of protein stability	2	2
M3	regulation of RNA splicing, nucleocytoplasmic transport, nuclear transport, RNA splicing	3	4
M4	positive regulation of heterotypic cell-cell adhesion, negative regulation of lipid catabolic process, negative regulation of catabolic process, regulation of heterotypic cell-cell adhesion, regulation of lipid catabolic process	53	50
M5	carboxylic acid transport, organic acid transport, microtubule organizing center organization, establishment of organelle localization, organic anion transport	13	10

Table 3.4: HumanBase FTD-MAPT down-regulated modules and associated biological processes.

Module	Top terms (max 5)	Genes	Terms
M1	neutrophil migration, neutrophil chemotaxis, granulocyte migration, granulocyte chemotaxis, leukocyte migration	293	135
M2	response to hormone, response to prostaglandin D, cellular response to prostaglandin D stimulus, endocytosis, canonical Wnt signaling pathway	250	334
M3	symbiont process, viral process, regulation of symbiosis, encompassing mutualism through parasitism, positive regulation of I-kappaB kinase/NF-kappaB signaling, regulation of viral process	308	360
M4	endothelial cell differentiation, I-kappaB kinase/NF-kappaB signaling, endothelium development, cellular response to tumor necrosis factor, regulation of endothelial cell differentiation	276	170
M5	endocrine system development, adrenal gland development, urogenital system development, renal system development, epithelium development	197	246
M6	T cell activation via T cell receptor contact with antigen bound to MHC molecule on antigen presenting cell, T cell activation involved in immune response, leukocyte cell-cell adhesion, lymphocyte activation involved in immune response, lymphocyte activation	19	30
M7	bicellular tight junction assembly, apical junction assembly, cell-cell junction assembly, cell junction assembly, cell-cell junction organization	8	11

Table 3.5: HumanBase FTD-GRN up-regulated modules and associated biological processes.

Module	Top terms (max 5)	Genes	Terms
M1	pantothenate metabolic process, Fc receptor signaling pathway, vitamin metabolic process, cilium organization, cilium assembly	139	59
M2	positive regulation of cell migration, positive regulation of cellular component movement, positive regulation of cell motility, transmembrane receptor protein tyrosine kinase signaling pathway, positive regulation of protein kinase B signaling	52	124
M3	epithelial cell differentiation, endothelial cell differentiation, endothelium development, epithelium development, regulation of T cell receptor signaling pathway	48	77
M4	positive regulation of ERK1 and ERK2 cascade, collagen metabolic process, regulation of ERK1 and ERK2 cascade, ERK1 and ERK2 cascade, formation of primary germ layer	24	68
M5	hormone metabolic process, regulation of hormone levels	9	2
M6	transmembrane receptor protein tyrosine kinase signaling pathway, regulation of cell morphogenesis, vascular endothelial growth factor receptor signaling pathway, regulation of blood pressure, cellular response to vascular endothelial growth factor stimulus	52	50
M7	response to insulin, odontogenesis of dentin-containing tooth, smoothened signaling pathway, response to glucocorticoid, response to corticosteroid	61	43
M8	endocytosis, import into cell, response to organonitrogen compound, response to lipid	7	4

Table 3.6: HumanBase FTD-MAPT up-regulated modules and associated biological processes.

Chapter 4

Regional transcriptional patterns in FTD

Disclaimer

This contents of this chapter are not published, a manuscript is under preparation. The idea for this chapter came from Peter Heutink and me. All analyses in this chapter have been done by me.

4.1 Abstract

Most studies on FTD focus on the frontal and temporal lobes, which are the most affected brain regions in this disease. Studying other brain regions, even if seemingly not affected, can yield valuable additional information. Here, we have analysed CAGE-seq data from seven different brain regions of FTD patients with mutations in GRN, MAPT, or C9orf72 and healthy controls. Using our transcriptomic data, we identified the caudate and putamen as regions with high transcriptional aberrations, additional to frontal and temporal lobes and the cerebellum as being most affected in FTD-C9orf72. Using gene co-expression analysis, we have identified region- and cell type-specific modules that are correlated with FTD or subtypes of it. Among them, we identified a microglia-specific inflammatory module that is highly associated with FTD-GRN and likely drives the neuroinflammation that is specific for this FTD subtype. We furthermore identified an FTD-MAPT associated module that is involved in microtubule function and might therefore be unique to FTD-MAPT. Examination of blood vessel-associated modules revealed up-regulation of endothelial cell specific genes in almost every region and disease subtype and identified LAMB2 as hub gene with potentially important function for this module. To our knowledge, our studies depicts the first transcriptional study of seven different brain regions and three different disease subtypes in FTD. Our results therefore provide novel insights into regional vulnerability in FTD and furthermore highlights several potential disease mechanisms.

4.2 Introduction

As outlined in the previous sections, the precise molecular mechanisms underlying FTD is a matter of active research. Apart from mechanisms at the cell level, it remains unclear why the frontal and temporal lobes are especially vulnerable to FTD. Indeed, the selective vulnerability of specific brain regions and cell types in many neurodegenerative disorders is actively researched and not fully understood [63]. Moreover, because most studies in neurodegenerative disease research are restricted to post-mortem human brain tissue or animal models, we lack knowledge about disease development and progression over time, starting from earliest stages. Recent studies have shown that atrophy in FTD spreads via neuronal connections to other brain regions, starting from epicenters [34]. Given the connectivity of the brain, it can therefore be assumed that many brain regions in FTD are indeed affected, albeit not as strong as the frontal and temporal lobes. To advance our understanding of FTD, it is therefore necessary to expand research activities beyond frontal and temporal lobes, and determine why other brain regions are less or not at all affected by neurodegeneration.

Here, we started to address these questions with the help of Cap Analysis of Gene Expression sequencing (CAGE-seq) data from the RiMod-FTD project of seven different brain regions (frontal, temporal and occipital lobes, hippocampus, putamen, cerebellum, caudate). The data stems from post-mortem human brain tissue samples collected from FTD patients with mutations in MAPT, GRN and C9orf72 and healthy controls. CAGE-seq can be used to study transcription start sites (TSSs), enhancers and gene expression. As frontal and temporal lobes suffer from major neurodegeneration in end-stage FTD, we reasoned that other brain regions could be used as proxies to earlier disease stages and help to advance our understanding of regional expression differences in FTD subtypes. Furthermore, region-specific gene expression can be measured and examined for disease-relevance. Finally, due to the increased sample size and heterogeneity of this dataset, we reasoned that it can be ideally used to apply weighted gene correlation networks analysis (WGCNA) [108]. Using WGCNA we have generated region and disease specific transcriptional modules in FTD with evidence from multiple brain regions. We have investigated these modules for their cellular functions and specificity to different regions, cell types and disease subtypes. Using these methods, we have identified several genes, modules and pathways that are of particular interest in FTD and promising subjects for further experimental studies.

4.3 Results

4.3.1 Transcriptional dysregulation is not restricted to frontal and temporal lobes in FTD

We performed CAGE-seq on samples from seven different brain regions: frontal, temporal and occipital lobes, hippocampus, caudate, putamen and cerebellum. Because frontal and temporal lobes are most affected in FTD, data collection was focused on these regions, resulting in the highest sample numbers (58 and 64, respectively, Fig. 4.1A). For all other regions between 21 and 28 samples were processed. All data was processed and mapped to the human genome (Methods), with subsequent identification of CAGE-seq tag clusters. Analysis of the genomic origins of tag clusters revealed that over 60% of clusters are located in the promoter region of genes, 7.4% at the 5'-UTR, 10.1% in exons and the remaining peaks at introns, the 3'-UTR or intergenic regions (Fig. 4.1B). As the prime interest of our study was in gene expression, we translated CAGE-seq cluster counts to gene expression by assigning counts to the closest annotated gene with the maximal distance of 3kb (Methods). To get an initial overview of transcriptomic diversity across regions and groups, we performed principal component analysis (PCA) after normalization and variance stabilization with DESeq2 [119]. Using all samples and coloring by region it becomes apparent that the cerebellum is highly different compared to the other brain regions (Fig. 4.1C). When performing PCA separately for each region, controls seem to cluster together in most regions except the occipital lobe, indicating transcriptomic aberrations in FTD across many regions (Fig. 4.1D). The cerebellum regional PCA contains two clusters that cannot be resolved by disease groups.

We next performed differential expression analysis, comparing FTD disease groups against controls for every region separately. As a word of caution we want to highlight that, owing to diverging and partly limited sample numbers for different regions, a direct comparison of transcriptomic changes between regions has to be considered carefully. For FTD-MAPT and FTD-GRN we detected the highest numbers of differentially expressed genes (DEGs) in the frontal and temporal regions (Fig. 4.1E). Interestingly, the largest number of DEGs for FTD-C9orf72 was detected in the caudate, followed by the cerebellum. The temporal lobe contains more DEGs (3378) than the frontal lobe (2828) in FTD-MAPT, which is in accordance with current knowledge, as this region is typically stronger affected in FTD-MAPT. The highest number of DEGs overall are observed for FTD-GRN. In all groups, we observed high numbers of DEGs in the putamen, indicating vulnerability of this region in FTD. While the hippocampus contains the highest number of DEGs in FTD-MAPT, the caudate is stronger affected in FTD-GRN and FTD-C9orf72, suggesting potential selective regional vulnerability of these regions to Tau-pathology and TDP-43-pathology, respectively.

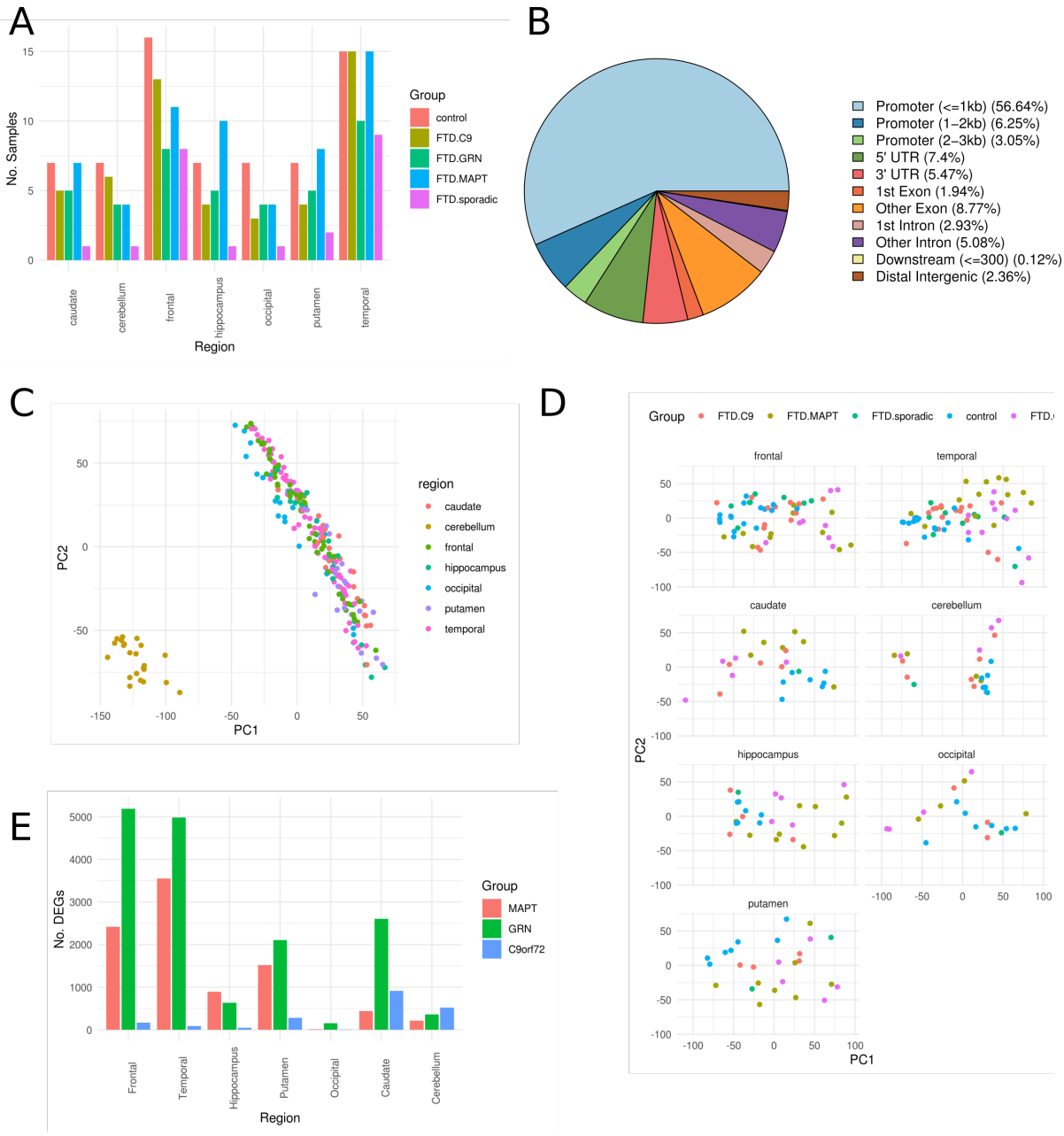

Figure 4.1: Principal component and differential gene expression analysis from CAGE-seq gene expression data. **A** Number of samples for each disease group and brain region. **B** Annotated features near CAGE-seq peaks. **C** PCA plot of all regions. **D** Region-wise PCA plots colored by disease group. **E** Number of differentially expressed genes in each disease group and region.

We then performed pathway enrichment analysis using go:Profiler [155] to identify cellular functions that are distinctly affected in specific regions, disease groups, or both.

Due to the large number of different comparisons, we can only highlight selected enrichment results here. For conciseness, we will furthermore only discuss gene ontology biological processes (GO:BP) here. In frontal and temporal lobes, we observe strong dysregulation of neural system related pathways in FTD-MAPT and FTD-GRN, as well as cellular transport, extracellular matrix and metabolic processes, while we only detected dysregulation of collagen metabolic processes in FTD-C9orf72 for these regions. The large numbers of DEGs and strong apparent dysregulation of neuronal system pathways is likely caused in part due to the prominent neuronal loss in frontal and temporal lobes. In the caudate, autophagy and metabolic processes are dysregulated in FTD-C9orf72 and neuronal system pathways in FTD-GRN, while we could not find a significant GO:BP term for FTD-MAPT. FTD-C9orf72 is the only group with enriched terms in the cerebellum (metabolic processes). Metabolic and autophagy processes are dysregulated in the hippocampus of FTD-GRN patients, and cellular component organization and cilium assembly in FTD-MAPT patients. We detected enrichment of localization for FTD-GRN as the only enriched term in the occipital lobe. Synaptic signaling pathways are dysregulated in the putamen of FTD-GRN patients, while cellular component organization pathways are dysregulated in FTD-MAPT patients.

Clearly, transcriptomic aberrations in FTD are not merely limited to the frontal and temporal lobes, which are mainly affected by neurodegeneration. Considerable numbers of DEGs could be detected in several regions and FTD subtypes. This suggests that, at least in the disease end-stage, changes in transcriptional regulation are abundant across the whole brain in FTD. Interestingly, especially autophagy pathways seem to be dysregulated in several brain regions. Autophagy is known to play important roles in many neurodegenerative diseases and also in FTD [49]. It is therefore tempting to speculate that dysfunctional autophagy could be an early event in disease progression, as observed here in many seemingly un- or weakly-affected brain regions.

4.3.2 Regional and disease specific co-expression modules in FTD

To better understand regional differences and to find key genes involved in determining FTD-subtype specific regional vulnerability, we performed weighted gene correlation network analysis (WGCNA) using all 239 CAGE-seq samples. After merging highly correlated modules (see Methods) we detected 23 co-expression modules across all samples and regions (Fig. 4.2A). We could find modules significantly associated with every region and disease phenotype (Fig. 4.2B). The strongest positive association was detected for the cerebellum-specific module greenyellow, which most likely represents the distinct transcriptional programs in this brain region. Several modules are significantly associated with FTD and FTD subtypes. For instance, the lightcyan module is associated with FTD-MAPT, the brown module with FTD-GRN and the brown4 module with FTD-GRN and FTD-C9orf72.

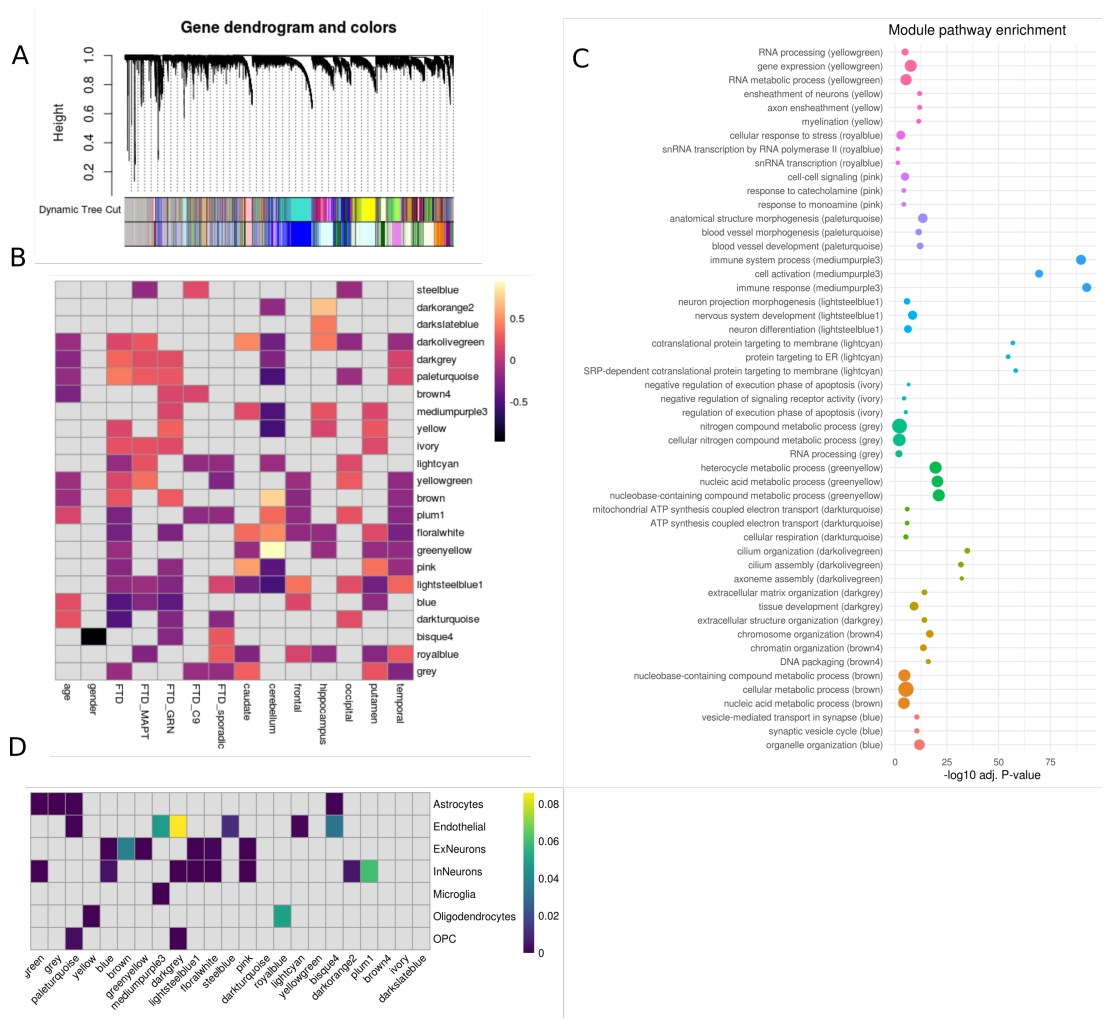

Figure 4.2: WGCNA of 239 CAGE-seq expression samples across different brain regions. **A** Gene-dendrogram generating by WGCNA including module colors before and after merging. **B** Module-trait association matrix. Non-significant associations are marked in grey, significant associations are colored from dark (negative association) to light (positive association). **C** Pathway enrichment of modules, with the negative logarithmized adj. P-value on the x-axis and the 3 most significantly enriched biological processes (BPs) for each module on the y-axis. Circles are sized according to number of BP member genes and colored according to different modules for easier visual inspection. **D** EWCE cell type enrichment analysis of modules.

The lightsteelblue1 (LS1) module is positively associated with frontal and temporal lobes and negatively with FTD, which suggests that it might reflect neurodegeneration in these regions. This was supported by pathway enrichment analysis, which revealed

neuronal system related biological processes (BPs) as highly enriched in this module (Fig. 4.2C). Other significant pathway enrichments include extracellular matrix organization and blood vessel development for darkgrey (DG) and paleturquoise (PT) modules, which are both positively associated with FTD and the temporal lobe. The FTD-GRN specific mediumpurple3 (MP3) module is highly enriched for immune system pathways, highlighting the strong immune system component in FTD-GRN. Intriguingly, the causal gene GRN is a member of this module, supporting the idea that the initial mutation in GRN is driving the observed dysregulation. The darkolivegreen (DOG) module, which is associated with FTD-MAPT, is highly enriched for cilium organization and assembly, which are microtubule-associated processes and might therefore indicate a connection to the causal MAPT mutation, as MAPT mutations can lead to destabilization of microtubules. Another FTD-MAPT associated module, lightcyan, is enriched for processes that are involved in protein transport. Other detected significant enrichments include cellular respiration (darkturquoise), metabolic processes (gray), cell-cell signaling (pink), cellular response to stress (royalblue), myelination (yellow) and RNA processing (yellowgreen). The causal genes MAPT and C9orf72 are members of the LS1 and greenyellow modules, respectively, which are both not exclusively associated with these disease subgroups. As mutations in MAPT and C9orf72 act via complex mechanisms, and not solely by aberrant expression levels, it is reasonable to assume that co-expression analysis will not necessarily associate these genes with the disease-driving modules. FTD-causing GRN mutations, on the other hand, directly act on expression levels, making the FTD-GRN associated MP3 module an interesting target for studying causal disease mechanisms in FTD-GRN.

Co-expression modules detected by WGCNA can be cell type-specific, as cell composition changes across samples introduce correlation trends among cell type-specific genes. WGCNA modules can therefore not only be interpreted as functionally connected gene networks, but cell type-specific gene modules. We performed expression weighted celltype enrichment (EWCE) [179] analysis to identify enrichment of cell type-specific genes in modules (Fig. 4.2D). As expected, most modules show significant enrichment for a certain cell type. The LS1 module is significantly enriched for neuronal specific genes, which supports the hypothesis that its negative correlation with FTD reflects neurodegeneration. The FTD-GRN specific MP3 module is exclusively enriched for microglial genes, and to a lesser extend endothelial genes, suggesting that the strong immune component in FTD-GRN is driven by microglia, which are also the cell type with the strongest expression of GRN. The blood vessel development-associated PT module is enriched for endothelial cells, astrocytes and oligodendrocyte precursor cells (OPCs). The lightcyan module, which is mainly specific to FTD-MAPT, is exclusively enriched for endothelial-specific genes.

4.3.3 Regional module activity is partly subtype specific

We next wanted to examine how the activity of modules changes across different brain regions and disease subtypes compared to healthy individuals. While frontal and temporal lobes are mainly affected in FTD, we hypothesized that other regions could be affected by neurodegeneration to a lesser extent, possibly serving as approximations to an earlier state of the disease. This idea was built mainly on recent studies, which have shown that neurodegenerative diseases spread from epicenters to affect other areas of the brain (3). To compare module activity across brain regions, we calculated the average module-wise gene expression and used the average across samples for visual inspection (Fig. 4.3A). While average expression profiles cannot capture the complicated intricacies of intra-modular transcriptomes, they can be easily interpreted and allow for direct comparison between groups.

For many modules, substantial differences of FTD groups to the baseline (controls) are observable, indicating a systematic change in module activity. For instance, the LS1 module, which we have identified to be associated with neurons and possibly neurodegeneration, has notably smaller activity scores in frontal and temporal lobes for all disease groups, and to a lesser extent in other brain regions. This indicates that neuronal pathways are affected across the whole brain in FTD patients, albeit to a substantially diverging degree. Expression trends of the microglia and FTD-GRN associated MP3 module show strongly increased activity compared to the baseline in temporal and frontal lobes, slightly increased activity in caudate and occipital lobes, but no increase in other regions. Interestingly, FTD-MAPT has the highest activity scores for this module in the hippocampus, and FTD-C9orf72 in caudate and putamen, indicating altered microglial activity for these FTD subgroups as well, albeit not in the frontal and temporal lobes. We also observed an increased activity score of the PT module for almost all regions and FTD subtypes. This module is enriched for genes involved in blood vessel development and positively associated with FTD, especially with the temporal lobe, pointing towards an increased activity of angiogenesis-related pathways, similar to what has been recently observed in an FTD mouse model [25].

For most modules, there is strong correlation between disease groups across different regions (Fig. 4.3B). This shows that regional variability is, for many modules, the biggest driver of transcriptional differences. However there are also several modules with negative or weak correlation between groups, which indicates disease group-specific transcriptional trends that are larger than regional differences. One example is the ivory module, whose activity increases greatly in caudate, putamen and occipital lobe for FTD-MAPT and FTD-GRN compared to both FTD-C9orf72 and controls (Fig. 4.3B). The ivory module contains genes involved in apoptosis regulation. Correlation between module activities reveals some blocks of highly correlated modules (Fig. 4.3C). For instance, the modules MP3, PT, darkgrey, DOG and darkslateblue are positively corre-

lated. Another block of high positive correlation is made up by the floralwhite, blue and darkturquoise modules. Interestingly, the MP3, PT and yellow modules are strongly negatively correlated with the mentioned modules. The strong positive and negative between-module correlation can be explained by either overlap in transcriptional regulation, or underlying, latent mechanisms that lead to similar transcriptional patterns. It is quite likely that abundant neurodegeneration will affect several modules in similar fashion, for example. Similarly, negative correlation between modules might be explained reciprocal relationships between cellular pathways that are represented by modules.

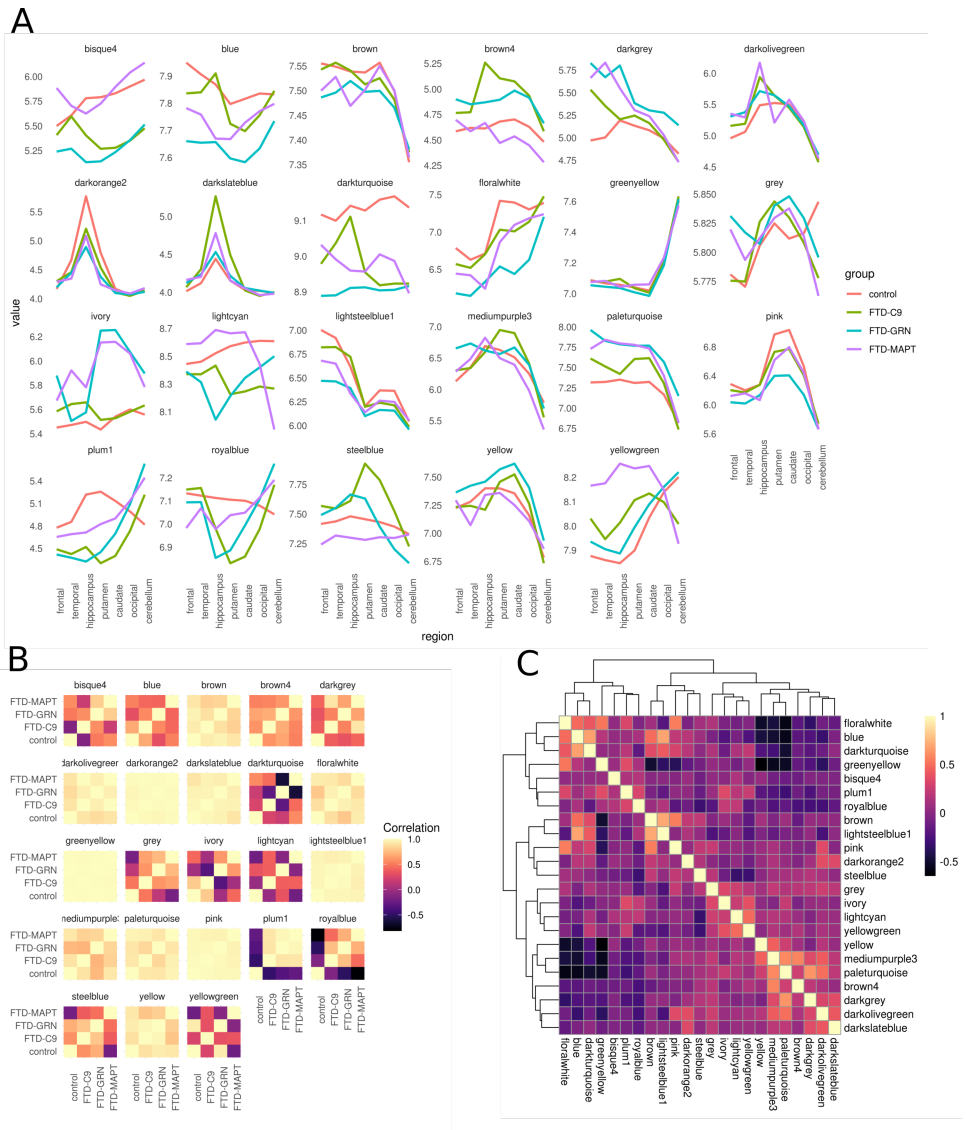

Figure 4.3: Module activity trends across regions measured by average module expression. **A** Raw module activity scores, colored and divided by groups and smoothed across samples. **B** Per-module Pearsons correlation disease group average module activity across regions. **C** Heatmap of Pearson correlation between module activity scores.

4.3.4 TNF up-regulation is specific to frontal and temporal lobes

The MP3 module is strongly and specifically associated with FTD-GRN, enriched for microglial genes and has the causal gene GRN as member. These aspects highlight the potential of MP3 to harbour genes and pathways of high relevance to FTD-GRN disease pathogenesis. Therefore, we determined which genes are driving the high activity in frontal and temporal lobes and to what extend MP3 activity is increased in other brain

regions. Module activity of MP3 is highly increased in the frontal and temporal lobes of FTD-GRN patients and only slightly increased in the caudate and the occipital lobe (Fig. 4.4A). For the other brain regions, no increase in module activity is detected. Next, we evaluated module membership and trait significance measurements of genes to detect potential module hub genes. Candidate genes with high values in both of these metrics include IL13RA1 and IL10RB, which encode interleukin receptors important for immune system functions, as well as RIPK3, a central protein for the necroptosis cell death pathway (Fig. 4.4B). The high potential of RIPK3 for being a hub gene indicates its importance to the module. The necroptosis pathway might therefore be either a driver, or a highly correlated downstream effect of increased module activity.

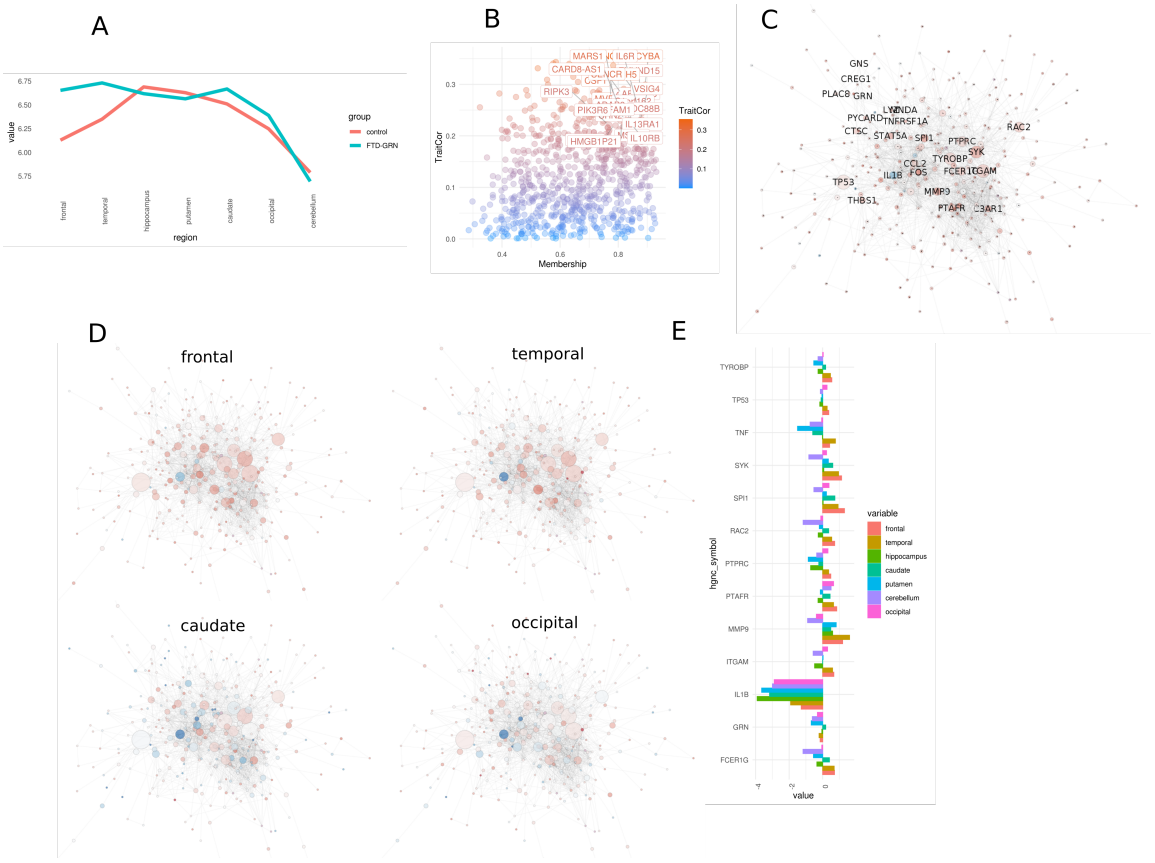

Figure 4.4: Driver genes in the MP3 module. **A** Module activity of MP3 in FTD-GRN and controls for all seven brain regions. **B** Module membership and correlation with the FTD-GRN trait for genes in MP3. **C** PPI network of MP3. Node sizes are dependent on betweeness centrality values. Labels of the genes with the highest BC values are shown, as well as GRN and its direct neighbors. **D** PPI network of MP3. Node colors correspond to log fold-changes in the depicted regions. **E** Log fold-changes values for selected genes of MP3 in FTD-GRN. Colors depict different brain regions.

We generated a protein-protein interaction (PPI) network using STRING-DB (Methods) to identify the most central genes in MP3 from a network perspective. Using the betweeness centrality (BC) as metric reveals several highly central genes to MP3 (Fig. 4.4C). The gene with the highest BC value is Tumor necrosis factor (TNF), which encodes a cytokine with multiple important roles for the immune system. This importance is captured by the high BC value. Other genes with high BC values include Matrix metalloproteinase 9 (MMP9), Interleukin 1 beta (IL1B), TYRO Protein Tyrosine Kinase Binding Protein (TYROBP), among others. Interestingly, while being central to MP3 function, IL1B has negative fold-changes in all brain regions for FTD-GRN, and it is the most strongly down-regulated gene in MP3 in the frontal and temporal lobes (Fig. 4.4 D E). IL1B is a pro-inflammatory cytokine, of which elevated levels have been reported

in other neurodegenerative diseases such as Alzheimer's disease (AD) [77, 174]. It is therefore surprising to see decreased levels of IL1B in all brain regions of FTD-GRN patients, especially as FTD-GRN is characterized by increased neuroinflammation (see also Chapter 3) and all other central genes of the MP3 module are up-regulated. The decrease of IL1B is particularly strong in regions other than frontal and temporal lobes, with log fold-changes as low as -4 (hippocampus). In frontal and temporal lobes, the decrease is less prominent.

The largest up-regulation of MP3 hub genes is observed for frontal and temporal lobes, followed by the caudate region for several genes (Fig. 4.4E), indicating increased inflammation in this region, as well. This fits with the finding of strong transcriptomic aberrations in this region in FTD-GRN, as we have observed earlier (Fig. 4.1E). Further confirmation for the hypothesis of vulnerability of the caudate in FTD comes studies that found atrophy in the caudate and putamen in FTD, with especially high severity in FTLD-TDP [72].
TNF, however, is only up-regulated in frontal and temporal lobes, and has close to zero or negative LFCs for all other brain regions (Fig. 4.4E). Given the central role of TNF in inflammation and cell death, the increased production of TNF in frontal and temporal lobes might be what causes the strong neurodegeneration in these regions. Other regions, while also affected by atrophy and partly inflammation, might be protected due to lower TNF levels. TNF, as a central regulator of neuroinflammation in frontal and temporal lobes in FTD-GRN, therefore depicts an interesting target for further investigation and potentially for therapeutic interventions.

4.3.5 DNALI1 is a hub gene of cilium-assembly module DOG in FTD-MAPT

The DOG module is significantly associated with FTD-MAPT and the hippocampus, and it is enriched for genes involved in cilium assembly, which is directly linked to microtubules and therefore MAPT. For these reasons, we decided to more closely examine this module. We used the module membership (MM) and trait correlation (TC, with FTD-MAPT) metrics computed by WGCNA to identify potential driver genes of the DOG module. We detected several candidate genes that have high MM values and show significant correlation with the FTD-MAPT trait (Fig. 4.5B). We additionally generated a PPI network using the STRING database [187] (Fig. 4.5C). We removed nodes not directly connected to the core network and calculated the BC values of all remaining nodes. Dynein Axonemal Light Intermediate Chain 1 (DNALI1) has the largest BC value and is among the genes with the highest MM and TC metrics in this module. We therefore considered this gene to be the best candidate for a module driver- or hub-gene.

As the DOG module is a potentially important module for FTD-MAPT pathogenesis,

we wanted to assess how module gene expression behaves in regions other than frontal and temporal lobes. We therefore calculated the log-fold change (LFC) of genes in FTD-MAPT compared to controls for every region and colored PPI nodes according to LFC (Fig. 4.5D). The cerebellum network, an apparently unaffected region, shows even distribution of positive and negative LFC values, and hence no up- or down-regulation. In all other regions, DNALI1 has a positive LFC, with the highest values in the hippocampus, putamen, frontal and temporal lobes. While there is no strong up-regulation observable in putamen, caudate and occipital lobe, there are more genes with positive LFCs compared to negative LFCs, indicating slight up-regulation of module genes. In hippocampus, temporal and frontal lobes, most genes show highly positive LFC values, as the DOG module is highly active compared to controls in these regions.

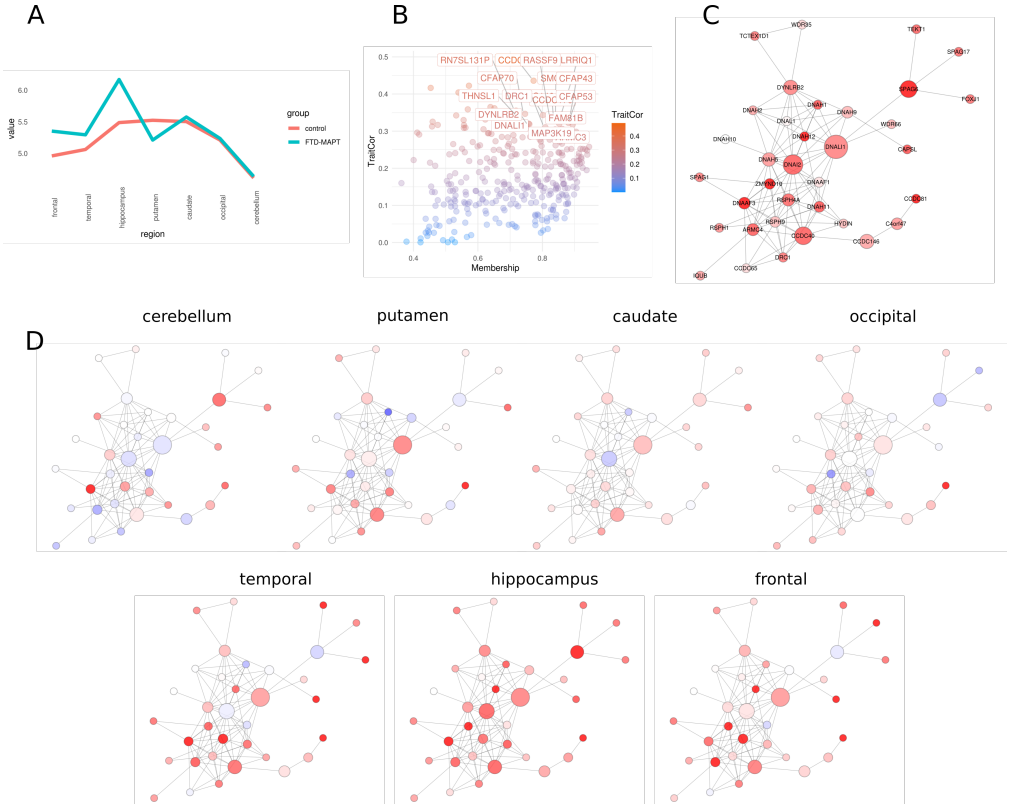

Figure 4.5: **A** Averaged activity values of DOG across brain regions. **B** Correlation with FTD-MAPT trait(y-axis) plotted against module membership for genes of the DOG module. **C** PPI network colored according to LFC of DOG module. Node size corresponds to betweenness centrality. **D** PPI networks similar to D, for all different brain regions.

DNALI1 was found to be up-regulated in post-mortem brain tissue from patients with frontotemporal lobar degeneration with ubiquitinated inclusions (FTLD-U) [128]. Another recent study investigated gene expression in the middle temporal gyrus in Alzheimer's

disease (AD) [149]. The authors found DNALI1 to be one of 13 genes that are significantly associated with neurofibrillary tangle (NFT) density. Moreover, DNALI1 was among the top four genes significantly associated with Braak stage. These studies provide additional evidence for a potential role of DNALI1 as driver gene in tau pathology. Further experiments will be necessary to precisely allocate the role of DNALI1 in tau pathology, and whether DNALI1 should be investigated as a potential drug target.

4.3.6 Blood vessel-associated gene expression is increased in multiple regions in FTD brains

Next, we dissected the paleturquoise (PT) module, which encompasses 1229 genes and is enriched for genes involved in blood vessel development (Fig. 4.6C). Genes specific to astrocytes and endothelial cells are significantly enriched in this module. Interestingly, average gene expression of the PT module is increased compared to controls in almost all regions and FTD subgroups (Fig. 4.6A). Solely in the cerebellum and occipital lobe of FTD-C9orf72 samples we did not detect an increase in expression. In FTD-MAPT the strongest increase in module activity was observed in putamen and temporal lobe, followed by hippocampus, frontal lobe and caudate (Fig. 4.6B). The frontal lobe is the region with the strongest increase in PT module activity in FTD-GRN, followed by the putamen, occipital and temporal lobes. We observed less strong increases in PT activation in FTD-C9orf72, where putamen, caudate, frontal and temporal lobes are the regions with strongest increase in activity. Notably, the increase in expression of PT module genes is not restricted to regions with excessive neurodegeneration, as is evident from the strong increase in the caudate (FTD-MAPT) or putamen (FTD-C9orf72), for example. Activity of the LS1 module is negatively correlated with activity of the PT module (FTD-MAPT: r=-0.47, FTD-GRN: r=-0.76, FTD-C9orf72: r=-0.66), albeit only significantly in FTD-GRN. It is therefore tempting to speculate that increased blood vessel development is an early feature of FTD, which might precede neurodegeneration and brain atrophy, as it is observed in regions relatively unaffected by these pathological hallmarks.

To further characterize the PT module, we generated a PPI network using STRING and evaluated module genes according to module membership and correlation with FTD. It becomes evident from the PPI network that the gene epidermal growth factor receptor (EGFR) is the most central node of the network according to the betweenness centrality measure (Fig. 4.6C). The EGFR gene has several functions, most importantly it is associated with inducing cellular proliferation, making it a prominent drug target in several cancers [218]. Gene module membership of the PT module is strongly correlated with FTD trait association (Fig. 4.6D). According to these metrics, EGFR is not among the top hub genes, indicating it might not function as a module driver gene in FTD, but is simply very central to module function. Laminin subunit beta 2 (LAMB2) is the gene

with the highest MM and trait correlation metrics. Average expression values of LAMB2 are increased in every brain region and every FTD subtype compared to levels in controls (Fig. 4.6E), with statistically significant increases in the majority of comparisons. Typically, LAMB2 is expressed at very basal levels in the brain compared to almost all other tissues of the human body, according to expression values from GTEx [1] (Fig. 4.6F). Elevated LAMB2 expression levels might therefore indeed be a pathological feature of FTD disease, that is not restrained to the majorly affected frontal and temporal lobes.

Vascular dysfunction is usually not considered as a prominent feature of FTD pathology. However, a recent study has identified blood vessel abnormalities in P301L transgenic mice, which is a causal FTD mutation [25]. More precisely, the authors found an increased number of blood vessels in brains of P301L transgenic mice, albeit with smaller diameter and often with obstructed capillary flow. A different study tried to identify the causes of regional vulnerability in FTD, and correlated regional gene expression from the Allen Brain Atlas with brain atrophy in brains from patients with FTD-MAPT, FTD-GRN and FTD-C9orf72 [6]. Interestingly, the authors found that genes expressed in endothelial cells, astrocytes and involved in circulatory system development are positively correlated with atrophy severity. It was previously shown that several brain regions are hypoperfused in FTD [81, 53]. It is therefore tempting to speculate that vasculature changes might be directly involved in FTD pathogenesis, opening up a new avenue for potential remedies.

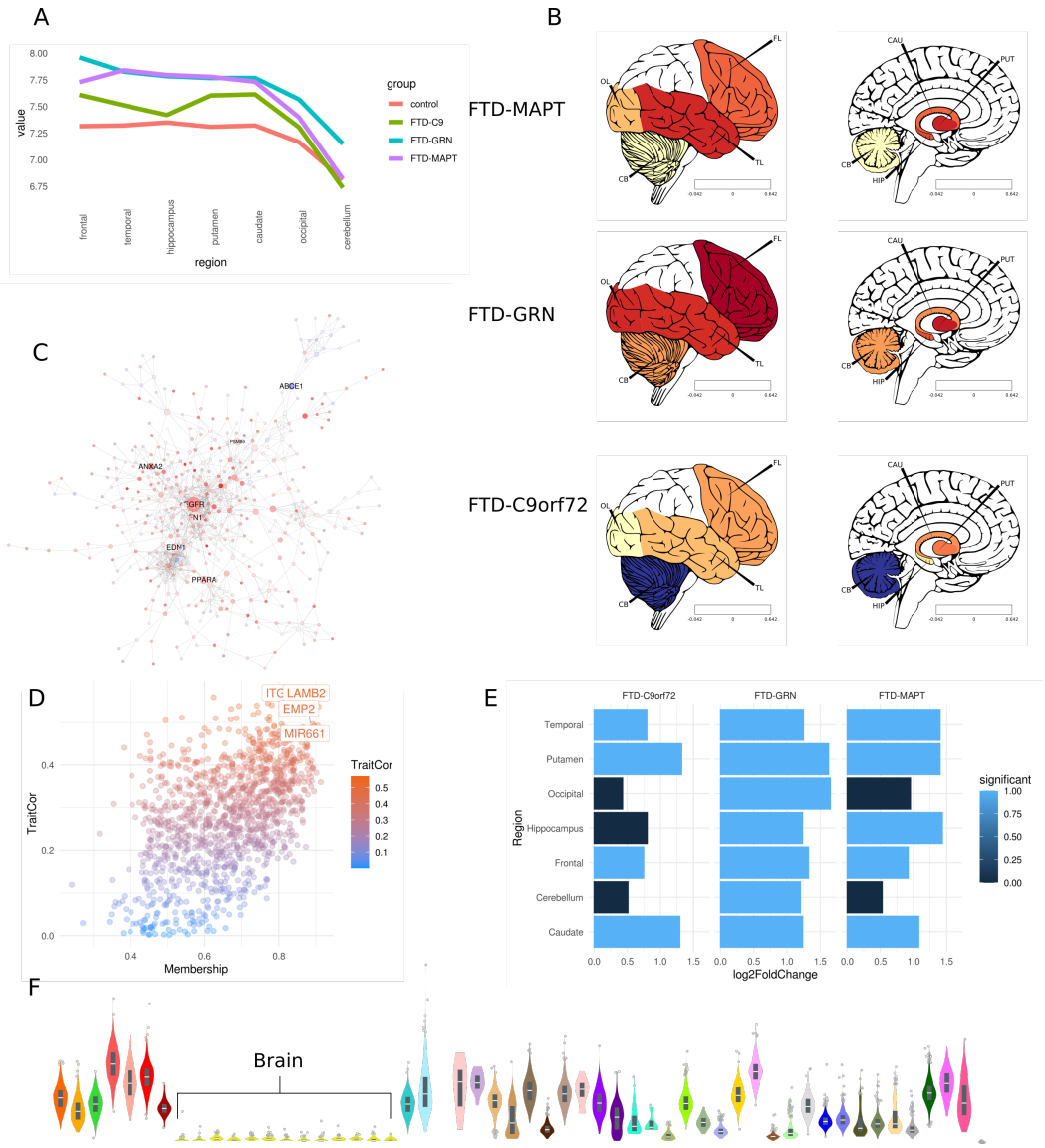

Figure 4.6: Expression of blood vessel-associated module PT in different brain regions. **A** Module activity changes over baseline in different brain regions for FTD subgroups. **B** Protein interaction network of PT module, coloured by LFC in FTD-MAPT. Node size corresponds to betweenness centrality. **C** Line Plot of module activity in different brain regions. **D** Module membership (x-axis) plotted against correlation with FTD (y-axis) for the paleturquoise module. **E** Log2 fold changes of LAMB2 in different brain regions and disease groups compared to control. Light blue indicates significant deregulation (adj. P-value < 0.05) **F** LAMB2 expression (TPM) across multiple tissues (GTEx). Violin plots corresponding to brain tissue are marked.

4.4 Discussion

Here, we have analyzed multi-regional human post-mortem brain data from FTD patients with causal mutations in GRN, MAPT and C9orf72. To our knowledge, this is the first transcriptomic experiment covering as many regions and different FTD subtypes. The experiment therefore provides valuable new information about transcriptomic states in regions other than the frontal and temporal lobes, which are commonly of interest due to their prime roles in FTD. DE analysis has confirmed that the largest transcriptomic aberrations can be observed in frontal and temporal lobes for FTD-GRN and FTD-MAPT. However, in FTD-C9orf72 the caudate appears to be stronger affected on the transcriptomic level, where we also observed a large number of DEGs for FTD-GRN. Furthermore, the putamen appears to be highly affected in FTD-GRN and FTD-MAPT. These two regions are anatomically close to each other and to the frontal and temporal lobes. It is therefore possible that FTD pathology primarily spreads to these regions once frontal and temporal lobes are affected strong enough. The hippocampus, which is also in close proximity to caudate and putamen, shows signs of transcriptomic aberrations as well. In contrary, the occipital lobe and the cerebellum, which are located on opposite sides of the brain, show almost no signs of transcriptomic aberrations. Our data therefore provides evidence that FTD pathology spreads to regions close to frontal and temporal lobes and that caudate and putamen are affected by the disease in end-stage FTD.

Using WGCNA, we have furthermore identified disease- and region-specific co-expression modules. Strong association of the immune system/microglia module MP3 with FTD-GRN confirms the increased neuroinflammation in this FTD subtype. Expression of the necroptosis mediator gene RIPK3 is highly correlated with the MP3 module, which suggests the necroptosis as a primary or at least important part of apoptosis in FTD-GRN. Furthermore, while many regions show signs of increased inflammation, the central inflammatory gene TNF is only up-regulated in frontal and temporal lobes. It is therefore tempting to speculate that TNF up-regulation might be the primary driver of neuroinflammation in frontal and temporal lobes, suggesting this pathways as a promising target for potential treatments. Unexpectedly, we observed down-regulation of the pro-inflammatory cytokine IL1B in all brain regions, which is generally known to be up-regulated in other forms of dementia. It would therefore be interesting to follow up on IL1B expression and function in FTD.

We have furthermore identified the DOG module as potentially specific to FTD-MAPT and tauopathy. The module is involved in cilium assembly, thereby linking it to microtubules and MAPT. DNALI1, the hub gene we have identified for this module, has been significantly associated with NFT density and Braak stage in AD. This link to tau pathology provides further evidence for the importance of DNALI1 and the DOG module in FTD-MAPT. Better characterization of the DOG module and its molecular functions as well as response to tau pathology are promising avenues to pursue to better understand

how FTD-MAPT progresses and to find fruitful drug targets.

Finally, we have identified modules involved in blood vessel development that are highly up-regulated in almost all brain regions in FTD. These findings confirm results from Chapter 3, where we have also identified evidence of increased blood vessel growth. Using network analysis, we have identified EGFR and LAMB2 as central genes for the PT module. Especially LAMB2, which is very lowly expressed in the healthy brain, depicts a novel candidate gene that might drive FTD pathogenesis. Given evidence from recent and older studies, we think the involvement of the vasculature system in FTD pathogenesis should be explored more deeply. As of yet, there exist very little studies that try to determine the precise roles of the vasculature in FTD. While it is possible that the activation of blood vessel-associated genes as observed here is a mere side effect in FTD pathogenesis, it is important to experimentally verify this.

The biggest limitation of this study is the low number of samples for FTD subgroups in regions other than the frontal and temporal lobes. Given the high heterogeneity of the data, the results therefore have to be considered with caution. Nevertheless, our study provides an important starting point for further experiments that examine interesting brain regions such as the caudate and putamen in more depth. Moreover, we have identified several interesting co-expression modules and genes that might be of interest for FTD pathogenesis, and thus provide promising starting points for future studies.

4.5 Methods

4.5.1 CAGE-sequencing and data processing

CAGE-seq libraries were prepared from 5 micrograms of RNA from frozen brain tissues according to a published protocol [188]. Libraries were sequenced on a HiSeq 2000 and/or HiSeq2500 on a 1x50 bp single read flow cell (Illumina) at an average of 20M reads/sample.

Sequencing adapters and barcodes in CAGE-seq FastQ files were trimmed using Skewer (v.0.1.126) [87]. Sequencing artefacts were removed using TagDust (v1.0) [109]. Processed reads were then aligned against the human genome hg38 using STAR (v.2.4.1) [52]. CAGE detected TSS (CTSS) files were created using CAGEr (v1.10.0) [70]. With CAGEr, we removed the first G nucleotide if it was a mismatch. CTSS were clustered using the 'distclu' method with a maximum distance of 20 bp. Exact commands used can be obtained from https://github.com/dznetubingen/cageseq-pipeline-mf.

CTSS count tables were then transformed into gene-wise count tables using ChIPseeker (v.1.20.0) [223]. Peaks were first transformed into BED file format, loaded with ChIPseeker

as peak file and transformed to a GenomicRanges object. Next, the *annotatePeak()* function was used to annotate the peaks, using up- and downstream TSS ranges of 3000 basepairs (bp). The annotated peaks were then transformed to a count table. Thus, every peak within a distance of 3000 bp up- or downstream of the promoter region of a gene, was assigned to this gene. The closest gene was used in case of collision. The resulting gene $\times sample count table was used for all downstream analyses$.

4.5.2 Differential gene expression and enrichment analysis.

Differential expression (DE) analysis was performed using DESeq2 [119]. Based on initial PCA analysis, we removed 5 samples that appeared as outliers. DE analysis was performed for each region separately, comparing the disease groups to the control group while controlling for gender.

Pathway enrichment analysis was performed using the R package *gprofiler2* (v0.1.9) [155]. For every DE analysis comparison, we considered genes with Benjamini-Hochberg-adjusted P-value below 0.05 as significant and tested them for enrichment using the *gostres()* function with domain scope 'annotated'. Terms and pathways with a P-valule below 0.05 were considered as significantly enriched. We specifically considered terms from the databases GO:BP (Biological Process), KEGG [89] and Reactome [86].

4.5.3 WGCNA analysis

We performed WGCNA analysis using he R package of version v1.69 [108]. As expression data, the count values were used after transformation with the DESeq2 function *varianceStabilizingTransformation()*. This transformation performs normalization of library sizes and yields approximately homoskedastic data. Next we calculated the gene-wise variance and removed genes with a variance below 0.1, as we considered them uninformative. This step removed 874 genes from the analysis. After visual inspection, we chose a soft threshold power of 10 to calculate the adjacency matrix. After calculation of the topological overlap matrix and hirerarchical clustering, dynamic tree cutting was performed using a minimum cluster size of 30. Then, close modules were merged using a module eigengene dissimilarity threshold of 0.3, which resulted in the final set of WGCNA modules. Module trait association heatmaps were created using the *pheatmap* package. Enrichment analysis was performed as described above using the *gprofiler2* package.

Visualization of module expression in brain regions

We used the cerebroViz R package (v1.0) [16] for visualization of module expression in different brain regions. The average expression value was used to show the general trend

of module expression. Correlation of modules and activity of modules in heatmaps was visualized with help of the pheatmap R package (v1.0.12) [101].

4.5.4 Protein interaction network analysis

PPI networks for modules were created using the STRING database [187]. We retained only interactions with high score (combined score > 0.7) and removed genes disconnected from the main network component. Network plots were created and betweeness centrality calculated using Cytoscape [175].

Chapter 5

Conclusions

The study of neurodegenerative diseases is inherently difficult due to their progressive development and the brain as affected tissue. To develop treatments for neurodegenerative diseases, we have to develop a precise understanding of the molecular mechanisms that lead to disease progression and outbreak over long periods of time and that affect only specific brain regions or neurons. A major hindrance toward this goal is that we cannot take samples from the brain, such as is possible from cancer tissue, for instance. It has therefore been extremely difficult to distinguish causes, consequences and bystander effects. Nevertheless, we have made great progresses in our understanding of increasingly prevalent diseases such as AD and FTD. To a large extent, these progresses can be attributed to new genomics technologies that enabled researchers to study genomic and transcriptomic variations and aberrations at high throughput and to acceptable prices, and to the development of sophisticated computational algorithms that extract knowledge from the generated data. In this thesis, I contributed to the genomics field and the field of neurodegenerative disease research by developing a novel algorithm for cell type deconvolution and by analysing FTD-related genomics data to create new hypotheses for disease mechanisms.

In chapter 2, I have presented a novel algorithm for cell type deconvolution that uses deep learning and scRNA-seq data to estimate cell type compositions from tissue expression profiles. While other algorithms for this task existed before, we have developed a completely different approach to cell type deconvolution, that performs better than conventional deconvolution algorithms in a variety of scenarios. While regression- and reference profile-based deconvolution algorithms have been developed almost to their fullest extent, with little improvements to be hoped for, we have developed an entirely different approach to the same problem. Our deep learning-based approach, while already performing equally or better compared to the best deconvolution algorithms, can likely be improved in further studies. For instance, as we have pointed out in the discussion of chapter 2, domain adaptation methods are likely to increase the performance of algorithms similar to Scaden. With the rapid progression of the deep learning field it is furthermore to be expected that new techniques for these algorithms will be developed that could be leveraged for future deconvolution algorithms. I therefore believe that our

contribution to the deconvolution field with Scaden is not only the algorithm itself, but the general idea of it, which I hope will spark new developments in this field that could lead to even better performance.

In chapter 3 and 4, I have extensively analysed multi-omics data from the frontal lobe and other brain regions of patients with FTD caused by mutations in GRN, MAPT and C9orf72. Apart from the concrete analysis results, we have presented the data from phase 1 of the RiMod-FTD resource, which aims to build a multi-omics resource of data from post-mortem human brain tissue, longitudinal mouse models and iPSC-derived cellular models. Disease-specific data resources can be extremely valuable assets for the research community, as has been proven by the ROSMAP project [3], for instance. The RiMod-FTD resource depicts the first such resource for the field of FTD and will be extended with additional datasets in the future. We believe that RiMod-FTD will help many researchers to develop new hypotheses in FTD research and therefore greatly help to advance the field.

With our integrative analysis of the RiMod-FTD datasets, we could highlight several molecular pathways and mechanisms of importance to disease subtypes and to FTD in general. For instance, using the deconvolution algorithm developed in chapter 2 and the RNA-seq dataset from RiMod-FTD, we could show that excitatory neurons seem to be primarily affected in FTD. Using the results from the deconvolution analysis, we could furthermore show striking enrichment of endothelial cells in all FTD subtypes and strong increase of microglial cell fractions in FTD-GRN. While it is known that microglia play a more important role in FTD-GRN, we are the first to show evidence for increased cell fractions using only transcriptomic data. To this day, it is furthermore very unclear what role the vascular system plays in FTD. Additional to the cell type enrichment from the deconvolution analysis, we could show in chapters 3 and 4 that pathways involved in blood vessel development are up-regulated in the frontal lobe and also in most other brain regions. These results therefore strongly indicate changes of the vascular system in FTD, and it will be important for future studies to determine whether these are bystanders of the disease or whether they are involved in the pathogenic process.
Apart from these high-level effects in FTD, we could highlight several TFs and miRNAs which are likely to play important regulator roles in FTD. We have specifically determined several TFs that might drive the neuroinflammation in FTD-GRN and therefore represent potential drug targets. Moreover, we have identified several miRNAs that target cellular trafficking pathways which we have experimentally verified in cellular models. By inhibiting important cellular transport mechanisms, these miRNAs likely contribute to FTD pathogenesis and should therefore be further verified and examined.

In conclusion, with this thesis, I have contributed a new computational algorithm to the field of bioinformatics and several new hypotheses and insights to the field of FTD research. I believe that these results will help other scientists to test their own hypotheses

or to select new research paths to pursue in order to advance our understanding of FTD. The biology of neurodegenerative diseases is incredibly complex, and it will require numerous other works such as this to completely understand it. However, the increasingly fast pace in the field of biology, and the astonishing progress made in the last decades and years, makes me hopeful that effective treatments for FTD and other neurodegenerative disease are achievable in the near future.

Abbreviations

AD	Alzheimer's Disease
AGD	Argyrophylic grain disease
ALS	Amyotrophic lateral sclerosis
ATAC-seq	Assay for Transposase-Accessible Chromatin sequencing
CBD	Corticcobasal degeneration
CBS	Corticobasal syndrome
CCC	Concordance correlation coefficient
CPU	Central processing unit X
DE	Differentially expressed
DEG	Differentially expressed gene
DL	Deep learning
DMP	Differentially methylated position
DNA	Deoxyribonucleic acid
DNN	Deep Neural Network
EWCE	Expression weighted celltype enrichment X
FFNN	Feed forward neural network
FTD	Frontotemporal dementia
FTLD	Frontotemporal lobar degeneration
GEP	Gene expression profile
GGT	Globular glial tauophathy
GPU	Graphics processing unit
ML	Machine learning
MLP	Multilayer perceptron
MMP	Matrix metalloproteinase
MND	Motorneuron disease
NCL	Neuronal ceroid lipofuscinosis
OLS	Ordinary least squares
PBMC	Perhipheral blood mononuclear cells
PPA	Primary progressive aphasia
PSP	Progressive supranuclear palsy
PiD	Pick's disease
RIN	RNA integrity value
RMSE	Root mean squared error
RNA	Ribonucleic acid

RNA-seq	RNA-sequencing
SG	Stress ganule
SNP	Single nucleotide polymorphism
TF	Transcription factor
TFBS	Transcription factor binding site
WGCNA	Weighted gene correlation network analysis
bvFTD	behavioural variant FTD
lncRNA	long-noncoding RNA
mRNA	messenger RNA
miRNA	micro RNA
nfvPPA	non-fluent variant PPA
scRNA-seq	single-cell RNA-sequencing
smRNA-seq	small RNA-sequencing

Bibliography

[1] dbGaP Study.

[2] D. A. Bennett, J. A. Schneider, Z. Arvanitakis, and R. S. Wilson. Overview and Findings from the Religious Orders Study. *Current Alzheimer Research*, 9(6):628–645, 6 2012.

[3] D. A. Bennett, J. A. Schneider, A. S. Buchman, L. L. Barnes, P. A. Boyle, and R. S. Wilson. Overview and Findings from the Rush Memory and Aging Project. *Current Alzheimer Research*, 9(6):646–663, 6 2012.

[4] A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H. F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE*, 4(7), 7 2009.

[5] Z. Ahmed, E. H. Bigio, H. Budka, D. W. Dickson, I. Ferrer, B. Ghetti, G. Giaccone, K. J. Hatanpaa, J. L. Holton, K. A. Josephs, J. Powers, S. Spina, H. Takahashi, C. L. White, T. Revesz, and G. G. Kovacs. Globular glial tauopathies (GGT): Consensus recommendations. *Acta Neuropathologica*, 126(4):537–544, 2013.

[6] A. Altmann, D. M. Cash, M. Bocchetta, C. Heller, R. Reynolds, K. Moore, R. S. Convery, D. L. Thomas, J. C. van Swieten, F. Moreno, R. Sanchez-Valle, B. Borroni, R. Laforce, M. Masellis, M. C. Tartaglia, C. Graff, D. Galimberti, J. B. Rowe, E. Finger, M. Synofzik, R. Vandenberghe, A. de Mendonça, F. Tagliavini, I. Santana, S. Ducharme, C. R. Butler, A. Gerhard, J. Levin, A. Danek, G. Frisoni, R. Ghidoni, S. Sorbi, M. Otto, M. Ryten, and J. D. Rohrer. Analysis of brain atrophy and local gene expression in genetic frontotemporal dementia. *Brain Communications*, 8 2020.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[8] S. Anders, P. T. Pyl, and W. Huber. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 1 2015.

[9] Y. Aoki, R. Manzano, Y. Lee, R. Dafinca, M. Aoki, A. G. L. Douglas, M. A. Varela, C. Sathyaprakash, J. Scaber, P. Barbagallo, P. Vader, I. Mäger, K. Ezzat,

M. R. Turner, N. Ito, S. Gasco, N. Ohbayashi, S. E. Andaloussi, S. . Ichi Takeda, M. Fukuda, K. Talbot, and M. J. A. Wood. C9orf72 and RAB7L1 regulate vesicle trafficking in amyotrophic lateral sclerosis and frontotemporal dementia. *Brain*, pages 887–897, 2017.

[10] T. Arai, M. Hasegawa, H. Akiyama, K. Ikeda, T. Nonaka, H. Mori, D. Mann, K. Tsuchiya, M. Yoshida, Y. Hashizume, and T. Oda. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochemical and Biophysical Research Communications*, 351(3):602–611, 12 2006.

[11] D. J. Arenillas, A. R. R. Forrest, H. Kawaji, T. Lassmann, T. F. FANTOM Consortium, W. W. Wasserman, and A. Mathelier. CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. *Bioinformatics (Oxford, England)*, 32(18):2858–60, 2016.

[12] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 5 2014.

[13] P. E. Ash, K. F. Bieniek, T. F. Gendron, T. Caulfield, W. L. Lin, M. DeJesus-Hernandez, M. M. Van Blitterswijk, K. Jansen-West, J. W. Paul, R. Rademakers, K. B. Boylan, D. W. Dickson, and L. Petrucelli. Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. *Neuron*, 77(4):639–646, 2 2013.

[14] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. *Jmlr*, 17(June):1–35, 6 2018.

[15] F. Avila Cobos, J. Vandesompele, P. Mestdagh, K. De Preter, and J. Wren. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 1 2018.

[16] E. Bahl, T. Koomar, and J. J. Michaelson. cerebroViz: An R package for anatomical visualization of spatiotemporal brain data. *Bioinformatics*, 33(5):72–763, 3 2017.

[17] H. D. Bain, Y. S. Davidson, A. C. Robinson, S. Ryan, S. Rollinson, A. Richardson, M. Jones, J. S. Snowden, S. Pickering-Brown, and D. M. Mann. The role of lysosomes and autophagosomes in frontotemporal lobar degeneration. *Neuropathology and Applied Neurobiology*, 45(3):244–261, 4 2019.

[18] M. Baker, I. R. Mackenzie, S. M. Pickering-Brown, J. Gass, R. Rademakers, C. Lindholm, J. Snowden, J. Adamson, A. D. Sadovnick, S. Rollinson, A. Cannon, E. Dwosh, D. Neary, S. Melquist, A. Richardson, D. Dickson, Z. Berger, J. Eriksen, T. Robinson, C. Zehr, C. A. Dickey, R. Crook, E. McGowan, D. Mann, B. Boeve, H. Feldman, and M. Hutton. Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature*, 442(7105):916–919, 8 2006.

[19] J. Bang, S. Spina, and B. L. Miller. Frontotemporal dementia. *The Lancet*, 386(10004):1672–1682, 10 2015.

[20] S. Bannwarth, S. Ait-El-Mkadem, A. Chaussenot, E. C. Genin, S. Lacas-Gervais, K. Fragaki, L. Berg-Alonso, Y. Kageyama, V. rie Serre, D. G. Moore, A. Verschueren, C. cile Rouzier, I. Le Ber, G. lle Augé, C. Cochaud, F. Lespinasse, A. de Septenville, A. Brice, P. Yu-Wai-Man, H. Sesaki, J. Pouget, and V. ronique Paquis-Flucklinger. A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *A JOURNAL OF NEUROLOGY*.

[21] F. E. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, 18(7):437–451, 7 2017.

[22] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360, 10 2016.

[23] V. V. Belzil, P. O. Bauer, M. Prudencio, T. F. Gendron, C. T. Stetler, I. K. Yan, L. Pregent, L. Daughrity, M. C. Baker, R. Rademakers, K. Boylan, T. C. Patel, D. W. Dickson, and L. Petrucelli. Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood. *Acta Neuropathologica*, 126(6):895–905, 2013.

[24] Y. Benjamini and Y. Hochberg. Controling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 1995.

[25] R. E. Bennett, A. B. Robbins, M. Hu, X. Cao, R. A. Betensky, T. Clark, S. Das, and B. T. Hyman. Tau induces blood vessel abnormalities and angiogenesis-related gene expression in P301L transgenic mice and human Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 115(6):E1289–E1298, 2 2018.

[26] A. Benussi, A. Alberici, E. Buratti, R. Ghidoni, F. Gardoni, M. D. Luca, A. Padovani, and B. Borroni. Toward a glutamate hypothesis of frontotemporal dementia. *Frontiers in Neuroscience*, 13, 3 2019.

[27] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer.

[28] M. H. U. Biswas, S. Almeida, R. Lopez-Gonzalez, W. Mao, Z. Zhang, A. Karydas, M. D. Geschwind, J. Biernat, E. M. Mandelkow, K. Futai, B. L. Miller, and F. B. Gao. MMP-9 and MMP-2 Contribute to Neuronal Cell Death in iPSC Models of Frontotemporal Dementia with MAPT Mutations. *Stem Cell Reports*, 7(3):316–324, 9 2016.

[29] L. G. Bodea, A. Eckert, L. M. Ittner, O. Piguet, and J. Götz. Tau physiology and pathomechanisms in frontotemporal lobar degeneration. *Journal of Neurochemistry*, 138(Suppl Suppl 1):71–94, 8 2016.

[30] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 9 1991.

[31] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 5 2016.

[32] F. Bright, E. L. Werry, C. Dobson-Stone, O. Piguet, L. M. Ittner, G. M. Halliday, J. R. Hodges, M. C. Kiernan, C. T. Loy, M. Kassiou, and J. J. Kril. Neuroinflammation in frontotemporal dementia. *Nature Reviews Neurology*, pages 1–16, 7 2019.

[33] M. Brkic, S. Balusu, C. Libert, and R. E. Vandenbroucke. Friends or Foes: Matrix Metalloproteinases and Their Multifaceted Roles in Neurodegenerative Diseases. *Mediators of Inflammation*, 2015, 2015.

[34] J. A. Brown, J. Deng, J. Neuhaus, I. J. Sible, A. C. Sias, S. E. Lee, J. Kornak, G. A. Marx, A. M. Karydas, S. Spina, L. T. Grinberg, G. Coppola, D. H. Geschwind, J. H. Kramer, M. L. Gorno-Tempini, B. L. Miller, H. J. Rosen, and W. W. Seeley. Patient-Tailored, Connectivity-Based Forecasts of Spreading Brain Atrophy. *Neuron*, 0(0), 9 2019.

[35] A. Burberry, N. Suzuki, J. Y. Wang, R. Moccia, D. A. Mordes, M. H. Stewart, S. Suzuki-Uematsu, S. Ghosh, A. Singh, F. T. Merkle, K. Koszka, Q. Z. Li, L. Zon, D. J. Rossi, J. J. Trowbridge, L. D. Notarangelo, and K. Eggan. Loss-of-function mutations in the C9ORF72 mouse ortholog cause fatal autoimmune disease. *Science Translational Medicine*, 8(347), 7 2016.

[36] J. R. Burrell, M. C. Kiernan, S. Vucic, and J. R. Hodges. Motor Neuron dysfunction in frontotemporal dementia Abbreviations: ALSFRS-R = Amyotrophic

Lateral Sclerosis Functional Rating Score-Revised; FTD = frontotemporal dementia; FUS = fused in sacrcoma; MND = motor neuron disease; MRCSS = Medical Research Counc. *A JOURNAL OF NEUROLOGY*.

[37] A. Cagnin, M. Rossor, E. L. Sampson, T. MacKinnon, and R. B. Banati. In vivo detection of microglial activation in frontotemporal dementia. *Annals of Neurology*, 56(6):894–897, 12 2004.

[38] J. N. Campbell, E. Z. Macosko, H. Fenselau, T. H. Pers, A. Lyubetskaya, D. Tenen, M. Goldman, A. M. Verstegen, J. M. Resch, S. A. McCarroll, E. D. Rosen, B. B. Lowell, and L. T. Tsai. A molecular census of arcuate hypothalamus and median eminence cell types. *Nature Neuroscience*, 20(3):484–496, 2 2017.

[39] D. M. Cash, M. Bocchetta, D. L. Thomas, K. M. Dick, J. C. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonça, S. Sorbi, M. N. Rossor, S. Ourselin, and J. D. Rohrer. Patterns of gray matter atrophy in genetic frontotemporal dementia: results from the GENFI study. *Neurobiology of Aging*, 62:191–196, 2 2018.

[40] R. Chen, X. Wu, L. Jiang, and Y. Zhang. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell Reports*, 18(13):3227–3241, 3 2017.

[41] Y. Chen, D. Mccarthy, M. Ritchie, M. Robinson, and G. K. Smyth. edgeR: differential expression analysis of digital gene expression data User's Guide, 2008.

[42] S. Cheng, M. Guo, C. Wang, X. Liu, Y. Liu, and X. Wu. MiRTDL: A Deep Learning Approach for miRNA Target Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(6):1161–1169, 2016.

[43] B. P. Chitramuthu, H. P. J. Bennett, and A. Bateman. Progranulin: a new avenue towards the understanding and treatment of neurodegenerative disease. *Brain*, 140(12):3081–3104, 2017.

[44] S. Y. Choi, R. Lopez-Gonzalez, G. Krishnan, H. L. Phillips, A. N. Li, W. W. Seeley, W. D. Yao, S. Almeida, and F. B. Gao. C9ORF72-ALS/FTD-associated poly(GR) binds Atp5a1 and compromises mitochondrial function in vivo. *Nature Neuroscience*, 22(6):851–862, 6 2019.

[45] M. Cruts, I. Gijselinck, J. Van Der Zee, S. Engelborghs, H. Wils, D. Pirici, R. Rademakers, R. Vandenberghe, B. Dermaut, J. J. Martin, C. Van Duijn, K. Peeters, R. Sciot, P. Santens, T. De Pooter, M. Mattheijssens, M. Van Den Broeck, I. Cuijt, K. Vennekens, P. P. De Deyn, S. Kumar-Singh, and

C. Van Broeckhoven. Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature*, 442(7105):920–924, 8 2006.

[46] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Gephart, B. A. Barres, and S. R. Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America*, 112(23):7285–7290, 6 2015.

[47] R. Dayanandan, M. Van Slegtenhorst, T. G. Mack, L. Ko, S. H. Yen, K. Leroy, J. P. Brion, B. H. Anderton, M. Hutton, and S. Lovestone. Mutations in tau reduce its microtubule binding properties in intact cells and affect its phosphorylation. *FEBS Letters*, 446(2-3):228–232, 3 1999.

[48] M. DeJesus-Hernandez, I. Mackenzie, B. Boeve, A. Boxer, M. Baker, N. Rutherford, A. Nicholson, N. Finch, H. Flynn, J. Adamson, N. Kouri, A. Wojtas, P. Sengdy, G.-Y. Hsiung, A. Karydas, W. Seeley, K. Josephs, G. Coppola, D. Geschwind, Z. Wszolek, H. Feldman, D. Knopman, R. Petersen, B. Miller, D. Dickson, K. Boylan, N. Graff-Radford, and R. Rademakers. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, 72(2):245–256, 10 2011.

[49] Z. Deng, P. Sheehan, S. Chen, and Z. Yue. Is amyotrophic lateral sclerosis/frontotemporal dementia an autophagy disease? *Molecular Neurodegeneration*, 12(1), 12 2017.

[50] A. Dhingra, J. Täger, E. Bressan, S. Rodriguez-Nieto, M. S. Bedi, S. Bröer, E. Sadikoglou, N. Fernandes, M. Castillo-Lizardo, P. Rizzu, and P. Heutink. Automated production of human induced pluripotent stem cell-derived cortical and dopaminergic neurons with integrated live-cell monitoring. *Journal of Visualized Experiments*, 2020(162):1–29, 8 2020.

[51] D. W. Dickson, N. Kouri, M. E. Murray, and K. A. Josephs. Neuropathology of frontotemporal lobar degeneration-Tau (FTLD-Tau). In *Journal of Molecular Neuroscience*, volume 45, pages 384–389. Springer, 11 2011.

[52] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013.

[53] A. T. Du, G. H. Jahng, S. Hayasaka, J. H. Kramer, H. J. Rosen, M. L. Gorno-Tempini, K. P. Rankin, B. L. Miller, M. W. Weiner, and N. Schuff. Hypoperfusion in frontotemporal dementia and Alzheimer disease by arterial spin labeling MRI. *Neurology*, 67(7):1215–1220, 10 2006.

[54] J. Duchi, E. Hazan, Y. Singer, and G. Research. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Technical report.

[55] M. Egeblad, E. S. Nakasone, and Z. Werb. Tumors as organs: Complex tissues that interface with the entire organism. *Developmental Cell*, 18(6):884–901, 3 2010.

[56] H. Ek Olofsson and E. Englund. A cortical microvascular structure in vascular dementia, Alzheimer's disease, frontotemporal lobar degeneration and nondemented controls: a sign of angiogenesis due to brain ischaemia? *Neuropathology and Applied Neurobiology*, 45(6):557–569, 10 2019.

[57] F. M. Elahi and B. L. Miller. A clinicopathological approach to the diagnosis of dementia. *Nature Reviews Neurology*, 13(8):457–476, 8 2017.

[58] G. Eraslan, Avsec, J. Gagneur, and F. J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, 7 2019.

[59] P. A. Ewels, A. Peltzer, S. Fillinger, J. Alneberg, H. Patel, A. Wilm, M. U. Garcia, P. D. Tommaso, and S. Nahnsen. nf-core: Community curated bioinformatics pipelines. *bioRxiv*, page 610741, 2019.

[60] G. B. Fields. The Rebirth of Matrix Metalloproteinase Inhibitors: Moving Beyond the Dogma. *Cells*, 8(9), 8 2019.

[61] P. J. Flannery and E. Trushina. Mitochondrial dynamics and transport in Alzheimer's disease, 7 2019.

[62] A. Frishberg, N. Peshes-Yaloz, O. Cohn, D. Rosentul, Y. Steuerman, L. Valadarsky, G. Yankovitz, M. Mandelboim, F. A. Iraqi, I. Amit, L. Mayo, E. Bacharach, and I. Gat-Viks. Cell composition analysis of bulk genomics using single-cell data. *Nature Methods*, 16(4):327–332, 4 2019.

[63] H. Fu, J. Hardy, and K. E. Duff. Selective vulnerability in neurodegenerative diseases. *Nature Neuroscience*, 21(10):1350–1358, 10 2018.

[64] H. Fu, A. Possenti, R. Freer, Y. Nakano, N. C. Villegas, M. Tang, P. V. Cauhy, B. A. Lassus, S. Chen, S. L. Fowler, H. Y. Figueroa, E. D. Huey, G. V. Johnson, M. Vendruscolo, and K. E. Duff. A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. *Nature Neuroscience*, 22(1):47–56, 1 2019.

[65] L. Gan, M. R. Cookson, L. Petrucelli, and A. R. La Spada. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nature Neuroscience*, 21(10):1300–1309, 10 2018.

[66] T. F. Gendron, K. F. Bieniek, Y. J. Zhang, K. Jansen-West, P. E. Ash, T. Caulfield, L. Daughrity, J. H. Dunmore, M. Castanedes-Casey, J. Chew, D. M. Cosio, M. Van Blitterswijk, W. C. Lee, R. Rademakers, K. B. Boylan, D. W. Dickson, and L. Petrucelli. 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):829–844, 2013.

[67] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve, F. Manes, N. F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B. L. Miller, D. S. Knopman, J. R. Hodges, M. M. Mesulam, and M. Grossman. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014, 3 2011.

[68] M. V. Greenberg and D. Bourc'his. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, 20(10):590–607, 10 2019.

[69] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, D. I. Chasman, G. A. Fitzgerald, K. Dolinski, T. Grosser, and O. G. Troyanskaya. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, 5 2015.

[70] V. Haberle, A. R. Forrest, Y. Hayashizaki, P. Carninci, and B. Lenhard. CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Research*, 43(8):e51–e51, 1 2015.

[71] A. R. Haeusler, C. J. Donnelly, and J. D. Rothstein. The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. *Nature Reviews Neuroscience*, 17(6):383–395, 6 2016.

[72] C. Halabi, A. Halabi, D. L. Dean, P. N. Wang, A. L. Boxer, J. Q. Trojanowski, S. J. Dearmond, B. L. Miller, J. H. Kramer, and W. W. Seeley. Patterns of striatal degeneration in frontotemporal dementia. *Alzheimer Disease and Associated Disorders*, 27(1):74–83, 1 2013.

[73] L. Handl, A. Jalali, M. Scherer, R. Eggeling, and N. Pfeifer. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. In *Bioinformatics*, volume 35, pages i154–i163. Oxford University Press, 7 2019.

[74] J. Hausser and M. Zavolan. Identification and consequences of miRNA-target interactions-beyond repression of gene expression. *Nature Reviews Genetics*, 15(9):599–612, 7 2014.

[75] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, 5 2010.

[76] M. Hemberg and V. Kiselev. scRNA-Seq Datasets.

[77] M. T. Heneka, M. P. Kummer, A. Stutz, A. Delekate, S. Schwartz, A. Vieira-Saecker, A. Griep, D. Axt, A. Remus, T. C. Tzeng, E. Gelpi, A. Halle, M. Korte, E. Latz, and D. T. Golenbock. NLRP3 is activated in Alzheimer's disease and contributes to pathology in APP/PS1 mice. *Nature*, 493(7434):674–678, 1 2013.

[78] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19:562–578, 2018.

[79] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. Technical report.

[80] R. Hrdlickova, M. Toloue, and B. Tian. RNA-Seq methods for transcriptome analysis. *Wiley interdisciplinary reviews. RNA*, 8(1), 2017.

[81] W. T. Hu, Z. Wang, V. M. Lee, J. Q. Trojanowski, J. A. Detre, and M. Grossman. Distinct cerebral perfusion patterns in FTLD and AD. *Neurology*, 75(10):881–888, 9 2010.

[82] E. Huntzinger and E. Izaurralde. Gene silencing by microRNAs: Contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, 12(2):99–110, 2 2011.

[83] M. Hutson. Has artificial intelligence become alchemy?, 5 2018.

[84] M. Hutton, C. L. Lendon, P. Rizzu, M. Baker, S. Froelich, H. Houlden, S. Pickering-Brown, S. Chakraverty, A. Isaacs, A. Grover, J. Hackett, J. Adamson, S. Lincoln, D. Dickson, P. Davies, R. C. Petersen, M. Stevens, E. de Graaff, E. Wauters, J. van Baren, M. Hillebrand, M. Joosse, J. M. Kwon, P. Nowotny, L. K. Che, J. Norton, J. C. Morris, L. A. Reed, J. Trojanowski, H. Basun, L. Lannfelt, M. Neystat, S. Fahn, F. Dark, T. Tannenberg, P. R. Dodd, N. Hayward, J. B. J. Kwok, P. R. Schofield, A. Andreadis, J. Snowden, D. Craufurd, D. Neary, F. Owen, B. A. Oostra, J. Hardy, A. Goate, J. van Swieten, D. Mann, T. Lynch, and P. Heutink. Association of missense and 5-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, 393(6686):702–705, 6 1998.

[85] I. P. Ioshikhes and M. Q. Zhang. Large-scale human promoter mapping using CpG islands. *Nature Genetics*, 26(1):61–63, 9 2000.

[86] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, 1 2020.

[87] H. Jiang, R. Lei, S. W. Ding, and S. Zhu. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(1):182, 6 2014.

[88] K. A. Josephs. Frontotemporal dementia and related disorders: Deciphering the enigma. *Annals of Neurology*, 64(1):4–14, 7 2008.

[89] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 1 2016.

[90] S. Kausar, F. Wang, and H. Cui. The Role of Mitochondria in Reactive Oxygen Species Generation and Its Implications for Neurodegenerative Diseases. *Cells*, 7(12):274, 12 2018.

[91] D. R. Kelley, J. Snoek, and J. L. Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, 7 2016.

[92] K. W. Kelley, H. Nakao-Inoue, A. V. Molofsky, and M. C. Oldham. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nature Neuroscience*, 21(9):1171–1184, 9 2018.

[93] M. Kiaei, K. Kipiani, N. Y. Calingasan, E. Wille, J. Chen, B. Heissig, S. Rafii, S. Lorenzl, and M. F. Beal. Matrix metalloproteinase-9 regulates TNF-α and FasL expression in neuronal, glial cells and its absence extends life in a transgenic mouse model of amyotrophic lateral sclerosis. *Experimental Neurology*, 205(1):74–81, 5 2007.

[94] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, 8 2019.

[95] Y. S. Kim and T. H. Joh. Matrix metalloproteinases, new insights into the understanding of neurodegenerative disorders. *Biomolecules and Therapeutics*, 20(2):133–143, 5 2012.

[96] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. 12 2014.

[97] G. Kleinberger, A. Capell, C. Haass, and C. Van Broeckhoven. Mechanisms of granulin deficiency: Lessons from cellular and animal models. *Molecular Neurobiology*, 47(1):337–360, 2013.

[98] C. Y. Ko, L. H. Chang, Y. C. Lee, E. Sterneck, C. P. Cheng, S. H. Chen, A. M. Huang, J. T. Tseng, and J. M. Wang. CCAAT/enhancer binding protein delta (CEBPD) elevating PTX3 expression inhibits macrophage-mediated phagocytosis of dying neuron cells. *Neurobiology of Aging*, 33(2):11–422, 2012.

[99] C. Y. Ko, W. C. Chang, and J. M. Wang. Biological roles of CCAAT/enhancer-binding protein delta during inflammation. *Journal of Biomedical Science*, 22(1), 1 2015.

[100] R. Köchl, X. W. Hu, E. Y. Chan, and S. A. Tooze. Microtubules facilitate autophagosome formation and fusion of autophagosomes with endosomes. *Traffic*, 7(2):129–145, 2 2006.

[101] R. Kolde. pheatmap. 2015.

[102] T. Konno, O. A. Ross, H. A. Teive, J. Sławek, D. W. Dickson, and Z. K. Wszolek. DCTN1-related neurodegeneration: Perry syndrome and beyond. *Parkinsonism and Related Disorders*, 41:14–24, 8 2017.

[103] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones. MiRBase: From microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, 1 2019.

[104] A. Krishnan, R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, A. Tadych, N. Volfovsky, A. Packer, A. Lash, and O. G. Troyanskaya. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 19(11):1454–1462, 10 2016.

[105] A. Kuhn, D. Thu, H. J. Waldvogel, R. L. M. Faull, and R. Luthi-Carter. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature Methods*, 8(11):945–947, 11 2011.

[106] B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, and K. Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science (New York, N.Y.)*, 352(6293):1586–90, 6 2016.

[107] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. The Human Transcription Factors. 2018.

[108] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 2008.

[109] T. Lassmann. TagDust2: A generic method to extract reads from sequencing data. *BMC Bioinformatics*, 16(1):24, 1 2015.

[110] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015.

[111] J. K. Lee, J. H. Shin, J. H. Suh, I. S. Choi, K. S. Ryu, and B. J. Gwag. Tissue inhibitor of metalloproteinases-3 (TIMP-3) expression is increased during serum deprivation-induced neuronal apoptosis in vitro and in the G93A mouse model of amyotrophic lateral sclerosis: A potential modulator of Fas-mediated apoptosis. *Neurobiology of Disease*, 30(2):174–185, 5 2008.

[112] J. T. Leek and J. D. Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9):e161, 2007.

[113] T. P. Levine, R. D. Daniels, A. T. Gatta, L. H. Wong, M. J. Hayes, and A. Bateman. The product of C9orf72, a gene strongly implicated in neurodegeneration, is structurally related to DENN Rab-GEFs. *BIOINFORMATICS*, 29(4):499–503, 2013.

[114] Y. Liao, G. K. Smyth, and W. Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 4 2014.

[115] S.-C. Ling, M. Polymenidou, and D. Cleveland. Converging Mechanisms in ALS and FTD: Disrupted RNA and Protein Homeostasis. *Neuron*, 79(3):416–438, 8 2013.

[116] F. Liu, H. Li, C. Ren, X. Bo, and W. Shu. PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. *Scientific Reports*, 6(1):1–14, 6 2016.

[117] S. Lorenzl, S. Narr, B. Angele, H. W. Krell, J. Gregorio, M. Kiaei, H. W. Pfister, and M. F. Beal. The matrix metalloproteinases inhibitor Ro 26-2853 extends survival in transgenic ALS mice. *Experimental Neurology*, 200(1):166–171, 7 2006.

[118] M. Lotfollahi, F. A. Wolf, and F. J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 8 2019.

[119] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.

[120] H. Lui, J. Zhang, S. R. Makinson, M. K. Cahill, K. W. Kelley, H. Y. Huang, Y. Shang, M. C. Oldham, L. H. Martens, F. Gao, G. Coppola, S. A. Sloan, C. L. Hsieh, C. C. Kim, E. H. Bigio, S. Weintraub, M. M. Mesulam, R. Rademakers, I. R. MacKenzie, W. W. Seeley, A. Karydas, B. L. Miller, B. Borroni, R. Ghidoni, R. V. Farese, J. T. Paz, B. A. Barres, and E. J. Huang. Progranulin Deficiency Promotes Circuit-Specific Synaptic Pruning by Microglia via Complement Activation. *Cell*, 165(4):921–935, 5 2016.

[121] I. R. A. Mackenzie, E. H. Bigio, P. G. Ince, F. Geser, M. Neumann, N. J. Cairns, L. K. Kwong, M. S. Forman, J. Ravits, H. Stewart, A. Eisen, L. McClusky, H. A. Kretzschmar, C. M. Monoranu, J. R. Highley, J. Kirby, T. Siddique, P. J. Shaw, V. M.-Y. Lee, and J. Q. Trojanowski. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis withSOD1 mutations. *Annals of Neurology*, 61(5):427–434, 5 2007.

[122] I. R. A. Mackenzie and M. Neumann. Molecular neuropathology of frontotemporal dementia: insights into disease mechanisms from postmortem studies. *Journal of Neurochemistry*, 138:54–70, 8 2016.

[123] M. Marouf, P. Machart, V. Bansal, C. Kilian, D. S. Magruder, C. F. Krebs, and S. Bonn. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, 11(1):1–12, 12 2020.

[124] M. E. McCauley, J. G. O'Rourke, A. Yáñez, J. L. Markman, R. Ho, X. Wang, S. Chen, D. Lall, M. Jin, A. K. M. G. Muhammad, S. Bell, J. Landeros, V. Valencia, M. Harms, M. Arditi, C. Jefferies, and R. H. Baloh. C9orf72 in myeloid cells suppresses STING-induced inflammation. *Nature*, pages 1–6, 8 2020.

[125] K. Menden, M. Marouf, A. Dalmia, P. Heutink, and S. Bonn. Deep-learning-based cell composition analysis from tissue expression profiles. *bioRxiv*, page 659227, 11 2019.

[126] K. Menden, M. Marouf, S. Oller, A. Dalmia, D. S. Magruder, K. Kloiber, P. Heutink, and S. Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, 7 2020.

[127] Z. A. Miller, K. P. Rankin, N. R. Graff-Radford, L. T. Takada, V. E. Sturm, C. M. Cleveland, L. A. Criswell, P. A. Jaeger, T. Stan, K. A. Heggeli, S. C. Hsu, A. Karydas, B. K. Khan, L. T. Grinberg, M. L. Gorno-Tempini, A. L. Boxer, H. J. Rosen, J. H. Kramer, G. Coppola, D. H. Geschwind, R. Rademakers, W. W. Seeley, T. Wyss-Coray, and B. L. Miller. TDP-43 frontotemporal lobar degeneration and autoimmune disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 84(9):956–962, 9 2013.

[128] M. Mishra, T. Paunesku, G. E. Woloschak, T. Siddique, L. Zhu, S. Lin, K. Greco, and E. H. Bigio. Gene expression analysis of frontotemporal lobar degeneration of the motor neuron disease type with ubiquitinated inclusions. *Acta Neuropathologica*, 114(1):81–94, 7 2007.

[129] S. Mizielinska, T. Lashley, F. E. Norona, E. L. Clayton, C. E. Ridler, P. Fratta, and A. M. Isaacs. C9orf72 frontotemporal lobar degeneration is characterised by frequent neuronal sense and antisense RNA foci. *Acta Neuropathologica*, 126(6):845–857, 10 2013.

[130] S. Mohammadi, N. Zuckerman, A. Goldsmith, and A. Grama. A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues. *arXiv*, 10 2015.

[131] T. Molnár, A. Mázló, V. Tslaf, A. G. Szöllősi, G. Emri, and G. Koncz. Current translational potential and underlying molecular mechanisms of necroptosis. *Cell Death and Disease*, 10(11):1–21, 11 2019.

[132] G. Monaco, B. Lee, W. Xu, S. Mustafah, Y. Y. Hwang, C. Carré, N. Burdin, L. Visan, M. Ceccarelli, M. Poidinger, A. Zippelius, J. Pedro de Magalhães, and A. Larbi. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*, 26(6):1627–1640, 2 2019.

[133] D. M. Moujalled, W. D. Cook, J. M. Murphy, and D. L. Vaux. Necroptosis induced by RIPK3 requires MLKL but not Drp1. *Cell Death and Disease*, 5(2):e1086, 2 2014.

[134] A. G. Murley and J. B. Rowe. Neurotransmitter deficits from fronto temporal lobar degeneration. *Brain*, 141(5):1263–1285, 5 2018.

[135] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. Technical report.

[136] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 3 1970.

[137] M. Neumann, R. Rademakers, S. Roeber, M. Baker, H. A. Kretzschmar, I. R. A. Mackenzie, and I. Mackenzie. A new subtype of frontotemporal lobar degeneration with FUS pathology. *A JOURNAL OF NEUROLOGY*.

[138] M. Neumann, D. M. Sampathu, L. K. Kwong, A. C. Truax, M. C. Micsenyi, T. T. Chou, J. Bruce, T. Schuck, M. Grossman, C. M. Clark, L. F. McCluskey, B. L. Miller, E. Masliah, I. R. Mackenzie, H. Feldman, W. Feiden, H. A. Kretzschmar,

J. Q. Trojanowski, and V. M. Lee. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, 314(5796):130–133, 10 2006.

[139] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 2015.

[140] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, page 1, 5 2019.

[141] J. G. O'Rourke, L. Bogdanik, A. Yáñez, D. Lall, A. J. Wolf, A. K. Muhammad, R. Ho, S. Carmona, J. P. Vit, J. Zarrow, K. J. Kim, S. Bell, M. B. Harms, T. M. Miller, C. A. Dangler, D. M. Underhill, H. S. Goodridge, C. M. Lutz, and R. H. Baloh. C9orf72 is required for proper macrophage and microglial function in mice. *Science*, 351(6279):1324–1329, 3 2016.

[142] F. Palese, E. Bonomi, T. Nuzzo, A. Benussi, M. Mellone, E. Zianni, F. Cisani, A. Casamassa, A. Alberici, D. Scheggia, A. Padovani, E. Marcello, M. Di Luca, A. Pittaluga, A. Usiello, B. Borroni, and F. Gardoni. Anti-GluA3 antibodies in frontotemporal dementia: effects on glutamatergic neurotransmission and synaptic failure. *Neurobiology of Aging*, 86:143–155, 2 2020.

[143] F. Panza, M. Lozupone, D. Seripa, A. Daniele, M. Watling, G. Giannelli, and B. P. Imbimbo. Development of disease-modifying drugs for frontotemporal dementia spectrum disorders. *Nature Reviews Neurology*, 16(4):213–228, 4 2020.

[144] E. Park, Z. Pan, Z. Zhang, L. Lin, and Y. Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American journal of human genetics*, 102(1):11–26, 2018.

[145] E. Patrick, M. Taga, A. Ergun, B. Ng, W. Casazza, M. Cimpean, C. Yung, J. Schneider, D. Bennett, C. Gaiteri, P. Jager, E. Bradshaw, and S. Mostafavi. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *bioRxiv*, page 566307, 3 2019.

[146] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Publishing Group*, 14, 2017.

[147] P. Paul, A. Chakraborty, D. Sarkar, M. Langthasa, M. Rahman, M. Bari, R. K. Singha, A. K. Malakar, and S. Chakraborty. Interplay between miRNAs and human diseases. *Journal of Cellular Physiology*, 233(3):2007–2018, 3 2018.

[148] D. H. Paushter, H. Du, T. Feng, and F. Hu. The lysosomal function of progranulin, a guardian against neurodegeneration. *Acta Neuropathologica*, 136(1):1, 7 2018.

[149] I. S. Piras, J. Krate, E. Delvaux, J. Nolz, D. F. Mastroeni, A. M. Persico, W. M. Jepsen, T. G. Beach, M. J. Huentelman, P. D. Coleman, and C. Combs. Transcriptome Changes in the Alzheimer's Disease Middle Temporal Gyrus: Importance of RNA Metabolism and Mitochondria-Associated Membrane Genes. *Journal of Alzheimer's Disease*, 70(3):691–713, 2019.

[150] R. Poplin, P. C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, and M. A. Depristo. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983, 11 2018.

[151] C. Pottier, T. A. Ravenscroft, M. Sanchez-Contreras, and R. Rademakers. Genetics of FTLD: overview and what else we can expect from genetic studies. *Journal of Neurochemistry*, 2016.

[152] K. K. Pramanik, S. Nagini, A. K. Singh, P. Mishra, T. Kashyap, N. Nath, M. Alam, A. Rana, and R. Mishra. Glycogen synthase kinase-3β mediated regulation of matrix metalloproteinase-9 and its involvement in oral squamous cell carcinoma progression and invasion. *Cellular Oncology*, 41(1):47–60, 2 2018.

[153] R. U. Rahman, A. Gautam, J. Bethune, A. Sattar, M. Fiosins, D. S. Magruder, V. Capece, O. Shomroni, and S. Bonn. Oasis 2: Improved online analysis of small RNA-seq data. *BMC Bioinformatics*, 19(1):54, 2 2018.

[154] K. Rascovsky, J. R. Hodges, D. Knopman, M. F. Mendez, J. H. Kramer, J. Neuhaus, J. C. van Swieten, H. Seelaar, E. G. P. Dopper, C. U. Onyike, A. E. Hillis, K. A. Josephs, B. F. Boeve, A. Kertesz, W. W. Seeley, K. P. Rankin, J. K. Johnson, M.-L. Gorno-Tempini, H. Rosen, C. E. Prioleau-Latham, A. Lee, C. M. Kipps, P. Lillo, O. Piguet, J. D. Rohrer, M. N. Rossor, J. D. Warren, N. C. Fox, D. Galasko, D. P. Salmon, S. E. Black, M. Mesulam, S. Weintraub, B. C. Dickerson, J. Diehl-Schmid, F. Pasquier, V. Deramecourt, F. Lebert, Y. Pijnenburg, T. W. Chow, F. Manes, J. Grafman, S. F. Cappa, M. Freedman, M. Grossman, and B. L. Miller. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain : a journal of neurology*, 134(Pt 9):2456–77, 9 2011.

[155] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 7 2019.

[156] P. Reinhardt, M. Glatza, K. Hemmer, Y. Tsytsyura, and C. S. Thiel. Derivation and Expansion Using Only Small Molecules of Human Neural Progenitors for Neurodegenerative Disease Modeling. *PLoS ONE*, 8(3):59252, 2013.

[157] R. G. Rempe, A. M. Hartz, and B. Bauer. Matrix metalloproteinases in the brain and blood-brain barrier: Versatile breakers and makers. *Journal of Cerebral Blood Flow and Metabolism*, 36(9):1481–1507, 9 2016.

[158] A. Renton, E. Majounie, A. Waite, J. Simón-Sánchez, S. Rollinson, J. Gibbs, J. Schymick, H. Laaksovirta, J. van Swieten, L. Myllykangas, H. Kalimo, A. Paetau, Y. Abramzon, A. Remes, A. Kaganovich, S. Scholz, J. Duckworth, J. Ding, D. Harmer, D. Hernandez, J. Johnson, K. Mok, M. Ryten, D. Trabzuni, R. Guerreiro, R. Orrell, J. Neal, A. Murray, J. Pearson, I. Jansen, D. Sondervan, H. Seelaar, D. Blake, K. Young, N. Halliwell, J. Callister, G. Toulson, A. Richardson, A. Gerhard, J. Snowden, D. Mann, D. Neary, M. Nalls, T. Peuralinna, L. Jansson, V.-M. Isoviita, A.-L. Kaivorinne, M. Hölttä-Vuori, E. Ikonen, R. Sulkava, M. Benatar, J. Wuu, A. Chiò, G. Restagno, G. Borghero, M. Sabatelli, D. Heckerman, E. Rogaeva, L. Zinman, J. Rothstein, M. Sendtner, C. Drepper, E. Eichler, C. Alkan, Z. Abdullaev, S. Pack, A. Dutra, E. Pak, J. Hardy, A. Singleton, N. Williams, P. Heutink, S. Pickering-Brown, H. Morris, P. Tienari, and B. Traynor. A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron*, 72(2):257–268, 10 2011.

[159] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 2015.

[160] S. Rivera, L. García-González, M. Khrestchatisky, and K. Baranger. Metalloproteinases and their tissue inhibitors in Alzheimer's disease and other neurodegenerative disorders. *Cellular and Molecular Life Sciences*, 76(16):3167–3191, 8 2019.

[161] R. D. Rodriguez and L. T. Grinberg. Argyrophilic grain disease: An underestimated tauopathy. *Dementia e Neuropsychologia*, 9(1):2–8, 2015.

[162] R. A. Romanov, A. Zeisel, J. Bakker, F. Girach, A. Hellysaz, R. Tomer, A. Alpár, J. Mulder, F. Clotman, E. Keimpema, B. Hsueh, A. K. Crow, H. Martens, C. Schwindling, D. Calvigioni, J. S. Bains, Z. Máté, G. Szabó, Y. Yanagawa, M. D. Zhang, A. Rendeiro, M. Farlik, M. Uhlén, P. Wulff, C. Bock, C. Broberger, K. Deisseroth, T. Hökfelt, S. Linnarsson, T. L. Horvath, and T. Harkany. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nature Neuroscience*, 20(2):176–188, 2 2017.

[163] D. R. Rosen, T. Siddique, D. Patterson, D. A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, J. P. O'Regan, H. X. Deng, Z. Rahmani, A. Krizus, D. McKenna-Yasek, A. Cayabyab, S. M. Gaston, R. Berger, R. E. Tanzi, J. J. Halperin, B. Herzfeldt, R. V. Den Bergh, W. Y. Hung, T. Bird, G. Deng, D. W. Mulder, C. Smyth, N. G. Laing, E. Soriano, M. A. Pericak-Vance, J. Haines, G. A. Rouleau, J. S. Gusella, H. R. Horvitz, and R. H. Brown. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature*, 362(6415):59–62, 3 1993.

[164] G. A. Rosenberg. Matrix metalloproteinases and their multiple roles in neurodegenerative diseases. *The Lancet Neurology*, 8(2):205–216, 2 2009.

[165] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[166] S. Saberi, J. E. Stauffer, J. Jiang, S. D. Garcia, A. E. Taylor, D. Schulte, T. Ohkubo, C. L. Schloffman, M. Maldonado, M. Baughn, M. J. Rodriguez, D. Pizzo, D. Cleveland, and J. Ravits. Sense-encoded poly-GR dipeptide repeat proteins correlate to neurodegeneration and uniquely co-localize with TDP-43 in dendrites of repeat-expanded C9orf72 amyotrophic lateral sclerosis. *Acta Neuropathologica*, 135(3):459–474, 3 2018.

[167] S. Sainsbury, C. Bernecky, and P. Cramer. Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3):129–143, 3 2015.

[168] N. Sakae, S. F. Roemer, K. F. Bieniek, M. E. Murray, M. C. Baker, K. Kasanuki, N. R. Graff-Radford, L. Petrucelli, M. Van Blitterswijk, R. Rademakers, and D. W. Dickson. Microglia in frontotemporal lobar degeneration with progranulin or C9ORF72 mutations. *Annals of Clinical and Translational Neurology*, 6(9):1782–1796, 9 2019.

[169] D. W. Sanders, S. K. Kaufman, S. L. DeVos, A. M. Sharma, H. Mirbaha, A. Li, S. J. Barker, A. C. Foley, J. R. Thorpe, L. C. Serpell, T. M. Miller, L. T. Grinberg, W. W. Seeley, and M. I. Diamond. Distinct tau prion strains propagate in cells and mice and define different tauopathies. *Neuron*, 82(6):1271–1288, 6 2014.

[170] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 5 2015.

[171] M. Schelker, S. Feau, J. Du, N. Ranu, E. Klipp, G. MacBeath, B. Schoeberl, and A. Raue. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature Communications*, 8(1):2032, 12 2017.

[172] Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. Smith, M. Kasper, C. Ämmälä, and R. Sandberg. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, 24(4):593–607, 10 2016.

[173] C. Sellier, M. Campanari, C. Julie Corbier, A. Gaucherot, I. Kolb-Cheynel, M. Oulad-Abdelghani, F. Ruffenach, A. Page, S. Ciura, E. Kabashi, and N. Charlet-Berguerand. Loss of C9 ¡scp¿ORF¡/scp¿ 72 impairs autophagy and synergizes with polyQ Ataxin-2 to induce motor neuron dysfunction and cell death. *The EMBO Journal*, 35(12):1276–1297, 6 2016.

[174] S. S. Shaftel, W. S. T. Griffin, and K. M. Kerry. The role of interleukin-1 in neuroinflammation and Alzheimer disease: An evolving perspective. *Journal of Neuroinflammation*, 5:7, 2 2008.

[175] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 11 2003.

[176] E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 9 2013.

[177] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. 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):15776–15781, 12 2003.

[178] N. G. Skene, J. Bryois, T. E. Bakken, G. Breen, J. J. Crowley, H. A. Gaspar, P. Giusti-Rodriguez, R. D. Hodge, J. A. Miller, A. B. Muñoz-Manchado, M. C. O'Donovan, M. J. Owen, A. F. Pardiñas, J. Ryge, J. T. R. Walters, S. Linnarsson, E. S. Lein, P. F. Sullivan, and J. Hjerling-Leffler. Genetic identification of brain cell types underlying schizophrenia. *Nature Genetics*, page 1, 5 2018.

[179] N. G. Skene and S. G. N. Grant. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Frontiers in Neuroscience*, 10:16, 1 2016.

[180] G. Skibinski, N. J. Parkinson, J. M. Brown, L. Chakrabarti, S. L. Lloyd, H. Hummerich, J. E. Nielsen, J. R. Hodges, M. G. Spillantini, T. Thusgaard, S. Brandner,

A. Brun, M. N. Rossor, A. Gade, P. Johannsen, S. A. Sørensen, S. Gydesen, E. M. Fisher, and J. Collinge. Mutations in the endosomal ESCRTIII-complex subunit CHMP2B in frontotemporal dementia. *Nature Genetics*, 37(8):806–808, 8 2005.

[181] K. R. Smith, J. Damiano, S. Franceschetti, S. Carpenter, L. Canafoglia, M. Morbin, G. Rossi, D. Pareyson, S. E. Mole, J. F. Staropoli, K. B. Sims, J. Lewis, W. L. Lin, D. W. Dickson, H. H. Dahl, M. Bahlo, and S. F. Berkovic. Strikingly different clinicopathological phenotypes determined by progranulin-mutation dosage. *American Journal of Human Genetics*, 90(6):1102–1107, 6 2012.

[182] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 3 1981.

[183] F. Spitz and E. E. Furlong. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 9 2012.

[184] W. Stoothoff, P. B. Jones, T. L. Spires-Jones, D. Joyner, E. Chhabra, K. Bercury, Z. Fan, H. Xie, B. Bacskai, J. Edd, D. Irimia, and B. T. Hyman. Differential effect of three-repeat and four-repeat tau on mitochondrial axonal transport. *Journal of Neurochemistry*, 111(2):417–427, 10 2009.

[185] K. H. Strang, T. E. Golde, and B. I. Giasson. MAPT mutations, tauopathy, and mechanisms of neurodegeneration. *Laboratory Investigation*, 99(7):912–928, 7 2019.

[186] V. Swarup, F. I. Hinz, J. E. Rexach, K.-i. Noguchi, H. Toyoshiba, A. Oda, K. Hirai, A. Sarkar, N. T. Seyfried, C. Cheng, S. J. Haggarty, M. Grossman, V. M. Van Deerlin, J. Q. Trojanowski, J. J. Lah, A. I. Levey, S. Kondou, and D. H. Geschwind. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. *Nature Medicine*, 25(1):152–164, 1 2019.

[187] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 1 2019.

[188] H. Takahashi, T. Lassmann, M. Murata, and P. Carninci. 5 end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols*, 7(3):542–561, 2 2012.

[189] Y. Tanaka, T. Matsuwaki, K. Yamanouchi, and M. Nishihara. Exacerbated inflammatory responses related to activated microglia after traumatic brain injury in progranulin-deficient mice. *Neuroscience*, 231:49–60, 2 2013.

[190] X. Tang, A. Toro, T. G. Sahana, J. Gao, J. Chalk, B. E. Oskarsson, K. Zhang, and K. Zhang. Divergence, Convergence, and Therapeutic Implications: A Cell Biology Perspective of C9ORF72-ALS/FTD. *Molecular Neurodegeneration*, 15(1):34, 6 2020.

[191] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. M. Sunkin, M. Hawrylycz, C. Koch, and H. Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335–346, 1 2016.

[192] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[193] M. E. Umoh, E. B. Dammer, J. Dai, D. M. Duong, J. J. Lah, A. I. Levey, M. Gearing, J. D. Glass, and N. T. Seyfried. A proteomic network approach across the ALS - FTD disease spectrum resolves clinical phenotypes and genetic vulnerability in human brain. *EMBO Molecular Medicine*, 10(1):48–62, 1 2018.

[194] E. M. Valente, A. R. Bentivoglio, P. H. Dixon, A. Ferraris, T. Lalongo, M. Frontali, A. Albanese, and N. W. Wood. Localization of a novel locus for autosomal recessive early-onset parkinsonism, PARK6, on human chromosome 1p35-p36. *American Journal of Human Genetics*, 68(4):895–900, 2001.

[195] F. Vallania, A. Tam, S. Lofgren, S. Schaffert, T. D. Azad, E. Bongen, W. Haynes, M. Alsup, M. Alonso, M. Davis, E. Engleman, and P. Khatri. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature Communications*, 9(1):4735, 12 2018.

[196] M. Van Blitterswijk, M. Dejesus-Hernandez, and R. Rademakers. How do C9ORF72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: Can we learn from other noncoding repeat expansion disorders? *Current Opinion in Neurology*, 25(6):689–700, 12 2012.

[197] P. Van Damme, A. Van Hoecke, D. Lambrechts, P. Vanacker, E. Bogaert, J. Van Swieten, P. Carmeliet, L. Van Den Bosch, and W. Robberecht. Progranulin functions as a neurotrophic factor to regulate neurite outgrowth and enhance neuronal survival. *Journal of Cell Biology*, 181(1):37–41, 4 2008.

[198] B. van Wilgenburg, C. Browne, J. Vowles, and S. A. Cowley. Efficient, Long Term Production of Monocyte-Derived Macrophages from Human Pluripotent Stem Cells under Partly-Defined and Fully-Defined Conditions. *PLoS ONE*, 8(8), 8 2013.

[199] R. E. Vandenbroucke, E. Dejonckheere, F. Van Hauwermeiren, S. Lodens, R. De Rycke, E. Van Wonterghem, A. Staes, K. Gevaert, C. López-Otin, and C. Libert. Matrix metalloproteinase 13 modulates intestinal epithelial barrier integrity in inflammatory diseases by activating TNF. *EMBO Molecular Medicine*, 5(7):1000–1016, 2013.

[200] S. Vatovec, A. Kovanda, and B. Rogelj. Unconventional features of C9ORF72 expanded repeat in amyotrophic lateral sclerosis and frontotemporal lobar degeneration. *Neurobiology of Aging*, 35(10):1–2421, 10 2014.

[201] A. R. Vaz, C. Cunha, C. Gomes, N. Schmucki, M. Barbosa, and D. Brites. Glycoursodeoxycholic Acid Reduces Matrix Metalloproteinase-9 and Caspase-9 Activation in a Cellular Model of Superoxide Dismutase-1 Neurodegeneration. *Molecular Neurobiology*, 51(3):864–877, 6 2015.

[202] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini. Separation of samples into their constituents using gene expression data. Technical report, 2001.

[203] C. Wang, M. A. Telpoukhovskaia, B. A. Bahr, X. Chen, and L. Gan. Endolysosomal dysfunction: a converging mechanism in neurodegenerative diseases. *Current Opinion in Neurobiology*, 48:52–58, 2 2018.

[204] S. M. Wang, S. W. Lim, Y. H. Wang, H. Y. Lin, M. D. Lai, C. Y. Ko, and J. M. Wang. Astrocytic CCAAT/Enhancer-binding protein delta contributes to reactive oxygen species formation in neuroinflammation. *Redox Biology*, 16:104–112, 6 2018.

[205] X. Wang, J. Park, K. Susztak, N. R. Zhang, and M. Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):1–9, 12 2019.

[206] Y. Wang and E. Mandelkow. Tau in physiology and pathology. *Nature Reviews Neuroscience*, 17(1):22–35, 12 2015.

[207] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 4 2004.

[208] J. L. Whitwell, S. D. Weigand, B. F. Boeve, M. L. Senjem, J. L. Gunter, M. Dejesus-Hernandez, N. J. Rutherford, M. Baker, D. S. Knopman, Z. K. Wszolek, J. E. Parisi, D. W. Dickson, R. C. Petersen, R. Rademakers, C. R. Jack, and K. A. Josephs. Neuroimaging signatures of frontotemporal dementia genetics: C9ORF72, tau, progranulin and sporadics. *Brain*, 135(3):794–806, 2012.

[209] R. R. Wick, L. M. Judd, and K. E. Holt. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, 20(1):129, 6 2019.

[210] F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 12 2018.

[211] B. Wolozin and P. Ivanov. Stress granules and neurodegeneration. *Nature Reviews Neuroscience*, 20(11):649–666, 11 2019.

[212] I. O. C. Woollacott and J. D. Rohrer. The clinical spectrum of sporadic and familial forms of frontotemporal dementia. *Journal of Neurochemistry*, 138:6–31, 8 2016.

[213] M. C. Wren, J. Zhao, C. C. Liu, M. E. Murray, Y. Atagi, M. D. Davis, Y. Fu, H. J. Okano, K. Ogaki, A. J. Strongosky, P. Tacik, R. Rademakers, O. A. Ross, D. W. Dickson, Z. K. Wszolek, T. Kanekiyo, and G. Bu. Frontotemporal dementia-associated N279K tau mutant disrupts subcellular vesicle trafficking and induces cellular stress in iPSC-derived neural stem cells. *Molecular Neurodegeneration*, 10(1):46, 9 2015.

[214] Z. Xi, L. Zinman, D. Moreno, J. Schymick, Y. Liang, C. Sato, Y. Zheng, M. Ghani, S. Dib, J. Keith, J. Robertson, and E. Rogaeva. Hypermethylation of the CpG Island Near the G 4 C 2 Repeat in ALS with a C9orf72 Expansion. *Am J Hum Genet*, pages 981–989, 2013.

[215] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metabolism*, 24(4):608–615, 10 2016.

[216] D. Xu, T. Jin, H. Zhu, H. Chen, D. Ofengeim, C. Zou, L. Mifflin, L. Pan, P. Amin, W. Li, B. Shan, M. G. Naito, H. Meng, Y. Li, H. Pan, L. Aron, X. Adiconis, J. Z. Levin, B. A. Yankner, and J. Yuan. TBK1 Suppresses RIPK1-Driven Apoptosis and Inflammation during Development and in Aging. *Cell*, 174(6):1477–1491, 9 2018.

[217] S. Yablonska, V. Ganesan, L. M. Ferrando, J. H. Kim, A. Pyzel, O. V. Baranova, N. K. Khattar, T. M. Larkin, S. V. Baranov, N. Chen, C. E. Strohlein, D. A. Stevens, X. Wang, Y. F. Chang, M. E. Schurdak, D. L. Carlisle, J. S. Minden, and R. M. Friedlander. Mutant huntingtin disrupts mitochondrial proteostasis by interacting with TIM23. *Proceedings of the National Academy of Sciences of the United States of America*, 116(33):16593–16602, 8 2019.

[218] T. Yamaoka, M. Ohba, and T. Ohmori. Molecular-targeted therapies for epidermal growth factor receptor and its resistance mechanismse, 11 2017.

[219] R. Yarwood, J. Hellicar, P. G. Woodman, and M. Lowe. Membrane trafficking in health and disease. *DMM Disease Models and Mechanisms*, 13(4), 4 2020.

[220] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding Neural Networks Through Deep Visualization. *arXiv*, 6 2015.

[221] A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso, J. van Swieten, B. Borroni, D. Galimberti, M. Masellis, M. C. Tartaglia, J. B. Rowe, C. Graff, F. Tagliavini, G. B. Frisoni, R. Laforce, E. Finger, A. de Mendonça, S. Sorbi, J. D. Warren, S. Crutch, N. C. Fox, S. Ourselin, J. M. Schott, J. D. Rohrer, D. C. Alexander, C. Andersson, S. Archetti, A. Arighi, L. Benussi, G. Binetti, S. Black, M. Cosseddu, M. Fallström, C. Ferreira, C. Fenoglio, M. Freedman, G. G. Fumagalli, S. Gazzina, R. Ghidoni, M. Grisoli, V. Jelic, L. Jiskoot, R. Keren, G. Lombardi, C. Maruta, L. Meeter, S. Mead, R. van Minkelen, B. Nacmias, L. Öijerstedt, A. Padovani, J. Panman, M. Pievani, C. Polito, E. Premi, S. Prioni, R. Rademakers, V. Redaelli, E. Rogaeva, G. Rossi, M. Rossor, E. Scarpini, D. Tang-Wai, H. Thonberg, P. Tiraboschi, A. Verdelho, M. W. Weiner, P. Aisen, R. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowki, A. W. Toga, L. Beckett, R. C. Green, A. J. Saykin, J. Morris, L. M. Shaw, Z. Khachaturian, G. Sorensen, L. Kuller, M. Raichle, S. Paul, P. Davies, H. Fillit, F. Hefti, D. Holtzman, M. M. Mesulam, W. Potter, P. Snyder, A. Schwartz, T. Montine, R. G. Thomas, M. Donohue, S. Walter, D. Gessert, T. Sather, G. Jiminez, D. Harvey, M. Bernstein, P. Thompson, N. Schuff, B. Borowski, J. Gunter, M. Senjem, P. Vemuri, D. Jones, K. Kantarci, C. Ward, R. A. Koeppe, N. Foster, E. M. Reiman, K. Chen, C. Mathis, S. Landau, N. J. Cairns, E. Householder, L. Taylor-Reinwald, V. Lee, M. Korecka, M. Figurski, K. Crawford, S. Neu, T. M. Foroud, S. Potkin, L. Shen, K. Faber, S. Kim, K. Nho, L. Thal, N. Buckholtz, M. Albert, R. Frank, J. Hsiao, J. Kaye, J. Quinn, B. Lind, R. Carter, S. Dolen, L. S. Schneider, S. Pawluczyk, M. Beccera, L. Teodoro, B. M. Spann, J. Brewer, H. Vanderswag, A. Fleisher, J. L. Heidebrink, J. L. Lord, S. S. Mason, C. S. Albers, D. Knopman, K. Johnson, R. S. Doody, J. Villanueva-Meyer, M. Chowdhury, S. Rountree, M. Dang, Y. Stern, L. S. Honig, K. L. Bell, B. Ances, M. Carroll, S. Leon, M. A. Mintun, S. Schneider, A. Oliver, D. Marson, R. Griffith, D. Clark, D. Geldmacher, J. Brockington, E. Roberson, H. Grossman, E. Mitsis, L. de Toledo-Morrell, R. C. Shah, R. Duara, D. Varon, M. T. Greig, P. Roberts, M. Albert, C. Onyike, D. D'Agostino, S. Kielb, J. E. Galvin, B. Cerbone, C. A. Michel, H. Rusinek, M. J. de Leon, L. Glodzik, S. De Santi, P. M. Doraiswamy, J. R. Petrella, T. Z. Wong, S. E. Arnold, J. H. Karlawish, D. Wolk, C. D. Smith, G. Jicha, P. Hardy, P. Sinha, E. Oates, G. Conrad, O. L. Lopez, M. A. Oakley, D. M. Simpson, A. P. Porsteinsson, B. S. Goldstein, K. Martin, K. M. Makino, M. S. Ismail, C. Brand, R. A. Mulnard, G. Thai, C. McAdams-Ortiz, K. Womack, D. Mathews, M. Quiceno, R. Diaz-Arrastia, R. King, M. Weiner, K. Martin-Cook, M. DeVous, A. I. Levey, J. J. Lah, J. S. Cellar, J. M. Burns, H. S. Anderson, R. H. Swerdlow, L. Apostolova, K. Tingus, E. Woo, D. H. Silverman, P. H. Lu, G. Bartzokis, N. R. Graff-Radford, F. Parfitt, T. Kendall,

H. Johnson, M. R. Farlow, A. M. Hake, B. R. Matthews, S. Herring, C. Hunt, C. H. van Dyck, R. E. Carson, M. G. MacAvoy, H. Chertkow, H. Bergman, C. Hosein, B. Stefanovic, C. Caldwell, G. Y. R. Hsiung, H. Feldman, B. Mudge, M. Assaly, A. Kertesz, J. Rogers, C. Bernick, D. Munic, D. Kerwin, M. M. Mesulam, K. Lipowski, C. K. Wu, N. Johnson, C. Sadowsky, W. Martinez, T. Villena, R. S. Turner, K. Johnson, B. Reynolds, R. A. Sperling, K. A. Johnson, G. Marshall, M. Frey, B. Lane, A. Rosen, J. Tinklenberg, M. N. Sabbagh, C. M. Belden, S. A. Jacobson, S. A. Sirrel, N. Kowall, R. Killiany, A. E. Budson, A. Norbash, P. L. Johnson, J. Allard, A. Lerner, P. Ogrocki, L. Hudson, E. Fletcher, O. Carmichael, J. Olichney, C. DeCarli, S. Kittur, M. Borrie, T. Y. Lee, R. Bartha, S. Johnson, S. Asthana, C. M. Carlsson, S. G. Potkin, A. Preda, D. Nguyen, P. Tariot, S. Reeder, V. Bates, H. Capote, M. Rainka, D. W. Scharre, M. Kataki, A. Adeli, E. A. Zimmerman, D. Celmins, A. D. Brown, G. D. Pearlson, K. Blank, K. Anderson, R. B. Santulli, T. J. Kitzmiller, E. S. Schwartz, K. M. Sink, J. D. Williamson, P. Garg, F. Watkins, B. R. Ott, H. Querfurth, G. Tremont, S. Salloway, P. Malloy, S. Correia, H. J. Rosen, B. L. Miller, J. Mintzer, K. Spicer, D. Bachman, S. Pasternak, I. Rachinsky, D. Drost, N. Pomara, R. Hernando, A. Sarrael, S. K. Schultz, L. L. Ponto, H. Shim, K. E. Smith, N. Relkin, G. Chaing, L. Raudin, A. Smith, K. Fargher, B. A. Raj, T. Neylan, J. Grafman, M. Davis, R. Morrison, J. Hayes, S. Finley, K. Friedl, D. Fleischman, K. Arfanakis, O. James, D. Massoglia, J. J. Fruehling, S. Harding, E. R. Peskind, E. C. Petrie, G. Li, J. A. Yesavage, J. L. Taylor, and A. J. Furst. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature Communications*, 9(1):1–16, 12 2018.

[222] J. J. Young, M. Lavakumar, D. Tampi, S. Balachandran, and R. R. Tampi. Frontotemporal dementia: latest evidence and clinical implications. *Therapeutic advances in psychopharmacology*, 8(1):33–48, 1 2018.

[223] G. Yu, L. G. Wang, and Q. Y. He. ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383, 7 2015.

[224] J. Yuan, P. Amin, and D. Ofengeim. Necroptosis and RIPK1-mediated neuroinflammation in CNS diseases. *Nature Reviews Neuroscience*, 20(1):19–33, 1 2019.

[225] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (New York, N.Y.)*, 347(6226):1138–42, 3 2015.

[226] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyan-

skaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 8 2018.

[227] J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 9 2015.

[228] Q. Zhu, J. Jiang, T. F. Gendron, M. McAlonis-Downes, L. Jiang, A. Taylor, S. Diaz Garcia, S. Ghosh Dastidar, M. J. Rodriguez, P. King, Y. Zhang, A. R. La Spada, H. Xu, L. Petrucelli, J. Ravits, S. Da Cruz, C. Lagier-Tourenne, and D. W. Cleveland. Reduced C9ORF72 function exacerbates gain of toxicity from ALS/FTD-causing repeat expansion in C9orf72. *Nature Neuroscience*, 23(5):615–624, 5 2020.

[229] M. T. Zimmermann, A. L. Oberg, D. E. Grill, I. G. Ovsyannikova, I. H. Haralambieva, R. B. Kennedy, and G. A. Poland. System-Wide Associations between DNA-Methylation, Gene Expression, and Humoral Immune Response to Influenza Vaccination. *PLOS ONE*, 11(3):e0152034, 3 2016.

[230] M. Zou, F. Wang, R. Gao, J. Wu, Y. Ou, X. Chen, T. Wang, X. Zhou, W. Zhu, P. Li, L. W. Qi, T. Jiang, W. Wang, C. Li, J. Chen, Q. He, and Y. Chen. Autophagy inhibition of hsa-miR-19a-3p/19b-3p by targeting TGF-β R II during TGF-β1-induced fibrogenesis in human cardiac fibroblasts. *Scientific Reports*, 6(1):1–15, 4 2016.