

***Development of Gene Set Enrichment and Imputation Methods  
for Transcriptomics and Proteomics:  
Application in the Study of Neurofibrillary Tangle-bearing  
Neurons in Alzheimer's Disease***

**Emir Turkes**

**PhD Thesis**

**University College London**

# Table of Contents

Abstract.....	5
Impact Statement.....	6
Aims.....	7
Hypothesis.....	8
Declaration Page.....	9
Research Declaration Form.....	10
1. Introduction.....	13
1.1 Transcriptomics.....	14
1.1.1 Overview.....	14
1.1.2 Polymerase Chain Reaction.....	15
1.1.3 Microarray Technology.....	15
1.1.4 Sanger Sequencing.....	17
1.1.5 Next-generation Sequencing.....	19
1.1.6 Single-cell Sequencing.....	21
1.1.7 Spatial and Long-read Sequencing.....	23
1.2 Proteomics.....	24
1.2.1 Overview.....	24
1.2.2 Mass Spectrometry.....	25
1.2.3 Top-down vs. Bottom-up.....	27
1.2.4 Labeled vs. Label-free Quantification.....	27
1.2.5 Data-dependent vs. Independent Acquisition.....	28
1.2.6 Targeted Proteomics.....	30
1.2.7 Spatial Proteomics.....	31
1.3 Alzheimer's Disease.....	32
1.3.1 Overview.....	32
1.3.2 Amyloid Cascade Hypothesis.....	33
1.3.3 Amyloid-beta Structure and Function.....	34
1.3.4 Amyloid-Tau Interaction.....	35
1.3.5 Tau Structure and Function.....	36
1.3.6 Pathological Tau Accumulation.....	37
1.3.7 Tau Post-translational Modifications.....	38
1.3.8 Amyloid and Tau Staging.....	39
1.3.9 Cell Death and Atrophy.....	41
1.3.10 Selective Vulnerability.....	42
1.3.11 Cell and Non-cell Autonomous Factors.....	44
1.3.12 Circuit Dynamics and Connectivity.....	44
1.3.13 Cellular Disease Response.....	45
1.3.14 Studies of Neurofibrillary Tangle-bearing Neurons.....	46
2. Methods.....	51
2.1 FACS-sorted Single-soma RNA Sequencing.....	51
2.2 Laser-capture Microdissection Mass Spectrometry.....	56
2.3 Analysis Organisation and Reproducibility.....	60
2.3.1 Data Hosting and Version Control.....	60
2.3.2 Docker and Singularity.....	60
2.3.3 R Markdown.....	61
2.3.4 Gene / Protein Identifiers.....	63
2.4 Single-cell RNA Sequencing Preprocessing.....	64
2.4.1 Data Cleaning.....	64

2.4.2 Quality Control.....	64
2.4.3 Dimensionality Reduction and Clustering.....	71
2.4.4 Sample Merging and Cell-type Annotation.....	74
2.5 Mass Spectrometry Preprocessing.....	81
2.5.1 Data Cleaning.....	81
2.5.2 Quality Control.....	81
2.5.3 Sample Removal.....	84
2.5.4 Normalisation.....	86
3. Development of ImputeFinder Imputation Method.....	93
3.1 Definition and Description of Missing Values.....	93
3.2 Prior Art: Imputation.....	95
3.3 Benchmarking of Existing Imputation Methods.....	96
3.4 The Lack of Hybrid Imputation Approaches.....	97
3.5 Methodology of ImputeFinder.....	99
3.6 Construction of Simulated Proteomics Dataset.....	104
3.7 Benchmarking in Simulated and Real Dataset.....	107
3.8 Sensitivity Analysis.....	109
3.9 Advantages of ImputeFinder.....	113
3.10 Disadvantages and Limitations.....	113
3.11 Implementation Details.....	115
4. Development of GeneFunnel Gene Set Enrichment Method.....	117
4.1 Definition and Description of Gene Set Enrichment.....	117
4.2 Prior Art: Functional Class Scoring.....	119
4.3 Methodology of GeneFunnel.....	123
4.4 Mathematical Proof of Non-negative Scores.....	126
4.5 Exploration of GeneFunnel Properties.....	130
4.6 Exploring GeneFunnel Alongside Other Functional Class Scoring Methods.....	134
4.7 Benchmarking of GeneFunnel Against Other Methods in Synthetic Data.....	138
4.8 Benchmarking of GeneFunnel Against Other Methods in Real Data.....	144
4.9 Benchmarking of Computational Efficiency of GeneFunnel Against Other Methods.....	155
4.10 Advantages of GeneFunnel.....	157
4.11 Disadvantages and Limitations.....	159
4.12 Implementation Details.....	160
5. Development of Downstream Analysis Pipeline.....	163
5.1 Integrated Transcriptomic/Proteomic Differential Expression Analysis.....	163
5.2 Development of Network Analysis and Hub Selection Approach.....	168
5.3 Development of Web Viewers.....	174
6. Results.....	177
6.1 Overview of Transcriptomic and Proteomic Analysis of Tangle-bearing Neurons.....	177
6.2 Top Differentially Expressed Proteins and Genes and Associated Pathways.....	179
6.3 Network Hub Analysis.....	183
6.4 Pathway Analysis of Hubs.....	191
6.4.1 NEFM.....	194
6.4.2 APP.....	199
6.4.3 SQSTM1.....	204
6.4.4 HSP90AA1.....	208
6.4.5 YWHAE.....	212
6.4.6 WASF1.....	216
6.4.7 CNTNAP1.....	219
6.4.8 GOT2.....	221

7. Discussion.....	224
7.1 ImputeFinder and GeneFunnel Resolve Obstacles in Real-world Analysis.....	224
7.2 Evidence of Successful Isolation of NFT-bearing Neurons on the Gene and Protein-level..	226
7.3 Similarly Designed Studies Support Findings in This Analysis.....	227
7.4 Caveats and Interpretational Considerations.....	229
7.5 Emergent Themes of Tangle-bearing Neuron Pathophysiology.....	230
7.5.1 Co-aggregation of Neurofilaments and Microtubule Destabilisation.....	231
7.5.2 Potentially Protective Role of the Non-Amyloigenic Pathway.....	233
7.5.3 p62 Accumulation and Dysregulation of Autophagy.....	236
7.5.4 Chaperone Co-factors and the Dual Roles of the HSP90 Complex.....	237
7.5.5 Sequestration of Tau Dephosphorylating Phosphatases.....	239
7.5.6 Recruitment of Neurotrophic Factors Through the WRC.....	241
7.5.7 Glial Involvement Through Paranodal Junctions.....	242
7.5.8 Compensation of Impaired Glutamate Recycling.....	244
8. Conclusion.....	247
8.1 Overall Limitations and Future Directions.....	248
9. References.....	250
10. Appendix.....	308

## Abstract

Transcriptomics and proteomics are high-throughput methods that assay gene expression and protein abundance in a biological sample at a given point in time. These datasets feature high dimensionality and technical noise which are routinely addressed using various computational methods. In particular, gene set enrichment is commonly used for measuring how enriched expression is for functionally defined subsets of the gene/protein profile, whereas imputation handles the replacement of missing data with predicted values.

In this thesis, I review the prior art of these methods and identify pitfalls that warrant investigation. I then introduce novel methods, GeneFunnel and ImputeFinder, with freely available software implementations (<https://github.com/eturkes/genefunnel> and <https://github.com/eturkes/imputefinder>) that attempt to address these pitfalls without imposing performance bottlenecks, stringent assumptions, or unintuitive reasoning. Although ImputeFinder did not have a comparable equivalent, GeneFunnel was benchmarked against leading methods in both synthetic and real-world data, showing superior analytic and computational performance across all metrics. An interactive web viewer of these benchmarks is available at <https://data.duff-lab.org/app/genefunnel-benchmarks-viewer>.

I deploy the methods in a real-world context by developing a pipeline for characterising neurofibrillary tangle-bearing neurons in Alzheimer's Disease. A previously available dataset of human post-mortem tissue, where tangle-bearing neurons were isolated from non-tangle-bearing neurons and subject to transcriptomic profiling, was reanalysed alongside a similarly designed in-house dataset that profiled proteins. The integrated analysis is complemented by interactive network visualisations and a web-based viewer, allowing in-depth exploration of the results at <https://data.duff-lab.org/app/tangle-bearing-neurons-viewer>.

The analysis focused on uncovering major drivers of biological pathways upregulated in tangle-bearing neurons in both the transcriptomic and proteomic datasets, identifying the pathway hubs NEFM, APP, SQSTM1, HSP90AA1, YWHAE, WASF1, CNTNAP1, and GOT2. Using informatics and literature review, I investigate the contribution of these hubs to distinct functional domains, laying the groundwork for a unified model of the pathophysiology of tangle-bearing neurons in Alzheimer's Disease.

## Impact Statement

The present work evaluates existing methods in the analysis of transcriptomics and proteomics data, highlighting problematic areas that are widespread and offering solutions. I focus on gene set enrichment and imputation methods in particular. Regarding gene set enrichment, I examine and address six primary issues. 1) the handling of missing and lowly expressed features. 2) consideration of dependencies between samples, between features, and interactions across the two. 3) retention of statistical properties of the input data. 4) consideration of complexity and assumptions. 5) compatibility with downstream handling of data and interpretation. 6) speed and scalability. Regarding imputation, I examine and address two primary issues. 1) detection and handling of mixed types of missing values within a dataset (e.g. missing at random vs. missing not at random). 2) incorporation of comparison group information to retain features with missing values of probable biological origin. For each of these methods, I provide software solutions for the named issues, GeneFunnel and ImputeFinder respectively.

In addition to benchmarking performed in a controlled manner across real and synthetic data, I apply these methods to datasets generated to investigate the molecular changes that occur in the context of neurons harbouring neurofibrillary tangles, one of two hallmark pathological features in Alzheimer's Disease. One dataset, generated in-house, utilises laser-capture microdissection (LCM) to isolate tangle-bearing neurons from non-tangle-bearing neurons within patient donors for proteomics profiling. Another dataset, previously available, utilises fluorescence-activated cell sorting (FACS) to sort tangle-bearing and non-tangle-neurons within patient donors for single-cell transcriptomics profiling. Application of the aforementioned gene set enrichment and imputation methods were additionally supplemented with general exploratory work, bespoke network analyses, and web development, facilitating deeper investigation and easier exploration.

Use of the newly developed methods were effective in tackling technical issues inherent to datasets of these nature, and improved biological interpretation of the processed data. These complementary datasets, covering both genes and proteins, are a highly valuable resource for understanding the molecular changes that occur in the neurofibrillary tangle-bearing neurons that define Alzheimer's Disease. I demonstrate recapitulation of known and hypothesised mechanisms underlying this pathological feature and prioritise a set of eight hub gene/proteins that may concisely, but comprehensively, represent the major drivers of distinct disease processes. In order to provide longevity and accession of these results, a web viewer is provided that allows generation of custom figures and searching of statistical information.

## Aims

This project aims to fully utilise highly dimensional data in targeted experimental contexts to better understand molecular processes in Alzheimer's Disease, specifically those that take place in neurons harbouring neurofibrillary tangles. In order to do so, research focus was directed towards computational methods development. In initial analyses of the datasets, a variety of issues were encountered that introduced obstacles in forming reliable conclusions. One set of issues concerned quality control, specifically the handling of missing values in proteomics data when using imputation. Due to the nature of experiments that isolate single-cells with specific pathology, a higher degree of technical noise was observed in comparison to conventional omics datasets, and attempts to address this using existing methods proved inadequate. The second set of issues relate to downstream analysis of the processed data when using gene set enrichment. I found that existing methods introduced biases and assumptions that either directed attention towards a narrow subset of changes while neglecting others, or produced results that were difficult to reason with altogether.

By creating generalised open-source software solutions in R and C++ to address these issues, I intend to not only advance understanding of these datasets and the Alzheimer's Disease field, but increase the availability of tools in the informatics space. The design decisions of these software also help inform and highlight prevalent issues in data processing, some of which I argue to be overlooked in most analyses. I demonstrate the utility of these tools, both in the real-world context they were developed for, and in synthetic datasets covering a wide range of hypothetical scenarios. These experiments are supplemented with bespoke network analyses and web viewers, allowing for easy exploration of output through custom figure generation and data search.

Because neurofibrillary tangles arise in a cell-type specific manner, single-cell methods are ideal for comparing tangle-bearing from non-tangle-bearing neurons. The technical challenge of these particular experiments mean that the availability of high-quality transcriptomics and proteomics data is highly limited, and therefore I aimed to perform a comprehensive characterisation of this crucial but insufficiently explored comparison. An additional aim was to replicate and expand upon existing datasets, prompting an in-house laser-capture microdissection experiment where tangle-bearing and non-tangle-bearing neurons within patient donors were isolated for proteomics profiling.

This thesis covers the following key questions:

- Applied to the transcriptomic and proteomic datasets at hand that compare neurofibrillary tangle-bearing and non-tangle-bearing neurons, how effective are existing computational pipelines for processing the data?
- More generally, what deficiencies and unaddressed problems can be identified in the methods comprising these pipelines?

- What solutions, if any, can be envisioned to address these issues and what are the limitations and drawbacks of such solutions?
- When using the most suitable methods available, including those novel to this work, what are the molecular changes taking place between tangle-bearing and non-tangle-bearing neurons on the transcriptomic and proteomic level?
- How do these molecular changes compare with existing literature and what directions do they suggest for future research and validation?

## Hypothesis

- Existing computational methods for transcriptomics and proteomics have drawbacks that make their application inadequate for the datasets of interest that compare neurofibrillary tangle-bearing neurons against non-tangle-bearing neurons in Alzheimer's Disease post-mortem tissue.
- Tangle-bearing neurons, compared to non-tangle-bearing neurons, exhibit significant changes on the transcriptomic and proteomic level across a range of biological pathways, some of which have been described by previous literature, others of which are novel or relatively unexplored given the novelty of these datasets and the application of newly developed computational methods.

## Declaration Page

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

## Research Declaration Form

# UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

*Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:*  
*have been uploaded to a preprint server*  
*are in submission to a peer-reviewed publication*  
*have been published in a peer-reviewed publication, e.g. journal, textbook.*

*This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.*

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

#### **What is the title of the manuscript?**

Tau filaments are tethered within brain extracellular vesicles in Alzheimer's disease

#### **Please include a link to or doi for the work**

10.1038/s41593-024-01801-5

#### **Where was the work published?**

Nature Neuroscience

#### **Who published the work? (e.g. OUP)**

Nature Publishing Group

#### **When was the work published?**

01/2025

#### **List the manuscript's authors in the order they appear on the publication**

Stephanie L. Fowler, Tiana S. Behr, Emir Turkes, Darragh P. O'Brien, Paula Maglio Cauhy, Isadora Rawlinson, Marisa Edmonds, Martha S. Foiani, Ari Schaler, Gerard Crowley, Sumi Bez, Elena Ficulle, Eliona Tsefou, Roman Fischer, Beth Geary, Pallavi Gaur, Chelsea Miller, Pasquale D'Acunzo, Efrat Levy, Karen E. Duff, Benjamin Ryskeldi-Falcon

#### **Was the work peer reviewed?**

Yes

#### **Have you retained the copyright?**

Yes

#### **Was an earlier form of the manuscript uploaded to a preprint server? (e.g. medRxiv). If 'Yes', please give a link or doi)**

<https://www.biorxiv.org/content/10.1101/2023.04.30.537820v1>

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:



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

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

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

Click or tap here to enter text.

**b) Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if ‘Yes’, please give a link or doi)

Click or tap here to enter text.

**c) Where is the work intended to be published?** (e.g. journal names)

Click or tap here to enter text.

**d) List the manuscript’s authors in the intended authorship order**

Click or tap here to enter text.

**e) Stage of publication** (e.g. in submission)

Click or tap here to enter text.

**3. For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

S.L.F. performed EV isolations and western blotting. S.L.F., T.S.B. and I.R. performed tau filament extraction. T.S.B. and I.R. performed dot blots. T.S.B. performed cryo-ET and subtomogram averaging. T.S.B. and B.R.-F. performed single-particle cryo-EM. S.L.F. and E. Turkes performed proteomic and NTA bioinformatic analyses. D.P.O. performed liquid chromatography–tandem mass spectrometry. S.L.F., P.M.C. and M.E. performed NTA. S.L.F. performed HEK biosensor cell seeding. A.S. and G.C. performed mouse stereotaxic injections. M.S.F. performed mouse brain immunostaining and analysis. S.B., E.F. and E. Tsefou assisted in EV isolation and analysis. D.P.O., R.F., B.G. and P.G. consulted on liquid chromatography–tandem mass spectrometry and data analysis. C.M. performed initial EV isolations from human tissue. P.D.A. and E.L. consulted on EVs. E.L., K.E.D. and B.R.-F. acquired funding. K.E.D. and B.R.-F. supervised the study. S.L.F., T.S.B., K.E.D. and B.R.-F. wrote the manuscript.

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

3.1

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

*Candidate*



Emir Turkes

*Date:*

20/03/2025

*Supervisor/ Senior Author (where appropriate)*



Karen E. Duff

*Date*

20/03/2025

## 1. Introduction

Transcriptomics and proteomics describe high-throughput methods that measure gene expression and protein abundance, respectively. The suffix specifies their categorisation under the greater umbrella of “omics” approaches, which broadly encompass any form of high-throughput molecular information gathered from a biological sample. For example, the original omics method, genomics, focuses on examination and comparison of precise nucleotide sequences from the DNA of samples of interest. A method like phosphoproteomics however, measures not protein abundance per se, but quantifies a reaction known as phosphorylation that takes place at various locations along a protein’s structure. As these examples show, omics covers a wide variety of assays, some quite basic to biology, others more nuanced. These assays play a crucial role in advancing our understanding of diseases like Alzheimer’s Disease (AD) and the development of more effective treatments.

While omics approaches in AD can provide detailed molecular profiles, it is important to consider the context. A long-standing question is whether tau tangles act as a primary driver of neuronal loss or arise as a downstream consequence of earlier events. Biomarker studies suggest that A $\beta$  deposition, synaptic dysfunction, and other molecular changes often precede detectable tau aggregation. On the other hand, the existence of primary tauopathies, such as progressive supranuclear palsy and corticobasal degeneration, shows that tau aggregation alone can initiate neurodegeneration. Moreover, in Alzheimer’s disease, subtle changes in tau may occur before overt tangle formation, with recent evidence suggesting these changes taking place before amyloid deposition, raising the possibility that tau contributes to disease initiation as well as progression. Ultimately however, there is insufficient evidence to determine whether tau tangles are a cause or a consequence of AD. For the purposes of this thesis, tau aggregation is assumed to occur downstream of amyloid pathology in most cases, though is likely the more significant contributor to overt neurodegeneration. This is likely the more common interpretation in current Alzheimer’s research, although it is not universally held.

For the purposes of methods development, the analysis was carried out in AD tissue, which contains both 3-repeat (3R) and 4-repeat (4R) tau isoforms. Simpler tauopathies such as Pick’s disease (3R) or progressive supranuclear palsy (4R) are single-isoform and primary tauopathies, occurring in the absence of A $\beta$  deposition and other co-pathologies. While these models offer greater biochemical uniformity, they do not reflect the mixed tau isoform composition or the multi-pathology environment of AD. The aim of this work was to develop and apply methods in the specific context of neurofibrillary tangle-bearing neurons as they occur in AD, where tau pathology coexists and interacts with other factors such as A $\beta$ . Using this model ensured that the computational framework was optimised for the unique characteristics of Alzheimer’s pathology, which would not be captured in single-isoform primary tauopathies.

Understanding tau pathology in AD also requires consideration of which neuronal populations are most affected and why. Selective vulnerability refers to the tendency of specific neuronal populations to be affected earlier or more severely by pathological processes than others within the same brain region. This may be determined by intrinsic factors such as molecular properties of certain populations or their connectivity, as well as extrinsic influences from the surrounding environment. In the context of AD, identifying the molecular features associated with selective vulnerability could help clarify why certain neurons develop tau pathology and degenerate while others remain relatively preserved. However, in post-mortem studies there is a potential for survivor bias, whereby the cells available for analysis may represent those that have resisted pathology for longer, and thus may reflect resilience rather than true vulnerability. Interpreting molecular differences between cell populations therefore requires caution, with consideration of whether observed features are drivers of degeneration or markers of survival.

## 1.1 Transcriptomics

### 1.1.1 Overview

Gene expression can be defined as the initiation of the sequence of steps that ultimately result in a functional gene product, typically, but not exclusively limited to, proteins (Buccitelli & Selbach, 2020). The initiating step itself is typically defined to be transcription which entails the production of an mRNA (messenger RNA) product that consists of a nucleotide sequence complementary to the DNA sequence of the gene being transcribed. From this mRNA, the next step, at a high-level, is called translation, where the mRNA sequence is translated into instructions that allow for the assembly of a corresponding protein product. Although a number of sometimes highly complex steps can take place surrounding these events, this two-step process defines what is known as the central dogma of molecular biology, where genetic information originates in DNA, is transcribed to mRNA, and is then translated into proteins. As proteins are the main functional units of the cells that make up an organism, measures of its abundance and the expression of genes that typically result in their production are among the most important functional readouts of a sample in molecular biology.

**Figure 1.**

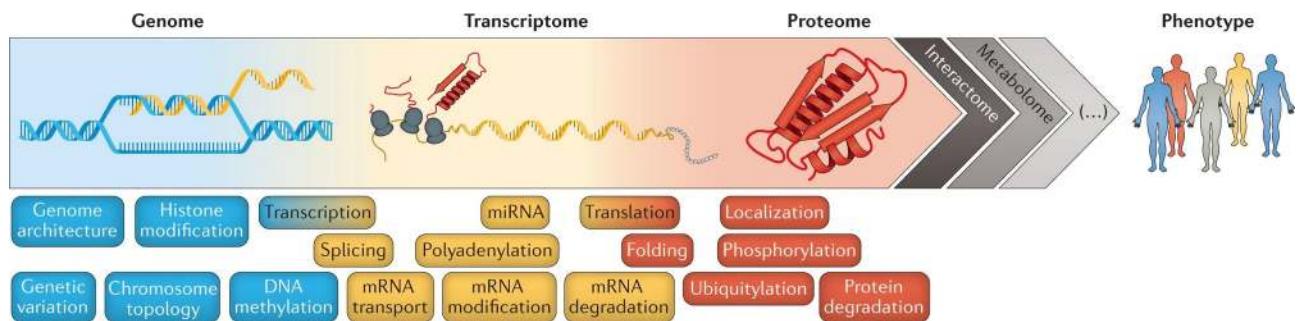


Figure 1: Schematic of the passage of genetic information from the genome to proteome. Genetic information is encoded in DNA and goes through the high-level processes of transcription and translation, as well as other less general processes, in the production of a protein, the main functional unit of a cell. The interaction of these processes across the genome and with external factors such as the environment is responsible for an organism's phenotype or observable characteristics. Figure reproduced from (Buccitelli & Selbach, 2020).

### 1.1.2 Polymerase Chain Reaction

As an omics approach, transcriptomics methods are high-throughput assays that attempt to quantify the abundance of mRNA “transcripts” that result from gene expression. The threshold for high-throughput is not clearly defined, though it is commonly agreed that such a method should at least aim for an unbiased readout of the genome. Quantitative measures of gene expression have existed since the 1990s with the development of RT-qPCR (real-time quantitative polymerase chain reaction) and quickly became ubiquitous (Bustin, 2000; Heid et al., 1996). However, RT-qPCR and derivative methods rely on the use of fluorescent probes for readout, posing significant spectroscopic challenges for multiplexing. Therefore, a single experiment has the capacity to cover distinct genes on the order of dozens, a minuscule fraction of the 20,000 or so genes in the human genome, not counting the tens of thousands of additional gene products outside the protein-coding genome (Venter et al., 2001). Such methods may not qualify as an omics approach due to the necessity of selection of genes of interest. It should be noted however, that in recent years, advances in gene measurement *in-situ*, that is, directly taking place on a tissue of interest, have begun to make possible panels that now cover thousands of distinct genes, with the eventual goal of covering the genome with fluorescent tag approaches (Janesick et al., 2023; R. Ke et al., 2013). These methods operate on similar principles as RT-qPCR albeit are more specifically related to FISH (fluorescence *in-situ* hybridisation), the central difference being that signal is read directly from probes that remain bound to their target *in-situ*, rather than from dissociated amplification products of the original mRNA.

### 1.1.3 Microarray Technology

In between the development gap of multiplexing RT-qPCR and FISH approaches, RNA microarrays emerged as a significant disruptive technology, and potentially satisfy conditions to be called the earliest transcriptomics method. Demonstrated in 1995 covering 45 unique transcripts (Schena et al., 1995), by 2002 the technology had already been well commercialised with Affymetrix’s seminal GeneChip U133 product covering 39,000 transcripts derived from the 2001 draft of the human genome (Constans, 2002). The years that followed saw an explosion in usage and development of microarrays across research, clinical, and commercial sectors (see Figure 2) (Lenoir & Giannella, 2006). Coupled with the explosion of throughput offered by microarrays, this time period also saw the emergence of software designed specifically for the idiosyncrasies of microarray data (Dudoit et al., 2003).

**Figure 2.**

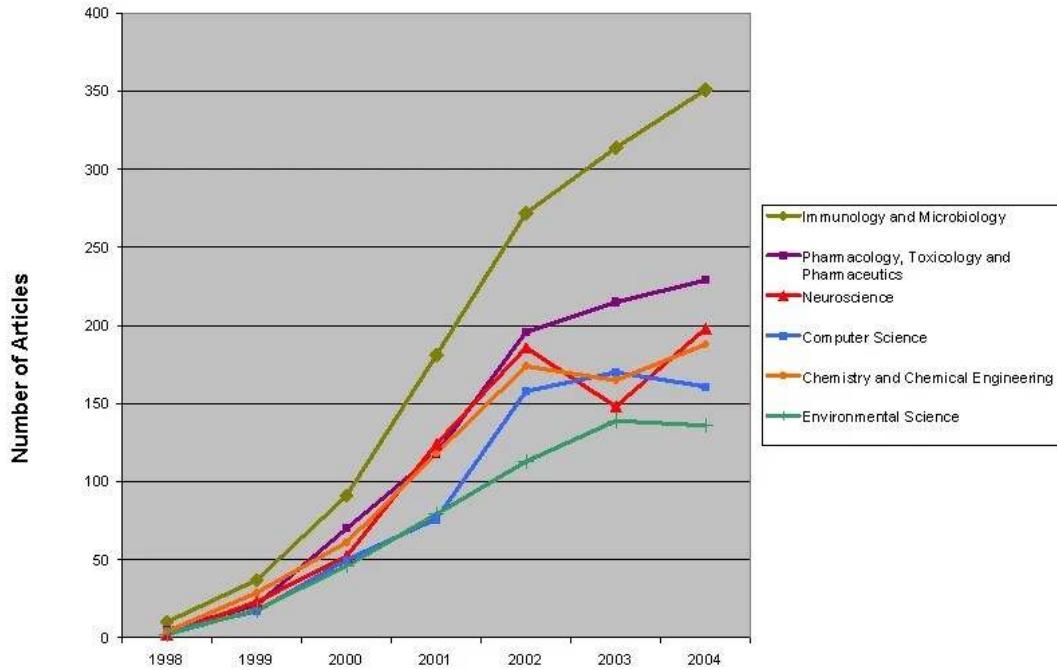


Figure 2: The number of published microarray articles by subject from 1998 to 2004. The increase in publications following the commercialisation of microarray technology sometime around the millennium is evident. Figure reproduced from (Lenoir & Giannella, 2006).

Microarrays fall within the category of multiplex lab-on-chip systems, in other words, miniature assays capable of measuring a wide variety of biological parameters simultaneously (Pham, 2018). Microarrays in particular are 2D arrays usually printed on glass or silicone, originally for the purpose of detecting mutations in a single or multiple genes, but was quickly expanded to include detection of transcripts, proteins, metabolites, and many others. The core operating principle is hybridisation, similarly to FISH. On each chip are many so-called spots, up to tens of thousands by the early 2000s and in recent years hundreds of thousands (Wöhrle et al., 2020). Each spot contains a unique set of fluorescently-tagged probes that hybridise selectively to the target of interest. Though the fluorescent dyes themselves are not unique between spots, the issue of spectral overlap hampering multiplexing in FISH-like methods is solved due to the physical distances between spots, with the trade-off being loss of spatial information. It is also possible to incorporate multiple dyes into each spot to allow comparisons between multiple samples on a single chip, with two being the convention and up to four having been demonstrated

(Staal et al., 2005). Furthermore, in the case of detecting RNA, a reverse transcription step typically takes place as in PCR, where the probes are in fact primers for the amplification of RNA into fluorescently-tagged cDNA (complementary DNA). These desirable properties, as well as relatively low cost of usage after only a few years of development (\$200 to \$1,200 per chip commercially, and less than \$150 from university core facilities (Rubenstein, 2002)), led to the quick dominance of microarrays as the transcriptomics method of the 2000s.

**Figure 3.**

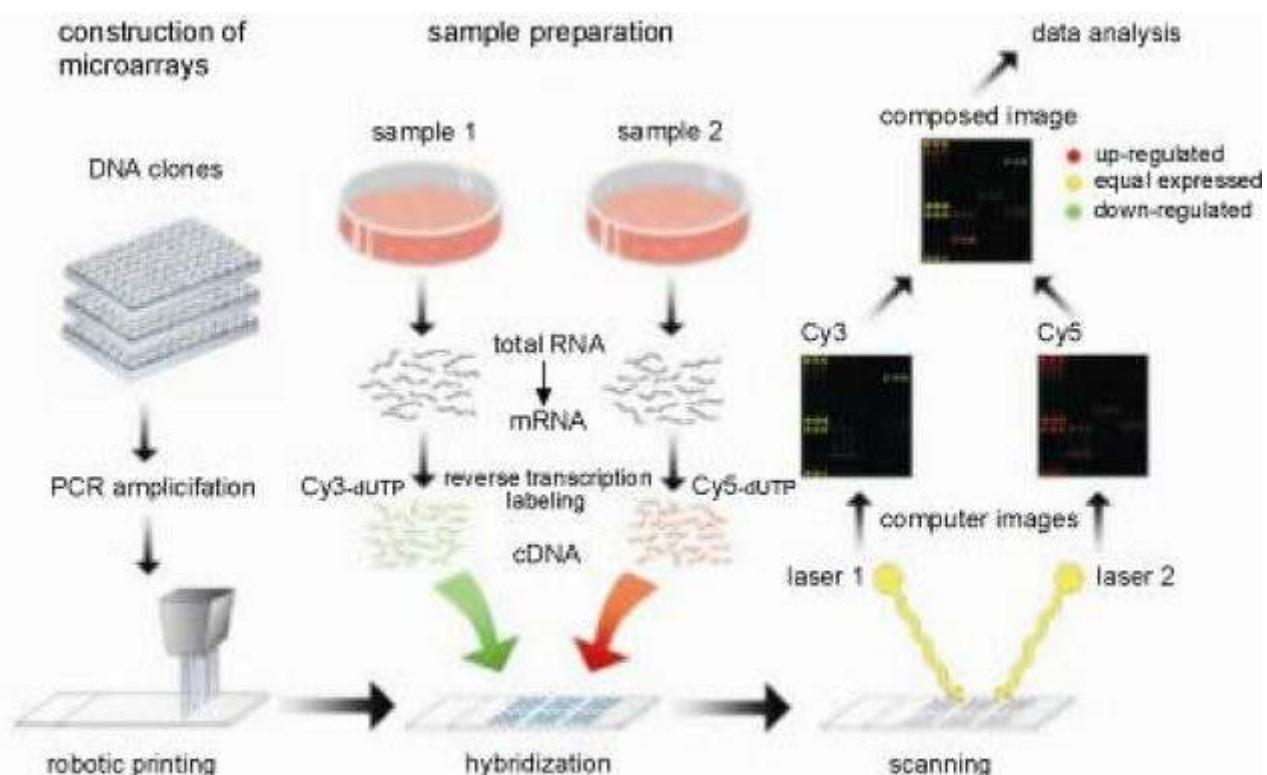


Figure 3: Schematic overview of the use of microarray for measuring gene expression. Shown are the general steps for construction of microarrays, sample preparation including hybridisation and conversion of RNA to fluorescently-tagged cDNA, and use of a laser-based scanner to detect fluorescence intensity in each dye to quantify differences between samples for each transcript included in the array. Figure reproduced from (Majtán et al., 2004).

#### 1.1.4 Sanger Sequencing

As the new millennium neared the end of its first decade, sequencing technology, where DNA and RNA sequences are resolved per base pair rather than the hybridisation-based approach of binding of complementary sequences, began to see massive advancements (Shendure & Ji, 2008). The foundations of sequencing is said to have been laid as early as 1977, with the introduction of Sanger sequencing by Frederick Sanger (Sanger et al., 1977). Until the late 2000s, this remained the gold standard for sequencing genomes and was famously used to create the first draft of the complete human genome in the year

2000 by the Human Genome Project (International Human Genome Sequencing Consortium et al., 2001). While crucial to biological research, the 13 year project, costing \$2.7 billion dollars, highlighted painful inefficiencies in the cost and throughput of Sanger sequencing for projects of this scale (Lewin et al., 2018).

One of the major limitations of Sanger sequencing is the fact that the process can only be carried out on sequences of between 600 and 1,000 base pairs in length, thus requiring fragmentation of the DNA of interest as the initiating step (Kircher & Kelso, 2010; Shendure & Ji, 2008). These fragments are then amplified in one of two ways. So-called “shotgun sequencing”, is useful when sequencing *de novo*, in other words, when sequencing is taking place in an organism that has never been sequenced before. It involves random fragmentation of the DNA followed by incorporation or cloning into the DNA of an actively reproducing bacterial species, typically *E. coli*. Another approach is to use PCR, where a primer flanking the target fragment combined with DNA polymerase allows rapid synthesis of cDNA complementary to the fragment; this requires some *a priori* knowledge of the target in order to design the primer.

Once the random fragments are amplified, a reaction takes place that labels the last nucleotide on the 3' end with one of four fluorophores, indicating whether the nucleotide is the chemical base adenine, cytosine, guanine, or thymine (A, C, G, or T). Finally, the fragments are sorted by molecular weight, a proxy for length, contemporarily using capillary electrophoresis. By scanning fluorophores of the sorted fragments, and repeating this process many times, it is then possible to sequence a genome. Essential to scaling Sanger sequencing, capillary electrophoresis remains a significant bottleneck. Though there exist systems for the processing of up to 384 sequences in parallel, such systems are rare (Kircher & Kelso, 2010). The more conventional 96-capillary systems are capable of sequencing about 6 million base pairs of DNA per day; at this rate the 3 billion base pairs of the human genome would take around 500 days.

**Figure 4.**

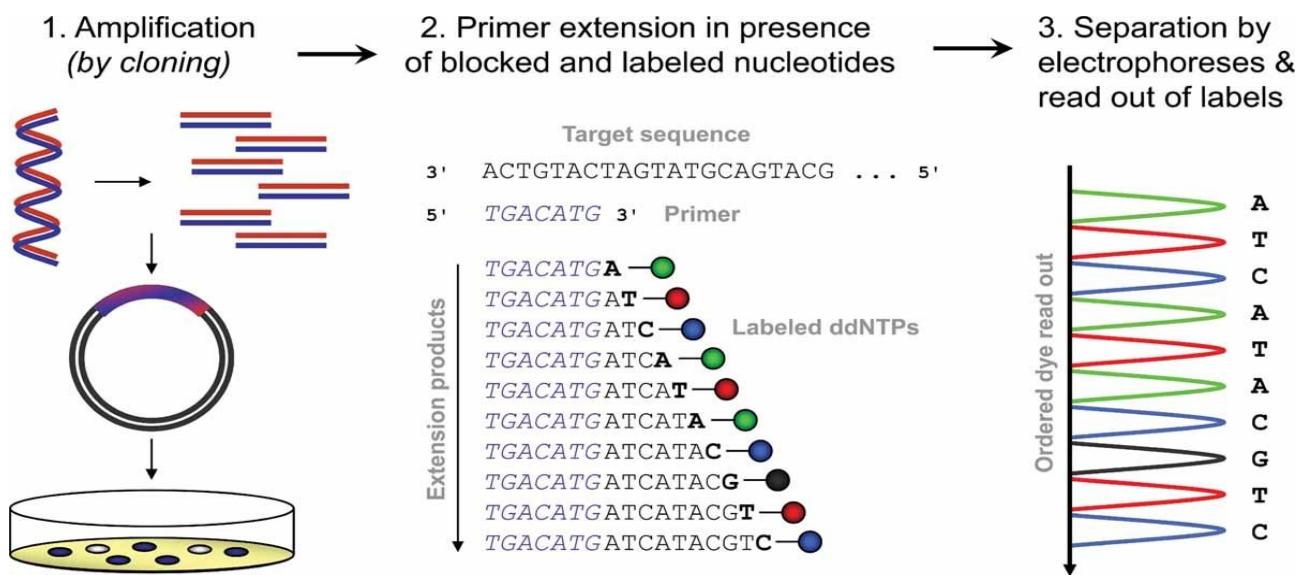


Figure 4: Schematic overview of Sanger sequencing. DNA is first randomly fragmented and amplified (shown is the shotgun sequencing approach using cloning for *de novo* sequences). A unique fluorophore is appended to the fragments for each possible nucleotide. Finally the fragments are sorted by weight and the sequence is assembled by reading out the fluorophores. Figure reproduced from (Kircher & Kelso, 2010).

### 1.1.5 Next-generation Sequencing

By the end of the 2000s, an improvement to the Sanger method called pyrosequencing saw commercialism by several large companies such as Illumina, dawning the era known as next-generation sequencing (NGS) or massively-parallel sequencing (MPS) that would underlie the most popular methods used today (Heather & Chain, 2016). Both Sanger and pyrosequencing share in common the sequence-by-synthesis (SBS) principle, that is the synthesis of cDNA and reading of the sequence in a base-wise manner (Uhlen & Quake, 2023). Rather than using fluorophores and electrophoresis, pyrosequencing works through the real-time conversion of pyrophosphate into ATP. As cDNA is synthesized, each of the four possible nucleotides are added one base at a time, extending the strand. The match of a nucleotide complementary to the target produces a base pair, releasing pyrophosphate which is then converted to ATP using ATP sulfurylase in the reaction mixture. Finally, ATP is used as a substrate for luciferase, a luminescent reaction that occurs in proportion to the amount of pyrophosphate (Nyrén & Lundin, 1985). This can be used to determine each nucleotide in the strand sequence, including those that are repeated. Commercialism of this approach broke through the unsolved bottlenecks of Sanger sequencing, and it is now possible to sequence a human genome in a day for less than \$1,000, in stark contrast to the 13 years and nearly \$3 billion spent by the Human Genome Project up until 2001 (Uhlen & Quake, 2023).

**Figure 5.**

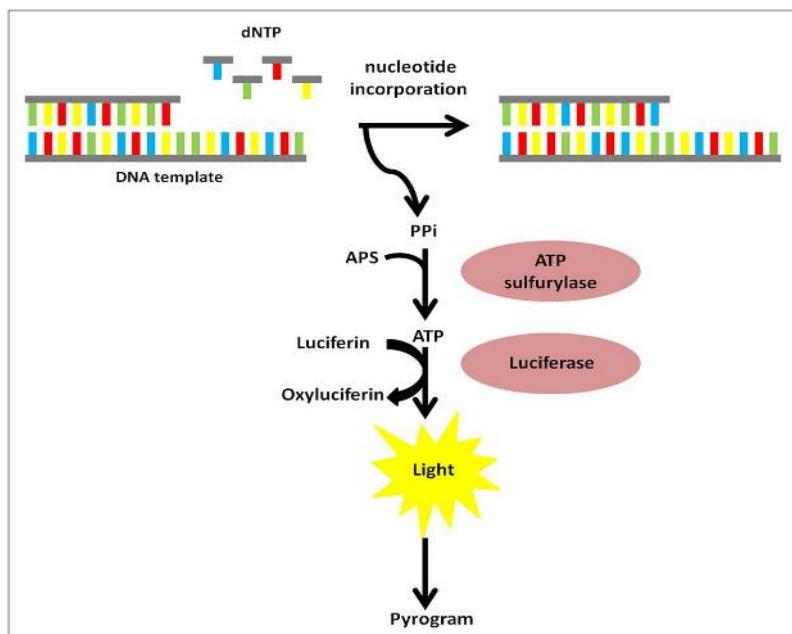


Figure 5: Simplified schematic of pyrosequencing. The diagram illustrates steps following the initial fragmentation step that is shared with Sanger sequencing. Each of the nucleotides are introduced to the DNA template mix, and those that are complementary to the template are incorporated into the growing synthesis strand, allowing a luciferase signal to be read from the released pyrophosphate. Figure reproduced from (Rybicka et al., 2016).

With high-throughput solutions for sequencing now widely available by the late 2000s, many researchers began to embrace this modality over microarray as the *de facto* gene expression method. The hybridisation basis of microarray necessitates *a priori* knowledge of the genes for quantification and detection is limited to the sequences defined by the probes deployed. In comparison, sequencing is applicable to cases where the genome of the target species is unavailable, as well as nuanced events such as RNA editing events or differential isoform usage (Malone & Oliver, 2011). Nevertheless, microarrays remain in usage today for its maturity and cost-effectiveness in answering targeted questions.

It can also be argued that biases related to microarray, at least in conventional usage for quantification of gene expression in well-annotated genes, have been largely solved, while those in sequencing continue to be an active area of research. Sources of variation in microarrays have been understood to be largely related to the manufacture of equipment and the conditions between labs; for instance differences between laser scanners or degradation of fluorescent dyes in relation to ozone (Malone & Oliver, 2011). Meanwhile, one of the most problematic issues that plague sequencing is that of sequencing depth. In regards to RNA sequencing (RNAseq), the reading of greater numbers of transcripts is required to capture the gene expression of more lowly expressed genes, inflating the costs of an experiment. This problem is exacerbated in the presence of very highly expressed genes, which compete for reads. The other major issue is heterogeneity in coverage along a transcript, for example the tendency for sequencing to underestimate GC-rich and poor fragments (Risso et al., 2011). This and other phenomena are not observed with microarrays and though many normalisation methods have been developed to address the issues of RNAseq, each have such nuanced advantages and trade-offs that tools have been developed just for their selection on a per-dataset basis (Scheepbouwer et al., 2023).

**Figure 6.**

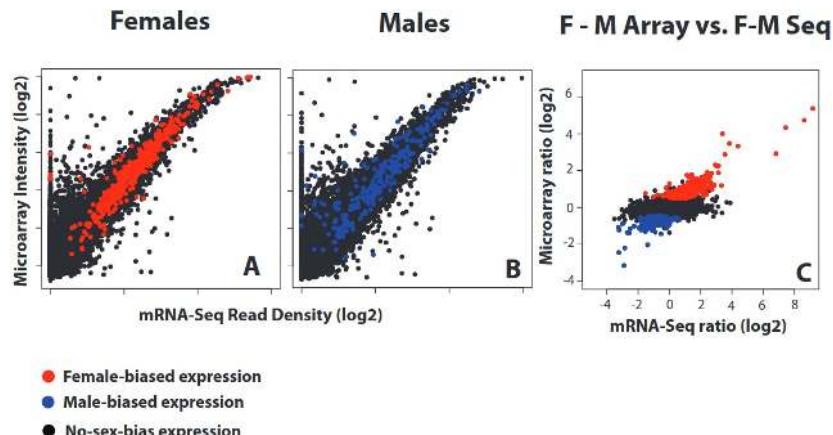


Figure 6: Correlation between microarray intensity and RNAseq reads in matched samples from *D. pseudoobscura*. The two assays show a near linear relationship, though microarray appears to saturate signal at higher intensity and have sensitivity for a number of genes of low intensity that could not be detected through RNAseq. Figure reproduced from (Malone & Oliver, 2011).

### 1.1.6 Single-cell Sequencing

In spite of the roadblocks unique to NGS, its throughput and suitability for exploring *de novo* structures have led to the technology's staying power. The benefit of further developments have been incremental advancements in sequencing sensitivity and accuracy but also larger leaps in the multiplexing and resolution of samples. Easily the most influential has been the development of single-cell sequencing (scRNASeq). The earliest example dates back to 2009, when Tang and colleagues managed to sequence mRNA from a single cell, detecting 75% more genes than from a microarray approach (Tang et al., 2009). The years that follow would see a proliferation of competing single-cell technologies, often differing wildly in approach (see Figure 7). Some companies such as 10x Genomics would centre focus around these technologies, while consortiums would form to create large atlases that aim to sample every type of cell in an organism. In 2013, single-cell sequencing would be named the "Method of the Year" by Nature ("Method of the Year 2013," 2014).

**Figure 7.**

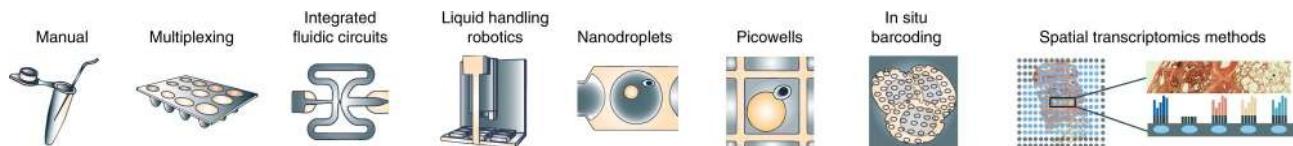


Figure 7: Rough timeline of the advancements in single-cell technologies. The latest technologies not only aim to preserve single-cell resolution but also resolution of the spatial localisation of the cell. Figure reproduced from (Aldridge & Teichmann, 2020).

The most popular single-cell technology, and the one used in this thesis, is droplet-based scRNASeq. Here, water-in-oil droplets encapsulate cells as they are passed through a microfluidics device, at frequencies of tens of thousands of droplets per second and scaling to millions of droplets (Salomon et al., 2019; X. Zhang et al., 2019). Several competing implementations exist today, the most ubiquitous being inDrop, Drop-seq, and 10X Genomics Chromium. All three share in common similar principles for encapsulation by droplets, as well as introduction of regents as droplets are passed. Crucially, they all feature usage of cell barcodes and unique molecular identifiers (UMIs). Cell barcodes are unique and predefined nucleotide sequences introduced to each droplet through gel beads that become associated with the cDNA that is reverse transcribed from the mRNA. UMIs are random sequences incorporated into the first strand of cDNA synthesis, such that all following cDNA that are PCR amplified carry the same UMI. The combination of these

modifications allow for the assignment of reads to a specific cell and the handling of PCR bias by distinguishing transcript abundance due to PCR as opposed to abundance due to gene expression. While each of these features are not necessarily unique to droplet based single-cell transcriptomics, their strengths and compatibility with the relatively straightforward technology cemented the method's success.

**Figure 8.**

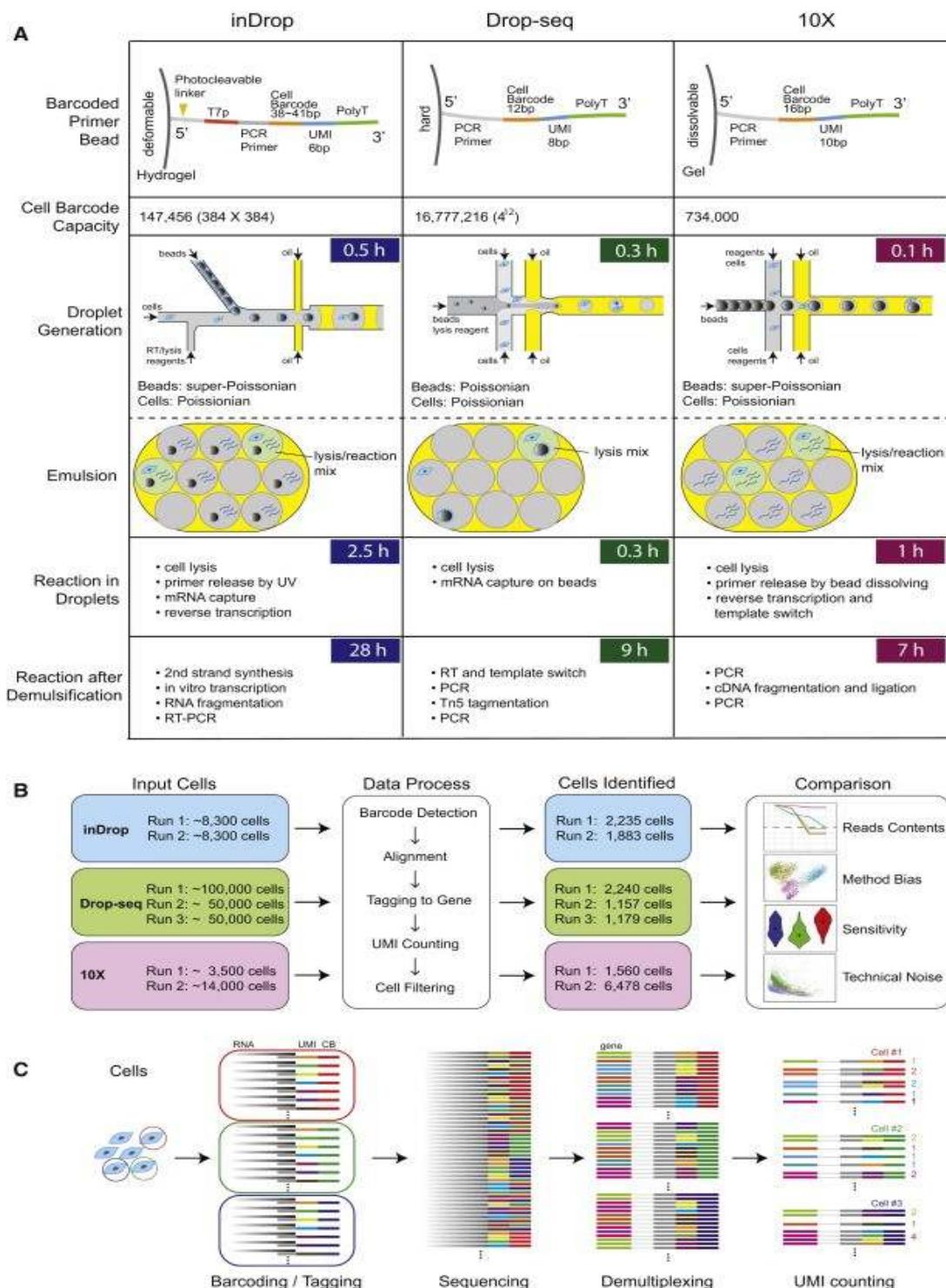


Figure 8: Comparison of three popular droplet-based scRNASeq pipelines. All share in common the use of barcodes for differentiating cells and UMIs for mitigating PCR bias. Figure reproduced from (X. Zhang et al., 2019).

### 1.1.7 Spatial and Long-read Sequencing

Today, the transcriptomics field is at a turning point in combining advantages of previously disparate technologies. These new methods aim to provide spatial localisation of transcripts, single-cell resolution, and genome-wide coverage; encompassed by a class of technologies dubbed spatially resolved transcriptomics (SRT), which would also go on to be named method of the year in 2021 (Marx, 2021). Perhaps the final missing layer is another method of the year called long-read sequencing (Marx, 2023). As already mentioned, a common limitation of all sequencing methods since the days of Sanger sequencing is the inability to process fragments longer than 150-200 basepairs. Long-read sequencing overcomes this, though at the cost of massively increased error rate which remains a significant problem area today. Methods such as spatial transcriptomics and long-read sequencing remain out of scope for this thesis work, which utilises single-cell sequencing data, however, their importance for future research in Alzheimer's Disease and other fields cannot be understated.

**Figure 9.**

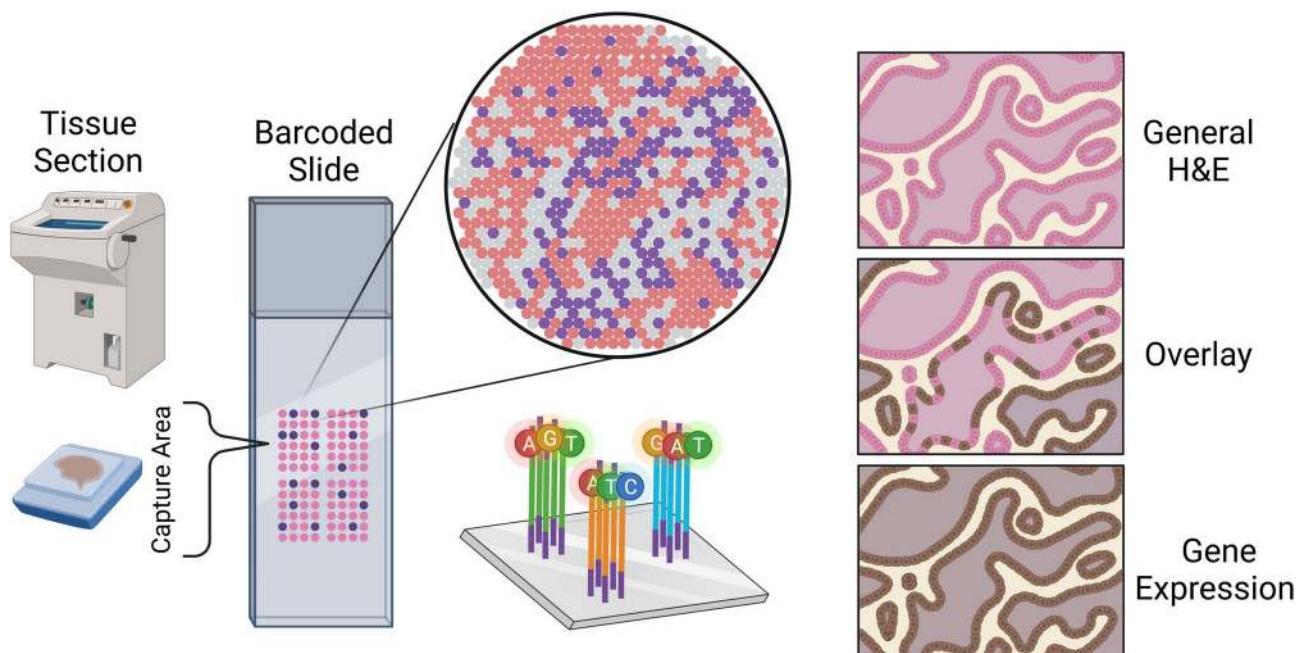


Figure 9: Illustrated schematic of the slide mounting process for spatial sequencing, accompanied by a simulated visualization of the data output from a spatial transcriptomics experiment. Figure reproduced from (Shireman et al., 2023).

## 1.2 Proteomics

### 1.2.1 Overview

The biological relevance of proteomics also cannot be understated, as proteins are a closer functional readout of an organism than transcripts. The most common application of proteomics is to measure the abundance of different proteins in a sample of tissue. This is conventionally carried out using mass spectrometry (MS), an instrument that measures mass-to-charge ( $m/z$ ) values and signal intensities of ions (Shuken, 2023). In most experiments, the machine operates on peptides, smaller chains of amino acids that together form the structure of a protein. In such kinds of bottom-up approaches, proteins are digested into peptide fragments using proteases and then further broken down into a gas phase of ions. These ions are sprayed into the mass spectrometer which measures their electrical properties. The data is compared to a database of peptide MS information, identified, and used to infer the likely protein composition of the sample.

**Figure 10.**

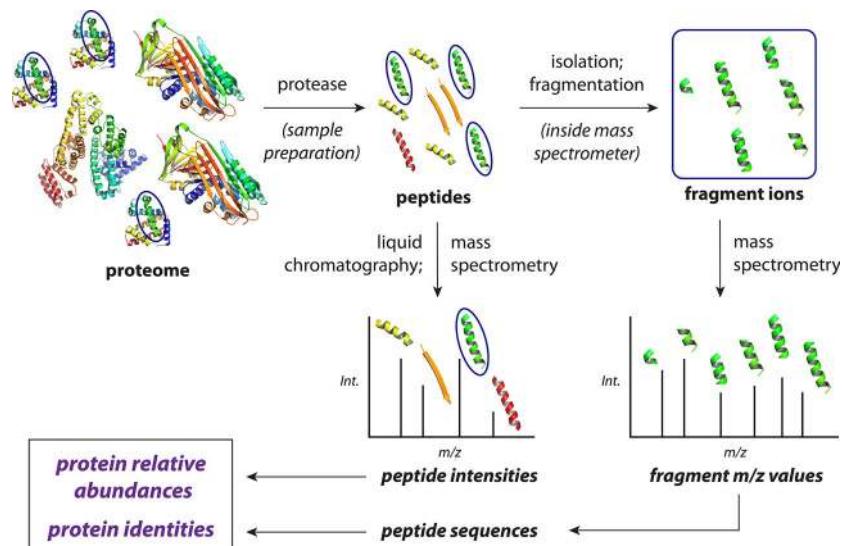


Figure 10: High-level schematic of the basic steps of conventional bottom-up mass spectrometry proteomics. Proteases are used to break down proteins into peptides, which are then ionised. Mass-to-charge or  $m/z$  values of the ions are measured by the spectrometer and are the primary data source for the identification of peptide intensities. The protein composition of the sample can then be deduced from the individual peptides. Figure reproduced from (Shuken, 2023).

Use of mass spectrometry in proteomics is said to have taken off in 1989, when electrospray ionisation became available and could be used to vaporise proteins (Fenn et al., 1989; Mann, 2016), eventually leading to a Nobel Prize in Chemistry in 2002. Prior to this development, methods used chemical peptide-sequencing methods like Edman degradation, which was limited in capability for the analysis of very small amounts of peptides in complex mixtures. The new method brought sensitivity to the femtomole level.

While the first wide-scale analysis using electrospray ionisation covered only 19 peptides (Hunt et al., 1992), in a period of 25 years the coverage would increase 1000-fold. This largely due to the assistance of large databases for MS/MS spectra, which continues to be the main driving force behind the power of mass spectrometry for proteomics today.

### 1.2.2 Mass Spectrometry

Three components comprise all mass spectrometers: the source of ions, the mass analyser, and the detector (Sinha & Mann, 2020). In order for peptides or proteins to be compatible with these components, they must be converted into a gaseous phase of ions. Through a process still not fully understood, liquid containing peptide or protein are passed through a small opening set to a high voltage of about 2-4 kV using high-performance liquid chromatography (HPLC), promoting the disintegration of the liquid contents into ions, which are then passed into the mass analyser for separation by their m/z values.

Quadrupole mass analysers, by far the most common type of analyser, operate on the principle of accelerating ions and measuring their trajectories along a quadrupole, an arrangement of electrically charged metallic rods within a vacuum (Wilkinson, 2021). A TOF or time-of-flight quadrupole analyser captures velocity differences on the order of sub-microseconds between acceleration at 20 kV and arrival time at the detector. In contrast, an Orbitrap quadrupole analyser uses oscillation frequency rather than velocity, as ions move along a metal spindle. Before arriving at the detector, ions may also undergo fragmentation in a special quadrupole known as a collision cell. This produces what is known as MS2 spectra, which some methods use to supplement the MS1 spectra that is produced from unfragmented ions being read by the detector (T. Huang et al., 2020).

**Figure 11.**

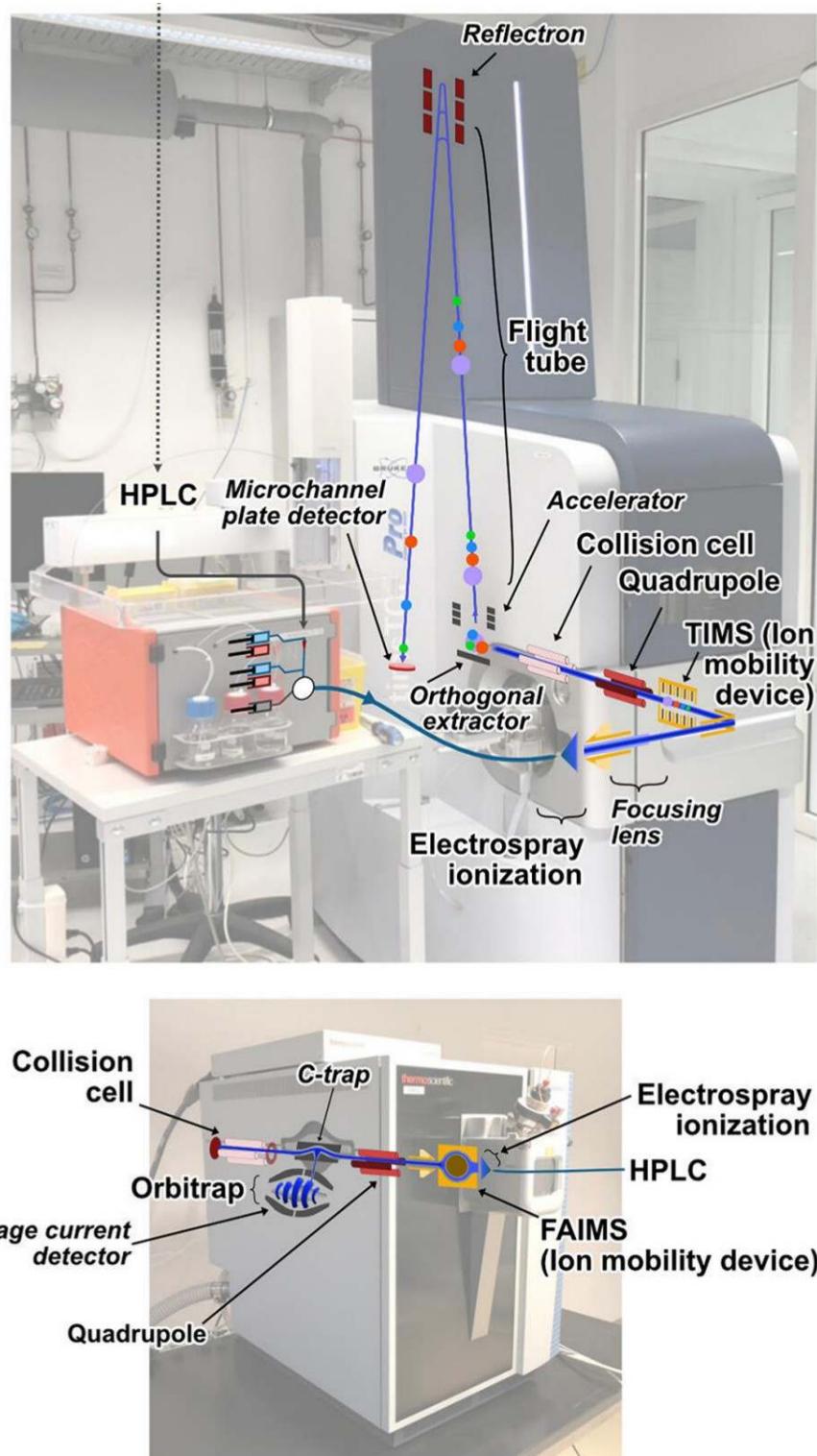


Figure 11: Components of a typical mass spectrometry machine. Proteins or peptides are passed to the machine through HPLC (high-performance liquid chromatography) allowing electrospray ionisation to take place. Following a series of quadrupoles, m/z values and signal intensities of the ions are recorded at the detector. Figure adapted from (Sinha & Mann, 2020).

### 1.2.3 Top-down vs. Bottom-up

Mass spectrometry proteomics can be approached from either a top-down or bottom-up perspective (Roberts et al., 2024). Significantly less common due to technical challenges, top-down approaches pass complete unfragmented proteins into the detector. State-of-the-art methods are still considered underdeveloped compared to bottom-up proteomics, but offer the promise of resolving proteoforms – variations of protein structure that include post-translational modifications (PTMs) and other *de novo* or genetically defined variants. A significant technological development enabling the feasibility of top-down proteomics was the incorporation of MALDI or matrix-assisted laser desorption/ionization (Hillenkamp et al., 1991). Compounds of larger molecular mass are traditionally difficult to ionise but with MALDI they are embedded in a matrix compound calibrated to absorb the wavelengths of laser light used to trigger the ionisation process.

**Figure 12.**

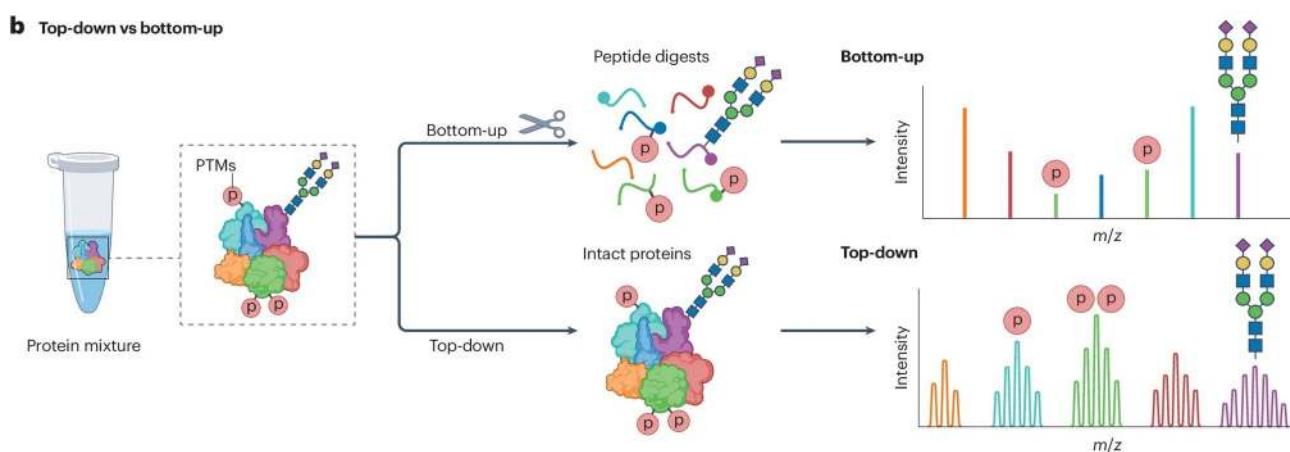


Figure 12: Schematic comparing bottom-up and top-down approaches to mass spectrometry proteomics. Whereas top-down approaches aim to maintain intact protein structure as it passes to the detector, bottom-up approaches incorporate a cleavage step to operate on the level of peptides. Figure adapted from (Roberts et al., 2024).

### 1.2.4 Labeled vs. Label-free Quantification

A categorisation that applies to both bottom-up and top-down approaches is whether the method uses label-free quantification (LFQ) or is label-based, the difference pertaining to the deconvolution of samples before quantification (Guo et al., 2022; Z. Wang et al., 2021). By labelling samples before separation of proteins/peptides during liquid chromatography, multiple samples can be multiplexed into a single run. The most popular approaches are those that use isobaric chemical labelling such as TMT (tandem mass tag) and iTRAQ (isobaric tag for absolute and relative quantification) (Sivanich et al., 2022). These methods bind the tag to stereotypical residues such as proline-rich areas of the N-terminus, limiting interference with the quantification and fragmentation functionality of the mass spectrometer. Isobaric chemical groups are those that differ negligibly in terms of molecular mass but have identifiable differences in their atomic structure, making them

ideal for labelling while reducing unwanted analytical impact. TMT and iTRAQ label peptides/proteins after extracting from the samples of interest, but when working with cell cultures, labelling can also take place as the cultures are grown. This is the basis of the popular method SILAC (stable isotope labelling by amino acids in cell culture) (Mann, 2006). Considered a metabolic rather than chemical approach, SILAC may allow for more complete and consistent labelling.

**Figure 13.**

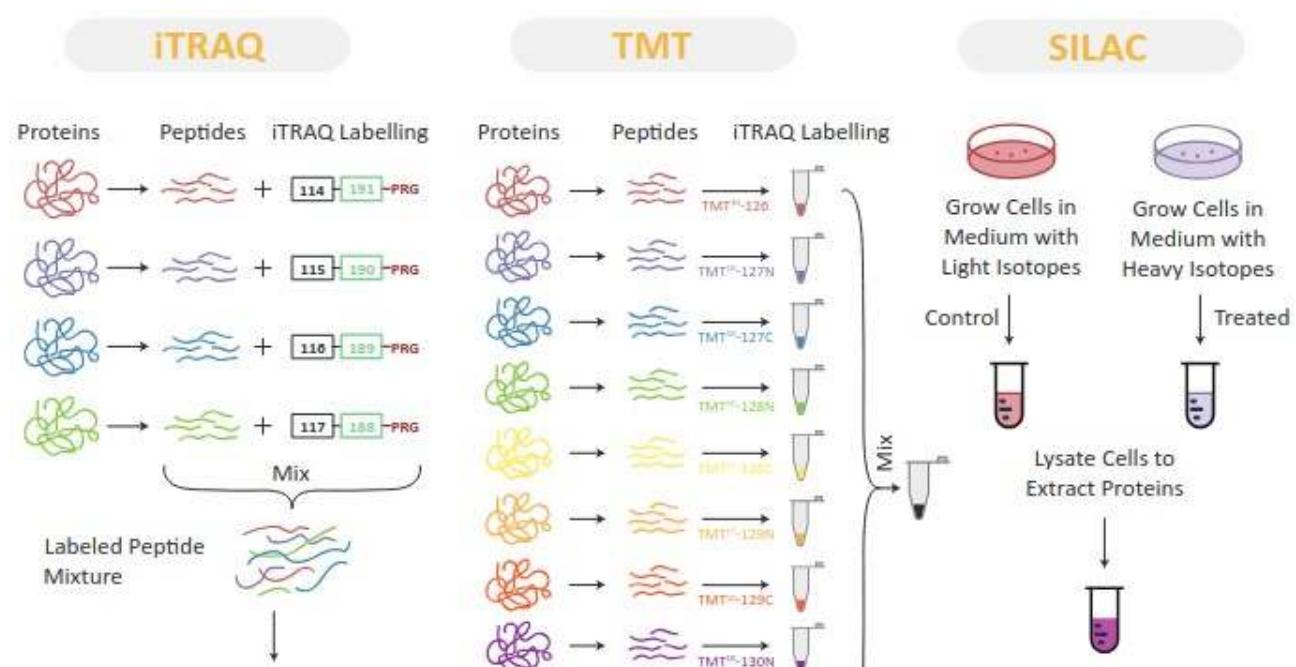


Figure 13: Schematic comparing three of the most popular label-based proteomic methods. Whereas iTRAQ and TMT apply isobaric tags to isolated proteins and peptides, SILAC uses tags in the media which cells are grown. Figure adapted from <https://www.creative-proteomics.com/pdf/Comparison-of-Three-Label-based-Quantification-Techniques-iTRAQ-TMT-and-SILAC.pdf>.

### 1.2.5 Data-dependent vs. Independent Acquisition

Bottom-up, label-free quantification approaches are usually divided into two main implementations, DDA or data-dependent acquisition and DIA or data-independent acquisition (Guan et al., 2020). DDA is the older of the two and considered a gold standard approach for proteomics. The key difference between the two methods lie in how peptides are selected for fragmentation as well as MS<sub>2</sub> data collection before peptide entry into the collision cell. DDA fragmentation uses a subset of the peptide data, selected by automated peak selection of ion intensity. By doing so, the selection criteria operates in a data-dependent manner. In contrast, DIA approaches set predefined m/z windows to partition peptides which are then batched together for MS<sub>2</sub> detection.

**Figure 14.**

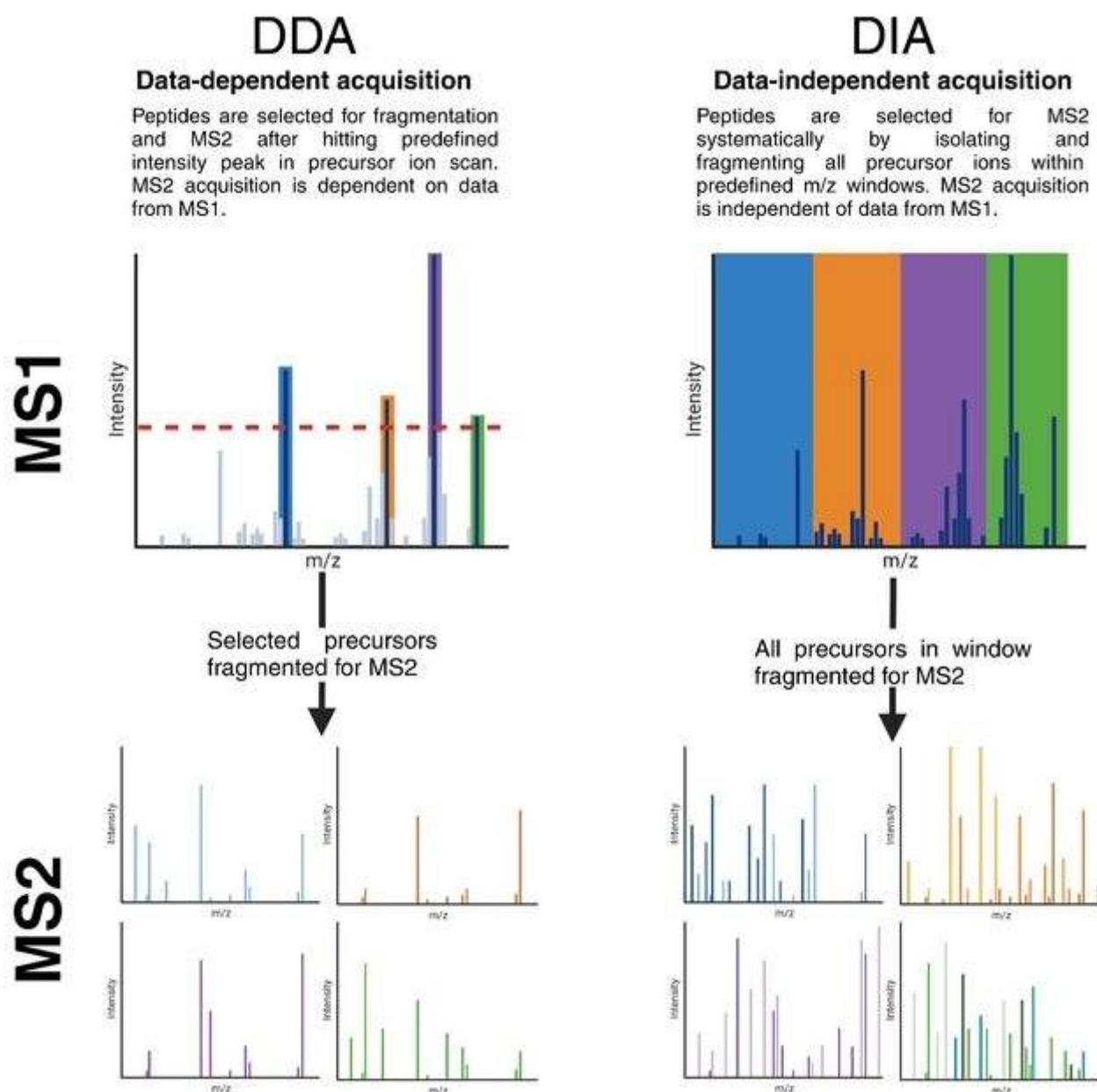


Figure 14: Comparison of DDA and DIA proteomics. The notable difference is the selection of peptides for MS2. DDA selects a subset of peptides on the basis of ion intensity peaks in a manner dependent on MS1 data. DIA on the other hand, passes all peptides to MS2 and partitions the peptides using predefined  $m/z$  windows. Figure reproduced from (Ward et al., 2024).

Since MS2 data from DIA methods cover a wider spectrum of peptides, the general consensus is that it outperforms DDA in terms of quantification reproducibility, specificity, accuracy, and situations of low protein availability; this has been demonstrated in experiments using “gold standard” samples where proteins have been spiked-in (Barkovits et al., 2020; Willems et al., 2021). Nonetheless, there is debate over the validity of such claims, largely owing to the fundamentally different analysis approaches utilised by the two

methods. DDA conventionally resolves proteins from peptides using a sequence database search while DIA depends on spectral library searches. However, it is possible to perform a DDA experiment using spectral library, and a controlled experiment by (Fernández-Costa et al., 2020) showed that DIA and DDA perform comparably when doing so. Though overall DIA still maintained a slight edge in terms of reproducibility, the authors argue that spectral library searches better optimised for DDA may eliminate the gap. Indeed, DDA offers the potential for operational efficiency, as the peptide selection approach attempts to minimise redundant peptide ion selection and increase depth of protein coverage (Bateman et al., 2014).

**Figure 15.**

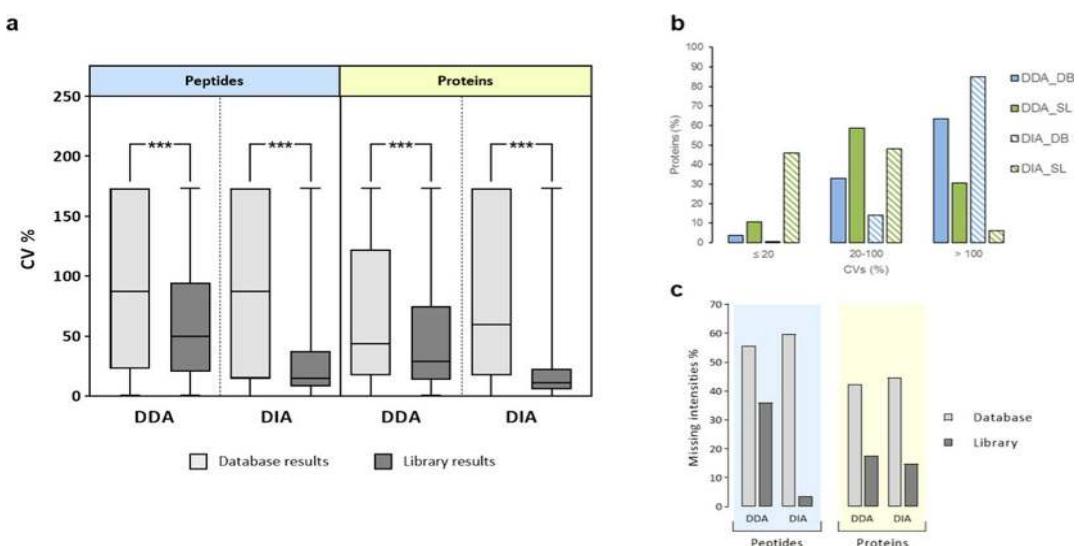


Figure 15: Comparison of coefficients of variation of DDA and DIA approaches with different database search methods in controlled samples. While DIA remains more reproducible, DDA methods are greatly improved by referencing ions through a spectral library rather than the convention of a sequence database. Figure reproduced from (Fernández-Costa et al., 2020).

### 1.2.6 Targeted Proteomics

The methods discussed thus far all aim to capture an unbiased survey of the proteome however, targeted approaches have also been developed for the analysis of singular or a subset of proteins (Borràs & Sabidó, 2017). These methods typically build off of existing mass spectrometry protocols but filter acquisition to the ions of interest, as is the case in selected ion monitoring or SIM. Besides improvements in the sensitivity for proteins of interest, targeted proteomics is ideal for studying proteoform variations such as post-translational modifications. Furthermore, it is more amenable for the derivation of absolute quantification, such the concentration of biomarkers in plasma (Uchida et al., 2013). It is predicted that with future developments, the distinction between targeted and unbiased proteomics will blur, as peptides become targetable on an individual basis but also in a highly multiplexed fashion (Kang et al., 2017).

### 1.2.7 Spatial Proteomics

In an analogous trend to the transcriptomics world, proteomics has been making way towards facilitating spatial resolution of the proteome, also garnering a Nature Method of the Year for 2024 ("Method of the Year 2024," 2024). One of the more popular methods is imaging mass cytometry (IMC) (Baharlou et al., 2019), where the tissue is coated with metal isotope-tagged antibodies and then ablated with a laser. The aerosolised and ionised matter is then fed into the mass cytometer for protein quantification. Using the isotopes, the location of the proteins can be mapped back to their coordinates on the original tissue. A variation of this method includes mass spectrometry imaging (MSI) (H. Zhang et al., 2023), where the ionised tissue is instead quantified using a mass spectrometer, consistent with conventional proteomics. This garners the advantage of building off the wealth of work in that area, in addition to the capacity for analysing complex proteoform structure such as post-translational modifications. Another method, more similar to the approaches popularised in spatial transcriptomics, is cyclic immunofluorescence (cycIF) (J. Lin et al., 2016), where antibodies targeting a panel of proteins are applied in a series of cycles and imaged. CycIF is currently able to achieve greater resolutions than IMC and MIBI but does not utilise mass spectrometry, limiting its application when analysing complex proteoforms. Finally, laser-capture microdissection (LCM), discussed in greater detail in later sections due to its use in this thesis work, brings forward the possibility of immunohistochemical staining on tissue followed by the precision extraction of single-cells, when can then be profiled using standard mass spectrometry, as well as transcriptomic methods. It is predicted that with future developments, spatial proteomics will be able to sample the entire proteome, including features such as PTMs, at a single-cell and spatially resolved resolution. Combined with spatial transcriptomics and epigenetics, holistic and fine-grained maps of physiology and disease may open the doors for a great leap in basic research and translational medicine.

**Figure 16.**

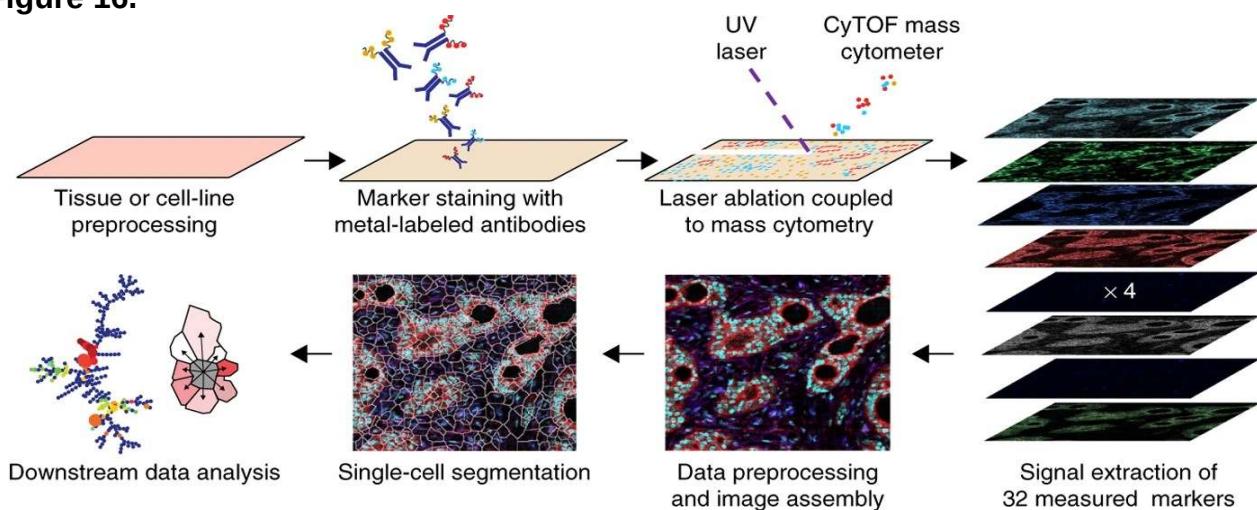


Figure 16: Schematic of the workflow of imaging mass cytometry (IMC). Tissue labelled with metal isotope-tagged antibodies is laser-ablated, allowing spatial quantification of proteins through a mass cytometer. Figure reproduced from (Giesen et al., 2014).

## 1.3 Alzheimer's Disease

### 1.3.1 Overview

Alzheimer's Disease (AD) is the most common tauopathy and form of dementia, accounting for 50-60% of the estimated 40-50 million dementia cases worldwide (Nichols et al., 2019), and is expected to almost double in prevalence every 20 years (Prince et al., 2013). The tauopathies are a class of neurodegenerative disorders characterised by the aggregation of pathological forms of microtubule-associated protein tau, encoded by the *MAPT* gene (Arendt et al., 2016; Spillantini et al., 1997). There remains no approved drugs that directly target tau at this time, though 164 trials assessing 127 drugs were underway as of 2024 (Cummings et al., 2024). Clinical presentation of the tauopathies vary (Josephs, 2017), but generally include some aspect of dementia, defined loosely as the progressive reduction in cognition and ability to live independently (Prince et al., 2013). The range of phenotypes among tauopathies is likely driven by their remarkable heterogeneity in spatial-temporal progression, cell-type specific effects, and predominant tau species, as exemplified in Figures 17 and 18.

**Figure 17.**

Disease	Predominant Tau isoform	Affected cell types	Affected brain regions	Pathology
Pick's disease	3R	Neurons and glia	Frontal, temporal, and parietal lobes; hippocampus	Pick bodies, neuropil threads, ramified astrocytes, and round aggregates
Corticobasal degeneration	4R	Neurons and glia	Frontal and parietal cortices; substantia nigra	Balloononed neurons, pretangles, astrocytic plaques, coiled bodies, and neuritic threads
Progressive supranuclear palsy	4R	Neurons and glia	Frontal cortices, subthalamic nucleus, brain stem	Neurofibrillary tangles, globose tangles, tufted astrocytes, coiled bodies
Globular glial tauopathy	4R	Neurons and glia	Frontal and temporal lobes	Globose oligodendrocyte inclusions
Argyrophilic grain disease	4R	Neurons and glia	Transentorhinal and entorhinal cortices, hippocampus	Argyrophilic grains, oligodendritic coiled bodies, neuronal pretangles
Alzheimer's disease	3R and 4R	Neurons	Entorhinal cortex, hippocampus, cortex	Neurofibrillary tangles, neuropil threads

Figure 17: Table demonstrating the neuropathological heterogeneity among various tauopathies. Figure reproduced from (Götz et al., 2019).

**Figure 18.**

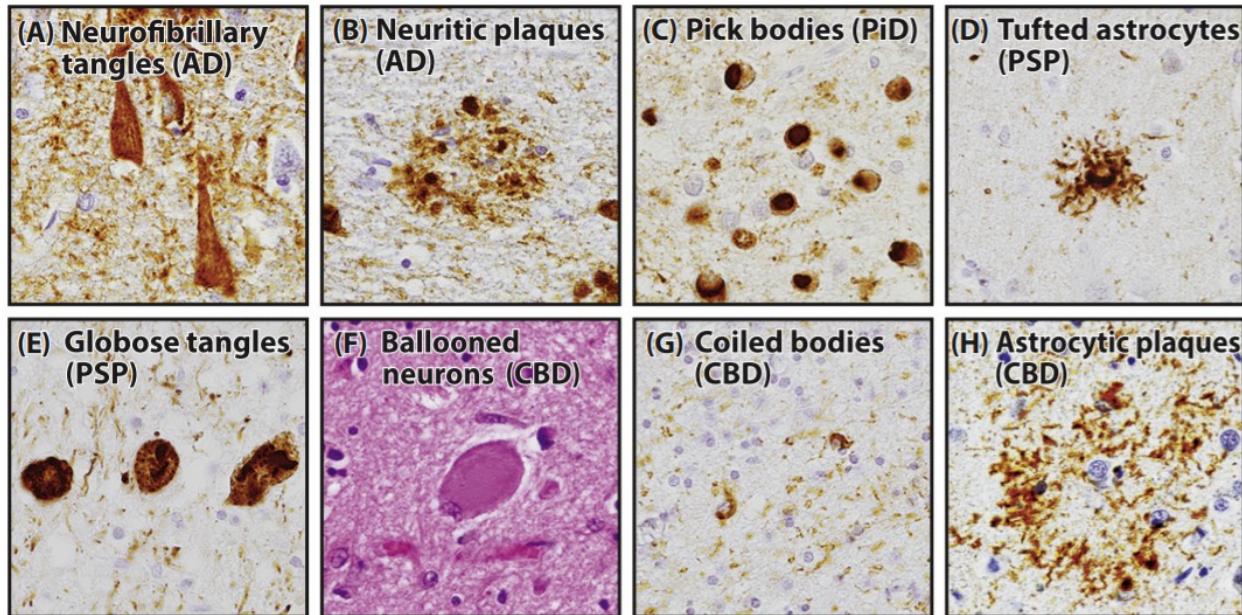


Figure 18: Immunohistochemistry in human post-mortem brain tissue showing the most common types of tau inclusions and the particular tauopathy that they are characteristic of. Figure reproduced from (Uemura et al., 2020).

AD is classified as a secondary tauopathy, as it features depositions of extracellular amyloid-beta (A $\beta$ ) plaques, derived from the *APP* gene, in addition to intraneuronal tau neurofibrillary tangles (NFTs) (Masters et al., 2015; Scheltens et al., 2021). AD can be divided into familial and sporadic forms as well as early and late-onset forms (EOAD and LOAD, respectively), distinguished by onset before or after age 65. While EOAD and LOAD are both largely sporadic, EOAD has a higher incidence of autosomal dominant mutations in *APP*, *PSEN1*, or *PSEN2* as well as rare variants in risk genes, and accounts for about 5% of all AD cases (Mendez, 2012; W. Zhang et al., 2020). Sex differences are also present in AD, with women reported as having 1.17 times the male prevalence rate (Nichols et al., 2019).

### 1.3.2 Amyloid Cascade Hypothesis

The amyloid cascade hypothesis (Hardy & Higgins, 1992; Selkoe & Hardy, 2016), first proposed in the early 1990s, remains one of the most widely studied models for the pathogenesis of AD. It proposes that the overproduction or impaired clearance of A $\beta$  peptides results in their aggregation into insoluble plaques, which disrupts neural communication and triggers neuroinflammation. As a consequence, tau proteins undergo hyperphosphorylation, causing them to form NFTs that compromise the structural integrity of neurons and impede intracellular transport mechanisms. However, this hypothesis is not without criticism and there is evidence suggesting that NFTs deposition precedes A $\beta$  plaques by as much as 10 years (Arnsten et al., 2020). Moreover, the primary tauopathies, which posit tau as the sole aggregate, suggest that tau alone is sufficient for pathogenesis.

**Figure 19.**

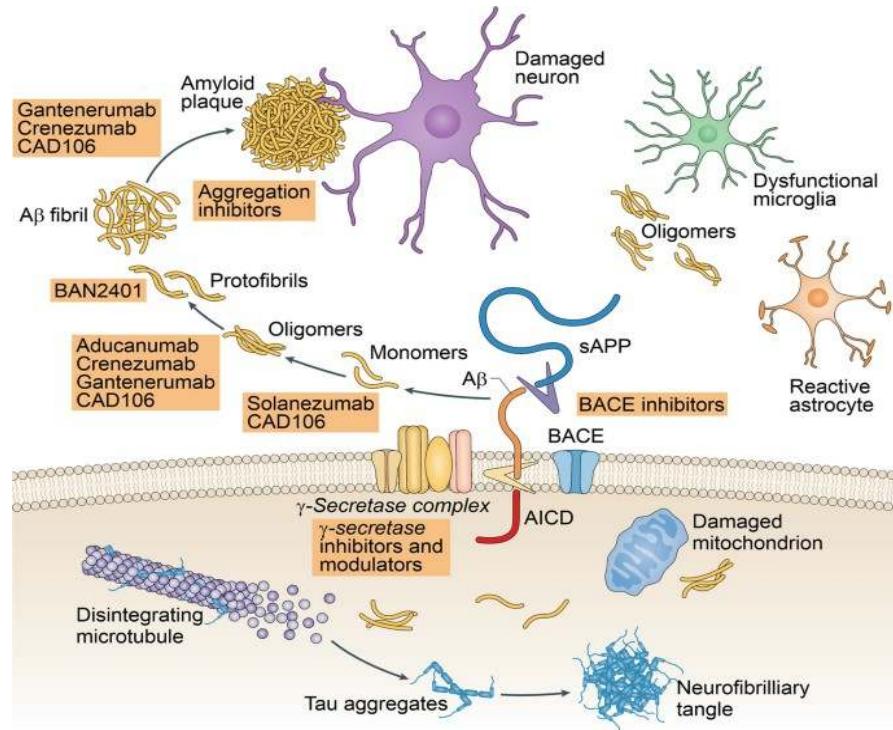


Figure 19: Diagram of the hypothesised pathophysiology of AD from the perspective of the amyloid cascade hypothesis. Shown also are trialled treatments that have been introduced to target various stage points of the disease pathway. Figure reproduced from (Panza et al., 2019).

### 1.3.3 Amyloid-beta Structure and Function

A $\beta$  is generated through the proteolytic processing of amyloid precursor protein (APP), a transmembrane glycoprotein, via a sequential cleavage mechanism involving  $\beta$ -secretase and  $\gamma$ -secretase. The  $\gamma$ -secretase complex, which utilises presenilin 1 or 2 (PSEN1 or PSEN2) as its catalytic subunit, produces A $\beta$  peptides of varying lengths, with A $\beta$ 40 and A $\beta$ 42 being the most abundant forms (G. Chen et al., 2017). Both APP and PSEN mutations are commonly associated with Familial Alzheimer's Disease (FAD), and may contribute to excessive A $\beta$  accumulation through various pathways, such as promoting A $\beta$  generation and disrupting autophagic degradation processes (Chong et al., 2018; Weggen & Beher, 2012). However, the underlying causes of Sporadic Alzheimer's Disease (SAD), which comprise over 90% of cases, remain less clear. It has been suggested that abnormal post-translational modifications of the amyloid- $\beta$  peptide enhance its neurotoxicity and promote aggregation, potentially triggering or accelerating the progression of SAD (Barykin et al., 2017). Additionally, genetic risk factors, particularly the apolipoprotein E (APOE)  $\epsilon$ 4 allele, play a significant role in the risk of developing SAD. Individuals carrying a single APOE  $\epsilon$ 4 allele face a 2 to 3-fold increased risk, while those with two copies experience up to a 15-fold greater likelihood of developing the disease (Yamazaki et al., 2019). Additionally, various other genetic risk factors, along with cardiovascular conditions such as diabetes and hypercholesterolemia, and lifestyle factors

including diet and sleep, have been the focus of extensive research in recent years for their potential influence on A $\beta$  metabolism in SAD (Oomens et al., 2021).

#### 1.3.4 Amyloid-Tau Interaction

Data thus far suggest that the interplay between A $\beta$  and tau aggregation, along with its impact on neuronal function, is a widespread and significant phenomenon (Busche & Hyman, 2020). A $\beta$  has been shown to promote the formation of tau oligomers, with both amyloid plaques and soluble A $\beta$  contributing to the spread and aggregation of paired helical filament (PHF) tau (He et al., 2018). Furthermore, A $\beta$  exposure renders tau more resistant to protease degradation (De Strooper, 2010). This suggests that A $\beta$  induces structural changes in tau, potentially through post-translational modifications, conformational shifts, or oligomerisation. Although multiple studies have suggested that pathological A $\beta$  and tau aggregates can co-localise within neurons and synaptic terminals (Manczak & Reddy, 2013), others have shown in human tissue and mouse models that such co-localisations occur in less than 0.02% synapses (Pickett et al., 2019). Another possibility is that A $\beta$  and tau influence each other indirectly by disrupting neuronal processes such as kinase regulation, glial activation, and neuroinflammatory responses (Busche & Hyman, 2020). Being a multifactorial disease, in addition to the central roles of A $\beta$  and tau, a range of other factors may play a role in AD pathology, including acetylcholine depletion, chronic neuroinflammation, oxidative stress, disruptions in metal ion homeostasis, glutamatergic dysregulation, insulin resistance, alterations in the gut microbiome, impaired cholesterol metabolism, mitochondrial dysfunction, and defects in autophagy (J. Zhang et al., 2024).

**Figure 20.**

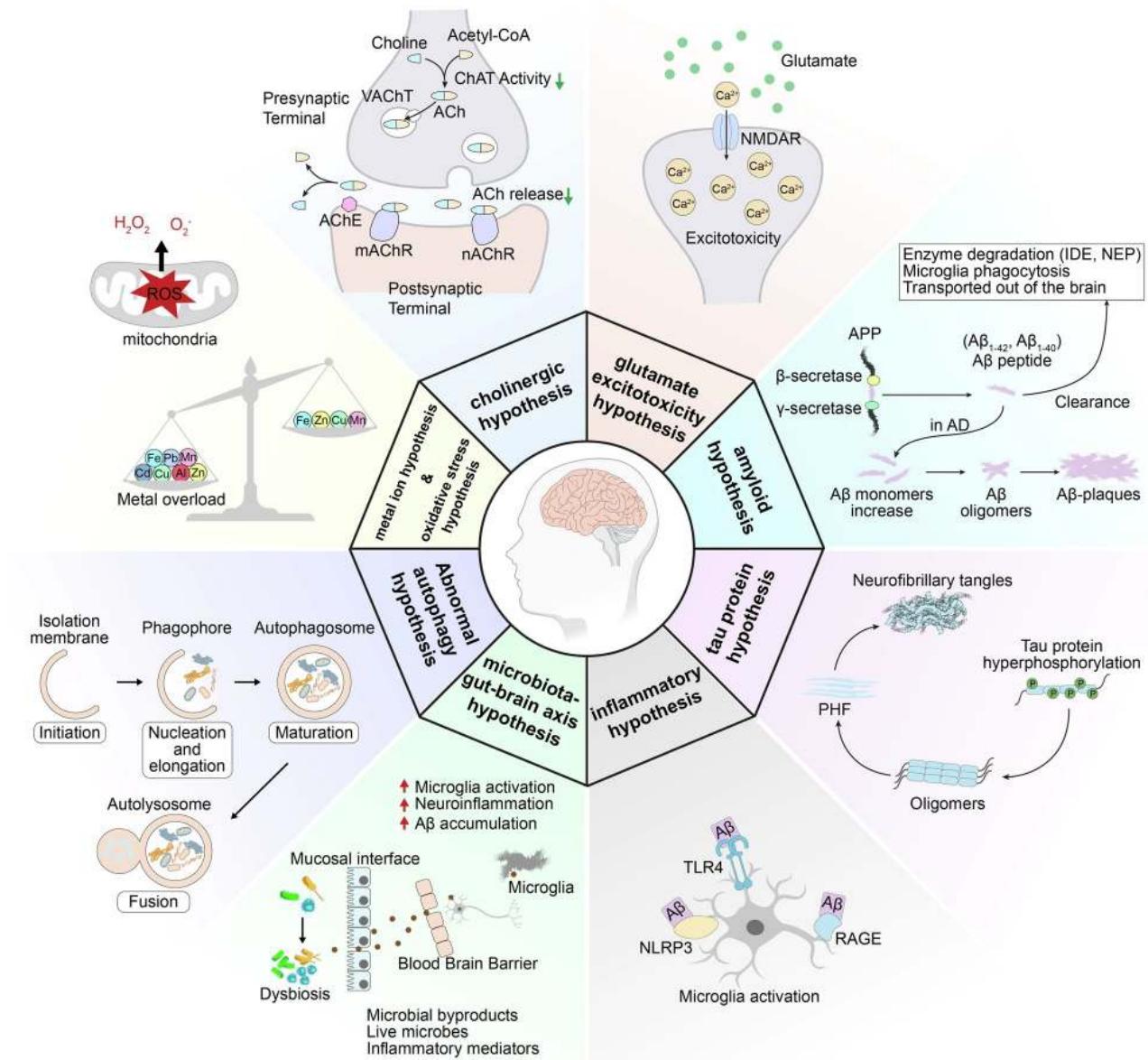


Figure 20: Schematic outlining the range of diverse mechanisms implicated in the pathology AD. The causality and interactions between them remain unclear, raising significant challenges in the development of effective treatments. Figure reproduced from (J. Zhang et al., 2024).

### 1.3.5 Tau Structure and Function

Tau is an intrinsically disordered protein (IDP) that lacks a stable three-dimensional structure under physiological conditions, allowing it to associate freely with microtubules (Stelzl et al., 2022). The human tau protein is encoded by the *MAPT* gene located on chromosome 17, and through alternative splicing, it produces six isoforms in the adult brain (Strang et al., 2019). These isoforms vary based on the inclusion or exclusion of

exons 2, 3, and 10, resulting in differences in the number of microtubule-binding repeat domains, either three (3R) or four (4R). The balance between these isoforms is developmentally regulated and is crucial for normal neuronal function. Alterations in this balance have been implicated in various neurodegenerative disorders beyond AD (Buchholz & Zempel, 2024).

**Figure 21.**

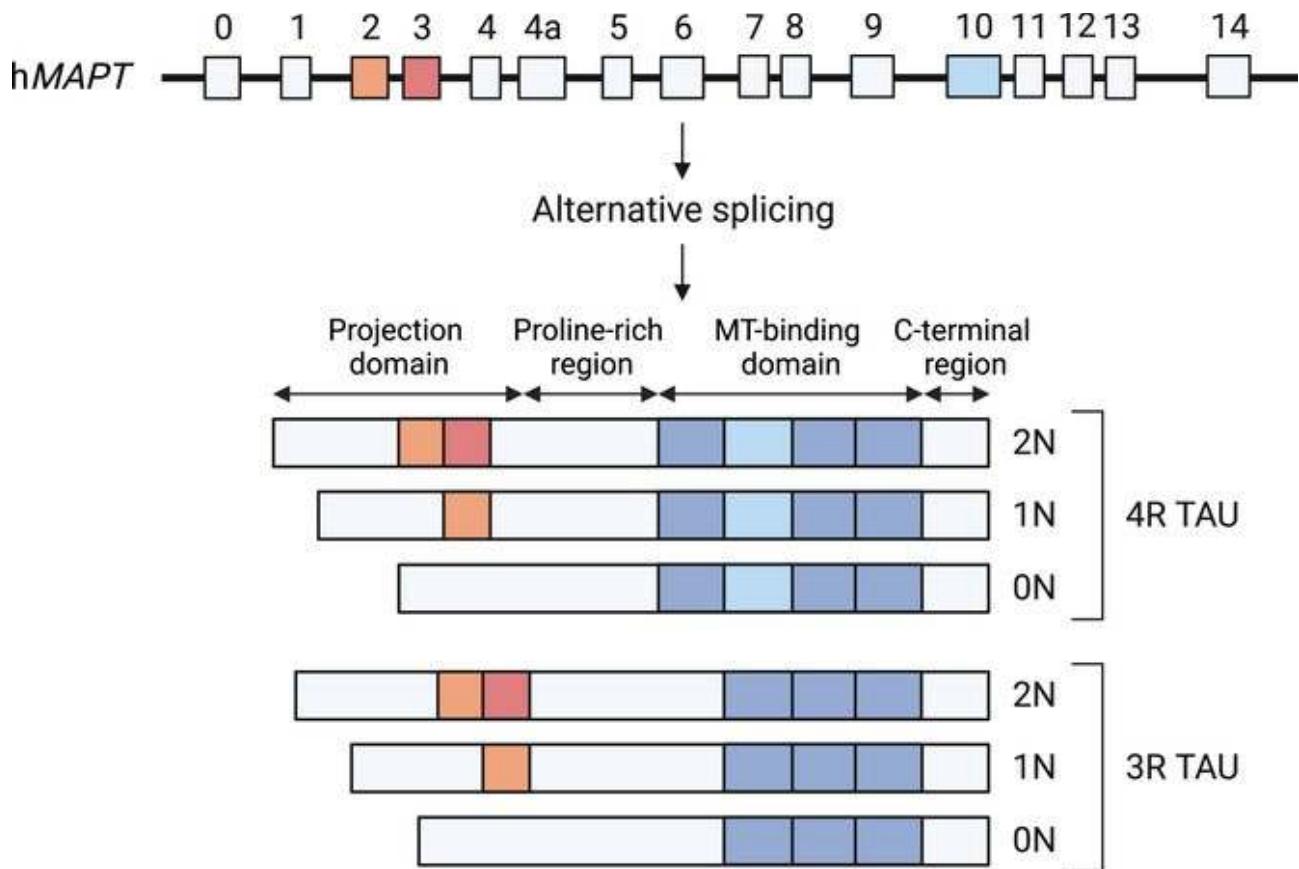


Figure 21. The six brain-specific human tau isoforms produced through alternative splicing of exons 2, 3, and 10. Tau can be structured into four distinct domains, of which the isoforms are defined by differences in the N-terminal projection domain and the C-terminal MT-binding domain. Figure reproduced from (Buchholz & Zempel, 2024).

### 1.3.6 Pathological Tau Accumulation

Tau protein follows a well-defined pattern of accumulation over time and across brain regions, closely mirroring the progression of clinical symptoms. This strong correlation makes tau a highly specific pathological indicator in Alzheimer's Disease (Braak & Braak, 1991; Malpas et al., 2020). Tau is primarily localised alongside microtubules in neuronal axons, though it is also detected at reduced levels in dendrites, the soma, and certain glial cells (Kanaan, 2024). Tau also contains multiple phosphorylation sites distributed across its N-terminal, C-terminal, and repeat domains, with their regulation dependent on the interplay between various kinases and phosphatases to preserve normal neuronal function.

(Drummond et al., 2020). In pathological conditions, dysregulated kinase and phosphatase activity causes tau to become hyperphosphorylated. This modification weakens tau's affinity for microtubules, leading to its dissociation and subsequent structural alterations. Mislocalised tau begins to accumulate, forming oligomers, PHFs, and NFTs within the cell body and dendrites (Goedert et al., 1991). These pathological changes progressively disrupt neuronal function, ultimately resulting in cell death (Alonso et al., 2018; Schneider et al., 1999). Pathological tau can manifest as a variety of forms, as can be seen in Figure 22, with mature PHF1-positive tangles being the form most conventionally associated with advanced Alzheimer's Disease (Moloney et al., 2021).

**Figure 22.**

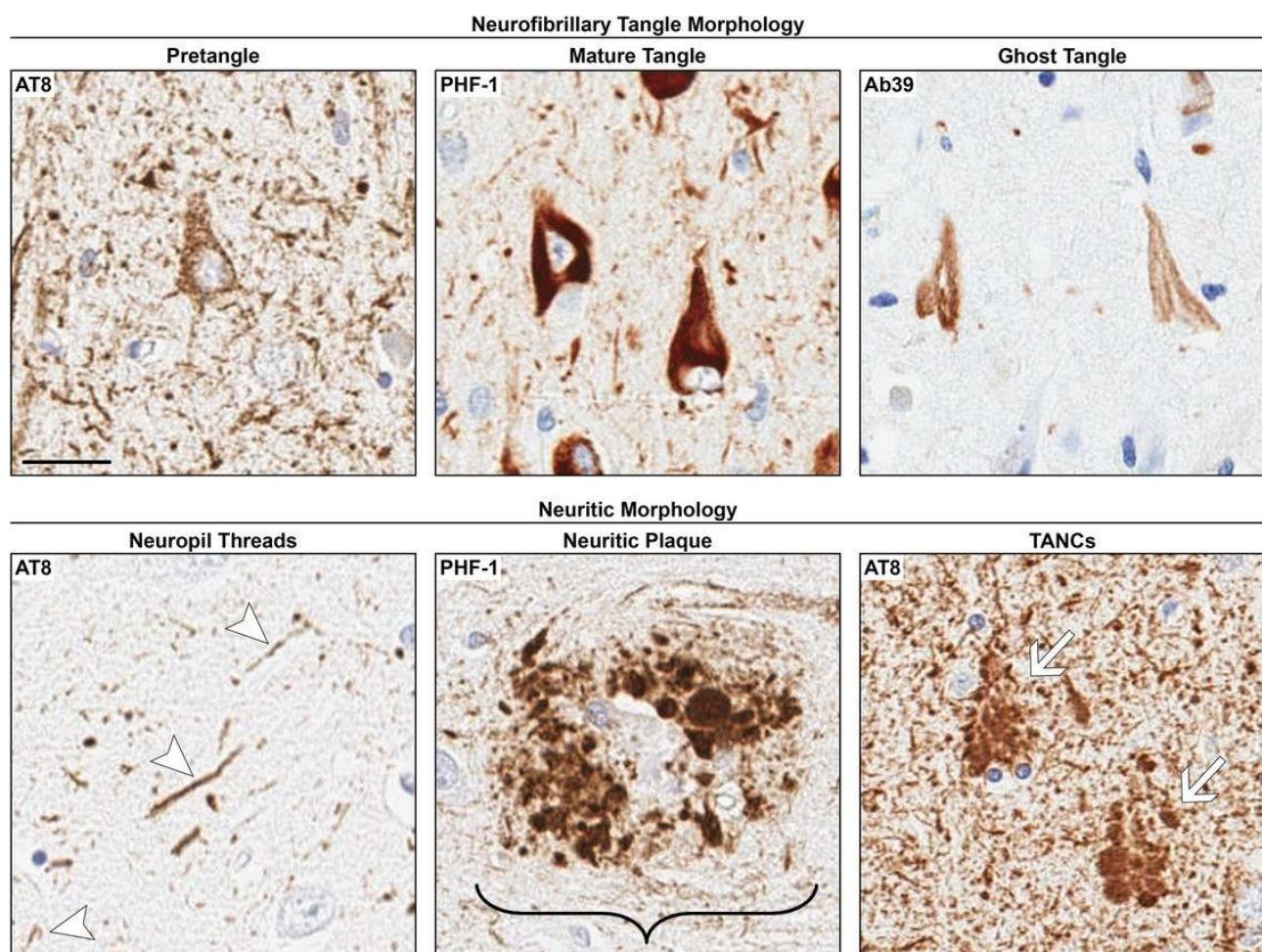


Figure 22. Various types of neurofibrillary tangle morphologies visualised through immunohistochemistry in human post-mortem AD tissue. Shown in all images is the CA1 subregion of the hippocampus. Figure reproduced from (Moloney et al., 2021).

### 1.3.7 Tau Post-translational Modifications

In addition to hyperphosphorylation, several other post-translational modifications of tau protein have been implicated in the promotion of tau aggregation and enhancement of its neurotoxicity. Proteolytic cleavage of tau, known as truncation, generates tau fragments

that are prone to aggregation (Boyarko & Hook, 2021). Glycosylation, the enzymatic addition of sugar moieties, and glycation, the non-enzymatic attachment of sugars, both influence tau's propensity to aggregate. Glycation, in particular, has been shown to promote tau polymerization and stabilise aggregated forms, contributing to NFT formation (Alquezar et al., 2021). The attachment of small ubiquitin-like modifier (SUMO) proteins to tau, termed sumoylation, also affects its solubility and degradation. Sumoylation has been observed to decrease tau solubility, potentially facilitating its aggregation and accumulation within neurons (H.-B. Luo et al., 2014).

**Figure 23.**

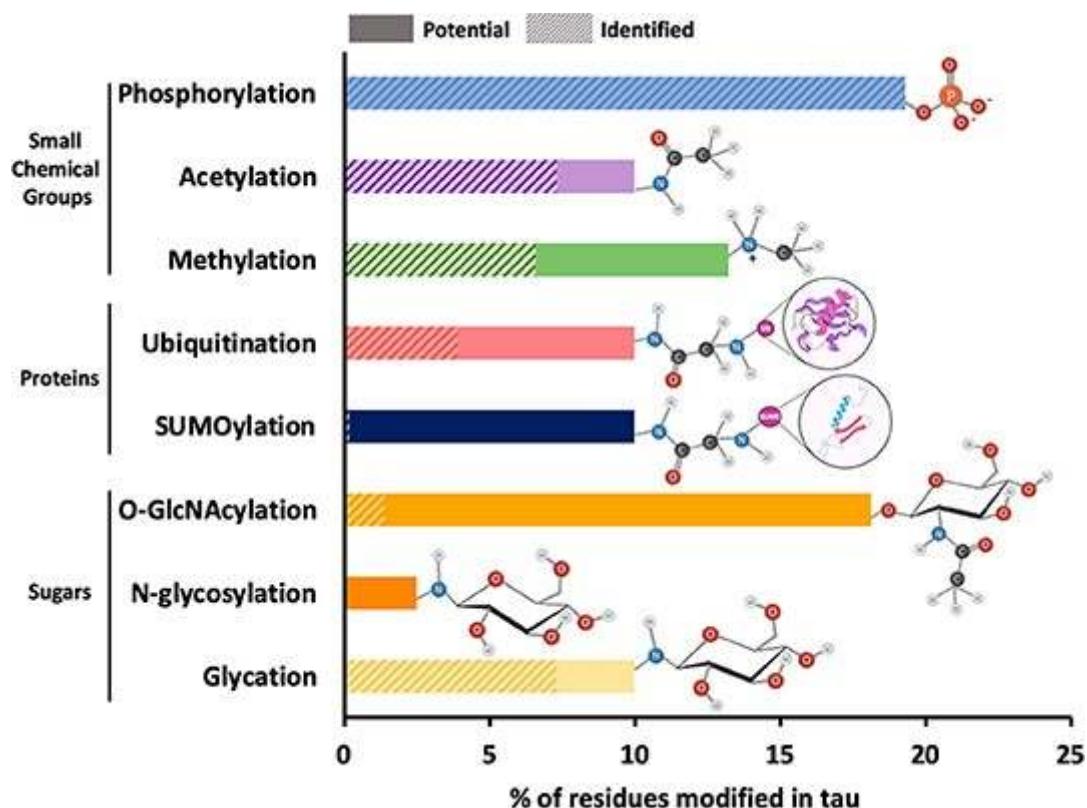


Figure 23: Relative frequency of various PTMs potentially and confirmed to be modified in tau. The contribution of each PTM to the toxicity and dysfunction of tau is a complex and highly active area of research. Figure reproduced from (Alquezar et al., 2021).

### 1.3.8 Amyloid and Tau Staging

AD follows a well characterized and stereotyped spatial/temporal pattern of A $\beta$  and NFT deposition, which have formed the basis of several widely used staging schemes. Shown in Figure 24 are regional distributions of Thal staging for A $\beta$  (Thal et al., 2002) and Braak staging for NFTs (Braak & Braak, 1991), where increasing stage generally corresponds with progression of the disease. The initial involvement of these two aggregates are notably different, with A $\beta$  first appearing across diffuse areas of the neocortex, while NFTs are first found in the entorhinal cortex (EC, seen in more detail in Figure 25). With progression, A $\beta$  plaques rapidly infiltrate the EC and hippocampus, while NFTs spread to

large pyramidal neurons of the CA1 hippocampal subregion and subiculum, the CA3, and later begin to appear in neocortical areas. Toward end stages of the disease, most neo and allocortical areas become involved, with the relative sparing of only the brainstem and cerebellum. Note however the very early appearance of NFTs in the locus coeruleus of the brainstem, the pathological relevance of which remains a matter of debate due to its ubiquity in non-demented individuals (K. Zhu et al., 2019).

**Figure 24.**

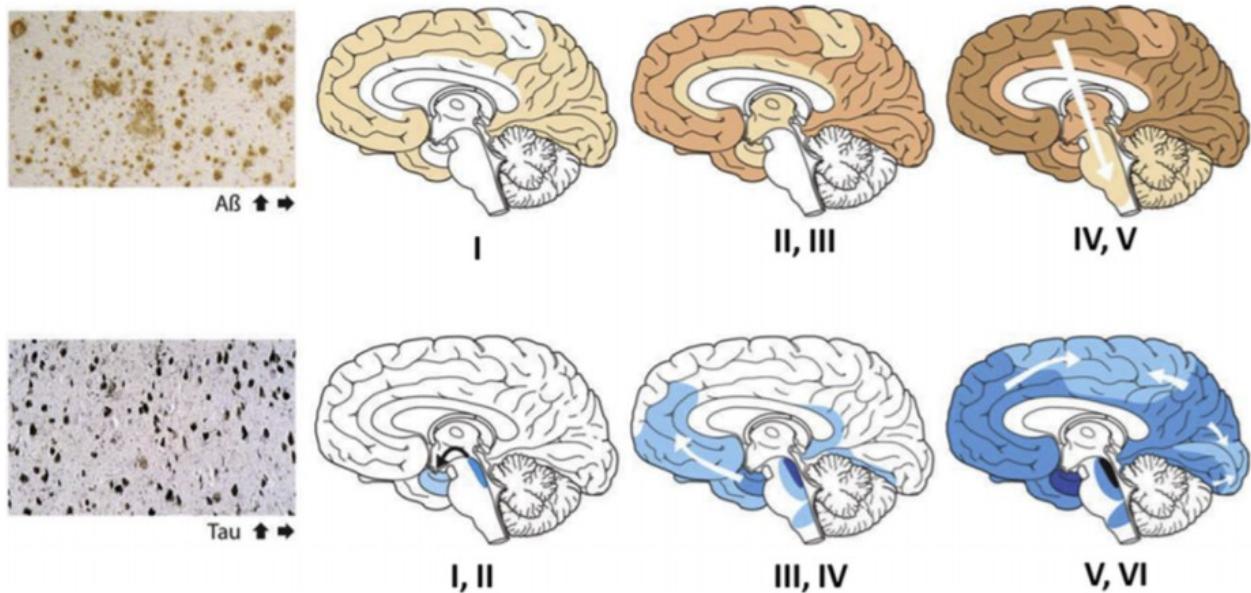


Figure 24: Top, Thal staging for Abeta. Bottom, Braak staging for NFTs. Figure reproduced from (Jouanne et al., 2017).

**Figure 25.**

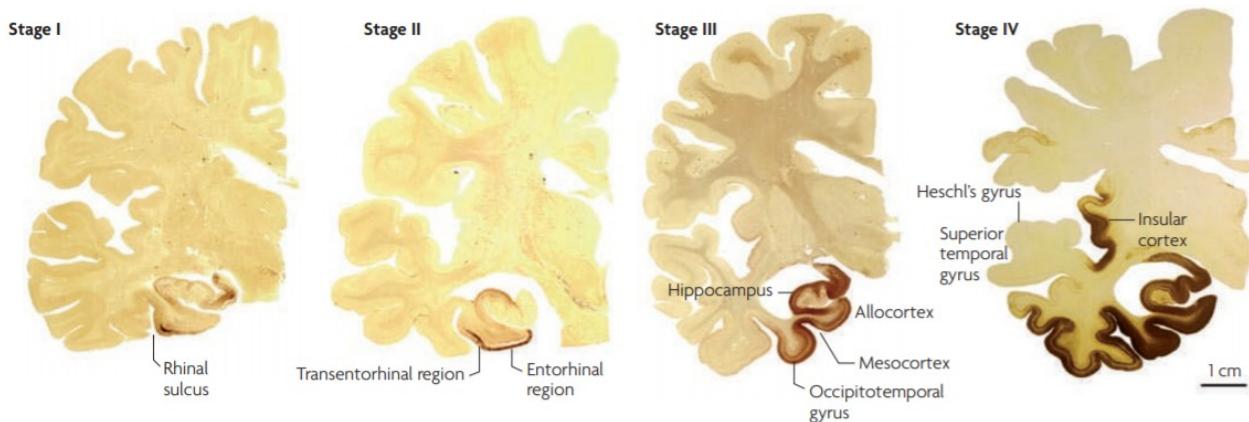


Figure 25: AT8 staining for pathological hyperphosphorylated tau in paraffin-embedded hemisphere sections, demonstrating the initial involvement of areas near the EC followed by involvement of the hippocampus and neocortex. Figure reproduced from (Kretzschmar, 2009).

### 1.3.9 Cell Death and Atrophy

Though no explicit staging system exists for tracking cell death and atrophy in AD, NFT aggregation is often a strong correlate of these endpoints and is a widely used to identify cells that have become directly involved in AD (Del Tredici & Braak, 2020). In Figure 26, there is a robust inverse relationship between the density of pathological tau and neuron count (Furcila et al., 2019). Figure 26 also demonstrates the defining feature of selective vulnerability – that certain regions/cell-types are preferentially affected over others. In this figure, the CA1 exhibits a markedly higher degree of both NFT density and cell loss compared to the CA3, a feature first described in the original Braak staging paper and replicated since (Braak & Braak, 1991; Mrdjen et al., 2019). It may also be the case that long-range and sparsely or unmyelinated axons (Braak et al., 2006) and those that are and that are neurofilament-rich (B. M. Morrison et al., 1998) are particularly vulnerable to AD.

**Figure 26.**

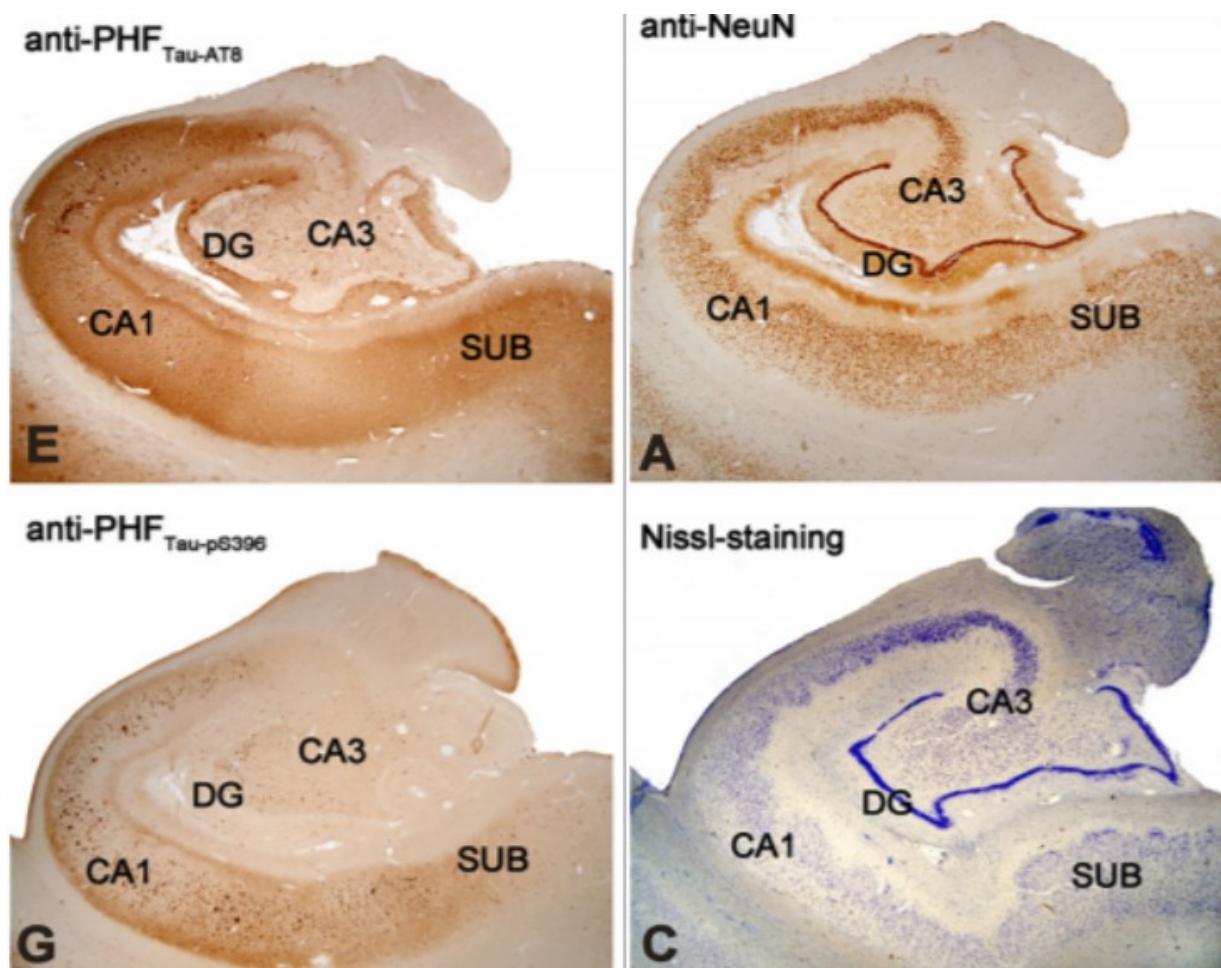


Figure 26: Immunohistochemical staining of hippocampal sections from human AD post-mortem tissue. Left, two stains for different conformations of pathological tau. Right, two stains marking the location of neuronal nuclei (top) and neuronal cell bodies (bottom). Figure adapted from (Furcila et al., 2019).

### 1.3.10 Selective Vulnerability

As reviewed in (Mrdjen et al., 2019), a number of regions/cell-types have been identified as being selectively vulnerable to NFTs in AD (see Figure 27). There is a fairly specific cell-type identified at the site of initial involvement, these being Reelin-expressing EC layer II pyramidal neurons (Chin et al., 2007; Stranahan & Mattson, 2010). High interest in the EC from a disease and functional standpoint have led to fine-grained efforts to map out cell-types in the EC and have begun to reveal the diverse population of subtypes in this region (Kobro-Flatmoen & Witter, 2019). For example, Reelin-expressing layer II neurons in the lateral EC (LEC) are characterised by distinct fan-shaped dendritic morphology while similar neurons in the medial EC (MEC) are more stellate in shape (Witter et al., 2017). Interestingly, recent work has shown that both subtypes locally innervate neurons of the same subtype only sparsely, instead preferring disynaptic inhibitory coupling, where excitation of one neuron indirectly inhibits another neuron through an intermediate inhibitory neuron (Nilssen et al., 2018). This mode of communication may have relevance to the early stages of AD, as A $\beta$  has been shown to perturb normal neural network activity, in particular those related to inhibitory control (Y. Xu et al., 2020).

**Figure 27.**

Brain region	Neuronal subtype	AD stage	Vulnerability to NFTs	Immunoreactivity
Throughout	Excitatory neurons	Throughout	Vulnerable	Glutamate receptors, especially NMDARI; VGLUT1
Entorhinal cortex layer II	Large pyramidal neurons	Early	Vulnerable	Reelin, SMI32, Rasgrp2, Sh3bgrl2
Hippocampus: subiculum	Large pyramidal neurons	Early	Vulnerable	Reelin, SMI32, Cartpt
Hippocampus: CA1	Large pyramidal neurons	Early	Vulnerable	Reelin, SMI32, Somatostatin receptor 4, NOV (CCN3)
Basal forebrain	Cholinergic neurons	Mid	Vulnerable	ChAT
Hippocampus: Dentate hilus	Mossy cells	Mid-late	Inconsistent	CGRP, PCP4
Hippocampus: CA3	Large pyramidal neurons	Late	Inconsistent	Reelin, SMI32, Gprin3, PKC-d
Hippocampus: CA2	Large pyramidal neurons	Late	Inconsistent	Reelin, SMI32, Caeng5, PCP4, IGFBP5, NT3
Hippocampus: Dentate gyrus layers III and V to VI	Granule neurons	Late	Relatively resistant	Prox1, calbindin, PCP4, IGFBP5, NT3
Locus coeruleus	Noradrenergic neurons	Late	Vulnerable	NET
Neocortex: layer III and V	Large pyramidal neurons	Late	Vulnerable	Reelin, SMI32
Primary visual cortex: layers V and VI	Pyramidal neurons	Late	Relatively resistant	Calca
Throughout	Inhibitory neurons	N/A	Relatively resistant	Parvalbumin, somatostatin, calbindin, calretinin; GAD; GABA receptors
Neocortex: layer II, upper III, VI	Small pyramidal neurons	N/A	Moderately resistant	Unknown
Neocortex: layer IV	Smooth stellate neurons	N/A	Relatively resistant	Unknown
Neocortex: layer IV	Spiny stellate neurons	N/A	Relatively resistant	Unknown
Cerebellum	Purkinje cells	N/A	Resistant	PCP4

Figure 27: Summary of a literature review of regions and cell-types shown to be selectively vulnerable in Alzheimer's Disease. Figure reproduced from (Mrdjen et al., 2019).

Although extensive characterization of selectively vulnerable cell-types on the subtype level will be the way forward in unravelling the nature of this phenomena, it is notable that on a broad level, glutamatergic (excitatory) neurons and GABAergic (inhibitory) neurons are generally vulnerable and resilient to NFTs in AD, respectively. Using tissue from human AD donors and the EC-tau mouse model of AD tau pathology (L. Liu et al., 2012), it has been previously shown that the misfolded tau antibody MC1 co-localises with excitatory but not inhibitory neurons, even at advanced Braak stage and age (Fu et al., 2019). Performing co-expression network analysis on publicly available single-cell RNAseq (scRNASeq) datasets, the authors identified that this difference is driven by the aggregation protector *BAG3*, which is more highly expressed in inhibitory vs. excitatory neurons in normal physiological conditions. The authors then validated its protective role by overexpressing the gene in excitatory neurons where they observed an attenuation in tau accumulation, demonstrating how intrinsic differences in gene expression can affect the relative vulnerability of a cell-type.

**Figure 28.**

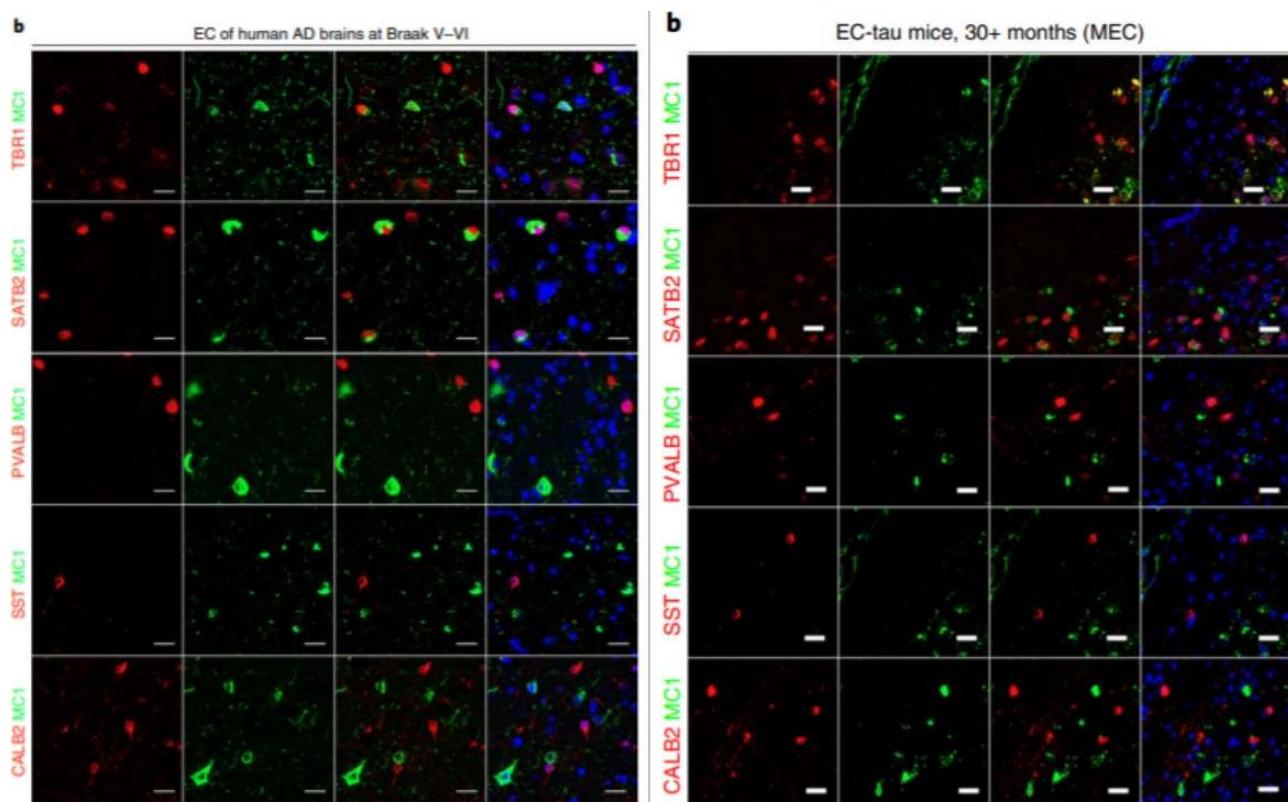


Figure 28: MC1 co-localisation in immunostained sections with excitatory neuron markers (TBR1, SATB2) and inhibitory neuron markers (CALB2, SST, PVALB). Left, human AD brains at Braak V-VI (BA9). Right, EC-tau mice at 30+ months age (MEC). Figure adapted from (Fu et al., 2019).

### 1.3.11 Cell and Non-cell Autonomous Factors

The mechanisms that may underlie selective vulnerability can be divided into two groups: cell-autonomous factors, those that operate independently for each cell, and non-cell-autonomous factors, those that are dependent on the status of other cells. AD has been observed to manifest through the contribution of both factors (Acosta et al., 2018; Z.-T. Wang et al., 2020). Many mechanisms considered to be cell-autonomous fall under the umbrella of homeostatically regulated processes. And in the context of selective vulnerability, researchers have explored cellular differences in the regulation of processes that include oxidative stress, metabolic and energy demands, intracellular calcium levels, excitotoxicity, proteolytic stress and protein folding, inflammatory reactions, unconventional translation, and ageing (Fu et al., 2018; Gan et al., 2018; Muddapu et al., 2020). Non-cell-autonomous processes on the other hand, are primarily associated with the transsynaptic spread of pathological proteins from one cell to another (Vogels et al., 2019), but also includes factors such as vasculature, inflammation, immune response, blood-brain barrier disruptions, and glial biology (Henstridge et al., 2019; Saxena & Caroni, 2011).

### 1.3.12 Circuit Dynamics and Connectivity

Circuit dynamics and anatomical connectivity are another important consideration, particularly when considering the prion-like properties of NFTs. Differentially vulnerable regions innervate one another in a laminar fashion through the trisynaptic loop (Amaral & Witter, 1989) (see Figure 28), wherein EC layer II neurons input into the DG and CA3 via the perforant path. The CA1 is then innervated by the CA3 through Schaffer collaterals, in addition to less prominent input from other sources including the EC (Witter & Moser, 2006). From the CA1, the subiculum is innervated through Alvear fibers. The EC, DG, and CA3 also project within themselves and the DG additionally projects to the CA3 through mossy fibres. It is notable that although the EC does project to the CA1 directly and indirectly, supporting transsynaptic models of tau spread, the also innervated CA3 and DG do not appear to accumulate tangles until later stages of AD. Additionally, there is evidence that the CA3 becomes hyperactive in AD (Haberman et al., 2017), which may be a mechanism for CA1 excitotoxicity. Interestingly, in a normal physiological state, overall CA1 firing rate is higher than that of the CA3 (Mizuseki et al., 2012), but this relationship reverses in ageing (Kanak et al., 2013; Oh et al., 2016).

**Figure 29.**

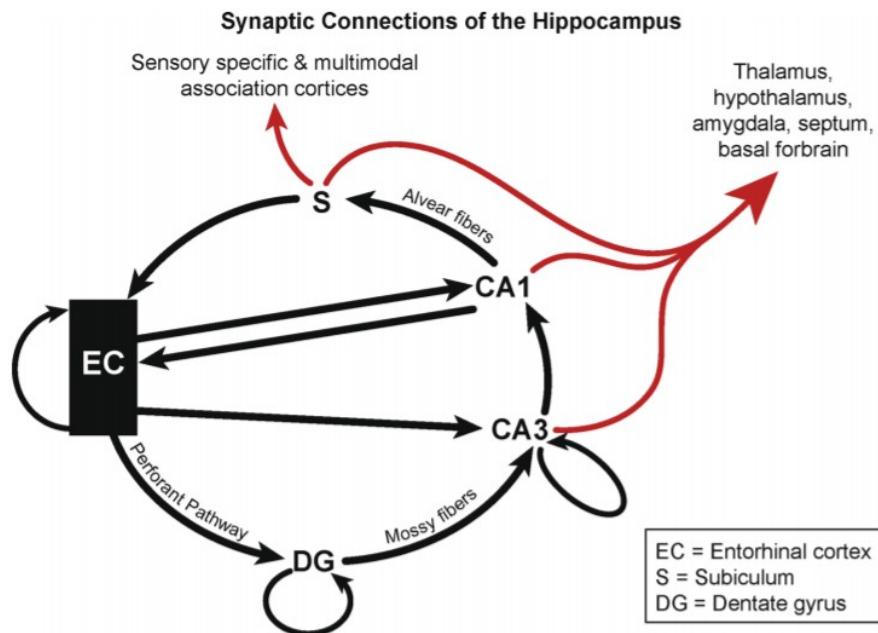


Figure 29: Synaptic connections of the hippocampus and EC, highlighting the trisynaptic loop. The particular connectivity of these regions may underlie the basis of the prion hypothesis and transsynaptic spread of tau in Alzheimer's Disease and other tauopathies. Figure reproduced from (Mrdjen et al., 2019).

### 1.3.13 Cellular Disease Response

Studies in disease states reveal that regions and cell-types that are selectively vulnerable differ in their disease response. For example, in rodent studies using AD models, CA1 neurons, compared to later affected hippocampal areas, exhibit higher ROS and superoxide production (X. Wang et al., 2005; Wilde et al., 2002), a greater abundance of nearby activated glial cells (Rodríguez et al., 2013), differential expression of NMDA receptors including overexpression of the apoptosis-inducing NR2B subunit (Z. Liu et al., 2012), high  $\text{Ca}^{2+}$  influx through L-type voltage gated calcium channels (L-VGCC) (Y. Wang & Mattson, 2014), an upregulation of kinases such as PRKCB and MAPK1 alongside a downregulation of phosphatases 1 and 2 (Gerschutz et al., 2014), and impaired autophagic lysosome function (Bordi et al., 2016). Likewise, in human subjects, MRI studies of individuals with AD and MCI have shown preferential atrophy and blood-brain barrier breakdown in the CA1 (Montagne et al., 2016), and it's been demonstrated that earlier affected cells carry a greater proportion of 3R rather than 4R tau inclusions (Hara et al., 2013; Iseki et al., 2006). Research in ageing and other stressful conditions have also produced valuable insights for investigating selective vulnerability in AD. In rodents, neurons from the CA1, compared to CA3 neurons, show a greater reduction of calbindin during ageing (Potier et al., 1994), and exhibit more severe mitochondrial damage post-ischemia (Radenovic et al., 2011) and from calcium-induced mitochondrial swelling (Mattiasson et al., 2003). Similarly, the CA1 undergoes greater calcium influx and calcium

deposition into mitochondrial from prolonged glutamate stimulation (Stanika et al., 2010). These specific examples are hardly exhaustive of the totality of research in this area, and aim to merely highlight the wide breath of findings that remain challenging to contextualise.

### 1.3.14 Studies of Neurofibrillary Tangle-bearing Neurons

A few studies using human AD post-mortem tissue have attempted to dissect individual neurons containing NFTs for further study using transcriptomics and proteomics, the approach of interest in this thesis work. The earliest found instance of this kind of work dates from (Ginsberg et al., 2000), where hippocampal sections underwent immunostaining to distinguish CA1 neurons containing tangles from those free of pathology for isolation and subsequent RNA amplification. NFT-bearing neurons were detected using the PHF1 antibody, whereas neurons without tangles were identified by staining for non-phosphorylated neurofilament proteins using the RMdO20 antibody. Large-scale cDNA GDA arrays were probed using radiolabelled RNA derived from single-cell isolates, comprising 20 NFT-bearing and 20 normal CA1 neurons. These neurons originated from five AD donors and five control donors, with samples pooled in groups of four neurons per array.

Among their findings, in CA1 neurons with NFTs, there was significant downregulation of mRNAs encoding protein phosphatase subunits, including subunits of PP1 and PP2A. Although mRNA expression for many tau-associated kinases such as CAM kinase, CDK2, and CDK5 remained unchanged, reductions were noted in the mRNA levels for GSK-3 $\beta$ , ERK1, and ERK2. Similar decreases were observed for cytoskeletal proteins, including all neurofilament subunits and  $\beta$ -tubulin, whereas  $\beta$ -actin, microtubule-associated proteins (MAP2, MAP1B), and tau isoforms remained unaffected. Additionally, NFTs were associated with decreased expression of mRNAs for proteins involved in synaptic transmission, notably the AMPA receptor subunits GluR1 and GluR2, as well as the NMDA receptor subunit NR2B. Several presynaptic vesicle related proteins, including synaptophysin, synaptotagmin, and synuclein also exhibited reduced mRNA expression in NFT-bearing CA1 neurons.

It was several years before a similarly designed study by (Dunckley et al., 2006) was carried out to study NFT-bearing neurons in AD. Like the proteomics dataset generated as part of this thesis work, the authors employed laser-capture microdissection (LCM), a technique that precisely isolates individual cells or cell populations from heterogeneous tissues using targeted laser ablation under microscopic visualisation. They used this method to selectively extract neurons containing neurofibrillary tangles, as well as histopathologically unaffected neurons, from layer II stellate cell islands in the entorhinal cortex of 19 individuals diagnosed with AD. RNA was subsequently obtained from these isolated neurons for use in microarray analyses. To uncover genes linked to NFT pathology, the authors conducted permutational paired t-tests, directly comparing matched samples of NFT-bearing and non-NFT neurons from each AD patient.

**Figure 30.**

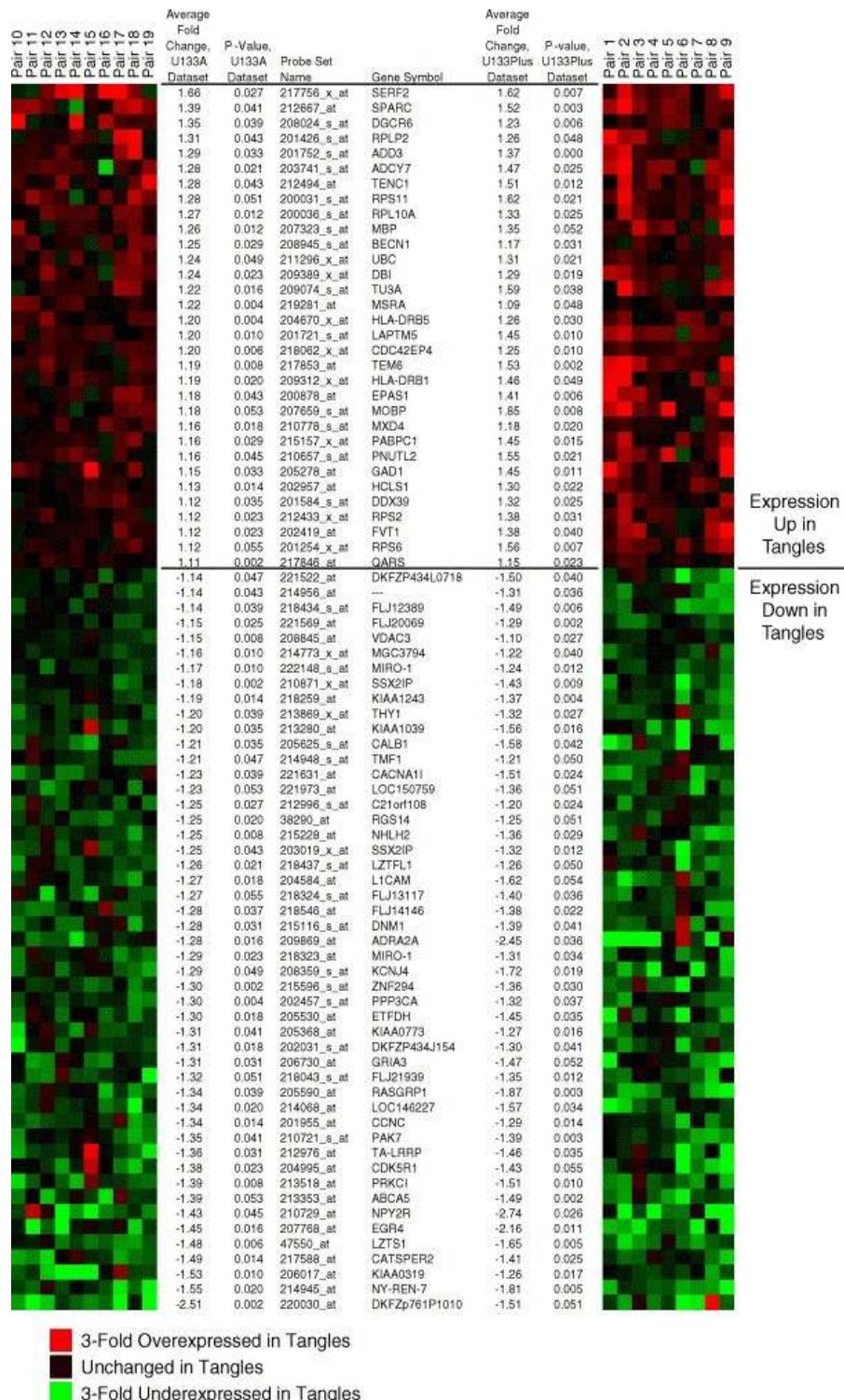


Figure 30: Heatmaps showing genes identified as significantly dysregulated, comparing matched pairs of neurons containing neurofibrillary tangles with those histopathologically unaffected. Each column represents an individual matched pair, with the leftmost heat map displaying data from the primary analysis performed using Affymetrix U133A arrays, and the right heat map representing corresponding findings from the subsequent validation dataset. Figure reproduced from (Dunckley et al., 2006).

The authors further validated initial findings from the expression analyses using immunohistochemistry, selecting genes based on their established roles in AD pathology or relevance to implicated biological processes. Specifically they validated upregulation in NFT-bearing neurons for the genes Apolipoprotein J (APOJ), casein kinase 2, beta polypeptide (CSNK2B), tissue inhibitor of metalloproteinase 3 (TIMP3), interleukin-1 receptor-associated kinase 1 (IRAK1), CD44 antigen, and member of the RAS oncogene superfamily RAP2A. Although of far lesser focus, they also validated downregulation of a few candidates, Calpain 7 (CAPN7) and p21-activated kinase 7 (PAK7). The authors conclude that the identified genes appear to be involved in a broad array of cellular processes, suggesting a diverse molecular landscape underlying Alzheimer's Disease pathology, and noted the difficulty in formulating an integrated model of NFT pathogenesis.

Studies of a similar kind experienced a dearth in research activity and no further work isolating tangle-bearing neurons for transcriptomic and proteomic analysis could be found until a pivotal study available as a preprint in 2020 and formally published in 2022 (Otero-Garcia et al., 2022), which is reanalysed as part of this thesis work. Here, the authors establish a novel protocol for high-throughput isolation using fluorescence-activated cell sorting sorting (FACS) followed by RNA sequencing of single neuronal somas from human AD post-mortem cortical tissue. Using the pathological hyperphosphorylated tau antibody AT8, NFT-bearing and non-NFT-bearing neurons were isolated within each of 8 donors, and analysis was performed to identify both shared and cell-type-specific molecular signatures associated with NFT-bearing neurons. Specific details of the methodology and associated dataset is explained in-depth in Section 2.1.

The study identified neuronal subtypes particularly susceptible to NFT pathology, notably the superficial-layer neurons expressing CUX2 and deeper-layer neurons expressing RORB and PCP4. Immunohistochemical validation confirmed these findings.

Transcriptomic analysis revealed substantial upregulation of genes associated with synaptic functions, including CALM1, ATP1B1, GRIN2B, CDK5R1, SYT4, CANX, and RTN4, as well as those linked to cytoskeletal structure and microtubule dynamics such as ACTG1, TUBB2A, PLPPR4, MAP1A, ENC1, and STMN2. Additionally, consistent upregulation was observed for stress-response genes, including JUN and ATF4, as well as the chaperone HSP90AA1, lysosomal protein PSAP, and genes related to iron metabolism, FTL and FTH1. Notably, APP and PRNP, implicated in amyloid pathology, were upregulated across multiple neuronal clusters, though PRNP expression showed variability.

Further analysis aimed at uncovering transcriptional regulatory mechanisms identified shared regulatory networks across affected neuronal subtypes, prominently involving REST, a transcription factor previously linked to neuronal function, ageing, and Alzheimer's Disease pathology. Commonly enriched functional pathways included synaptic transmission, calcium signalling, microtubule assembly and transport, axonal and dendritic structural remodelling, and intracellular trafficking pathways (Figure 31). Interestingly,

pathways directly related to apoptosis and neuronal cell death were less prominently enriched and showed a balanced representation of pro and anti-apoptotic regulators, including downregulation of FAIM2 and MIF and upregulation of ATF4, BAD, BNIP3, and HIF1A. Additionally, genes implicated in mitochondrial permeability transition were found upregulated, notably BAD, BNIP3, HSPA1A, and multiple members of the 14-3-3 family of phospho-serine/threonine-binding proteins (YWHAE, YWHAH, YWHAG, YWHAZ, YWHAB).

**Figure 31.**

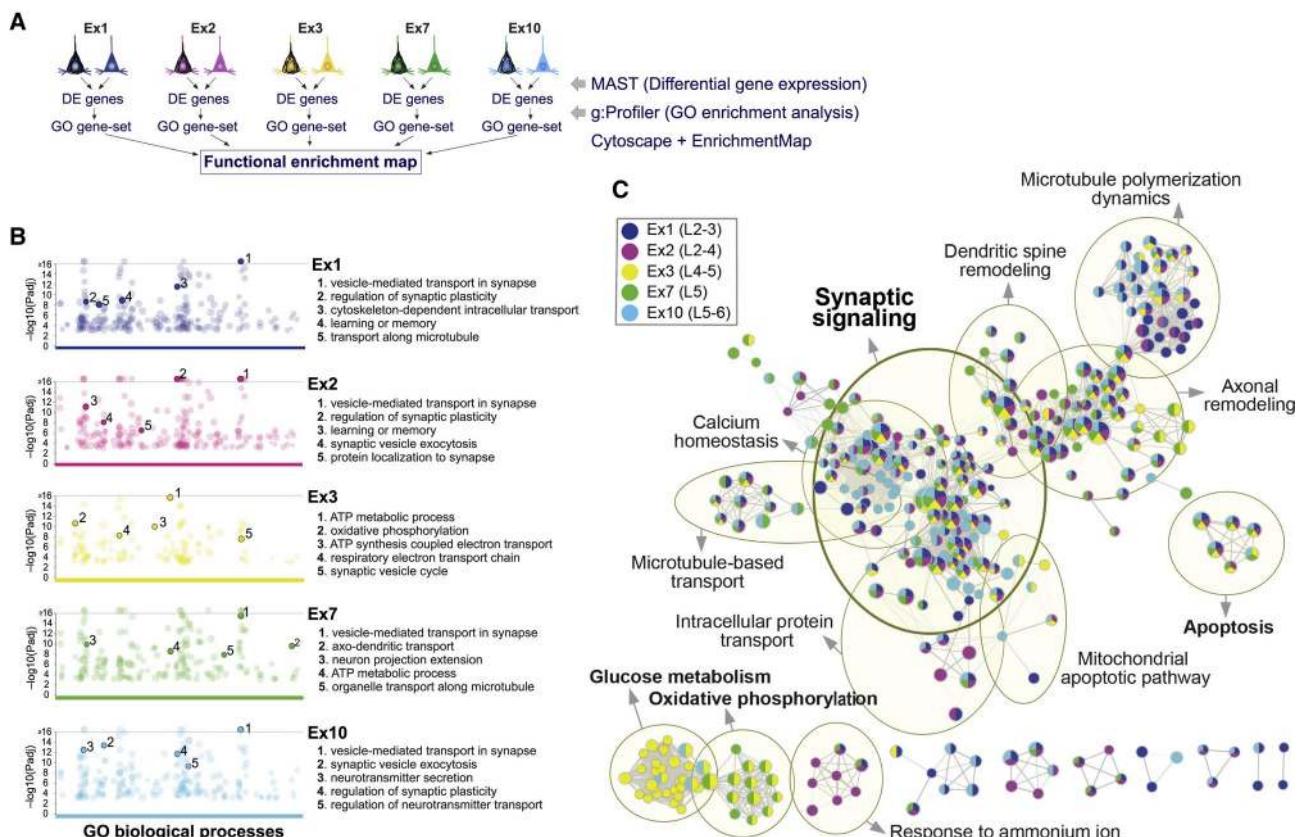


Figure 31: Shared and distinct pathways altered in terms of transcriptome across several excitatory neuron cell-types found to harbour NFTs in the prefrontal cortex of AD patient donors. A) summarises the approach taken to identify pathways that are either broadly shared or uniquely enriched in neuronal subtypes most vulnerable to NFTs, focusing on five clusters characterised by high cell counts and pronounced NFT accumulation. B) illustrates the Gene Ontology (GO) biological processes significantly enriched when comparing NFT-containing neurons to those free of NFT pathology within each cluster, visualised as Manhattan plots generated by g:profiler. C) provides a functional enrichment map in which each node represents a GO biological process, colour-coded according to neuronal subtype cluster, to highlight both overlapping and subtype-specific functional contributions, with related gene sets grouped by thematic clusters. Figure reproduced from (Otero-Garcia et al., 2022).

Around the same time as the original preprint of (Otero-Garcia et al., 2022) was the publication of a study that used LCM with mass spectrometry to analyse tangle-bearing neurons, similarly to the in-house dataset of this thesis work, and possibly the first of its kind to perform this experiment to assay the proteome (Hondius et al., 2021). In this study, neurons with GVD (granulovacuolar degeneration) or NFTs were separately isolated from post-mortem hippocampal tissue of AD patients ( $n = 12$ ), using laser-capture microdissection guided by immunohistochemical markers casein kinase (CK)1 $\delta$  for GVD and phosphorylated tau (AT8) for NFTs. Additionally, control neurons were similarly isolated from cognitively normal subjects ( $n = 12$ ). Proteomic profiling via label-free LC-MS/MS was then conducted on these neuron populations, of which 115 proteins showed significantly altered abundances in GVD-containing neurons, whereas 197 proteins were differentially expressed in NFT-bearing neurons compared to controls. Functional categorisation of the significantly altered proteins in GVD neurons indicated involvement in several key biological processes, notably protein folding, proteasomal degradation, endolysosomal pathways, cytoskeletal integrity, RNA processing, and glycolytic metabolism. Interestingly, NFT-bearing neurons shared many of these functional alterations but additionally exhibited pronounced reductions in ribosomal proteins and further disruptions in protein folding-related factors.

The current body of research examining NFT pathology and selective neuronal vulnerability in Alzheimer's Disease remains relatively sparse, highlighting a critical gap in our understanding of the disease's underlying mechanisms. Existing studies underscore the complexity and variety of molecular, cellular, and anatomical factors influencing why certain neurons succumb to NFT pathology while others remain resilient. Expanding this understanding will necessitate rigorous replication of current findings across diverse patient cohorts to validate the reproducibility and generalisability of observed effects. Such goal represents the motivation for the current thesis work, which takes a comprehensive multi-modal approach by integrating the high-quality transcriptomic data by (Otero-Garcia et al., 2022) with an in-house LCM proteomics dataset. While the work of (Hondius et al., 2021) presents another LCM proteomics dataset, with a similar aim, of crucial difference is their decision to use control non-tangle-bearing neurons from a separate non-demented donor population. Unlike the other studies in this section, this does not facilitate a within-donor analysis, and therefore the effects of tangle-bearing pathology is necessarily confounded with the general effects of AD. As a result, our group elected to create our own dataset, focused more specifically on NFTs solely and not GVD or other pathologies, and with a within-donor experimental design like the dataset from (Otero-Garcia et al., 2022) that it would be integrated with.

## 2. Methods

This thesis focuses on two datasets, one publicly available, and another generated in house. The publicly available dataset was published by (Otero-Garcia et al., 2022), where the authors established techniques for high-throughput isolation and transcriptomic analysis of individual neuronal somas containing NFTs from human AD brains. The other uses laser capture microdissection combined with mass spectrometry for a proteomic analysis of NFTs from human AD brains. In both datasets, neurofibrillary tangles were identified with the same antibody, mouse monoclonal AT8 (anti-phospho-Tau [Ser202, Thr205], ThermoFisher Cat# MN1020). AT8 recognises tau when phosphorylated at Ser202 and Thr205, is not isoform-specific (binds 3R and 4R tau when the epitope is phosphorylated), and is widely used to detect pretangles and tangles in Alzheimer's disease tissue. Each dataset was processed with a custom computational pipeline. Conventional procedures are detailed in this Methods section while analysis-specific developments are presented throughout Sections 3-5.

### 2.1 FACS-sorted Single-soma RNA Sequencing

The method published by (Otero-Garcia et al., 2022) introduced a novel protocol for high-throughput fluorescence-activated cell sorting (FACS) and RNA sequencing of individual neuronal somas containing NFTs from human AD brains. Their approach enabled the quantification of NFT susceptibility and neuronal loss across 20 distinct neocortical subtypes, revealing both common and cell-type-specific molecular signatures. A schematic overview of the approach taken is visualised in Figure 32.

**Figure 32.**

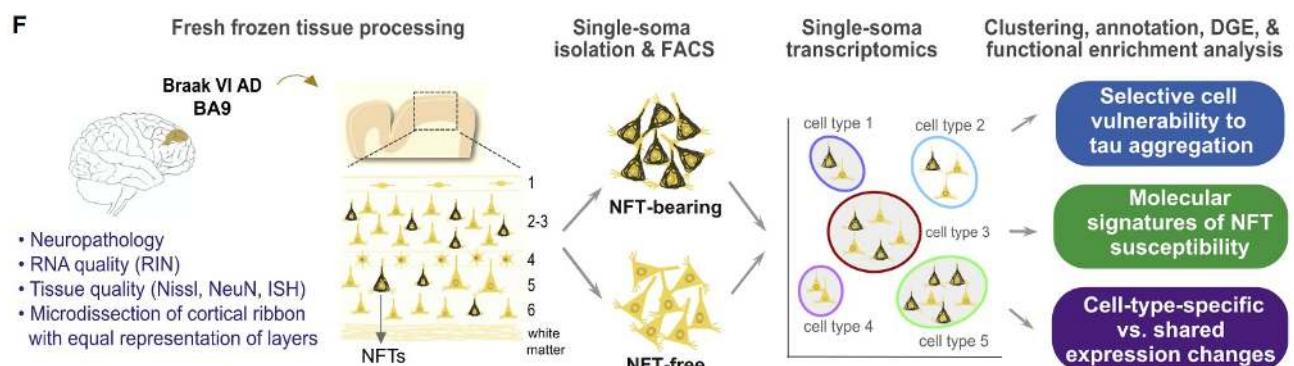


Figure 32: Schematic of the processing approach of NFT-bearing and non-NFT-bearing neurons derived within AD patient post-mortem cortical tissue. The authors used FACS with the AT8 antibody to separate the neuron populations, which then underwent a modified single-soma transcriptomics pipeline. The authors analysed the data with several aims, including selective vulnerability to tau aggregation, molecular signatures of NFT susceptibility, and cell-type-specific and shared expression changes in the neurons. Figure adapted from (Otero-Garcia et al., 2022).

The protocol described by the authors is summarised as follows. Fresh-frozen human brain tissue samples were first sourced from publicly available repositories (UCLA-Easton Center, the NIH Neurobiobank, and the Stanford Alzheimer Disease Research Center). The AD cohort for single-soma transcriptomic analysis included eight donors (five females, three males) who had been diagnosed with dementia and met the criteria for AD neuropathology. All AD cases had a Braak stage of VI/VI and an ABC score of A3B3C3. Donor ages ranged from 66 to 93 years, with a mean of  $76.9 \pm 12.4$  years. The postmortem interval (PMI) varied between 1 and 33 hours, with averages of  $12.8 \pm 7.6$  hours. RNA integrity number (RIN) values ranged from 5.7 to 7.8, with mean scores of  $6.5 \pm 0.4$  for AD samples.

**Figure 33.**

Donor ID #	Age	Sex	PMI (hr)	RIN	Brain weight	Brain region	AD stage (NIA-AA)	Braak stage	Other pathology
1	93	F	4	6.5	1,150	BA9	A3B3C3	VI	CVD
2	79	F	19.5	7.0	1,270	BA9	A3B3C3	VI	none
3	81	M	16	6.5	1,360	BA9	A3B3C3	VI	none
4	57	M	14	6.8	1,160	BA9	A3B3C3	VI	none
5	81	F	24	6.1	960	BA9	A3B3C3	VI	none
6	73	F	13	5.7	1,300	BA9	A3B3C3	VI	none
7	89	F	1	6.5	1,130	BA9	A3B3C3	VI	CVD
8	62	M	11	6.6	1,250	BA9	A3B3C3	VI	none

Figure 33: Sample profile of the AD donors included in this study. All NFT vs. non-NFT neurons were sampled within the same subject. Figure adapted from (Otero-Garcia et al., 2022).

Brain tissue blocks were brought from -80°C to -12°C to facilitate the dissection of thick sections (approximately 500 µm). For each experiment, a cortical section weighing around 200 mg was extracted. White matter and leptomeninges were removed, and the remaining tissue was finely chopped into fragments smaller than 1 mm<sup>3</sup> using a razor blade. RNA quality was evaluated from a 10 mg sample using the RNeasy kit (Qiagen, Cat#74134), following the manufacturer's protocol. The RNA integrity number (RIN) was measured using an Agilent Bioanalyzer 2100 RNA Nano chip (Agilent Technologies, Cat#5067-1511), following the manufacturer's protocol. To minimize RNA degradation during soma isolation, all processing steps were carried out on ice under RNase-free conditions.

Tissue homogenization was performed using a Potter-Elvehjem tissue grinder, which features a wider clearance (0.1–0.15 mm) between the pestle and tube compared to grinders commonly used for nuclear dissociation. This design allows for more effective dissociation of intact somas while minimizing mechanical damage. Each tissue sample was processed in 2.4 mL of homogenization buffer and 0.2 U/mL RNase inhibitor. The resulting homogenate was filtered through a 100-µm cell strainer and then divided into two 1.5-mL Eppendorf tubes for further processing.

The sample underwent further purification using iodixanol gradient centrifugation. The resulting supernatant was aspirated and discarded, while the pellet was gently resuspended in 200 µL of chilled homogenization buffer. All homogenates were then pooled into a single tube, and the total volume was measured and adjusted to 450 µL with homogenization buffer. To achieve a final iodixanol concentration of 21%, an equal volume of 42% iodixanol medium was added and mixed via pipetting. This mixture was layered onto 900 µL of pre-chilled 25% iodixanol medium in a 2-mL Eppendorf tube. The gradient was then centrifuged at 8,000 × g for 15 minutes at 4°C, causing intact neuronal somas to sediment at the bottom, while the upper layers contained myelin and cellular debris. The top fraction and supernatant were removed to avoid contamination of the soma-enriched pellet.

To resuspend the pellet, 50 µL of immunostaining buffer and 0.2 U/mL RNase inhibitor was added, followed by transfer into a fresh tube. The sample was then resuspended to a final volume of 200 µL in immunostaining buffer and incubated at 4°C for 15 minutes with gentle rocking. Primary antibodies were then introduced, including a mouse monoclonal anti-phospho-Tau (Ser202, Thr205) antibody (AT8, 1:150, ThermoFisher Cat#MN1020) and a rabbit polyclonal anti-MAP2 antibody (1:40, MilliporeSigma Cat#AB5622), and the suspension was incubated for an additional 40 minutes at 4°C. Following primary antibody incubation, 500 µL of immunostaining buffer was added, and the samples were mixed before being centrifuged at 400 × g for 5 minutes at 4°C. The supernatant was removed, and the pellet was resuspended in 600 µL of immunostaining buffer. For secondary staining, Alexa Fluor-conjugated antibodies were added, including goat anti-mouse Alexa Fluor 350 (1:500) and goat anti-rabbit Alexa Fluor 647 (1:500), along with a nuclear stain (SYTOX Green, 1:40,000). The solution was incubated at 4°C for 30 minutes with gentle rocking.

High-quality samples were characterised by a suspension primarily composed of individual neuronal somas and free nuclei, with minimal aggregation and cellular debris. The proportion of cells retaining well-preserved somas ranged from 20% to 50% of the total sample. On average, processing 100 mg of cerebral cortex tissue yielded approximately 0.5 to 1.5 million intact somas. FACS was utilised to isolate individual neuronal somas containing tau aggregates. Sorting was conducted using either a BD FACSAria II or a Sony SH800S, equipped with four excitation lasers (488 nm, 405 nm, 638 nm, and 561 nm). Single soma suspensions were isolated from Brodmann area 9 (BA9) of Braak stage VI AD donors ( $n = 4$ ). To isolate and profile neurons containing pathological tau aggregates, AT8 was chosen as the primary marker. Two populations of the populations were used for my re-analysis, neuronal somas with tau aggregates (MAP2<sup>+</sup>/AT8<sup>+</sup>) from AD brains, and adjacent neurons without detectable tau pathology (MAP2<sup>+</sup>/AT8<sup>-</sup>) from the same AD samples.

FACS gating was established using a series of sequentially applied parameters to accurately identify and isolate target cell populations. Side scatter (SSC) area was plotted against SYTOX Green to distinguish intact cells from debris and dead cells. Alexa Fluor

350 was plotted against Alexa Fluor 647 to differentiate subpopulations based on tau pathology and neuronal identity. The parameters were adjusted to exclude the smallest particles and large aggregates, ensuring that only intact neuronal somas were analysed. To minimize background signal and reduce the risk of false positives from non-specific binding or autofluorescence, a series of controls were included, consisting of unstained cells, samples treated with only secondary antibodies, and those labelled with individual primary antibodies. Sample acquisition was also kept below 30% droplet occupancy. The number of somas recovered ranged from 1,600 to 37,000 for AT8<sup>+</sup> neurons and over 300,000 for MAP2<sup>+</sup> neurons.

**Figure 34.**

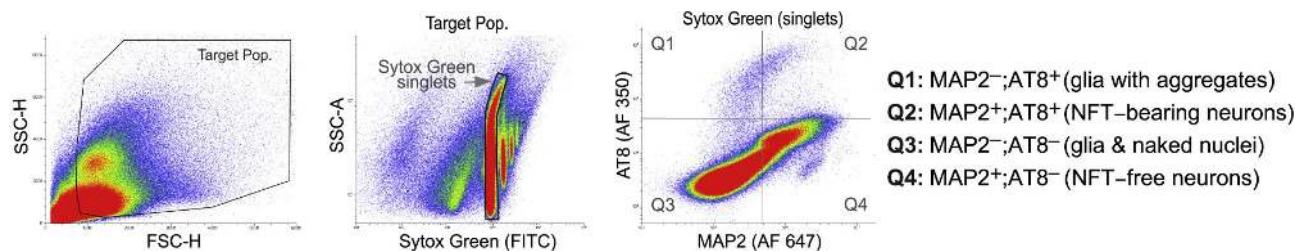


Figure 34: Example of the FACS gating approach used to separate neuronal populations based on fluorescence of Sytox Green and the MAP2 and AT8 antibodies. Figure adapted from (Otero-Garcia et al., 2022).

Single-soma mRNA capture and library preparation was conducted using the 10x Genomics Chromium Single Cell 3' v2 or v3 platforms. Cell counts were determined using a hemocytometer, and cell integrity was assessed via fluorescence microscopy. The number of cells loaded per experiment ranged from 1,400 to 11,000, with an upper limit of 5,000 cells per sample. Subsequent steps followed the manufacturer's protocols. The number of PCR cycles used for cDNA amplification ranged from 13 to 15. For library construction, the sample index PCR cycles were set between 12 and 13, adjusted based on the quantified cDNA input. Paired-end libraries were sequenced using the Illumina NovaSeq 6000 platform. Libraries derived from AD donors were pooled and processed together within a single sequencing run. Sample concentrations were normalised based on the total number of cells to ensure uniform read distribution across all samples. Each cell was sequenced at an average depth of 72,000 reads, achieving an approximate sequencing saturation of 85%. Paired-end sequencing data were processed using the Cell Ranger software suite (version 3.1) from 10x Genomics. The Cell Ranger count pipeline was employed with default settings to align reads to the prebuilt GRCh38 reference genome, as well as to perform quality control steps, including filtering, barcode identification, and unique molecular identifier (UMI) quantification.

**Figure 35.**

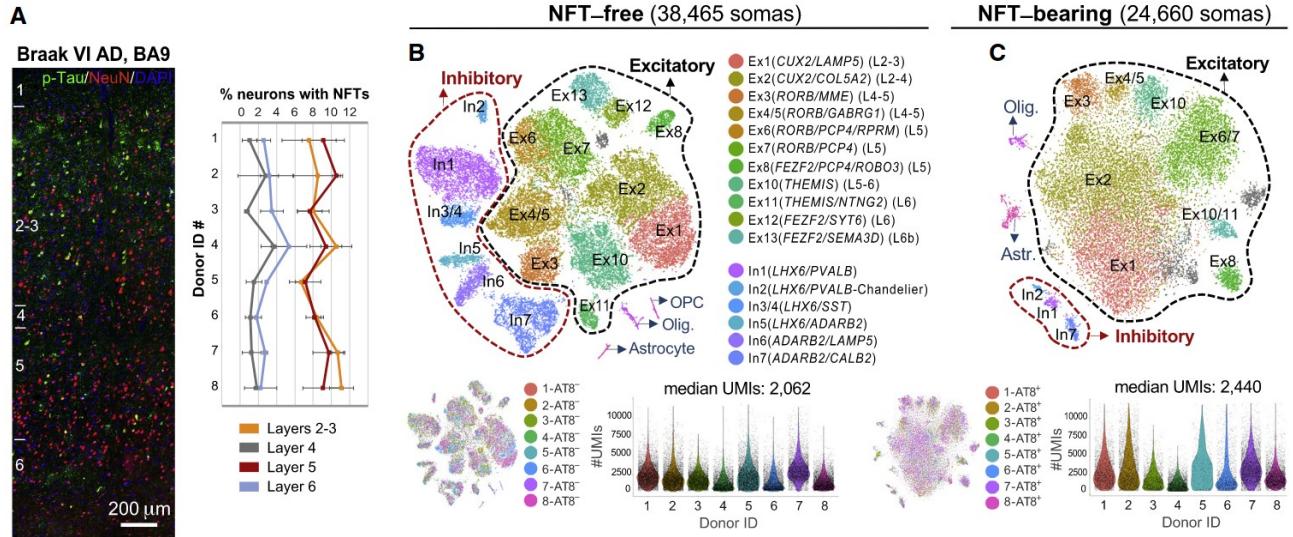


Figure 35: Outcome of the protocol employed by the authors which obtained 38,465 NFT-free somas and 24,660 NFT-bearing somas. The authors demonstrate that each population had a similar number of median UMLs, which were of sufficient quality for analysis. Both populations also reveal a variety of subcell-types within the neuronal populations, of which separation of clusters was finer-grained in the NFT-free population. Figure adapted from (Otero-Garcia et al., 2022).

## 2.2 Laser-capture Microdissection Mass Spectrometry

In addition to the publicly available dataset, a dataset was created in-house to supplement the transcriptional changes with proteomics. In order to separate tangle-bearing neurons we used laser capture microdissection (LCM), a method that uses a precise laser to cut out cells or areas of tissue from sections of previously stained tissue. Using human post-mortem AD brain tissue from the BA9 region, the tissue was stained with AT8 (phosphorylated tau) to identify tangle-bearing neurons and counterstained with cresyl-violet, aimed at showing the morphology of neurons. Individual AT8-positive or cresyl-violet-positive (AT8-negative) neurons were dissected from sections that were 8 µm in thickness. While this thickness does not exceed the diameter of a neuronal soma, the possibility of capturing material from adjacent cells in the z-plane cannot be fully excluded due to the nature of the technique. However, proteomic profiling indicated an enrichment for neuronal proteins, with the majority of proteins identified as mapping to neuronal populations when analysed in tandem with the single-soma FACS RNA-sequencing dataset. Although cells were selected on a single-cell basis, due to the detection limits of mass spectrometry, a minimum of 300 cells of the same type and from the same donor were pooled together, to ensure adequate starting material. Label-free proteomic analysis was then performed using the ultrasensitive timsTOF pro mass spectrometry, through a collaboration with Raja Nirujogi in Dundee University. The data acquisition was carried out by Dr Martha Foiani (part of the Duff lab) whereas I performed the analysis. Ten pathologically confirmed Alzheimer's Disease cases were acquired from Queen Square Brain Bank, UCL. AD cases met current diagnostic criteria (Braak & Braak, 1991; McKhann et al., 2011; Montine et al., 2012; Thal et al., 2002), all reaching a final ABC score of A3B3C3.

**Figure 36.**

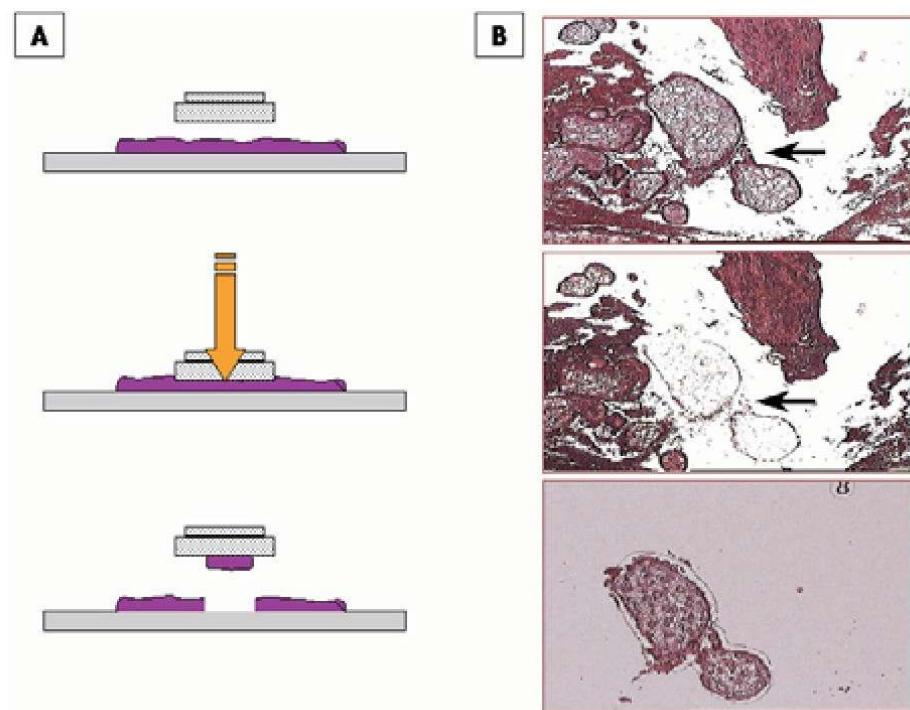


Figure 36: Generalised diagram of the LCM workflow. A tissue section is placed on a glass slide, and an LCM cap with a transfer film is carefully aligned over the target region. A laser pulse is then applied, selectively activating the transfer film, causing cells within the designated area to adhere to the cap. As the cap is lifted, the captured cells detach from the slide, while the remaining tissue section remains intact. Figure reproduced from (Budimlija et al., 2005).

Serial sections of formalin-fixed paraffin embedded 8 µm thickness were cut using a microtome and mounted onto glass slides. Slides were stained for AT8 and counterstained for cresyl-violet (Nissl staining), in order to separate cortical layers in the BA9 region of the cortex. Sections were deparaffinised by immersion in three changes of xylene and rehydrated with three steps of absolute alcohol, for 3 minute each. Endogenous peroxide activity was inhibited using methanol and 0.3% H<sub>2</sub>O<sub>2</sub> for 10 minutes, followed by tissue immersion in PBS for at least 10 minutes. Sections underwent antigen retrieval, which was performed by placing slides for 10 minutes in citrate buffer (0.45 g citric acid, 5.8 g tri-sodium citrate, 2 litres deionised H<sub>2</sub>O, pH 6.0) and heated in a microwave. Non-specific protein binding was blocked by submerging slides in 10% milk in PBS (0.5 M pH 7.2) for 30 minutes at RT. Following a washing step, tissue sections were incubated in AT8-biotinylated antibody at 4°C overnight. Following several steps of washing, sections were incubated in avidin-biotin complex (ABC, DAKO) incubation for 30 minutes. Binding of the antibodies was visualised by submerging slides in 3,3-Di-aminobenzidine (DAB, Sigma) activated by H<sub>2</sub>O<sub>2</sub> (500 µg/100 mL PBS). Neuronal cell morphology was obtained through cresyl-violet acetate counterstain (Sigma).

**Figure 37.**

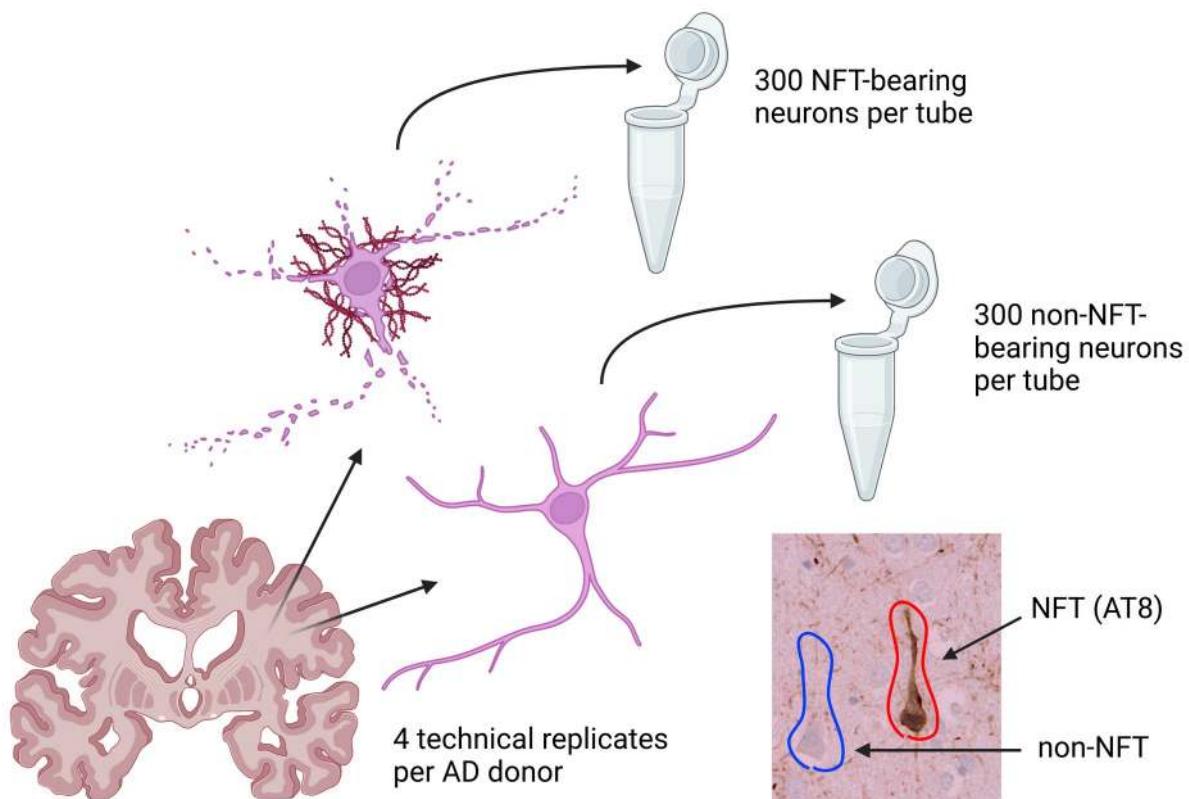


Figure 37: Schematic overview of the sampling procedure. Each cap pooled together 300 neurons from the same donor and of the same population (tangle-bearing vs. non-tangle-bearing), with 4 technical replicates per donor, across 10 AD donors. Created with BioRender.com.

Following the completion of staining, slides were air-dried before being promptly processed for laser microdissection. Laser capture microdissection was carried out using a Leica DM6000B laser capture microdissection microscope. To minimise contamination from adjacent cells, cutting outlines were carefully traced around each individual neuron, separating two distinct populations: those with AT8 staining and those without. Cresyl violet allows for the identification of layer II neurons in the cortex, so careful attention was put into capturing only layer II tau-bearing and non-tau-bearing-neurons, as they are the earliest affected neurons in AD. 300 AT8 positive/negative cells (4 technical replicates and 10 biological replicates) were cut and collected in the lids of separate 0.2 mL Eppendorf tubes containing 10 $\mu$ L of RIPA buffer for protein digestion. Samples were briefly spun down (20 seconds) at maximum speed and an additional 10 $\mu$ L of RIPA were added to the tube. Samples were stored at -80°C until shipment to Scotland.

Samples were processed in Dundee according to an optimised protocol, based on (Hughes et al., 2019). Samples were solubilised in 2% (m/v) SDS buffer for an unbiased protein retrieval followed by high power Bioruptor-based sonication with 15 cycles each cycle with 30 sec ON and 30 sec OFF. Samples were reduced by adding 5mM DTT incubated at 56°C for 30 min and alkylated by adding 20mM Iodoacetamide incubated in dark for 30 min. Further, SP3 (single-pot solid-phase- enhanced sample preparation) workflow was employed, allowing protein purification, removal of SDS and on-bead trypsin digestion, prior to LC-MS/MS analysis. EvoSep LC system with disposable trap columns were used to circumvent carry-over between samples. MS data was acquired with the highly sensitive timsTOF pro mass spectrometer by employing dia-PASEF (Data Independent Acquisition Parallel Accumulation and Serial Fragmentation) strategy to achieve near 100% duty cycle, which increased the sensitivity and coverage of proteomes derived from lower number of neuronal cells (Meier et al., 2018). Raw MS data was processed for database searches using DIA-NN 1.8 version as library free direct DIA strategy by allowing default settings against human Uniprot database.

**Figure 38.**

Case #	Braak stage	Sex	Age at death	Age at symptom onset	PMI (h)	Fixed hemisphere	Pathological diagnosis
1	VI	F	63	48	20	Left	AD
2	VI	F	63	59	24	Left	AD
3	VI	F	64	64	32	Left	AD
4	VI	F	65	56	49	Left	AD
5	VI	F	68	52	13	Left	AD
6	VI	M	68	52	9	Left	AD
7	VI	M	69	65	20	Left	AD
8	VI	M	69	52	12	Left	AD
9	VI	M	69	58	7	Left	AD
10	VI	M	86	72	10	Left	AD

Clinical diagnosis	APOE status	TDP43 limbic	Thal	B&B tau	CERAD	A	B	C	ABC	AD level
AD	33	No	5	6	3	3	3	3	A3B3C3	High
CBS due to AD	#N/A	No	5	6	3	3	3	3	A3B3C3	High
AD	#N/A	No	5	6	3	3	3	3	A3B3C3	High
AD	#N/A	No	5	6	3	3	3	3	A3B3C3	High
Pick's disease or frontal variant of FTD, bvFTD	34	No	5	6	3	3	3	3	A3B3C3	High
Sporadic young onset AD	#N/A	No	5	6	3	3	3	3	A3B3C3	High
CBS	#N/A	No		5	3	3	3	3	A3B3C3	High
AD	33	No	5	6	3	3	3	3	A3B3C3	High
PPA (FTD)	#N/A	No	5	6	3	3	3	3	A3B3C3	High
nfvPPA (FTD)	#N/A	No	5	5	3	3	3	3	A3B3C3	High

Figure 38: Table of clinical data of cases. All cases were acquired from Queen Square Brain Bank, UCL.

## 2.3 Analysis Organisation and Reproducibility

### 2.3.1 Data Hosting and Version Control

Both datasets underwent custom computational analysis pipelines based on best practices and tailored to suit the dataset at hand. Each pipeline is documented with reproducible instructions at the following links: <https://github.com/eturkes/otero-garcia-2022-ssRNAseq>, <https://github.com/eturkes/NFT-LCM-N8>. In order to ensure reproducibility and provenance of the analysis, all code was closely version controlled using Git. Git is a distributed version control system designed to track changes in code, facilitate collaboration, and manage software development projects efficiently. Github was used to host the code, which is a cloud-based platform that provides hosting for Git repositories, enabling developers to collaborate on projects, track changes, and manage version control. Because Github does not support hosting of large files, all raw data and results are stored in Dropbox at the following link:  
<https://www.dropbox.com/scl/f0/dx1xmpdvq8paam9uirte3/AMUozK4aieT9fLPaCxHQsj0?rlkey=k1f74ght8299gz87pxjbuc3sb&st=uqccic97&dl=0>. Finally, all code was licensed under the GNU General License Version 3 (GPLv3), a widely used open-source software license that ensures users have the freedom to run, modify, and distribute software.

### 2.3.2 Docker and Singularity

In order to enhance reproducibility of the analyses, Docker and Singularity was used to control the analysis environments. Docker is a container system available on Linux systems, harnessing the power of cgroups to create isolated namespaces (Merkel, 2014). This allows for lightweight virtualisation where full Linux distributions can be quickly be created and destroyed. These containers are usually used in an ephemeral and read-only manner, where there are guarantees in place that the environment is unchanged regardless of the host system. Using a Dockerfile, programming language versions and packages can be explicitly defined, for example R 4.3.3 in this analysis, running in an Ubuntu 24.04 environment. Because Docker requires root privileges on the host machine, a script to run the environment under Singularity is also provided. Singularity is an alternative container management system that can run Docker containers without root privilages, for that reason it has become popular in the high-performance computing (HPC) space (Kurtzer et al., 2017). After building the Docker and Singularity containers locally, they were uploaded to Docker Hub and Singularity Hub. These sites allow for the easy download of pre-built containers, ensuring users run the same environment as originally used in the analysis. Instructions for pulling these pre-built images are provided in the Github README of each analysis.

**Figure 39.**

The screenshot shows a GitHub page for a Dockerfile. At the top, it says "otero-garcia-2022-ssRNaseq / Dockerfile". Below that is a navigation bar with "Code", "Blame", "56 Lines (53 loc) · 2.23 KB", and "Code 55% faster with GitHub Copilot". To the right are "Raw", "Edit", "Download", and other icons. The main content is a text-based Dockerfile with numbered lines from 18 to 56. The file starts with `FROM rocker/rstudio:4.3.3` and includes various `RUN` commands for package installation, including R packages like `stringr`, `conflicted`, `Seurat`, `VIRIDIS`, `DT`, `hd5r`, `flexdashboard`, `networkR3`, `BioManager`, `remotes`, and `praise`. It also includes system package installations like `libglpk0`, `zlib1g-dev`, `patch`, `liblzma-dev`, `libbz2-dev`, and `Rscript` with specific arguments. The file ends with `apt-get clean` and `rm -rf` commands to remove temporary files.

```
18
19 FROM rocker/rstudio:4.3.3
20
21 LABEL org.opencontainers.image.authors="Enir Turkes enir.turkes@eturkes.com"
22
23 RUN apt-get update \
24     && apt-get install -y --no-install-recommends \
25         libglpk0 \
26         zlib1g-dev \
27         patch \
28         liblzma-dev \
29         libbz2-dev \
30
31 && Rscript -e "install.packages('rmarkdown')" \
32     -e "install.packages('stringr')" \
33     -e "install.packages('conflicted')" \
34     -e "install.packages('Seurat')" \
35     -e "install.packages('VIRIDIS')" \
36     -e "install.packages('DT')" \
37     -e "install.packages('hd5r')" \
38     -e "install.packages('flexdashboard')" \
39     -e "install.packages('networkR3')" \
40     -e "install.packages('BioManager')" \
41     -e "install.packages('remotes')" \
42     -e "BioManager::install('glimmpsp')" \
43     -e "BioManager::install('GSEAbase')" \
44     -e "BioManager::install('scuttle')" \
45     -e "BioManager::install('edgeR')" \
46     -e "BioManager::install('OSVA')" \
47     -e "BioManager::install('biomart')" \
48     -e "BioManager::install('IMF')" \
49     -e "BioManager::install('ComplexHeatmap')" \
50     -e "BioManager::install('DropletUtils')" \
51     -e "BioManager::install('scater')" \
52     -e "remotes::install_github('immunogenomics/praise')" \
53
54 && apt-get clean \
55 && rm -rf /var/lib/apt/lists/ \
56     /tmp/downloaded_packages/ \
57     /tmp/.rds
```

Figure 39: Screenshot of the Dockerfile hosted on the Github of one of the analyses. The Dockerfile builds off of a publicly available container from Docker Hub housing R version 4.3.3 and specifies the installation of several operating system and R packages.

### 2.3.3 R Markdown

R Markdown was used for organisation and documentation of all code, alongside a small utilities file in plain R (R Core Team, 2022). R Markdown is among the recent data analysis friendly formats such as Jupyter Notebooks that allow documentation and image visualisation directly alongside code (Allaire et al., 2022; Xie et al., 2018, 2020). R Markdown supports a wide variety of outputs including Word documents, PDF, and HTML documents. I opted to used HTML to harness interactive features when viewed in a web browser. The R Markdown files were organised in a modular and iterative manner. For example, the first R Markdown file to be run within a directory is suffixed with *01*, to imply that its run precedes others. Because some calculations are computationally expensive, the process also uses on-disk caching. A helper script is provided to run all files in sequence, alongside another command to spin up a Dockerised R Studio environment for interactive analysis. These can be run locally, or on a remote server using a reverse SSH proxy (instructions provided in Github repository).

**Figure 40.**

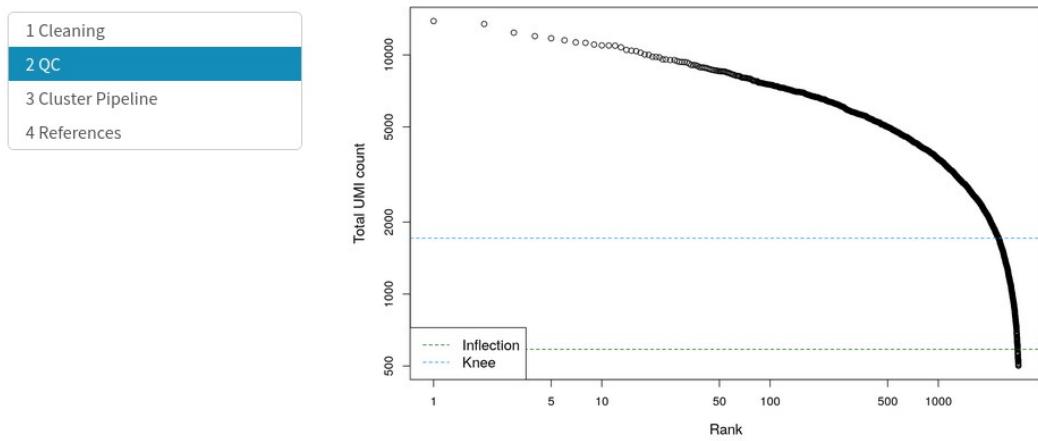
```

19 # This script runs all files in the analysis.
20 # Sections can be commented out as needed.
21
22 setwd(dirname(parent.frame(2)$ofile)) # Move to location of this file.
23
24 xfun::Rscript_call(
25   rmarkdown::render,
26   list(
27     file.path("individual", "1_MAP2_01prep.Rmd"),
28     output_file = file.path(
29       "..", "..", "results", "individual", "1_MAP2_01prep.html"
30     ),
31     envir = new.env()
32   )
33 )
34 xfun::Rscript_call(
35   rmarkdown::render,
36   list(
37     file.path("individual", "2_MAP2_01prep.Rmd"),
38     output_file = file.path(
39       "..", "..", "results", "individual", "2_MAP2_01prep.html"
40     ),
41     envir = new.env()
42   )
43 )

```

Figure 40: Excerpt of the *run\_all* script written in R for one of the analyses, demonstrating the modular nature of the R Markdown documents.

**Figure 41.**



We remove cells low in unique features and total counts and calculate percentage of mitochondrial and ribosomal reads and other genes commonly used as indicators of low-quality cells. We also use mitochondrial reads as a proxy for ambient RNA contamination.

```
mito <- grep("MT-", gene_anno$symbol)
datatable_download(gene_anno[mito, ])
```

HIDE

Copy Print Download Show 10 entries Search:

symbol	ensembl	
geneids33653	MT-ND1	ENSG00000198888
geneids33654	MT-ND2	ENSG00000198763
geneids33655	MT-CO1	ENSG00000198804
geneids33656	MT-CO2	ENSG00000198712
geneids33657	MT-ATP8	ENSG00000228253

Figure 41: Excerpt of an R Markdown file used in one of the analyses. R Markdown allows for the incorporation of code directly in a document which can support plain text, images, and interactive elements through a Javascript backend. After writing an R Markdown file, it can be compiled, producing a standalone HTML file that can be easily distributed. Precompiled R Markdown documents are provided for all analyses in the Dropbox link.

**Figure 42.**

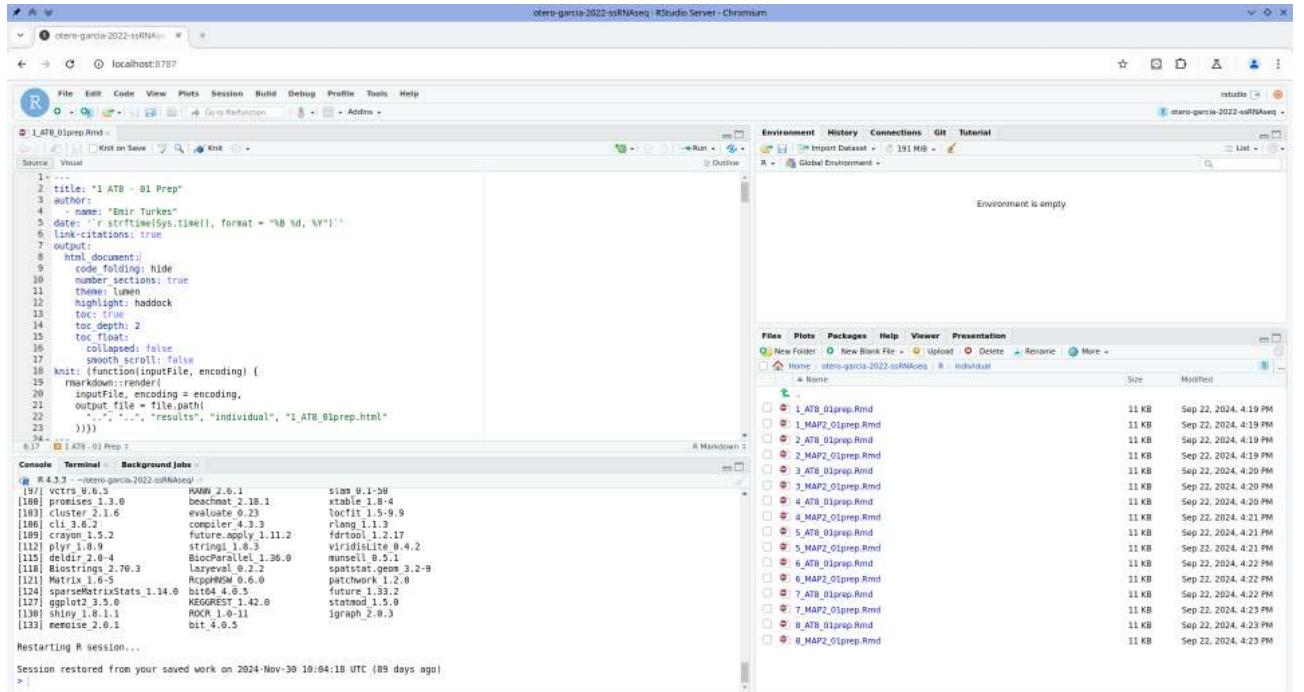


Figure 42: RStudio Server environment provided with the analysis. Using the Docker/Singularity images in the Github repositories, users can download an identical R Studio environment as that used in the analysis, allowing for easy inspection of the data and replication of results. From here, the R Markdown files can be recompiled if desired.

### 2.3.4 Gene / Protein Identifiers

Finally, in all analyses, ENSEMBL IDs were preferred over gene symbols for all internal processing where possible. ENSEMBL IDs provide the benefit of unique unchanging identifiers whereas symbols, while more easily human-interpretable, can be obsoleted or ambiguous. After internal processing, ENSEMBL IDs were converted to their closest gene symbol for visualisations and other results. Interchange between the identifiers was done as needed to facilitate the analysis.

## 2.4 Single-cell RNA Sequencing Preprocessing

### 2.4.1 Data Cleaning

The authors of (Otero-Garcia et al., 2022) provide a preprocessed Seurat object, which is a standard format for R-based single-cell transcriptomics analysis (Butler et al., 2018; Hao et al., 2021, 2024; Satija et al., 2015; Stuart et al., 2019). These objects however held data far downstream in their processing pipeline. For greater control over the analysis, I opted to work with the Cell Ranger h5 files available at the article's GEO accession location (GSE129308). Cell Ranger h5 files are Hierarchical Data Format (HDF5) files generated by 10x Genomics Cell Ranger pipeline, used for storing processed scRNAseq data (Satpathy et al., 2019). These files contain structured information such as gene expression matrices, barcodes, and UMI counts. Commonly, two types of h5 files are produced: *filtered\_feature\_bc\_matrix.h5*, which contains high-confidence cell-associated data, and *raw\_feature\_bc\_matrix.h5*, which includes all droplets, including those likely devoid of any cells. The h5 files were provided in a granular manner, with a unique file for each donor by condition (for example Donor 1, AT8 positive). My pipeline was written to operate on each h5 file individually, with an R Markdown file for each h5 file. The initial step of each analysis was to load the h5 file into a fresh Seurat object.

### 2.4.2 Quality Control

A distinctive feature of droplet-based data is the absence of prior information about whether a given library (represented by a cell barcode) originates from a droplet containing a cell or from an empty droplet. Consequently, distinguishing genuine cells from empty droplets must rely solely on observed expression profiles. This task is challenging because even empty droplets can capture ambient RNA from the extracellular environment, leading to detectable sequencing reads and non-zero expression counts in libraries that lack actual cellular content. The filtering AKA cellcalling step of Cell Ranger, which produces *filtered\_feature\_bc\_matrix.h5*, aims to differentiate cell-containing droplets from empty droplets by implementing the EmptyDrops algorithm described in (Griffiths et al., 2018; Lun et al., 2019). In this algorithm, ambient RNA background is modelled as a multinomial distribution derived from the ambient gene expression profile. Considering barcodes with a greater than 500 UMI (transcript) count, the selected barcodes are statistically evaluated against the ambient RNA background model. Those whose RNA expression significantly differs from this background are subsequently classified as genuine cells.

I opted to use the filtered matrix produced by the Cell Ranger cell-calling algorithm after careful inspection of its output. The general approach to confirming successful cell-calling is to visualise the sample with a barcode rank plot, implemented into the pipeline using the *barcodeRanks* function from the *DropletUtils* package in R (Griffiths et al., 2018; Lun et al., 2019), which contains the reference implementation of EmptyDrops. A barcode rank plot of a library after successful removal of empty droplets is characterised by a lack of cells below a certain total UMI count, generally around 500 especially when using Cell Ranger with default parameters. When plotting total UMI against the rank of each barcode sorted

by UMI count, this should furthermore produce a visually smooth curve, implying that the number of cells increase in a predictable fashion with reduction of total UMI count. This curve is predicted by the modelling of the cell-to-UMI relationship as a negative binomial distribution, the prevailing model in the scRNAseq field (W. Chen et al., 2018).

**Figure 43.**

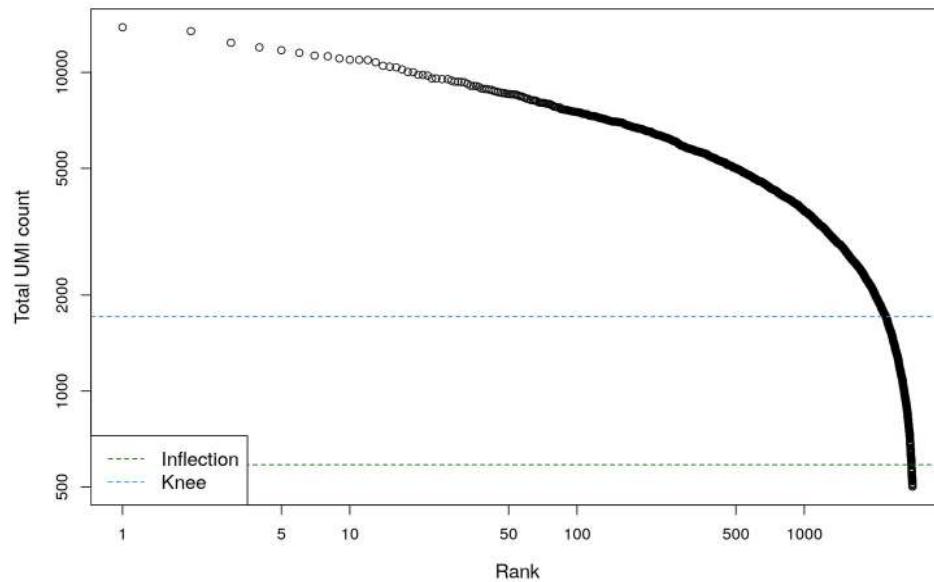


Figure 43: A barcode rank plot of a random sample in the dataset. On the x-axis is the rank of each barcode, where a larger number indicates lower total UMI count. On the y-axis is total UMI count of each barcode. A lack of barcodes below a certain UMI count, and a smooth curve showing a steady increase in cells with lower total UMI implies that no further cell-calling is needed.

**Figure 44.**

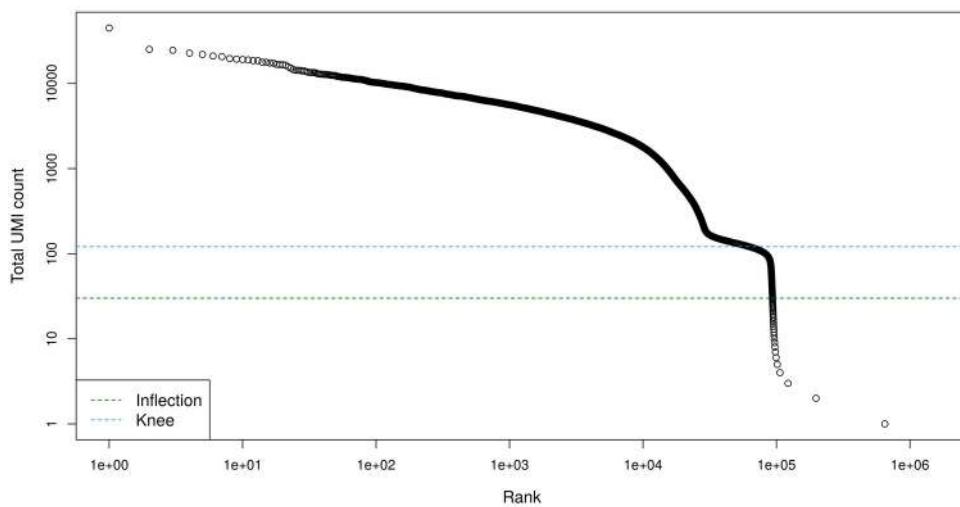


Figure 44: A barcode rank plot showing a random sample from an unrelated dataset where cell-calling had not been run. The x and y axes follow those in Figure 43. It can be observed that shape of the curve differs substantially from Figure 43, indicating that cell-calling is needed. Shown is an approximate inflection point preceding a characteristic “knee”, that suggests an appropriate total UMI count cutoff for effective cell-calling.

After determining that all samples underwent cell-calling by Cell Ranger successfully, I next aimed to perform routine quality control on the number of unique features (genes), total UMI counts, and percentages of mitochondrial, ribosomal, and *MALAT1* counts, all of which are commonly used as indicators of low-quality cells (Luecken & Theis, 2019; Osorio & Cai, 2021). In brief, the number of unique features helps distinguish robust transcriptomes from cells with poor RNA capture or degradation. Total UMI counts provide an overall measure of sequencing depth per cell, ensuring sufficient detection sensitivity. Mitochondrial read percentages act as an indicator of cell stress or apoptosis, as damaged cells often exhibit disproportionately high mitochondrial RNA content. Similarly, ribosomal RNA percentages can signal contamination or technical artefacts, while *MALAT1* count levels can help identify cells with nuclear retention bias.

**Figure 45.**

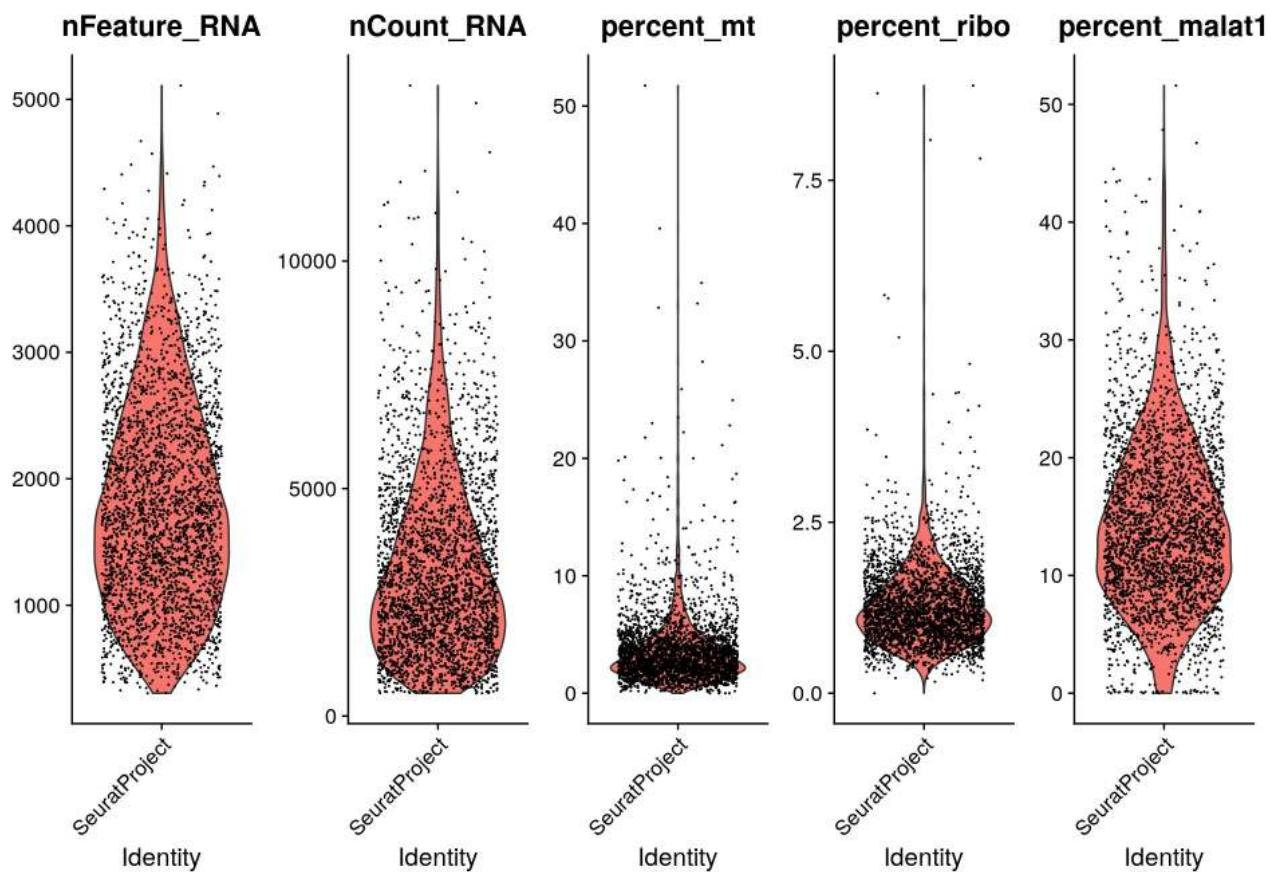


Figure 45: Violin plots showing various QC metrics on a random sample from the dataset before any thresholding or removal. Each point represents a cell, with the y-axis showing the cells' number of counts or percentage of counts, as determined by the plot title.

Although I obtain these QC metrics at this stage, I withhold any thresholding until after doublet removal. This is because scDblFinder, the doublet removal method of choice in this analysis, operates under the assumption that empty droplets have been removed, but further QC had not been performed yet (Germain et al., 2022). Likewise, certain approaches to QC can become skewed if calculations take place after doublet removal. The clearest case of this is when using an outlier approach like median absolute deviation (MAD) for number of unique features and total counts. This type of filtering is commonly used to remove a small number of outlier cells at the upper extreme of these metrics. The rationale is that these cells, while not necessarily low-quality, are not essential to a population-level understanding of the dataset (assuming they do not constitute a real distinct cell-type population of sufficient quantity). However, they can have considerable detrimental effects to the quality of steps such as normalisation and dimensionality reduction (Luecken & Theis, 2019). Because doublet removal primarily operates on cells in these upper extremes, thresholds for statistics such as MAD can substantially shift, leading to the adverse effect of removal of larger numbers of cells that may normally be considered informative.

Regarding the assumptions of scDblFinder, as per the vignette of the R package: “the input to scDblFinder should not include empty droplets, and it might be necessary to remove cells with a very low coverage (e.g. <200 or 500 reads) to avoid errors. Further quality filtering should be performed downstream of doublet detection, for two reasons: 1. the default expected doublet rate is calculated on the basis of the cells given, and if you excluded a lot of cells as low quality, scDblFinder might think that the doublet rate should be lower than it is. 2. kicking out all low quality cells first might hamper our ability to detect doublets that are formed by the combination of a good quality cell with a low-quality one” (<https://bioconductor.org/packages/release/bioc/vignettes/scDblFinder/inst/doc/scDblFinder.html>). The key operating principle underlying scDblFinder is the simulation of artificial doublets generated from the input data, which is used to train a classifier on both the original and artificial data. There are a variety of doublet removal methods that elect similar or different approaches, and scDblFinder was selected based on its strong performance in its reference publication, where it was benchmarked against most competing methods at the time.

scDblFinder was run with default parameters, producing a vector marking predicted doublets in each sample. To visualise the distribution of doublets, I marked doublets on a UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) created on each sample. UMAP is a dimensionality reduction technique that preserves the local and global structure of high-dimensional data, making it particularly useful for visualising complex datasets such as scRNAseq and machine learning embeddings (Becht et al., 2018; McInnes et al., 2020). The details for generating the UMAP will be discussing in the proceeding paragraphs, as it is ran again after all QC steps are completed.

**Figure 46.**

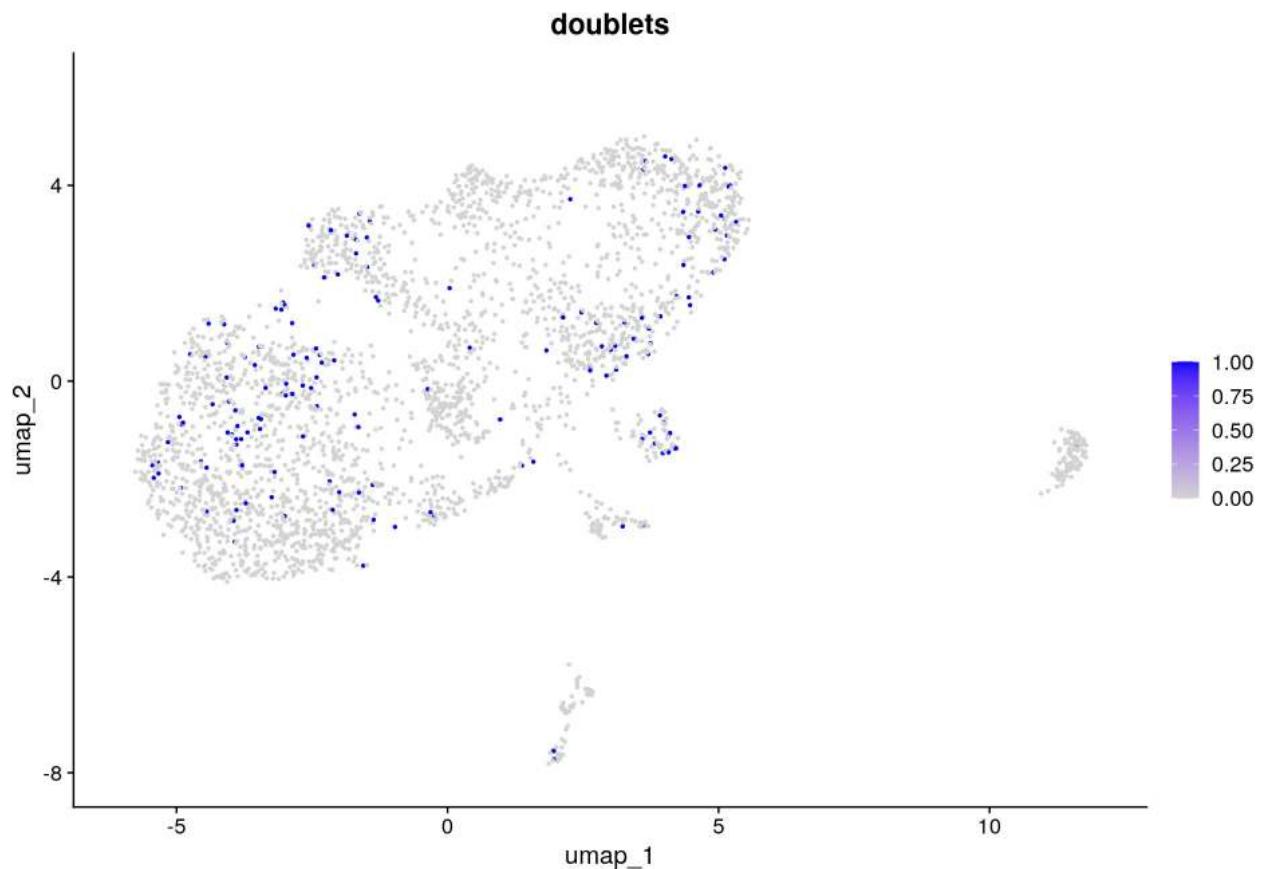


Figure 46: A UMAP projection of a random sample showing the distribution of doublets as identified by scDblFinder. In this particular sample, 159 doublets (5.3% of all cells) were called.

While the distribution of doublets in each sample did not seem to single out a potential real cell-type that might be of interest (in other words, doublets were evenly distributed rather than clustered), this approach may have been a bit conservative. According to the scDblFinder authors and 10X Genomics documentation, 10X Chromium droplet data, which was the basis of the protocol used in this dataset, should produce doublets on the rate of 0.8% per 1,000 cells. Figure 46 shows a sample with 2,995 cells, which by that calculation should then have a doublet rate of about 2.37%. In that sample, scDblFinder marked 5.3% of cells as doublets, more than double the expected rate. This was the case generally across all samples. I choose to go forward with this doublet removal, noting that it is conservative but not biased towards specific clusters and thus satisfactory for further analysis. It is possible that further optimisation may yield a doublet detection rate closer to the expected proportions, although it is also possible that this dataset simply contains more doublets than a typical dataset. This assumption is not unreasonable considering that the dataset was produced using a modified protocol combining FACS sorting to separate tangle-bearing and non-tangle-bearing neurons.

After doublet removal, the QC metrics (number of unique features, number of total counts, percentage mitochondrial, ribosomal, and *MALAT1* counts) were re-assessed (but not recalculated), and a decision was made regarding QC thresholding. Instrumental to this process was the creation of density plots on the metrics that help show abnormal peaks at range extremes that may indicate that filtering is needed. This is a common approach in many scRNAseq QC pipelines (for example [https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-scrna-seq/Seurat\\_QC\\_to\\_Clustering/](https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-scrna-seq/Seurat_QC_to_Clustering/) or [https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionIV/lessons/SC\\_quality\\_control\\_analysis.html](https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionIV/lessons/SC_quality_control_analysis.html)). Figure 47 shows for instance, the number of unique features per cell in a randomly selected sample. One should note the contrast with Figure 48, which is the same metric in an unrelated dataset. The major difference between the two is the number of peaks; for the data in this analysis, a single smoothed peak is observed at the mid-to-upper range of number of features, whereas the unrelated dataset has multiple peaks and most importantly, its largest peak by far is in the lower range of number of features. This information suggests that the thesis dataset (Figure 47) may not require further QC on this metric, while the unrelated dataset (Figure 48) may benefit from it, specifically a threshold on lower values around the vertical line shown.

**Figure 47.**

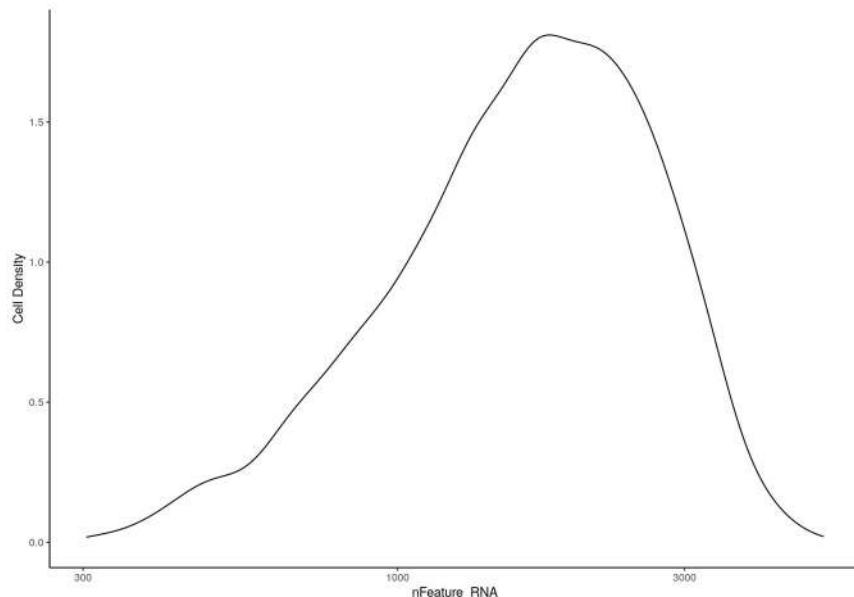


Figure 47: QC density plot on a random sample from the FACS ssRNAseq dataset. On the x-axis is the number of unique features per cell. On the y-axis is the density of cells at each x-axis value.

**Figure 48.**

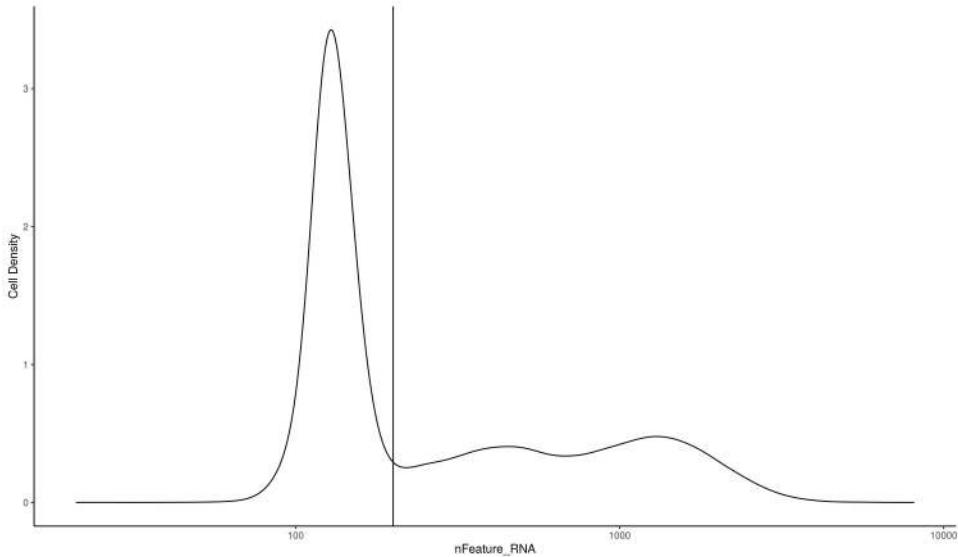


Figure 48: QC density plot from a sample in an unrelated dataset. The x and y axes are equivalent to those in Figure 47. A vertical line is drawn which suggests a potential cutoff to remove cells below a number of unique features.

Examination of density plots across QC metrics and across samples suggested that significant quality control was unneeded. It is important to consider that removal of empty droplets and doublet removal can often already substantially filter on the extremes of both sides of the range regarding total number of counts and number of unique features. In particular, doublet removal on this dataset was conservative and may have already removed more cells than necessary. That being said, a decision was made to apply a conventional filter on percentage of mitochondrial counts. As seen in Figure 49 of a density plot of this metric on a random sample from the dataset, a singular peak is mostly observed, but some peakiness does emerge at around the 10% mark. Interestingly, in (Osorio & Cai, 2021) the authors systematically analysed 1,349 scRNAseq datasets from human tissue with the aim of determining an ideal hard threshold for percent mitochondrial counts that is generally applicable across most datasets; they concluded that this standardised threshold should be 10%. Adaptive thresholding, for example removing cells with greater than 3 MAD in terms of mitochondrial counts, as implemented by the commonly used R package scater (McCarthy et al., 2017), is another option. However, I argue that this approach is difficult to interpret. By setting a hard threshold, a definition can be formed for what constitutes a low-quality cell (i.e. more than 10% of the counts coming from mitochondria), whereas adaptive thresholds hold no such meaning and can remove cells with wildly different mitochondrial percentages between datasets. In a very clean dataset, an adaptive approach may remove cells with mitochondrial percentages as low as 1 or 2%. It may be argued that even if those cells are outliers, they may not necessarily be low-quality.

Finally, it is important to note that I do not remove mitochondrial genes (nor ribosomal genes nor *MALAT1*) from the gene count matrix, though these genes were used for QC. The presence of non-zero counts in these genes are expected and can be a result of true biology; in the case of transcripts derived from mitochondria, they are known to be influenced by disease biology and are modulated by changes in nuclear gene expression (Muir et al., 2016). It is just the case that above a certain enrichment of these genes, their presence may be linked to quality issues rather than true biology, and rather than remove the genes, it is advisable to remove those overly-enriched cells instead. This is the case with single-cell (or in this case single-soma) data; when working with single-nucleus data, it makes more sense to remove both cells with enrichment of mitochondrial counts, as well as the genes from the matrix itself, as there is no biological reason for mitochondrial transcripts to reside in the nucleus; their capture likely arises purely from quality issues.

**Figure 49.**

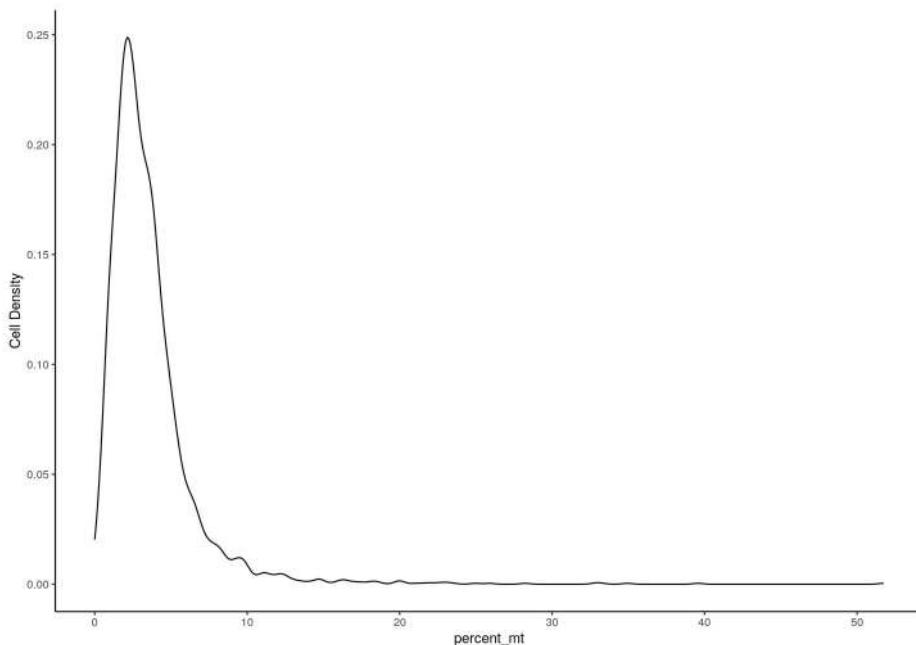


Figure 49: A density plot showing percentage mitochondrial counts in a random sample from the dataset. The peak near the 10% mark, as well as systematic research on optimal thresholding of this metric (Osorio & Cai, 2021), suggests that removing cells with greater than 10% mitochondrial counts is sensible in this dataset.

#### 2.4.3 Dimensionality Reduction and Clustering

Following the sole additional filtering of percentage mitochondrial counts, each sample underwent a conventional dimensionality reduction and clustering pipeline, following standard practice as advised by the Seurat developers. Note that the per-sample pipeline differs from the pipeline that will be applied to all samples when combined together, which will be described in upcoming sections. For the per-sample pipeline, it begins with normalisation using SCTransform, using the latest “v2” model (Choudhary & Satija, 2022;

Hafemeister & Satija, 2019). SCTransform is a variance-stabilising normalisation method for scRNASeq data that models transcript counts using a regularised negative binomial regression framework. It aims to account for technical confounders while preserving biological variation. By default, the only confounder that SCTransform regresses against is sequencing depth and it is possible to specify additional potential confounds (latent variables), a common one being percentage mitochondrial counts. I opted to not add any additional latent variables, as I decided that remaining expression of such genes, after QC, may reflect true biology and their influence on dimensionality reduction and other downstream steps is desirable.

Following SCTransform, principle components (PCs) were calculated, retaining the first 30 PCs as advised by the Seurat and SCTransform authors. This step can be further optimised for the selection of more or less PCs, and this optimisation was performed when processing on the combined samples, but left to defaults for the current steps on individual samples, as the analysis is strictly exploratory at this stage. The 30 PCs were then fed as input into the UMAP algorithm, set with default parameters. Finally, clustering was performed using Seurat's default louvain clustering algorithm. This algorithm is a graph-based community detection method that identifies clusters of similar cells by constructing a shared nearest neighbor (SNN) graph, where cells are represented as nodes and edges reflect their transcriptomic similarity.

Only one change was made to the default settings for clustering; the clustering resolution parameter was set to an adaptable value based on the number of cells in the sample. The author authors provide little guidance towards optimisation of this value, just that they find that a value of 0.8 is generally suitable for a dataset of about 3,000 cells. A larger value results in fine-grained clustering, while a lower one is coarser, and a desirable clustering outcome is largely dependent on the interests of the researcher as well as visually identifiable clusters on the UMAP projection. While methods exist for more empirical determination of clustering resolution, using well-established cluster stability metrics like silhouette scores and clustering trees (Zappia & Oshlack, 2018), for the purposes of exploratory analysis this was deemed unnecessary. Instead, I found that simply dividing the number of cells by 10,000 provides a clustering resolution value that is lightly adaptable to small changes in the number of cells between samples, producing clusters that align well with each sample's UMAP (Figure 50). For the analysis on combined samples, I elect for a different approach, which is described in following sections.

At this stage, the QC metrics described prior were replotted, including on a per-cluster level (Figure 51). Some of these metrics do show cluster-level differences, which may arise from biology, but may also indicate that some clusters are driven by minor differences in quality. I elected for no additional action in this area, but it demonstrates an area that could be of further research interest.

**Figure 50.**

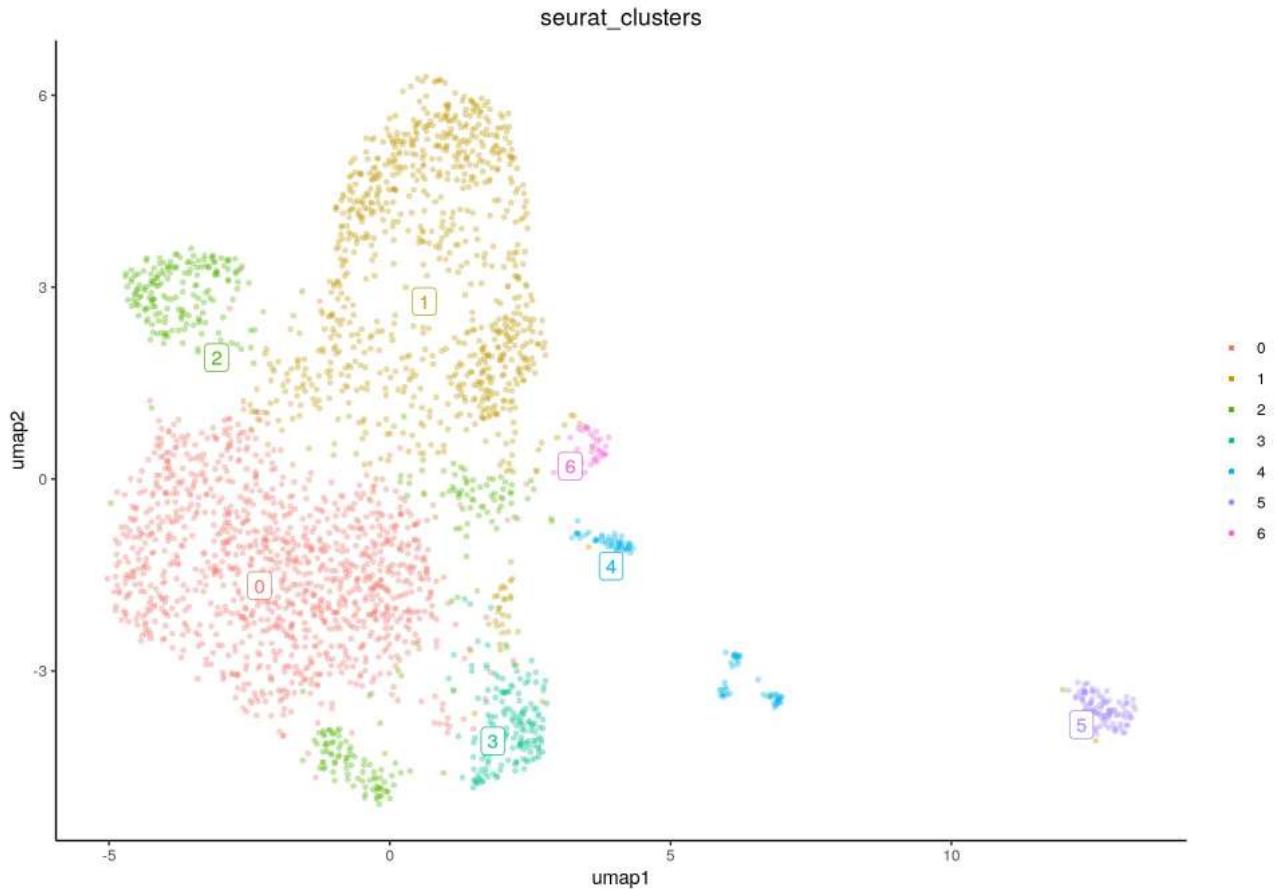


Figure 50: Final UMAP projection and clustering of a random sample in the dataset. These UMAPs were used solely for exploratory confirmation of the initial preprocessing before a combined analysis of the samples.

**Figure 51.**

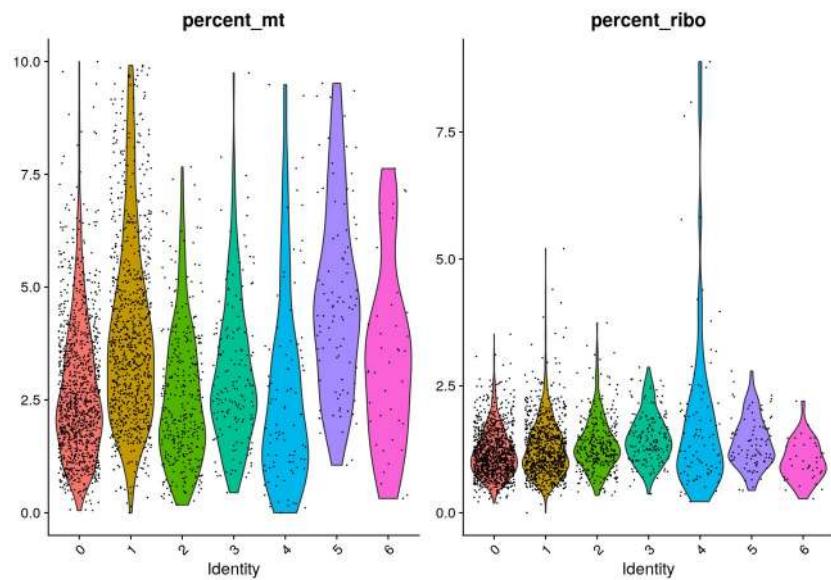


Figure 51: All QC metrics were replotted on a cluster-level basis, shown here are percentage mitochondrial and ribosomal counts. Although not further explored in this analysis, cluster-level differences are apparent, which may be biological or technical.

#### 2.4.4 Sample Merging and Cell-type Annotation

Following QC on each sample, another set of R Markdown files were used for a combined analysis. Seurat facilitated straightforward merging of the Seurat objects produced on the per-sample processing described above. Besides merging of raw counts, the SCTransform normalised data was also merged and corrected for library size between samples using *PrepSCTFindMarkers*. As per the documentation, “this function uses minimum of the median UMI (calculated using the raw UMI counts) of individual objects to reverse the individual SCT regression model using minimum of median UMI as the sequencing depth covariate” (<https://satijalab.org/seurat/reference/prepsctfindmarkers>). Due to difficult interpretability of this transformation, downstream analysis (gene set enrichment and differential expression) did not end up using this transform, just the raw counts, which were then normalised at the pseudobulk level (described in Section 5.1). In theory, working with this transformation may be a superior approach, as it should allow valid statistical testing between groups while utilising finer-grained single-cell normalisation, but the current implementation was deemed to be of a too experimental state.

After merging raw counts across all samples, I was able to obtain a combined Seurat object with 33,694 features across 51,955 somas from 8 donors, each with tangle-bearing and non-tangle-bearing populations; no exclusion of samples was required based on examination of the per-sample exploratory analysis described prior. To visualise the combined dataset, the dimensionality reduction and clustering pipeline previously described was run again, with any differences noted here. To better refine the most relevant PCs, this time I inspected an elbow plot (Figure 52), choosing to select the first 33 PCs for downstream analysis. This is a common and easy to understand approach for PC selection, where the aim is to identify an inflection point wherein additional PCs contribute little extra variance and the additional information is more likely to contain technical noise than biological variation (Zhuang et al., 2022). The other change made to the pipeline was selection of a resolution value. Because at this stage I am working with a single data structure of the combined data, automated high-throughput approaches are not needed, and it was deemed superior to manually select a value based on visual inspection. I decided on a low value of 0.005 given the number of cells, to simply separate excitatory neurons, which are the cell-types of interest, from inhibitory neurons and glial cells (confirmed through marker gene analysis, Figures 55 and 56). In any case, fine-grained cell-type annotation would instead use the labels provided by dataset authors using a reference-query approach (described below), so this clustering was only used for exploratory purposes and would not be used for actual subsetting of specific cell-types. The results of these described steps, visualising various properties of the dataset, is shown in Figures 52 through 57.

**Figure 52.**

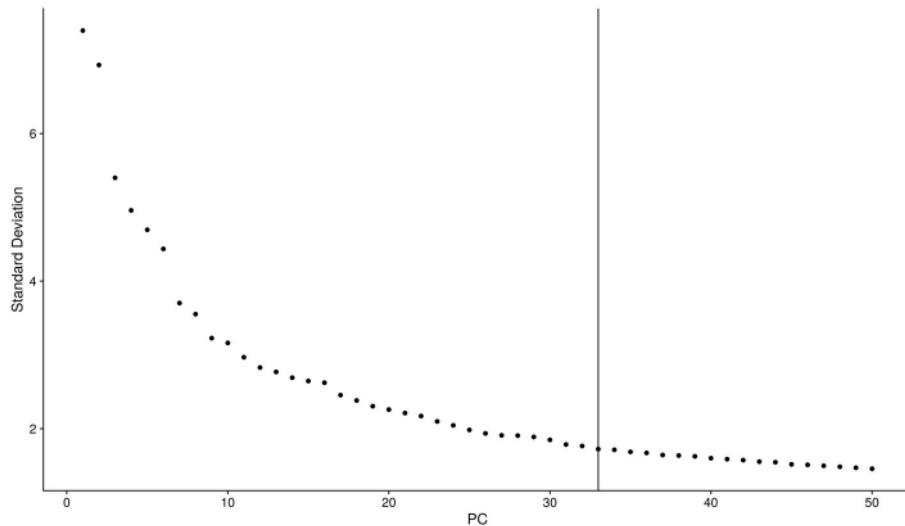


Figure 52: Elbow plot used for the selection of principal components (PCs) to retain for downstream steps. On the x-axis is each PC and on the y-axis is the standard deviation explained by each PC. PC 33 was determined to be a suitable point of inflection on the dataset with combined samples, and all PCs up to and including PC 33 was retained for downstream analysis.

**Figure 53.**

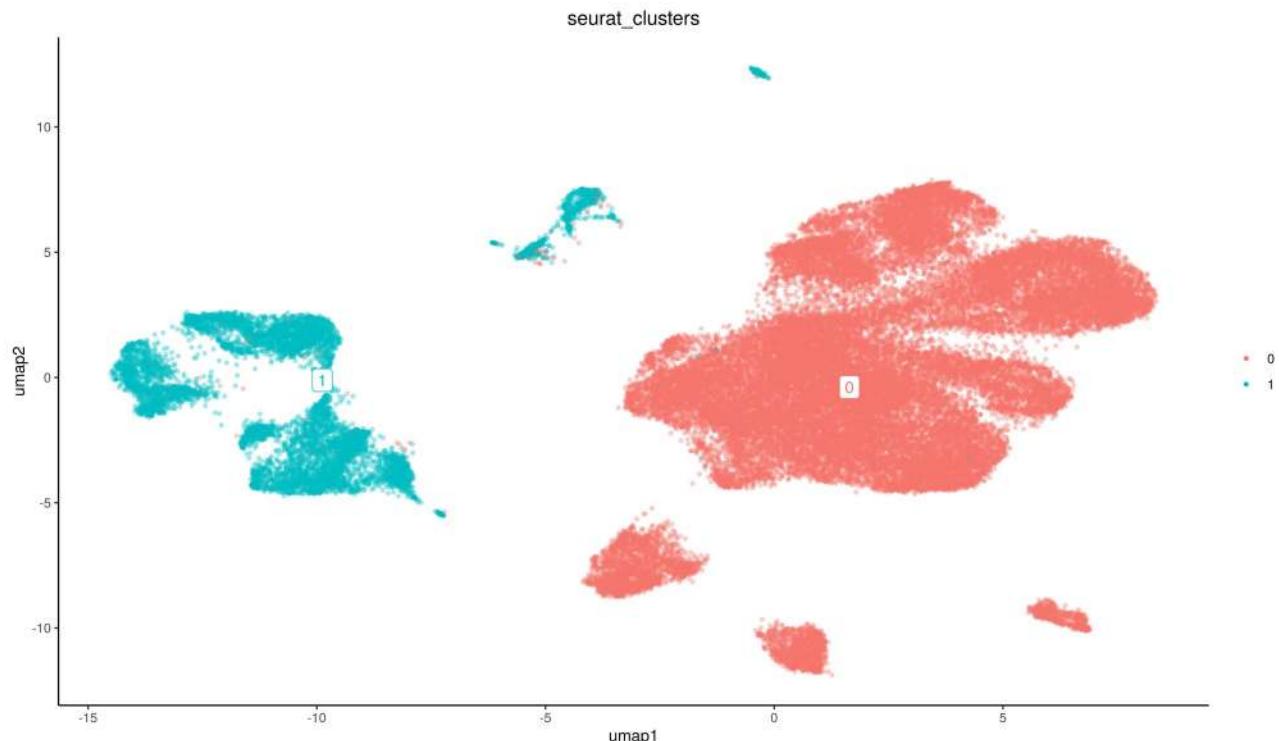


Figure 53: UMAP projection of all cells across all donors. A low clustering resolution value (0.005) was set to simply divide cells into inhibitory and excitatory neuron populations (confirmed through marker gene analysis in Figures 55 and 56).

**Figure 54.**

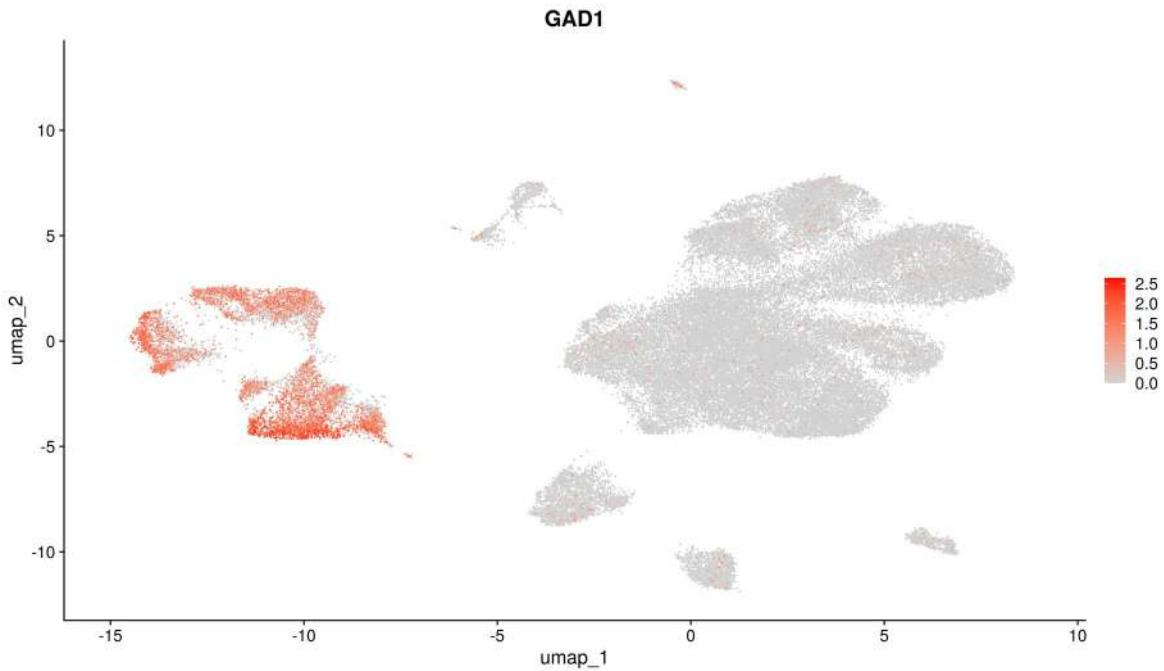


Figure 54: Confirmation of inhibitory neuron populations through plotting the normalised expression of the well-established inhibitory neuron marker gene *GAD1*.

**Figure 55.**

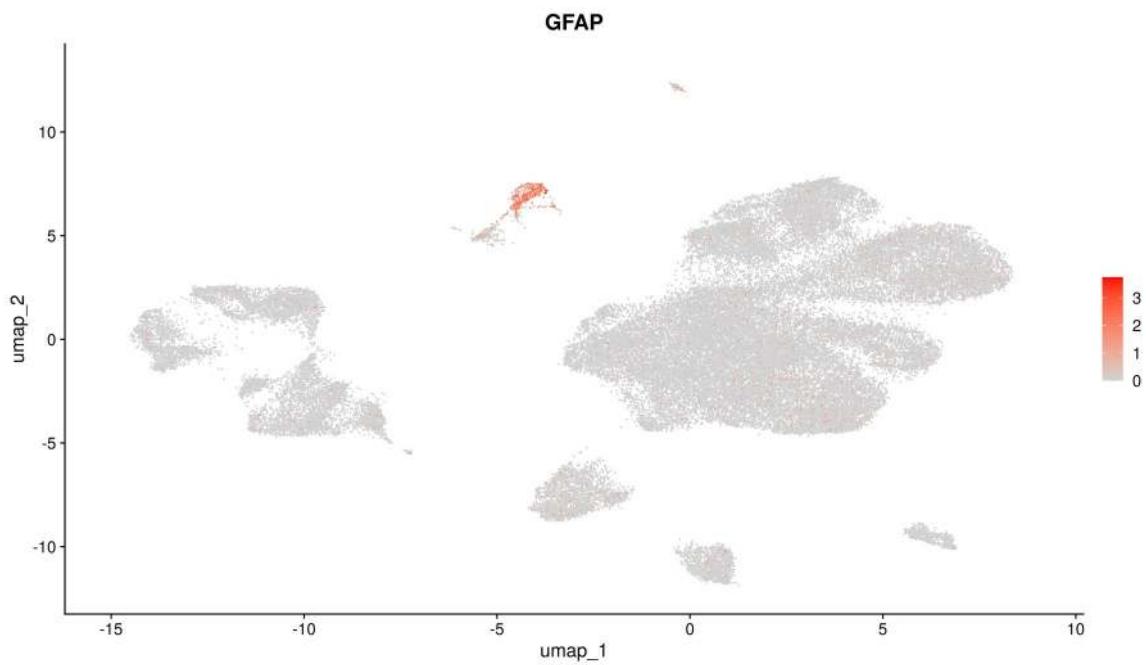


Figure 55: Confirmation of minor populations of glial cells, in this case astrocytes, using the well-established marker gene *GFAP*.

**Figure 56.**

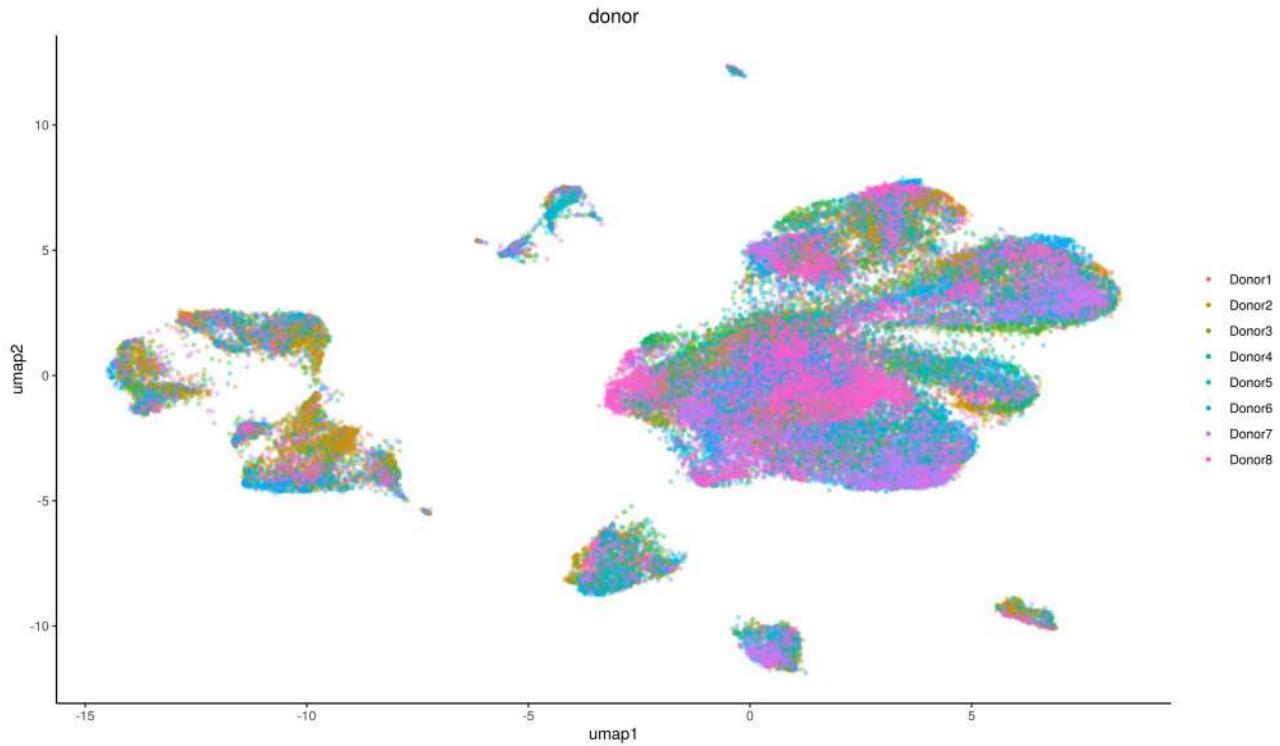


Figure 56: Visualisation of the distribution of the 8 donors among the cells, regardless of NFT status.

**Figure 57.**

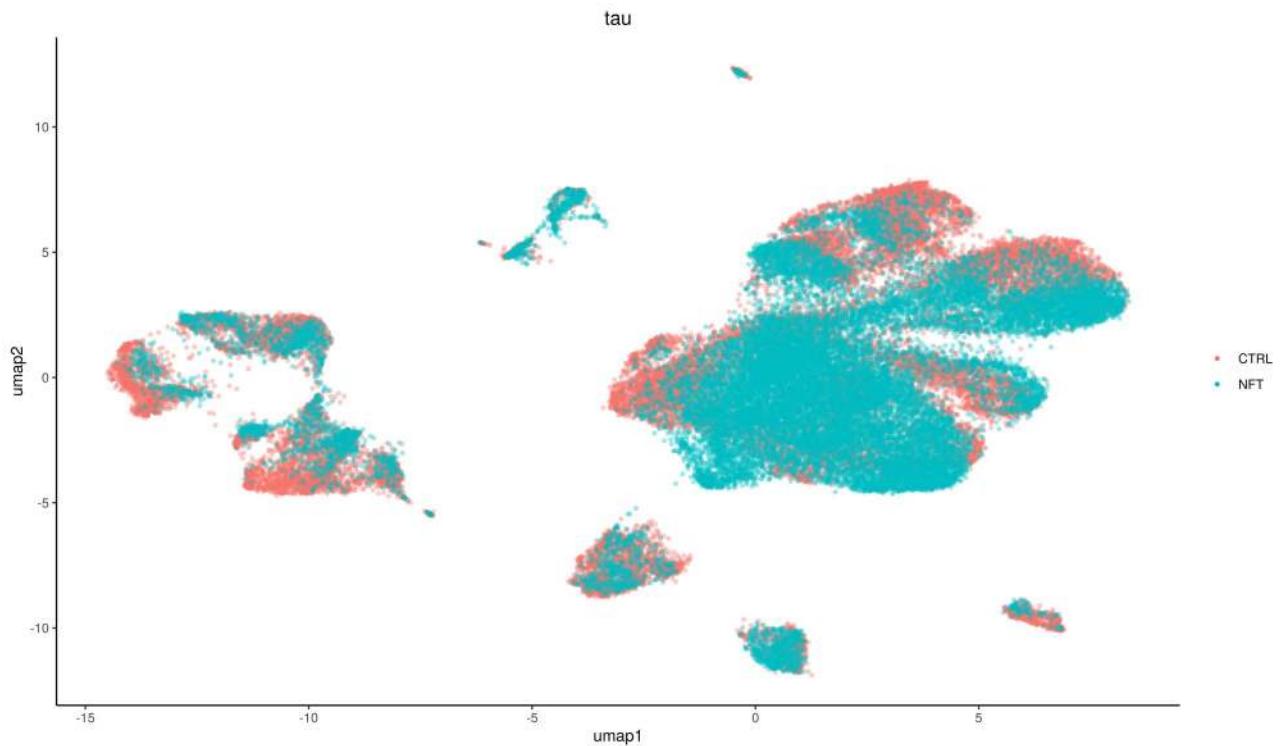


Figure 57: Visualisation of the distribution of the primary condition among the cells, where CTRL refers to non-tangle-bearing cells and NFT refers to tangle-bearing cells.

**Figure 58.**

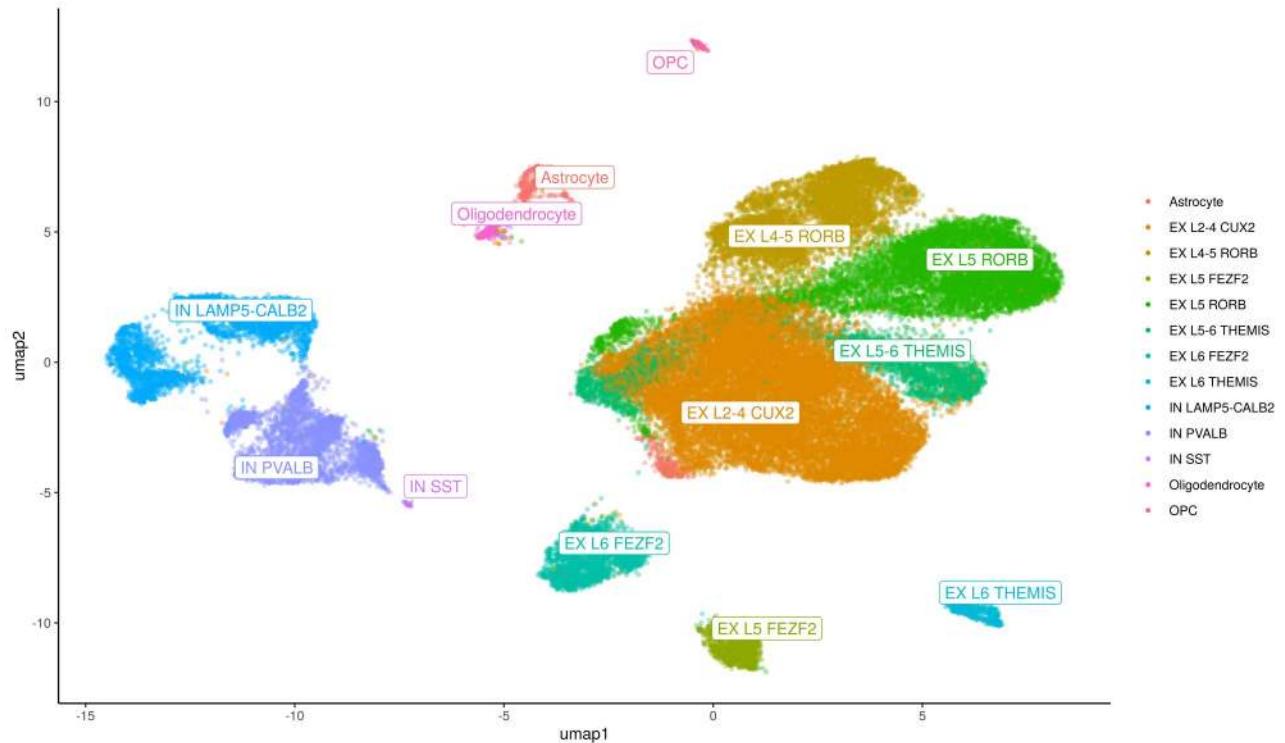


Figure 58: Final labelling of cell populations using the labels provided by the dataset authors after applying the Seurat reference-query approach for alignment (described below). Clusters of cells towards the left size of the UMAP were labelled as inhibitory neurons, in agreement with the marker gene analysis in Figure 55. Likewise, a few small glial populations are labelled. All other cells, composing the majority of the data, were annotated as various subpopulations of excitatory neurons.

As can be seen in Figure 58, the pipeline produced a dimensionally reduced dataset that aligns well with the author's provided labels, confirmed through the visual consistency of labels among nearby cells, forming well-defined clusters with little overlap between them. Furthermore, it was determined that no form of advanced integration was required to harmonise differences any among donors. Such methods have become highly popular for finding shared cell-types between donors or experimental batches by aligning shared biological variation while mitigating technical artefacts. However, they should be applied cautiously, as the integration procedure introduces dependencies between data points, which may fail to preserve the magnitude of relative expression between genes or even direction of change, resulting in artificial agreement between donors or batches or conversely, the masking of biological heterogeneity of interest (W. Chen et al., 2020). The choice to use integration, like many aspects of a bioinformatics pipeline, is ultimately up to the discretion of the researcher, and considering the drawbacks, I generally elect to avoid use of integration when a dataset exhibits minimal undesirable heterogeneity without it. This was determined through examination of Figure 56, which shows the distribution of cells among donors. It can be seen that all excitatory neuron populations, the populations of interest, are well distributed, regardless of source donor. This figure does however show

evidence of greater heterogeneity within inhibitory neurons; investigation of this population may warrant use of integration techniques, but the population was not within the scope of this analysis.

It should be noted that although the FACS strategy targeted MAP2<sup>+</sup> neurons, a small proportion of events in the sorted populations expressed astrocytic or oligodendrocyte precursor cell (OPC) markers. This may reflect technical factors such as non-specific MAP2 staining or carry-over of neuronal material into closely associated glia during dissociation. Subsets of these populations also exhibited reactivity with AT8. AT8 reactivity in glia has been reported in several contexts. Astrocytic tau pathology detectable with phosphorylation-dependent antibodies, such as AT8, is recognised as aging-related tau astrogliopathy (ARTAG), which can co-occur in brains with Alzheimer's pathology (Nolan et al., 2019). Meanwhile oligodendroglial tau inclusions ("coiled bodies") are also AT8-positive and have been described in AD (Kovacs, 2016).

In order to align the datasets, the Seurat reference-query approach was used (Stuart et al., 2019). Seurat provides functionality for projecting reference data or metadata onto a query dataset, a process that shares similarities with data integration but with key differences. Unlike integration, data transfer does not alter or correct the expression values of the query dataset, lessening the risk undesirable data manipulation. Instead, Seurat offers the option to project the PCA structure of the reference dataset onto the query to learn a joint structure. Principal components are purely linear transformations of data, preserving the original structure in a mathematically interpretable way while reducing dimensionality. Once anchor points between the datasets in the joint PCA are identified, the *TransferData* function is used to classify query cells based on reference labels, returning a matrix containing predicted cell identities and confidence scores, which can then be incorporated into the query metadata.

The labelling of the combined dataset concludes the conventional portions of the RNA sequencing pipeline and produces a Seurat object that is later used for incorporation into GeneFunnel and statistical analysis (Section 5.1). For those downstream analyses I focused exclusively on the excitatory neurons within the largest contiguous clustering space, specifically EX L2-4 CUX2, EX L4-5 RORB, EX L5 RORB, and EX L5-6 THEMIS (Figure 59). These cell-types were analysed extensively in (Otero-Garcia et al., 2022) as they were found to harbour the largest proportion of NFTs among the cell-types, implicating that they are selectively vulnerable. Several of these cell-types, such as the *RORB* expressing populations have been shown to be selectively vulnerable in other human AD scRNASeq studies as well (Leng et al., 2021). As the remaining pipeline of downstream analysis focuses on novel methods developed in this thesis, it is discussed in their associated sections rather than here in Methods.

**Figure 59.**

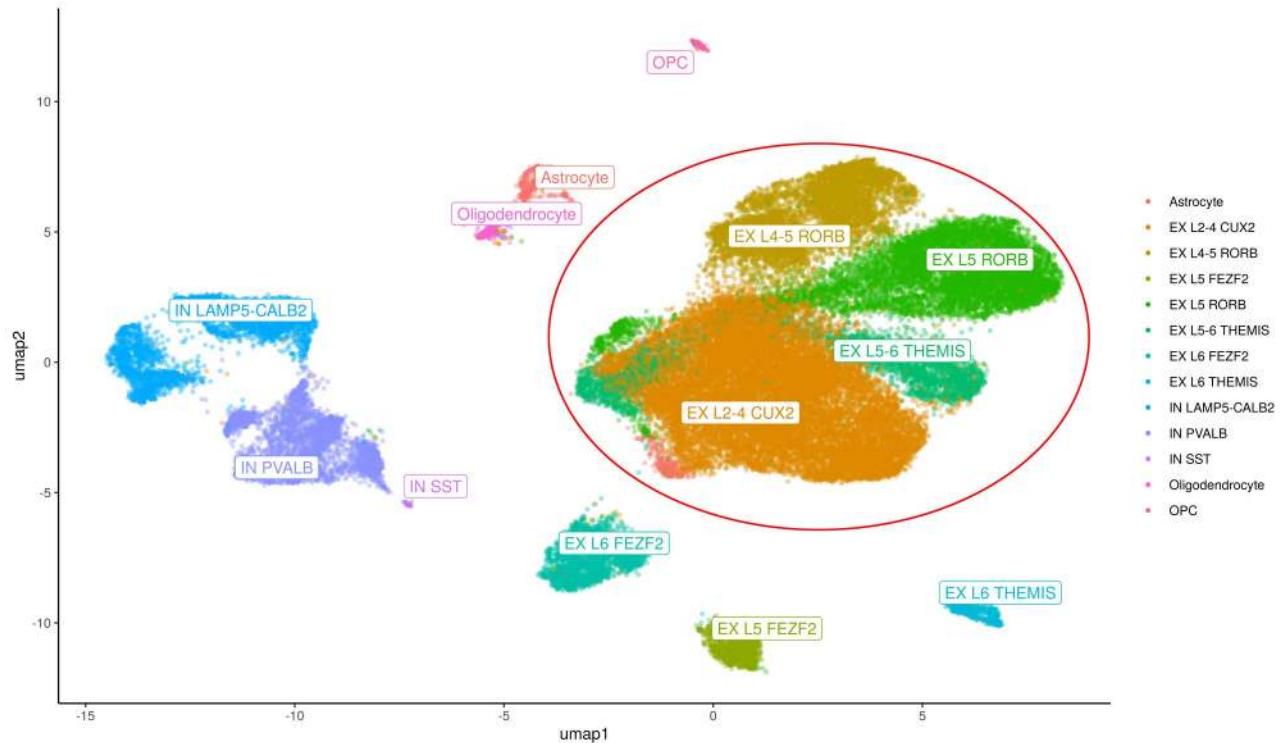


Figure 59: UMAP of the processed and annotated dataset of all combined samples of the FACS ssRNAseq dataset. Circled in red are the cell-type populations retained for all following downstream analyses (described in Section 5.1), specifically EX L2-4 CUX2, EX L4-5 RORB, EX L5 RORB, and EX L5-6 THEMIS. These populations were selected to retain excitatory neuron cell-types that fell within the largest contiguous cluster of excitatory neurons, in a sense the most representative set of cells of the dataset. Importantly, in the original dataset publication, the authors demonstrate that these populations also harbour the largest proportion of NFTs among cell-types, suggesting that they are the most selectively vulnerable cell-type populations in the dataset (Otero-Garcia et al., 2022).

## 2.5 Mass Spectrometry Preprocessing

### 2.5.1 Data Cleaning

The LCM mass spectrometry dataset was developed in-house, therefore the pipeline does not reference an external resource, unlike the transcriptomics pipeline. The foundational package driving this analysis was DEP (Differential Enrichment Analysis of Proteomics), available from the R Bioconductor repositories (X. Zhang et al., 2018). This package offers a substantial workflow for analysing mass spectrometry-based proteomics data and accepts tabular input formats, such as text files, generated by quantitative proteomics software like MaxQuant (Cox & Mann, 2008), which was distributed by the mass spectrometry service provider.

The pipeline I employed begins with the reading of the *proteinGroups.txt* file generated as output from MaxQuant. This file contains a comprehensive list of identified and quantified proteins from the raw mass spectrometry data, consolidating peptide-level information into protein-level results. As a form of raw input, this file generally requires some cleaning within R for greater usability. Therefore sample names were tidied; proteins without gene annotations were removed, and proteins with multiple gene symbols were collapsed into the most likely single symbol. Standard ways to approach these initial steps are documented by the DEP authors. After cleaning, the DEP SummarizedExperiment object could be created.

### 2.5.2 Quality Control

The first QC step of the pipeline is a simple histogram of log2 protein intensity values with all samples combined (Figure 60). The values are normally distributed as expected, however, a tail in the higher intensities suggest that it may be slightly skewed. Another QC plot showing the number of proteins per sample (Figure 61) suggests that this may be due to considerable differences in protein capture among the samples, including technical replicates. Likewise, there is significant variance in the pattern of missing values (Figure 62). This formed a considerable hurdle in the beginning steps of the analysis.

**Figure 60.**

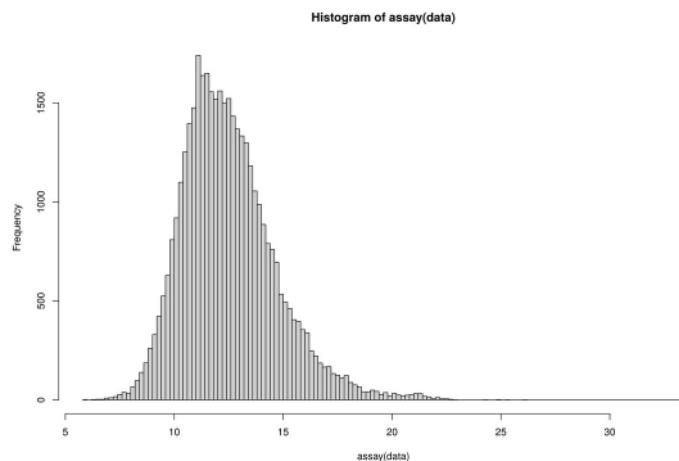


Figure 60: Histogram of log2 protein intensity values across the whole dataset. The data appears normally distributed as expected, but may contain a slight negative skew.

**Figure 61.**

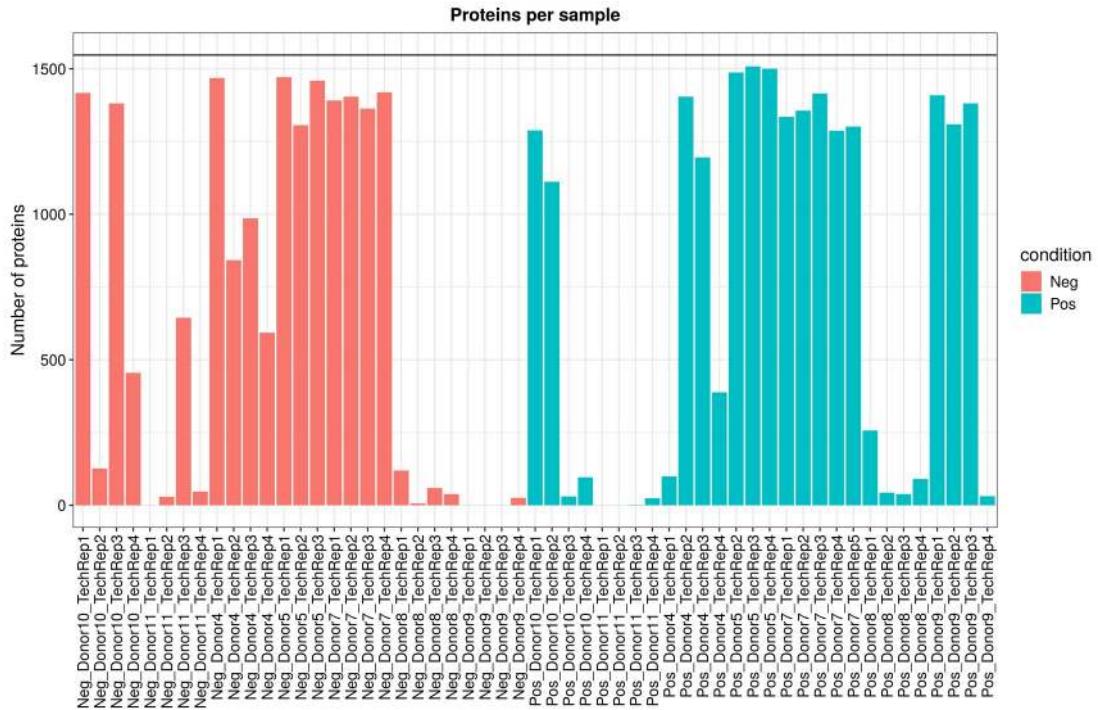


Figure 61: Plot of protein intensities per sample. The plot indicates that the efficiency of protein capture varies substantially between donors and technical replicates.

**Figure 62.**

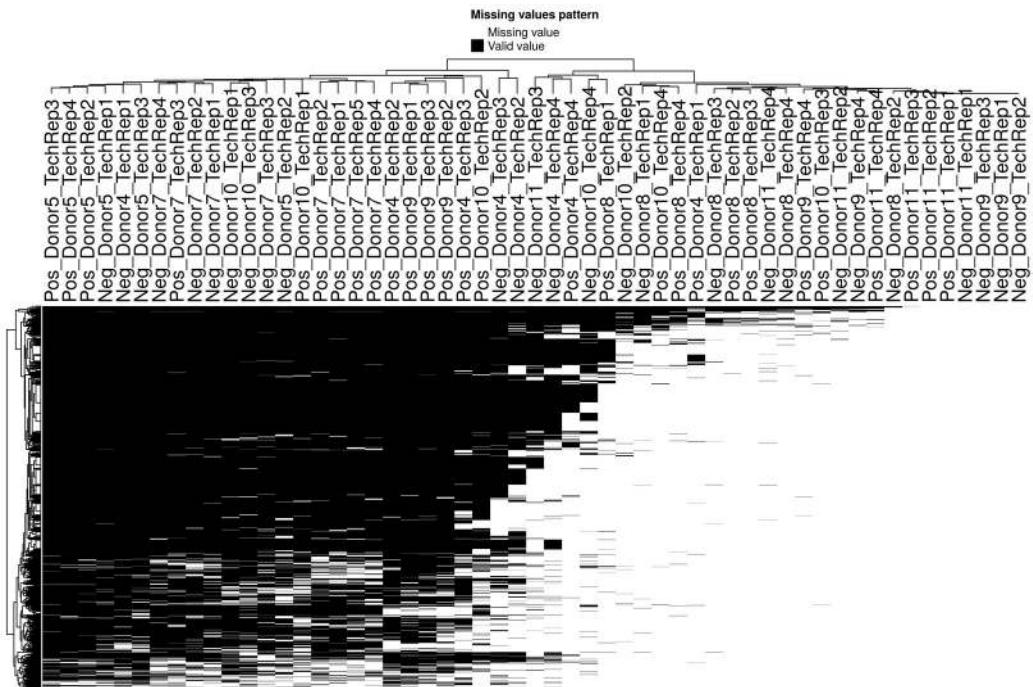


Figure 62: Plot of missing value distribution among samples, where black areas indicate proteins with missing values and white values indicate complete proteins. Like Figure 61, missing value distribution varied substantially between samples.

In order to explore the data further, I initially proceeded without any sample removal, though the QC steps so far suggest this may be necessary. As suggested by the DEP authors, variance stabilising normalisation (VSN) (Huber et al., 2002) was applied to the dataset. As seen in Figure 63, however, this normalisation failed to centre median of protein intensities across samples, likely due to the great heterogeneity between the samples. Nevertheless, following the final preprocessing step of the DEP pipeline, I applied imputation using the default k-nearest neighbor (kNN) algorithm (Gatto & Lilley, 2012) on the normalised data. Figure 64 shows the results of the imputation; abnormalities in the distribution of log<sub>2</sub> intensities further suggests that the default pipeline was not ideal.

**Figure 63.**

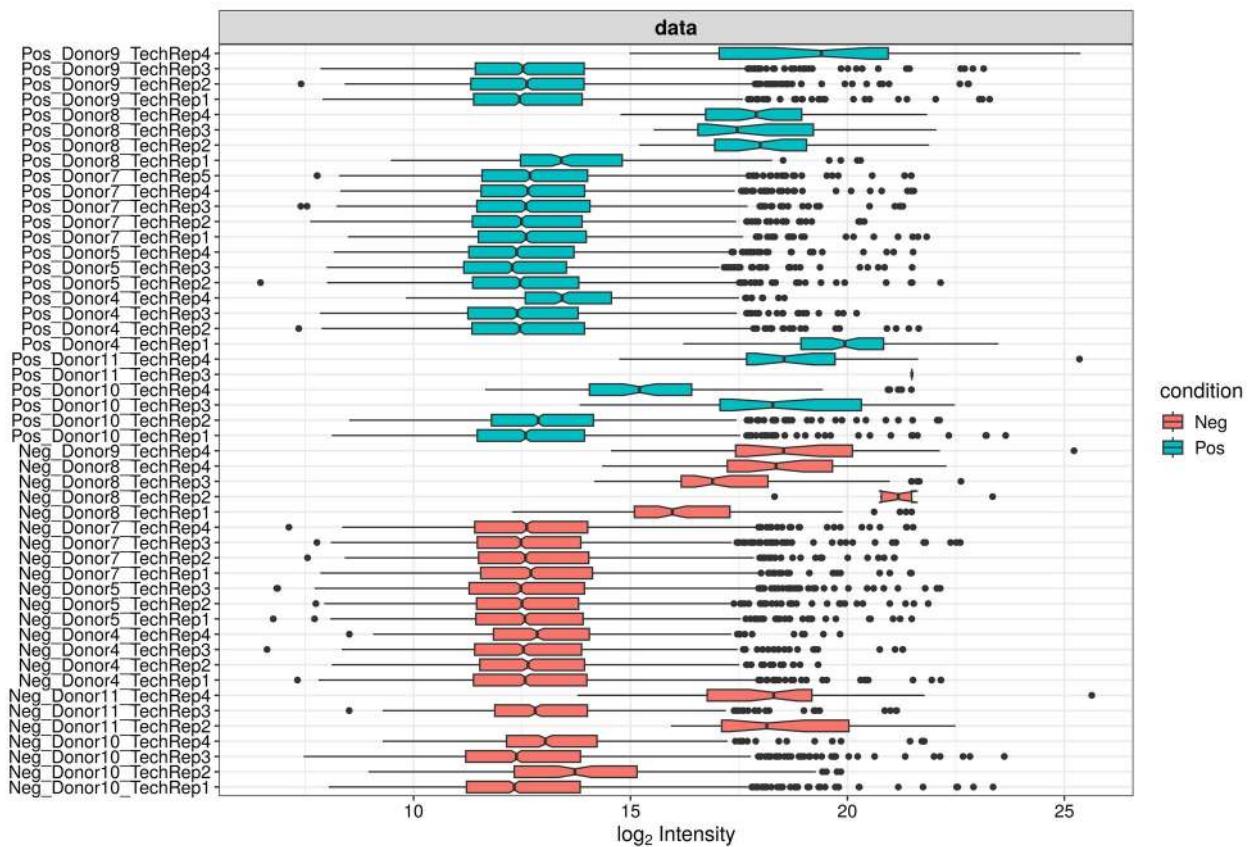


Figure 63: Standard VSN normalisation applied to the dataset without removal of samples. The failure to centre the medians of protein intensities suggests that the method was insufficient for normalisation of the dataset.

**Figure 64.**

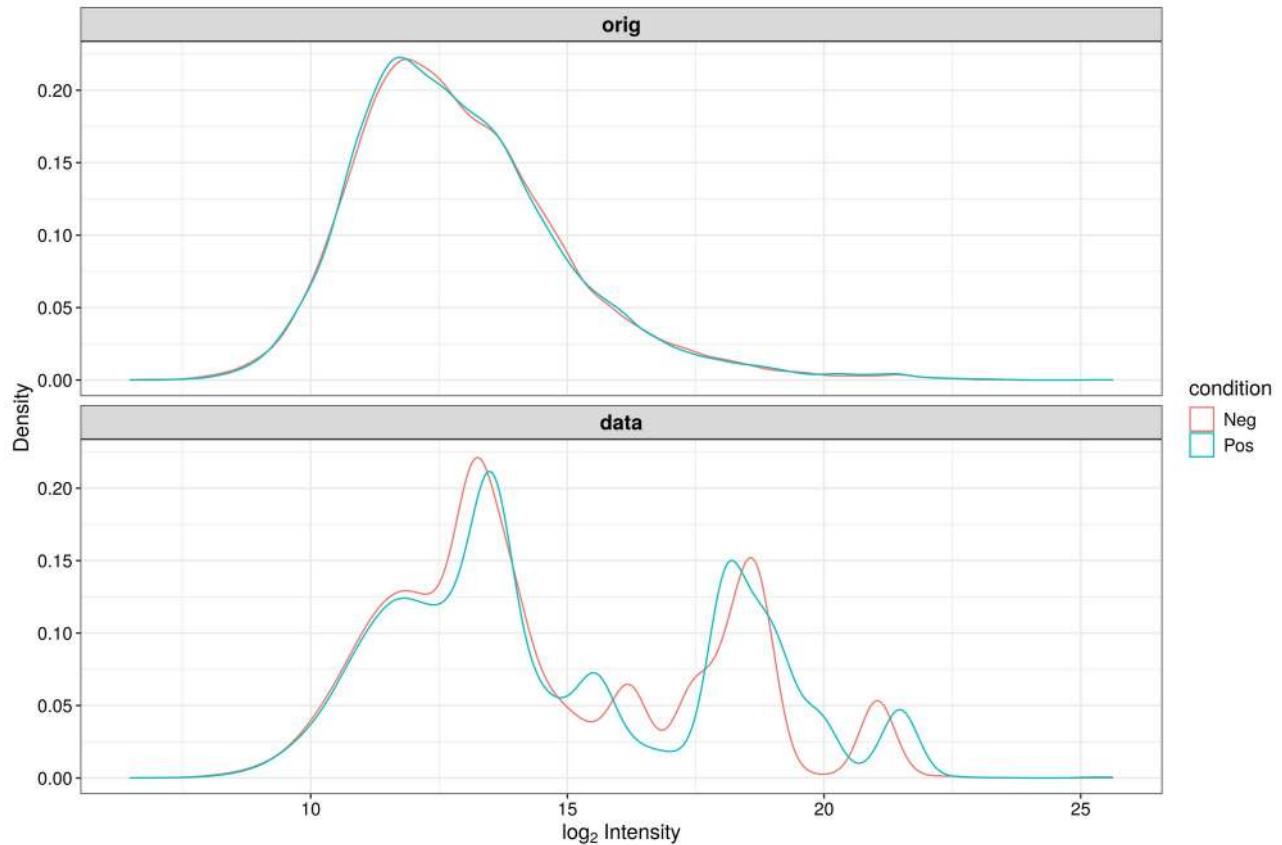


Figure 64: Result of kNN imputation (bottom) on the dataset without removal of samples. The lack of a smooth distribution, like the data before imputation (top), suggests that the method performed poorly in the naive first-pass pipeline described thus far.

### 2.5.3 Sample Removal

A PCA plot was created to inspect the results of this initial pipeline on all samples (Figure 65). The plot shows that the pipeline failed to separate the samples by condition, which is crucial for a successful analysis. Separation of donor effects were also unclear, with the distribution of points appearing more-or-less random. However, further inspection of the PCA clusters revealed that the first PCs appear to capture differences in proteins detected per sample, reflecting Figure 61. Therefore, by thresholding the samples to the coordinates where  $\text{PC1} < 0$  and  $\text{PC2} > 0$  (the upper left quadrant), I found a quantitative approach to subsetting to those samples that were of greater quality. This was confirmed by recreating Figure 61 with those samples only (Figure 66). I additionally removed one donor (Donor 9) because it failed to have a non-tangle-bearing neuron sample after removal of low-quality samples.

**Figure 65.**

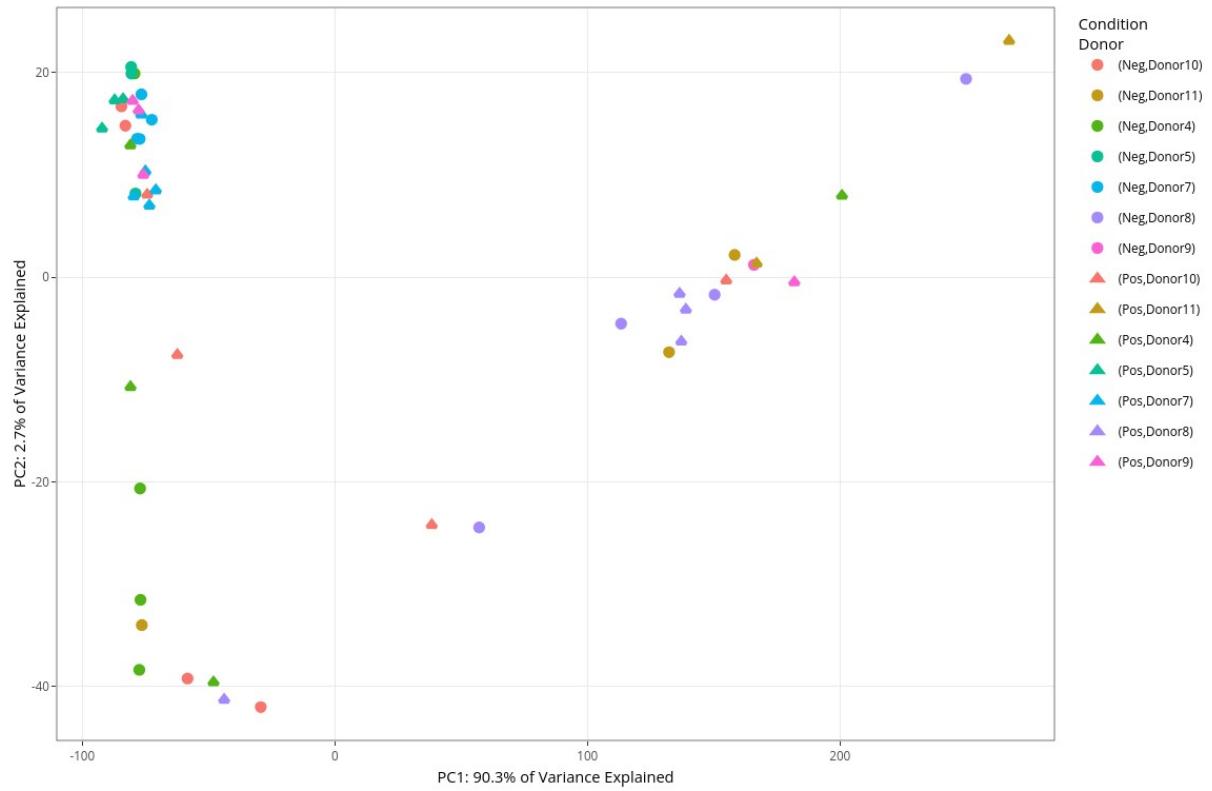


Figure 65: PCA of all samples after running the initial DEP pipeline. The first two components capture large variance but fail to separate samples by condition nor donor.

**Figure 66.**

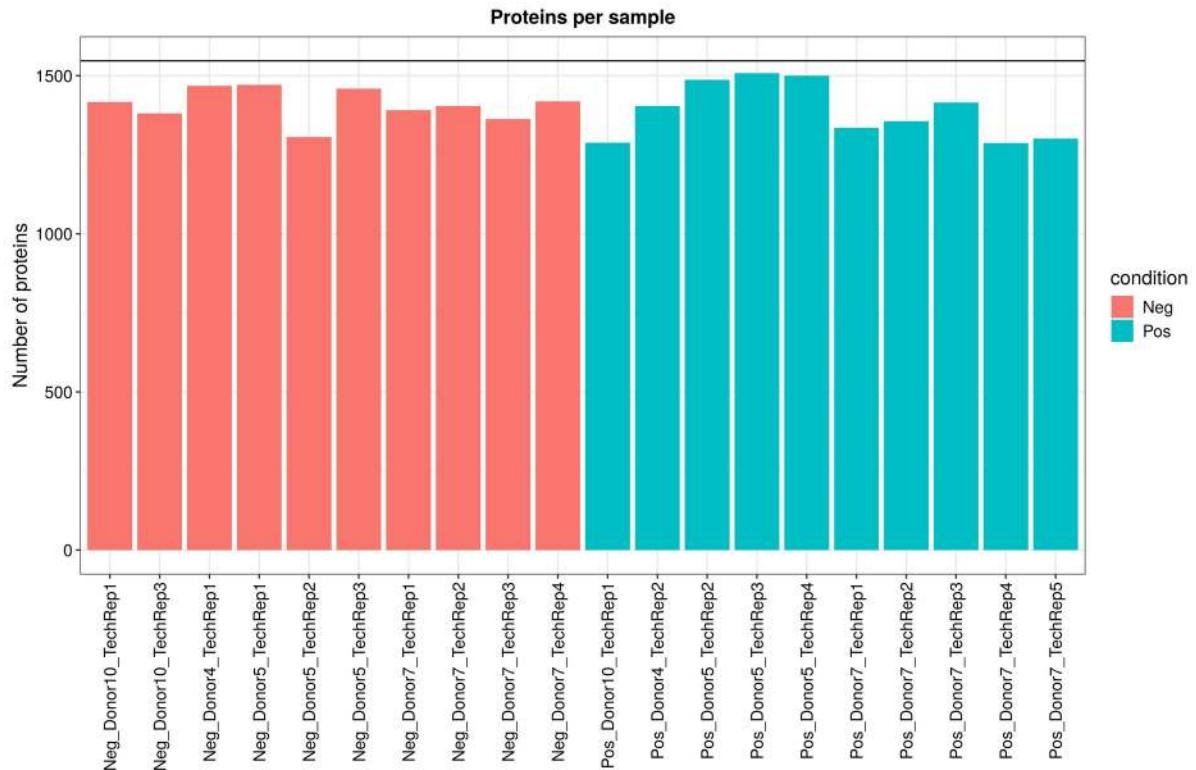


Figure 66: Proteins detected per sample after removal of low-quality samples by subsetting to those samples in the upper-left quadrant of the PCA plot in Figure 65. Also removed is Donor 9 for failing to retain a non-tangle-bearing neuron sample.

#### 2.5.4 Normalisation

The data was then reprocessed using the higher-quality samples and a more refined pipeline. Crucially, I also apply a novel imputation method I developed called ImputeFinder, described in greater detail in Section 3 and also published in (Fowler et al., 2025). Furthermore, a more adaptable normalisation method called EigenMS was employed (Karpievitch et al., 2014), which resulted in more effective stabilisation of variance than VSN, as evidenced by a flatter trend line when plotting standard deviation of protein intensities against their mean (Figures 68 and 69). Similarly, median-centering of intensities were observed to be mildly improved (Figures 70 and 71). The results of imputation by ImputeFinder, which contains steps that take place both before and after normalisation, is shown in Figure 72, and is suggestive of successful imputation and successful overall preprocessing pipeline. Originally designed for use in metabolomics data, which is also often mass spectrometry based, as described by the authors, “EigenMS works in several stages. First, EigenMS preserves the treatment group differences in the metabolomics data by estimating treatment effects with an ANOVA model (multiple fixed effects can be estimated). Singular value decomposition of the residuals matrix is then used to determine bias trends in the data. The number of bias trends is then estimated via a permutation test and the effects of the bias trends are eliminated” (Karpievitch et al., 2014).

The particular advantage of EigenMS for this dataset is that it allows for precise selection of bias trends to be removed, without removing those trends that correlate with the comparison of interest, in this case tangle-bearing vs. non-tangle-bearing neurons. Figure 67 shows a figure produced by EigenMS on this dataset that demonstrates this process. On the x-axis of each subfigure is each sample in the dataset, totalling the 20 remaining after sample removal. The y-axes describe SVD trends for each of the samples. The left-hand figures, titled “Raw Data” summarise these SVD trends for each overall bias trend, with the first three shown. The right-hand figures, titled “Residual Data”, summarise the SVD trends after removal of the bias trend on the corresponding left-hand side. The key point of this process is to identify the number of bias trends that should be removed before the remaining SVD trends reflect the comparison of interest. Because the samples are arranged such that the first 10 are non-tangle bearing samples, and the remaining 10 are tangle-bearing, bias trend removal should be iterated until a “Residual Data” plot is produced that shows SVD trends corresponding to these comparison groups, such that the SVD trend directions of change are clearly pointed in opposite directions. This point is reached in the final right-hand figure, indicating that the removal of 3 bias trends is sufficient for removal technical noise such that the remaining largest source of variation corresponds to the comparison of interest.

**Figure 67.**

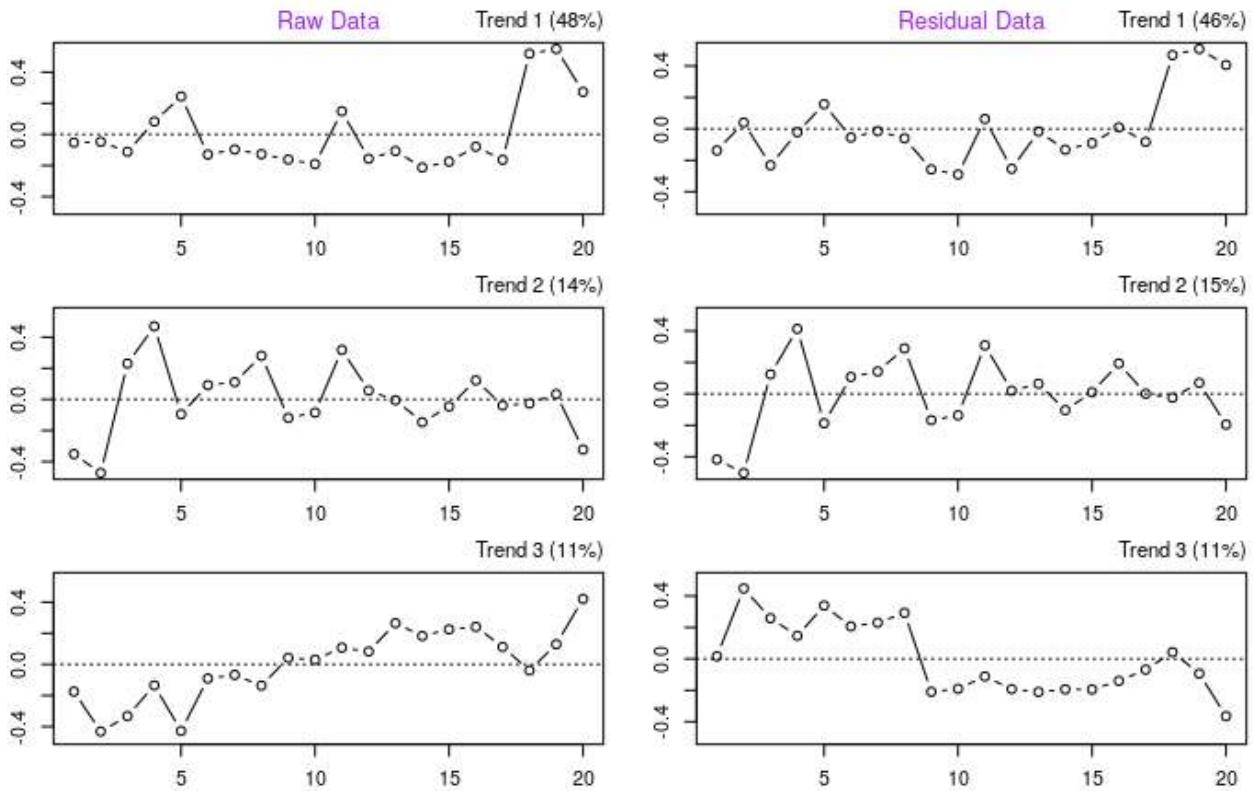


Figure 67: Output from EigenMS on the dataset to aid in bias trend removal. The x-axis of each subfigure shows each sample, where the first 10 are non-tangle-bearing samples and the last 10 are tangle-bearing samples. For each sample, the y-axes shows its corresponding SVD trend. The left-hand figures, titled “Raw Data” shows the SVD trend associated with each overall bias trend, while the right-hand figures, titled “Residual Data” show the remaining SVD trends after removal of the associated bias trend. After removal of a bias trend, the next set of trends shows the bias removal process with the prior trends removed. A researcher aims to select the minimum number of bias trends required for removal before the SVD trends align with the comparison of interest. That point in this dataset was determined to be 3 bias trends, as reflected in the final “Residual Data” figure.

**Figure 68.**

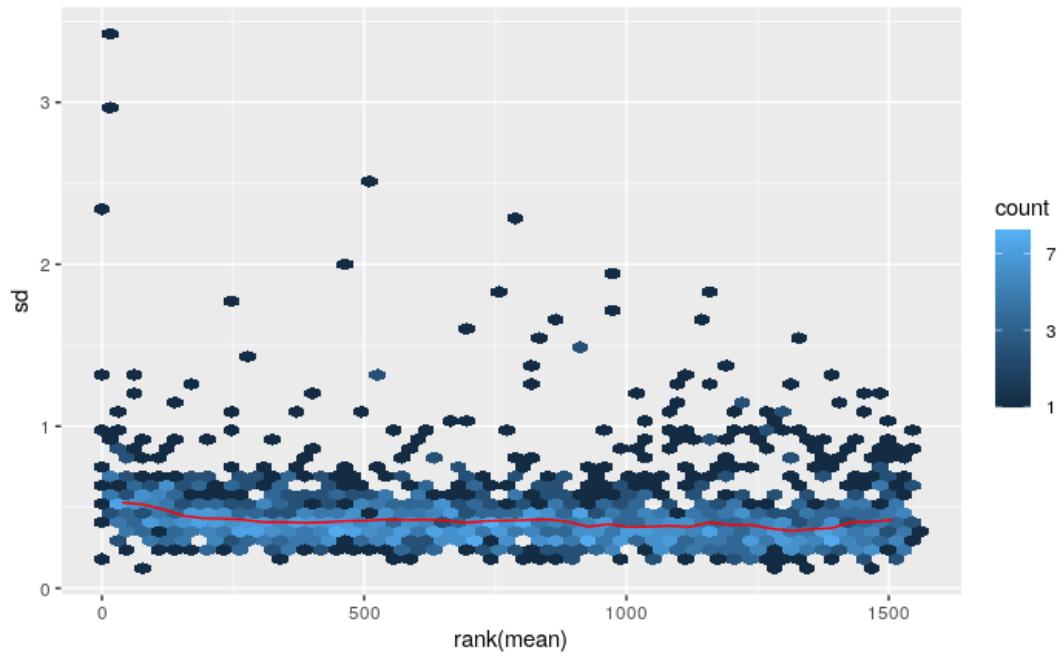


Figure 68: Result of EigenMS normalisation on variance stabalisation of the subsetted data before imputation.

**Figure 69.**

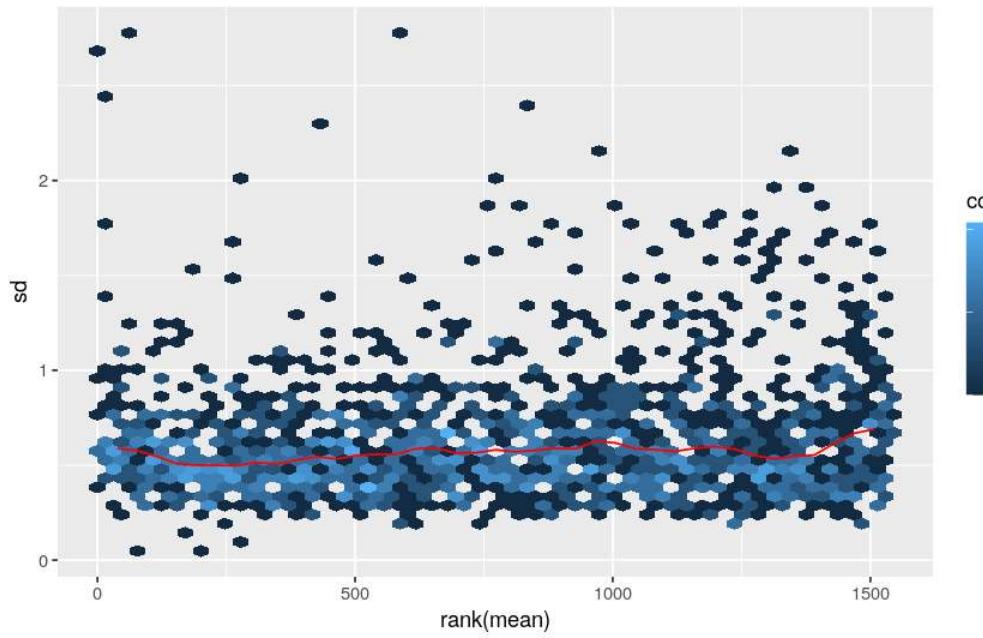


Figure 69: Result of standard VSN normalisation on variance stabalisation of the subsetted data before imputation.

**Figure 70.**

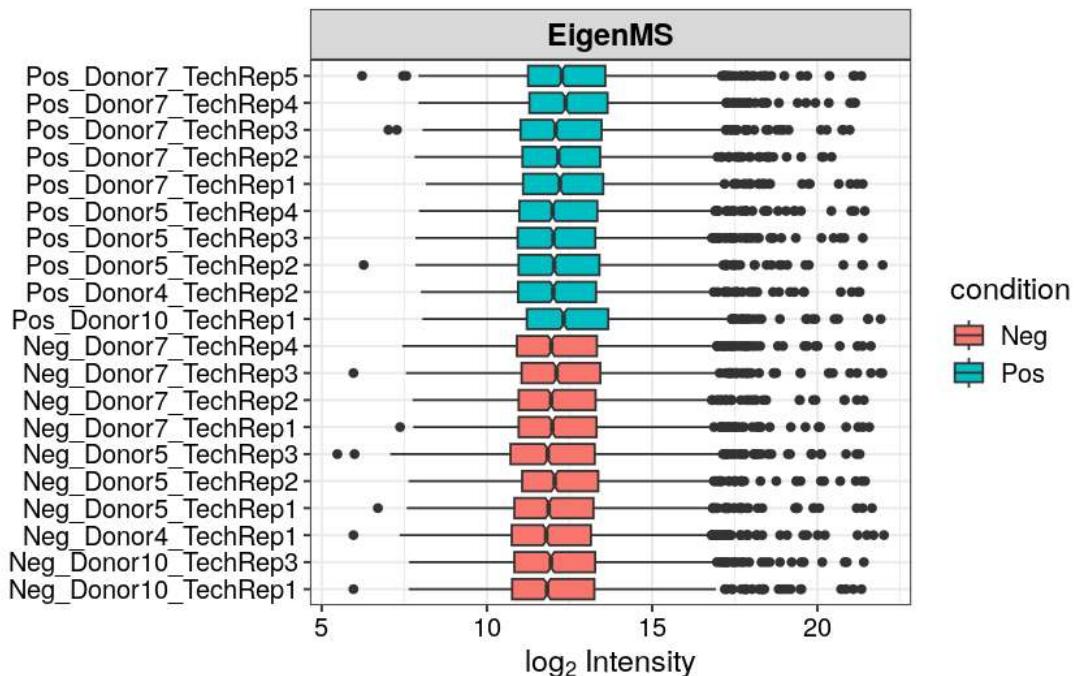


Figure 70: Box-plot of protein intensities on subsetted data after EigenMS normalisation.

**Figure 71.**

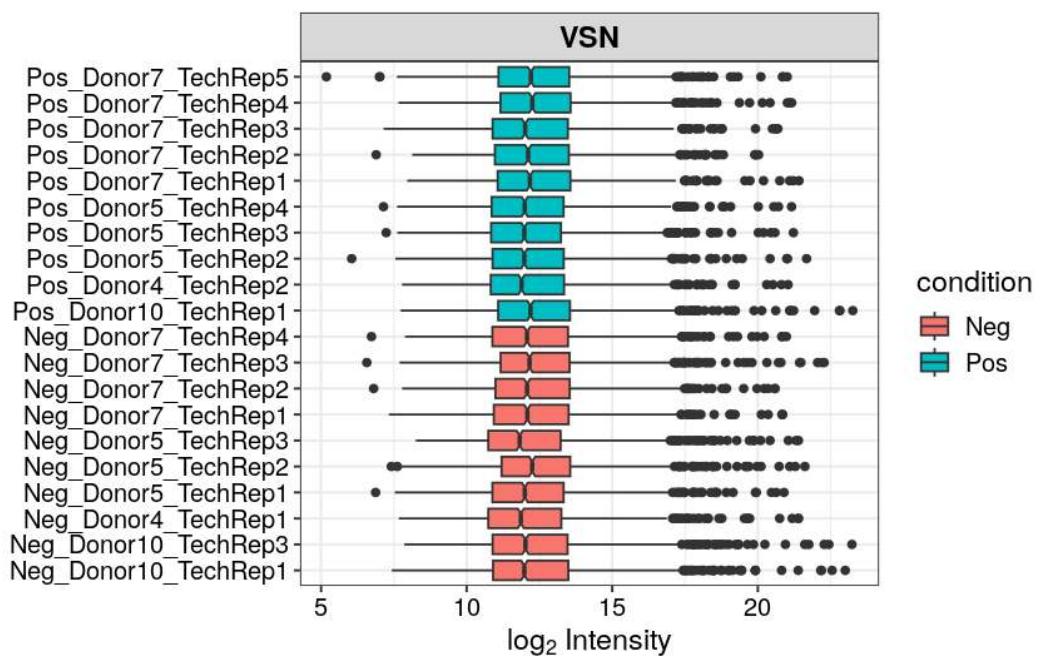


Figure 71: Box-plot of protein intensities on subsetted data after VSN normalisation.

**Figure 72.**

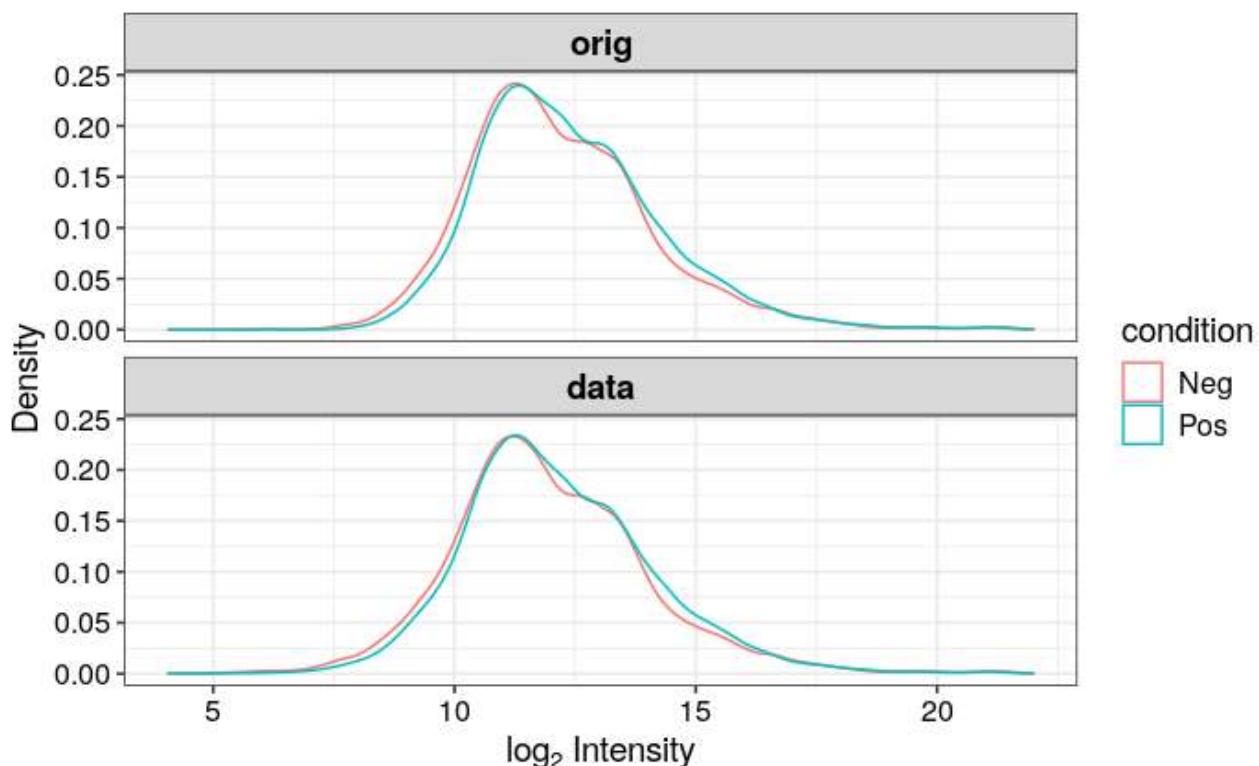


Figure 72: Protein intensity density plots before (top) and after (bottom) imputation using ImputeFinder coupled with EigenMS normalisation on subsetted samples. The close resemblance in distributions suggest that the imputation does not skew statistical assumptions required of downstream analysis.

I also evaluated whether the “Raw Data” eigentrends, that is the per-sample SVD score vectors before bias-trend removal (EigenMS Trends 1-3), reflected measured clinical variables. I focused on covariates that showed analysable variation between donors, namely Age at Death, Age at Onset, PMI and Sex (see Figure 38 for full table of clinical data). Fields with no useful variation were omitted, for example APOE status or ABC scores, because they were invariant or nearly so across the included donors. For each retained covariate I overlaid its standardised profile with the corresponding eigentrend across all 20 samples. Each series was z-scored across samples so they share a common scale. Sex was coded M = 1 and F = 0. As seen in Figure 73, no visual concordance is evident between the covariate profiles and the EigenMS trends. On this basis, the leading eigentrends in the Raw Data panel can be interpreted as technical structure rather than measured biology. Potential contributors include variation in peptide loading or sample preparation, source contamination, and other residual batch effects. In line with Figure 67 I elected to remove three bias trends with EigenMS and used the resulting Residual Data for downstream analyses. Further removal of trends would begin to remove true biological signal and reduce the contrast between tangle-bearing and non-tangle-bearing neurons.

**Figure 73.**

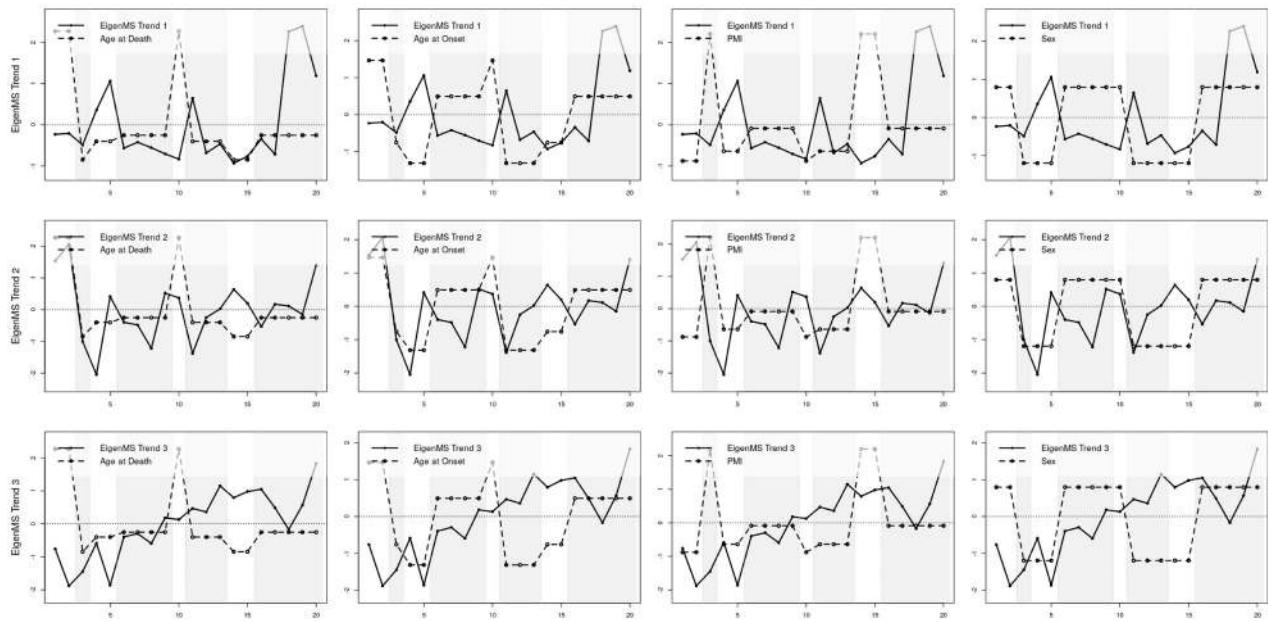


Figure 73: EigenMS eigentrend overlays with clinical covariates. Rows show EigenMS Trends 1–3 from the Raw Data. Columns show Age at Death, Age at Onset, PMI and Sex. In each panel the eigentrend for the 20 samples is plotted as a solid line and the selected covariate as a dashed line in the same sample order. Series are standardised within themselves by z-scoring so they share a scale. Sex is coded M = 1 and F = 0. Shaded bands indicate donor blocks. No visual concordance is evident between the covariate profiles and the EigenMS trends, consistent with technical rather than clinical structure.

Rerunning PCA after this new pipeline on the higher-quality samples (Figure 74) produces a dramatically different result than the initial run. On this new PCA, the samples now primarily separate by condition, that is tangle-bearing neurons vs. non-tangle-bearing neurons, suggesting that this data is viable for downstream analysis in these comparison groups. Nonetheless, there remains some minor degree of donor separation that can be seen in PC2. When substituting EigenMS with VSN, the samples fail to cleanly separate by condition. This demonstrates the practical benefit of EigenMS for this analysis, with the caveat that what was identified as technical noise could not be explained by an available covariate. Therefore interpretation warrants caution over possible overfitting of the normalisation procedure. Bearing that consideration in mind, the result of the EigenMS + ImputeFinder preprocessing serves as the foundation for downstream analysis that includes differential abundance analysis and GeneFunnel. These downstream steps, alongside imputation with ImputeFinder, are described in detail in dedicated sections as they constitute novel work.

**Figure 74.**

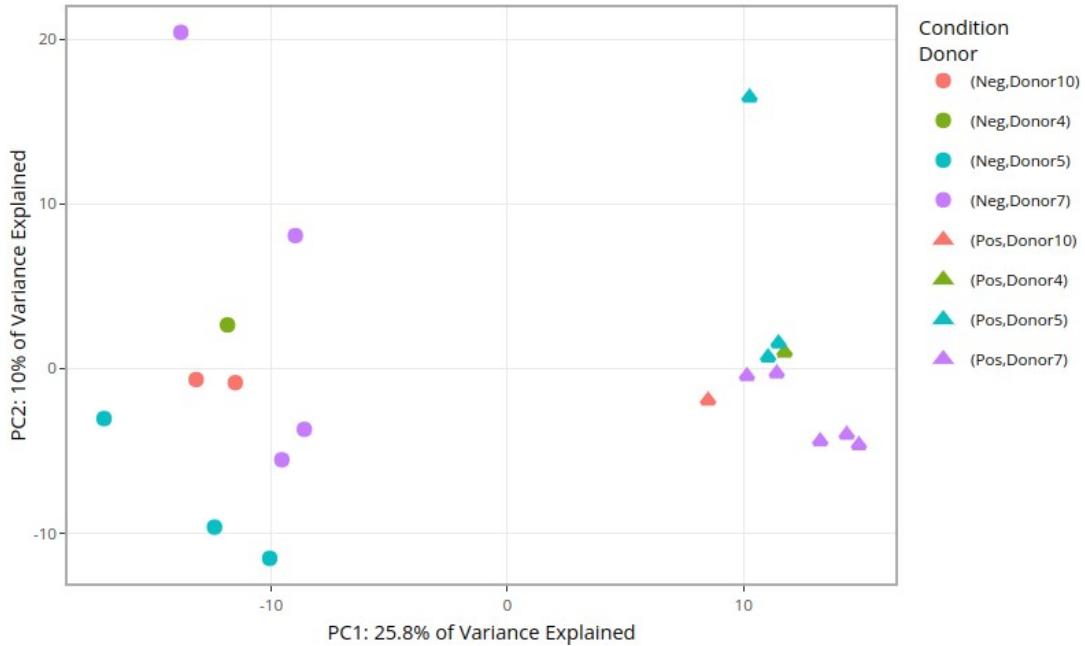


Figure 74: PCA analysis after running the EigenMS and ImputeFinder pipeline on the subsetted high-quality samples. The plot shows that PC1 effectively separate the samples by condition (tangle-bearing neurons vs. non-tangle-bearing neurons). PC2 on the other hand, still captures a minor degree of donor heterogeneity. This data serves as the basis for further downstream analysis using GeneFunnel and for differential abundance testing.

**Figure 75.**

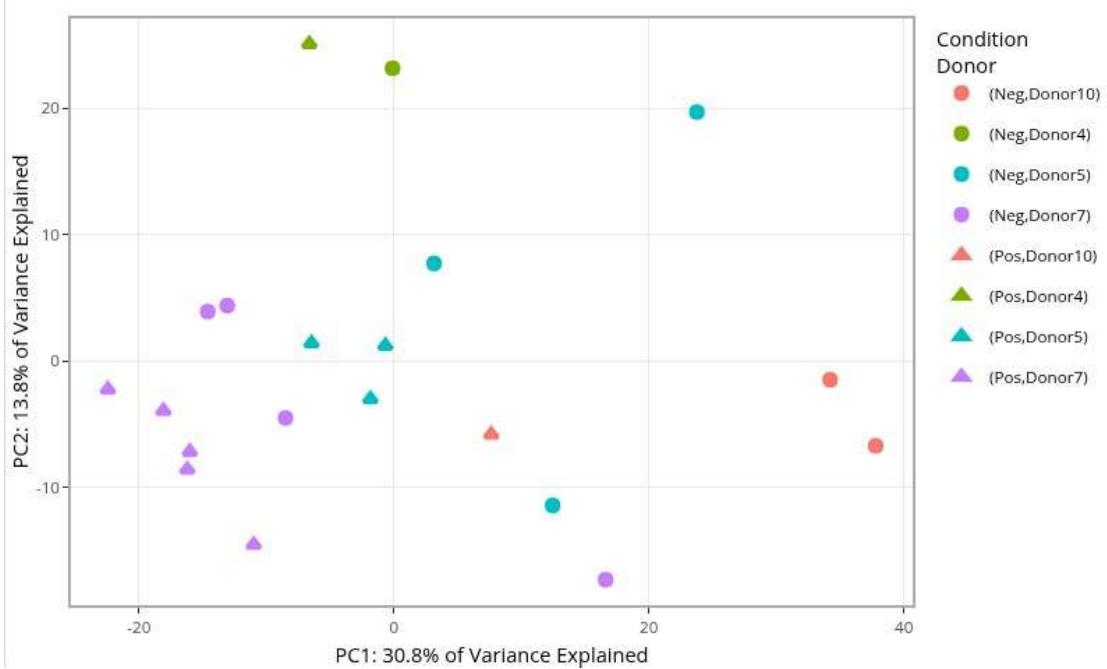


Figure 75: PCA analysis after running the VSN and ImputeFinder pipeline on the subsetted high-quality samples. Compared to Figure 74, featuring EigenMS, the primary condition is markedly less well separated.

### 3. Development of ImputeFinder Imputation Method

I developed ImputeFinder to handle the issue of missing values that plague proteomics data. This issue is routinely addressed using imputation, however, when reviewing existing methods for imputation, I came across the following unaddressed problem areas that prompted the development of a new method: 1) detection and handling of mixed types of missing values within a dataset (e.g. missing at random vs. missing not at random). 2) incorporation of comparison group information to retain features with missing values of probable biological origin. To-date, I was able to find one comparable method, called MI SFI-hybrid approach (Gardner & Freitas, 2021). However, the method, which will be further discussed, primarily highlights these problem areas and does not provide a software solution for handling them in real data. I first developed ImputeFinder to analyse extracellular vesicles derived from human AD tissue and it was later published in (Fowler et al., 2025). An R package for using ImputeFinder can be found at <https://github.com/eturkes/imputefinder>. It is in preparation for submission to the Bioconductor repository of bioinformatics tools for R.

#### 3.1 Definition and Description of Missing Values

The issue and handling of missing values in proteomics has been reviewed extensively (Kong et al., 2022; Lazar et al., 2016; M. Li & Smyth, 2023; M. Liu & Dongre, 2021). In proteomics, a missing value differs fundamentally from a zero value in transcriptomics due to differences in data acquisition methods and biological interpretation. Such missing values commonly arise due to instrumental and technical limitations rather than the true biological absence of a protein. Proteins may be present in a sample but go undetected due to factors such as ionisation efficiency, dynamic range constraints, and general stochastic processes in the acquisition methods.

In contrast, zero values in transcriptomics more likely reflect a biological absence or extremely low expression of a gene rather than a technical limitation. RNA sequencing uses deep sequencing coverage to count transcript reads, and the observed expression values are generally considered more complete and quantitative than proteomics data. A gene assigned a zero count in transcriptomics generally means that there was no measurable RNA transcript detected in that specific sample, rather than an artifact of instrument sensitivity or stochastic measurement variability. As a result, transcriptomic zero values are more likely biologically meaningful, whereas proteomic missing values require careful interpretation and statistical handling.

In proteomics, missing values in mass spectrometry data are generally categorized into two types: Missing at Random (MAR) and Missing Not at Random (MNAR), each with distinct implications for data interpretation and statistical analysis. In proteomics, MAR often refers to instances where a protein's non-detection is due to stochastic variability in measurement, meaning that the likelihood of missing data is unrelated to the actual abundance of the protein. It is often linked to technical error from instrumentation or biases

in detecting certain peptide fragments. MAR is sometimes further differentiated into Missing Completely at Random (MCAR), when source of missingness is believed to be completely unidentifiable. In contrast, MNAR occurs when missing values are systematically associated with protein abundance, often due to detection limits of the instrument. Low-abundance proteins are more likely to be missing because their signal falls below the instrument's sensitivity threshold, making their absence non-random and biased toward weaker signals.

**Figure 76.**

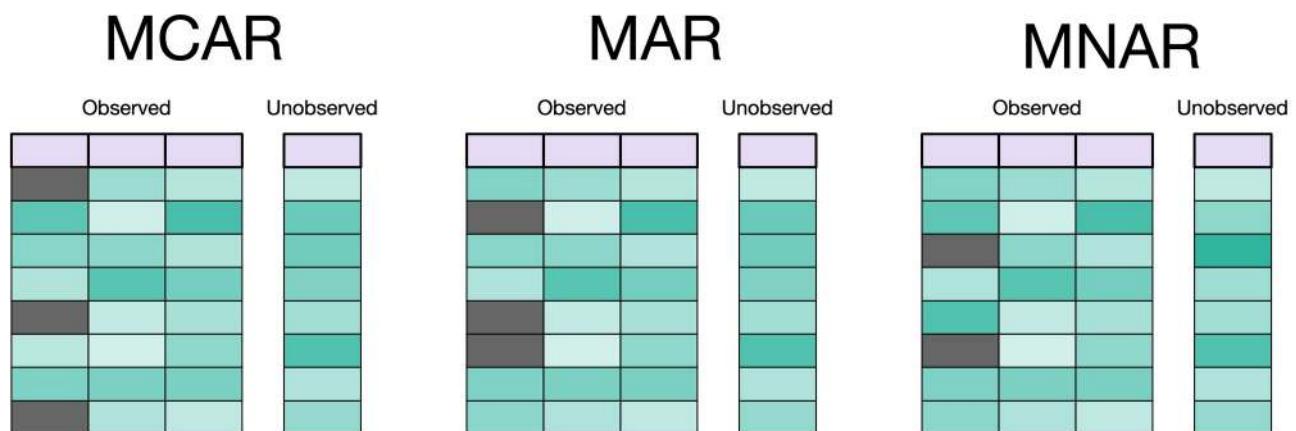


Figure 76: Three categories of missing data are illustrated in the first column of the “Observed” data matrix, where missing values are represented as gray squares. In the MCAR scenario, the missing values appear randomly distributed without any discernible pattern. For MAR, the absence of data corresponds with lighter values in the second column, suggesting the contribution of some factor. In contrast, the MNAR missing values tend to align with darker regions of the unobserved data, suggesting it is influenced primarily by limits of quantification. Figure reproduced from <https://feaz-book.com/missing>.

Distinguishing between MAR/MCAR and MNAR is critical in proteomics data analysis because improper handling of missing values can introduce bias in downstream statistical comparisons. MAR data can sometimes be ignored or handled using generalised imputation methods, whereas MNAR data is better addressed with specialized approaches, such as imputing missing values with low-intensity estimates to reflect their likely biological presence as below the detection limit. The issue is compounded when considering the likelihood that certain types of missing values, MNAR in particular, correlates with a comparison condition of interest. This is likely the case when certain proteins are robustly suppressed in disease states such as AD. Therefore, differentiating and addressing these types of missing values appropriately, in a condition-specific manner, is the key focus of ImputeFinder that I was unable find a suitable solution for in any existing method. ImputeFinder does not aim to reimplement individual MAR and MNAR methods, which have robust and long-standing support across a variety of fields, but rather provide a framework for appropriate application of such methods.

## 3.2 Prior Art: Imputation

The simplest type of imputation fills all missing entries with a fixed value, such as the overall mean or a low intensity constant. A common variant in proteomics is minimum value replacement, where each missing entry is replaced by a small constant (for example, half of the lowest observed intensity). Perseus’s “replace missing with noise” method draws from a normal distribution centred at a low intensity (effectively a randomised min substitution) (Tyanova et al., 2016). Low-intensity replacements (deterministic minimum, MinDet, or probabilistic minimum, MinProb) assume missing signals are left-censored values below detection (Gatto et al., 2021; Gatto & Lilley, 2012). MinDet uses a fixed small value (e.g. 1st percentile of each sample’s data), whereas MinProb adds random noise by sampling from a Gaussian centred at the minimal observed value with a small standard deviation. Such left-censoring methods are fast but can introduce bias if applied to values missing at random. They are best suited for MNAR missingness, and tend to underperform when a large fraction of data is MAR.

Another class of statistical methods leverages similarity across features or samples. These methods generally originate from work with microarrays. K-nearest-neighbor (kNN) imputation identifies peptides or proteins with expression profiles similar to the one with missing values and uses their measured intensities to infer the missing entry (Troyanskaya et al., 2001). A related approach is local least squares (LLS) imputation, which fits a small linear model using a subset of the most correlated features to estimate a missing value (Kim et al., 2005). These local methods preserve the multivariate structure of data and often yield more realistic values than global constants. However, neighbor-based methods can struggle if too many values are missing for a protein/peptide or if the data contains distinct clusters with little overlap.

Methods that exploit global data structure such as principal component analysis (PCA) and its Bayesian variant, Bayesian PCA (BPCA), treats missing value estimation as an inference problem by assuming that the data can be transformed into a lower-dimensional subspace. BPCA iteratively refines missing values by sampling from a posterior distribution of PCA model parameters (Oba et al., 2003). Similarly, expectation-maximization (EM) algorithms like MLE imputation use iterative estimation by first filling missing entries with initial guesses (e.g. means), perform PCA or calculate covariance, then re-estimate missing values until convergence (Hippel & Bartlett, 2019). A downside of these methods are computational cost and performance degradation if underlying assumptions (linear relationships, roughly normal data) are violated or if missing values are not random.

Ensemble machine learning can capture nonlinear relationships in proteomics data for imputation. A leading example is missForest, a random forest (RF) algorithm that iteratively trains a regression tree model to predict each feature’s missing values using all other features (Stekhoven & Bühlmann, 2012). Building off of these concepts, deep learning has been applied to proteomics imputation with promising results. Autoencoders, neural networks trained to reconstruct their input, can learn latent patterns from complete cases

and use that knowledge to infer missing values. Denoising autoencoders and variational autoencoders were combined in a proteomics imputation method from 2024 known as PIMMS (Webel et al., 2024). Another cutting-edge example is PEPerMINT, also published in 2024 (Pietz et al., 2024). PEPerMINT constructs a graph of peptide-protein relationships and uses a graph neural network to borrow strength from peptides of the same protein and from peptide sequence information. Such methods, while potentially complex, data-intensive, and difficult to interrogate, may be promising for providing adaptable, generalisable solutions for a number of case conditions.

### 3.3 Benchmarking of Existing Imputation Methods

Given the overwhelming variety of methods available for imputation, it is of crucial importance to benchmark them effectively. Imputation accuracy is often measured by how well an algorithm can recover artificially removed values. (Jin et al., 2021) introduced missing values into a complete proteomics dataset at varying levels (20% MAR with 20% MNAR, 20% MAR with 50% MNAR, and 20% MAR with 80% MNAR) and compared 7 methods. They found accuracy degraded markedly as MNAR missingness increased for all methods, but methods differed in resilience. Notably, RF, LLS, and BPCA had consistently lower error than simpler methods like single-value replacement. Another study in metabolomics (Wei et al., 2018) compared 8 methods and similarly recommended RF for general use, and their own QRILC method for heavily left-censored (MNAR majority) situations.

The ultimate test of an imputation method in proteomics however, is how it affects the identification of differentially abundant proteins. (Jin et al., 2021) also evaluated each method's impact on true positive (TP) detection and false discovery rate (FDR) using a controlled spike-in on real data. They reported that RF imputation produced the highest TP rate and kept FDR below 5%, whereas simpler methods had notably lower TP rates. Importantly, they further showed pathway enrichment results differed after different imputation methods, indicating biological interpretations can shift based on how missing data were handled.

It is important to note that benchmarking papers sometimes show conflicting and even contradictory conclusions. For instance, while the work by (Jin et al., 2021) suggests that LLS performs effectively, another (Bramer et al., 2021) advises against using LLS for imputation. These discrepancies likely stem from variations in the characteristics of the evaluation datasets, as well as differences in data processing and transformation procedures. Such pitfalls are discussed further in (Kong et al., 2022), where the authors performed a meta-review of benchmarking studies in proteomics imputation. They come to the conclusion that there is no “one-size-meets-all” method and the appropriate method should be decided after careful consideration by the research. The work is collated into the opinionated decision tree shown in Figure 77.

**Figure 77.**

Decision Chart

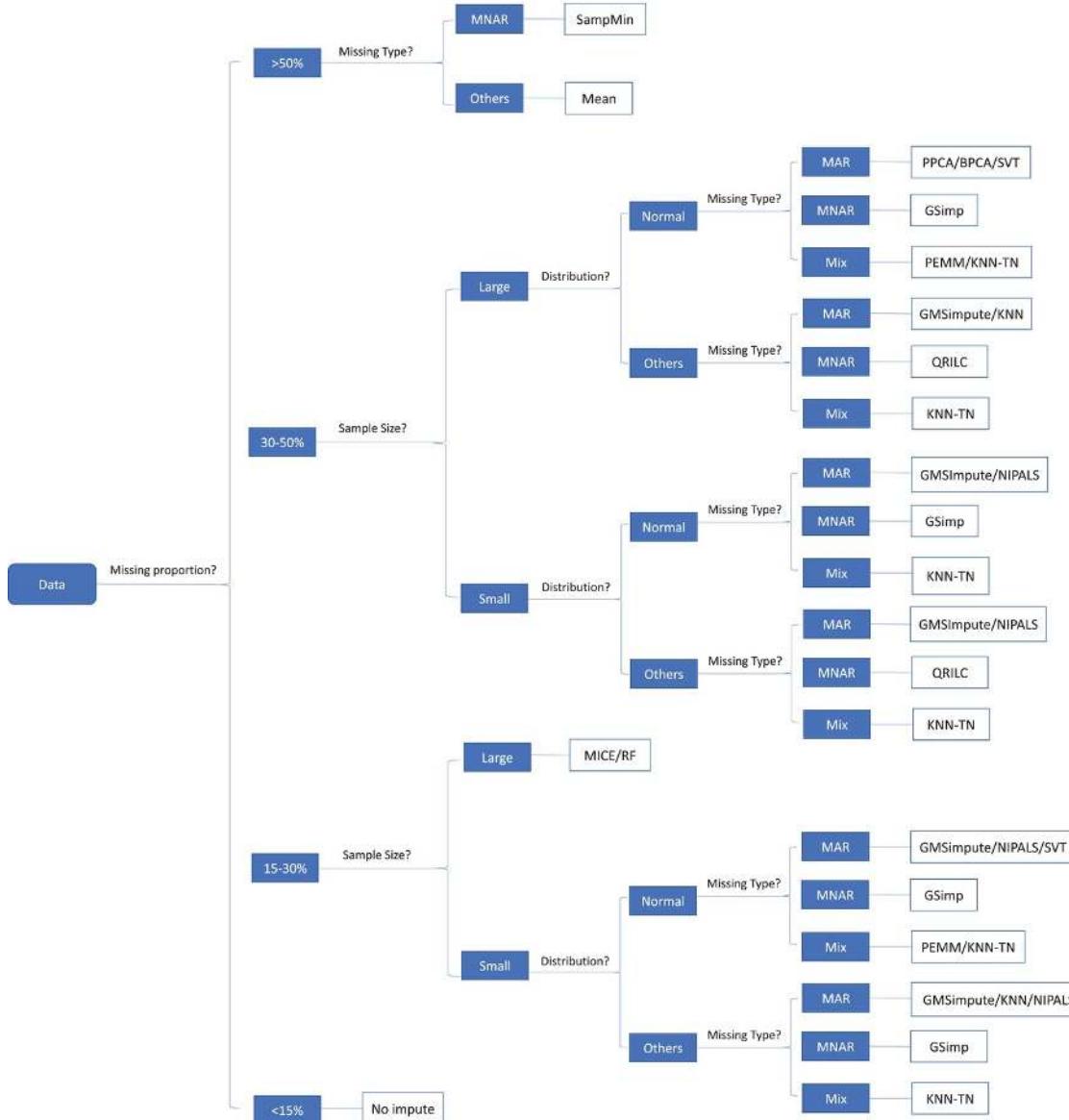


Figure 77: A decision chart that showcases the large variety of imputation methods available and conditions for which different methods may be considered ideal. The authors collated information from multiple review papers and benchmarking studies to arrive to their conclusions. Figure reproduced from (Kong et al., 2022).

### 3.4 The Lack of Hybrid Imputation Approaches

The above mentioned approaches are generally designed to deal with a single type of missing value (MAR, MNAR, etc.) or aim to be reasonably performant in mixed missingness situations. However, (Gardner & Freitas, 2021) provide strong justification for the separation of missing values within a dataset and application of multiple imputation techniques. They investigated a range of imputation strategies designated as MAR

approaches (kNN, SVD, MLE), and MNAR (MinDet, MinProb, QRILC), in addition to their own hybrid approach dubbed SFI-hybrid that combines kNN and QRILC. The effectiveness of these methods were evaluated on a simulated dataset based on real data, where missing values were induced (amputated), covering a range of scenarios from minimal missing values with comparable protein expression profiles to extensive missingness patterns often observed in presence/absence proteomics.

They drew several conclusions from the analysis. For one, they claim that single-method MAR or MNAR imputation strategies are only suitable when the underlying missingness mechanism is well characterised. Applying these methods without understanding the nature of the missing data can introduce bias or yield unreliable results. Secondly, when a protein is entirely absent from a treatment group, single-method MAR or MNAR imputation is not advisable as imputed values may converge towards more complete cases in the competing group, which may not reflect the reality of the group at hand. Lastly, the performance of single-method MAR and MNAR strategies deteriorate as the amount and complexity of missing data increases, whereas their hybrid approach maintained robustness.

**Figure 78.**

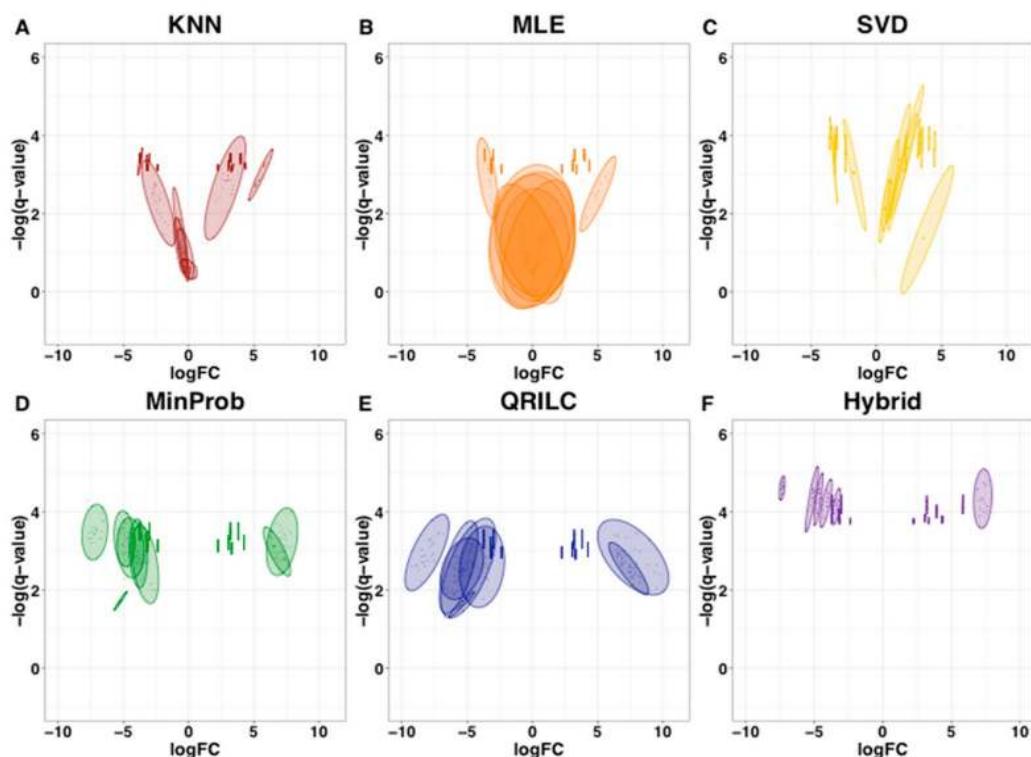


Figure 78: Spread plots of  $\log FC$  and  $-\log(q\text{-value})$  for amputated proteins in the simulated dataset. The figure suggests that the Hybrid approach maximises effect size and significance compared to the tested methods. Figure reproduced from (Gardner & Freitas, 2021).

While the work in (Gardner & Freitas, 2021) highlights the value in hybrid imputation approaches, as well as consideration of imputation within comparison groups, it is incomplete in the provision of a method for real-world use. Though the authors release a codebase to replicate the work, it does not contain a software solution. Furthermore, the actual designation of MAR vs. MNAR missing values remains out of scope of the publication. Working exclusively with amputation, throughout navigating their results, the authors are aware of which proteins were induced MAR and MNAR. This is valuable for an advocacy of the concept of hybrid imputation, however, in real-world data it is unclear which values are likely MAR and MNAR, presenting a gap in their work. Unfortunately, I was unable to find an existing solution for assigning missingness type, nor other methods that build off of a hybrid imputation concept. Therefore, I elected to develop ImputeFinder for the separation of mixed missingness in datasets and application of multiple imputation in different comparison groups.

### 3.5 Methodology of ImputeFinder

Figure 79 provides a schematic overview of the workflow of ImputeFinder. First discussing briefly, starting from the top-left of the diagram, the initial step entails replacement of proteins missing entirely in a condition to be compared. Next, a protein intensity cutoff for designating MNAR proteins is found empirically. This cutoff is then used to select MNAR proteins for each condition. Some filtering is then applied to remove MNAR proteins that are deemed too missing for further analysis. Finally, these lists of proteins are joined so that appropriate methods can be applied for each set of proteins.

**Figure 79.**

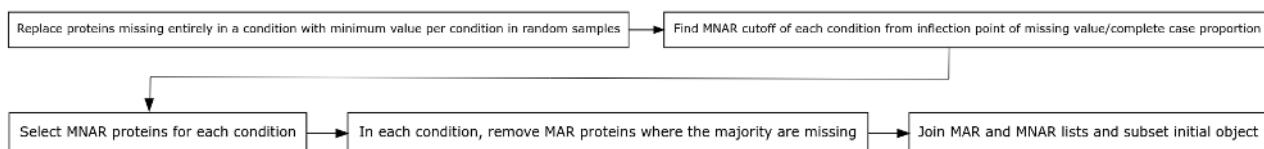


Figure 79: Schematic representing the workflow of ImputeFinder for differentiating and separating types of missing values in proteomics for the application of mixed imputation approaches.

In further detail, discussing the first step, proteins missing entirely in a condition are replaced with a minimum value per condition. The explicit approach, applied to a SummarizedExperiment object from DEP, is shown the code block of Figure 80. This step contains an important distinction that the protein must be present in at least one condition to be analysable at all, otherwise it is discarded. This is based on the rationale of a binary biological “on-off” protein; in one condition that protein may be present, but in another condition, protein intensity may be suppressed to the point of being below limits of quantification, in other words MNAR.

When replacing proteins fully missing in a condition, a minimal replacement procedure is performed. For each fully missing protein, only one sample out of the samples in the condition is replaced and by using the minimum intensity detected within a random sample of the condition. This is due to the minimum of one value being required within a condition to perform MNAR imputation. A random sample is selected in order to not introduce bias and better reflect the real-life stochastic reality that a random sample may have a protein intensity that barely passes the threshold of detection, and is thus detected with a low intensity value.

**Figure 80.**

```
rds <- file.path(cache_dir, "data_NA_replaced_by_fraction.rds")
if (file.exists(rds)) {
  data <- readRDS(rds)
} else {
  for (fraction in unique(data$condition)) {
    min <- min(assay(data)[ , which(data$condition == fraction)], na.rm = TRUE)
    replace <- which(is.na(rowMeans(assay(data)[ , which(data$condition == fraction)]), na.rm = TRUE)))
    set.seed(1)
    col <- sample(which(data$condition == fraction), size = length(replace), replace = TRUE)
    for (i in seq_along(replace)) {
      assay(data)[replace[i], col[i]] <- min
    }
  }
  saveRDS(data, file = rds)
}
```

Figure 80: Code block showing the first step in the ImputeFinder pipeline. Note that it is wrapped in a caching method as the approach as written can be time consuming. Also note the setting of a random seed to ensure reproducibility.

The next step modifies the *plot\_detect* function of DEP to create stacked probability density plots (Figure 81). These plots can be used to determine an “MNAR cutoff”, where proteins below a certain intensity value can likely be called MNAR. The rationale comes from the understanding that as intensity values decrease, they approach the limits of quantification in the mass spectrometer (M. Li & Smyth, 2023), and therefore, these proteins become increasingly likely to be MNAR. On the x-axis of the plot, which is created per-condition, the mean intensity of each protein is plotted, while on the y-axis, the relative proportion of missing values of each protein is plotted. In both the dataset in which this method was published (Fowler et al., 2025) and the LCM mass spectrometry data of this thesis, mean intensity decreases in a near linear fashion with increasing proportion of missing values, reaching an asymptote of 100% missing. It can be inferred that the missing values of these proteins are becoming increasingly MNAR and approaching limits for detection. On the other hand, as intensity increases, the proportion of missing values decrease. As the missing values approach 0%, at a certain inflection point, the relative proportion of missing values start to exhibit distinctive randomness. It is inferred that this is due to stochastic technical noise and thus any missing values of proteins above the inflection point’s intensity is more likely to be MAR.

Currently, the MNAR cutoff is managed through hand selection, however, automated selection of inflection point is under development for the software. The ideal location of the cutoff as being towards the rightward skew of the slope was determined through simulation, as discussed in the next subsection. Note also the importance of performing this procedure per-condition, as a protein can be called MNAR in one condition but MAR in another, particularly if it is an “on-off” protein dependent on the condition of interest. The inflection point is also currently decided per condition, but it is debatable as to whether a global inflection point is more sensible. Furthermore, it also seems logical that inflection points may be specific to the mass spectrometer in use, as instrumentation is a major driver of missingness (McGurk et al., 2020). Such information may be considered in future developments of the method.

**Figure 81.**

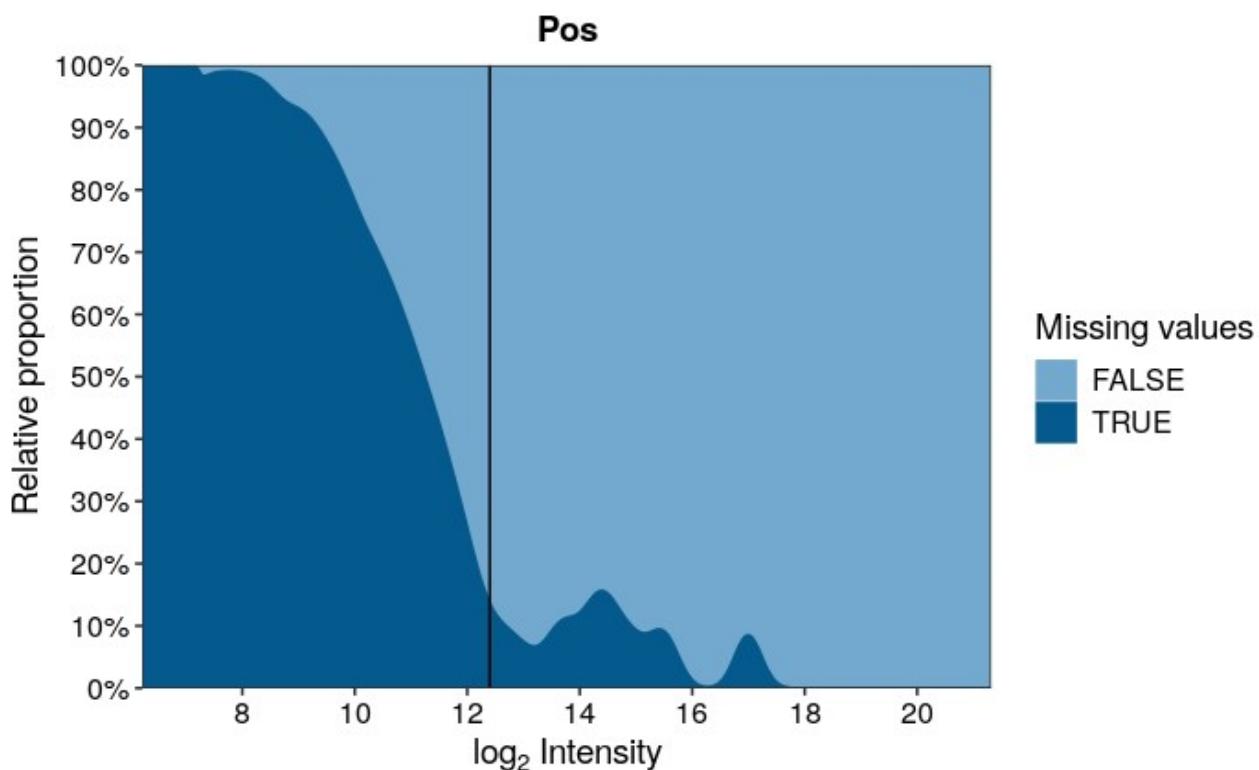


Figure 81: Stacked probability density plot produced by ImputeFinder on tangle-bearing neurons from the LCM Mass Spec dataset. The x-axis represents the mean intensity of each protein in the condition, while the y-axis is each protein's relative proportion of missing values, also isolated to the condition. The vertical line shows the inflection point, where proteins whose mean intensities within a condition are below the value are designated as having MNAR missing values, while those above are MAR.

Following inflection point selection is a housekeeping step, where proteins in each condition are ordered by their mean and separated into MNAR and MAR proteins. The mean was decided over median for this step, as a high-intensity value in one or more samples of a condition group may indicate that the protein is more likely MAR and using

the mean may push the protein towards being classed so. The rationale for this stems from the fact that mass spectrometry proteomics exhibits heteroscedasticity in the mean-variance relationship of proteins, though to a milder degree than RNA sequencing (Arneberg et al., 2007). This is the phenomenon that as mean protein intensity increases, so does the variance, therefore the presence of high-intensity outlier values may signal that the true intensities of the protein of interest, had there been no missing values, likely lean higher. Assuming this is true, then any missing values of that protein should be more appropriately classed as MAR.

After separation of missing value types, there is another important housekeeping step, which is to remove MAR proteins where the majority are missing. As discussed in (Gardner & Freitas, 2021) and reflected in my own testing, MAR imputation with a majority of missing values can lead to suspect imputation due to a lack of information. MNAR imputation however, does not seem to suffer from this limitation, as they generally employ simpler algorithms that focuses on imputing min values, which can operate reasonably with as a few as one complete value. Furthermore, it is logical that MNAR proteins would have a higher rate of missing values, as inferred from Figure 81. By imposing strict limitations on the number of complete values for MNAR proteins, the result would be near-complete loss of these proteins when their missing values are in fact the easier of the two to impute.

The final housekeeping step involves a set of intersections in order to ensure that all treatment conditions have MAR proteins with majority non-missing values. Out of the proteins that are majority non-missing in a single condition (for example 5 out of 8 samples of the condition), or have no missing values – when considering all other conditions, we only keep those that either satisfy these criteria as well, or are MNAR. In other words, we want to avoid retaining proteins that are MNAR in all groups, as this provides insufficient information for determining that a protein was detected in an experiment at all. This step aims to maximise retention of as many proteins as possible for downstream analysis, while removing those that lack the minimum necessary complete values. After this step, the data across all groups are subset to intersected proteins and separate objects are created that mark MNAR and MAR proteins for each treatment condition. The user can then apply whichever imputation algorithms they desire, the decision of which is out of scope for this tool.

For the LCM mass spec analysis, I opted to use k-nearest neighbor (kNN) for MAR proteins and the Minimum Probability method (MinProb) for MNAR proteins from the R package MSnbase (Gatto et al., 2021; Gatto & Lilley, 2012), both well established algorithms as discussed in Prior Art. As advocated throughout the ImputeFinder methodology, imputation was applied separately for tangle-bearing neurons and non-tangle-bearing neurons. This is highly important, particularly in the case of MAR proteins. Consider for instance a truly differentially abundant protein between the two groups, and that that protein has MAR missing values on both sides. Applying an MAR imputation method like kNN across all samples will impute values somewhere between the ranges of

the two groups, an outcome unreflective of the biological reality and that reduces the likelihood that it will be detected differentially abundant. But by applying the imputation within each group separately, the imputed values should lie within the ranges of that group solely, and have the intended effect of preserving differential abundance of the protein between the groups. Besides this point, it is also impossible to apply MAR imputation on one group and MNAR imputation on another group when applying imputation across all samples, a situation that I previously discuss to be biologically likely in the case of proteins that exhibit “on-off” dynamics due to a disease condition.

**Figure 82.**

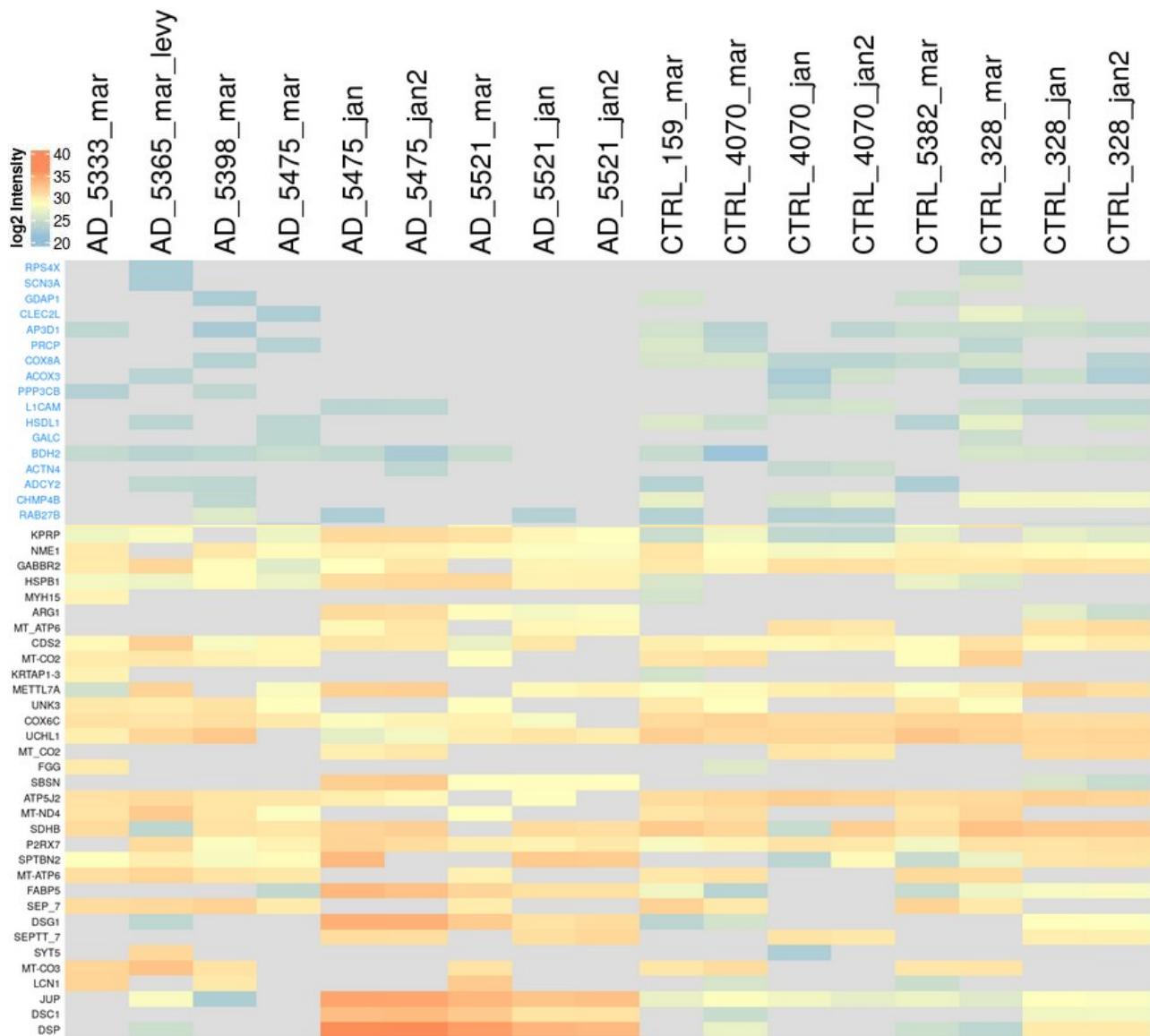


Figure 82: Visualisation of a subset of log<sub>2</sub> protein intensities across samples in a dataset associated with (Fowler et al., 2025). The data is subset to several proteins from the lowest range of values and several proteins from the highest range. Proteins designated MNAR are highlighted in blue and it can be seen that there are more missing values and lower intensity values in general. In contrast, proteins on the higher end of the intensity

range contain relatively fewer missing values. Visualisation of the entire set of proteins (not shown) would have the appearance that at the lower range of intensities, missing values decrease relatively linearly, until the inflection point, at which they become relatively more random.

**Figure 83.**

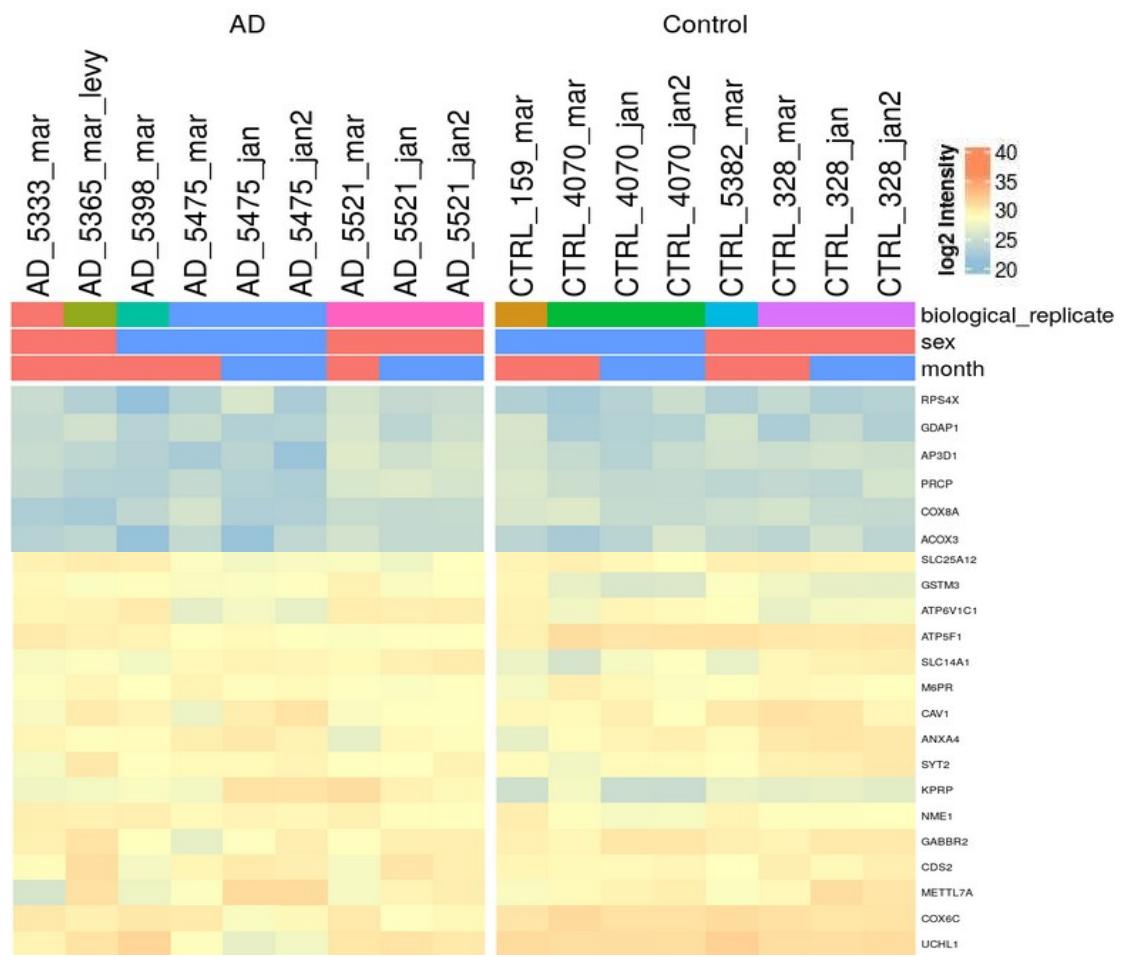


Figure 83: Visual assessment of kNN imputation for MAR and MinProb methods for MNAR proteins, applied per group, in a dataset associated with (Fowler et al., 2025). Like Figure 82, the data is subset to several proteins from the lowest range of values and several proteins from the highest range. The general consistency of imputed values within their ranges suggest that the imputation is reasonable.

### 3.6 Construction of Simulated Proteomics Dataset

As previously discussed, there was no comparable method to benchmark ImputeFinder against. While SFI-hybrid from (Gardner & Freitas, 2021) is comparable in concept, and lays groundwork for ImputeFinder, it lacks a software implementation, nor is an explicit method for designation of MAR and MNAR values within a dataset. Therefore, in order to better understand and test the validity of ImputeFinder, I designed a comprehensive simulation experiment that was published in (Fowler et al., 2025). In the experiment, a simulated dataset was generated to include 3,000 manually designated differentially expressed proteins (DEPs AKA DAPs), incorporating both MAR and MNAR missing

values. This dataset was built using a matrix of 6,105 complete-case log2 intensity values by creating a normal distribution on the mean and standard deviation of proteins in the real dataset (see Figure 84A i for histogram of the real data, and Figure 84B i for the simulated). To introduce DEPs into the dataset, intensity values for 250 random proteins were increased by a factor of 1.5 in specific groups, while 750 proteins in other groups had their intensity values halved (Figure 84B ii). The rationale for doing so was to simulate both upregulated proteins and proteins that are downregulated to the point of being below limits of quantification in a mass spectrometer.

MAR missing values were then simulated using the ampute function from the R package mice (Buuren & Groothuis-Oudshoorn, 2011), applying a constant 5% random missing data rate across the whole population of proteins. Simulation of MNAR missing values was applied to proteins with a mean intensity below 12, with greater sampling weight to lower-intensity proteins to reflect the asymptote towards complete missingness seen in real data. The histogram of the simulated data after introduction of these missing values is seen in Figure 84B iii and stacked probability density plots in Figure 85. An intensity of 12 was chosen as the amputation threshold to not bias all MNAR values towards only the induced DEPs that had values reduced. As highlighted by the vertical line in Figure 84B ii, there is a notable valley in the histogram around 8.5 due to the induced DEPs, therefore the amputation threshold should be somewhere above it to affect some proteins beyond these DEPs, better reflecting a real-world dataset.

In the final step, ImputeFinder is applied (using kNN for MAR and MinProb for MNAR) on the simulated, DEP induced, amputated data, with the result seen in Figure 84B iv. Inspection of the final imputed Figure 84B iv reveals that the log2 intensity distribution is similar to both Figure 84B ii (simulated with DEPs but no missing values) and Figure 84A ii (real data after imputation). This suggests that the imputation restored the state of having DEPs but no missing values, like in Figure 84B ii, implying that ImputeFinder succeeded in imputing missing values while retaining the induced DEPs. Furthermore, the experiment produced a final distribution that is similar to ImputeFinder applied on real data, implying that the simulated experiment is reflective of the set of processes observed in real data.

**Figure 84.**

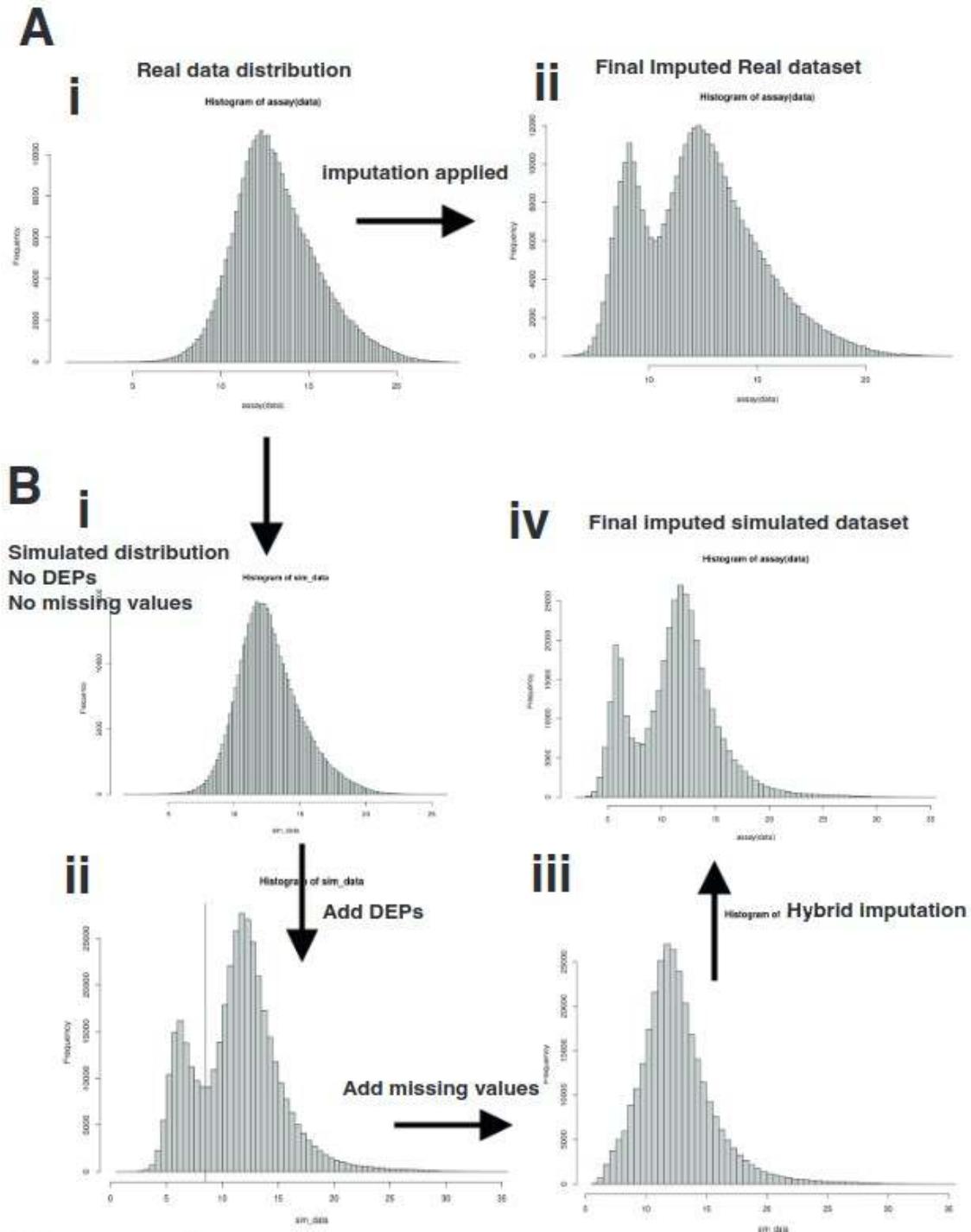


Figure 84: Sequence of transformations in the simulation experiment of ImputeFinder. The pipeline starts with real data in A i, which is simulated to produce the data in B I. DEPs are then introduced to the simulated data, reflected by the histogram in B ii. Next, missing values are introduced, producing the histogram in B iii. Finally, ImputeFinder is applied to produce data with the distribution in B iv. The distribution in B iv is similar to the distribution in A ii, which is real data after the ImputeFinder workflow. This similarity implies validity of the simulation method, while similarity to B ii implies validity in the use of ImputeFinder to address missing values while retaining DEPs.

**Figure 85.**

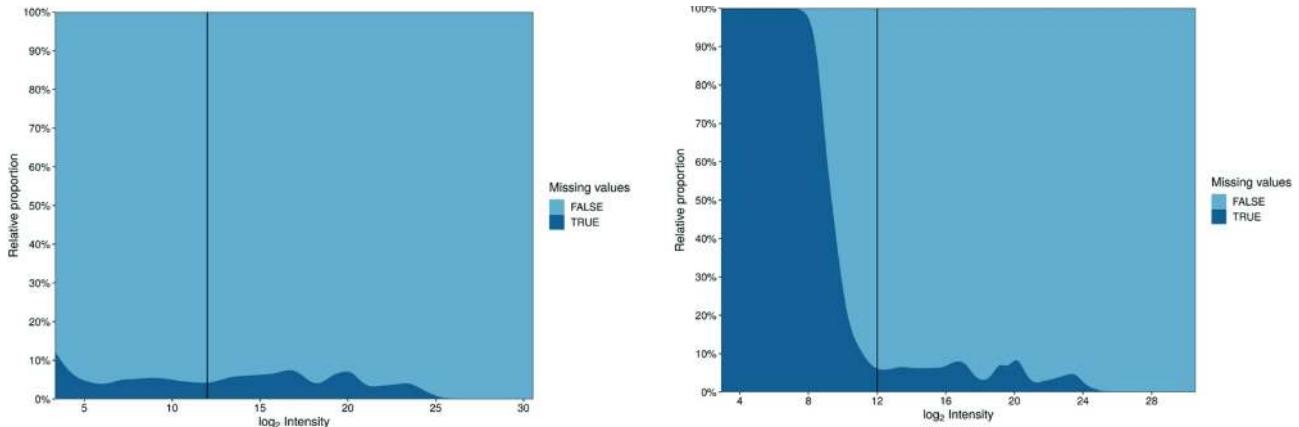


Figure 85: Stacked probability density plots of the simulated data at various steps of amputation of missing values. Left, is the simulated data after amputing randomly distributed MAR missing values. Right, is the data on the left after additional amputation of MNAR missing values, producing a final figure with an MNAR inflection point similar to that of real data (see Figure 81) and reflective of the MNAR amputation threshold value set at mean log<sub>2</sub> intensity of 12.

### 3.7 Benchmarking in Simulated and Real Dataset

Finally, I benchmarked the ability of ImputeFinder to identify the true DEPs introduced into the simulated dataset. For DEP testing, I used the R package limma (Phipson et al., 2016; Ritchie et al., 2015), as was used on the real dataset in (Fowler et al., 2025) and as recommended by the authors of DEP. limma (Linear Models for Microarray and RNA-Seq Data) is a widely used R package designed for the differential expression analysis of high-throughput gene expression data, including microarray, RNA-seq, and proteomics datasets. It employs an empirical Bayes approach to shrink variance estimates, with the core methodology being based on linear modeling and moderated t-statistics. Using the default limma workflow, I tested for DEPs (adjusted F p-value < 0.05) in the simulated dataset after introduction of DEPs and missing values under three conditions: 1) after application of ImputeFinder, 2) without imputation, and 3) after filtering to only those proteins without any missing values (complete-cases). I also carried a comparison of these three conditions in the real dataset. The results are summarised in Figure 86.

Effectively, in the simulated data, the results reveal that ImputeFinder is highly performant in capturing close to all of the true DEPs, with 92.97% captured, compared to 69.07% in the unimputed case and 24.03% in the complete cases. Furthermore, it was highly accurate, as 93.56% of the DEPs were those that were induced, though the unimputed and complete case conditions were also relatively accurate, at 85.80% and 93.27%, respectively. By maximising retention of likely DEP proteins with missing values, while removing those unlikely to be DEP, ImputeFinder also achieved the smallest averaged adjusted F p-values among the true DEPs, at 0.03333 compared to 0.2976, and 0.759 in the unimputed and complete case conditions, respectively. Finally, in Figure 86D, when

applying the three conditions to the real dataset, ImputeFinder discovered the largest number of DEPs, at 66.60% of the total proteins in the dataset vs. 53.05% without imputation and 24.03% when only looking at complete cases. Note that due to being a real dataset, it is not possible to verify whether a DEP is a real or not. However, it can be inferred from the simulation study that the DEPs uncovered in all cases are likely accurate, with ImputeFinder being potentially the most accurate.

**Figure 86.**

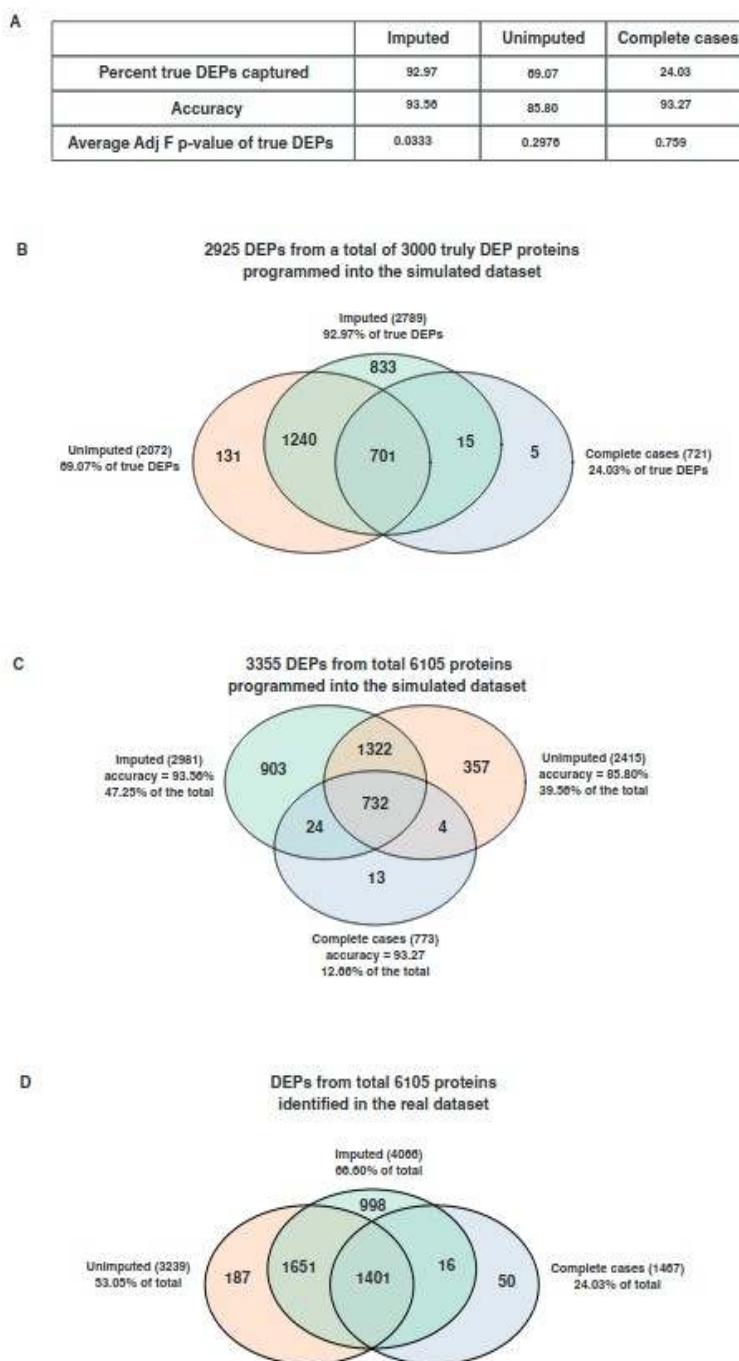


Figure 86: Benchmarking of the hybrid imputation strategy. A) For the simulated dataset containing 3,000 manually designated DEPs, the proportion of correctly identified DEPs, detection accuracy, and the average adjusted F p-values were assessed across imputed,

unimputed, and complete case conditions. B) A Venn diagram illustrating the overlap of DEPs identified from the set of 3,000 true DEPs in imputed, unimputed, and complete-case datasets. C) A Venn diagram showing DEPs detected from the full dataset of 6,105 proteins across imputed, unimputed, and complete-case datasets. These DEPs were compared against the 3,000 true DEPs to assess detection accuracy. D) A Venn diagram depicting DEPs identified from 6,105 total proteins in the real dataset across imputed, unimputed, and complete-case conditions. In all cases, DEPs were determined using a conventional limma protocol as described.

### 3.8 Sensitivity Analysis

In order to more clearly demonstrate the benefit of separating MAR and MNAR values using ImputeFinder, as well as benchmark the method against a variety of imputation methods, another simulation experiment was created. Starting with the same initial sample distribution established prior, DEPs were again introduced into the first 3,000 proteins for fraction groups F1-3, F4-6, and F7-8. Unlike the first simulation, an equal proportion of downregulated and upregulated proteins were introduced. For example, for the first 500 proteins, protein intensities in all samples of F1-3 were divided by 2, while for the next 500 they were increased by 1.5, and so on. 1.5 was selected as a realistic upregulation factor that produced a histogram without a noticeable shift in the overall intensities distribution (Figure 87). MAR amputation was increased from 5% of all proteins to 25% to demonstrate a more severe case of missing values that necessitate the need to separate MAR and MNAR missing values (Figure 88). All other properties were kept the same as the previous simulation.

**Figure 87.**

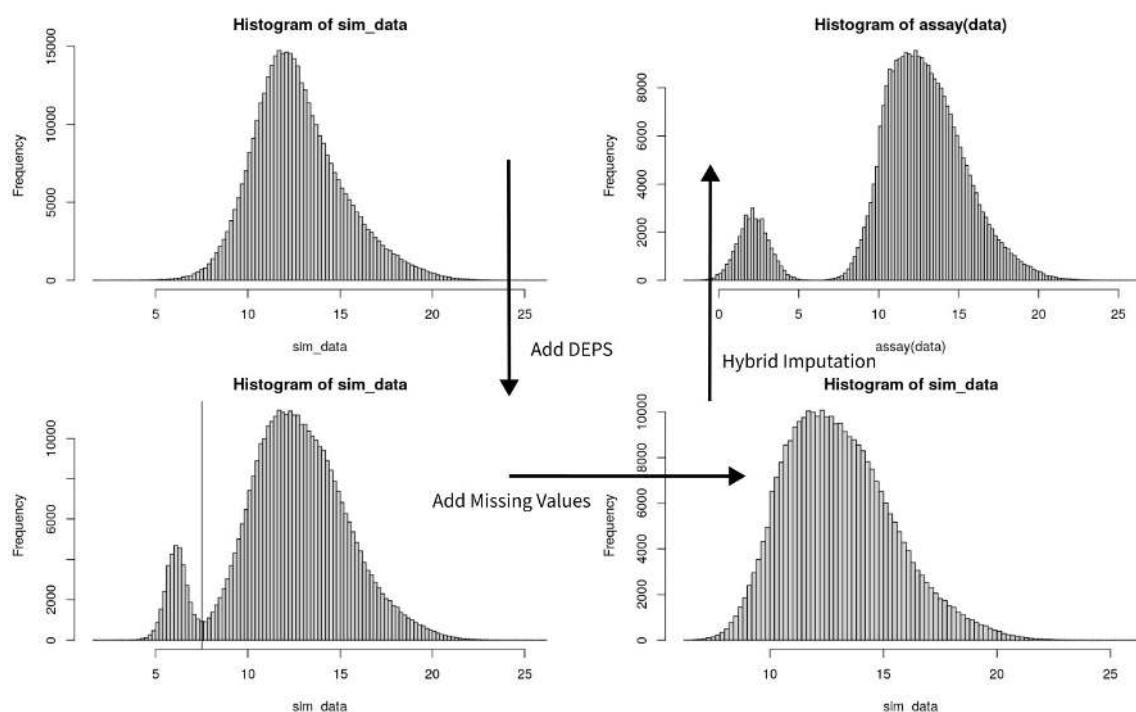


Figure 87: Schematic of synthetic dataset construction and outcome of imputation for the purposes of sensitivity analysis. The approach resembles that of Figure 84 but increases the number of upregulated proteins and introduces more MAR missing values.

**Figure 88.**

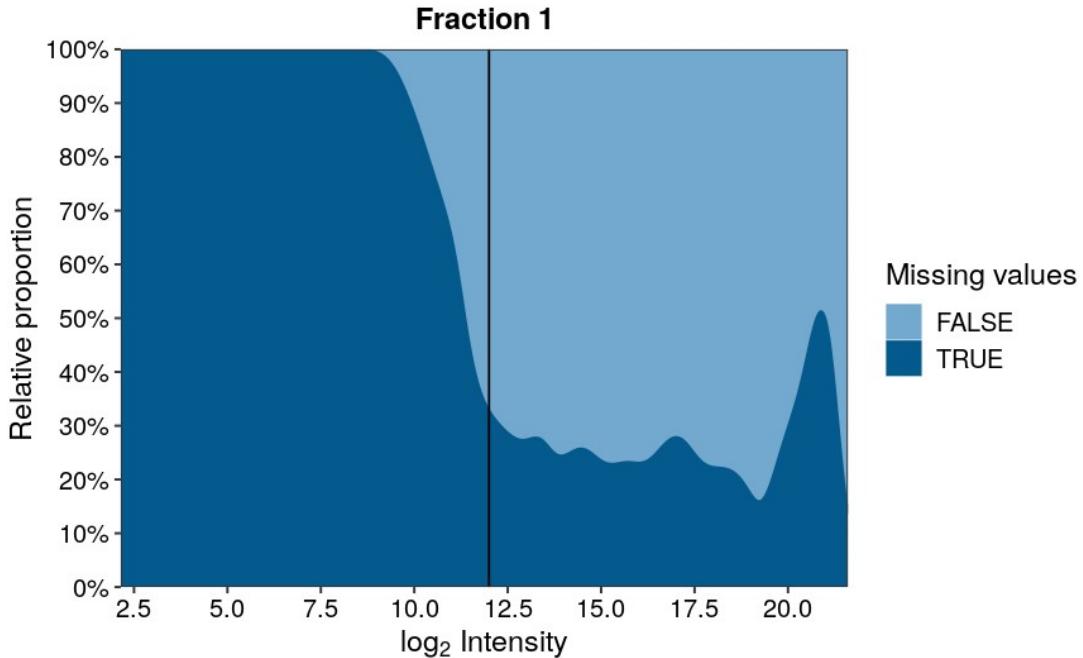


Figure 88: Stacked probability density plot on the sensitivity analysis simulation dataset. Note the inflated proportion of missing values past the inflection point.

To make the effect of separating MAR and MNAR clearer, a sweep of MNAR cut-off settings from 8 to 14 was run on the same simulated data. I chose this range to span the full transition, or “cliff,” in the relationship between intensity and missingness, where the probability of missingness rises steeply as intensity falls. The position of this cliff is visible in the stacked density plot with a vertical reference line.

For comparison I used the default imputation options in the DEP package. BPCA reconstructs missing entries with a low-rank Bayesian PCA model. KNN imputes from the nearest neighbours in expression space. QRILC performs quantile regression tailored to left-censored MNAR values. MLE fits a censored-normal model via the EM algorithm. MinDet replaces values with a small detection-limit estimate per sample. MinProb draws small values from a left-shifted distribution to reflect detection uncertainty. Min substitutes a simple small constant. Together these cover model-based and minimum-replacement strategies.

The figures that follow present three concise two-column tables, one per metric: percent true DEPs captured, accuracy, and the average adjusted F p-value among true DEPs. For each MNAR cut-off I pair the ImputeFinder result with the corresponding reference method, and I include unimputed and complete values for context.

**Figure 89.**

Percent True DEP Captured	
<i>Unimputed = 47.37 · Complete = 26.60</i>	
ImputeFinder	Reference Method
MNAR Cutoff 8 - 92.50	BPCA - 51.87
MNAR Cutoff 9 - 93.63	KNN - 62.17
MNAR Cutoff 10 - 93.53	QRILC - 79.73
MNAR Cutoff 11 - 93.77	MLE - 60.37
MNAR Cutoff 12 - 93.73	MinDet - 83.27
MNAR Cutoff 13 - 88.87	MinProb - 80.90
MNAR Cutoff 14 - 85.93	Min - 76.00

Figure 89: Performance of ImputeFinder in capturing true DEPs in the sensitivity analysis synthetic dataset across several MNAR cut-off settings (8–14), benchmarked against single reference methods (BPCA, KNN, QRILC, MLE, MinDet, MinProb, Min) and unimputed and complete cases.

**Figure 90.**

Accuracy	
<i>Unimputed = 99.44 · Complete = 99.38</i>	
ImputeFinder	Reference Method
MNAR Cutoff 8 - 85.89	BPCA - 92.90
MNAR Cutoff 9 - 85.33	KNN - 99.47
MNAR Cutoff 10 - 85.13	QRILC - 99.54
MNAR Cutoff 11 - 89.81	MLE - 95.87
MNAR Cutoff 12 - 93.17	MinDet - 99.56
MNAR Cutoff 13 - 91.40	MinProb - 99.47
MNAR Cutoff 14 - 92.40	Min - 99.35

Figure 90: Performance of ImputeFinder in capturing true positives in the sensitivity analysis synthetic dataset across several MNAR cut-off settings (8–14), benchmarked against single reference methods (BPCA, KNN, QRILC, MLE, MinDet, MinProb, Min) and unimputed and complete cases.

## Figure 91.

### Average Adjusted F P-value in True DEPs

Unimputed = 0.44380 · Complete = N/A

ImputeFinder	Reference Method
MNAR Cutoff 8 - 0.01814	BPCA - 0.29350
MNAR Cutoff 9 - 0.01329	KNN - 0.12870
MNAR Cutoff 10 - 0.01367	QRILC - 0.07212
MNAR Cutoff 11 - 0.01241	MLE - 0.12520
MNAR Cutoff 12 - 0.00795	MinDet - 0.06155
MNAR Cutoff 13 - 0.02121	MinProb - 0.06518
MNAR Cutoff 14 - 0.03629	Min - 0.09046

Figure 91: Performance of ImputeFinder in increasing statistical power to detect true DEPs in the sensitivity analysis synthetic dataset across several MNAR cut-off settings (8–14), benchmarked against single reference methods (BPCA, KNN, QRILC, MLE, MinDet, MinProb, Min) and unimputed and complete cases.

Across all MNAR cut-off settings, ImputeFinder was shown to be more sensitive for capturing true DEPs than any of the reference methods (Figure 89). The unimputed baseline performs poorly in this regard, with the lowest score of the reference methods aside from using complete cases only. Of the reference methods, MinDet and MinProb, methods designed to explicitly tackle MNAR missing values using simple approaches, captured the largest number of true DEPs, though these methods did not surpass ImputeFinder across any of the MNAR cutoff values chosen. Accuracy (Figure 90) shows a different pattern. Values close to 99% were observed for the unimputed and complete cases, as well as for several reference methods. Because ImputeFinder is markedly more sensitive, a modest reduction in accuracy is expected, since calling more positives has a tendency to increase both true and false positives. Accuracy should therefore be considered alongside the marked differences in sensitivity between methods. A better balance between ImputeFinder’s sensitivity and accuracy may be achieved through future optimisations of the method’s filtering steps.

The average adjusted F-value among true DEPs (Figure 91) is lowest for ImputeFinder, which indicates stronger statistical evidence for the signals it recovers. The minimum occurs at MNAR cut-off 12 (about 0.00795) while sensitivity remains high and, importantly, this setting also gives the highest accuracy among the ImputeFinder runs. Cutoffs towards the top or middle of the MNAR “cliff” in the stacked density plot likely leave too many MNAR values treated as MAR, which dilutes the imputation benefit. Whereas cutoffs past

the cliff classify too many values as MNAR, which can introduce inaccurate replacements and reduce performance, seen as the drop in sensitivity and worsening p-values at 13-14. Taken together, the three tables support using an MNAR cutoff around 12 for this dataset. Using the stacked density plot as visual guidance, this suggest a general rule of thumb for assigning the cutoff at the bottom of the MNAR cliff. Nevertheless, in this experiment, ImputeFinder was shown to be relatively robust to cutoff placement relative to the performance of the reference methods.

### 3.9 Advantages of ImputeFinder

ImputeFinder offers a structured and systematic approach to handling missing values in proteomics data by distinguishing between Missing at Random (MAR) and Missing Not at Random (MNAR) before imputation. Unlike some existing methods that impose specific imputation algorithms, ImputeFinder remains flexible and modular by providing a framework rather than implementing new imputation techniques. Researchers can integrate their preferred MAR and MNAR imputation methods, allowing them to tailor the workflow to the unique characteristics of their dataset. This targeted classification enables researchers to apply different imputation strategies tailored to each type of missingness, improving the accuracy and interpretability of downstream analyses.

Another key advantage of ImputeFinder is its per-condition imputation framework. Since the missingness mechanism for a given protein can vary across experimental conditions, being MNAR in one condition but MAR or fully observed in another, ImputeFinder ensures that imputation decisions are made in a condition-specific manner. This is in fact a common biological situation, particularly when a disease condition suppresses proteins in a particular group to below a mass spectrometer's limit of quantification. By separating types of missing values, this reduces the risk of overgeneralisation, preventing inappropriate assumptions about missing data mechanisms and preserving the biological validity of the dataset.

Additionally, ImputeFinder is implemented as an R package (discussed further in Implementation Details), making it accessible to the bioinformatics and proteomics research communities. The package is designed with usability in mind, enabling researchers to easily incorporate it into their workflows without requiring extensive computational expertise. The structured nature of the framework also enhances reproducibility by ensuring that the same missing value classification and imputation strategy can be consistently applied across multiple datasets.

### 3.10 Disadvantages and Limitations

While ImputeFinder introduces a useful framework for distinguishing between MAR and MNAR missing values and applying targeted imputation strategies, there are several limitations and areas that require further validation and refinement. These constraints primarily stem from the challenges associated with benchmarking, the empirical

determination of missingness classification thresholds, and the lack of comparable methodologies.

One major limitation is that ImputeFinder has only been benchmarked using a simulated dataset and a real-world proteomics dataset, though both of which demonstrate promising results. However, it was not possible to benchmark the method against comparable approaches, as no existing method explicitly separates MAR and MNAR missing values before applying distinct imputation strategies. This lack of direct comparison means that while the framework appears effective based on current testing, its performance relative to hypothetical alternative implementations remains unknown. Further independent validation across diverse datasets and experimental conditions would be beneficial to confirm the method's generalisability and robustness.

Another key limitation lies in the empirical establishment of an intensity cutoff for MNAR classification. The framework requires manually determining an inflection point in protein intensity, below which missing values are classified as MNAR, and above which they are assumed to be MAR (or potentially complete). While the benchmarking experiments suggest that an inflection towards the bottom of the slope in the stacked probability density plot results in high accuracy and sensitivity for differentially abundant proteins, an exhaustive exploration determining the optimal and precise placement of this inflection point is an area of future investigation. Furthermore, while references support a likely linear relationship between missingness and protein intensity in MNAR cases, this remains a subject of debate in the field (M. Li & Smyth, 2023; R. Luo et al., 2009; O'Brien et al., 2018). The lack of a rigorous, universally accepted model for defining MNAR thresholds means that some subjectivity is involved in the current classification approach. This is particularly relevant in datasets where the intensity-missingness relationship deviates from expected trends, potentially affecting classification accuracy.

Additionally, the current method does not automate the selection of the MNAR inflection point, requiring users to determine it manually for each dataset. While this provides flexibility, it also introduces a degree of user-dependent variability, which could lead to inconsistencies in classification across studies. Automating this step, potentially through data-driven approaches such as breakpoint detection, Bayesian modeling, or machine learning-based classification, could increase reproducibility and reduce potential bias introduced by manual selection.

Finally, ImputeFinder does not introduce new imputation algorithms but instead relies on existing MAR and MNAR imputation methods. While this modularity is advantageous, it also means that the effectiveness of the framework is dependent on the quality of the selected imputation strategies. In cases where imputation methods do not perform well for a given dataset, the framework itself cannot compensate for this limitation. Future developments could explore guiding users toward optimal imputation choices based on dataset characteristics, perhaps through internal benchmarking within the package.

### 3.11 Implementation Details

ImputeFinder is currently being prepared for submission to Bioconductor, with its package structure and documentation generated using the R package `biocthis` to ensure compatibility with Bioconductor's standards. This means that the package has been designed with ease of integration in mind, allowing researchers to incorporate it into existing proteomics workflows seamlessly. Its implementation is lightweight, requiring minimal dependencies and prioritizing flexibility in handling missing data.

**Figure 92.**

Icolladotor Merge pull request #16 from milanmlft/master ...		
 .github	Also use plural here	17 days ago
 R	Let use_bioc_readme_rmd() add README.Rmd to .Rbuildignore	4 days ago
 actions	Add Marcel Ramos as a contributor thanks to #11 ^^	2 months ago
 dev	Improve workflow steps after working through them	6 months ago
 inst	Also use plural here	17 days ago
 man	Add the 'report_bioc' argument to use_bioc_description() to support th...	2 months ago
 tests	Resolve #13	18 days ago
 vignettes	Add a footnote about badger. Related to #6	2 months ago
 .Rbuildignore	Add the GHA workflow from leekgroup/derfinderPlot@4124541. Relate...	7 months ago
 .gitignore	add vignette template	7 months ago
 DESCRIPTION	v1.1.2 -- bump version after recent changes	17 days ago
 NAMESPACE	Add the biocthis_example_pkg() function to mask usethis::proj_set() f...	6 months ago
 NEWS.md	Update the NEWS.md file with the current information	7 months ago
 README.Rmd	Match r-lib/actions@8198dc0	18 days ago
 README.md	Match r-lib/actions@8198dc0	18 days ago
 codecov.yml	Add badges	7 months ago

Figure 92: The project structure advocated by the R package `biocthis` for submission to Bioconductor. Figure reproduced from [https://dzhang32.github.io/biocthis\\_workshop/](https://dzhang32.github.io/biocthis_workshop/).

A key feature of ImputeFinder is its reliance on `ggplot2` (Wickham, 2016) as its only major dependency. The package includes a modified missingness-intensity plot originally derived from the `DEP` package, but it does not require `DEP` itself, keeping the dependency footprint small while still providing meaningful visualizations. This ensures that users can benefit from informative graphical representations of missing data patterns without the need for extensive additional installations.

The input required for ImputeFinder is straightforward. Users provide a matrix of unnormalised log2 protein intensities, where rows represent proteins and columns correspond to samples. Additionally, users supply a data frame of group assignments, defining the condition or experimental group for each sample. Using this information, ImputeFinder classifies missing values per group, ensuring that a protein's missingness type is assessed within its specific experimental context rather than across the entire dataset. Beyond this, ImputeFinder contains no tweakable parameters nor additional input, an intentional design decision to reduce complexity for the user and limit harmful practices like data dredging.

The output of ImputeFinder is structured as a list, where each element corresponds to a distinct experimental group. Within each group-specific entry, the method provides two key outputs: the list of proteins classified as MNAR and the set of MAR or complete cases. This classification enables users to refine their dataset based on missingness type, ensuring that MAR and MNAR values can be handled separately in downstream analysis.

Currently, ImputeFinder does not perform imputation itself but instead provides a structured classification of missing values, allowing researchers to apply their own imputation strategies. Users are responsible for following ImputeFinder with their normalisation strategy of choice, subsetting the protein matrix based on the intersection of proteins across groups, and then selecting the appropriate imputation techniques for each type of missingness. This modular approach ensures that researchers retain full control over the overall pipeline while benefiting from the improved accuracy that results from distinguishing between MAR and MNAR values. Future iterations of ImputeFinder may expand to further streamline the workflow by automating certain steps. Enhancements could include selection of common normalisation methods, automated filtering to subset the protein matrix across groups, built-in imputation pipelines to handle MAR and MNAR values separately, and improved visualisation and reporting functionalities for missing data patterns.

### Figure 93.

```
impute1 <- data[ , which(data$condition == "Neg")]
impute_vector <- rownames(impute1) %in% Neg_MNAR # Logical vector specifying MNAR.
set.seed(1)
impute1 <- impute(impute1, "mixed", randna = !impute_vector, mar = "knn", mnar = "MinProb")

impute2 <- data[ , which(data$condition == "Pos")]
impute_vector <- rownames(impute2) %in% Pos_MNAR
set.seed(1)
impute2 <- impute(impute2, "mixed", randna = !impute_vector, mar = "knn", mnar = "MinProb")
```

Figure 93: Code showing how separated MNAR and MAR values can be imputed using different methods from the *impute* function of DEP. This example is directly lifted from the analysis featured in this thesis work on the LCM Mass Spec dataset. It demonstrates the application of mixed imputation on the tangle-bearing negative and tangle-bearing positive groups separately.

## 4. Development of GeneFunnel Gene Set Enrichment Method

GeneFunnel was developed to address various deficiencies in existing gene set enrichment methods. It falls within the subset of methods known as functional class scoring (FCS), originating from single-sample GSEA (ssGSEA) (Barbie et al., 2009), which aim to produce enrichment results per-sample, generally producing a matrix of gene sets by samples that resemble the original gene/protein by sample matrix of which it is derived. However, I found that all existing methods fail to retain independence between samples and produce results that are unintuitive and difficult to reason with. Explicitly, the following issues among these methods are discussed in detail in this section: 1) the handling of missing and lowly expressed features. 2) consideration of dependencies between samples, between features, and interactions across the two. 3) retention of statistical properties of the input data. 4) consideration of complexity and assumptions. 5) compatibility with downstream handling of data and interpretation. 6) speed and scalability. In order to overcome these issues, I developed a new R package called GeneFunnel with a performant C++ backend available at <https://github.com/eturkes/genefunnel>. It is in preparation for submission to the Bioconductor repository of bioinformatics tools for R.

### 4.1 Definition and Description of Gene Set Enrichment

Gene set enrichment, reviewed extensively in (Bayerlová, 2015; Bull et al., 2024; Candia & Ferrucci, 2024; Das et al., 2020; Geistlinger et al., 2020; Khatri et al., 2012; Maleki et al., 2020; Wijesooriya et al., 2022), is a widely used computational method designed to identify biologically meaningful patterns in gene expression data by assessing whether predefined gene sets are significantly overrepresented in a dataset. Unlike traditional differential expression analysis that examines individual genes (or proteins) in isolation, gene set enrichment evaluates groups of genes that share functional relationships, such as involvement in metabolic pathways, cellular processes, or disease mechanisms. These gene sets are typically sourced from curated databases like KEGG (Kyoto Encyclopedia of Genes and Genomes), GO (Gene Ontology), Reactome, and the Molecular Signatures Database (MSigDB), which classify genes based on shared biological roles, molecular functions, or regulatory pathways. By focusing on collective gene behaviour rather than single-gene changes, gene set enrichment provides deeper insights into the molecular mechanisms underlying phenotypic differences.

There are several approaches to gene set enrichment analysis, with over-representation analysis (ORA) being one of the most straightforward. ORA involves comparing a predefined list of differentially expressed genes against curated gene sets to determine statistical overrepresentation, typically using Fisher's exact test or a hypergeometric test. However, ORA has limitations as it relies on arbitrary cutoffs for selecting DEGs, potentially overlooking biologically relevant changes below statistical thresholds. To overcome this, Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) uses a rank-based method that considers the entire dataset rather than a predefined cutoff. GSEA ranks genes based on their correlation with a phenotype and calculates an enrichment

score (ES), which quantifies the degree of overrepresentation of a gene set within the ranked list. This method enables the detection of subtle but coordinated gene expression changes, which are often biologically significant but might be missed by conventional differential expression approaches. Statistical significance is assessed through permutation testing, generating a null distribution of ES values to compute FDR-adjusted p-values.

Another widely used extension of GSEA is single-sample GSEA (ssGSEA), generalised as functional class scoring (FCS), which incorporates gene-level information into pathway-level enrichment scores. These methods aim to calculate enrichment scores per-sample, allowing for continuous comparisons across conditions rather than binary classifications of enriched or non-enriched pathways. And unlike traditional GSEA, it does not rely on a ranking of genes across predefined groups. This results in a matrix of enrichment scores that can be further analysed using clustering, dimensionality reduction, statistical testing, or correlation with phenotypic traits. Additionally, ssGSEA and other FCS-based approaches may better account for subtle variations in pathway activity within heterogeneous datasets by independently calculating enrichment scores for each sample, rather than relying on predefined case-control comparisons. This per-sample scoring enables the detection of gradual or condition-specific pathway activation patterns that may not be apparent in population-wide differential expression analyses.

Topology-based methods in gene set enrichment analysis extend traditional enrichment approaches by incorporating the structural relationships between genes within biological pathways. Unlike classical overrepresentation or functional class scoring methods, which primarily evaluate gene sets as simple lists, topology-based approaches account for the connectivity, interactions, and hierarchical organisation of genes within pathways. These methods leverage pathway graphs, where nodes represent genes and edges denote regulatory or signalling interactions, to refine enrichment calculations by weighting genes based on their topological significance. By incorporating pathway structure, topology-aware enrichment methods may provide a more biologically meaningful interpretation of gene expression changes.

Gene set enrichment has a wide range of applications across biomedical research. It is frequently used to identify dysregulated pathways in diseases, such as pinpointing key signalling cascades in neurodegenerative disorders like Alzheimer's Disease. It also plays a crucial role in drug discovery, where gene expression signatures from treated samples can be compared to pathway databases to infer potential mechanisms of action. And in the context of functional genomics studies, gene set enrichment can enhance the interpretation of high-throughput screening results by contextualising gene expression changes within known biological processes.

Despite its advantages, gene set enrichment is not without limitations. Its accuracy depends on the quality and completeness of gene set databases, meaning that less well-characterised pathways may be overlooked. Additionally, gene sets often contain highly

correlated genes, which can introduce bias in enrichment scoring. Another challenge is the reliance of some methods on ranking methods, of which there are many and can drive dramatic differences in the outcome of enrichment testing beyond the enrichment test itself (Zyla et al., 2017). Nonetheless, gene set enrichment remains an essential and common tool in the analysis of various omics assays, fuelling a very active field of research within bioinformatics.

**Figure 94.**

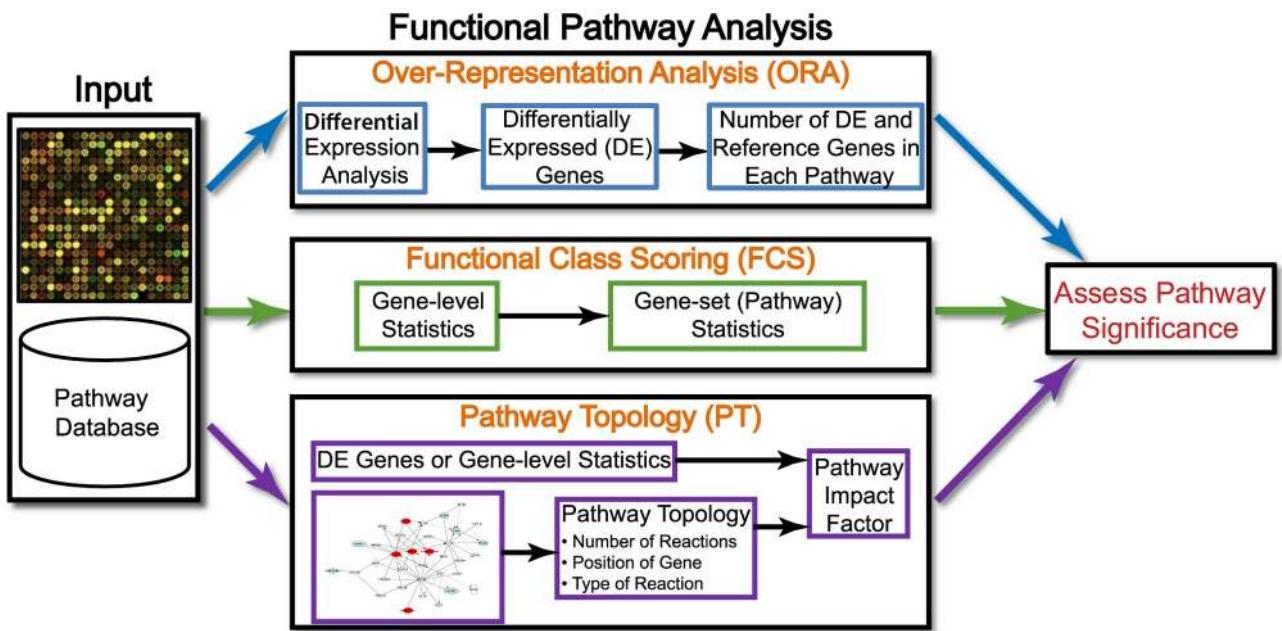


Figure 94: Overview of the major classes of gene set enrichment methods, covering those of Over-Representation Analysis (ORA), Functional Class Scoring (FCS), and Pathway Topology (PT). Figure reproduced from (Khatri et al., 2012).

## 4.2 Prior Art: Functional Class Scoring

GeneFunnel classifies as a functional class scoring (FCS) method, so I will focus this review on other leading FCS methods, all of which I later benchmark against GeneFunnel. As mentioned in the prior section, FCS leverages continuous expression data from all genes without needing a significance cutoff to compute pathway-level scores, producing a per-sample (or per-cell) enrichment score. The output being a matrix was a key appeal from my perspective, as it opens the possibility for further analysis using general purpose methods that operate on matrices such as linear modelling and dimensionality reduction, provided statistical assumptions are met. The different FCS methods elect a variety of computational strategies, the assumptions and pitfalls of each I discuss below.

Gene Set Variation Analysis (GSVA) is a popular FCS method that transforms gene expression data from a gene level matrix into a gene set level matrix of enrichment scores

(Hänelmann et al., 2013). GSVA uses a kernel-based, non-parametric approach to estimate the cumulative distribution function (CDF) of expression for each gene across the sample population. In practice, GSVA replaces the expression values with kernel-smoothed ranks (CDF estimates) and then calculates a KS-like (Kolmogorov-Smirnov) enrichment score per-sample by comparing the distribution of expression values inside the gene set to those outside. This yields a continuous pathway activity score for each sample without requiring class labels. Essentially, GSVA assesses how up or downregulated a gene set is in a given sample relative to the overall dataset.

GSVA was introduced to handle heterogeneous data and subtle expression changes. It has been shown to increase power for detecting modest but coordinated pathway shifts, and has been shown to work on both microarray and RNAseq data (after appropriate normalisation). A key feature is its unsupervised nature, it can be applied to a single cohort of samples to reveal variation in pathway activity across conditions or continuous phenotypes. However, one important aspect is that GSVA borrows information across samples as the kernel estimation uses the entire sample set as context. This means GSVA scores are relative to the given dataset; if the composition of samples changes (e.g. adding or removing samples), the scores can shift. As a result, GSVA performs well with sufficiently large sample sizes, but can become unstable in very small cohorts (since the CDF estimation for each gene is less reliable). For scRNAseq, GSVA can be applied by treating each cell as a sample, and though it has been used in single-cell studies, the computational cost can be significant for large cell numbers, as I will later show through benchmarking. Moreover, in extremely sparse single-cell data, many genes have zero counts in most cells, which can make the kernel-based estimation less informative. Despite these challenges, GSVA remains a widely-used baseline for single-sample gene set scoring due to its robustness in detecting subtle pathway variation.

**Figure 95.**

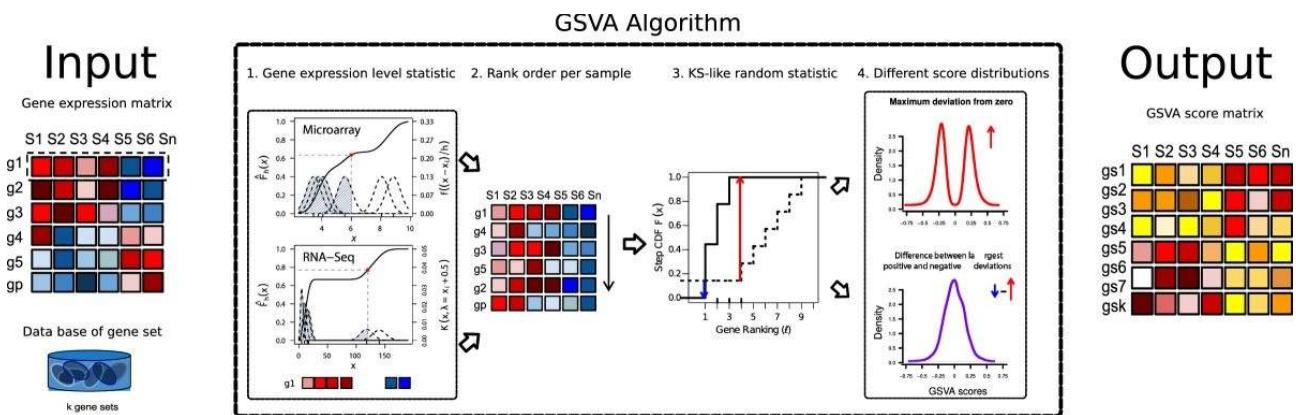


Figure 95. Schematic of the GSVA algorithm. The GSVA algorithm takes as input a gene expression matrix, typically composed of log2-transformed microarray expression values or RNAseq count data, along with a collection of predefined gene sets. For each gene set, a KS-like rank statistic is computed to assess its enrichment. The algorithm generates an

output matrix in which each entry represents a pathway enrichment score corresponding to a specific gene set and sample. Figure reproduced from (Hänelmann et al., 2013).

Single-sample GSEA (ssGSEA) is a sample-specific adaptation of the classic GSEA method (Barbie et al., 2009). Instead of comparing two groups, ssGSEA computes an enrichment score for each gene set for each sample independently, using that sample's ranked gene expression profile. The method ranks all genes by expression within a given sample, then calculates an enrichment score by a cumulative rank distribution comparison, essentially measuring if genes in the set tend to appear at the top (highly expressed) or bottom (lowly expressed) of that sample's rank list. This is done by computing the difference between two empirical CDFs (genes inside the set vs. outside) like the original GSEA KS statistic, yielding a score that can be positive or negative.

Because ssGSEA is rank-based, it is naturally robust to outliers in gene expression and differences in measurement scale; it relies only on the relative ordering of genes within a sample. It also does not strictly require multiple samples and in principle one can compute ssGSEA on a single sample or cell. However, standard ssGSEA implementations (e.g. in the GSVA R package) perform a final normalisation that uses the score distribution across all samples, which can introduce some inter-sample dependency. Truly single-sample versions, like in (Barbie et al., 2009) where it was first described, avoid using other samples as a reference. In practice, ssGSEA yields an enrichment score per-sample that is intuitively similar to the original GSEA's NES (normalised enrichment score) but on a per-sample basis.

In terms of performance, ssGSEA has been widely used for bulk RNAseq and has also been applied to single-cell data by computing per-cell scores. However, ssGSEA can be very computationally intensive for large datasets because it effectively performs a sort and cumulative sum for each sample and each gene set. Benchmarking studies, including my own, have noted that ssGSEA tends to be one of the slowest methods at large scale (X. Wang et al., 2024). Additionally, because ssGSEA (in its usual form) yields scores that may depend on the overall expression distribution of a dataset, its robustness in small sample sets is not ideal, i.e. when few samples are available, the score normalisation can be biased.

Pathway Level Analysis of Gene Expression (PLAGE) takes a different approach to single-sample gene set scoring by using matrix factorisation to estimate pathway activity (Tomfohr et al., 2005). PLAGE assumes that if a pathway is coherently activated, the genes in that set will show coordinated expression across samples. The PLAGE algorithm first standardises the expression matrix for the gene set (z-scoring each gene across all samples) so that all genes contribute equally regardless of their absolute expression level. It then performs a singular value decomposition (SVD) on this standardised matrix. The first singular vector represents the dominant expression pattern shared by those genes across the samples, and the values of that vector for each sample are taken as the pathway activity level in that sample. PLAGE collapses the gene set into a single latent

factor that best explains the expression variation of the gene set, and uses that factor score as the enrichment score for the pathway.

One advantage of PLAGUE is that by using SVD it inherently accounts for correlations between genes in the set. Genes that consistently co-vary will contribute strongly, whereas uncorrelated noise will be deemphasised. This can improve sensitivity for pathways where many genes change modestly in unison. However, PLAGUE also has some limitations. Because it uses all samples to perform the SVD, it depends on having a reasonably large sample set to get stable estimates, and in very small datasets, the first SVD may not be reliably estimated or could pick up random variation. For instance, it has been shown that PLAGUE's performance deteriorates with small sample size, showing unstable scores when the number of samples is low, particularly in comparison to other methods (Figure 96) (Foroutan et al., 2018). This is likely because with few samples, the co-expression structure is hard to distinguish from noise. Another consideration is that PLAGUE's assumption of one dominant factor may not hold if a pathway has multiple independent modes of variation.

**Figure 96.**

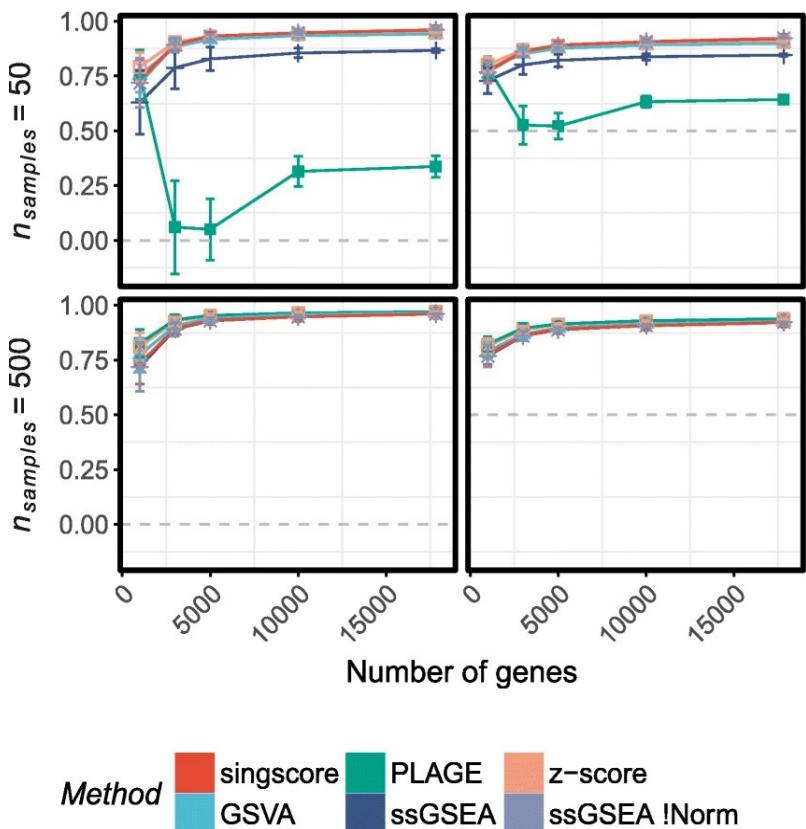


Figure 96: Excerpt of a stability analysis comparing several FCS methods in different transcriptomic datasets when altering sample size and number of genes. The left-most figures show Spearman's correlation coefficients and the right-most figures show concordance index, where higher values in each indicate greater robustness to the changing conditions. Figure adapted from (Foroutan et al., 2018).

In contrast to those discussed thus far, a straightforward, but not uncommon method, is Z-score summation (implemented in the GSVA package). This approach simply standardizes expression of each gene across samples and sums the z-scores of genes in the set for each sample. This yields a gene set score roughly indicating how many standard deviations each member feature is above or below its mean. While simple and fast, this approach assumes independence of features and can be sensitive to one or two highly expressed features. It is generally less sophisticated than GSVA/ssGSEA, but provides a quick heuristic pathway score. It can be applied to single cells, but zero-inflated data can abnormally push many z-scores to negative values. The z-score method was found to have intermediate stability, more stable than PLAGE in small-sample scenarios but still influenced by dataset composition. It often serves as a baseline for more complex methods.

#### 4.3 Methodology of GeneFunnel

The novel functional class scoring algorithm I propose in this work, GeneFunnel, is relatively straightforward, and is similar to subtracting the Mean Absolute Deviation (MAD) from the sum of values in a gene set. GeneFunnel iterates through each gene set for each sample, introducing no dependency between samples or gene sets. For each gene set, in the current sample, the sample's genes/proteins (AKA features) are subset to those in the gene set. The sum and mean are taken. Then, for each feature in the gene set, the feature's expression level is subtracted from the mean and the absolute value is taken. These "deviances" are then summed. A scaling factor is then derived, defined as the size of the gene set divided by twice the residual gene set size (1 minus the gene set size). Finally for each gene set, the summed deviance is multiplied by the scaling factor and then subtracted from the sum, yielding the score. This is done for all gene sets in a sample, before iterating through the rest of the samples, producing a gene set by sample matrix resembling the original gene/protein by sample matrix (Figure 97). The algorithm is expressed in mathematical notation below and an excerpt of the Rcpp (C++ interface for R) (Eddelbuettel & Balamuta, 2018; Eddelbuettel & François, 2011) implementation in Figure 98.

**Figure 97.**

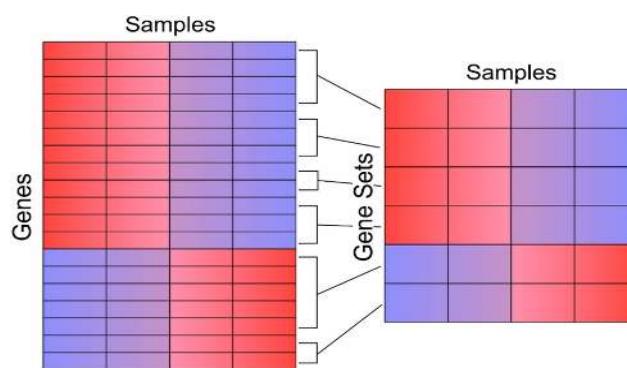


Figure 97. High-level schematic of the intent of GeneFunnel. From an initial matrix of genes (or proteins) by samples, and with the provision of an object containing gene sets, the input matrix is transformed into a gene set by sample matrix.

### **Mathematical Description of the GeneFunnel Algorithm:**

$$\text{score}_{k,j} = \sum_{i \in G_k} X_{i,j} - \left( \frac{|G_k|}{2(|G_k| - 1)} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}| \right)$$

Where:

- $X_{i,j}$  is the expression level of feature  $i$  in sample  $j$ ,
- $\sum_{i \in G_k} X_{i,j}$  is the sum of expression for the features in gene set  $G_k$  for sample  $j$ .
- $\bar{X}_{G_k,j}$  is the mean expression of the features in gene set  $G_k$  for sample  $j$ .
- $\sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}|$  is the sum of the absolute deviations from the mean.
- $\frac{|G_k|}{2(|G_k| - 1)}$  is the scaling factor, which accounts for the number of features in the gene set and adjusts the influence of deviation.

**Figure 98.**

```
NumericMatrix calculateScores(
    const arma::sp_mat& orig_mat, CharacterVector row_names, List gene_
) {
    int ncol_mat = orig_mat.n_cols;
    int nrow_list = gene_ids.size();

    NumericMatrix mat(nrow_list, ncol_mat);

    std::unordered_map<std::string, uword> row_map;
    for (uword i = 0; i < row_names.size(); ++i) {
        row_map[as<std::string>(row_names[i])] = i;
    }

    for (int j = 0; j < ncol_mat; ++j) {
        for (int i = 0; i < nrow_list; ++i) {
            CharacterVector gene_set = gene_ids[i];
            std::vector<uword> indices;

            for (int m = 0; m < gene_set.size(); ++m) {
                std::string gene = as<std::string>(gene_set[m]);
                if (row_map.find(gene) != row_map.end()) {
                    indices.push_back(row_map[gene]);
                }
            }

            vec idx_values(indices.size());
            for (size_t k = 0; k < indices.size(); ++k) {
                idx_values[k] = orig_mat(indices[k], j);
            }

            double sum_values = sum(idx_values);
            double var_values = sum(abs(idx_values - mean(idx_values)));

            size_t size = idx_values.size();
            double factor = static_cast<double>(size) / (2.0 * (size - 1));
            double score = sum_values - (var_values * factor);

            double epsilon = 1e-9;
            if (fabs(score) < epsilon) {
                score = 0.0;
            }

            mat(i, j) = score;
        }
    }

    return mat;
}
```

Figure 98: Rcpp implementation of the singular GeneFunnel function, *calculateScores*. The function is highly optimised for performance, using RcppArmadillo linear algebra libraries (Eddelbuettel & Sanderson, 2014).

At the core of GeneFunnel's scoring method is its use of both the sum and deviation of feature expression levels within a gene set. By first summing expression values, the method captures the overall activity level of a pathway, akin to approaches that rely on simple averaging or summation. However, instead of assuming that all features contribute equally, GeneFunnel then computes deviance scores for each feature, measuring how much each feature's expression deviates from the mean expression of the set. This deviation-aware component ensures that pathways with highly variable expression across member features are penalised, preventing scenarios where a small number of highly expressed features dominate the enrichment score.

The scaling factor applied to the summed deviance accounts for gene set size, ensuring that scores remain comparable across gene sets of different sizes. This is important because raw summation-based approaches can be biased toward larger gene sets, simply due to the additive nature of their scoring. By normalising the deviation penalty relative to gene set size, GeneFunnel maintains a balance between total expression and internal variability, making it more sensitive to pathway coherence rather than just overall expression magnitude.

Another important consideration in GeneFunnel's design is that it treats gene sets independently, meaning that the score for one pathway is not influenced by the composition of other gene sets or the overall dataset structure. This makes it particularly well-suited for applications where absolute (or as close to it as possible) pathway activity is the desired measure. Ideally, this may allow GeneFunnel scores to be compared across disparate datasets, and well as have suitability for meta-analyses. Unlike methods that rely on ranking or distribution-based transformations, GeneFunnel's approach remains resistant to dataset size changes, such as the addition or removal of samples, ensuring that scores remain interpretable even when analysing a single sample in isolation.

In summary, GeneFunnel provides a functional class scoring method that integrates total pathway activity with an internal consistency check through deviation scoring, ensures independence across samples and gene sets, and incorporates a scaling factor for size correction. These design choices make it particularly advantageous in settings where existing enrichment methods struggle with dataset-wide dependencies, small sample sizes, or variable gene set sizes.

#### 4.4 Mathematical Proof of Non-negative Scores

A fundamental requirement for GeneFunnel is that its scores remain non-negative, ensuring compatibility with common downstream analyses in functional genomics. The goal is to transform a standard gene/protein by sample expression matrix into a gene set by sample matrix, preserving key properties that allow established bioinformatics techniques, such as dimensionality reduction, normalisation, and differential expression analysis, to be applied seamlessly. Many of these methods, including log-transformation, require non-negative inputs, making it essential that GeneFunnel does not yield negative scores. Intuitively, a negative pathway activity score would be biologically meaningless, as it would imply an inversion of expression that contradicts the additive nature of gene set aggregation. Instead, the method is designed such that the minimum possible score is zero, which occurs in two biologically interpretable cases: when all features in the set have zero expression or when the set exhibits maximal internal deviation, meaning that the expression values are so dispersed that the deviation term fully offsets the total summed expression (i.e. the case when a single value is non-zero). Proving that GeneFunnel always produces non-negative scores formally validates that it is a proper transformation of gene expression data, ensuring interpretability and compatibility with standard computational workflows.

## Theorem: GeneFunnel Scores Cannot be Negative

Let  $X_{i,j}$  be the expression level of feature  $i$  in sample  $j$ , and let  $G_k$  be a predefined gene set containing  $|G_k|$  features. The GeneFunnel score for gene set  $G_k$  in sample  $j$  is given by:

$$\text{score}_{k,j} = \sum_{i \in G_k} X_{i,j} - \left( \frac{|G_k|}{2(|G_k| - 1)} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}| \right)$$

Then, for all  $k$  and  $j$ :

$$\text{score}_{k,j} \geq 0.$$

### Proof:

We begin by expanding the sum of values in the gene set (found left-hand-side or LHS of the parenthesis), where in a general case, the sum of values is equal to the mean of values times the number of values:

$$\sum_{i \in G_k} X_{i,j} = |G_k| \bar{X}_{G_k,j}$$

Substituting this into the scoring equation, located LHS of the parenthesis, we obtain:

$$\text{score}_{k,j} = |G_k| \bar{X}_{G_k,j} - \left( \frac{|G_k|}{2(|G_k| - 1)} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}| \right)$$

Factoring out  $|G_k|$  from that substitution, and from within the parenthesis, simplifies the score to:

$$\text{score}_{k,j} = |G_k| \left( \bar{X}_{G_k,j} - \frac{1}{2(|G_k| - 1)} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}| \right)$$

Looking within the parenthesis, we form the following inequality, stating that the mean of values is always greater than the sum of absolute deviances from the mean multiplied by the scaling factor:

$$\bar{X}_{G_k,j} \geq \frac{1}{2(|G_k| - 1)} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}|$$

Note that omitting the scaling factor from the RHS yields the equation for Mean Absolute Deviation (MAD):

$$\frac{1}{|G_k|} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}|$$

As shown in (Aghili-Ashtiani, 2021):

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Where  $x_1, x_2, \dots, x_n \in \mathbb{R}$  is a set of real numbers.

We therefore reformulate the problem as follows, noting however that the inequality that the mean of values as always being greater than or equal to the MAD does not hold:

$$\bar{x} \not\geq \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

In fact, in maximally deviating sets, where there is only a single non-zero value, the ratio of the MAD to the mean approaches 2 with increasing set size. Let's assume a vector  $x = [a, 0, 0, \dots, 0]$  of length  $n$ , where only the first value is non-zero. Then:

$$\bar{x} = \frac{a}{n}$$

and (complete derivation available in Appendix Derivation A):

$$\text{MAD} = \frac{2a(n-1)}{n^2}$$

The ratio between the MAD and mean can then be written as:

$$\frac{\text{MAD}}{\bar{x}} = \frac{2a(n-1)/n^2}{a/n} = \frac{2(n-1)}{n}$$

Finally, as set size approaches infinity:

$$\lim_{n \rightarrow \infty} \frac{2(n-1)}{n} = \lim_{n \rightarrow \infty} \left( 2 - \frac{2}{n} \right) = 2$$

Therefore:

$$\bar{x} \not\geq \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

These findings helped influence discovery of the appropriate GeneFunnel scaling factor. With the scaling factor applied, the above evaluates as follows:

$$\text{MAD}_{\text{scaled}} = \frac{1}{2(n-1)} \sum_{i=1}^n |x_i - \bar{x}|$$

$$\text{MAD}_{\text{scaled}} = \frac{1}{2(n-1)} \cdot \frac{2a(n-1)}{n} = \frac{a}{n}$$

$$\frac{\text{MAD}_{\text{scaled}}}{\bar{x}} = \frac{a/n}{a/n} = 1$$

Therefore:

$$\bar{x} \geq \frac{1}{2(n-1)} \sum_{i=1}^n |x_i - \bar{x}|$$

Substituting in the definitions for GeneFunnel:

$$\bar{X}_{G_k,j} \geq \frac{1}{2(|G_k|-1)} \sum_{i \in G_k} |X_{i,j} - \bar{X}_{G_k,j}|$$

We show that the inequality is satisfied for GeneFunnel scores, and conclude that the scaling factor combined with MAD is necessary to ensure that for all  $k$  and  $j$ :

$$\text{score}_{k,j} \geq 0.$$

Thus GeneFunnel always produces non-negative scores.

In addition to proving that GeneFunnel always produces non-negative scores, this proof also shows that GeneFunnel evaluates to zero when a feature set is maximally deviant, that is, the set of features contain only one non-zero value. In the context of gene set enrichment analysis, it is important to ensure that pathway-level activity scores reflect coherent biological signals rather than arbitrary fluctuations in individual gene expression. A key principle of GeneFunnel is that a gene set should only be considered enriched if its member features exhibit a coordinated expression pattern. However, if a gene set is highly inconsistent, where some features are highly expressed while others are completely silenced, then it is biologically uninformative to assign it a high enrichment score. In the case of maximal deviance, a gene set will always receive a score of zero due to the scaling factor. This is a desirable outcome because a gene set is defined by the activity of two or more features and therefore a gene set with expression of a single feature, no matter how high its expression, should not result in that gene set being enriched. As shown in the proof, using MAD alone will dampen such gene sets as well, but at the cost of producing negative values incompatible with many kinds of downstream analyses.

The scaling factor produces another interesting property when considering a gene set where half the features are non-zero, but equal to one another. This gene set will receive a score of half of its sum, and if split into two gene sets, the expressing portion would be equal to its sum while the non-expressing portion would equal zero. Effectively, the activity in this situation is best explained by one of the smaller gene sets, so it should receive the highest score, even if the gene set size is smaller. This is beneficial for narrowing down the specific aspects of pathway activity that are most present, penalising overly general pathways that contain features that are lowly or non-active in the dataset.

## 4.5 Exploration of GeneFunnel Properties

To thoroughly evaluate the behaviour of GeneFunnel and understand its scoring properties, I first conducted an exploration of its theoretical and practical characteristics before benchmarking it against existing methods. To facilitate this process, I developed a Shiny web application (<https://data.duff-lab.org/app/genefunnel-benchmarks-viewer>), which provides an interactive interface for investigating how GeneFunnel responds to different input scenarios. While I discuss the technical aspects of the web app's development in Section 5.3, here I focus on how it was used as an exploratory tool for assessing the properties of the algorithm.

**Figure 99.**

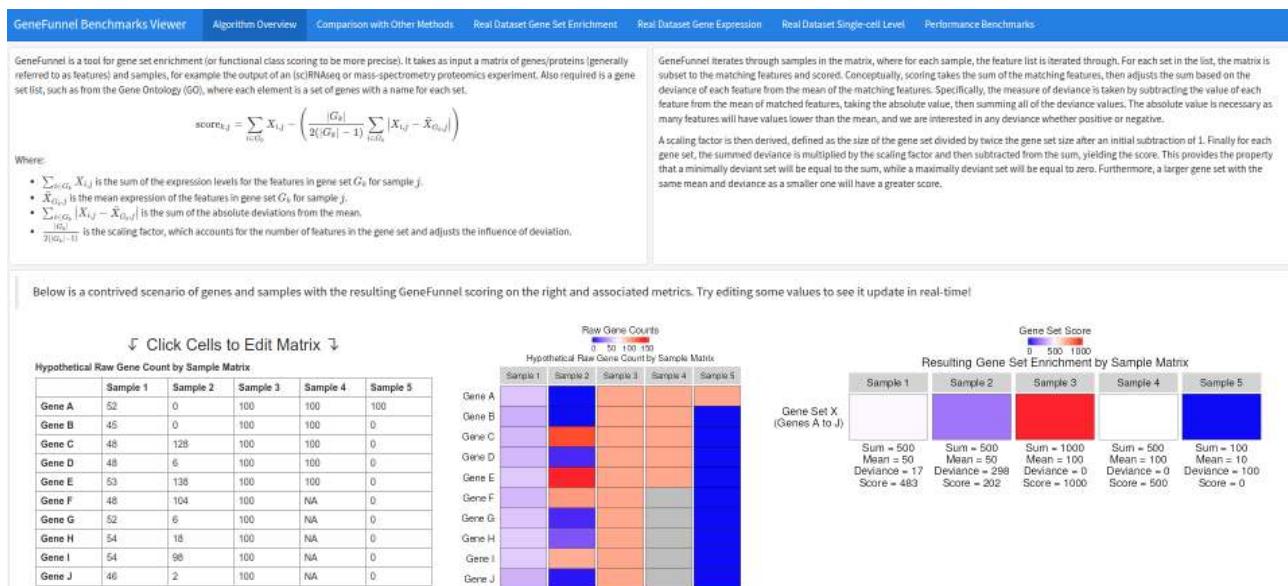


Figure 99: Screenshot of the landing page of the GeneFunnel Benchmarks Viewer Shiny app (<https://data.duff-lab.org/app/genefunnel-benchmarks-viewer>).

A key component of this exploration involved constructing a hypothetical gene by sample matrix to simulate different patterns of gene expression (Figure 100). This synthetic dataset allowed for precise control over the relationships between genes, enabling a systematic examination of how GeneFunnel assigns scores under various conditions. Within the web app, users can interactively modify values within this matrix, effectively simulating different gene expression profiles (Figure 101). Each change is processed in real time, with GeneFunnel recomputing scores for all gene sets dynamically. The results are displayed as a heatmap of the gene set by sample matrix, providing immediate visual feedback on how alterations in individual genes affect pathway-level enrichment scores (Figure 102). This interactive approach not only aids in validating theoretical expectations, such as the behaviour of GeneFunnel under extreme cases, but also helps intuitively illustrate how the method differs from traditional functional class scoring approaches.

**Figure 100.**

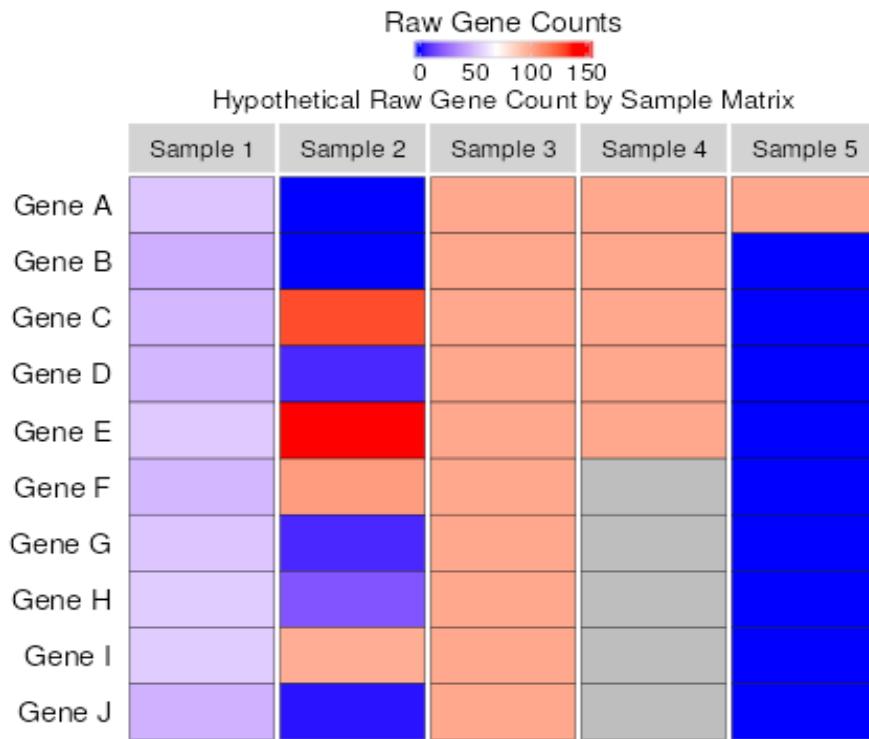


Figure 100: Heatmap of hypothetical gene by sample matrix to simulate different patterns of gene expression. Gray cells indicate NA values. Figure created using ComplexHeatmap R library (Gu et al., 2016).

**Figure 101.**

Hypothetical Raw Gene Count by Sample Matrix

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gene A	52	0	100	100	100
Gene B	45	0	100	100	0
Gene C	48	128	100	100	0
Gene D	48	6	100	100	0
Gene E	53	138	100	100	0
Gene F	48	104	100	NA	0
Gene G	52	6	100	NA	0
Gene H	54	18	100	NA	0
Gene I	54	98	100	NA	0
Gene J	46	2	100	NA	0

Figure 101: The gene count values underlying the heatmap in Figure 100. The table uses the shinyMatrix R library to allow users to edit values within the web app and update GeneFunnel output in real-time.

**Figure 102.**

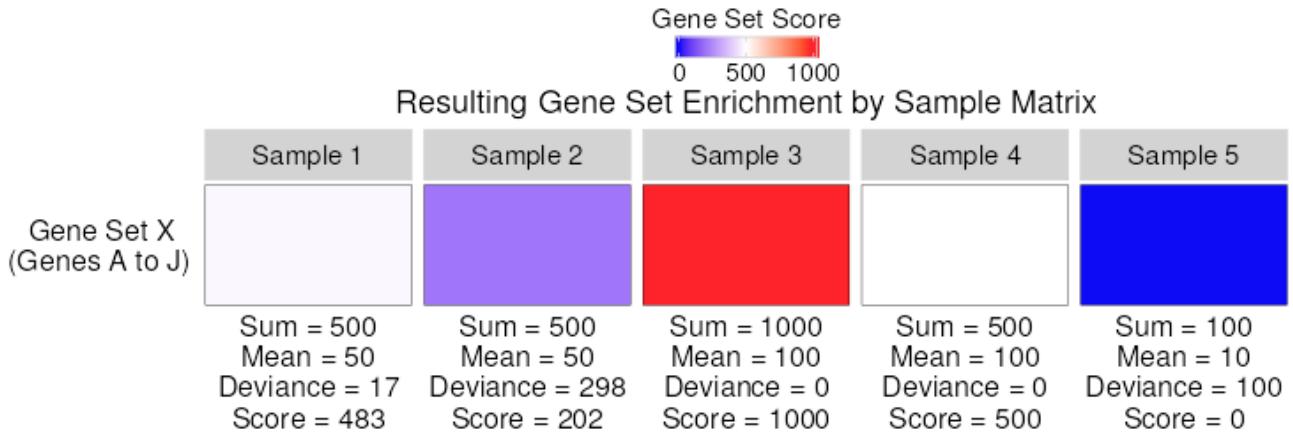


Figure 102: Heatmap coloured by GeneFunnel scores from the hypothetical data in Figures 100 and 101. The entire set of genes (rows) were considered to be a single gene set. This results in a collapse of the original 10 row by 5 column to the 1 row by 5 column matrix seen here. The heatmap contains information below the cells corresponding to the GeneFunnel algorithm: the sum of the gene set, the mean, the deviance (including scaling factor), and the final score.

The selection of values for the originating gene by sample matrix was very deliberate, to try to cover the broad range of situations GeneFunnel was designed to excel in, within a minimal example. Starting with the first column of Figures 100 and 101, with Sample 1, it can be seen that all of the values hover around the arbitrary expression value of 50. These values also include a small degree of noise or jitter, or within the terminology of GeneFunnel, deviance. Upon examination of Figure 102, the metrics below the Sample 1 cell confirm these properties. The mean is precisely 50, and with 10 values, this also results in a sum of 500. The small amount of deviance is also captured, which when subtracted from the mean results in a final value of 483.

The values in Sample 2 were specifically selected to contrast with Sample 1. Examining the original values, it is clear that this column contains much more deviance, with values above 100 and several values recorded as 0. This also begins to highlight another key point of GeneFunnel – zero values in the input data are never omitted. This is in stark contrast to existing methods; as I will review later, all of the benchmarked methods discard zero values from the calculation of enrichment scores. I will argue that this is improper handling of this case, as zero values are valuable information for determining if the totality of gene set is in fact enriched. In any case, even with the inclusion of zero values, Sample 2 was designed to have the same sum and mean as Sample 1, which is confirmed in Figure 102. However, the large deviance of 298 brings a significant penalty to the final score, dropping it from 500 to just 202. This is in contrast to Sample 1, which has a final score of 483. Comparing these two samples reveals why naive implementations of gene set enrichment that are overly reliant on sum or mean alone fail to capture interesting nuances in the data.

Arguably, the gene set contained in Sample 1 would be of greater interest in a real biological setting. All of its genes are expressed at a detectable level and within a close range of one another, raising confidence that all components of the gene set's function are active. Beyond this, focusing on the gene set in this sample increases the probability of successful downstream functional work, such as biochemical or molecular analysis. Nevertheless, one must remember that this is an assumption of the method. The counterargument against the utility of this assumption is the undisputable fact that expression of genes and abundance of proteins vary greatly in their dynamic ranges (Buccitelli & Selbach, 2020). Just because some components of gene set are measured at a low level while others at a higher level does not necessarily mean the gene set is inactivate. At present, GeneFunnel, nor any other reviewed method, has a solution to this reality; they all operate under the assumption that in general, the greater the value of more components of a gene set, the greater the final enrichment score.

Sample 3 showcases a much simpler test case than the first two samples. It simply intends to confirm that when all features are equal in value, the deviance is zero. This is confirmed in Figure 102. As the values in this sample now centre around 100 rather than 50, the mean is now 50 while the sum is 1,000. With a lack of deviance, this results in simply an enrichment score equal to the sum.

Sample 4 differs substantially from the others in that there are NA values in the original matrix. Similar to the argument for the inclusion of zero values, GeneFunnel also maintains a special stance for NA or missing values. Whereas a zero is treated as a measurement at the minimum of the range, an NA is considered to provide no information as to the expression level of the feature at hand. In practice, this represents the only situation where a feature will be excluded in the calculation of the score for a gene set, similarly to how other methods treat zeros. Therefore, GeneFunnel will happily accept NA and missing values, whereas the benchmarked methods fail. Noting this special treatment however, a researcher may still elect to remove these special values, and the general recommendation still is to do so. For the purposes of this hypothetical matrix, NA values effectively reduce the gene set size, which is useful for testing purposes.

As a result, the gene set in Sample 4 is actually treated as a gene set with a size of 5 rather than 10. This changes the scaling factor. Whereas the other samples have a scaling factor of  $\frac{10}{18}$ , the scaling factor for Sample 4 changes to  $\frac{5}{8}$ . This would normally have an effect on the deviance score, though in this example, there is no deviance to begin with, so it remains zero. However, what is noteworthy is that the mean of Sample 4 is equal to Sample 3, but the final enrichment score and sum is 500 rather than 1,000. This indicates GeneFunnel's preference for scoring larger gene sets higher, a deliberate design decision that users should be aware of. The argument for preferring larger gene sets is again driven by pragmatic interest – an enriched larger gene set is more likely to be of biological interest than smaller ones, particularly of the many gene sets in Gene Ontology that are comprised of fewer than 5 genes.

While a naive implementation of this will bias results simply towards larger gene sets, a drawback of many existing methods (Geistlinger et al., 2020; Simillion et al., 2017), it is balanced out by the deviance calculation. As a gene set grows larger, the probability that a dataset exhibits deviance among the features increase. The scaling factor, however, is designed modulate the severity of this correction. For instance, in the case of Sample 4,

$$\frac{5}{8} \quad \frac{10}{18}$$

the scaling factor of  $\frac{5}{8}$  is larger than  $\frac{10}{18}$ . Therefore, as gene set size approaches the

$$\frac{2}{2}$$

minimum of 2, the scaling factor becomes  $\frac{2}{2}$  and the deviance score is maximally applied. On the other hand, as the gene set size approaches infinity, the scaling factor approaches 0.5. The end result is that for a small gene set to have a high score, the few features it contains should have minimal deviance because the scaling factor offers little reduction of the penalty, whereas a larger gene set can have a bit more leeway and this is balanced out by the fact that larger gene sets will have a higher probability of deviance to begin with.

The final column, Sample 5, showcases a situation where values are maximally deviant, producing a score of zero as supported by the proof in preceding sections. Containing a single non-zero value, the sum is fully cancelled out by an equivalent deviance value. This would be the case in all gene set sizes containing a single non-zero value. As the proportion of non-zero values increase, the enrichment score gradually increases until an equilibrium where half of values are non-zero. Assuming no additional deviance, the final enrichment score in such case would be half of the sum.

## 4.6 Exploring GeneFunnel Alongside Other Functional Class Scoring Methods

I next aimed to build off the exploratory approach in the preceding section and apply it to several other functional class scoring methods. I elected to compare GeneFunnel with methods discussed as Prior Art: GSVA (testing both Poisson and Gaussian kernels), ssGSEA, PLAGE, and Z-score. Like the last analysis, the results are wholly contained in the web app in the next tab section. The first series of explorations again focus on a hypothetical gene by sample matrix, constructed similarly as the first with slight modifications (Figure 103). After running each of the models, the results are condensed into enrichment heatmaps (Figure 104) and tables of the raw values (Figure 105).

All methods were run as recommended by their authors for all benchmarking. Importantly, all data input into GSVA Gaussian, ssGSEA, PLAGE, and Z-score underwent a  $\log_2 + 1$  transformation, with GSVA Poisson (and GeneFunnel) being the only methods taking the raw data. The minimum set size was also set to 2 for all methods. Finally, the normalisation step in ssGSEA was turned off, as the method is no longer a single-sample method with it applied (Barbie et al., 2009; Hänelmann et al., 2013). All methods were ran with parallel processing through BiocParallel.

**Figure 103.**

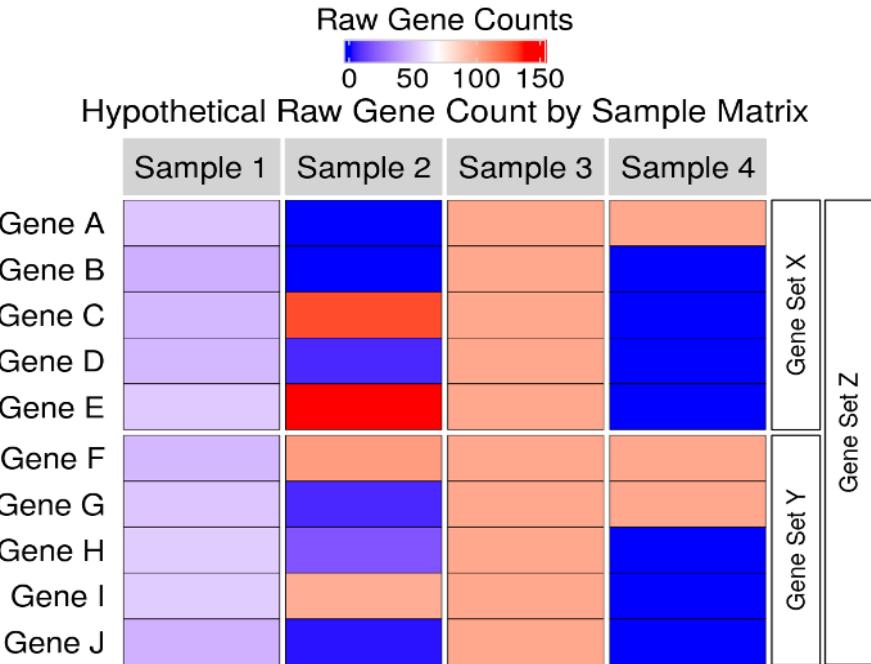


Figure 103: Another hypothetical gene by sample matrix for use with benchmarking various FCS methods against GeneFunnel. It is identical to the one in Figure 100 aside from three key points. 1) The sample containing NA values is removed, as all the tested methods fail to run when the input matrix contains NA or missing values. 2) Any gene sets that would have all zeros are modified to have at least one non-zero value (Sample 4), as the tested methods discard such gene sets. 3) During testing, the first and second half of the genes are evaluated as separate gene sets (designated as Gene Set X and Y in the right-side annotations), as well as a gene set encompassing all genes (Gene Set Z).

**Figure 104.**

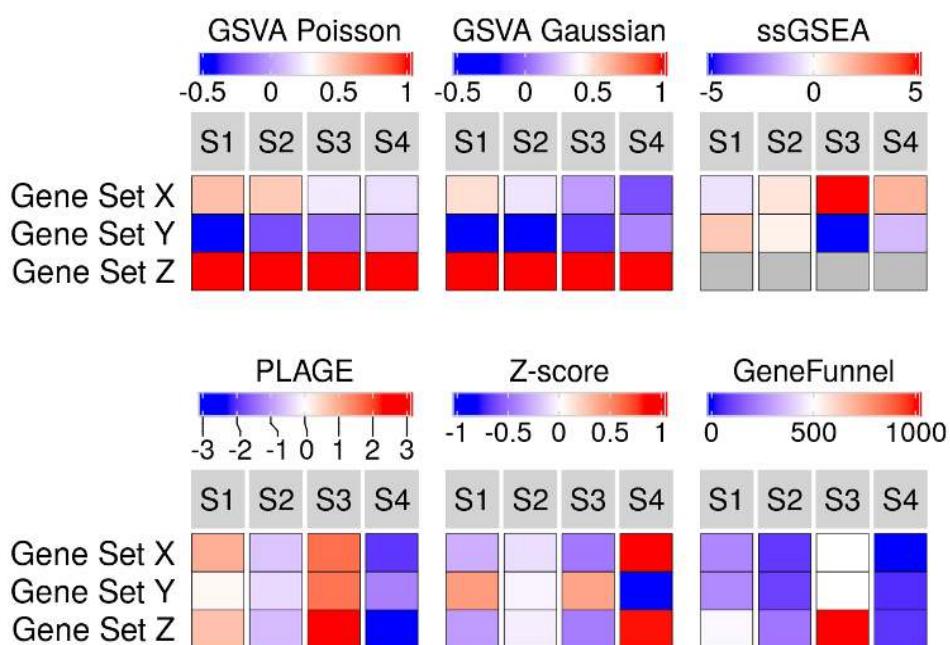


Figure 104: Output of various function class scoring methods, including GeneFunnel, on the hypothetical matrix in Figure 103. Gene sets correspond to the groupings shown in the right-side annotation of Figure 103. Gray cells indicate NA values produced as output.

**Figure 105.**

GSVA Poisson				
	---Sample 1---	---Sample 2---	---Sample 3---	---Sample 4
Gene Set X	0.5272727	0.4857143	0.2307692	0.2000000
Gene Set Y	-0.4142857	-0.2545455	-0.1500000	0.0307692
Gene Set Z	1.0000000	1.0000000	1.0000000	1.0000000
GSVA Gaussian				
	---Sample 1---	---Sample 2---	---Sample 3---	---Sample 4
Gene Set X	0.5	0.3333333	0.1428571	-0.0545455
Gene Set Y	-0.2	-0.2000000	-0.1272727	0.0857143
Gene Set Z	1.0	1.0000000	1.0000000	1.0000000
ssGSEA				
	---Sample 1---	---Sample 2---	---Sample 3---	---Sample 4
Gene Set X	-0.6097222	0.6736724	5	1.954856
Gene Set Y	1.4094127	0.3650516	-5	-1.478317
Gene Set Z	NaN	NaN	NaN	NaN
PLAGE				
	---Sample 1---	---Sample 2---	---Sample 3---	---Sample 4
Gene Set X	0.9495725	-0.5899948	1.643413	-2.002991
Gene Set Y	0.0773811	-0.3881970	1.594036	-1.283220
Gene Set Z	0.7261658	-0.6916860	2.289222	-2.323702
Z-score				
	---Sample 1---	---Sample 2---	---Sample 3---	---Sample 4
Gene Set X	-0.2722034	-0.1010296	-0.4637861	0.8370190
Gene Set Y	0.4322311	-0.0267202	0.4014973	-0.8070082
Gene Set Z	-0.3380642	-0.0431403	-0.4462624	0.8274669
GeneFunnel				
	---Sample 1---	---Sample 2---	---Sample 3---	---Sample 4
Gene Set X	237.7500	75.5000	500	0.00000
Gene Set Y	244.5000	89.5000	500	50.00000
Gene Set Z	483.3333	202.2222	1000	66.66667

Figure 105: Table of values produced from the functional class scoring methods tested on the hypothetical matrix in Figure 103. Values correspond to heatmap colours in Figure 104.

Before diving into details, it is apparent from a cursory glance that each method produces starkly different results, aside from the two variations of the GSVA method. This alone highlights the pervasive problem that plagues gene set enrichment – the massive heterogeneity in results produced between methods. As gene set enrichment is often used as a guiding tool during exploratory analysis that sets the groundwork for extensive and often costly downstream work, it is alarming that methods even within the same category (functional class scoring) show so little concordance. This known issue (Geistlinger et al., 2020; Wijesooriya et al., 2022) was a major motivator for the development of GeneFunnel, not necessarily to add to the pile of methods, but establish a method that is intuitive and can be easily reasoned back to the original data. It is also why initial benchmarking began with small, contrived, hypothetical datasets like the ones discussed here.

Beginning with Sample 1, I claim that GeneFunnel is the only method to sensibly score this sample. For the most basic test, both Gene Set X and Y should be more-or-less similar, as the data contained in each are nearly identical. GeneFunnel produces scores that reflect this, with 237.75 and 244.50 for Gene Set X and Y, respectively. All other methods show noticeable and generally large differences between them, especially with GSVA. GSVA in particular attempts to distribute its output along a range of -1 and 1, similar to a Z-score, making it the most inappropriate for assessing just a few gene sets. While this dataset is indeed a contrived example, it is not inconceivable to be interested in only scoring a few select gene sets in a real-world situation. GSVA was however, the only method other than GeneFunnel to attribute the highest score to Gene Set Z, the largest gene set encompassing all features in the test dataset. This was the second property that I deemed sensible for scoring Sample 1.

In Sample 2, I expected to see generally lower scores than in Sample 1, while following the same pattern of Gene Set X and Y being comparable, and Gene Set Z having the largest scores. GeneFunnel fulfilled this criteria, while all others failed. Most of the methods showed similar patterning as in Sample 1, while Z-score appeared to similarly score Gene Set Y and Gene Set Z (the largest gene set) this time, which could not be explained.

Sample 3 is the most straightforward of the samples. With no deviance at all, Gene Set X and Y should be identical, with Gene Set Z at least being identical or larger. This time at least, there were two methods that could be considered comparable to expected output seen in GeneFunnel. PLAGE showed very sensible results in that all gene sets of Sample 3 were the largest scoring sets in the whole dataset. Furthermore, the scores for Gene Set X and Y are quite similar (1.643413 vs. 1.594036), though not precisely identical like GeneFunnel. While Sample 3 Gene Set Z is indeed the highest score in the entire dataset for GeneFunnel as well, Gene Sets X and Y are more similar to Gene Set Z of Sample 1. It is a matter of debate and subjective viewpoint as to whether GeneFunnel or PLAGE appears more sensible regarding Gene Set Z of Sample 1 in relation to Sample 3. The other method that seemed to perform decently in Sample 3 was GSVA Gaussian, as Gene Set X and Y are more similar to one another compared to those sets in other samples.

Gene Set Z also received the highest score, though the methodology of GSVA makes it so that there is no difference in the score of Gene Set Z between samples; it converges towards 1 in all cases.

Finally, I expected Sample 4 to produce scores that increase in value slightly from Gene Set X, to Gene Set Y, to Gene Set Z, as the proportion of zeros to non-zero values decrease. GeneFunnel reflected this, although the difference between Gene Set Y and Z were small and hard to see on the heatmap (50 vs. 66.66). The only other method that had the correct trend was GSVA Gaussian, however, like other samples, the gap between Gene Set Z compared to the other gene sets is extreme. While PLAGE didn't show the expected pattern per se (Gene Set Z was the lowest scoring), it did correctly show Gene Set X as less enriched than Gene Set Y, which is an undebatable expectation. Furthermore, as a whole, the values in Sample 4 are the lowest in the entire dataset, which should also be expected.

In conclusion, at least in this contrived scenario, aside from GeneFunnel, all of the tested methods performed poorly. While it may be the case that none of these methods were constructed to work with such small test cases, it is still a significant drawback. After all, a very useful approach for exploratory work into understanding how a method works and interacts with changing parameters are through small, controlled experiments like these. Incomparability with such scenarios bring about major limitations to the adoption of these methods. Outside of this, not every real-world experiment is high-throughput especially when working with emerging technology such as spatial omics. It is important for bioinformatic methods to be robust to a range of dataset sizes and I demonstrate here that at least within small datasets, GeneFunnel performs sensibly.

## 4.7 Benchmarking of GeneFunnel Against Other Methods in Synthetic Data

To test whether the small-panel results were an artifact of the setup rather than the methods, I built another synthetic benchmark comparing two groups on a large gene catalog and added formal statistical testing alongside a broader mix of approaches, including approaches that cover the main families of gene-set inference: ORA, camera, fgsea, and GSVA/ssGSEA. ORA (over-representation analysis) takes the final list of differentially expressed genes and asks, via an enrichment test against the background gene catalogue, whether each set contains more hits than expected by chance; this is the generalised approach taken by the popular g:Profiler (Raudvere et al., 2019), but this implementation permits an arbitrary set catalogue and background, which is necessary for synthetic benchmarks. Camera, from limma (Phipson et al., 2016; Ritchie et al., 2015), is a competitive test that fits a linear model per gene and then evaluates whether genes in a set show stronger differential expression than genes outside the set. fgsea, is an R implementation of GSEA (Subramanian et al., 2005), which operates on a ranked list of genes and computes an enrichment score that reflects whether set members concentrate near the top or bottom of the ranking (Korotkevich et al., 2016). GSVA and ssGSEA are

among the functional class scoring methods used in the prior benchmark, computing a per-sample score for each set without using group labels, similar to GeneFunnel. For these, as well as GeneFunnel, I apply a stock limma-trend pipeline, similar to that of the main thesis analysis on tangle-bearing neurons, to test for differential enrichment between groups. This mixture of methods allow for a comparison of hit-list enrichment (ORA), model-based competitive testing (camera), rank-based enrichment (fgsea), functional class scoring (GSVA and ssGSEA), and the proposed functional class scoring method (GeneFunnel) under one evaluation protocol.

The simulation uses a 20,000 gene matrix partitioned into 1,000 non-overlapping gene sets (20 genes per set) and 10 samples (5 in group A, 5 in group B). In each experiment, 50 sets are designated signal and the remaining 950 null. Counts are drawn from a negative-binomial model with realistic library-size variation, then normalised with edgeR TMM before set-level scoring and testing across the gene matrix. Signals were injected under three patterns that isolate different behaviours of gene set enrichment methods and expose different dynamic ranges of gene set activity. In “spike”, only half of genes in a signal set is perturbed, but strongly, while the remainder is left untouched, which probes a method’s tolerance to partial activation and within-set heterogeneity (Figure 106). In “variance”, the set mean of the signal set is preserved while the within-set dispersion is deliberately reduced in one group, testing the ability of the methods in assessing within-set consistency (Figure 107). In “coordinated”, a small but consistent log fold change is applied to all genes in a signal set in one group, producing the most classic example of gene set enrichment but within a small dynamic range so as to stress the sensitivity of each method (Figure 108). I evaluate each setting at FDR 0.05, using statistical testing intrinsic to each method or limma-trend otherwise, recording sensitivity, specificity, precision and other common benchmarks. In contrast with the last simulation, which functions as an exploration of functional class scoring properties under precisely defined but ultimately unrealistic scenarios, this simulation study intends to more comprehensively cover the various approaches to gene set enrichment in a dataset with realistic properties and signal structures. It furthermore provides a clearer picture of where GeneFunnel’s design, which rewards both signal magnitude and within-set consistency, confers advantages or exposes limitations in practical use.

**Figure 106.**

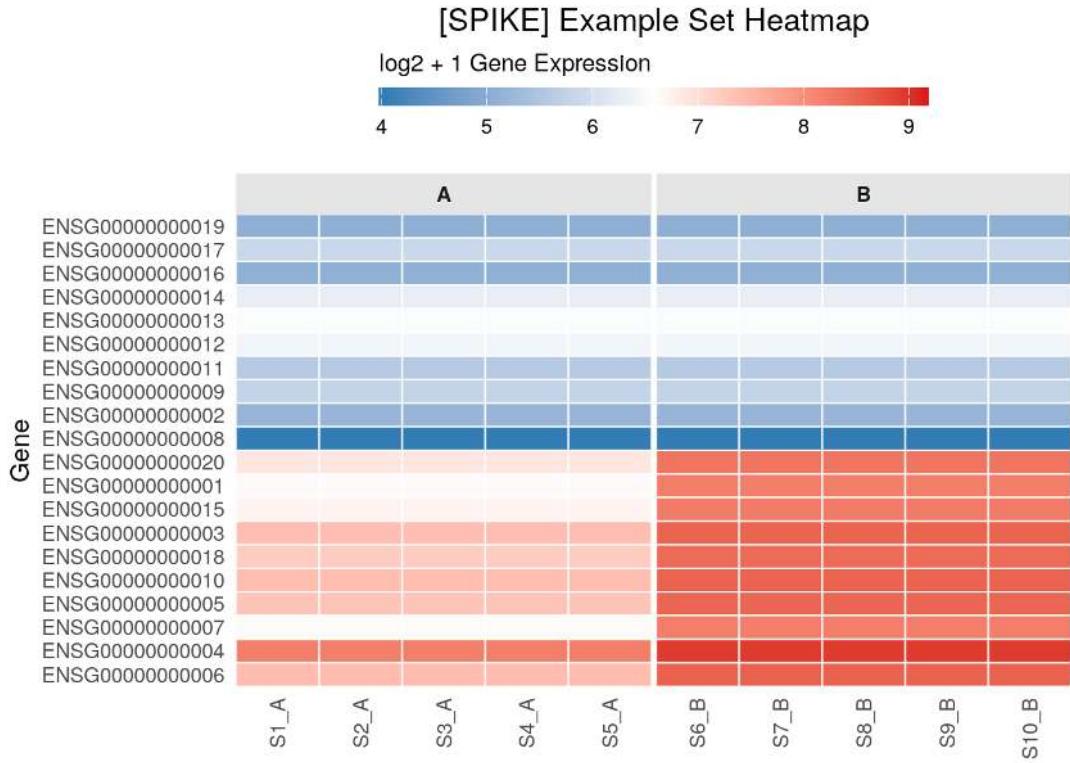


Figure 106: An example of the perturbations for one randomly selected signal set. In the “spike” paradigm, half of genes of the signal set in one group have 200 counts added to their signal while the rest remain unchanged. Columns are split by group, rows are clustered within the set, and the colour bar shows centred log2+1 expression.

**Figure 107.**

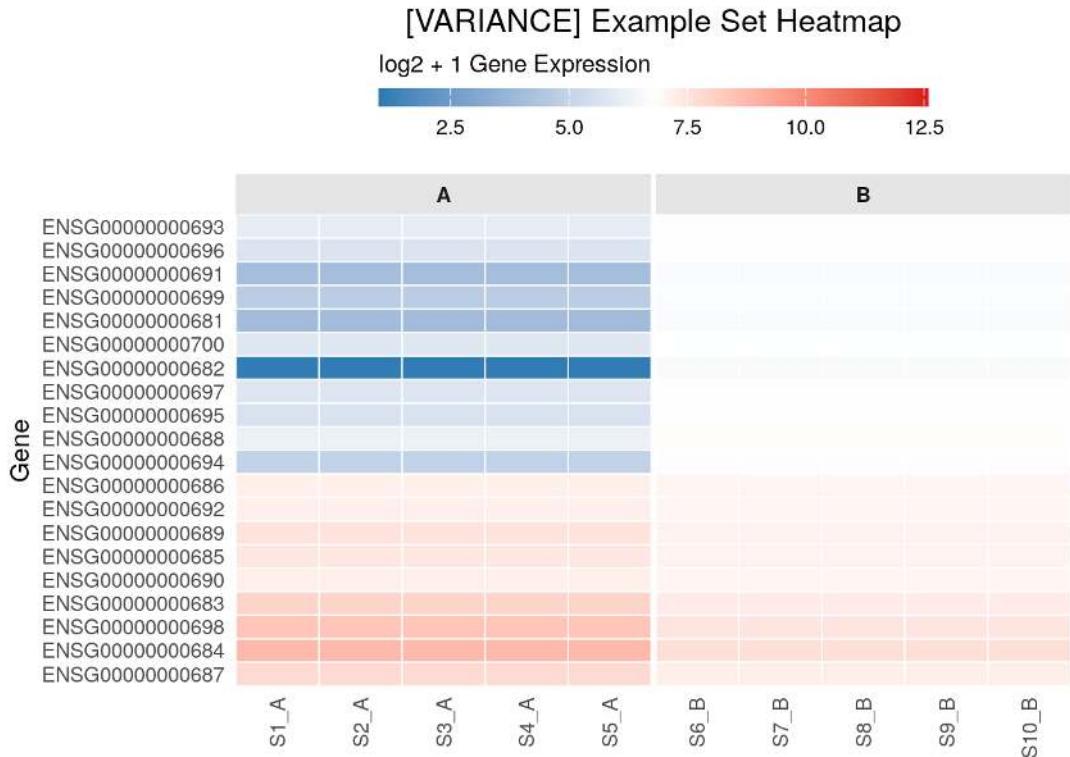


Figure 107: An example of the perturbations for one randomly selected signal set. In the “variance” paradigm, set means are preserved but intergene variance of the genes in the signal sets of one group is reduced to 25% of its original value. Columns are split by group, rows are clustered within the set, and the colour bar shows centred log2+1 expression.

**Figure 108.**

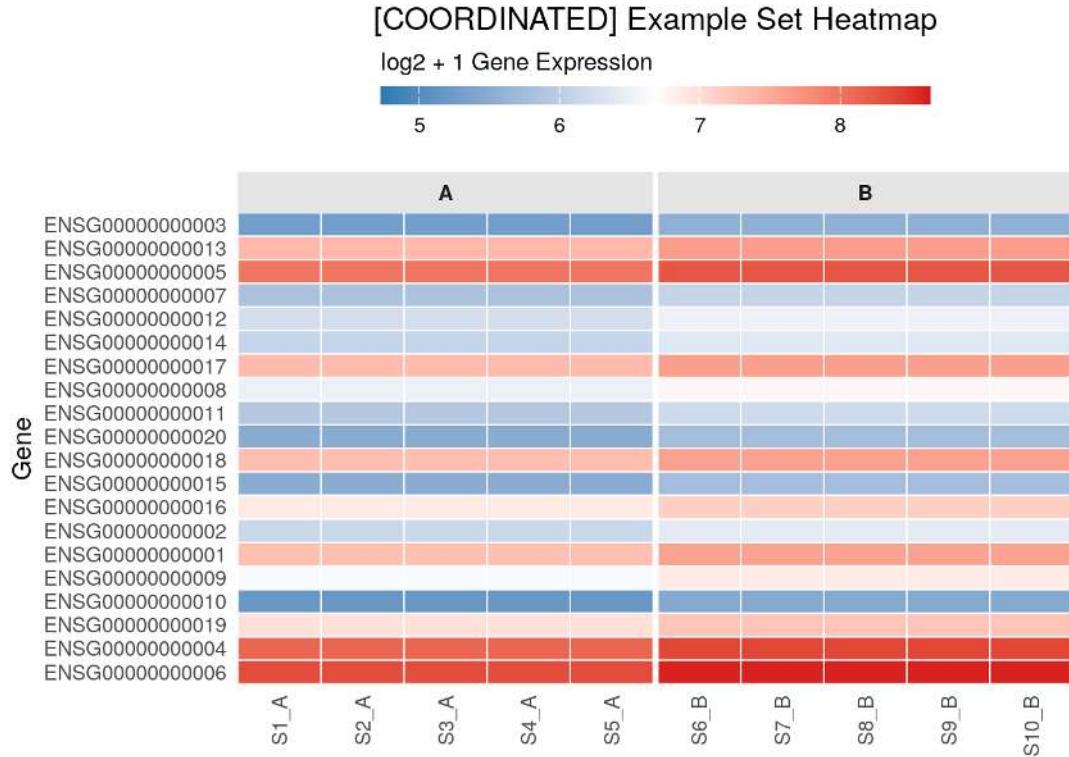


Figure 108: An example of the perturbations for one randomly selected signal set. In the “coordinated” paradigm, all genes in the signal set shift by a small (0.25 logFC with 0.1 standard deviation) same-direction amount in one group. Columns are split by group, rows are clustered within the set, and the colour bar shows centred log2+1 expression.

The tables below summarise method performance at the threshold of FDR (BH adjusted p-value) 0.05 for each perturbation paradigm. With 50 signal sets and 950 null sets per experiment, TP (true positive) counts signal sets correctly detected, FN (false negative) the missed signal sets, TN (true negative) the correctly rejected null sets, and FP (false positive) the null sets falsely called. From these I report sensitivity (TP/P), specificity (TN/N), precision (TP/(TP+FP)), accuracy ((TP+TN)/(P+N)), and F1 (the harmonic mean of precision and sensitivity). I also report the average FDR for the signal sets, defined as the mean BH adjusted p-value across all 50 signal sets for each paradigm. Higher is better for all rates except average FDR, where lower is better.

**Figure 109.**

**Detection metrics for SPIKE ( $\alpha = 0.05$ )**

20k genes; 1000 sets  $\times$  20 genes; 50 signal sets; 5 vs 5 samples

Method	Counts				Rates					
	TP	FN	TN	FP	sensitivity	specificity	precision	F1	accuracy	avg FDR
GeneFunnel	50	0	950	0	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b><math>1.14 \times 10^{-6}</math></b>
camera	50	0	950	0	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	$8.17 \times 10^{-4}$
fgsea	50	0	948	2	<b>1.000</b>	0.998	0.962	0.980	0.998	$2.00 \times 10^{-2}$
GSVA	45	5	949	1	0.900	0.999	0.978	0.938	0.994	$1.83 \times 10^{-2}$
ssGSEA	50	0	947	3	<b>1.000</b>	0.997	0.943	0.971	0.997	$3.00 \times 10^{-4}$
ORA	15	35	950	0	0.300	<b>1.000</b>	<b>1.000</b>	0.462	0.965	$5.19 \times 10^{-1}$

Figure 109: Per-method performance for the “spike” paradigm at the FDR threshold of 0.05. Rows list methods, columns report detection counts (TP, FN, TN, FP) and the derived rates (sensitivity, specificity, precision, F1, accuracy, average FDR). Colours correspond to magnitude while bold font marks the highest values within a column.

**Figure 110.**

**Detection metrics for VARIANCE ( $\alpha = 0.05$ )**

20k genes; 1000 sets  $\times$  20 genes; 50 signal sets; 5 vs 5 samples

Method	Counts				Rates					
	TP	FN	TN	FP	sensitivity	specificity	precision	F1	accuracy	avg FDR
GeneFunnel	11	39	950	0	<b>0.220</b>	<b>1.000</b>	<b>1.000</b>	<b>0.361</b>	<b>0.961</b>	<b><math>2.87 \times 10^{-1}</math></b>
camera	7	43	950	0	0.140	<b>1.000</b>	<b>1.000</b>	0.246	0.957	$6.14 \times 10^{-1}$
fgsea	0	50	950	0	0.000	<b>1.000</b>	—	0.000	0.950	$5.96 \times 10^{-1}$
GSVA	0	50	950	0	0.000	<b>1.000</b>	—	0.000	0.950	$8.51 \times 10^{-1}$
ssGSEA	0	50	950	0	0.000	<b>1.000</b>	—	0.000	0.950	$8.37 \times 10^{-1}$
ORA	8	42	950	0	0.160	<b>1.000</b>	<b>1.000</b>	0.276	0.958	$4.89 \times 10^{-1}$

Figure 110: Per-method performance for the “variance” paradigm at the FDR threshold of 0.05. Rows list methods, columns report detection counts (TP, FN, TN, FP) and the derived rates (sensitivity, specificity, precision, F1, accuracy, average FDR). Colours correspond to magnitude while bold font marks the highest values within a column.

**Figure 111.**

**Detection metrics for COORDINATED ( $\alpha = 0.05$ )**

20k genes; 1000 sets  $\times$  20 genes; 50 signal sets; 5 vs 5 samples

Method	Counts				Rates					
	TP	FN	TN	FP	sensitivity	specificity	precision	F1	accuracy	avg FDR
GeneFunnel	9	41	950	0	<b>0.180</b>	<b>1.000</b>	<b>1.000</b>	<b>0.305</b>	<b>0.959</b>	<b><math>4.34 \times 10^{-1}</math></b>
camera	5	45	950	0	0.100	<b>1.000</b>	<b>1.000</b>	0.182	0.955	$5.94 \times 10^{-1}$
fgsea	0	50	950	0	0.000	<b>1.000</b>	—	0.000	0.950	$5.58 \times 10^{-1}$
GSVA	8	42	950	0	0.160	<b>1.000</b>	<b>1.000</b>	0.276	0.958	$4.78 \times 10^{-1}$
ssGSEA	2	48	950	0	0.040	<b>1.000</b>	<b>1.000</b>	0.077	0.952	$5.30 \times 10^{-1}$
ORA	0	50	950	0	0.000	<b>1.000</b>	—	0.000	0.950	1.00

Figure 111: Per-method performance for the “coordinated” paradigm at the FDR threshold of 0.05. Rows list methods, columns report detection counts (TP, FN, TN, FP) and the derived rates (sensitivity, specificity, precision, F1, accuracy, average FDR). Colours correspond to magnitude while bold font marks the highest values within a column.

In the “spike” paradigm, where only half of the genes in a signal set are perturbed, though robustly, all methods perform well except ORA. GeneFunnel and camera achieve perfect detection (50/50 true positives with no false positives). The rank-based and unsupervised scoring approaches are close to this ceiling; fgsea detects all 50 signal sets with two false positives, GSVA recovers 45 of 50 with one false positive, and ssGSEA detects all 50 with three false positives. ORA remains highly specific but has low sensitivity (15/50), likely because partial activation leaves too few genes surpassing the differential expression threshold to trigger over-representation at the set level. Although the 50 signal sets do not overlap any other sets, several methods are sensitive to the overall distribution of gene-level statistics or ranks. The robust perturbation of the signal sets may have shifted this background slightly, resulting in false positives for those methods. GeneFunnel is only susceptible to this issue at the statistical testing stage, i.e. limma, as the scoring mechanism itself operates on each gene set and sample in isolation.

In the “variance” paradigm, where the mean is preserved and only within-set dispersion is altered, procedures that target location differences lose power. GeneFunnel retains the highest sensitivity because its scoring emphasises within-set consistency as well as magnitude, allowing reduced variability to register as a stronger, more coherent pattern despite the lack of mean change. Camera and ORA identify a smaller fraction of signal sets, and the rank-based and unsupervised scoring methods detect none at the chosen threshold. Though no other method claims to measure inter-gene variance, individual changes to gene counts to reduce inter-gene variance pushes some genes past the significance threshold for regular differential expression testing. It is likely that when several such genes occur in the same set, methods that aggregate gene-level evidence, such as camera or ORA, can incidentally report enrichment despite not explicitly including criteria for within-set consistency. Across methods, average FDRs are higher than in “spike”, reflecting weaker evidence when the signal resides in dispersion rather than in the mean.

In the “coordinated” paradigm, where a very small (0.25 logFC with 0.1 standard deviation), but consistent log-fold change is applied to all members of each signal set, GeneFunnel again achieves the best combination of sensitivity and F1. GSVA is second, in line with its design to capture coordinated per-sample shifts, and camera detects fewer sets at this subtle effect size. fgsea and ORA do not register signal sets at all here, indicating that the per-gene effects are too small to accumulate sufficient ranked-list or hit-list evidence at an FDR of 0.05. This paradigm is the most standard formulation of gene set enrichment and serves as a direct test of method sensitivity to small but coherent shifts. With the current effect size and 5 vs 5 samples the signal is intentionally challenging, so power concentrates in methods that aggregate weak, consistent changes across all genes in a set.

Across these simulations GeneFunnel shows the most consistent power across the three alternatives. It reaches the ceiling in the spike setting, retains the highest sensitivity when the signal is variance only, and remains competitive for small coordinated shifts. This

matches the design goal of the method, which produces per-sample set scores that reward both effect magnitude and within-set consistency, so partial activation, tighter dispersion and subtle coordinated changes can each yield a detectable set-level signal. GeneFunnel works within a standard limma-trend workflow, gives interpretable profiles at the sample level, and maintains low false-positive rates at the stated FDR.

A few caveats remain. The current simulation uses 5 vs. 5 samples, non-overlapping sets of size 20 and a single catalogue of genes. Performance could change with larger or smaller cohorts, different set sizes, heavy set overlap or highly redundant catalogues, and stronger gene-gene correlation. In this experiment, GeneFunnel and other functional class scoring methods, relied on limma-trend for statistical testing, and ensuring its proper calibration is non-trivial. Furthermore, there are other downstream testing frameworks that can significantly affect the performance of these methods. Finally, the evidence here remains fully synthetic, and while the proceeding section covers usage in real-world data, testing in biological “ground truth” data, such as those utilising RNA spike-ins may be of value, though such datasets still contain non-trivialities in generation and interpretation.

## 4.8 Benchmarking of GeneFunnel Against Other Methods in Real Data

Having run two synthetic experiments, one exploratory within the FCS family and one spanning method families with formal testing, I now move to real data. Because ground truth is unknown in this setting, I restrict the comparison to FCS methods to create a more like-for-like testing framework, which makes qualitative comparisons more interpretable. The dataset of choice was the transcriptomics dataset that encompasses the main results of this thesis work: the FACS-sorted ssRNAseq dataset (Otero-Garcia et al., 2022). I started with the annotated Seurat object obtained from the completion of the pipeline described in Methods. This object is a single cell by gene matrix which I then pseudobulked to resemble a bulk RNAseq matrix using the `aggregateAcrossCells` function from (McCarthy et al., 2017). The parameters for pseudobulking were set to produce a simple two column output, aggregating cells into either a tangle-bearing or non-tangle-bearing group while ignoring donor label information. In the following section I describe a number of controlled transformations of these objects and test the ability of each functional class scoring methods to capture these transformations.

The first transformation was to arbitrarily select a single column, in this case the tangle-bearing neurons, and alter the gene expression of genes corresponding to specific gene sets. To begin with, I choose two particular gene sets: NELF Complex and Trace-amine Receptor Activity. These gene sets were chosen because they have no gene overlap with other gene sets in the testing set, therefore, any changes detected should only be in these two sets (Figure 112). NELF Complex was modified to reduce variability, that is, all genes of the set were transformed into the sum of the gene counts divided by the total number of genes in the set. Trace-amine Receptor Activity was simply modified to have increased counts; all genes in the set had 100 counts added to them. A column containing these modifications was added to the original object, while leaving the original column

unmodified. I then ran each FCS method on the modified and unmodified columns. The result of each FCS output is shown in Figure 113. I subset to the top two gene sets sorted by greatest absolute difference between the modified and unmodified columns. If a method successfully captured the induced modifications, then the altered gene sets should be the ones present in the sorted data.

**Figure 112.**

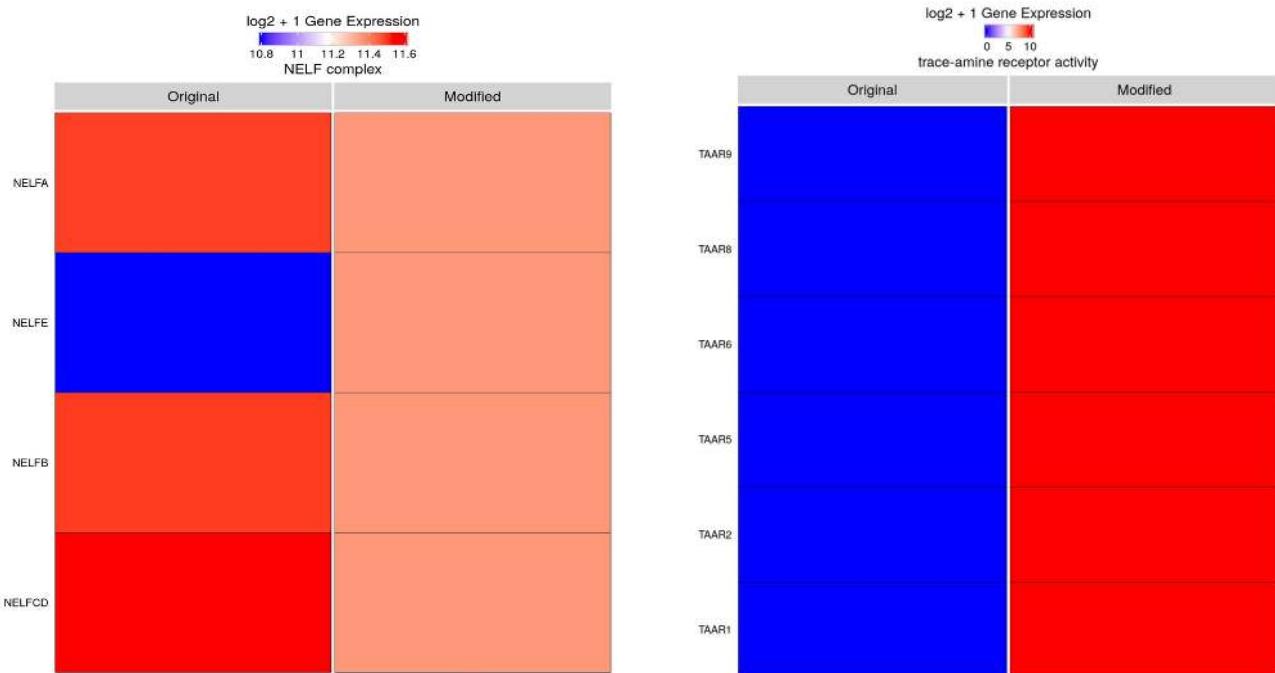


Figure 112: Heatmaps showing controlled modifications of specific gene sets in pseudobulked data from (Otero-Garcia et al., 2022), the FACS ssRNAseq dataset. In each, I introduce a Modified column where the counts for genes in NELF Complex were altered to reduce variability, and the counts for genes in Trace-amine Receptor Activity were increased, as described in the text. The data is  $\log_2 + 1$  transformed before plotting.

**Figure 113:**

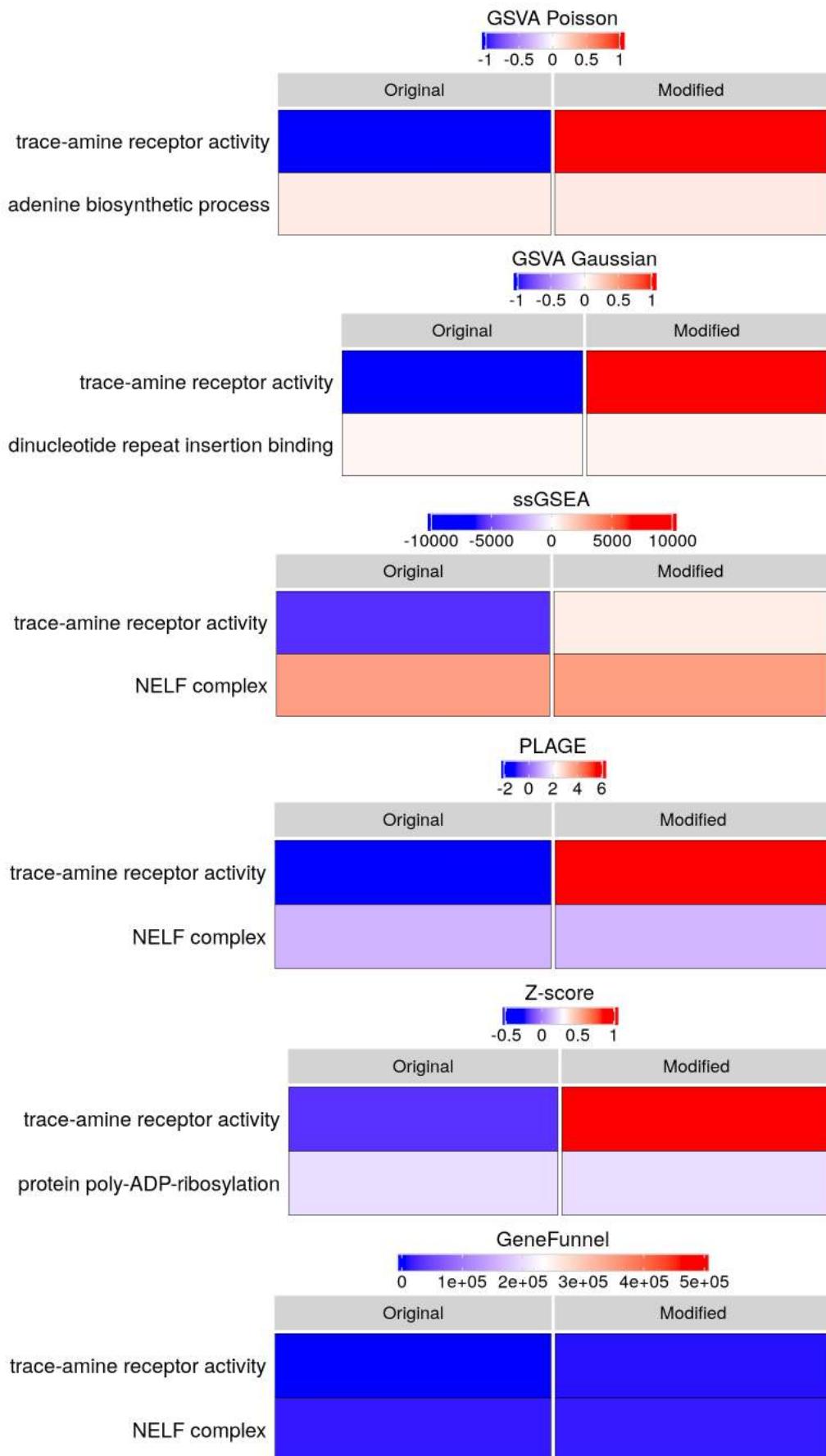


Figure 113: Comparison of the six FCS methods in the controlled modifications of the pseudobulked FACS ssRNASeq dataset. The “Original” column contains unmodified data, while the “Modified” had counts of NELF Complex modified to reduce variability while the counts of Trace-amine Receptor Activity were increased. Shown are the top two gene sets for each method after sorting by greatest absolute difference between the “Modified” and “Original” columns.

As can be seen in Figure 113, all methods successfully captured the change in Trace-amine Receptor Activity, which simply had counts of associated genes increased by 100 counts. This demonstrates that all methods have the capacity to capture simple linear changes in expression level. However, only three methods also showed NELF as being among the top two hits: ssGSEA, PLAGUE, and GeneFunnel. This shows that these methods are sensitive, at least to some extent and whether incidental or not, to changes in the variability (or deviance in GeneFunnel methodology), even without changes in overall expression levels.

This example is also noteworthy when considering the magnitude of changes expressed in the heatmaps. No scaling was applied to any heatmap and the colour bar is set to encompass the entire range of values produced by each method. For all methods except for GeneFunnel, the changes are visually apparent, while in GeneFunnel, the changes are so subtle and the colour bar range so wide that it is essentially invisible. This is a feature of GeneFunnel, not a bug. Recall that the changes were induced subtly, for instance, an increase of only 100 counts in each gene of NELF Complex. GeneFunnel reflects that this change is small, but still detectable. This is advantageous because it preserves magnitude of relative expression between the genes, offering great dynamic range. One may argue that this reduces sensitivity for statistical testing, however, as shown when exploring synthetic data in Figures 103 through 105, GeneFunnel does not introduce artificial variance between what should be similarly scored gene sets or samples, unlike the tested methods. While other methods may increase power by artificially inflating effect sizes, even if unintentionally, increased sensitivity when using GeneFunnel is derived from robust, stable scoring that closely reflects the source data.

Using the dynamic range of GeneFunnel, one can infer in Figure 113 that both gene sets contain genes that are either lowly expressed in the context of the dataset, or a set of genes that are highly variable. Effectively, whether or not the gene sets are differentially expressed, in both samples these gene sets are relatively lowly enriched. One can make no such inferences using the other methods. Nevertheless, if one wanted to strictly highlight changes between groups, it is up to the discretion of the researcher to limit the colour bar to the range of values shown in the heatmap, or apply scaling, which will produce a GeneFunnel heatmap more similar to the other methods in Figure 113. An example of this is shown in Figure 114. In general however, I advocate against such approaches as they may be misleading and ultimately results in a loss of information.

**Figure 114.**

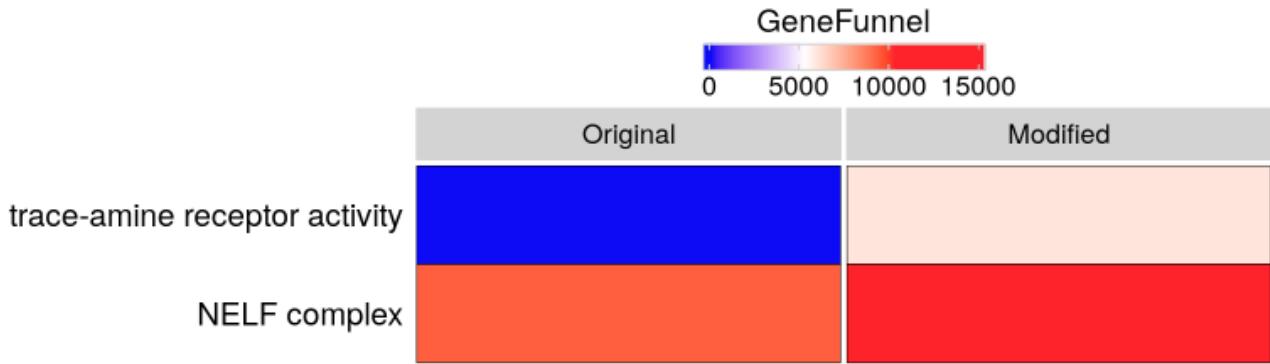


Figure 114: The same GeneFunnel heatmap shown in Figure 113, however with the colour set to the range of values encompassed by the two-row heatmap, rather than the range of values across the entire profile of GeneFunnel scores.

I continued with this line of testing with further variations variations on the theme. In a similar fashion, I next modified the gene set Tau Protein Binding, which was modified to reduce variability, that is, all genes are set to the sum of the gene counts divided by the total number of genes. The key difference in this experiment is that this gene set overlaps with other gene sets, so it may not necessarily be the highest hit, but should still be ranked near the top. Figure 115 shows what this modification looks like, and Figure 116 shows the output of testing on the six FCS methods.

**Figure 115.**

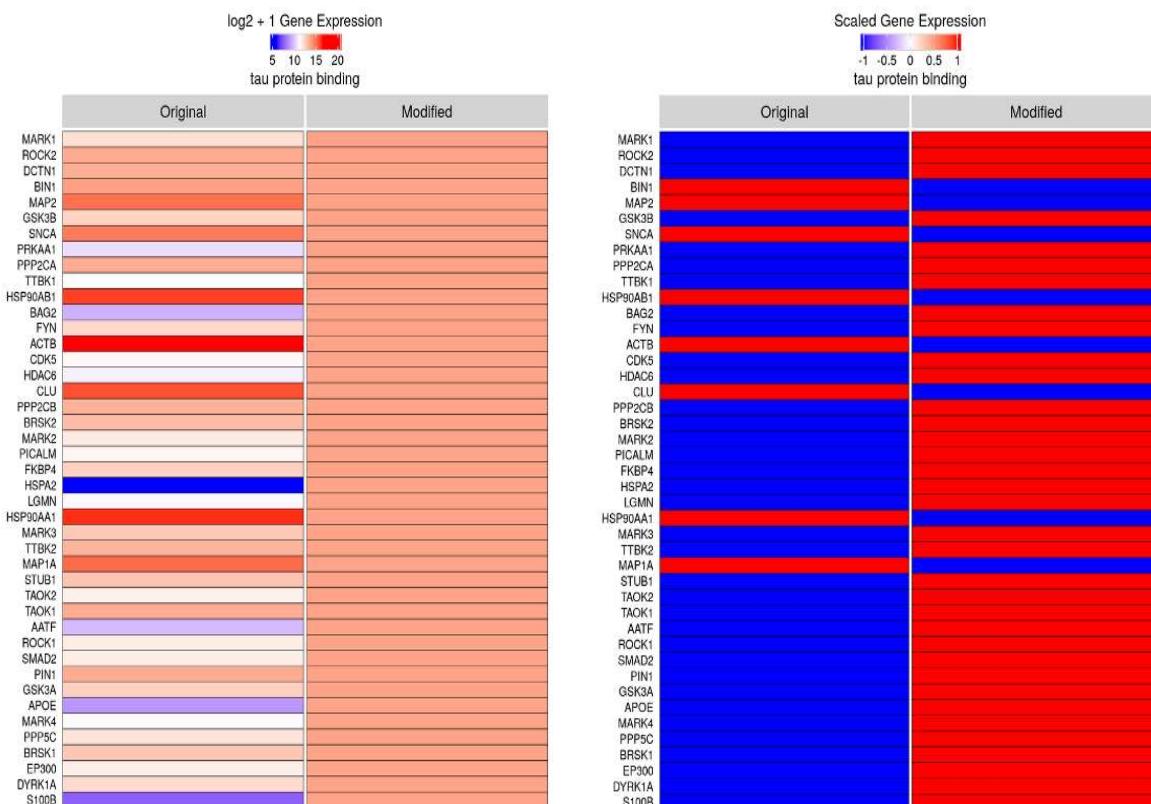


Figure 115: Modifications performed on the genes in the gene set Tau Protein Binding in the FACS ssRNAseq dataset. In this instance, only variability was reduced, as described. Shown left is the data after log2 + 1 transformation, and right is the data after per-row-scaling for added visualisation purposes.

**Figure 116:**

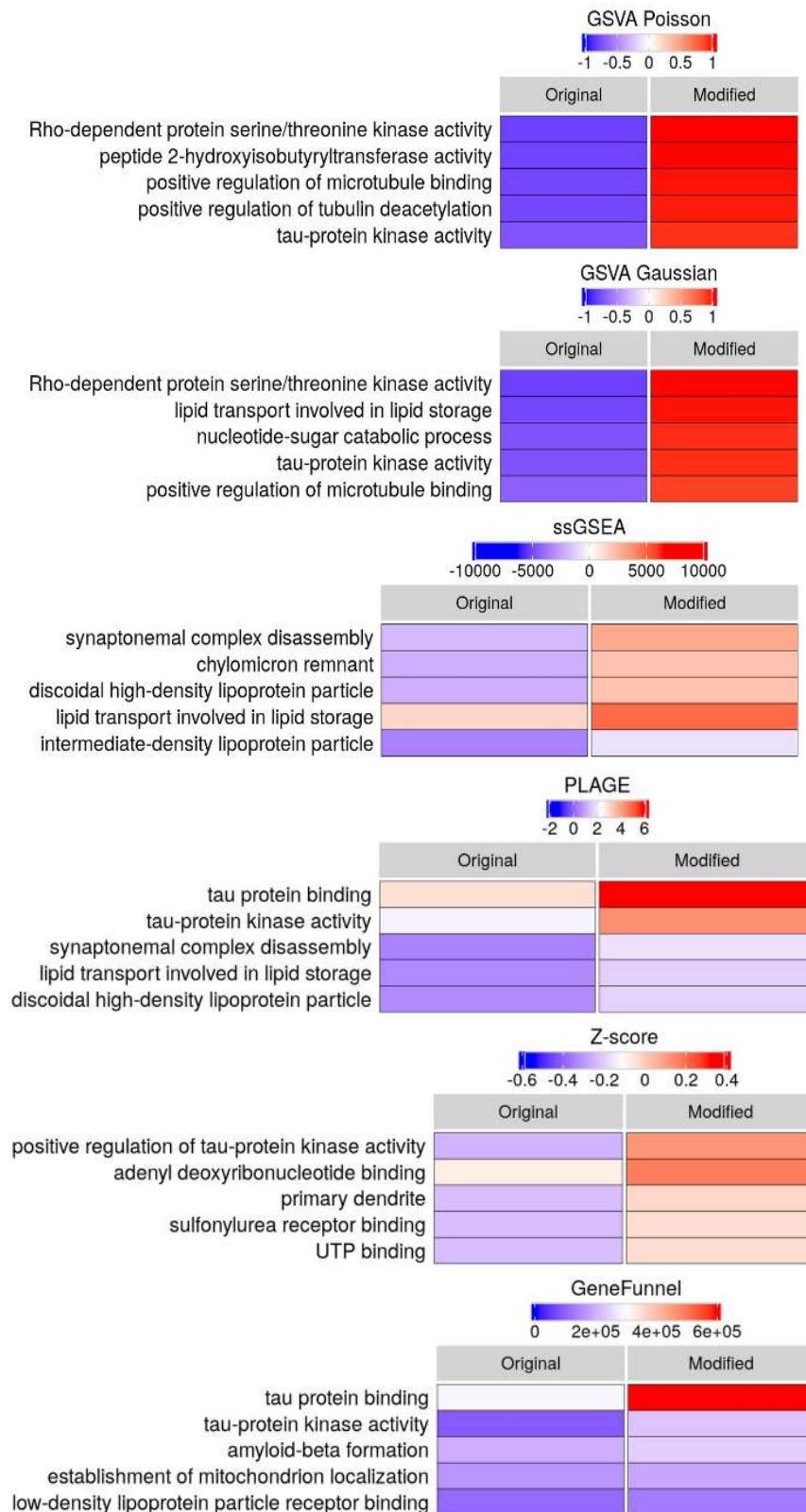


Figure 116: Result of FCS testing when modifying the gene set Tau Protein Binding to reduce variability. Because Tau Protein Binding overlaps with other gene sets, the top five gene sets, sorted by absolute difference between Modified and Original columns are shown, as Tau Protein Binding may not necessarily be the top hit.

The results of this experiment shows that two methods successfully capture the Tau Protein Binding term, with both selecting it as the top hit: PLAGE and GeneFunnel. These methods also rank terms with many overlapping genes highly, such as Tau-protein Kinase Activity and Amyloid-beta Formation. Noteworthy is that in other methods that did not capture Tau Protein Binding, they often showed Tau-protein Kinase Activity (or some variation of it) in the top five hits; these methods being the two GSVA methods and Z-score. ssGSEA failed to capture a gene set that appears to have immediate relevance to Tau Protein Binding. Unlike the last experiment (Figures 112 through 114), this one also produced a larger magnitude of change in GeneFunnel, requiring no scaling approaches to visualise the differences. This is likely the result of the larger gene set size of Tau Protein Binding; when modifying more genes, a greater difference between the two groups is produced. GeneFunnel reflects this, while the other methods appear to adversely produce effect sizes similar to those in Figure 113.

Like the previous experiments, I performed the benchmarking procedure again, this time modifying the counts of the Neurofibrillary Tangle gene set to have increase counts; all genes had 100 counts added to them. Like Tau Protein Binding, this gene set overlaps with other gene sets. Figure 117 shows what the modification looks like, while Figure 118 shows the results of FCS testing.

**Figure 117.**

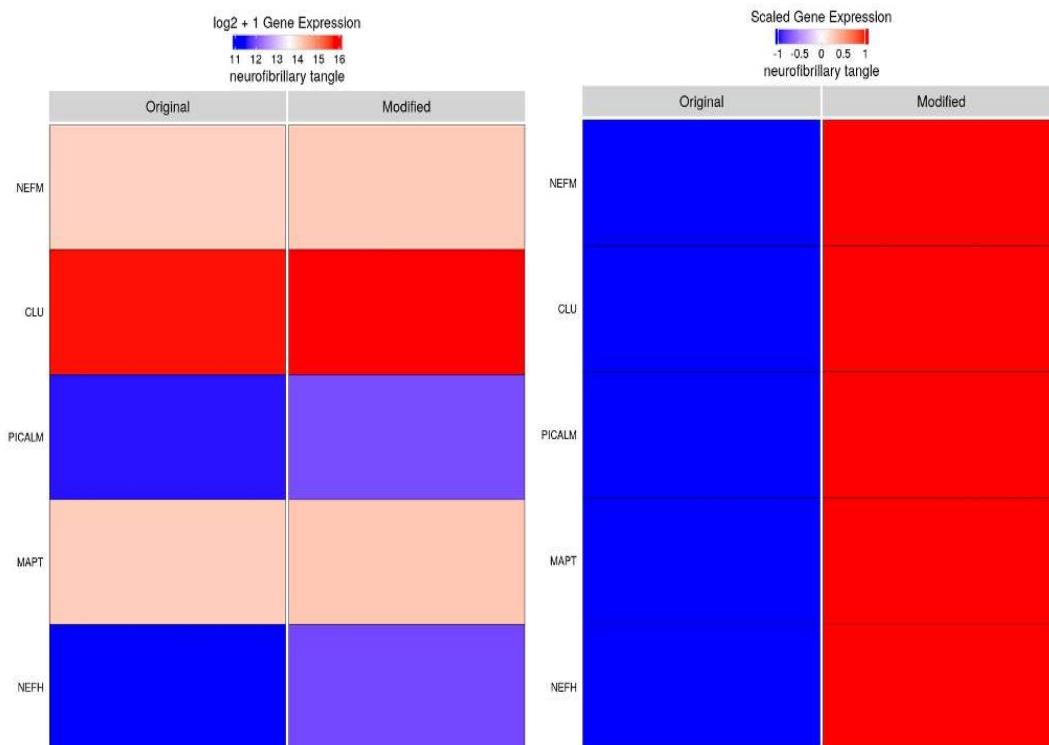


Figure 117: Modifications performed on the genes in the gene set Neurofibrillary Tangle in the FACS ssRNAseq dataset. In this instance, counts were increased, as described. Shown left is the data after  $\log_2 + 1$  transformation, and right is the data after per-row-scaling for added visualisation purposes.

**Figure 118.**

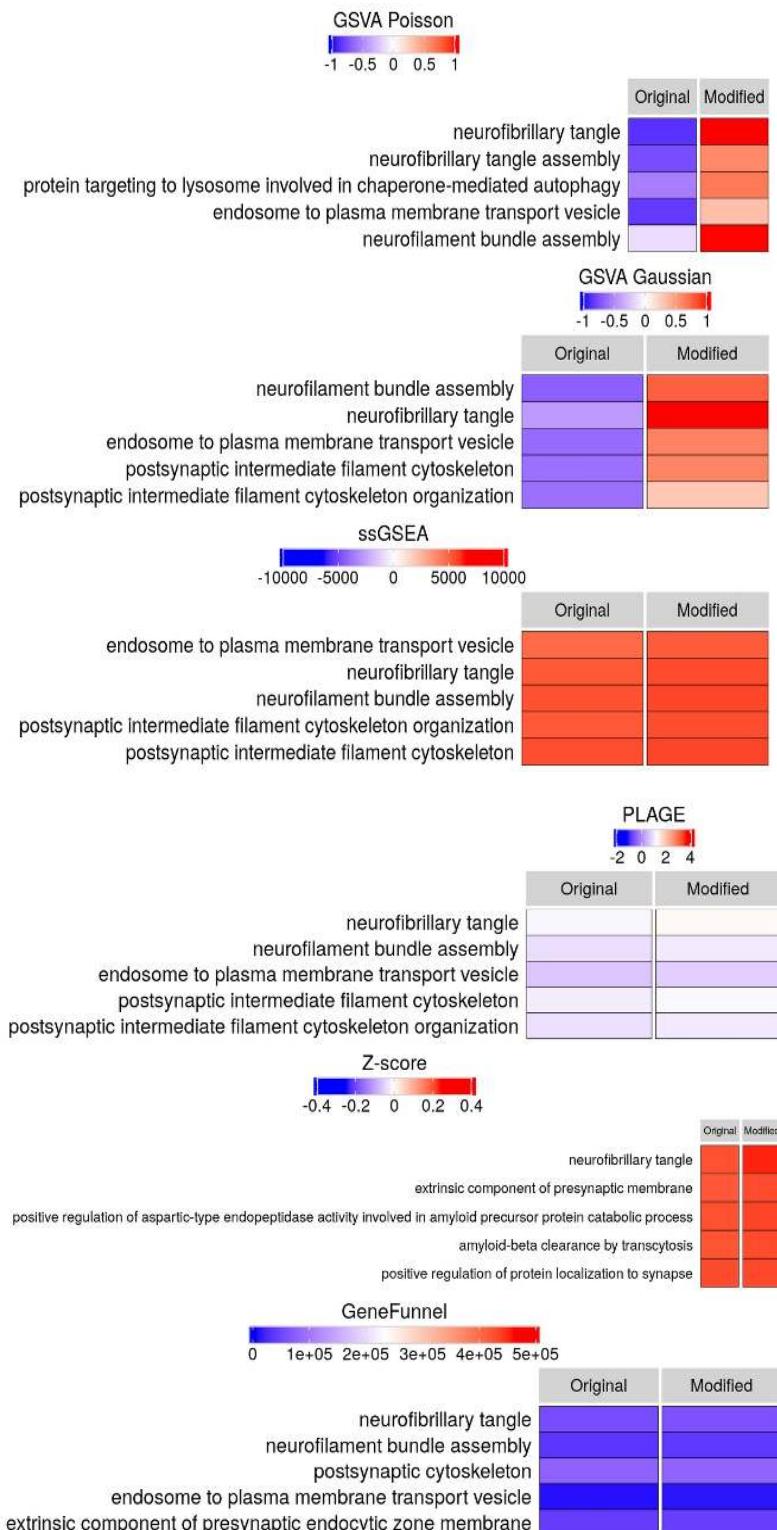


Figure 118: Result of FCS testing when modifying the gene set Neurofibrillary Tangle to have increased counts. Because Neurofibrillary Tangle overlaps with other gene sets, the top five gene sets, sorted by absolute difference between Modified and Original columns are shown, as Neurofibrillary Tangle may not necessarily be the top hit.

The results of this experiment are much more promising across methods than the last; all of the tested FCS methods show Neurofibrillary Tangle as among the top five hits, although GSVA Gaussian and ssGSEA only list it as the second hit. Nonetheless, GSVA Gaussian at least ranks Neurofilament Bundle Assembly as the top hit, which is a highly overlapping gene set. These results show that in general, all methods can capture simple changes in gene set expression, but only a few have the capacity to detect the changes in variability introduced in the prior experiment (Figures 115 and 116).

As a final test, I examined the ability of each method to detect changes in the condition-level pseudobulked dataset without modifications, in other words, a comparison of the gene set composition of tangle-bearing vs. non-tangle-bearing neurons. In order to make this comparison as straightforward as possible, all the donors are pooled together and no statistical testing is performed. The hypothesis is that when sorting gene sets by the absolute difference between the two conditions, as done in the prior tests, gene sets relevant to Alzheimer's Disease should rise to the top. If not, then manual inspection of the top gene sets should at least reveal that they are reasonable and reflect likely real changes. The results of gene set enrichment for this experiment is shown in Figure 119, along with inspection of the genes within some of the gene sets in Figure 120.

Out of the benchmarks performed thus far, this last benchmark appears to produce the largest divergence between GeneFunnel and the other methods. Comparisons with other methods aside, GeneFunnel does appear to highlight gene sets of immediate relevance to AD: containing terms such as Tau Protein Binding (gene expression shown in detail in Figure 120) and Positive Regulation of Tau-protein Kinase Activity. Neither of these terms are shown among the top five for the other methods. In regard to term overlap between GeneFunnel and other methods, there is a term related to dendrites in both GeneFunnel and GSVA Poisson and a term related to synapses in both GeneFunnel and the Z-score method, with neither overlaps being exact matches.

**Figure 119.**

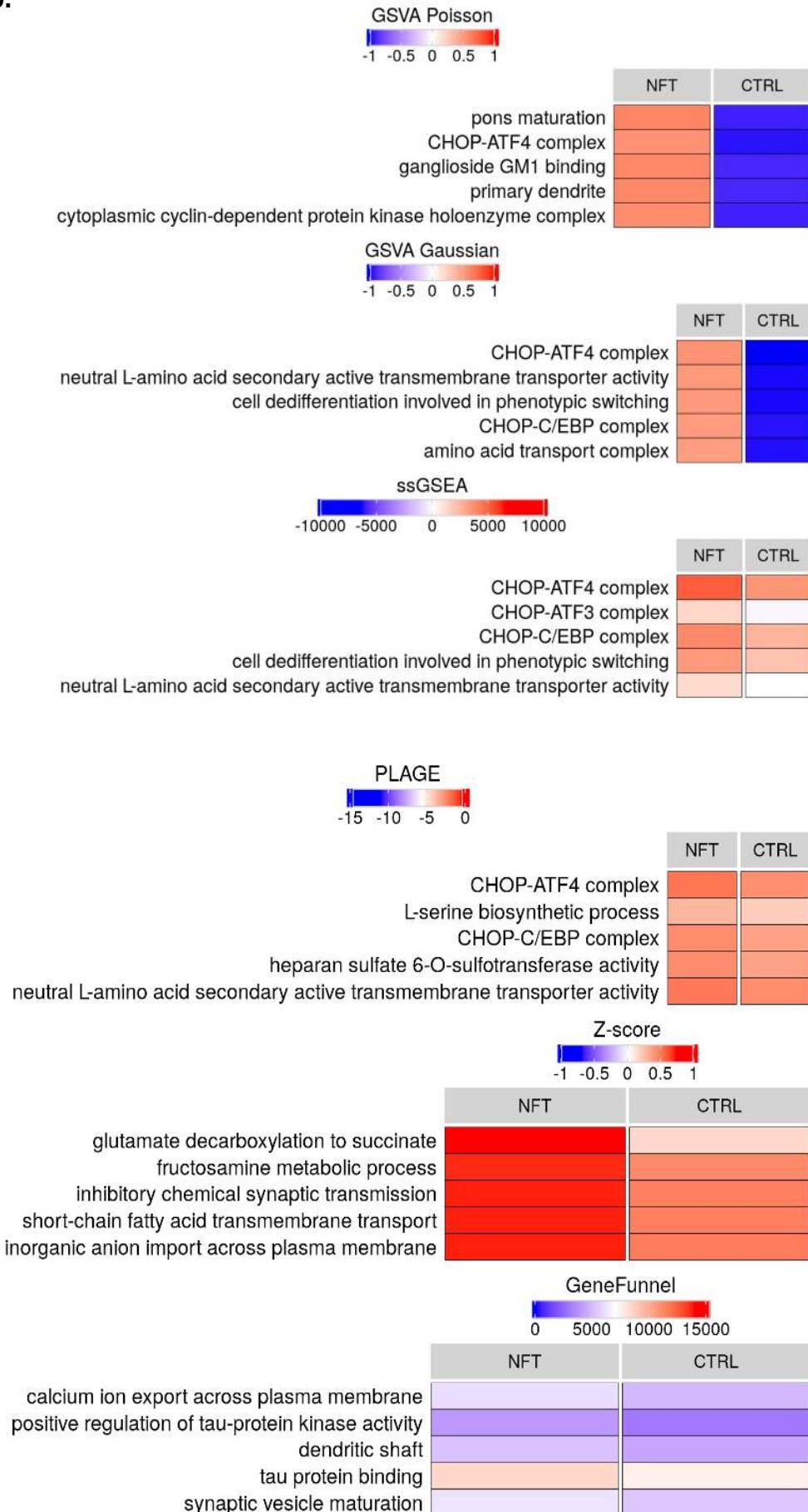


Figure 119: Gene set enrichment results across the methods when comparing tangle-bearing vs. non-tangle-bearing neurons in the FACS ssRNAseq dataset without modifications. The data was pseudobulked by donor to ensure ease of comparison. The top five gene sets sorted by absolute difference between NFT and CTRL columns is shown for each method.

**Figure 120.**

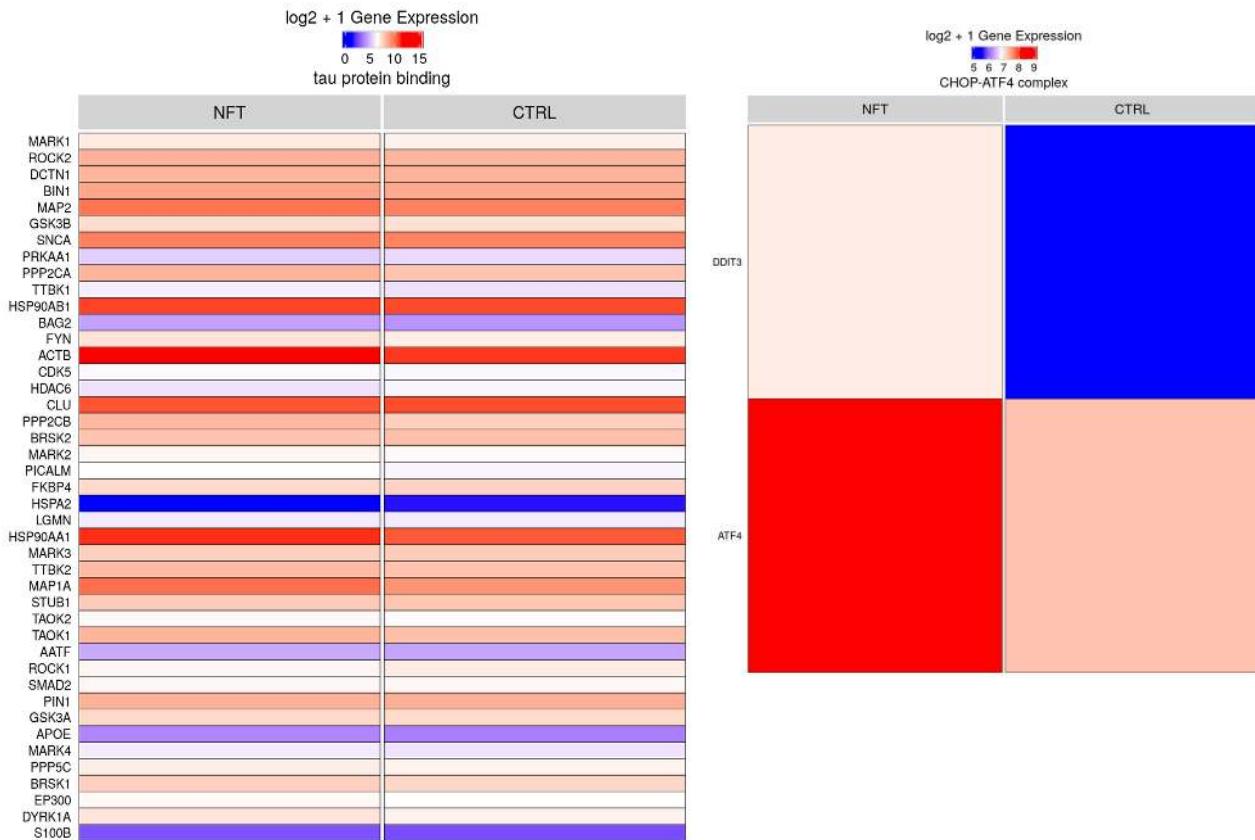


Figure 120: The expression of genes in selected gene sets highlighted in the experiment depicted in Figure 119, where tangle-bearing and non-tangle-bearing neurons were contrasting using the FCS methods in the unmodified FACS ssRNAseq data.

Aside from the Z-score method and GeneFunnel, the other methods do appear to have a noticeable level of alignment with one another. In particular, they all seem to focus highly on processes related to CHOP, a factor known to interact with the C/EBP family of transcription factors (Sok et al., 1999). In AD, CHOP is implicated to protect neurons from ER stress (Aceves et al., 2024), so this finding may indeed warrant further inspection. However, examination of the actual gene set raises suspicion for the reasons behind its prioritisation by various methods. As can be seen in Figure 120, this is a very small gene set, composed of just two genes. One gene, *ATF4*, is highly expressed, and is likely solely dependent for driving the large difference in enrichment between the NFT and CTRL conditions. As described in prior sections, GeneFunnel is designed to balance gene set size, increasing the weight of deviance penalty for small gene sets, with gene sets

specifically of the size 2 carrying the most weight. Indeed, there is a large difference between *ATF4* and the only other gene *DDIT3* and GeneFunnel uses this difference to penalise the gene set highly. This allows for the higher prioritisation of gene sets like Tau Protein Binding, where the magnitude of no singular gene change is comparable to *ATF4*, but across the 25 genes comprising the gene set, many are increased by some degree in NFT vs. CTRL. I argue that such gene sets are of greater research interest and thus GeneFunnel provides greater utility over other methods in this instance. They are also easier to interpret, as tau related processes are the direct mechanism involved in this comparison, so GeneFunnel provides important positive confirmation as to the validity of the experimental data.

## 4.9 Benchmarking of Computational Efficiency of GeneFunnel Against Other Methods

Even if a method has high analytical performance compared to others, that method may not be feasible to use if runtime or memory usage is excessive. This area received significant attention in GeneFunnel, prompting an implementation in Rcpp (C++ interface to R) (Eddelbuettel & Balamuta, 2018; Eddelbuettel & François, 2011) with the highly optimised RcppArmadillo linear algebra libraries (Eddelbuettel & Sanderson, 2014). In order to compare computational efficiency across methods, I took the original FACS ssRNAseq data and replicated samples or cells to different sizes and then passed each method through the function mark from the R package `bench`. All tests were performed with 5 iterations to ensure robustness. Furthermore, when comparing serial and parallel processing, the same framework was used in all methods: `BiocParallel`.

Three variations of this approach were recorded. For the first, I used a pseudobulked version of the FACS ssRNAseq dataset with 6 total samples and ran all methods using serial processing. The output is summarised in Figure 121. Next, I took this same data and reran the methods with 60 samples using parallel processing. This output is summarised in Figure 122. Finally, for the last test, I went back to the original single-cell data without pseudobulking. Using parallel processing, I tested each method on a maximum of 600 cells alongside various subsets of the data. Using six subsets, at each subset the number of cells was halved. For example, whereas the sixth subset contained 600 cells, the fifth subset contained 300, and so-on. The results of this experiment is captured in Figure 123.

**Figure 121.**

expression	min	median	'itr/sec'	mem_alloc	'gc/sec'
<chr>	<bch:tm>	<bch:tm>	<dbl>	<bch:byt>	<dbl>
1 GSVA Poisson	2.91s	2.94s	0.275	8.57GB	0.659
2 GSVA Gaussian	2.8s	2.85s	0.347	8.57GB	0.833
3 ssGSEA	39.06s	39.64s	0.0251	10.02GB	0.126
4 PLAGUE	4.67s	4.72s	0.212	23.67MB	0.763
5 Z-score	1.7s	1.72s	0.578	76.97MB	1.16
6 GeneFunnel	375.91ms	379.54ms	2.57	2.24MB	0.514

Figure 121: Runtime and memory usage across various functional class scoring methods when using serial processing on 6 pseudobulked samples from the FACS ssRNAseq dataset. Benchmarking performed using the `mark` function from the R package `bench`.

**Figure 122.**

expression	min	median	`itr/sec`	mem_alloc	`gc/sec`
<chr>	<bch:tm>	<bch:tm>	<dbl>	<bch:byt>	<dbl>
1 GSVA Poisson	5.35m	5.68m	0.00297	NA	0.00297
2 GSVA Gaussian	2.02m	2.13m	0.00789	NA	0.00789
3 ssGSEA	17.22m	17.26m	0.000964	NA	0
4 PLAGUE	1.36m	1.37m	0.0122	NA	0.0244
5 Z-score	2.12m	2.13m	0.00782	NA	0
6 GeneFunnel	12.45s	12.58s	0.0794	NA	0

Figure 122: Runtime when performing the same experiment as Figure 121 but with parallel processing and 60 rather than 6 samples. Note that when using parallel processing, memory usage cannot be captured using the framework provided by the R package `bench`.

**Figure 123.**

Benchmarking Tools on Different Data Subsets

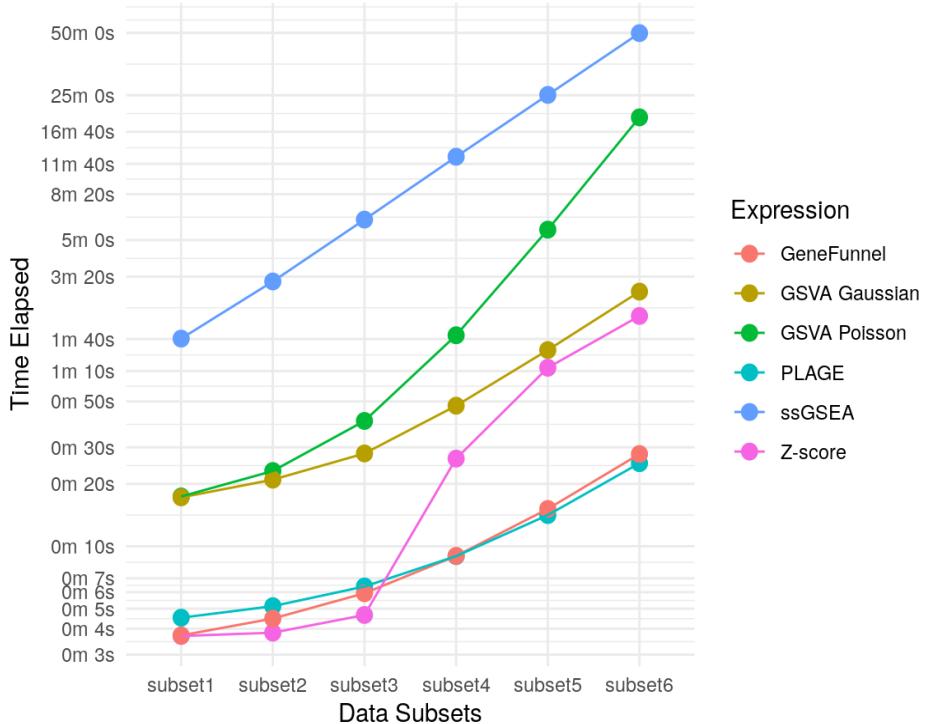


Figure 123: Runtime benchmarking of the six methods on various subsets of samples from the unmodified non-pseudobulk version of the FACS ssRNAseq dataset. At the largest subset, subset 6, 600 cells are used. At each prior subset, the number of cells is halved; 300 cells at subset 5, 150 cells at subset 4, etc.

Inspection of the figures shows that GeneFunnel is the leader in computational efficiency in both runtime and memory usage in all scenarios, although PLAGUE is comparable when

comparing runtime in single-cell data (Figure 123). When using serial processing, the median runtime and memory usage of GeneFunnel is 379.54ms and 2.24MB, respectively. The next most performant methods, PLAGUE and Z-score have runtimes measured in the seconds and several tens of megabytes of memory usage. ssGSEA was notably unoptimised, taking almost 40 seconds and consuming 10GB of memory. The GSVA methods, while reasonably quick (~2 seconds), also consumed about 8GB memory. When using parallel processing and increasing the number of samples by a factor of 10 (6 to 60 samples), the same efficiency rankings hold true (Figure 122). GeneFunnel is the quickest by far, taking a median of 12.58s, with PLAGUE being the next quickest at 1.37m and Z-score and GSVA Gaussian tied at 2.13m. Similarly to the first experiment, ssGSEA took an excessively long time, at a median of 17.26m to the time of completion.

In Figure 123, I reverted to analysis on the original single-cell version of the FACS ssRNAseq dataset, to measure how feasible the various methods are for single-cell analysis. Throughout the different subset of data, both GeneFunnel and PLAGUE rank at the top, with both taking under 30s when evaluating 600 single cells. The two methods track closely in runtime in this instance, though were more divergent in the pseudobulk data. This can be explained through the handling of dropouts among the methods. GeneFunnel is the only method that fully retains zero values in all calculations and never discards data. In contrast, PLAGUE, and others, drop much of this information, lessening the gap with GeneFunnel particularly in the sparse single-cell data. Although not within the scope of the GeneFunnel algorithm, if GeneFunnel were to drop values in a similar value, the gap would likely widen again to a margin similar to the pseudobulk tests.

Regarding the other methods in the final test, towards the 600 cell mark they all begin approach runtimes into the minutes, with ssGSEA taking almost 1hr per run at this point. Interestingly, the Z-score method starts out comparable to PLAGUE and GeneFunnel initially, but diverges significantly from the fourth subset (150 cells). Considering that typical single-cell datasets these days contain tens to hundreds of thousands of cells, it is clear that many of these methods are likely completely infeasible to run on a single-cell level. Overall, these efficiency benchmarks support the optimisations garnered by GeneFunnel's C++ implementation, simple algorithm, and parallel processing capability.

#### 4.10 Advantages of GeneFunnel

GeneFunnel introduces a novel approach to functional class scoring that directly addresses limitations in existing methods by incorporating deviation-aware scoring while maintaining sample independence. One of its most significant advantages is that it ensures pathway-level enrichment scores reflect coordinated gene expression rather than being driven by a few highly expressed genes. Many existing methods, such as GSVA and ssGSEA, operate on the assumption that total expression within a gene set is a sufficient proxy for pathway activity. However, this can lead to inflated scores for gene sets where only a subset of genes are highly expressed while others are inactive, producing misleading conclusions about pathway activation. GeneFunnel overcomes this by

introducing an internal deviation penalty, which ensures that gene sets exhibiting extreme dispersion do not receive high scores. This property makes it particularly well-suited for cases where internal consistency within a pathway is biologically relevant, such as distinguishing truly co-regulated gene sets from those that are only partially activated.

**Figure 124.**

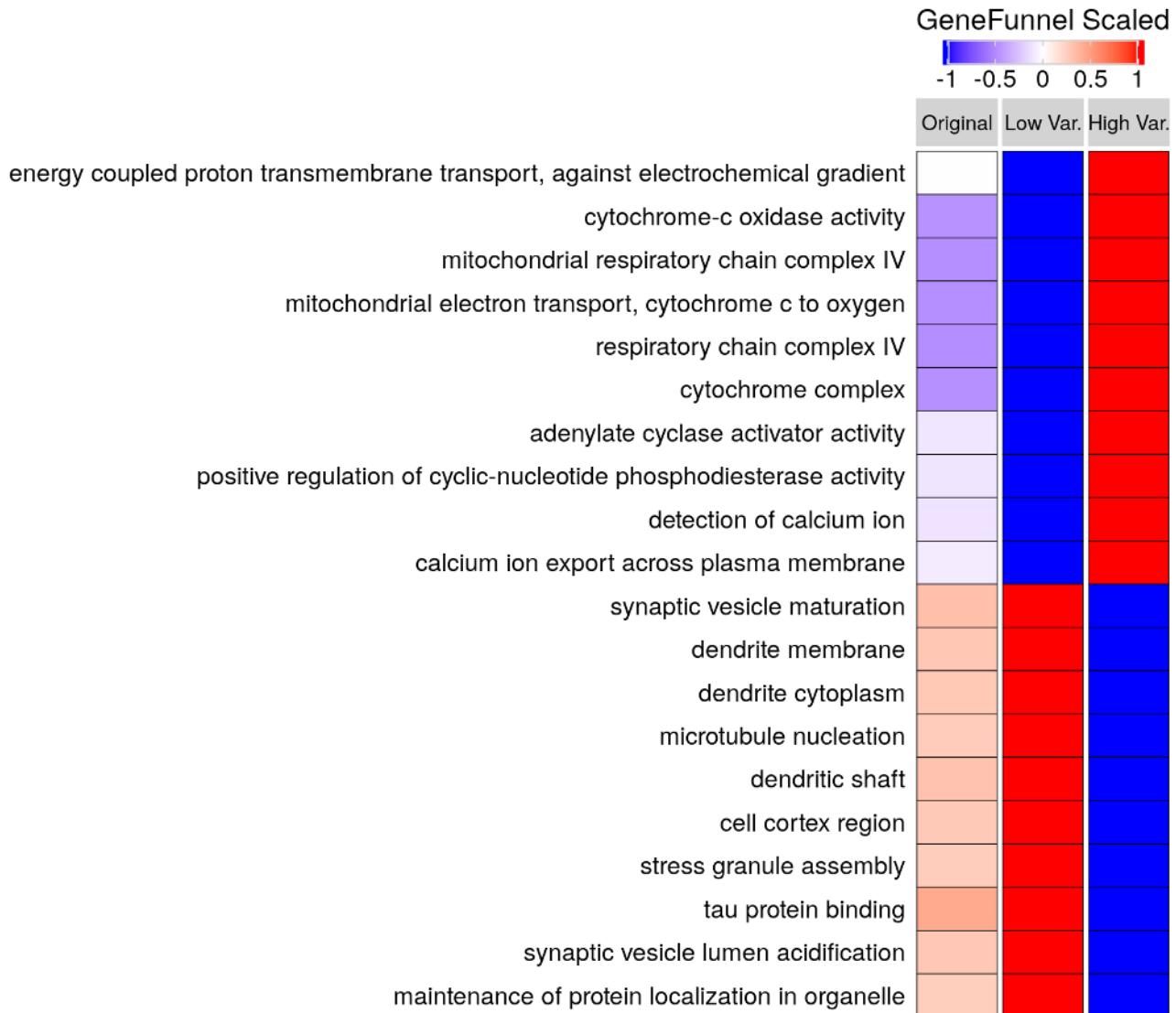


Figure 124: An example of GeneFunnel's powerful ability to detect changes in variability in datasets. Using the pseudobulked tangle-bearing neurons of the FACS ssRNASeq data, the data was modified to produce two new columns. In the Low Var. column, the overall counts profile of the dataset was altered to reduce variability while retaining total library size. In the High Var. column, variability was increased while retaining total library size. After running the three columns through GeneFunnel, the output was sorted to highlight gene sets most changed in the variability modified columns, highlighting GeneFunnel's ability to detect these changes. Per-row scaling is applied for visual purposes to emphasise differences.

The benchmarking performed in this work supports the predictability of GeneFunnel in the controlled scenarios, a major advantage over other methods that often use more complicated algorithms, making intuitive reasoning and troubleshooting difficult. In all of the test cases, GeneFunnel outperformed others significantly in capturing differentially enriched gene sets. In addition, it always maintains independence between samples and gene sets. This is important, particularly as datasets are re-analysed, expanded, or meta-analysed with other datasets. Furthermore, while no method can be absolutely quantifiable when working with data that is intrinsically relative, GeneFunnel retains the original range of genes composing gene sets, i.e. a gene set composed of relatively lowly expressed genes will receive a low expression score. This in contrast to methods that solely focus on differential expression like GSVA. Finally, GeneFunnel carries out its function in an efficient manner, ranking above far above peers in terms of runtime and memory usage.

#### 4.11 Disadvantages and Limitations

Despite its advantages, GeneFunnel comes with certain theoretical and practical limitations that warrant further consideration. One of the most significant concerns is whether variability within a gene set is truly biologically meaningful. While GeneFunnel penalises pathway scores when gene expression is highly inconsistent within a set, it is important to recognize that gene expression levels exist within intrinsic biological ranges. A gene expressed at low levels relative to others in a pathway is not necessarily inactive, its expression may be at the upper limit of its normal dynamic range, even if its absolute expression is much lower than other genes in the set (Buccitelli & Selbach, 2020). Without a comprehensive corpus of predefined gene/protein expression ranges, which is largely lacking in current databases, it is difficult to determine whether observed variability is genuinely reflective of biological dysregulation or simply an artifact of individual gene expression constraints. Addressing this issue would require integrating GeneFunnel with external datasets containing reference expression ranges, an area of research (M. Wang et al., 2012), though currently sparse.

Another key limitation is that GeneFunnel prioritises maintaining resemblance to the original data distribution, ensuring that gene sets with low expression levels correspond to lower scores. While this is useful for interpretability, it may come at the cost of detection sensitivity, particularly when compared to methods such as GSVA, which focus exclusively on detecting relative enrichment between conditions rather than maintaining absolute expression information. By incorporating absolute expression into its scoring mechanism, GeneFunnel may miss subtle cases where differential expression is the primary signal rather than overall expression magnitude. This trade-off makes it particularly important to carefully consider whether absolute or relative expression is more relevant for a given study.

Like all gene set enrichment methods, GeneFunnel's accuracy is inherently dependent on the biases and completeness of the gene set database being used. The gene sets in this study were exclusively derived from Gene Ontology (GO), meaning that the benchmarks

primarily reflect GO-specific enrichment performance. Since pathway definitions vary across different gene set collections, it remains unclear how well GeneFunnel generalizes beyond GO. Future evaluations should test GeneFunnel's performance with alternative gene set databases, such as KEGG, Reactome, or MSigDB, to ensure that its properties hold across different biological ontologies.

Although GeneFunnel never introduces dependencies between samples or gene sets, it does have one key exception where results may become obsolete over time: if the composition of a gene set changes, its score will change. This is because scores are inherently tied to the structure of the gene set itself, meaning that if a pathway definition is updated in future gene set releases, previous analyses may not be fully reproducible. While this issue is unavoidable in any enrichment method tied to evolving biological databases, it highlights a potential drawback for long-term reproducibility in GeneFunnel analyses.

Another important consideration is that the balance between gene set size, deviance penalties, and summation terms were under continuous development, and it remains unclear whether the current weighting scheme is truly optimal. While the method was carefully designed to balance these components, fine-tuning the exact contribution of each factor could further refine GeneFunnel's ability to detect meaningful pathway activity. Related to this, it is also debatable whether summation of expression values and explicit gene set size normalisation is the ideal approach. While most comparable functional class scoring methods incorporate similar normalization strategies, whether intended or not, it is not universally accepted that it is best practice for gene set enrichment scoring.

Finally, it remains uncertain whether the final distribution of GeneFunnel scores are inherently well-suited for downstream statistical analyses such as differential expression testing and dimensionality reduction. While the scoring method was designed to be interpretable and comparable to the original feature by sample expression matrix, the actual statistical properties of the resulting values, such as their distributional assumptions, variance scaling, and impact on downstream modelling, have not been rigorously tested. Many commonly used downstream statistical techniques, such as t-tests, log-transforms, or PCA, make implicit assumptions about data distribution that may not perfectly align with the output of GeneFunnel. This issue is currently assumed rather than proven, making it an area of future investigation to ensure that GeneFunnel scores can be seamlessly integrated into standard transcriptomic workflows without introducing unintended biases.

## 4.12 Implementation Details

Like ImputeFinder, GeneFunnel is being actively prepared for submission to Bioconductor at the following URL: <https://github.com/eturkes/genefunnel>, ensuring that it adheres to best practices for reproducible and well-documented bioinformatics software. The biocthis package was used to structure the package according to Bioconductor guidelines, facilitating smooth integration into the Bioconductor ecosystem. This will enable users to

easily install, update, and incorporate GeneFunnel into their workflows while benefiting from Bioconductor's extensive infrastructure for version control and dependency management.

To ensure efficient computation, GeneFunnel is implemented in Rcpp, leveraging the high-performance RcppArmadillo library for optimised matrix operations. The core algorithm is contained within a single function, which is wrapped in an R interface to maintain accessibility while taking advantage of low-level C++ speed improvements. Additionally, the function is compatible with BiocParallel, allowing users to efficiently compute GeneFunnel scores in parallel across multiple samples, significantly improving runtime for large-scale transcriptomic datasets.

Like ImputeFinder, GeneFunnel does not contain any tweakable parameters beyond an input gene/protein by sample matrix and a list object containing gene sets and the genes contained therein. This is by design to reduce complexity for the user and limit harmful practices such as a data dredging. GeneFunnel is compatible with both raw, untransformed, and unnormalised data, as well as more highly processed data. As GeneFunnel outputs a gene set by sample matrix based on the simple assumptions described here, it is up to the user to decide the direction of further processing and downstream analysis. Where possible however, raw data is preferred as it is less likely that data is removed from the original matrix. Recall that zero values have meaning in GeneFunnel but that it is also unadvisable to retain these values for common preprocessing steps such as normalisation in the source data. The solution therefore is to run GeneFunnel as early as possible in a pipeline and then apply identical parallel pipelines to both the source data and GeneFunnel output. This is the approach taken in this thesis work. However, I demonstrate its use on both raw and processed data, with the FACS ssRNASeq dataset using raw counts and the LCM Mass Spec dataset using log-transformed, normalised, imputed data as necessitated to address the greater quality issues in that dataset. While this is less ideal, I later show that both datasets produce comparable GeneFunnel results.

Regarding the list of gene sets, it can derive from any source, such as GO or Reactome as previously discussed, or user-created gene sets. There are no assumptions regarding overlap of gene sets and the smallest gene set size can be 2. GeneFunnel makes no recommendations regarding filtering on gene set size and aims to be robust against this, unlike GSEA for example, which by default ignores gene sets that contain fewer than 15 features or more than 500 features. That being said, a user may decide to ignore gene sets containing features that may confound an experiment, for example sex-related features. These may be removed before or after running GeneFunnel as GeneFunnel has no dependencies between gene sets. Alternatively, one can remove these features from the gene sets, as long as they are aware that this has a direct impact on the score. The same effect can be achieved by removing these features from the source matrix, as missing and NA values are excluded from calculations (but not zero values). If electing to remove features, or if not all features in the gene sets are in the source data, one may also decide to prune away gene sets that are deemed insufficiently covered by the source data.

For example, because the LCM Mass Spec dataset only covers a few hundred proteins, I choose to only analyse gene sets where at least 50% of the features were in the dataset, as scoring a gene set using only a small fraction of its features did not appear sensible. Such decisions are left to the discretion of the user.

For benchmarking and interactive exploration of GeneFunnel's performance, an online web app is available at <https://data.duff-lab.org/app/genefunnel-benchmarks-viewer>. This platform enables users to visually compare GeneFunnel's scoring behaviour with alternative functional class scoring methods, examine real-time pathway scoring results, and assess its performance under various input conditions. By providing an interactive interface, one can develop a more intuitive understanding of how GeneFunnel processes gene expression data and how it compares to existing enrichment methods. Finally, while not yet available, it is in the roadmap for make available a web app where users can submit datasets for server-side processing. The main web viewer associated with this thesis work (<https://data.duff-lab.org/app/tangle-bearing-neurons-viewer>) provides a glimpse into what such an app may resemble. There remain many possibilities in this front, such as integration with gene, protein, and pathway databases, as well as AI integration, for efficient and seamless exploration and interpretation of results.

## 5. Development of Downstream Analysis Pipeline

### 5.1 Integrated Transcriptomic/Proteomic Differential Expression Analysis

To integrate findings across both transcriptomic and proteomic datasets, a differential expression analysis was performed following imputation using ImputeFinder and functional class scoring using GeneFunnel. This approach enabled the identification of differentially expressed genes (DEGs), differentially expressed proteins (DEPs AKA DAPs), and differentially enriched gene sets associated with tangle-bearing neurons. By leveraging limma (Phipson et al., 2016; Ritchie et al., 2015), a widely used method for linear modeling in high-throughput expression analysis, differential expression testing was conducted across both omics modalities, allowing for comparison of transcriptomic and proteomic alterations in Alzheimer's Disease pathology.

Firstly, it is important to describe the composition of gene sets that were used for GeneFunnel analysis. Gene sets were downloaded from the g:Profiler (Raudvere et al., 2019) website on 2024/08/25, corresponding to the Ensembl 111 release. g:Profiler was selected as the source due to be consistently up-to-date and organised in their procurement of gene sets. The sets chosen were those from Gene Ontology, and sets from all three ontologies – cellular component, molecular function, and biological process – were then combined into a single set. The sets housing ENSEMBL IDs were selected to reduce ambiguity in gene symbols.

Some sets that were considered hard to interpret or confounding in this experiment was removed prior to GeneFunnel analysis to reduce multiple testing burden downstream. Duplicates were removed, and while GeneFunnel should be robust to gene set size, I limited gene set size to 45, to reduce scope towards more specific processes and allow for easier interrogation of gene set enrichment. Though this may introduce a degree of bias, 45 was chosen as it includes all terms of interest related to AD pathophysiology. Since many other unrelated terms are included within this cutoff, the bias was deemed negligible, but it would be indeed the case that statistical testing only on AD related gene sets would be considered invalid due to loss of type I error control (Bourgon et al., 2010). The minimum gene set size was 2, as required by GeneFunnel.

Next, I removed removed redundancies regarding gene sets prefixed with “regulation” or “selection”, opting only to keep the “positive regulation” and “positive selection” variants of them. This reduces ambiguity as terms that are just prefixed with “regulation” for instance, without direction of effect, are unclear as to whether they refer to up or downregulation. In order to not exclude these regulatory terms entirely, the “positive” variants of them were considered easiest to interpret and reduces redundancy. The “negative” variants in particular are problematic as many of these terms differ from their parent term by a single feature; in any gene set enrichment method, these terms can easily be marked enriched even if the key feature is lowly expressed, leading to false interpretations. I then removed

all gene sets containing features on the Y-chromosome as well as the sex-specific features *TSIX* and *XIST*, as the datasets were mixed sex and sex differences were not of interest.

Finally, for the FACS ssRNAseq dataset, the gene sets were subset to only those where all features were in the input matrix, including those with zero values in all cells. GeneFunnel would be run on raw counts without removal of any genes from the initial matrices given by Cell Ranger. Since the LCM Mass Spec dataset had considerably fewer features, on the order of several hundreds of proteins, I only included gene sets where at least 50% of features could be found in the input matrix after imputation. Without doing so, many gene sets would be scored using a very small subset of its features, which would not lead to sensible scoring. Though GeneFunnel was run after normalisation and imputation, the log-transform was reversed before running GeneFunnel.

As both the count matrices and GeneFunnel scoring from the FACS ssRNAseq dataset were at the single-cell level at this stage, the next step was to perform pseudobulking on both the count matrix and scores. Pseudobulking is an increasingly common practice where single cells are aggregated by a metadata label of choice, most commonly by donor as was done here (Zimmerman et al., 2021). The most obvious reason for doing so is to avoid pseudoreplication, an ill-advised practice where single-cells are treated as independent biological replicates during statistical testing. This has the effect of drastically inflating significance scores such as p-values and masking information regarding heterogeneity in true biological replicates such as donors. The second benefit is that mimicking a bulk RNAseq dataset offers a practical solution for overcoming several statistical hurdles specifically associated with single-cell data, particularly sparsity and heteroscedasticity. Pooling across cells is a viable approach for averaging out dropouts (L. Lun et al., 2016) and bulk RNAseq normalisation methods are considered more robust, with more relaxed statistical assumptions compared to their single-cell counterparts (Cole et al., 2019).

Next, filtering and normalisation steps took place, first in the GeneFunnel scores. This pipeline is largely based on a standard bulk RNAseq pipeline frequently suggested when working with pseudobulked data, particularly in preparation for limma-trend analysis (Y. Chen et al., 2016). For filtering, the *filterByExpr* function from EdgeR was used (Robinson et al., 2010). This is a function that accepts a design matrix or other designation of the contrasts of interest. By doing so, it aims to remove features that have little chance of being called DE. Specific scenarios that *filterByExpr* excels over more naive approaches include those where a feature is expressed in one treatment group, but absent in another. Less optimised approaches might require the feature have minimum expression across a number of samples, without taking into account group information, but *filterByExpr* only requires that minimum expression be in at least one of the groups.

*filterByExpr* has a number of parameters but it is intended to have sensible defaults, so no changes were made to these defaults. However, it was applied in a specific way alongside normalisation, as suggest by one of its authors Aaron Lun on the Bioconductor forums

(<https://support.bioconductor.org/p/116351/#116369>). Essentially *filterByExpr* was run twice, one time in order to calculate normalisation factors which are added back to the original, unfiltered object, and then the final filtering done with the normalisation factors at hand. The reasoning behind this workflow is that accurate calculation of normalisation factors (done through the *calcNormFactors* function in EdgeR) requires that very low counts (as well as GeneFunnel scores) are first removed, so this is done using *filterByExpr*. However, rather than use this filtered object as the final object, the normalisation factors are used to inform a more performant filtering that may better handle compositional biases in the data. Internally, *filterByExpr* transforms the data into counts-per-million (CPM) and in the absence of normalisation factors, it simply uses the library size of each sample. But with the normalisation factors available, it will use this information, which in theory should result in a superior, or at least more informed, filtering approach.

After the final filtering, normalisation factors are calculated once again, before transformation into the final log2CPM normalisation, as recommended for limma-trend analysis. A notable parameter at this stage is the choice of prior count. In order to avoid taking the log of zero, a small CPM value is added before log transformation. It is recommended to optimise this parameter with inspection of an SA plot, which plots residual standard deviation against average log expression. When using limma-trend, limma attempts to fit a trend against this relationship and includes a boolean parameter for robustness which ignores outliers. An optimisation is to adjust prior count, which is by default set to 2, to minimise the number of outliers. These outliers are typically in the lowest log expression range, as these features tend to have the highest variances. Increasing the prior count has the effect of clamping down on these variances. Increasing prior count should be done judiciously though, as the SA plot should still exhibit a trend showing that low expressing features have greater variances – this trend should not flat as limma-trend does not expect complete variance stabilisation, just a reduction in outliers. Furthermore, too large of a prior count has the effect of artificially inflating the values of the original counts. This process could be performed without issue on GeneFunnel scores, and a value of 6 was selected as a suitable prior count. Figure 125 shows the SA plot of GeneFunnel scores with this prior count, and Figure 126 shows the SA plot of the counts for comparison. The counts pipeline is very similar and described in the next paragraph.

The counts pipeline was performed in an identical fashion to the scores, with one major exception. Before running *filterByExpr* and the following steps, the counts were filtered to only those genes that are present in the GeneFunnel gene sets after the GeneFunnel scores were filtered and normalised, as described above. In order to reduce multiple testing burden as much as possible, I elected to only test for features that were also in gene sets to be tested. Since the filtering procedure for GeneFunnel scores thus far only uses non-specific filtering and the initial corpus is derived from the complete Gene Ontology, the gene sets should not be biased towards AD-related terms. This is a key point, as doing so would invalidate statistical assumptions for both testing of the scores and the features that would be filtered from the scores (Bourgon et al., 2010). With this in

mind, I filtered the genes as described and ran them through an identical pipeline as the scores, with the only difference being that the ideal prior count was found to be 4 rather than 6.

**Figure 125.**

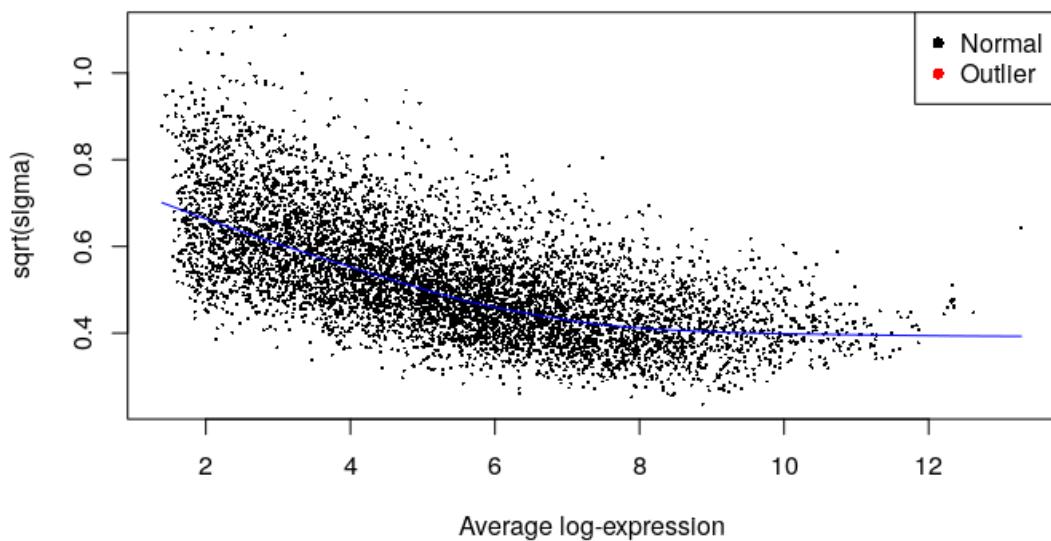


Figure 125: SA Plot of GeneFunnel scores in the FACS ssRNAseq dataset showing standard deviation against average log expression. A prior count of 6 before logCPM transformation was found to minimise outliers when fitting the trend, while retaining the characteristic mean-variance relationship that the limma-trend pipeline expects.

**Figure 126.**

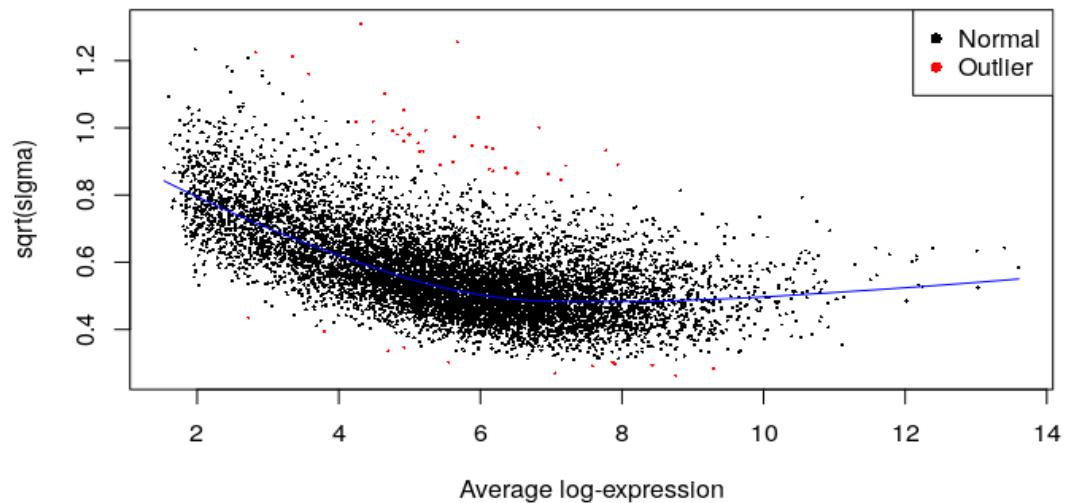


Figure 126: SA plot of the counts in the FACS ssRNASeq dataset. A prior count of 4 was found to be ideal in this case. Overall characteristics between the SA plot for GeneFunnel score and the originating counts are similar, suggesting that the scores have statistical properties similar to the input data and can be passed through a parallel analysis pipeline.

At the point of GeneFunnel scoring, the LCM Mass Spec data was already filtered, transformed, normalised, and imputed, in contrast to the raw counts for the FACS ssRNASeq dataset. Therefore, its pipeline for limma-trend preparation is far more minimal. Notably, only the scores received a single *filterByExpr* step on scores with default parameters, as GeneFunnel produced an abundance of low scoring gene sets. There was no calculation of normalisation factors as the input data was already normalised. In any case, no compositional biases are expected as the data does not originate from a single-cell level, where even after pseudobulking individual cells are expected to contribute some degree of variance that can be corrected for. Like with the FACS ssRNASeq dataset, after filtering the GeneFunnel scores, the protein matrix was subset to those matching features of the gene sets retained in the score matrix. Prior count was optimised for both scores and proteins, but the default setting of 2 was found to be sufficient for both.

After completing these parallel pipelines, genes, proteins, and gene sets were passed through a conventional limma-trend pipeline, as described in the limma documentation. limma-trend was chosen for its superior performance in a comprehensive benchmarking study of 36 differential expression methods for single-cell analysis (Soneson & Robinson, 2018). Limma is also highly flexible and being based on straightforward linear modelling, does not impose strict statistical assumptions. It has been shown to be compatible with testing of functional class scoring output (Hänzelmann et al., 2013) as well as mass spectrometry proteomics (X. Zhang et al., 2018). The limma-trend variant is designed to work with any data that exhibits a mean-variance trend, which is observed both in transcriptomics and proteomics, and so the same pipeline can be applied to both, making it ideal for integrated analysis.

Only one modification was made to the default limma-trend pipeline, that is the use of the *duplicateCorrelation* function prior to fitting the linear model. This function adapts the modelling to account for repeated measures. Recall that each pair of tangle-bearing vs. non-tangle-bearing neurons are sampled within the same patient donors. While this can be analysed naively, i.e. with every sample as a unique replicate, one can afford more power by incorporating such sampling information into the statistical design. *duplicateCorrelation* achieves this by treating donor as a random effect, with equal magnitude across all features. While effective, this does suggest a point of future optimisation. Because features may interact differently with this random effect, a linear mixed model may be the basis of a superior analysis. limma has been adapted to support mixed models with the dream package (Hoffman & Roussos, 2021), which is under consideration for future work with these experimental designs.

Finally, after statistical testing and Bayesian inference using limma-trend with default parameters (aside from using the robustness parameter to exclude outliers in the fitting of the trend, see Figures 125 and 126), I obtained a table of the test results comparing tangle-bearing neurons with non-tangle-bearing neurons in genes, proteins, and their corresponding GeneFunnel scores (four total tables). The next and final crucial step of this pipeline was the implementation of multiple testing correction. While each table could be corrected independently, as a default and potentially naive approach would entail, it would be considered more valid to account for all testing performed in the entire experiment at once. In such cases, especially when using FDR, this can in fact lead to increase in power while still controlling for error, as discussed in the limma documentation (<https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf>). Therefore, I elected to concatenate all unadjusted p-values from the four tables and correct them using the default BH (Benjamini-Hochberg correction) together. A significant feature was then defined as being below the adjusted-p < 0.05 cutoff. By applying a parallel pipeline across the different modalities and carefully accounting for the total testing performing, this integrated analysis framework allowed for a direct comparison of transcriptomic and proteomic alterations, providing a comprehensive view of Alzheimer's Disease associated molecular changes in tangle-bearing neurons.

## 5.2 Development of Network Analysis and Hub Selection Approach

In order to explore the large breath of gene set enrichment and differential expression results, extensive interactive network graphs were built using VisNetwork, a Javascript library with an R implementation (<https://github.com/datastorm-open/visNetwork>). The main goal was to visualise differentially expressed features, easily assessing if features were differentially expressed/abundant in the FACS ssRNAseq dataset, the LCM Mass Spec dataset, or both. Furthermore, I aimed to connect features on the basis of shared differentially enriched gene sets. Because doing so with all differentially expressed/enriched features resulted in networks too large to be navigable, I had to establish metrics for the pruning of the networks. This was implemented as a slider, allowing for an interactive range of network sizes. The networks are available at the landing page for the main analysis of this thesis: <https://data.duff-lab.org/app/tangle-bearing-neurons-viewer>, with code available at <https://github.com/eturkes/tangle-bearing-neurons>.

**Figure 127.**

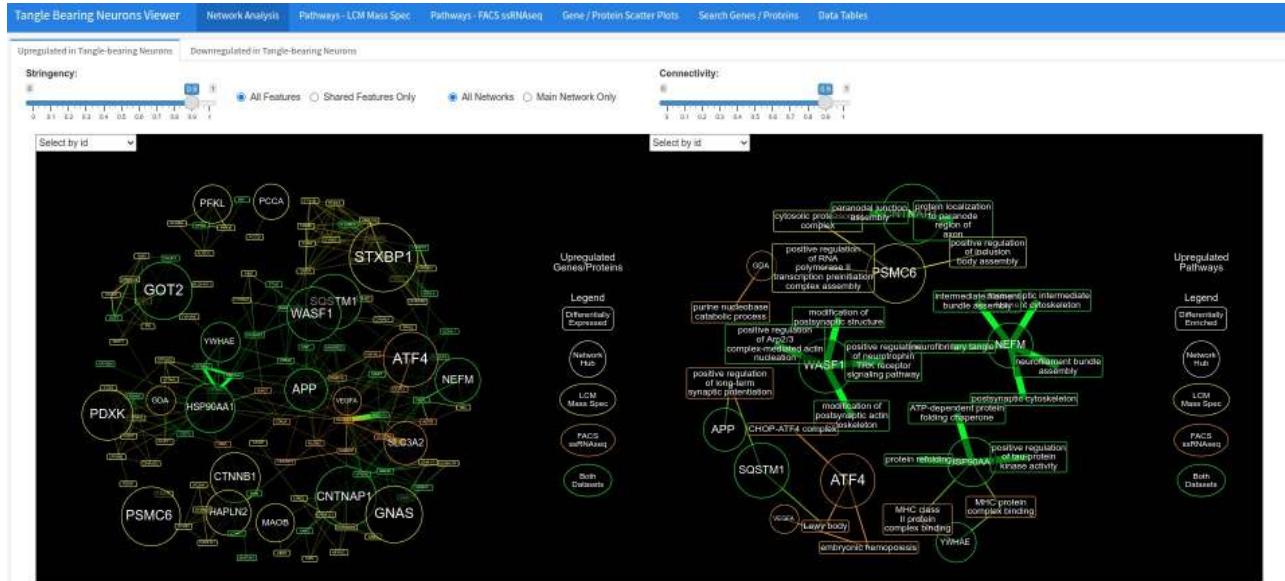


Figure 127: Landing page of the viewer associated with the main work of the thesis, an integrated transcriptomic and proteomic analysis of tangle-bearing neurons in AD. The VisNetwork figures are the first aspects of the analysis a user interfaces with.

The network is built by first collating the four differential expression/enrichment tables output by limma in the last section (DEGs and GSE for the FACS ssRNAseq dataset; DAPs and GSE for the LCM Mass Spec dataset). When associating gene sets with features, if the gene set was enriched in the opposite direction as the feature, it was discarded. This is due to the ultimate design of the network as being built to find hub features. These hubs, as will be described, are highly prioritised through concordant evidence across gene sets. It was considered to apply a penalty on hub features when discordant gene sets are present, with the highest penalty being assigned when an equal amount of gene sets are enriched in both directions, cancelling out the hub feature candidate score. Ultimately, for this analysis, it was decided that feature differential expression/abundance is a more reliable signal over gene set enrichment, which is sensitive to inaccurate/incomplete annotations and requires the use of less robust methods. Therefore, the direction of feature was established as “ground truth” and gene set evidence was accumulated on only a positive basis. This explicitly prioritises feature-level signals over pathway-level trends when they conflict, and it is noted that this bias may be better addressed with a more sophisticated signing and weighting system in the future.

Though antagonistic sign information regarding gene sets were not incorporated into this analysis, sign was taken into account by creating separate networks for upregulated and downregulated features / gene sets. In each, features were ranked by adjusted p-value, and scaled such that the range lies between 0 and 1 to facilitate relative thresholding of the values using the “Stringency” slider, which is set to 0.9 by default. At this stage, the

LCM Mass Spec dataset comprised 218 DAPs while the FACS ssRNASeq dataset comprised 1,207 DEGs, making it challenging to equally represent both datasets on a single network. In order to do so, filtering of the features was designed to be adaptive to these sizes. For both filtering of DEGs and DAPs, the smaller of the lists was divided by the larger, in other words, 218 divided by 1,207 which equals 0.1806131. This was then either multiplied by the Stringency value for filtering the larger list of DEGs, or multiplied by the inverse (1 minus Stringency) for the smaller DAPs. Finally, that value was added to 0.1806131. In effect, at the default Stringency of 0.9, the scaled adjusted p-value rank of DAPs had to be above 0.1986744 to be kept, while DEGs had to be above 0.9819387. This was found effective for more equally representing both datasets in a single network, while giving intuitive meaning to the Stringency slider. Essentially, as the slider value approaches 1, more DEGs are filtered and more DAPs are kept, whereas as it approaches 0, more DAPs are filtered and more DEGs are kept. This explains the effectiveness of the default value of 0.9, it balances out the greater number of DEGs in this dataset.

With the features filtered, a co-occurrence matrix is built to count the number of differentially enriched gene sets a feature appears within. This is used to build a weighted, undirected graph using igraph (Csardi & Nepusz, 2006), where features are nodes and edges are gene sets. Networks are created individually for DAPs and DEGs, but they are not visualised at this stage. Instead, they are used to assign some features as network hubs. A three-part criteria is used for this, with each criteria scaled so that they contribute equally when multiplied together. The first part consists of the adjusted p-values for each feature, described prior, though without ranking this time to retain magnitude of relative differences between features like the other parts of this criteria. The second part is to count the number of enriched gene sets a feature is enriched in; in the co-occurrence network, this corresponds to the number of unique edges of each node. However, this simple count is modified to account for issues such as gene annotation bias (Haynes et al., 2018). A feature such as APP may appear in many gene sets solely on the virtue of it being highly studied. To correct for this, a simple but effective approach is to find the proportion of gene sets the feature is enriched in, out of all the tested gene sets the feature is a member of, and this was the approach taken for this second criteria. The third part of the criteria is to count the number of unique features that share enriched gene sets with the original feature; in the co-occurrence network, this corresponds to the number of unique neighbour nodes to the original node. No further corrections are needed for this part of the criteria. After multiplying these parts together and creating scores for each feature, the scores are scaled, and this time ranked.

The Stringency value was again used to designate the top scores as hubs. This time, a function was written to weight the Stringency about the midpoint of 0.5. The function is such that when Stringency is 1, the upper value is 0.625 and the lower value is 0.375. When Stringency is 0, these values are reversed, and when it is 0.5, both the values are 0.5 as well. The upper value is used as a filter for DAPs, the smaller of the two lists, so that a protein is a hub if its hub score is above this value. On the other hand, the lower value is used to filter DEGs, and a hub must be above that value. Note that this is inverse

from the initial filtering using Stringency; this time, the more strict criteria is being applied to the smaller list, which initially had looser filtering. This seems counter-intuitive but was found effective for balancing the ratio of DE features and hubs in both datasets.

With the features filtered, and network hubs selected, the co-occurrence network to be visualised was created using data from both datasets. A toggle switch is available, allowing the network to be created using all features or only those features that are DE in both datasets. This can be used alongside another toggle, that cuts down the network to only those in the “main network”. This main network is defined as nodes connected to the largest contiguous network of nodes, excluding those isolated from it. Note that neither of these have an effect on the selection of hubs, which takes place before this step. The layout of nodes uses a more advanced algorithm than those in igraph, using the Fruchterman Reingold algorithm (Fruchterman & Reingold, 1991) from the qgraph package (Epskamp et al., 2012). As node layout only has implications for visualisation, this is not a very important implementation detail, it was just found to subjectively improve initial placement of nodes, clustering them neatly nearby nodes with shared differentially enriched gene sets. Moreover, visNetwork allows a user to drag and move nodes to their discretion or as needed for better readability.

**Figure 128.**

```
hub_mat <- as.data.frame.matrix(table(net_data[, c(3, 2)]))
hub_mat <- hub_mat > 0
hub_mat <- hub_mat %*% t(hub_mat)
diag(hub_mat) <- 0
graph <- graph_from_adjacency_matrix(
  hub_mat, mode = "upper", weighted = TRUE
)

keep_idx <- which(V(graph)$name %in% c(keep2, keep3))
keep_degree <- unique(
  unlist(
    neighborhood(graph, order = map_range(inverse), nodes = keep_idx)
  )
)
keep_degree <- unique(c(keep_idx, keep_degree))
graph <- induced_subgraph(graph, vids = keep_degree)

if (input[[paste0("shared_up", 23)]] == "All Features") {
  graph <- subgraph(graph, which(degree(graph) > 0))
}

if (input[[paste0("comp_up", 23)]] == "Main Network Only") {
  graph <- subgraph(graph, which(components(graph)$membership == 1))
}
```

Figure 128: Code block showing the creation of the co-occurrence network using data from both datasets, followed by subsetting the network to those used for hub selection, and then evaluating the toggle switch for the plotting of all features vs. the main network only. Figure 129 shows an example of the final co-occurrence network on upregulated features

and their enriched gene sets. Describing the aesthetic features, hubs were shaped into circles, while all other DE features were shaped as rectangles. They were also coloured in accordance to the dataset they are associated with, with yellow for LCM Mass Spec only, orange for FACS ssRNAseq, and green for both. Hubs were sized in proportion to their hub score, while all other features were given a static size for readability. Finally, edges were sized according to the number of shared differentially enriched gene sets between the nodes. visNetwork also allows for extensive interactivity which was fully utilised. Users can hover over a node to see what enriched gene sets they are associated with, along with which dataset they are enriched in. This can be used in combination with clicking on a node, which highlights its neighbours. The highlighting method is also set to reveal the names of second degree neighbours without highlighting them, allowing for multiple layers of information. A second degree neighbour is that which is not directly connected to the original node, but is connected through another node that the original node is connected to. The clicking of nodes can alternatively be performed through a drop-down menu by searching for and clicking on the feature name.

**Figure 129.**

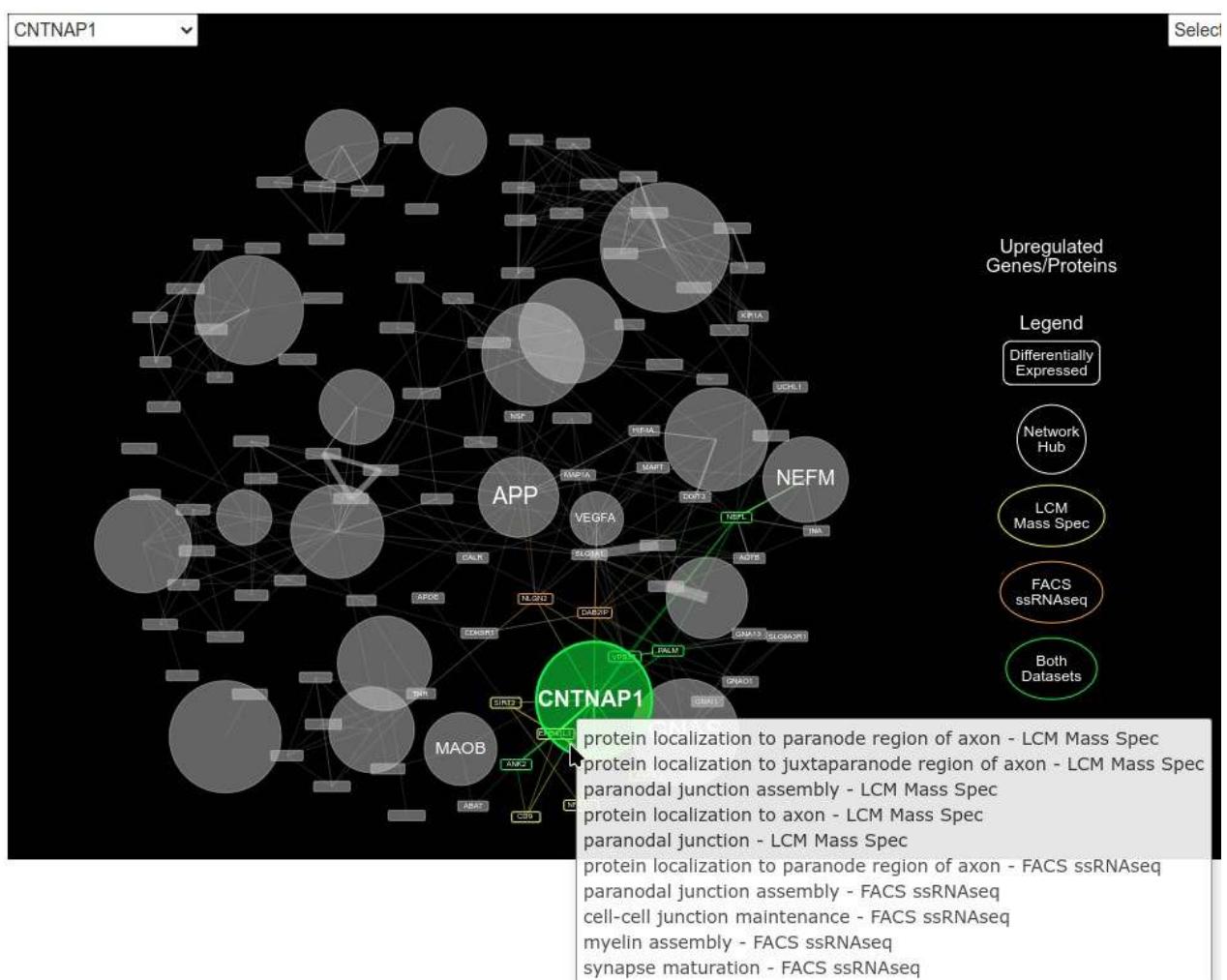


Figure 129: Demonstration of the aesthetics and interactive features of the graph network. In this example, a node was clicked on, which highlights all of its neighbours. Additionally,

second degree neighbours are shown with their labels but without highlighting. This can also be accomplished by selecting the node's label in the drop-down menu in the top-left. The mouse cursor is also left to hover over the node – after a few seconds, this produces the nearby menu, showing enriched gene sets for that node and the dataset from which the enrichment derives.

To complement the main networks, an additional network focusing more specifically on gene sets is produced alongside it in the viewer (the right-hand network in Figure 127). Being a bipartite rather than co-occurrence network, it shows both genes/proteins and gene sets as nodes. This graph is dependent on the output of the previous network, using only those features designated as hubs. It features an additional slider called “Connectivity”. This slider implements a simplified version of the filtering approach of the Stringency slider and is also set to a default of 0.9. The adjusted p-values of gene sets are ranked and then scaled, and then any gene sets greater than or equal to the Connectivity value is selected for network creation. No separate filtering for each dataset is performed, as there is no hub section for this network. With the features selected by the previous network and the gene sets selected with the Connectivity slider, network aesthetics are defined in a similar way as previously described, however without a hover-over function. An example of this network is seen in Figure 130.

**Figure 130.**

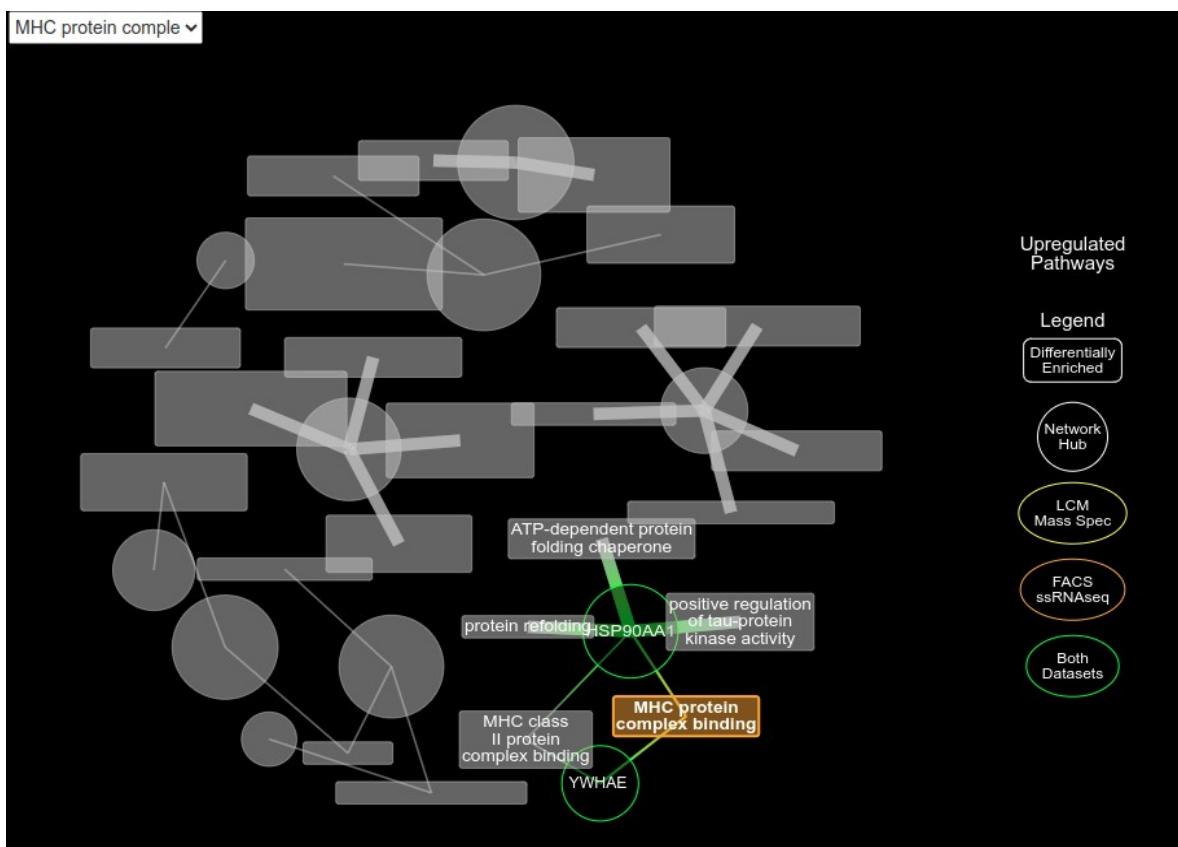


Figure 130: The complementary bipartite network showing both network hubs and a

subset of their enriched gene sets, pruned using the Connectivity slider. Note that in this network, gene sets can also be selected, showing their first and second degree neighbours. However, no functionality is currently implemented when hovering-over a node.

The same process described above was implemented for downregulated features and gene sets, though with different initial values for Stringency and Connectivity (both set to 0.1 rather than 0.9). As will be discussed in the coming sections, very few proteins and gene sets were downregulated in the LCM Mass Spec dataset, and no combination of parameters could produce a balanced network that didn't result in an overly small network. With these parameters set low, the networks appeared to benefit from more representation from the FACS ssRNAseq dataset at least, without much loss in information of the already small information coming from the LCM Mass Spec dataset (an advantage of using scaled ranking, which will never cause the filtering to discard all of the data). In any case, it was decided that upregulated features and gene sets would be the focus of this thesis work, so this downregulated network is available more for completeness and additional exploration.

**Figure 131.**

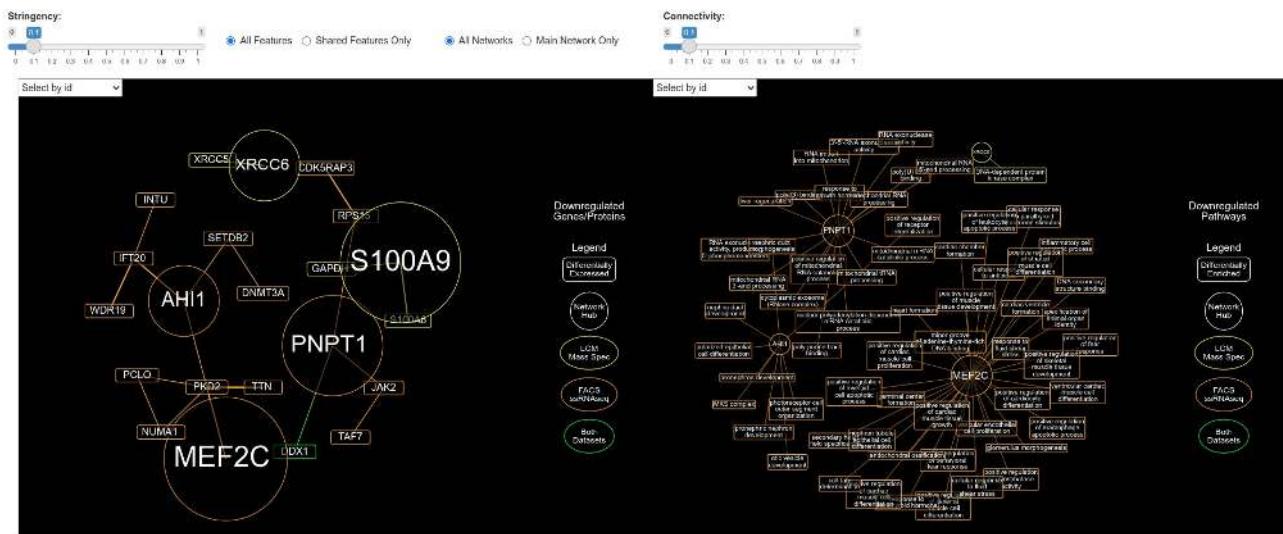


Figure 131: Networks of downregulated features and gene sets from the web viewer. Note that the Stringency and Connectivity parameters are set to 0.1 rather than 0.9. A more balanced network between both datasets could not be achieved due to the very downregulated proteins, so the parameters were set to allow more representation of genes as a compromise.

### 5.3 Development of Web Viewers

Given the large volume of results generated in this study, an interactive data viewer was developed to facilitate efficient exploration and visualization of key findings. The sheer number of differentially expressed genes, enriched gene sets, and their interconnections made it impractical to rely solely on static figures or tables for interpretation. Instead, a

dynamic web-based solution was implemented, allowing users to interactively query and explore the data in real time.

To ensure that the application could be securely hosted and accessed remotely, a dedicated Debian VPS server was rented, and a custom domain name (<https://data.duff-lab.org/>) was registered for ease of access. A Dockerised setup was selected as the deployment strategy due to its advantages in reproducibility, scalability, and ease of maintenance. Docker image files were specifically designed to run ShinyProxy, a Java-based server that enables the self-hosting of ShinyApps with containerized session management. The configuration for the server can be found at <https://github.com/duff-lab-team/shinyproxy-docker-compose>. Web applications themselves were developed using R Markdown with flexdashboard and a Shiny runtime, ensuring a balance between interactivity and structured reporting.

To improve performance and optimize resource utilization, a containerised approach was implemented. When a user accesses the application, the R Markdown file is dynamically compiled from scratch, and a unique Docker container is assigned to the session. To mitigate the computational overhead of on-demand rendering, computationally expensive steps, such as preprocessing, differential expression analysis, and enrichment calculations, were pre-cached in advance. This significantly reduces processing time while ensuring that users still have access to the most up-to-date results. Additionally, zram, a Linux-based memory compression tool, was deployed to allow for aggressive in-memory compression of up to 3x the available RAM, enabling the server to handle concurrent users beyond its raw memory capacity.

The main interactive analysis portal can be accessed at <https://data.duff-lab.org/app/tangle-bearing-neurons-viewer>, providing a comprehensive interface for examining the core results of this study. For benchmarking and testing of GeneFunnel's functional class scoring performance, a separate web viewer is available at <https://data.duff-lab.org/app/genefunnel-benchmarks-viewer>.

While this Dockerised approach works well for smaller analyses, it was found to scale poorly for the work in this thesis, primarily the main analysis, as the R Markdown document is compiled upon each request. To better accommodate users in accessing the data quickly, an experimental Linux kernel optimization called Checkpoint/Restore in Userspace (CRIU) was utilised. CRIU allows for freezing and restoring running processes, effectively enabling the storage of an already-initialized web viewer instance inside the Docker image. Instead of launching from scratch, a user request now restores a pre-frozen session in just a few seconds, significantly reducing the startup delay.

This technique was implemented for the main analysis portal, where standard initialization takes up to 10 minutes due to the large dataset size and overhead of R Markdown compilation. The experimental fast-loading version of the app is accessible at <https://data.duff-lab.org/app/tangle-bearing-neurons-viewer-quick>. While this method

greatly accelerates access, the current implementation is hacky and not fully functional. For instance, the source code of various dependencies including Shiny itself had to be modified to create a prototype version. Improvement of the CRIU-enabled viewer is an area of active development.

By combining scalable cloud-based hosting, containerised execution, memory optimization, and experimental process freezing techniques, this web-based solution provides a powerful and flexible means for researchers to interactively explore transcriptomic and proteomic results while keeping computational demands manageable. Future improvements will focus on enhancing the stability of the CRIU-based system, further optimising memory efficiency, and expanding the options for user query of results.

**Figure 132.**

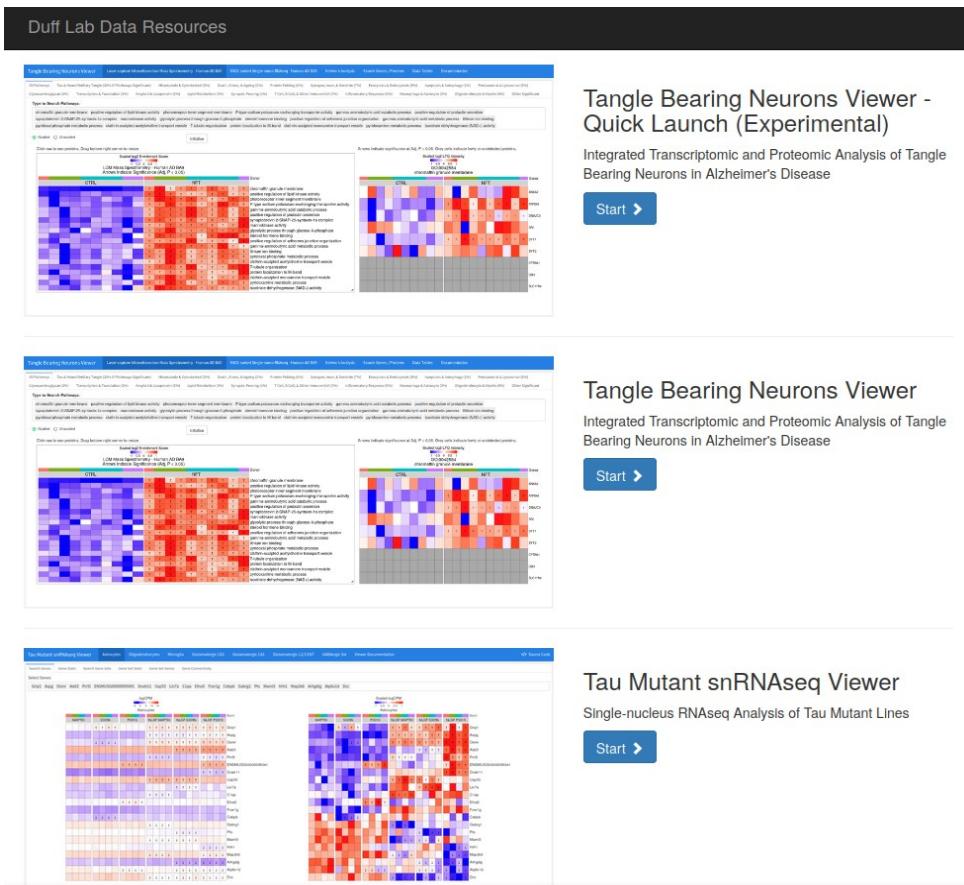


Figure 132: Screenshot of the Duff Lab website, which I developed to host various exploratory analyses both within and outside the scope of this thesis work.

## 6. Results

### 6.1 Overview of Transcriptomic and Proteomic Analysis of Tangle-bearing Neurons

In this study, the molecular profile of tangle-bearing and non-tangle-bearing neurons from post-mortem human prefrontal cortex tissue in Alzheimer's Disease donors were compared. Two datasets were used in this study, one using laser-capture microdissection coupled with mass spectrometry (LCM Mass Spec) to generate a proteomics profile, and another using FACS sorting coupled with single-soma RNA sequencing (FACS ssRNASeq) to generate a transcriptomics profile. The LCM Mass Spec dataset was generated in-house while the FACS ssRNASeq dataset had been previously available. Both datasets underwent their correspondent pre-processing pipelines described in Methods, and then were analysed using novel or highly tailored methods described throughout Sections 3-5.

The results of differential gene/protein (feature) expression/abundance and differential gene set (pathway) enrichment in the groups of interest in the two datasets are summarised in Figure 133.

**Figure 133.**

Dataset	Pathway Sets Tested	Genes/Proteins Tested	Pathway Sets Differentially Enriched	Genes/Proteins Differential Expressed	Pathway Sets DE UP	Pathway Sets DE DOWN	Genes/Proteins DE UP	Genes/Proteins DE DOWN	Percentage Pathway Sets DE	Percentage Genes/Proteins DE
LCM Mass Spec	362	665	191	323	171	20	262	61	53%	49%
FACS ssRNASeq	6,871	8,950	2,549	3,669	1,279	1,270	1,771	1,898	37%	41%

Figure 133. Basic summary metrics of differential gene/protein expression/abundance and differential pathway enrichment between tangle-bearing and non-tangle-bearing neurons in the LCM Mass Spec and FACS ssRNASeq datasets.

Figure 133 demonstrates robust molecular changes taking in tangle-bearing neurons on both the proteomic and transcriptomic level. It also highlights large differences in coverage and bias between the two datasets. Notably, the first two columns tally the number of pathways and genes/proteins (features) that were tested in each dataset, and while the FACS ssRNASeq dataset assessed 6,871 pathways and 8,950 features, the LCM Mass Spec dataset could only cover 362 pathways and 665 features. These numbers are derived from the pathways and features used prior to null hypothesis testing by limma and after all QC and filtering for low enrichment/expression. The numbers reflect a typical contrast between mass spectrometry and NGS sequencing approaches, which differ greatly in coverage. Nonetheless, looking at the last two columns, both approaches yielded large percentages of pathways and features that were called DE at a standard FDR cutoff of 0.05. Interestingly, though proteomics coverage was lower, a higher percentage of its pathways and features were found DE, at 53% and 49% respectively,

compared to 37% and 41% in the transcriptomics experiment. Finally, the middle columns show the exact numbers of pathways and features DE between the two datasets, reflective of the total tested and percentage called DE. While the number of upregulated and downregulated pathways and genes did not differ greatly in the FACS ssRNAseq dataset, there was a large difference in the LCM Mass Spec dataset, with 171 and 262 pathways and proteins upregulated but only 20 and 61 pathways and proteins downregulated. The overlap of pathways and features between the two datasets are summarised in the Venn diagrams of Figure 134.

**Figure 134.**

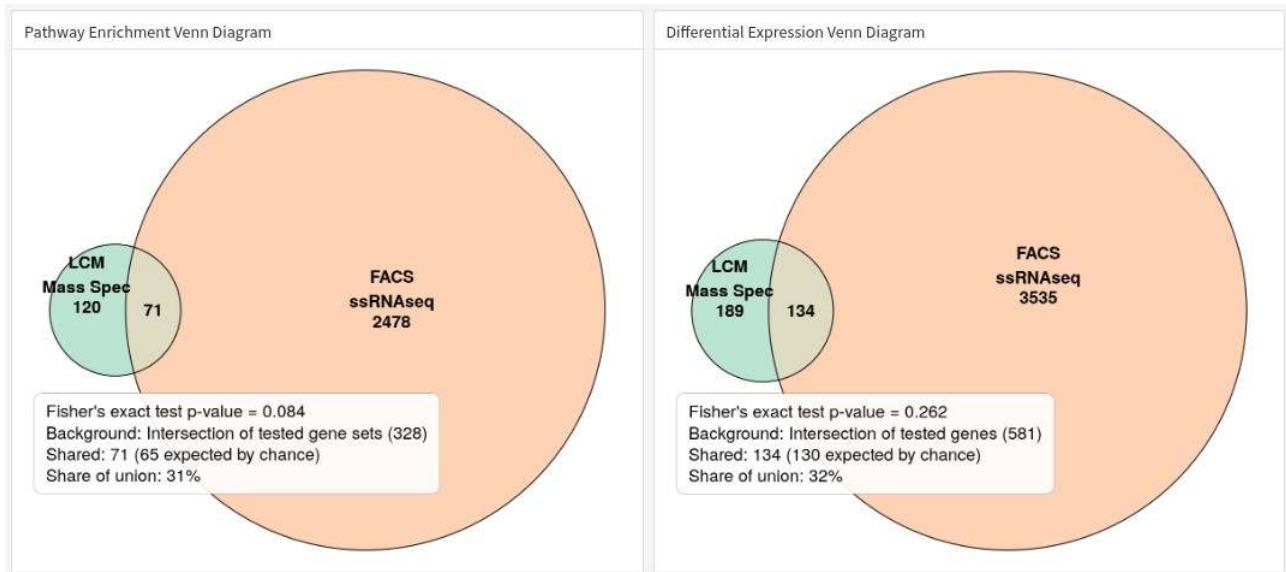


Figure 134. Venn diagrams showing the overlap of pathways and features between the LCM Mass Spec and FACS ssRNAseq datasets, alongside one-sided Fisher's exact test results assessing whether the overlap is greater than expected by chance.

Despite large differences in the number of discoveries in each dataset, 71 pathways and 134 features are shared. Using a one-sided Fisher's exact (hypergeometric) test with the background defined as the intersection of items tested across datasets ( $N = 328$  gene sets;  $N = 581$  genes), the expected overlaps by chance were 65 and 130, respectively. The observed overlaps are only slightly above expectation (pathways:  $p = 0.084$ , features:  $p = 0.262$ ) and thus not statistically significant below a threshold of 0.05. The share of the union overlapped is 31% for pathways and 32% for features. In practical terms, the assays have some findings in common, but each also contributes a sizeable set of unique pathways/features, reflecting notable divergence in either assay biases or transcriptomic/proteomic signals.

## 6.2 Top Differentially Expressed Proteins and Genes and Associated Pathways

The summary metrics demonstrate statistically significant changes taking place in tangle-bearing neurons across hundreds of proteins and thousands of genes. A common approach for highlighting key changes is to sort the list of features by a summary statistic such as adjusted p-value and to focus on the top N features. Such approaches are potentially naive, as effect size and consistency among replicates may not necessarily be informative of changes on a mechanistic level, inspiring the development of the pathway enrichment methods in this work. Furthermore, even if top features do appear to reflect mechanistic changes, it is common for them to only reveal a subset of all changes. For instance, many features co-regulate with others mechanistically, and a list of top N features may only comprise these features and not contain groups of co-regulated features further down the list. Nevertheless, top N approaches are easy to interpret and still very important in exploratory understanding of the data. Figure 135, shows one such figure of the top 25 DE features between tangle-bearing and non-tangle-bearing neurons in the LCM Mass Spec and FACS ssRNASeq datasets.

**Figure 135.**

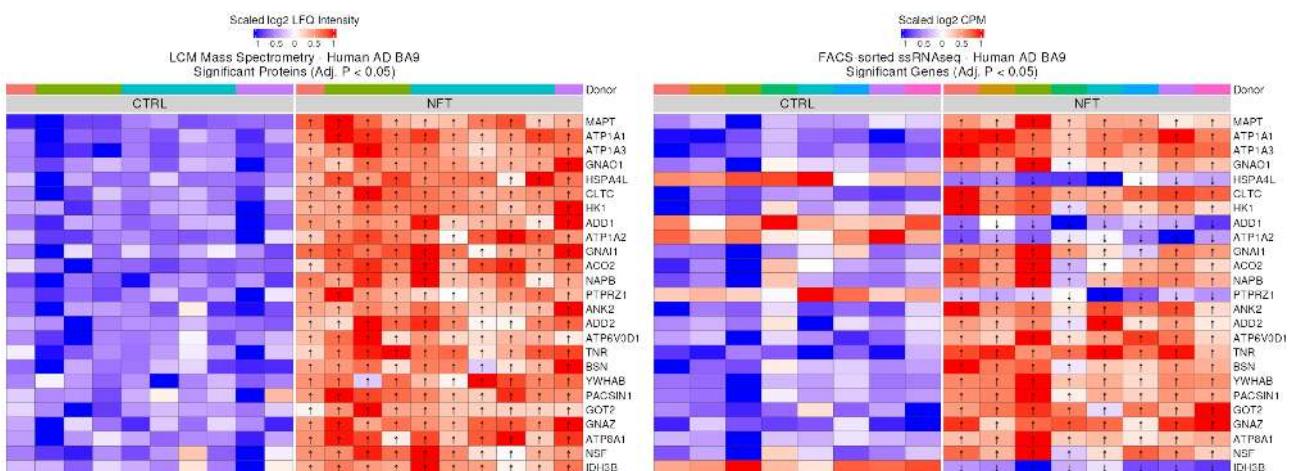


Figure 135. Heatmaps of the top 25 DE features when comparing tangle-bearing and non-tangle-bearing neurons (labeled NFT and CTRL, respectively) in the LCM Mass Spec and FACS ssRNASeq datasets. Cells corresponding to DE features are marked with an up or down arrow indicating up or downregulation at an adjusted p-value cutoff of < 0.05. Features are scaled from -1 to 1 within each row to highlight differences between groups. Biological replicates are indicated by the Donor annotation at the top of each heatmap. Donors were not such shared between datasets.

The most striking aspect of Figure 135 is the fact that when both datasets are subset to shared features, MAPT tops the list in adjusted p-value ranking for both proteins and genes. This is a key finding that confirms the validity of both experiments, as both centre around single-cell capture of tangle-bearing neurons, which were defined by positive

nuclear staining of the phosphorylated form of the *MAPT* gene product tau using the ATA8 antibody. The effect size of this change can be further investigated in the unscaled version of the heatmaps (Figure 136). Figure 136 shows that the change is strongest in the proteomics experiment and more subtle on the transcriptomics level. This aligns with known understanding of disease biology, where substantial increases in tau protein is readily detected in late-stage AD but changes in *MAPT* gene expression remain uncertain, perhaps due to the low effect size exhibited in this study.

**Figure 136.**

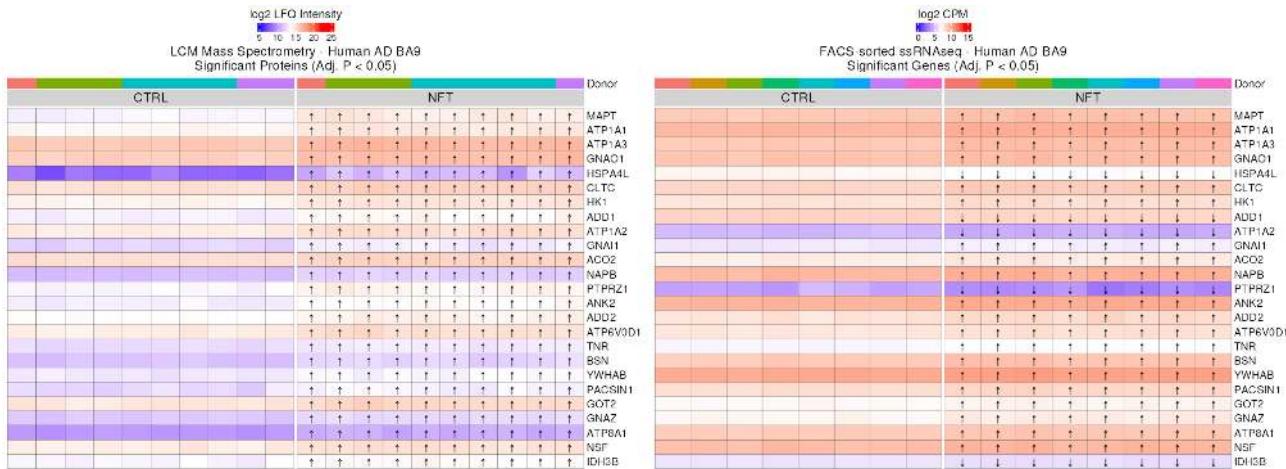


Figure 136. Equivalent figure to Figure 135, however no scaling is applied and visualisation shows the native log2 LFQ intensities and log2 CPM values. The colour range is set to cover the entire range of values of each dataset including those outside of the heatmap.

Regarding other features in the top 25, most change in the same direction between the proteomics and transcriptomics datasets. While all of the top features are upregulated in the proteomics dataset, 4 are downregulated in the transcriptomics – *HSPA4L*, *ATP1A2*, *PTPRZ1*, and *IDH3B*. This shows that while the two modalities largely align, it cannot be assumed that all changes take place in an equivalent fashion on the protein and gene level. Indeed, by analysing the top features of each dataset separately, a more nuanced story begins to emerge. Figures 137 and 138 show heatmaps of features, as well as associated pathways, when looking at the top N features of each dataset separately.

**Figure 137.**

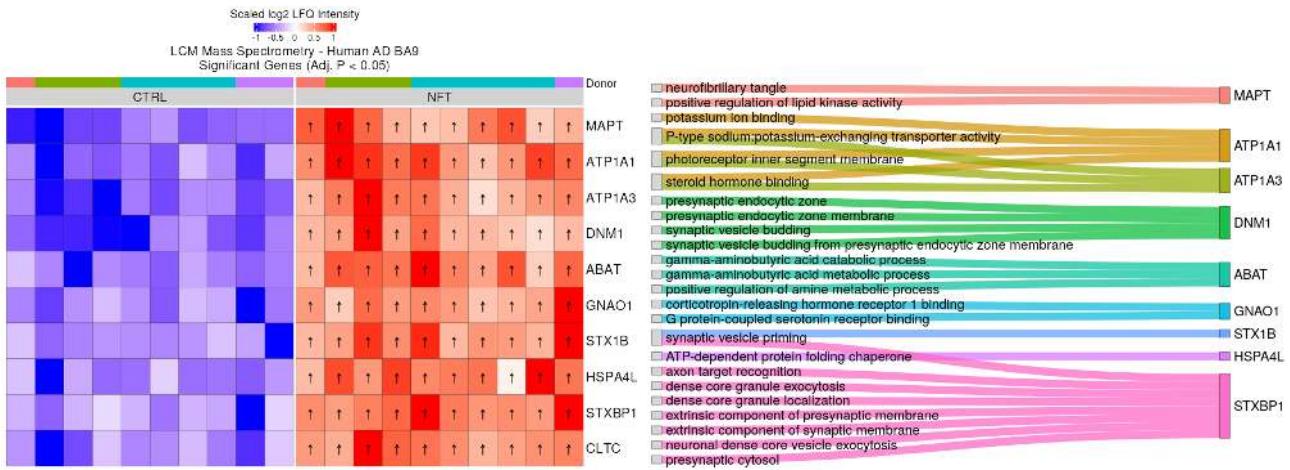


Figure 137. Top 10 proteins DE between tangle-bearing and non-tangle-bearing neurons sorted by adjusted p-value and with an adjusted p-value cutoff of < 0.05. Also shown are DE pathways (adjusted p-value < 0.05) tested using the same null hypothesis testing framework. The feature heatmap is scaled per-row between -1 and 1. While the pathways are visualised as a Sankey diagram showing the membership of the top 10 DE proteins within each pathway.

**Figure 138.**

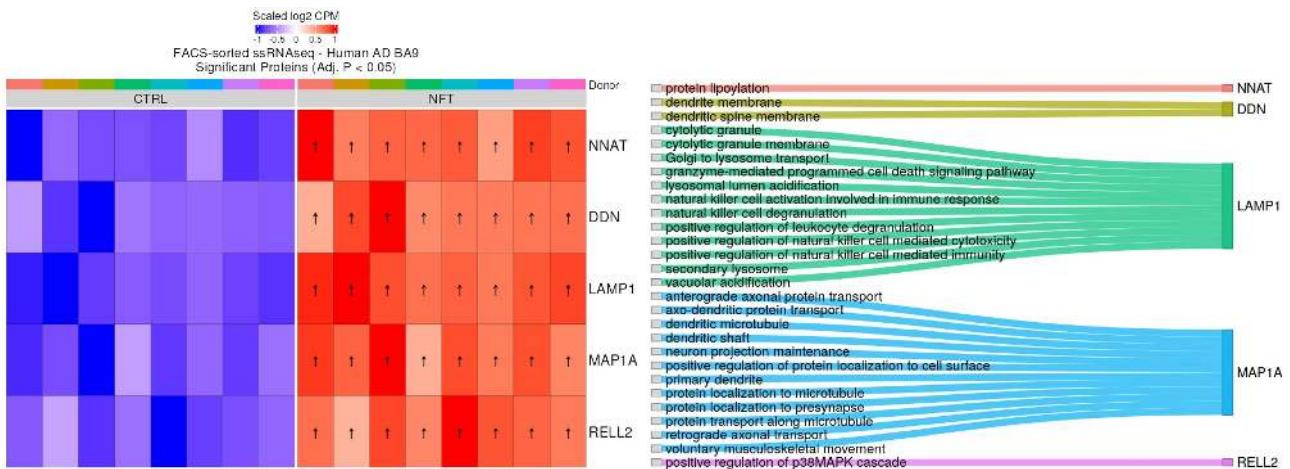


Figure 138. The same figure as Figure 137, however using the FACS ssRNAseq dataset. Only the top 5 genes are shown to allow better readability of the Sankey diagram.

It can be seen that when sorting the LCM Mass Spec dataset on its own, the results align well with sorting features of both datasets together. MAPT remains the top hit and its relevance to the dataset is further confirmed with the DE pathway “neurofibrillary tangle”, describing the neurofibrillary tangle-bearing neurons that were captured in the experiment. The rest of the top hits appear to be members of several distinct mechanisms. ATP1A1 and ATP1A3 are likely closely co-regulated and part of the same class of

sodium:potassium transporters, with additional implicated roles in steroid hormone binding. DMN1 is shown to be part of synaptic pathways such as “synaptic vesicle budding”, alongside STXBP1, though with seemingly more impact on endocytosis-related sub-functions. As impairment of synaptic pathways are highly implicated in AD (Dorostkar et al., 2015), and changes in these pathways provide additional evidence of capture of NFT-positive neurons. The remaining proteins, ABAT, GNAO1, and HSPA4L seem to be related to GABAergic pathways, hormone binding, and protein folding respectively, mechanisms that are less clear in the context of AD but may shed further insight in the pathophysiology of the disease.

In contrast to the heatmaps in Figures 135 and 136, which uses the intersection of features between the two datasets, when analysing the FACS ssRNAseq dataset on its own, it does not feature *MAPT* as a top hit, instead being supplanted by *NNAT*. Curiously, *NNAT* did not have many associated pathways called DE, with the only one being “protein lipoylation”, a post-translational modification known to be a key player in cell death (C.-H. Lin et al., 2024). Interestingly, several other genes support the theme of cell stress, with *LAMP1* being associated with changes in several inflammation related pathways and *RELL2* being associated with the p38MAPK cascade, a well-known coordinator of stress response (Canovas & Nebreda, 2021). Again, while microtubule associated protein tau (*MAPT*) is not in the top 5, *MAP1A* appears alongside a number of microtubule related pathways, reflecting that the strongest modulations in the microtubule space, at least in terms of gene expression, is capitulated more by players outside of *MAPT*. *DDN* is the remaining gene in the top 5, and appears to play a role in dendritic spines, which are notably pruned in AD (Dorostkar et al., 2015).

### 6.3 Network Hub Analysis

As expressed in the previous section, approaches that use the top N features have pitfalls that are less than ideal for comprehensively summarising changes in datasets with thousands of positive results. Therefore, a bespoke network analysis method was developed, as described in Section 5.2, for prioritising pathways and features that are changed in tangle-bearing neurons between the LCM Mass Spec and FACS ssRNASeq datasets. Figure 139 shows the final network for upregulated features in the two datasets.

**Figure 139.**

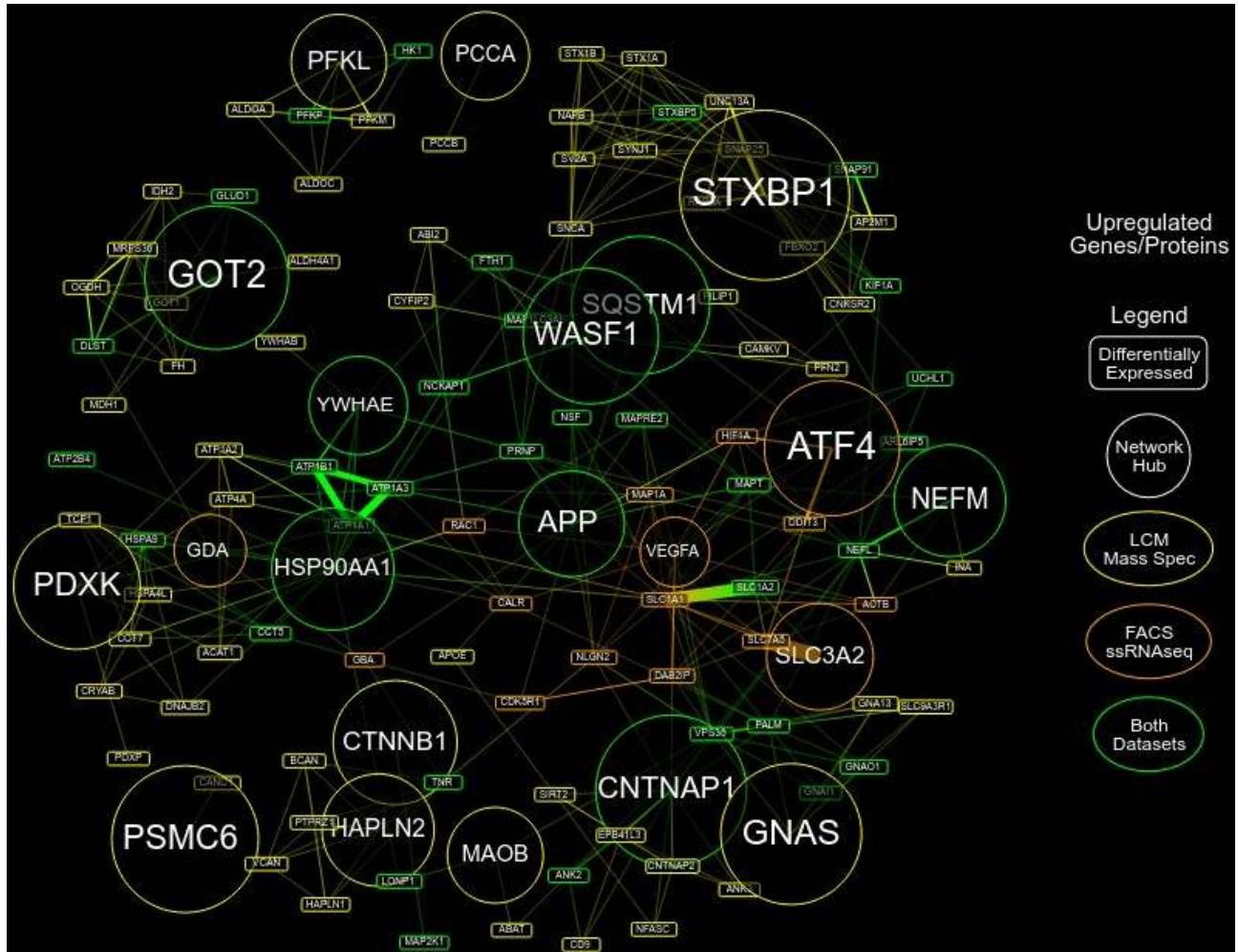


Figure 139. Upregulated features in tangle-bearing neurons in the LCM Mass Spec and FACS ssRNASeq datasets prioritised using a bespoke network analysis method (see Section 5.2). Only DE features (adjusted p-value < 0.05) are shown. Features designated as hubs are encircled. Hubs were called on the basis of mixed criteria related to the proportion of associated pathways called DE, and sized proportional to this score. This three-part criteria is described in detail in Section 5.2 and in brief equally weights the adjusted p-values of DEGs, the number of enriched gene sets associated with a DEG, and number of unique features associated with a DEG on the basis of shared enriched gene

sets. Colours indicate dataset in which a feature is DE and edges connecting nodes are sized proportional to number of shared DE pathways between features.

The network, which was generated using stringent parameters to prune the network from the large number of DE features, showcase a handful of hub genes alongside features connected on the basis of shared DE pathways. The hub genes have little overlap with the selected features of the previous section, only GOT2 and STXBP1, though many features appear as non-hubs in the network, such as MAPT. 8 hubs are DE in both datasets, while 9 are unique to the LCM Mass Spec and 4 are unique to the FACS ssRNAseq. Almost all hubs are contiguously connected to the largest unbroken network (the main network), except for 4 LCM Mass Spec hubs – PFKL, PSMC6, PCCA, and CTNNB1.

Interpretation of this network is best explored by interactive exploration of select features and hubs. Although not a hub, the product of MAPT is the protein that is pathologically aggregated in AD and the top shared feature between datasets in terms of adjusted p-value. Therefore MAPT was used as a starting point for exploring the network. Figure 140 shows the output of zooming in, clicking on, and hovering the cursor over MAPT in the online viewer for the analysis, revealing key details of its context in the network.

**Figure 140.**

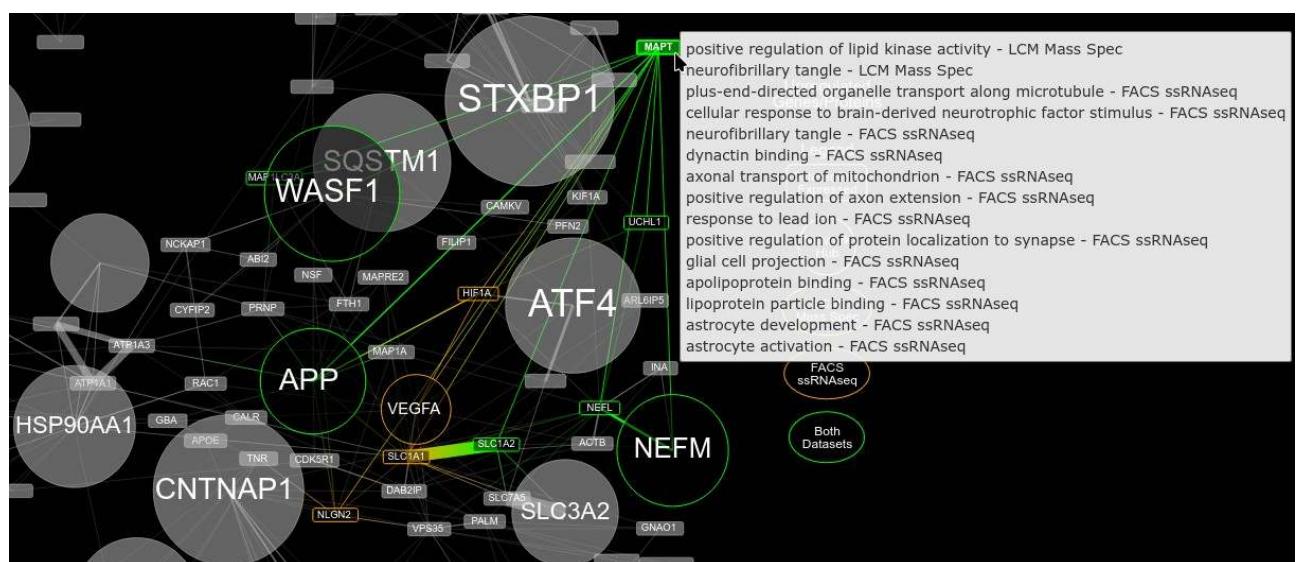


Figure 140. A view of the network analysis of upregulated features from the online viewer. The network is zoomed in and MAPT is selected, highlighting all direct connections in terms of features with shared DE pathways. Features in greyscale but labelled are second degree connections – not directly connected with MAPT but connected with a MAPT-connected feature. MAPT is also moused-over, showing a pop-up window of DE pathways containing MAPT in each dataset. Note that the layout of some features were manually moved to allow all to be visible in the zoomed in view.

As shown in the previous section, MAPT is DE in both datasets, as indicated by green colour. The tooltip showing DE pathways for MAPT confirms its relationship to neurofibrillary tangles in the context of the analysis, alongside other potentially relevant pathways related to microtubule function, lipid processing, axonal and synaptic functions, and apolipoprotein binding, a well-studied risk factor in AD. Though this study is in neurons, the transcriptomics dataset also shows several pathways that may impact glial response, another stereotypical feature of AD. Regarding other features, MAPT has first and or second degree connections with many hubs, notably APP, HSP90AA1, NEFM, and SQSTM1, all of which are highly implicated for involvement in AD.

To demonstrate the properties of a hub gene, the same procedure is applied to APP in Figure 141, a feature of high interest in AD. Similarly to the highlighting of neurofibrillary tangle pathways when looking at MAPT, the relationship of APP to the current study is confirmed with the top hit of “positive regulation of amyloid fibril formation” in the LCM Mass Spec dataset and “cellular response to amyloid-beta” in the FACS ssRNAseq dataset. A large number of additional pathways are linked with the transcriptomics dataset, covering cellular replication and reproduction, synaptic pathways, metabolism, and glial response. APP is a highly studied feature that's been shown to touch a wide variety of biological functions. As previously mentioned, the designation of APP as a hub warrants caution as the large number its large number of annotations by the GO consortium introduces bias. However, in the hub selection process I attempt to correct for this, ensuring that a hub has a large proportion of its potential pathways called DE and that the pathways have little duplication of features between them. Like MAPT, APP is directly and indirectly connected with many other hubs, as well as other well studied features in AD such as MAPT itself, APOE, MAP1A, and PRNP.

**Figure 141.**

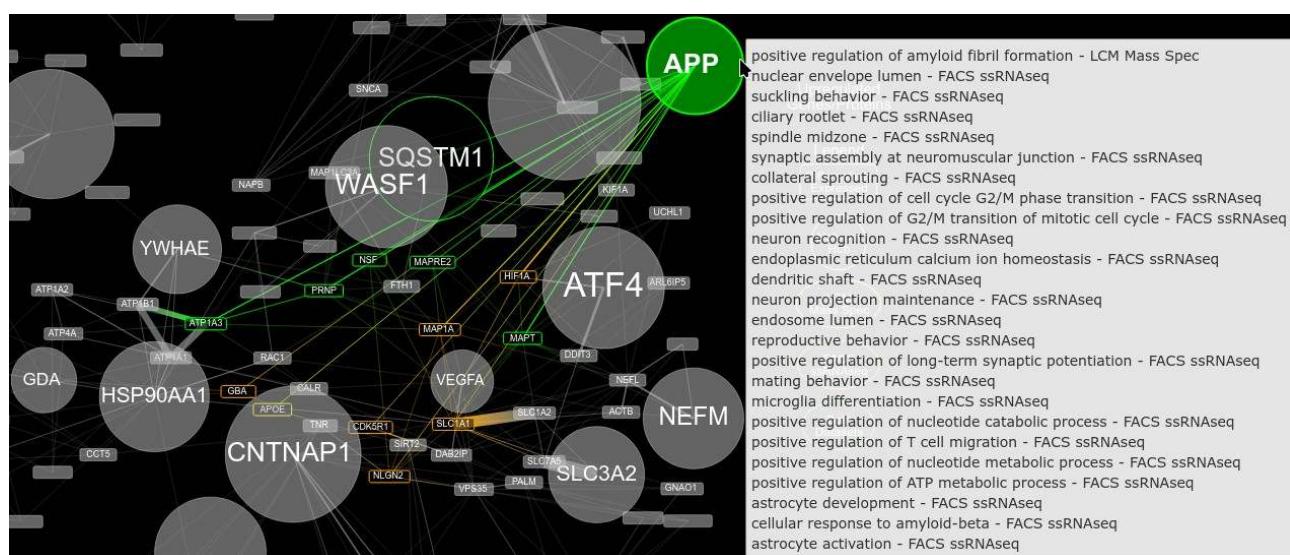


Figure 141. A zoomed in view and highlighted properties of APP, performed using a similar procedure as Figure 140.

Another type of network figure was created for the purpose of inspecting the sharing of pathways between hubs and datasets more finely. The final network of this kind for upregulated features in tangle-bearing vs. non-tangle-bearing neurons is shown in Figure 142. The network is directly derived from the overall network shown in Figure 139 and contains a subset of its hubs.

**Figure 142.**

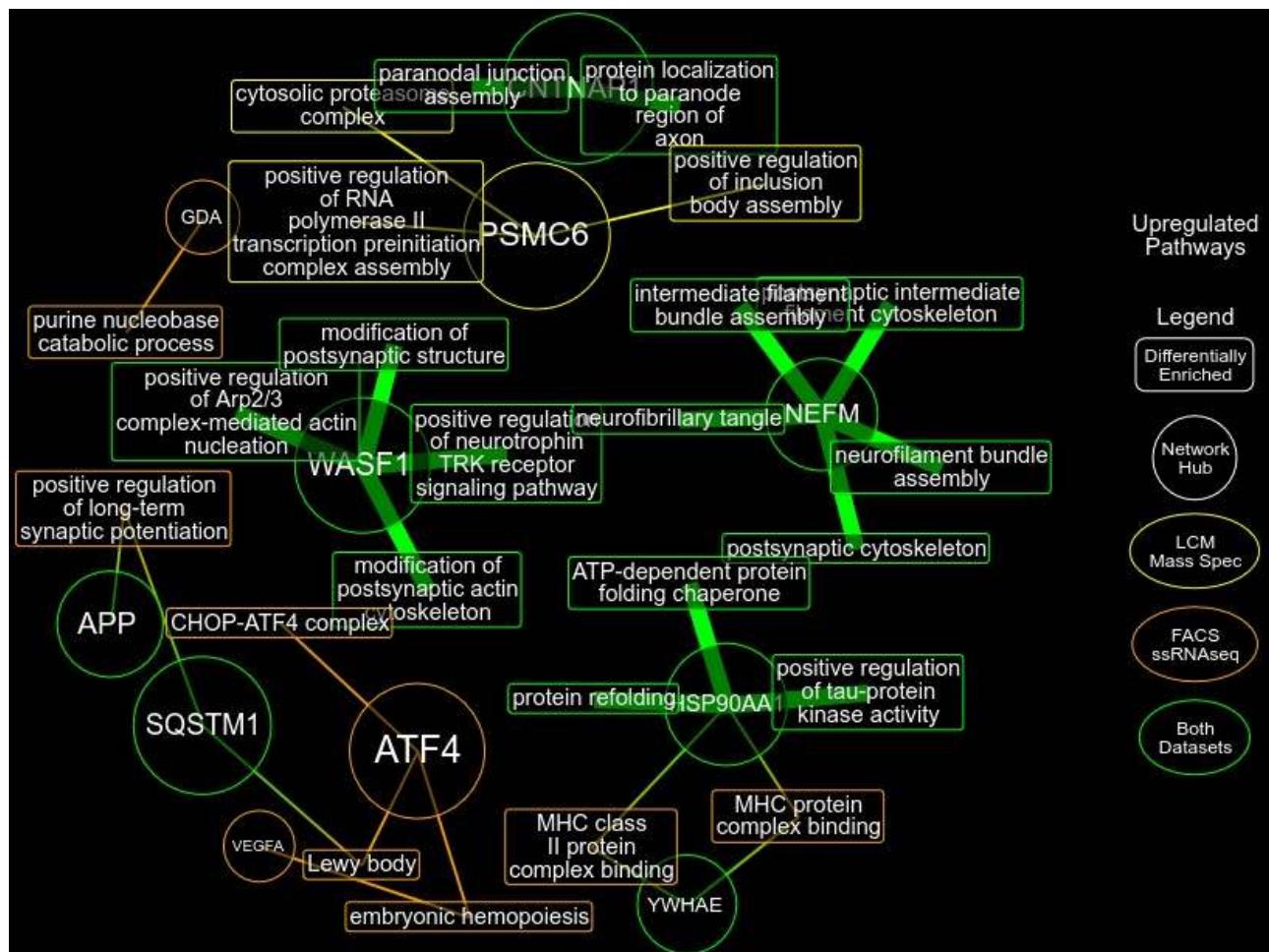


Figure 142. A variant of the network plot that highlights shared pathways between hubs and datasets. The figure is coupled with information from the overall network in Figure 139 and uses a subset of hubs designated by that network.

The network in Figure 142 provides insight in two ways. Firstly, it shows the specific DE pathways shared between datasets in the hubs. For example, adding additional confidence to NEFM as a hub, is the encircling neurofibrillary tangle and cytoskeletal related pathways, all of which are coloured green and shown with enlarged edges indicating that the pathways are DE in both datasets. HSP90AA1 shows this as well in terms of protein folding pathways and a tau-protein kinase pathway. And the pathways surrounding WASF1 are indicative of synaptic/cytoskeletal pathways, and Arp2/3 and TRK signaling pathways. This information is not only useful for finding shared pathways

between datasets but also proves effective to pruning down the mechanistic relevance of each hub in the context of disease, out of the many pathways most features are typically involved in. The second insight gleamed from this network are DE pathways shared between hubs rather than datasets. For instance, a relationship that APP seems to share with SQSTM1 is the upregulation of the “positive regulation of long-term synaptic potentiation” pathway. However, as indicated by the pathway’s orange colour and smaller edge size, this pathway is DE only in the FACS ssRNASeq dataset, though APP and SQSTM1 are upregulated in both datasets. This information is of limited utility in this particular study however, as no shared DE pathways in both datasets could be found connecting those hubs.

The procedure described above was also performed for downregulated features and pathways, as shown in Figures 143 and 144. In the case of this study, far fewer hubs could be found in either dataset. This is further exacerbated by the fact that far fewer downregulated features were called DE in the LCM Mass Spec dataset compared to upregulated features (Figure 133). Though network parameters were modified in an attempt to make network analysis of downregulated features and pathways as informative as the upregulated ones (described in detail in Section 5.2), no shared hubs and only a single shared feature, DDX1, was generated by the network. It can therefore be inferred that in comparison to upregulation, far fewer downregulated features and pathways are shared between the LCM Mass Spec and FACS ssRNASeq datasets. Furthermore, feature connectivity even within the distinct hubs, is far more limited. This additionally suggests far fewer “master regulators” of downregulated mechanisms in tangle-bearing neurons as fewer features likely co-regulate with one another. Finally, no hubs nor features comprise those traditionally associated with AD, aside from the gene *MEF2C*. Interestingly, it has been reported that *MEF2C* transcriptional upregulation in human post-mortem tissue is associated with resilience to neurodegeneration (Barker et al., 2021). This supports the direction of change in this study, where *MEF2C* is downregulated in tangle-bearing neurons, suggesting loss of resiliency to pathology.

**Figure 143.**

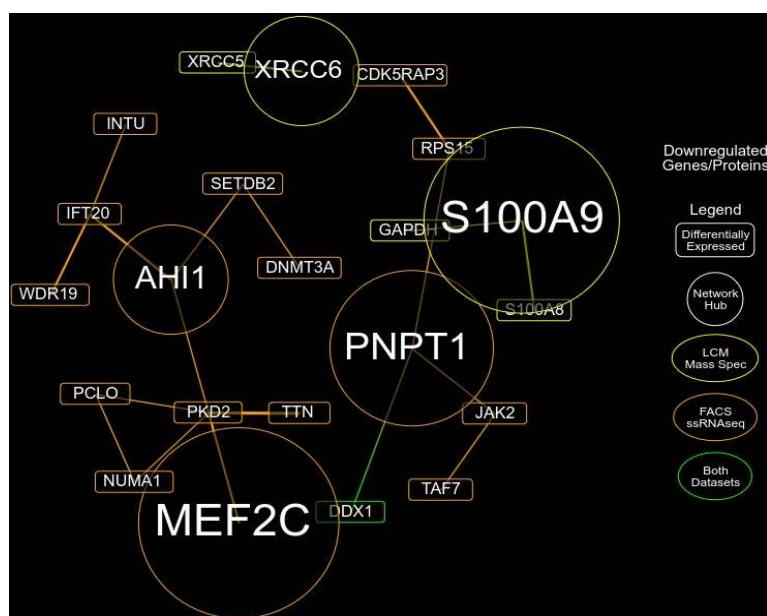


Figure 143. Network plot showing hub and feature connectivity, similar to Figure 139 but for those downregulated in tangle-bearing vs. non-tangle-bearing neurons. Some parameters were modified to account for the far fewer downregulated features and pathways (described in Section 5.2).

Figure 144 shows shared downregulated pathways between hubs and datasets, similar to Figure 142. Note that this network also has modified parameters, such that the criteria for showing a pathway is looser. This resulted in a network where a large number of DE pathways are shown as associated with a few hub features. It is possible to achieve a similar result when only looking at upregulation in this dataset, but the figure would be overcrowded and unreadable. Instead it is often more informative to produce figures balanced between shown hubs and pathways, but in the case when analysing downregulation, I was unable to achieve this without producing a very small network of lesser value. So Figure 144 is representative of this compromise and it was generally decided that downregulation would be out of scope for the present analysis due to the greater difficulty in interpreting it. Indeed, the logical interpretation of this figure is that while a large number of pathways are DE for some of hubs, the features comprising the pathways are not well represented by the pathways. In other words, many of the features are possibly ubiquitous and related to many mechanisms outside of the DE pathways. In practice, this may make them poor targets for translational medicine, as modulation of a highly ubiquitous gene or protein may have undesirable side effects not related to the pathophysiology of the disease. Further evidence of this stems from inspection the pathways of Figure 144, many which are related to development and hard to interpret in the context of AD. The few that appear more directly relevant include apoptotic and inflammatory processes associated with *MEF2C* and mitochondrial pathways with *PNPT1*.

**Figure 144.**

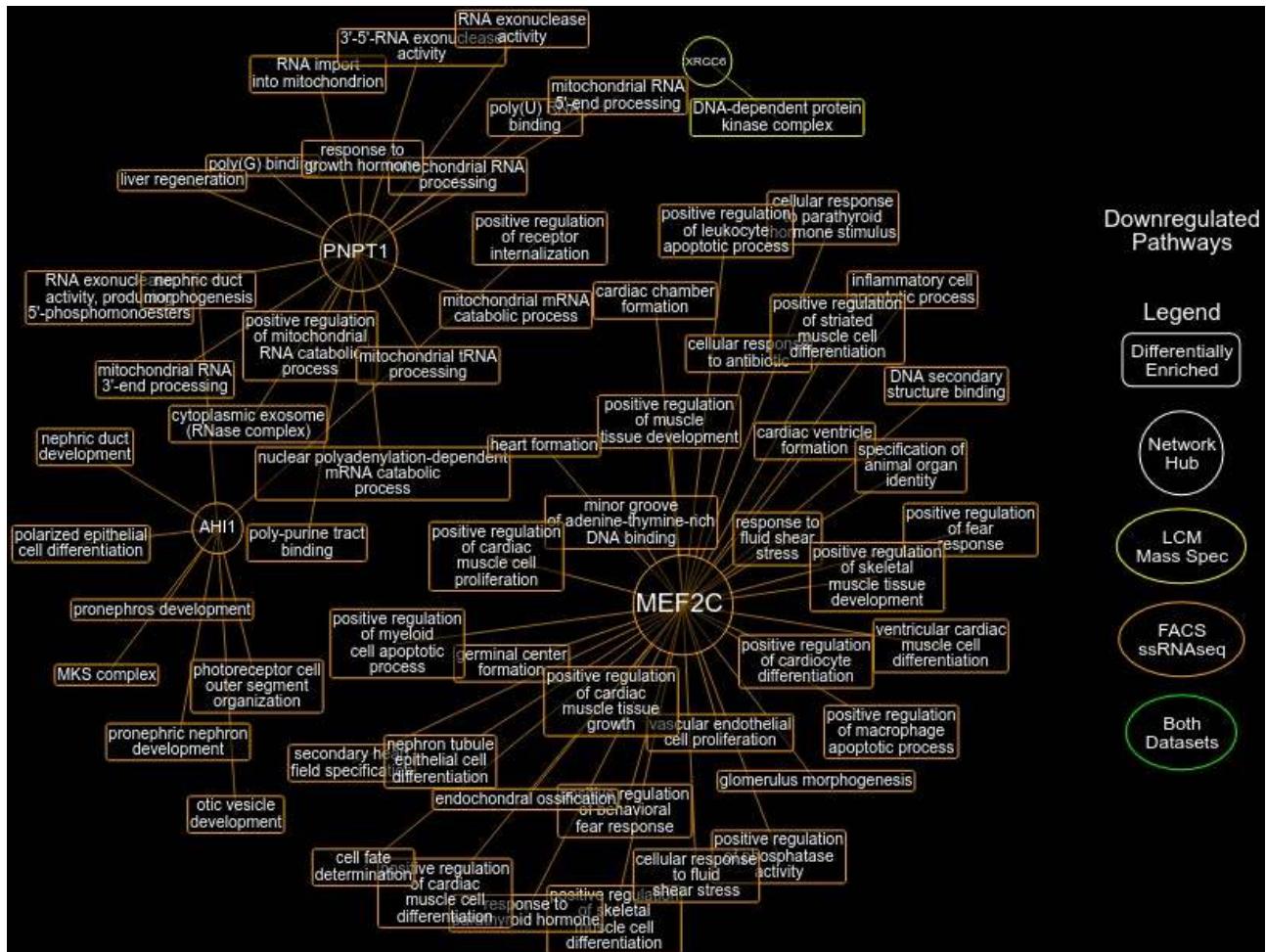


Figure 144. Network plot showing sharing of pathways between hubs and datasets, similar to Figure 142 but for those downregulated in tangle-bearing vs. non-tangle-bearing neurons. Parameters were adjusted to account for the far fewer downregulated features and pathways.

As a final sanity check, scatter plots of all features in the LCM Mass Spec and FACS ssRNASeq datasets were plotted (Figure 145), with corresponding expression values in tangle-bearing and non-tangle-bearing neurons in each axis. Each of the shared hubs shown in Figure 139 are also highlighted as a blue point. This figure can be further explored in the online viewer, where mousing-over on a point reveals the feature name and expression values. The scatter plots shows an overall strong correlation between the features of tangle-bearing and non-tangle-bearing neurons, as indicated by the close fit of genes/proteins along a diagonal line drawn across the axes. Those that deviate substantially from the fitted line were likely called DE during statistical testing. This is difficult to assess when plotting all features, but two properties suggest that this is the case. Recall that in the LCM Mass Spec dataset that far more of the DE proteins were upregulated rather than downregulated; this is reflected in Figure 145, where more of the proteins deviating from the diagonal are above the diagonal. Secondly, and perhaps more

importantly, all of the shared network hubs, shown as blue points, deviate from the diagonal in a visually apparent way. What's more is that they are all in the upregulated direction in both datasets, providing confirmation of the validity of the hub section approach. In summary, the data presented in this section describe a sophisticated approach to describing the changes taking place in the dataset and appears to align correctly with more simplified views like that of the scatter plots in Figure 145.

**Figure 145.**

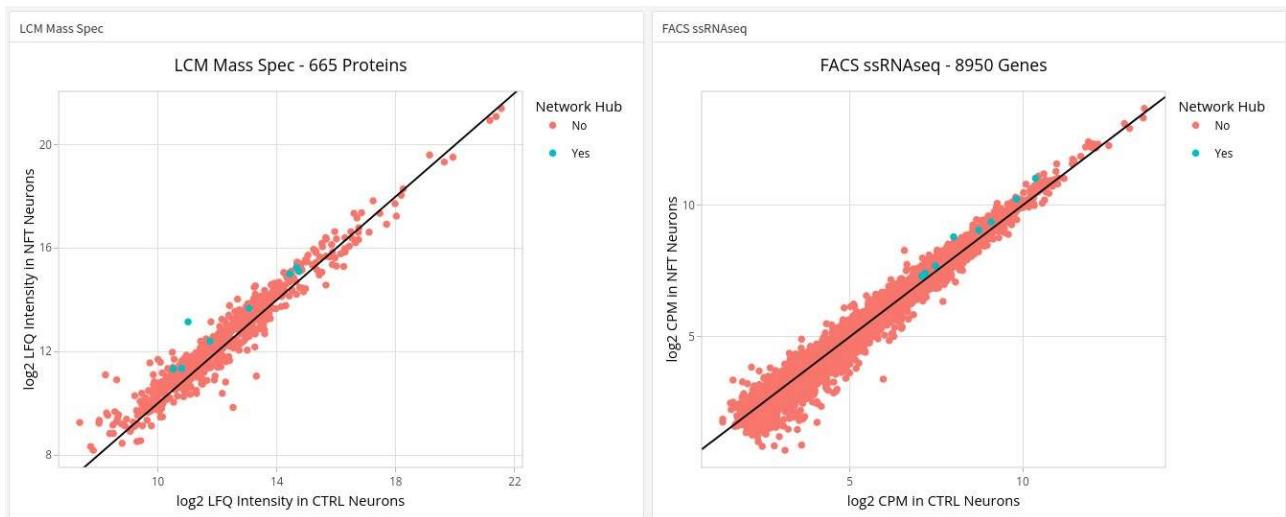


Figure 145. Scatter plot of the expression values of all features between NFT and CTRL neurons. As each dataset was comprised of multiple biological replicates, expression values were taken from the mean expression across replicates. Shared network hubs are highlighted as blue points. Information about feature name and expression values can be obtained by mousing-over a point in the online viewer.

## 6.4 Pathway Analysis of Hubs

In order to investigate further, while at the same time narrowing the scope of the analysis, the pathways and associated features of hubs were examined comprehensively. Out of the hubs highlighted in the network analysis, the focus was placed on those shared between the LCM Mass Spec and FACS ssRNAseq datasets. This therefore also restricted the analysis to upregulated pathways and features. The following sections thus focus on each of these hubs, specifically NEFM, APP, SQSTM1, YWHAE, WASF1, CNTNAP1, and GOT2. To help introduce the format of the following sections and initially analyse the data more unbiasedly, shown in Figures 146 and 147 are the top 10 DE pathways of each dataset, sorted by adjusted p-value with an adjusted p-value cutoff of  $< 0.05$ .

**Figure 146.**

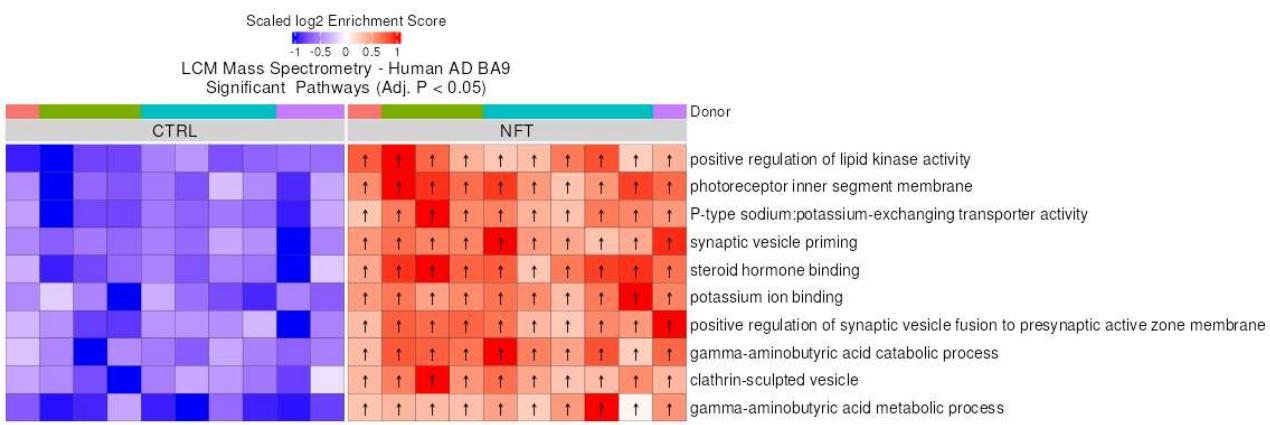


Figure 146. Heatmaps of the top 10 DE pathways comparing tangle-bearing and non-tangle-bearing neurons (labeled NFT and CTRL, respectively) in the LCM Mass Spec dataset. Cells corresponding to DE features are marked with an up or down arrow indicating up or downregulation at an adjusted p-value cutoff of  $< 0.05$ . Features are scaled from -1 to 1 within each row to highlight differences between groups. Biological replicates are indicated by the Donor annotation at the top of each heatmap.

**Figure 147.**

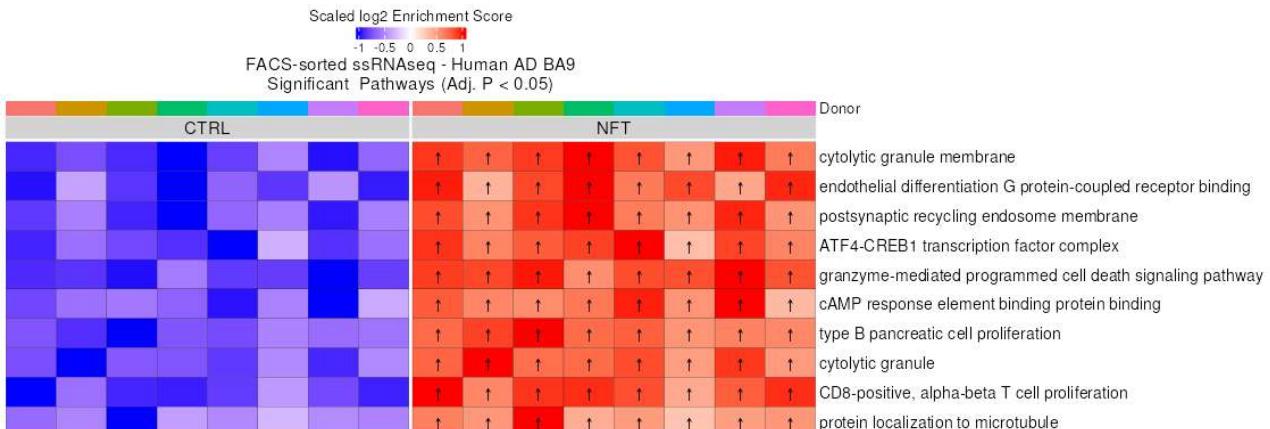


Figure 147: A similar heatmap to Figure 146, except for the FACS ssRNAseq dataset.

The pathways shown have a level of correspondence with the top DE features sorted by adjusted p-value. For instance, in the LCM Mass Spec dataset, the top pathway is “positive regulation of lipid kinase activity”, which also appears in Figure 137 as co-upregulated with the top DE protein MAPT. Indeed, plotting all proteins comprising this pathway includes MAPT, alongside the other components EEF1A2 and F2 (Figure 148). Similarly, in Figure 147, the top pathway in the FACS ssRNAseq dataset, “cytolytic granule membrane”, contains the gene *LAMP1*, which is also a top DE gene as seen in Figure 138. This provides strong real-data evidence that DE testing of pathways using GeneFunnel shows high correspondence with traditional DE testing of genes and proteins.

In general, the top 10 pathways of the LCM Mass Spec dataset covers synapse-associated pathways most frequently, in addition to lipid kinase activity, sodium:potassium transporters, hormonal functions, and GABAergic pathways, all of which are also reflected by the gene-level analysis in Figure 137. The FACS ssRNAseq data shows comparatively less correspondence with the gene-level analysis, covering cytosolic granules, a microtubule pathway, and cell death but not dendritic pathways nor p38MAPK pathways. In place are pathways related to ATF4-CREB1 transcription factors, cAMP response, and several immune cell pathways.

**Figure 148.**

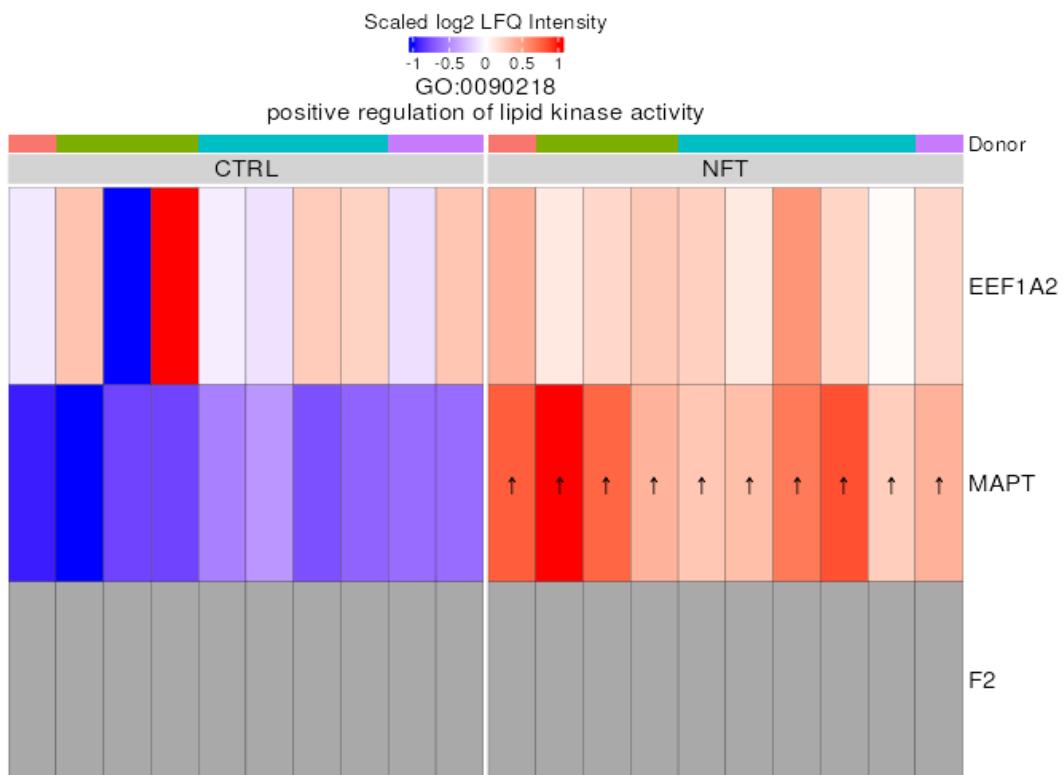


Figure 148. Heatmap of proteins comprising the “positive regulation of lipid kinase activity” pathway in the LCM Mass Spec dataset. The heatmap properties are similar to that of Figure 135, with significantly DE proteins defined with adjusted p-value < 0.05. Proteins absent from the dataset or too lowly abundant for analysis are shown in grey.

**Figure 149.**

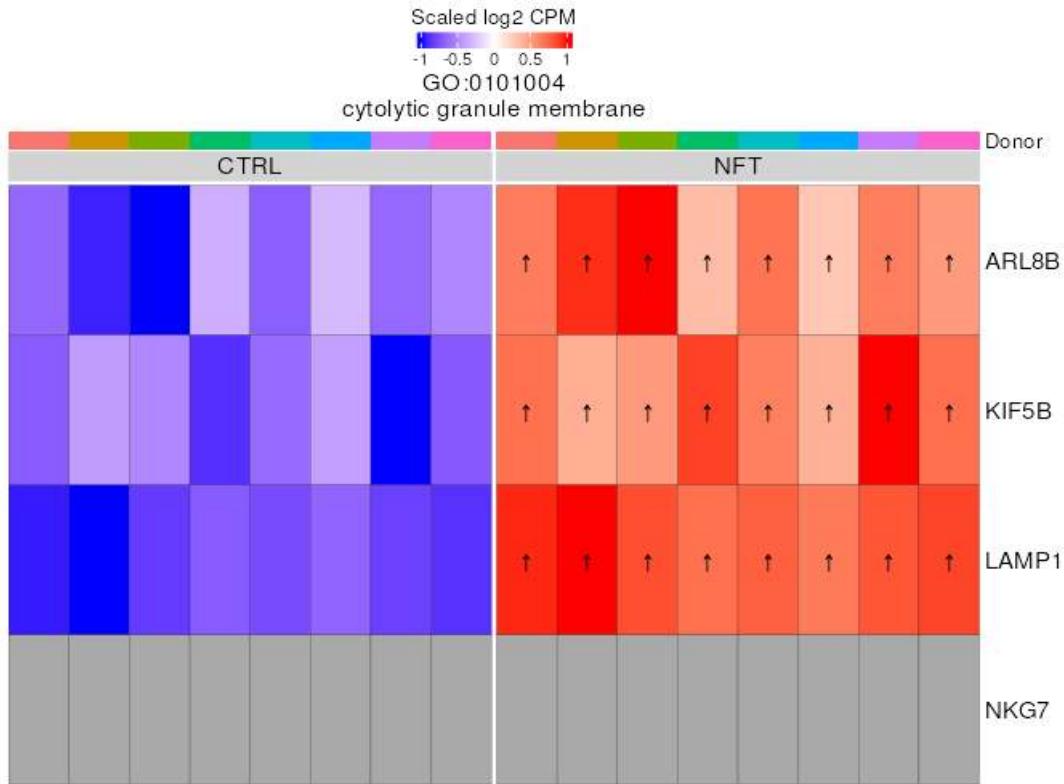


Figure 149. A heatmap similar to Figure 148, however showing the “cytolytic granule membrane” pathway in the FACS ssRNASeq dataset.

As mentioned earlier, top N approaches to analysis come with caveats that apply generally, regardless of if the targets are individual genes/proteins or if they are pathways. In the case of the pathway-level analysis performed here, there is a tendency of overlap between functionally related pathways. This results in the duplication of near-identical terms, for instance, the multiple GABAergic signaling pathways observed in Figure 146. There are several ways to address this. One way is to discard pathways on the basis of overlap with other pathways, but this presents difficulties in deciding which of the pathways to keep, i.e. the more general or more specific ones. Furthermore, pathway scoring can be very sensitive to this in the case of pivotal features that are absent or present between otherwise similar pathways. For this reason, I do not implement filtering of pathways based on overlap. Instead, it was the motivation for developing a hub analysis based on networks of the association between features and pathways. In theory, a well implemented approach to this may better highlight the diversity of changes that might be taking place in the dataset. This also massively increases complexity of the analysis and the introduction of undesirable assumptions, which is why this analysis also features simplified exploratory analysis based on simple procedures like top N, as seen here. But a network based approach with carefully defined metrics has the potential to converge towards full dataset characterisation while maximising concision of the results. This is the ultimate intent of this analysis and the next sections dive in-depth into each hub feature that was selected for this purpose, narrowing scope to only shared changes between the datasets.

#### 6.4.1 NEFM

Neurofilament medium polypeptide (NEFM) plays a fundamental role in axonal structure, cytoskeletal stability, and intracellular transport, acting as a key component of the neurofilament network that supports neuronal integrity (A. Yuan & Nixon, 2021). NEFM is upregulated in both the LCM Mass Spec and FACS ssRNAseq datasets, alongside another neurofilament, neurofilament light polypeptide (NEFL). These proteins work in conjunction with microtubules, regulated by microtubule-associated protein tau (MAPT), which itself was found to be upregulated in NFT-bearing neurons on both the transcriptomic and proteomic level. In Alzheimer's Disease, and particularly in tangle-bearing neurons, disruptions in tau homeostasis, neurofilament dysregulation, and cytoskeletal instability contribute to progressive neuronal dysfunction and degeneration.

**Figure 150.**

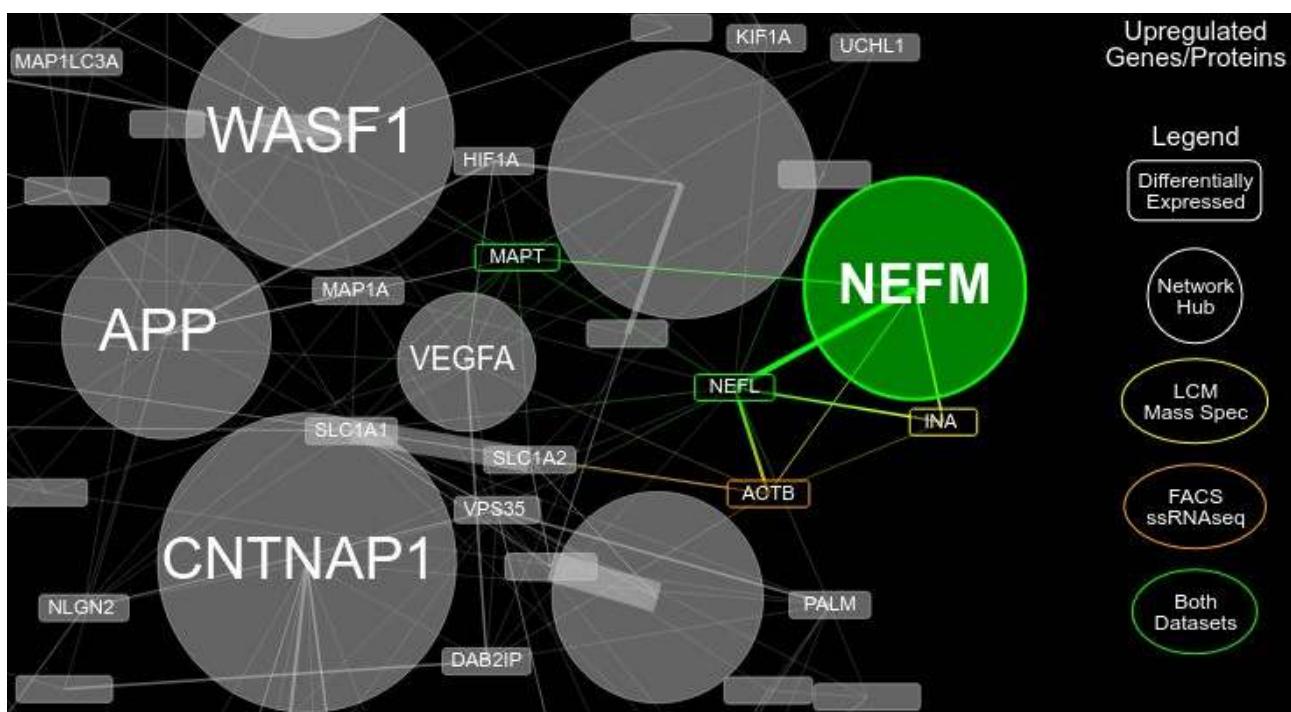


Figure 150. Feature network of NEFM. Figure format previously described in Figure 150.

Tau is essential for stabilising microtubules, ensuring the proper transport of organelles, vesicles, and signalling molecules throughout the neuron (Stamer et al., 2002). Under physiological conditions, tau binds to microtubules to regulate their assembly and disassembly (Barbier et al., 2019), complementing the structural role of neurofilaments such as NEFM and NEFL. However, in AD, tau undergoes hyperphosphorylation, leading to its detachment from microtubules and subsequent aggregation into neurofibrillary tangles. This loss of microtubule stability disrupts axonal transport, shifting the burden of structural support onto neurofilaments like NEFM (Yadav et al., 2016).

NEFL is the smallest neurofilament subunit, critical for initiating neurofilament assembly and regulating the fidelity of axons. Together with NEFM and NEFH, it forms

heteropolymers that provide structural stability to axons (Campos-Melo et al., 2018). In tangle-bearing neurons, however, neurofilament homeostasis is disrupted, potentially leading to mislocalisation, accumulation, and altered phosphorylation patterns of NEFM and NEFL (J. Wang et al., 2001). This neurofilament pathology has also been observed in dystrophic neurites surrounding amyloid plaques and within degenerating axons (Dickson et al., 1999), where aberrant neurofilament aggregation contributes to neuronal dysfunction. Interestingly, NEFH was not found to be differentially expressed in both datasets and is furthermore lowly expressed compared to the other neurofilaments (Figure 151), the complete set of which forms the differentially enriched GO gene set “neurofilament bundle assembly” (Figure 152). This may contribute to the overall pathological effect of neurofilaments in AD, by disrupting homeostasis between the trio of neurofilaments, a particular event that has been discussed in the context of neurodegenerative disease (Capano et al., 2000).

**Figure 151.**

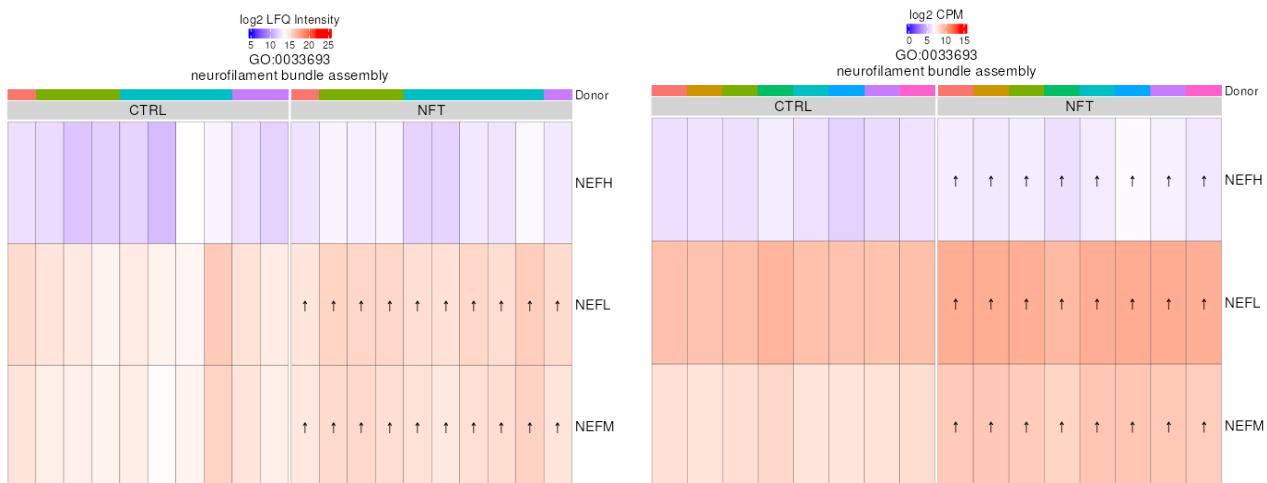


Figure 151: Heatmaps showing the unscaled expression of the trio of neurofilaments in the LCM Mass Spec (left) and FACS ssRNAseq (right) datasets. The set of these 3 features also comprise complete membership of the GO gene set (pathway) “neurofilament bundle assembly”, which is differentially enriched in both datasets. Figure format previously described in Figure 148.

Additionally, neurofilament levels in cerebrospinal fluid (CSF) and plasma have been identified as biomarkers of neurodegeneration in AD, with elevated NEFL levels correlating with axonal injury and disease progression (Giuffrè et al., 2023). Given that NEFM and NEFL expression patterns shift in response to cytoskeletal stress, their dysregulation in NFT-bearing neurons may reflect an ongoing neurodegenerative process driven by tau pathology. Closer analysis of enriched pathways associated with NEFM confirms its involvement in neurofibrillary tangles (Figures 152 to 154). This is alongside several

pathways related to neurofilaments, the cytoskeleton, as well as a pathway describing the “postsynaptic cytoskeleton”.

**Figure 152.**

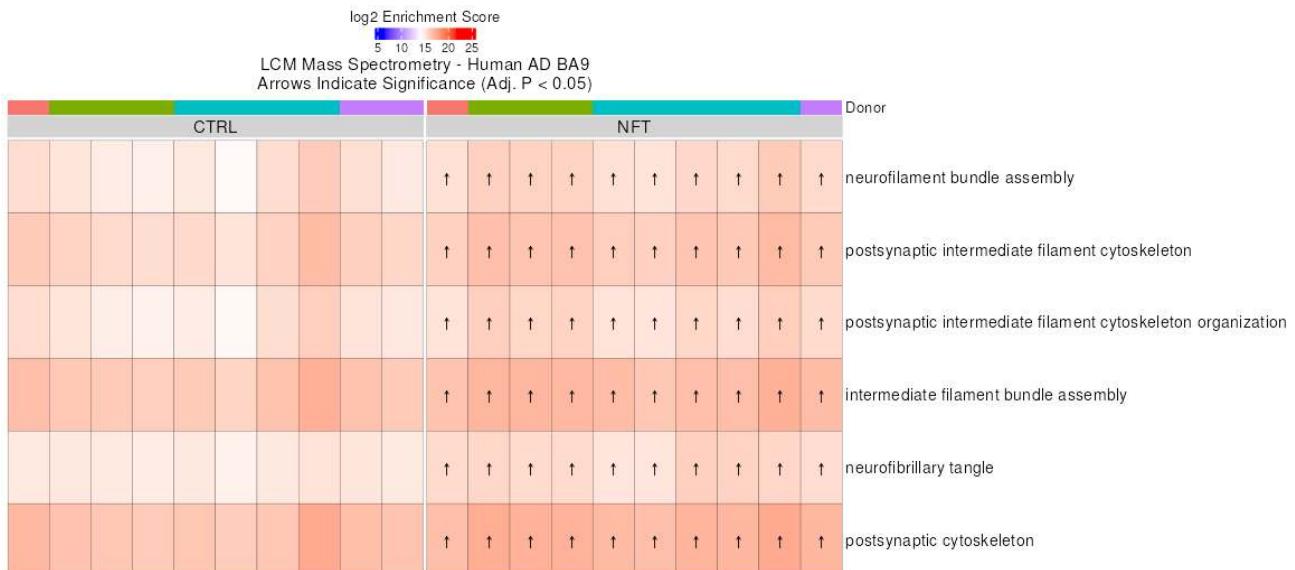


Figure 152: Heatmap showing enrichment of all differentially enriched pathways containing NEFM in the LCM Mass Spec dataset. Figure format previously described in Fig. 146.

**Figure 153.**

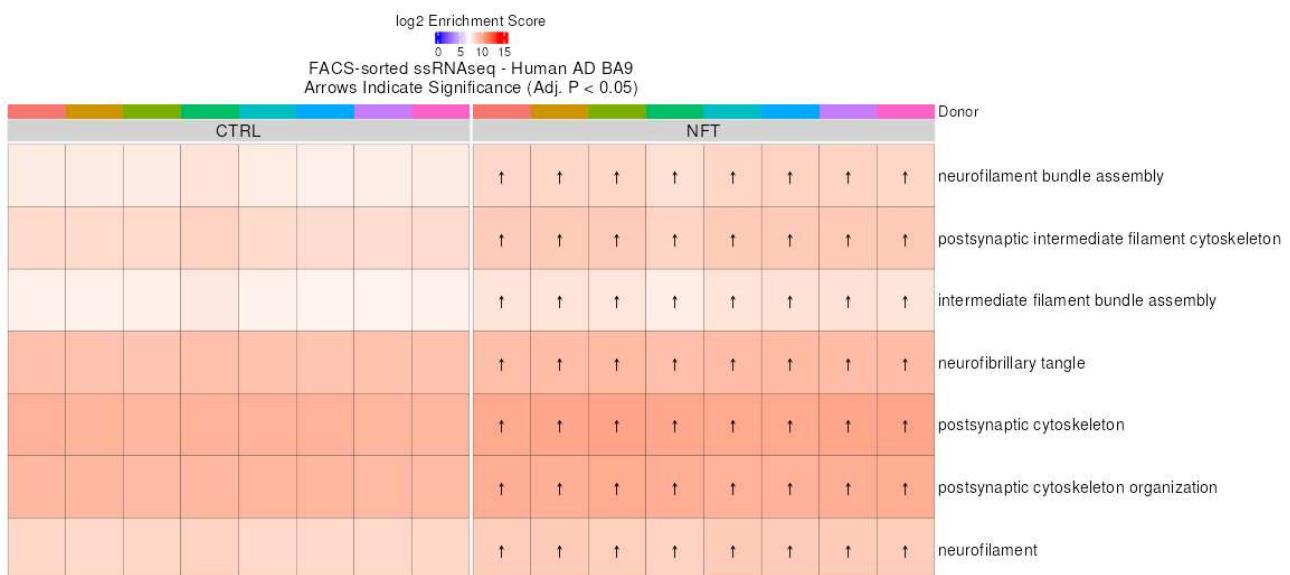


Figure 153: Heatmap showing enrichment of all differentially enriched pathways containing NEFM in the FACS ssRNAseq dataset. Figure format previously described in Fig. 146.

**Figure 154.**

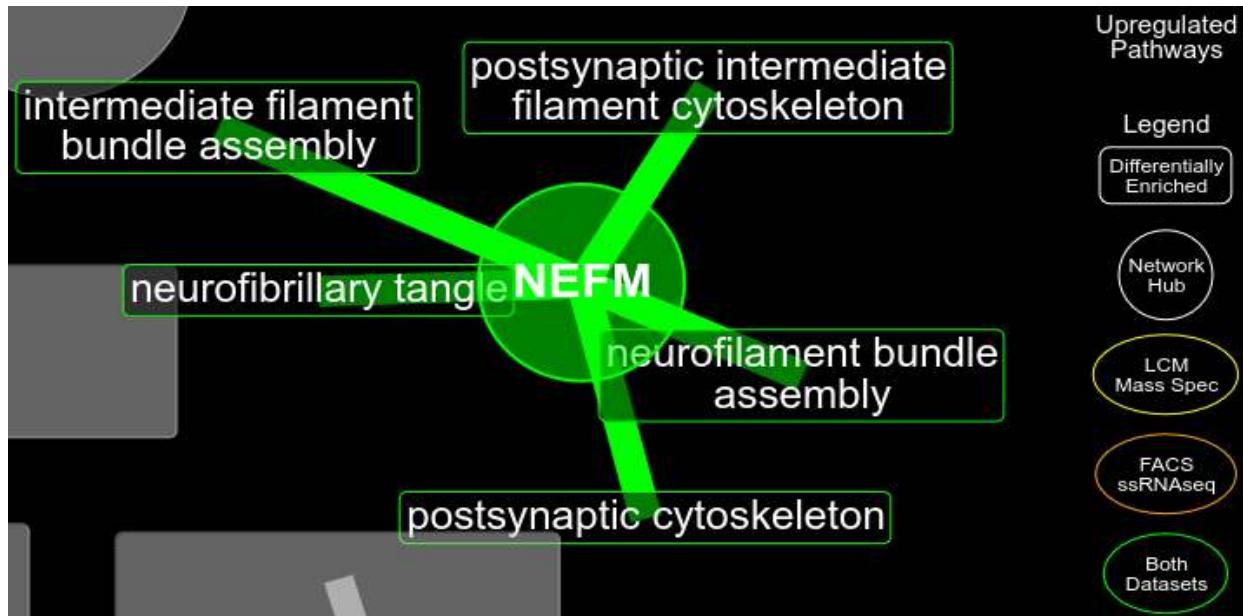


Figure 154: Pathway network of NEFM. Figure format previous described in Figure 142.

The “neurofibrillary tangle” pathway is defined by the Gene Ontology group as composed of NEFM and NEFH, though curiously not NEFL (Figure 155). This may constitute an inaccuracy on the group’s part, considering NEFL has been specifically reported to be present in the proteome of tangle-bearing neurons in AD (Hondius et al., 2021). If this is the case, it is a prime example of the caveats associated with the reliance on gene sets when performing gene set enrichment. The pathway is also defined by the inclusion of CLU and PICALM, both classic AD risk-genes more generally known to interact with A $\beta$  and clathrin-mediated endocytosis, respectively (Carrasquillo et al., 2010).

**Figure 155.**

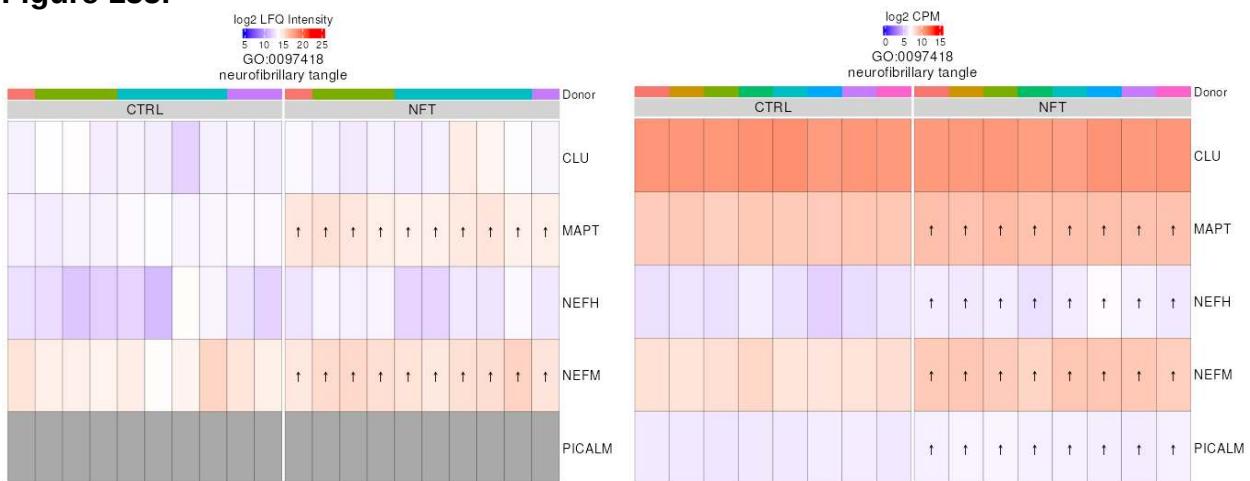


Figure 155: Heatmaps showing the unscaled expression of features within the “neurofibrillary tangle” pathway in the LCM Mass Spec (left) and FACS ssRNAseq (right) datasets. Figure format previously described in Figure 135.

NEFM is a major constituent of intermediate filament bundles, which provide structural support to axons and dendrites (A. Yuan et al., 2012). The assembly of intermediate filament bundles is crucial for maintaining neuronal integrity and resistance to mechanical stress. NEFM further plays a key role in modulating neurofilament spacing and cross-linking, ensuring proper axonal caliber and function (Ding & Kumar, 2024). However, in AD, tau hyperphosphorylation leads to microtubule destabilisation, forcing neurofilaments to bear a greater structural burden. The impairment of neurofilament bundle assembly in tangle-bearing neurons may accelerate neuronal breakdown, leading to neurodegeneration and cognitive impairment.

Beyond its roles in axonal support, NEFM is also dysregulated alongside pathways related to the postsynaptic cytoskeleton, where it helps regulate dendritic spine stability and synaptic function (A. Yuan et al., 2009). The postsynaptic cytoskeleton is essential for maintaining synaptic strength and plasticity, processes that are progressively impaired in AD. In tangle-bearing neurons, synaptic cytoskeletal components become disorganised, leading to synaptic weakening and loss (Otero-Garcia et al., 2022). NEFM dysregulation, coupled with tau pathology, likely contributes to the postsynaptic cytoskeletal instability observed in AD, further exacerbating neuronal communication deficits and cognitive decline.

**Figure 156.**

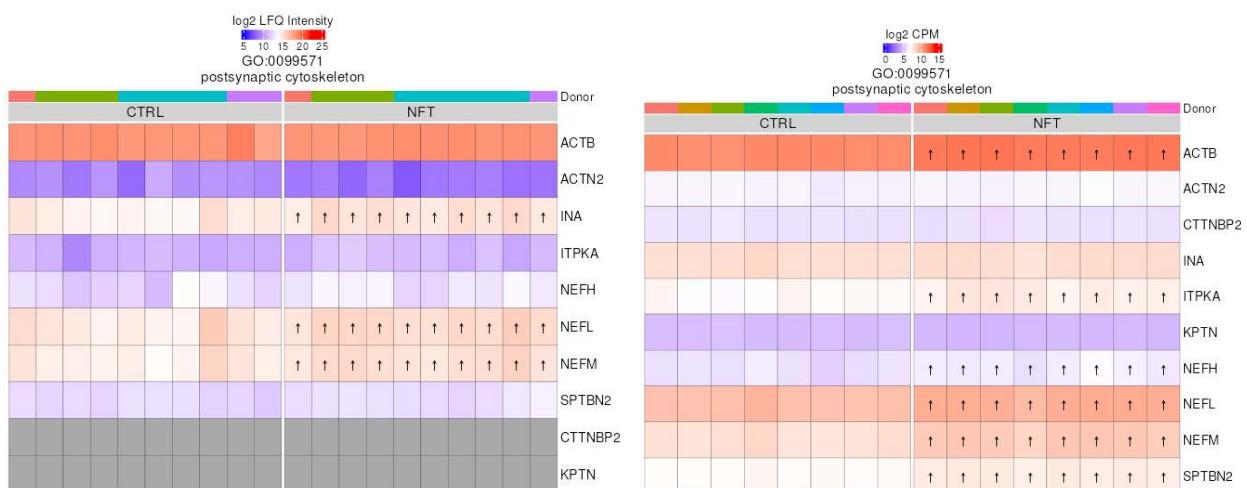


Figure 156: Heatmaps showing the unscaled expression of features within the “postsynaptic cytoskeleton” pathway in the LCM Mass Spec (left) and FACS ssRNAseq (right) datasets. Figure format previously described in Figure 135.

## 6.4.2 APP

Amyloid precursor protein (APP) plays a central role in AD pathology, where dysregulated APP processing leads to the accumulation of amyloid-beta (A $\beta$ ) peptides, which, in conjunction with tau aggregation, contribute to neuronal dysfunction and degeneration. APP's involvement extends beyond A $\beta$  production, as Figure 157 suggests it interacts with a variety of other genes and proteins, including ATP1A3, PRNP, NSF, MAPRE2, and MAPT, all of which have been implicated in neurodegenerative processes associated with NFT-bearing neurons. It also shares enriched pathways with another network hub, SQSTM1 across both datasets; this association will be discussed in the section dedicated to that hub.

**Figure 157.**

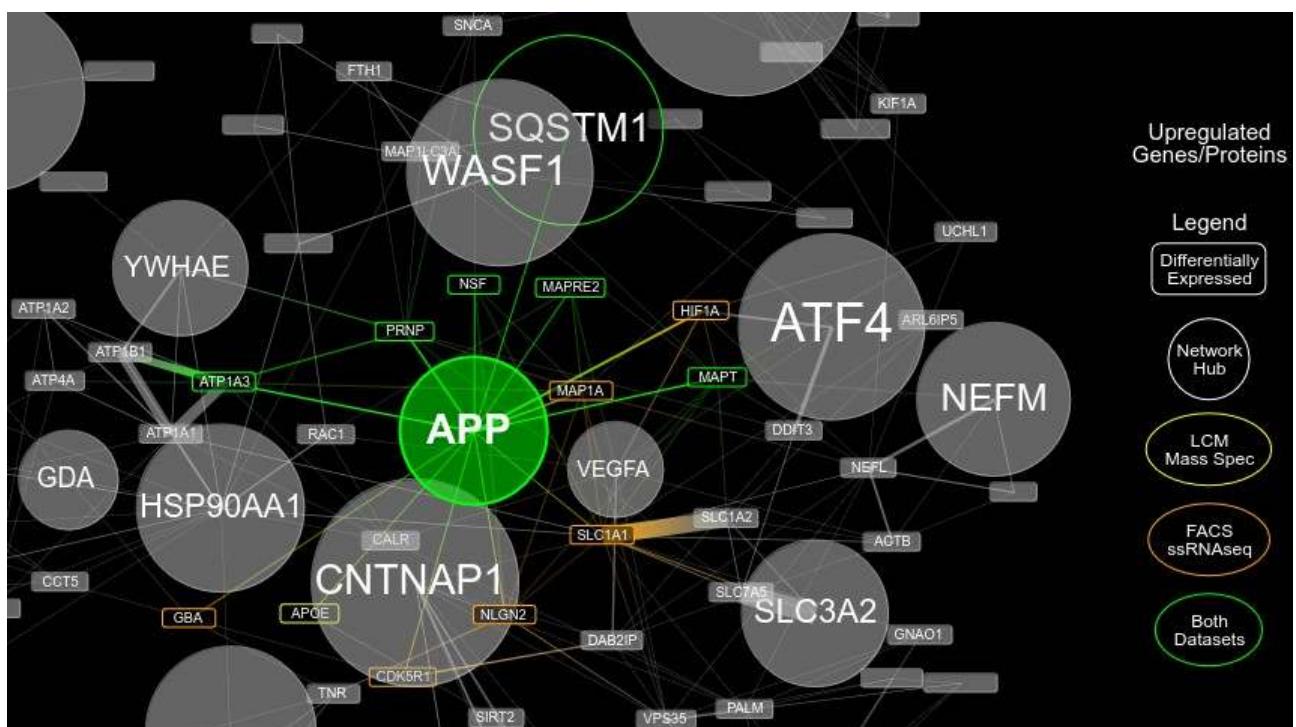


Figure 157: Feature network of APP. Figure format previously described in Figure 140.

ATP1A3 encodes the  $\alpha 3$  subunit of Na+/K+-ATPase, a critical enzyme responsible for maintaining neuronal ion homeostasis and action potential propagation. In AD, A $\beta$  and APP dysfunction have been shown to impair Na+/K+-ATPase activity, leading to altered neuronal excitability and energy metabolism (Adzhubei et al., 2022). ATP1A3 mutations are associated with neurological disorders (Vezyroglou et al., 2022), and its dysfunction in AD may contribute to neuronal hyperexcitability, impaired synaptic transmission, and eventual neurodegeneration. The connection between APP and ATP1A3 suggests that ion transport dysregulation could be a key factor in neuronal vulnerability, particularly in tangle-bearing neurons that are already under metabolic stress.

The prion protein (PRNP) has been implicated in protein misfolding diseases, including Creutzfeldt-Jakob disease and AD, and shares pathological similarities with A $\beta$  and tau aggregation (Calero et al., 2011). PRNP interacts with APP and has been proposed to act as a cellular receptor for A $\beta$  oligomers (Nygaard & Strittmatter, 2009), potentially amplifying tau pathology in NFT-bearing neurons. Dysregulated PRNP expression in AD may exacerbate synaptic dysfunction and neuronal toxicity, further accelerating the disease process. Additionally, both A $\beta$  and PRNP have been linked to oxidative stress (Castle & Gill, 2017), which is a contributing factor to neuronal damage in tangle-bearing neurons.

N-ethylmaleimide-sensitive factor (NSF) is a key regulator of SNARE-mediated vesicle trafficking and synaptic transmission (Y. Yang et al., 2018). APP interacts with vesicle transport pathways, and in AD, abnormal APP processing can disrupt synaptic vesicle cycling, leading to impaired neurotransmission and synaptic loss. Alterations in NSF function have been linked to defects in synaptic plasticity, which are exacerbated in NFT-bearing neurons where tau aggregates further impair microtubule-based transport. Given that NSF dysfunction contributes to synaptic vesicle misregulation, it is likely that APP and NSF dysregulation together accelerate synaptic failure in AD.

Microtubule-associated protein RP/EB family member 2 (MAPRE2) plays a crucial role in microtubule stabilization and intracellular transport, directly influencing cytoskeletal dynamics (McKitney et al., 2019). APP is trafficked along microtubules (T. Lin et al., 2021) and MAPRE2 dysfunction may contribute to instability in its transport, compounding the deleterious effects of APP in a disease state.

The APP-MAPT relationship is central to AD pathology. While APP cleavage generates A $\beta$  peptides that trigger neurotoxicity, tau undergoes hyperphosphorylation, forming NFTs that disrupt intracellular transport and neuronal homeostasis. Studies suggest that A $\beta$  oligomers can drive tau phosphorylation by activating kinases such as GSK3 $\beta$  and CDK5, linking APP dysfunction to NFT formation (Engmann, 2009). Furthermore, tau aggregates impair APP transport along axons (Stamer et al., 2002), potentially altering its cleavage and exacerbating A $\beta$  accumulation.

Pathway enrichment using the current methods failed to share much further light into the context of these features when considering both datasets together. While the analysis on the FACS ssRNASeq dataset found many differentially enriched APP related pathways, the LCM Mass Spec dataset could only uncover one. However, that single pathway did help provide confirmation into obvious roles of APP in Alzheimer's Disease, with the term being "positive regulation of amyloid fibril formation" (Figure 158). Furthermore, this gene set includes one of the most consistent and well-studied AD risk genes – APOE, which was found to be differentially abundant on the protein-level, though not on the gene-level (Figure 159). Note that in both datasets these features are relatively lowly expressed, suggesting that even if changes were detected, it may not be as robust as those related to NEFM, for instance.

**Figure 158.**

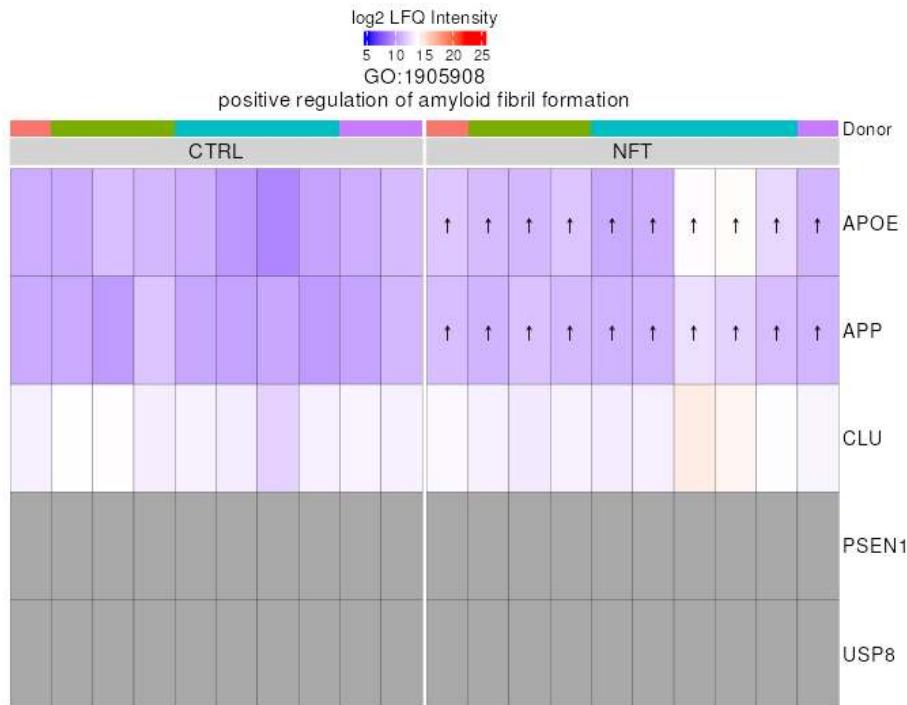


Figure 158: Heatmap showing the unscaled abundance of proteins within the “positive regulation of amyloid fibril formation” pathway in the LCM Mass Spec dataset. Figure format previously described in Figure 135.

**Figure 159.**

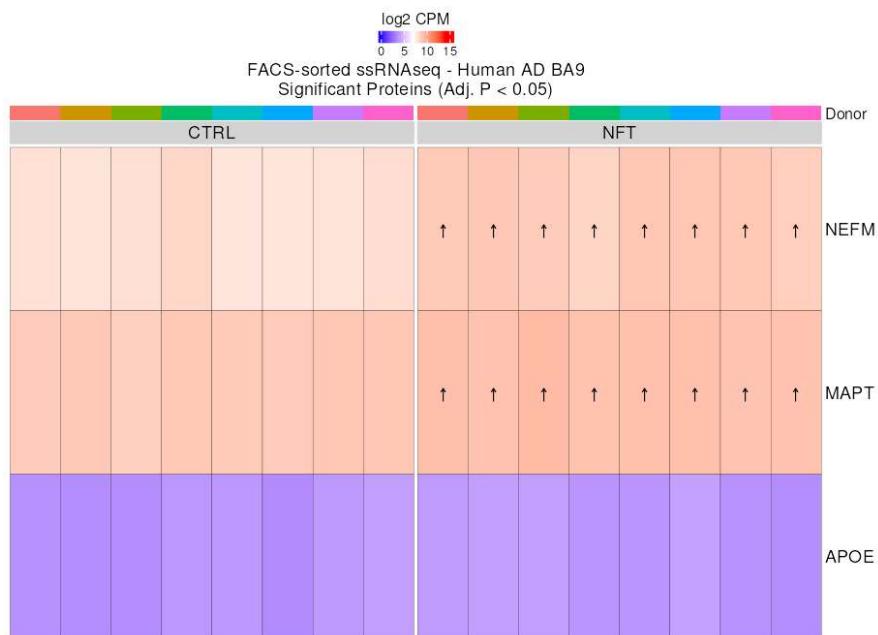


Figure 149: Heatmap showing the unscaled expression of *APOE* alongside two other more highly expressed genes in the FACS ssRNAseq data. Figure format previously described in Figure 135.

**Figure 160.**

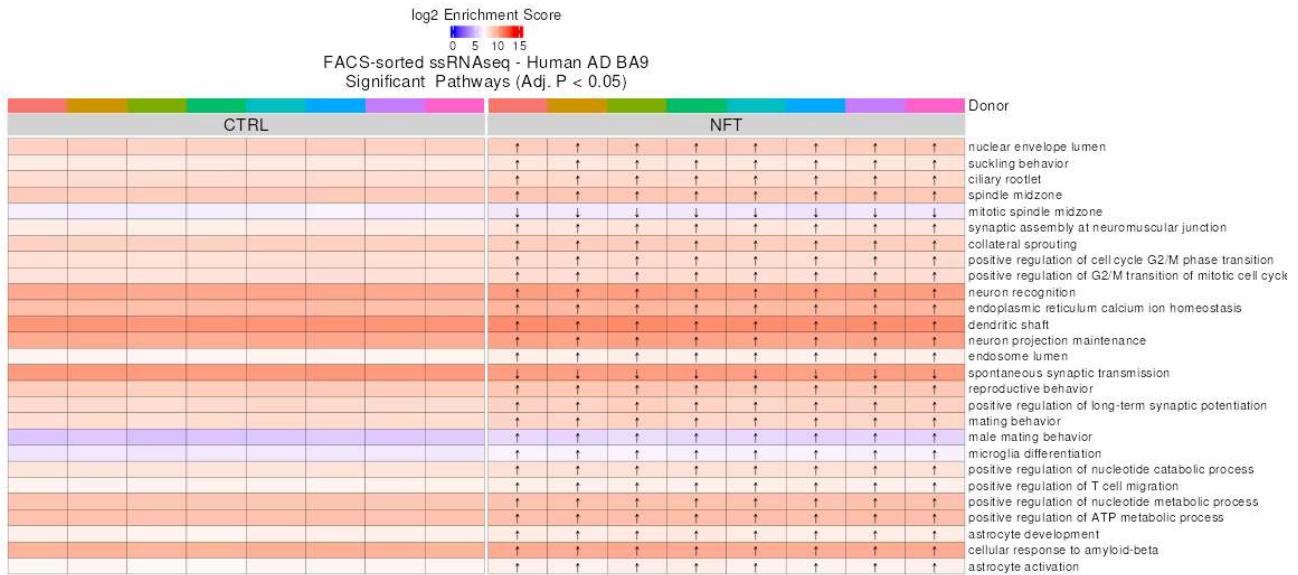


Figure 160: Heatmap showing enrichment of all differentially enriched pathways containing APP in the FACS ssRNASeq dataset. Figure format previously described in Figure 146.

The differentially enriched pathways in Figure 160 suggests that a wide variety of processes that interact with APP are upregulated on the transcript-level, with most being well-enriched in the dataset. This is supportive of findings regarding the interaction of genes from Figure 157, which appear to touch upon many domains. The apparent ubiquity of APP indeed makes further interpretation of these results difficult. Note however that another amyloid-related term is present in this list, “cellular response to amyloid-beta”. Further examination of this term (Figure 161) shows that is substantially larger in scope than the “positive regulation of amyloid fibril formation” previously visualised.

**Figure 161.**

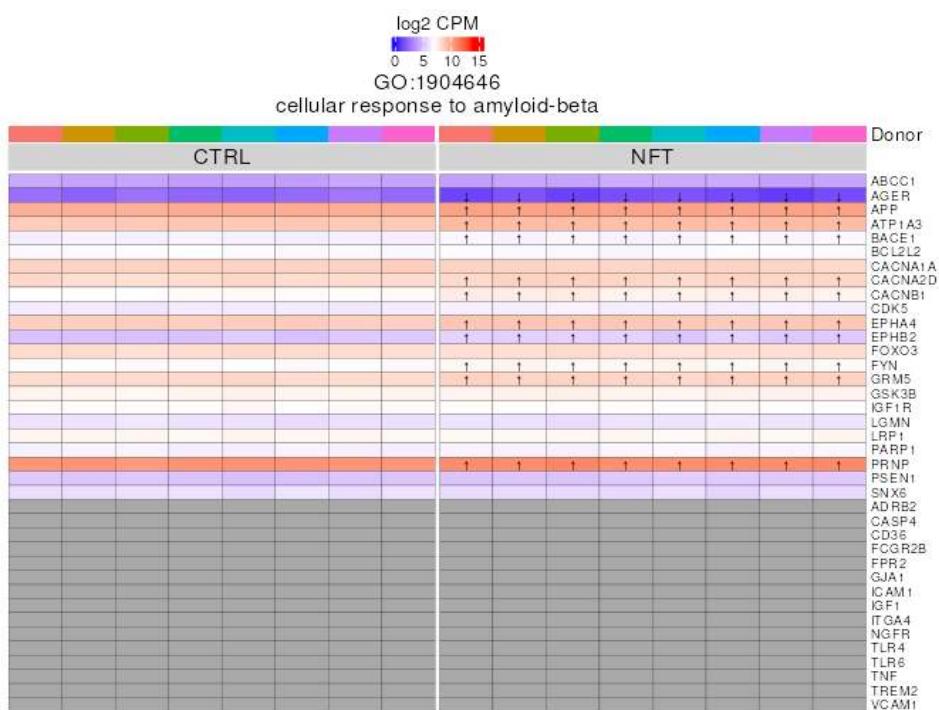


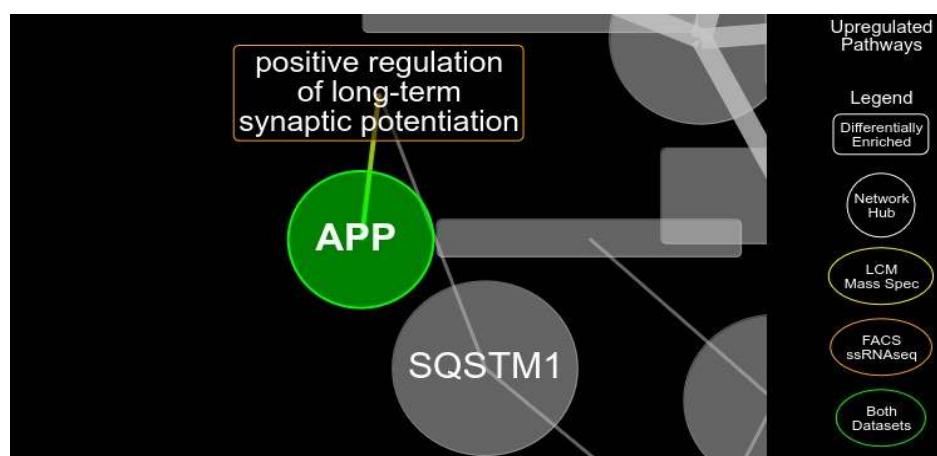
Figure 161: Heatmap showing the unscaled expression of genes within the “cellular response to amyloid-beta” pathway in the FACS ssRNAseq dataset. Figure format previously described in Figure 135.

In order to nail down a particular focus of the many enriched pathways on the transcriptomic level, I turned to the pathway network. With the pruning parameters set, the pathway-level network of APP (Figure 162) converged upon a single term, “positive regulation of long-term synaptic potentiation”. The gene contents of this term is further visualised in Figure 163. Interestingly, this term is shared with another selected hub – *SQSTM1*, which will be further discussed in the next section. Long-term synaptic plasticity is widely regarded as the fundamental mechanism underlying learning and memory, and impaired potentiation and depression of synaptic plasticity is a key factor in many neurodegenerative disorders, including Alzheimer’s Disease (Mango et al., 2019).

**Figure 162.**

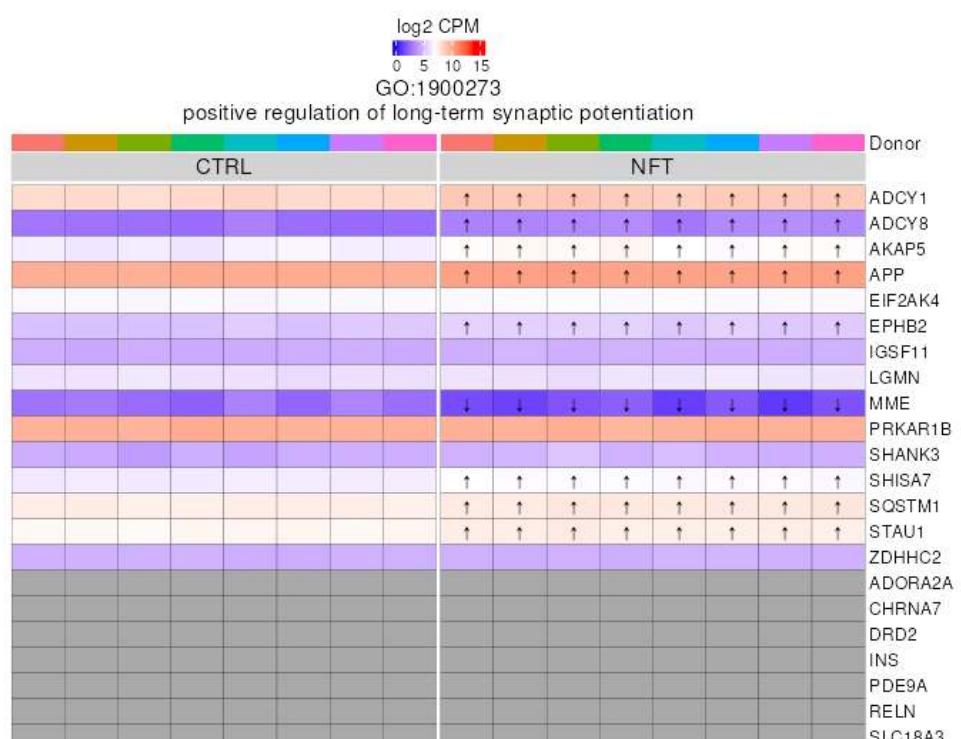
Figure 162: Pathway network of APP.

Figure format  
previously described  
in Figure 127.



**Figure 163.**

Figure 163: Heatmap showing the unscaled expression of genes within the “positive regulation of long-term synaptic potentiation” pathway in the FACS ssRNAseq dataset. Figure format previously described in Figure 135.



### 6.4.3 SQSTM1

Sequestosome 1 (SQSTM1, also known as p62) is a key regulator of autophagy and protein degradation pathways, particularly through its interactions with the ubiquitin-proteasome system (UPS) and the autophagy-lysosomal pathway (ALP) (Kumar et al., 2022). In AD, SQSTM1 plays a crucial role in the clearance of misfolded and aggregated proteins, including pathological tau (Y. Xu et al., 2019). However, in AD autophagy becomes dysfunctional (J.-H. Liang & Jia, 2014), and whether as a cause or consequence, p62 aggregation has been directly observed in neurofibrillary tangle-bearing neurons (Kuusisto et al., 2002). The relationship between SQSTM1 and other key proteins and genes detected in this analysis, such as FTH1, MAP1LC3A, and APP, provides further insight into how disruptions in proteostasis contribute to NFT pathology.

**Figure 164.**

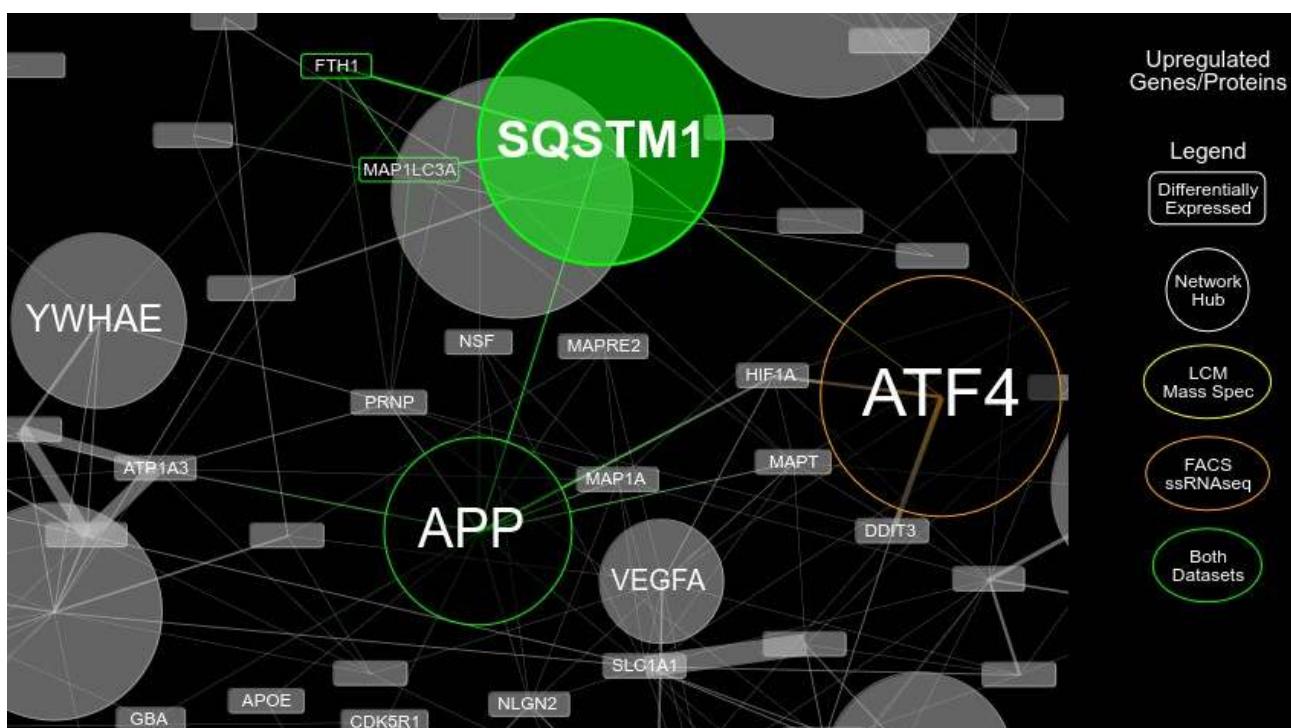


Figure 164. Feature network of SQSTM1. Figure format previously described in Fig. 140.

Ferritin heavy chain 1 (FTH1) is a key iron-storage protein that maintains cellular iron homeostasis and protects neurons from oxidative stress (Di Sanzo et al., 2022). Recent studies have identified iron dysregulation as a major contributor to neurodegeneration in AD (Ru et al., 2024), where iron accumulation promotes oxidative damage and may enhance tau aggregation. SQSTM1 has been implicated in ferritinophagy (Fang et al., 2025), a selective autophagy process responsible for degrading ferritin and maintaining iron balance. The dysfunction of SQSTM1-mediated ferritinophagy may exacerbate the oxidative environment in NFT-bearing neurons, accelerating tau pathology and neuronal degeneration.

Microtubule-associated protein 1 light chain 3 alpha (MAP1LC3A, commonly referred to as LC3A) is a core component of the autophagosome membrane (Bonam et al., 2020), essential for autophagy initiation and cargo degradation. SQSTM1 serves as an autophagy receptor, directly binding to LC3 through its LC3-interacting region (LIR) to facilitate the selective degradation of ubiquitinated protein aggregates (Kraft et al., 2016). In healthy neurons, this interaction enables the efficient removal of damaged proteins and prevents the accumulation of toxic aggregates. However, in AD autophagy is dysregulated (J.-H. Liang & Jia, 2014), leading to the accumulation of SQSTM1/p62 in tangle-bearing neurons (Kuusisto et al., 2002). The build-up of p62 may contribute to a broader failure of autophagosome formation and function, impacting components such as MAP1LC3A.

SQSTM1 has been linked to the metabolism of APP (another network in this analysis), as it can interact with ubiquitinated APP and facilitate its degradation via the autophagy-lysosomal system (Ma et al., 2019). However, in AD, p62 accumulation and autophagic dysfunction may impair APP homeostasis, leading to increased A $\beta$  production and deposition. Contributing to this feedback, A $\beta$  itself inhibits autophagy (M. Yuan et al., 2023), compounding proteostasis failure. As shown in the pruned pathway-level network in Figure 165, within the analysis APP and SQSTM1 (on the gene-level only) are linked by the GO term “positive regulation of long-term synaptic potentiation”. SQSTM1 has been implicated in synaptic plasticity through its role in AMPA receptor trafficking and mice deficient in p62 have been shown to exhibit impaired LTP in the hippocampal CA1 region (Jiang et al., 2009).

**Figure 165.**

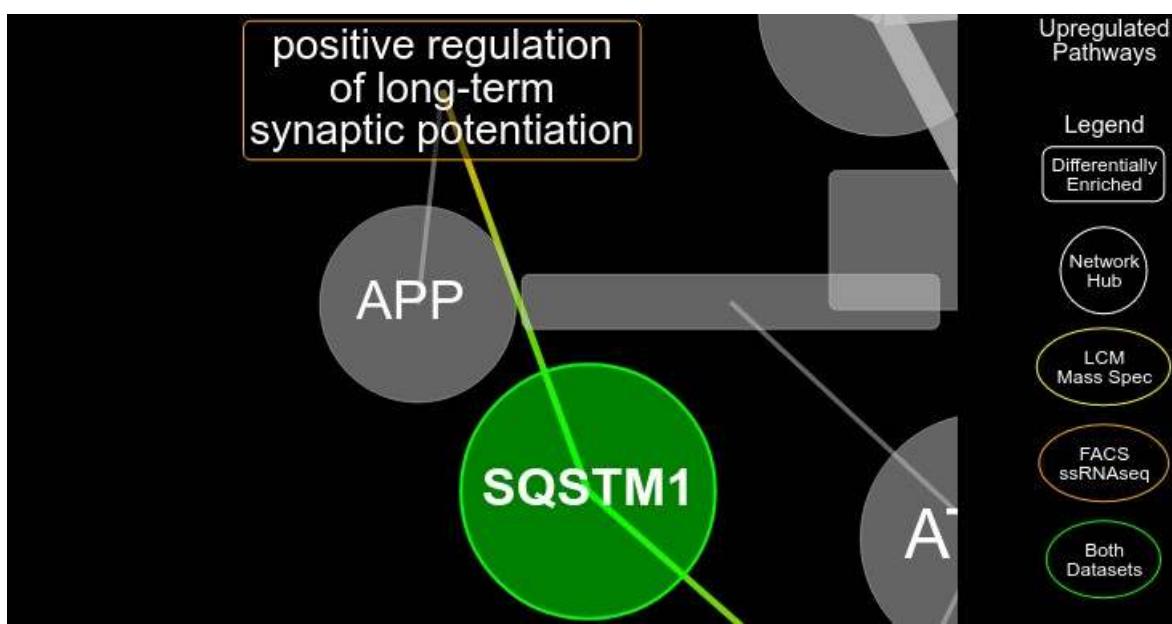


Figure 165: Pathway network of SQSTM1. Figure format previous described in Figure 142.

Figures 166 and 167, showing all differentially enriched pathways in each dataset confirms the role of SQSTM1 in autophagy in the present analysis. Further examination of some of these pathways also show membership of the previously discussed features, with the

“autolysosome” gene set in the LCM Mass Spec dataset showing the upregulation of FTH1 and MAP1LC3A (Figure 168). In the transcriptomic dataset, it is interesting to see the upregulation of Lewy body pathways in the tangle-bearing neurons. Lewy bodies describe protein aggregates primarily, but not exclusively, composed of  $\alpha$ -synuclein and neurofilaments, and is more commonly associated with Parkinson’s Disease and Lewy Body Dementia (Trojanowski, 1998). However, there are case reports of both NFTs and Lewy bodies coexisting within the same neuron (Iseki et al., 1999). While we did not employ immunohistochemistry in the current experiment to confirm this, the current transcriptomic results suggest that this may be of value for future experiments. A closer look at the genes in the “Lewy body” gene set is shown in Figure 169.

**Figure 166.**

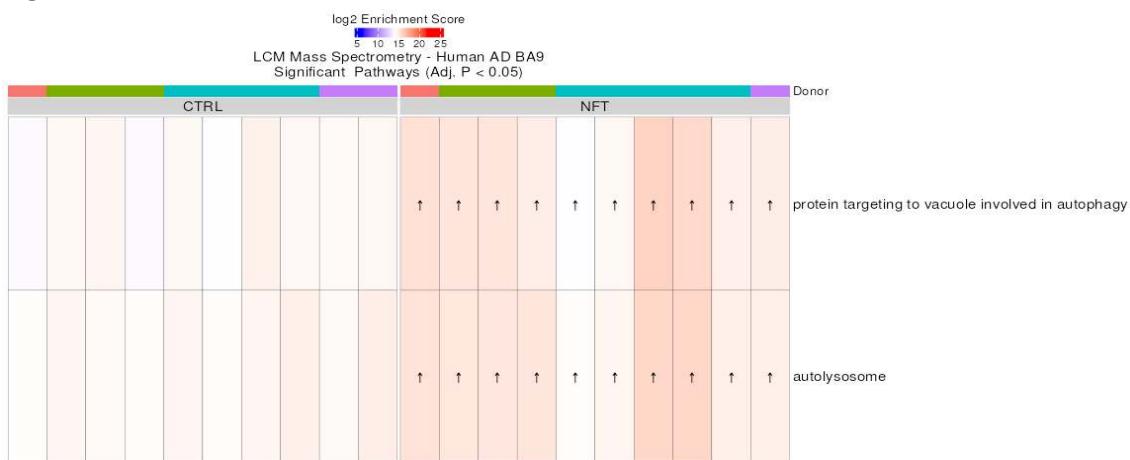


Figure 166: Heatmap showing enrichment of all differentially enriched pathways containing SQSTM1 in the LCM Mass Spec dataset. Figure format previously described in Fig. 146.

**Figure 167.**

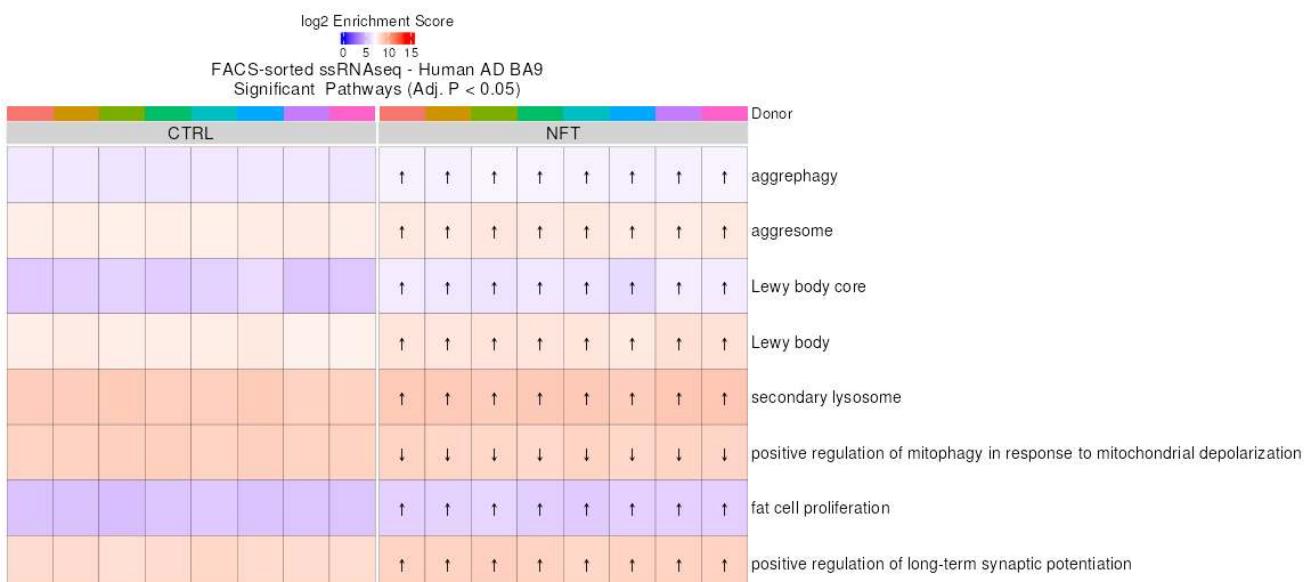


Figure 167: Heatmap showing enrichment of all differentially enriched pathways containing SQSTM1 in the FACS ssRNAseq dataset. Figure format previously described in Fig. 146.

**Figure 168.**

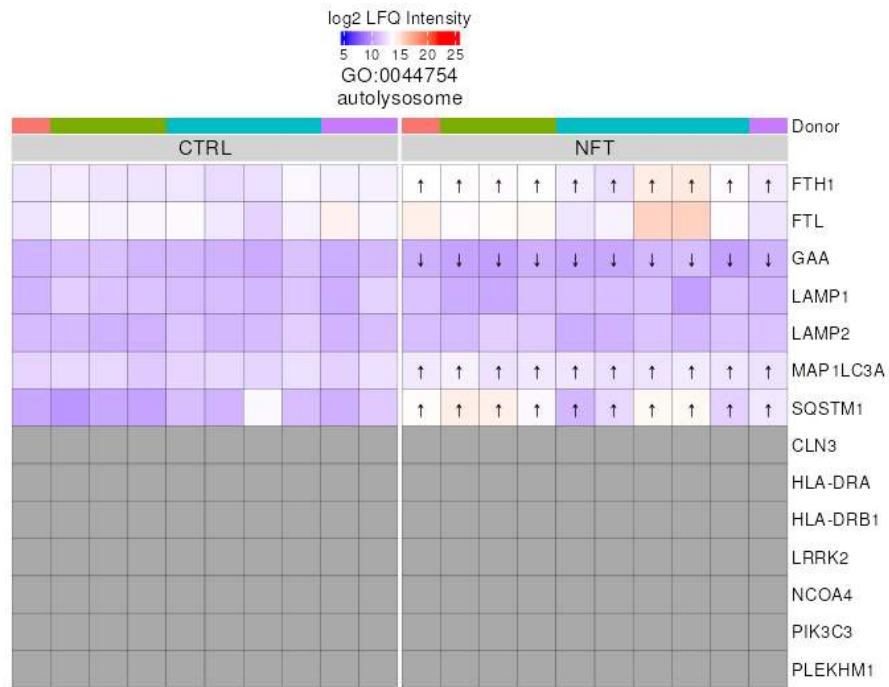


Figure 168: Heatmap showing the unscaled abundance of proteins within the “autolysosome” pathway in the LCM Mass Spec dataset. Figure format previously described in Fig. 135.

**Figure 169.**

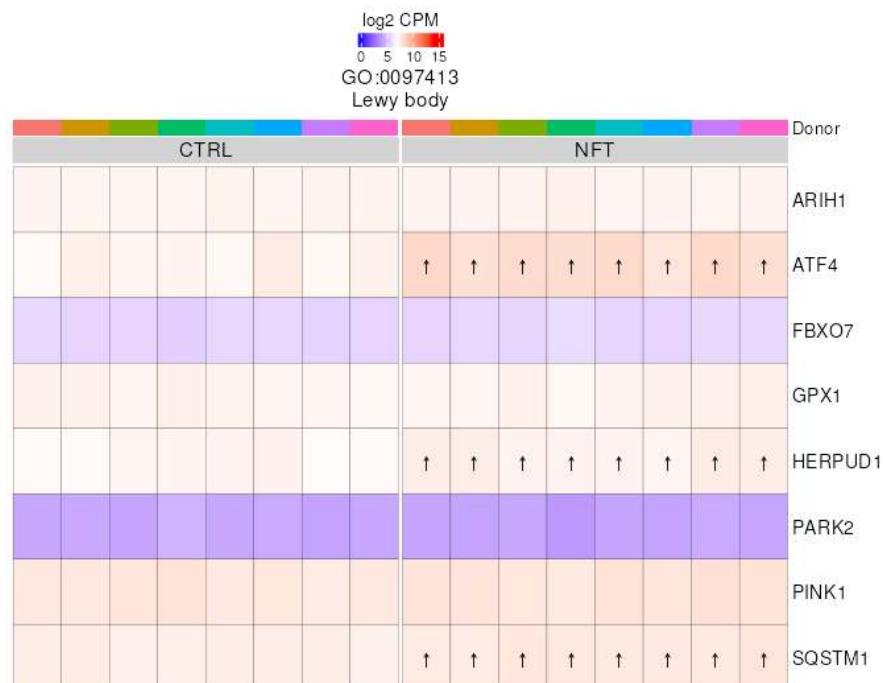


Figure 169: Heatmap showing the unscaled expression of genes within the “Lewy body” pathway in the FACS ssRNAseq dataset. Figure format previously described in Fig. 135.

#### 6.4.4 HSP90AA1

Heat shock protein 90 alpha, class A member 1 (HSP90AA1) is a molecular chaperone that plays a crucial role in protein folding, stability, and degradation under both normal and pathological conditions (Zuehlke et al., 2015). In AD, HSP90AA1 is implicated in tau homeostasis and NFT formation, as it interacts with hyperphosphorylated tau and regulates kinase activity involved in tau phosphorylation (Bohush et al., 2019). However, in Alzheimer's Disease, HSP90AA1 protein abundance and gene expression is often dysregulated (Astillero-Lopez et al., 2024; X.-L. Wang & Li, 2021), contributing to the accumulation of misfolded proteins, synaptic dysfunction, and cellular stress responses. Among the key interactors with HSP90AA1 in this analysis are YWHAE (a network hub discussed in its own section), ATP2B4, HSPA9, CCT5, LONP1, ATP1B1, ATP1A1, and NCKAP1.

**Figure 170.**

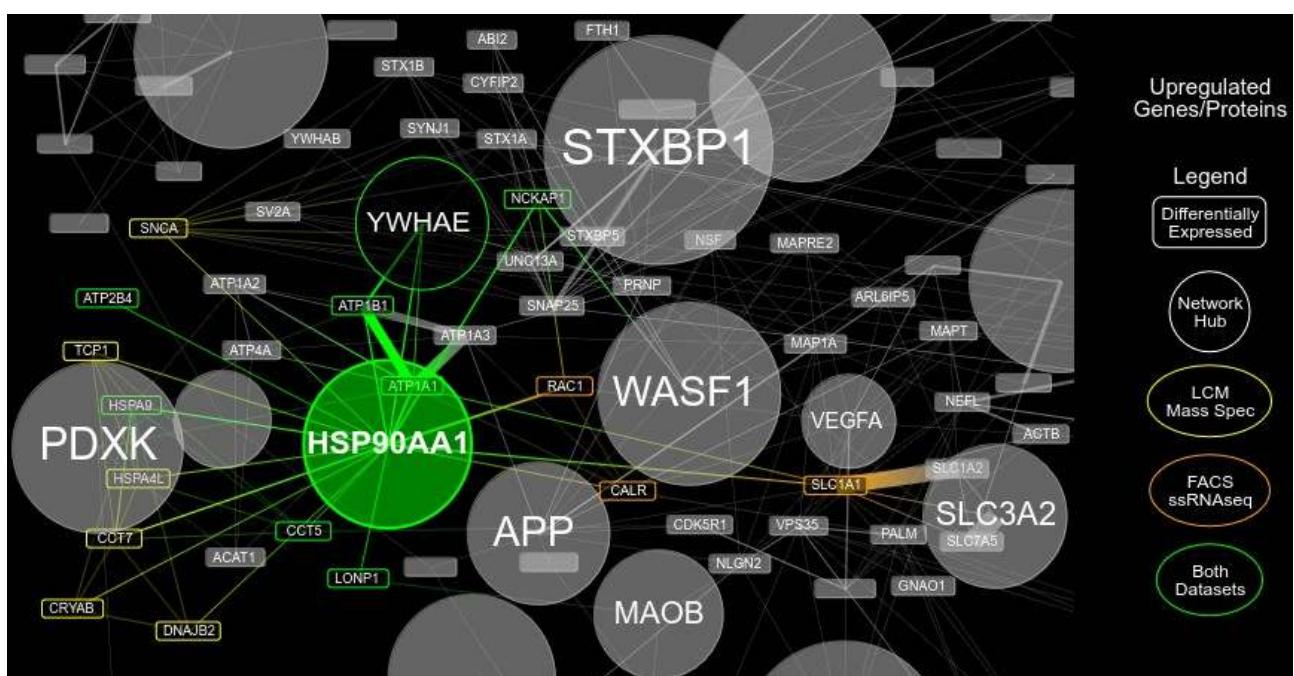


Figure 170. Feature network of HSP90AA1. Figure format previously described in Fig. 140.

To give a brief overview of these features: ATP2B4 (Plasma membrane calcium-transporting ATPase 4, PMCA4) is crucial for neuronal calcium homeostasis (Zámbó et al., 2017). While ATP1B1 and ATP1A1 are subunits of the Na<sup>+</sup>/K<sup>+</sup>-ATPase, which is essential for neuronal excitability and ion homeostasis (Sahoo et al., 2016). HSPA9 (mortalin, a mitochondrial chaperone) is another heat shock protein that interacts with HSP90AA1 to regulate mitochondrial stability and oxidative stress responses (Ferré et al., 2021). CCT5 (chaperonin-containing TCP-1 subunit 5) is a component of the CCT/TRiC complex, which is responsible for assisting in the folding of cytoskeletal proteins and microtubules (Grantham, 2020). LONP1 (Lon protease 1) is a mitochondrial protease that degrades misfolded mitochondrial proteins and plays a role in mitochondrial quality control

(Matsushima et al., 2021). Finally, NCKAP1 (Nck-associated protein 1) is involved in actin cytoskeleton remodeling, playing a role in synaptic plasticity and neuronal structure maintenance (Han & Ko, 2023a). The interaction of these many genes and proteins, across a variety of functions, in conjunction with the protein stability roles of HSP90AA1, may be strong contributors to the pathological processes at play in tangle-bearing neurons.

**Figure 171.**

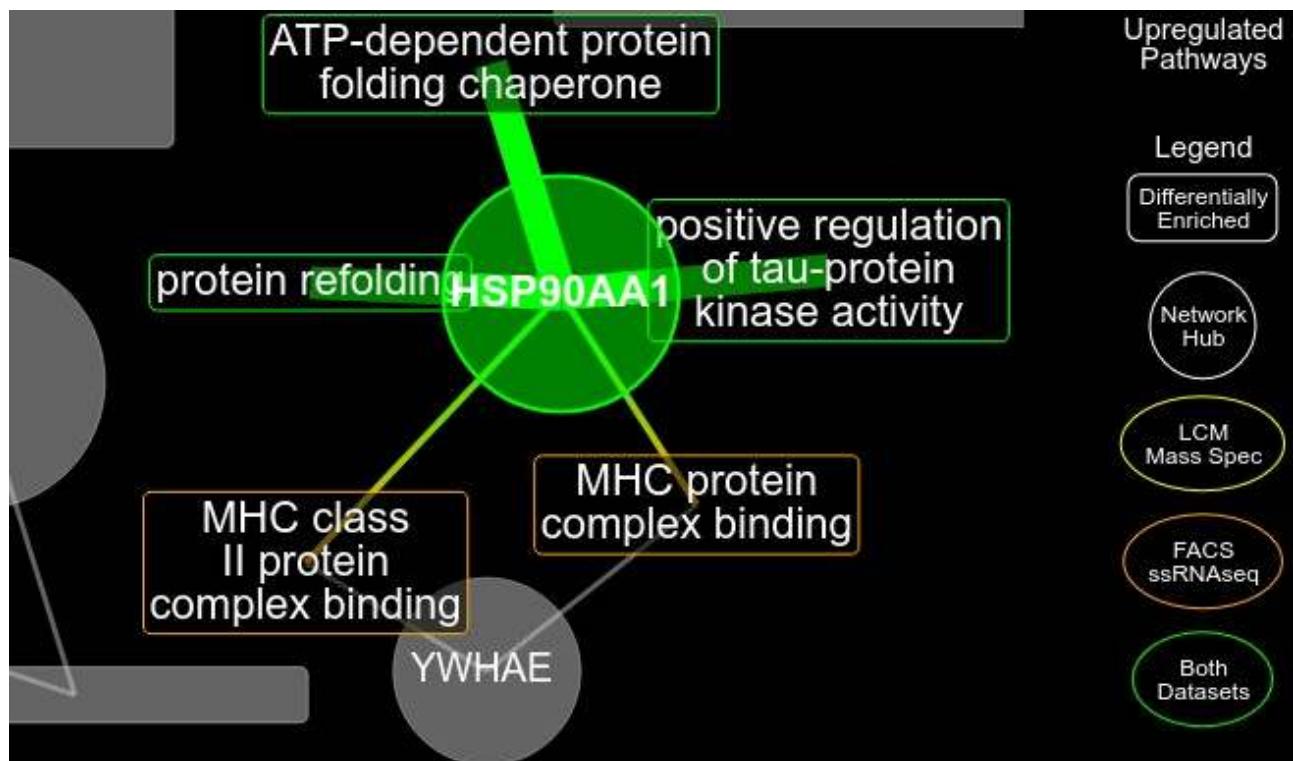


Figure 171. Pathway network of HSP90AA1. Figure format previous described in Fig. 142.

Figure 171 shows the pathway network for HSP90AA1 in this analysis, which may help narrow down the many implied mechanisms to those that may be most pertinent. In line with the most commonly associated function of HSP90AA1 are the two pathways “protein refolding” and “ATP-dependent protein folding chaperone”, the later of which is visualised in both datasets in Figure 172. These figures highlight the dispersion of protein folding mechanisms across many heat shock related features, as well as the CCT/TRiC complex. It is also interesting to observe the considerable variance in the features comprising this pathway, with some, like HSP90AA1 being relatively highly expressed/abundant, while others are a low, and a few were even detected as DE in the opposite direction of the overall pathway. This is a relevant example that supports usage of gene set enrichment methods; though not all components are DE in a consistent direction, enough of them are such that the sum of all parts suggest that the pathway as a whole is changed.

**Figure 172.**

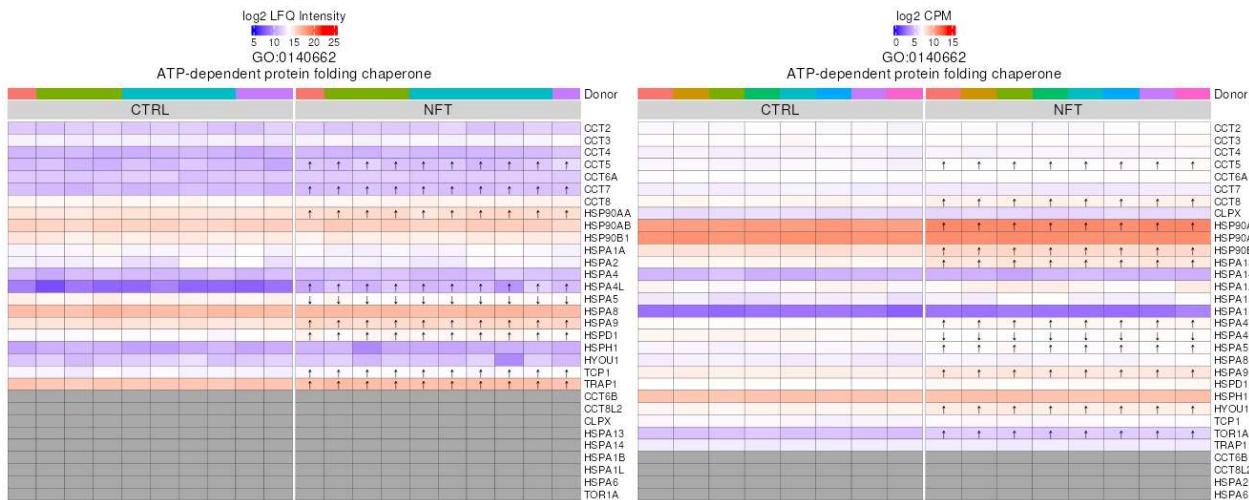


Figure 172: Heatmaps showing the unscaled expression of features within the “ATP-dependent protein folding chaperone” pathway in the LCM Mass Spec (left) and FACS ssRNASeq (right) datasets. Figure format previously described in Figure 135.

The other immediately relevant pathway shown in Figure 171 is the “positive regulation of tau-protein kinase activity” pathway. Unfortunately, Gene Ontology defines very few features within this set – only HSP90AA1 and HSP90AB1 (which was detected as DE on its own in either dataset. This makes it difficult to take actionable steps for further functional research, outside of the already known fact that HSP90AA1 (and seemingly HSP90AB1) interact with tau kinases. This itself, as previously discussed, has large implications for the progression of tau hyperphosphorylation, and accordingly, tau aggregation. That being said, visualisation of shared and distinct enriched pathways for HSP90AA1 and HSP90AB1 for the LCM Mass Spec dataset implicate additional roles of the two proteins together in tangle-bearing neurons in nitric-oxide synthase and telomerase activity regulation (Figure 173).

**Figure 173.**

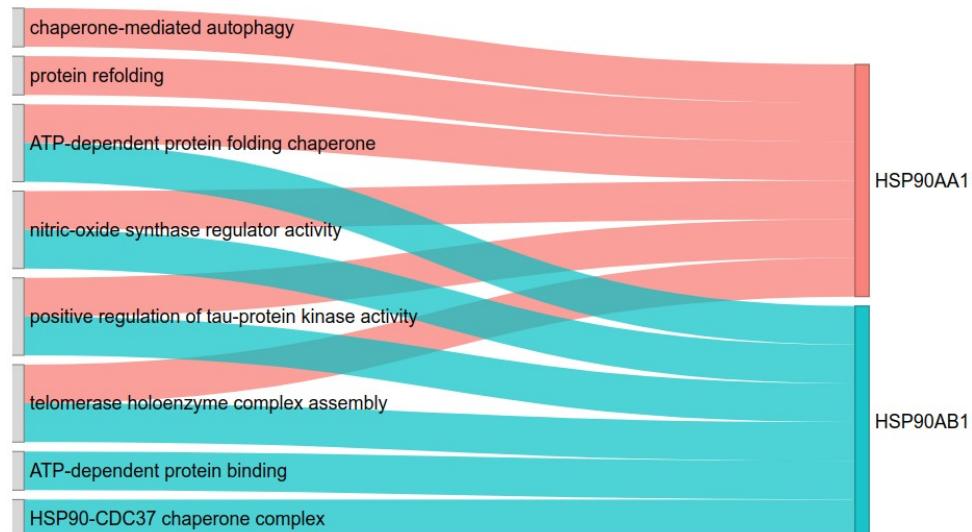


Figure 173: Sankey plot showing shared and distinct enriched pathways for HSP90AA1 and HSP90AB1 in the LCM Mass Spec dataset.

On the gene-level, the pathway network in Figure 171 also implicates the involvement of HSP90AA1 in the major histocompatibility complex (MHC). As this pathway is shared with YWHAE, another network hub, it will be discussed in that hub's section. Finally, beyond the pathways discussed thus far, visualisation of all differentially enriched pathways on the gene-level shows involvement of HSP90AA1 across many different domains (Figure 174). Of note, are a number of vascular-related pathways and those involved with basic molecular processing of compounds such as pyrimidines. The later is more difficult to interpret in the context of disease, but vascular dysfunction is well-known component of Alzheimer's Disease pathophysiology, inspiring widely publicised initiatives to classify AD as "type 3 diabetes" (Peng et al., 2024).

**Figure 174.**

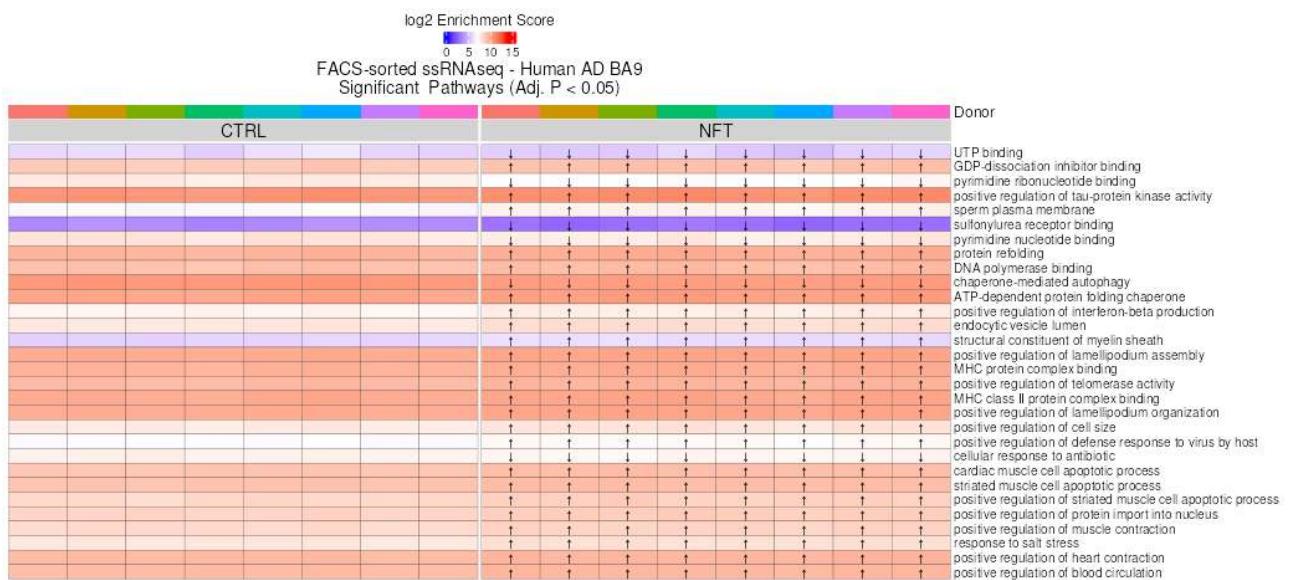


Figure 174: Heatmap showing enrichment of all differentially enriched pathways containing HSP90AA1 in the FACS ssRNAseq dataset. Figure format previously described in Fig. 146.

#### 6.4.5 YWHAE

YWHAE (14-3-3 $\epsilon$ ) is a scaffolding and signaling protein that plays a crucial role in regulating tau phosphorylation, synaptic function, and cellular stress responses (Foote & Zhou, 2012). As part of the 14-3-3 protein family, YWHAE binds to phosphorylated tau and influences its affinity for microtubules, thereby contributing to neuronal stability and cytoskeletal organization (Y. D. Ke et al., 2019). In this dataset, YWHAE interacts with a variety of genes and proteins involved in neuronal integrity and stress response, including HSP90AA1, ATP1A1, ATP1B1, PRNP, and FTH1.

**Figure 175.**

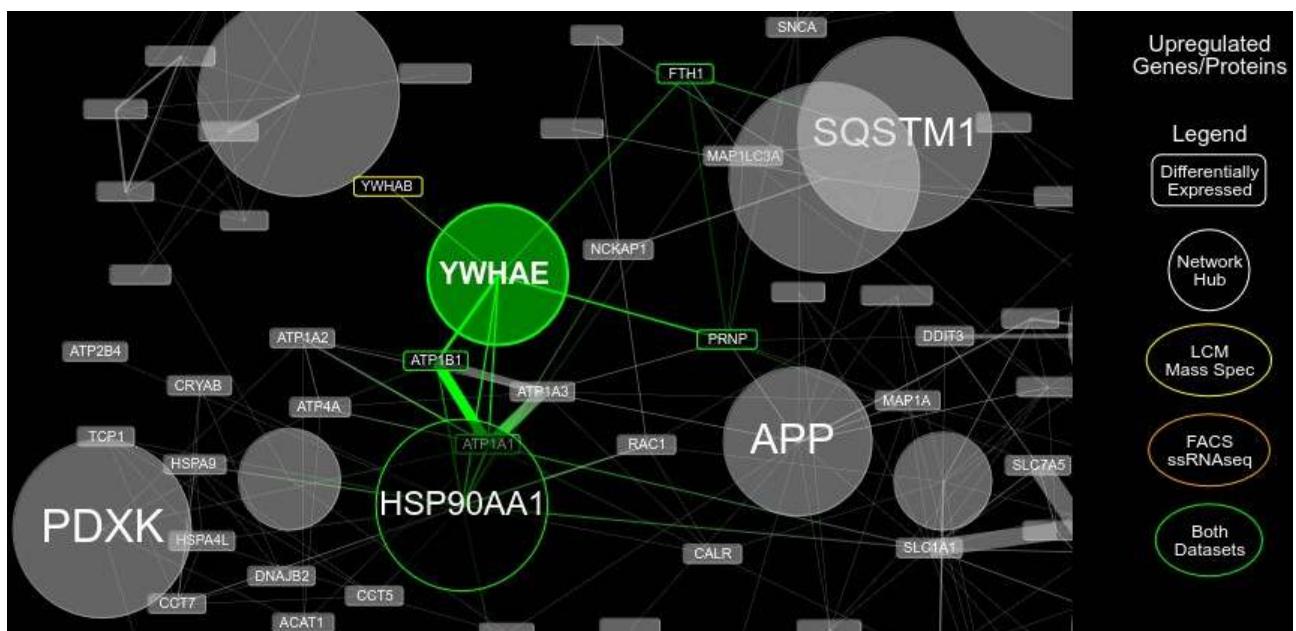


Figure 175. Feature network of YWHAE. Figure format previously described in Fig. 140.

YWHAE and HSP90AA1 share a functional relationship in regulating tau homeostasis and protein stability. HSP90AA1 is a molecular chaperone that assists in the folding and degradation of misfolded proteins, including tau. Meanwhile YWHAE directly binds to phosphorylated tau, influencing its detachment from microtubules. The dysregulation of each of these genes and proteins may work synergistically to potentiate tau aggregation. That being said, the present analysis, at least with current pruning parameters, did not highlight this association between the two features, instead implicating mechanisms involving major histocompatibility complex (MHC) (Figure 176). Upon further inspection, it is possible that this is an unreliable signal, as many of the genes involved are very lowly expressed, particularly those of the CD class (i.e. CD74) (Figure 177). Recall that those cells in the heatmap that are coloured gray are also lowly expressed genes, though they are counted in the gene set scoring by GeneFunnel, just excluded from differential expression testing and visualisation. Indeed, I could not find relevant research linking these genes together in relation to MHC. Gene Ontology cites (Buschow et al., 2010) as

the sole reference, which examined exosomes, a type of secreted vesicle, which while relevant to AD biology (Fowler et al., 2025), is not solid evidence of this activity within cells.

**Figure 176.**

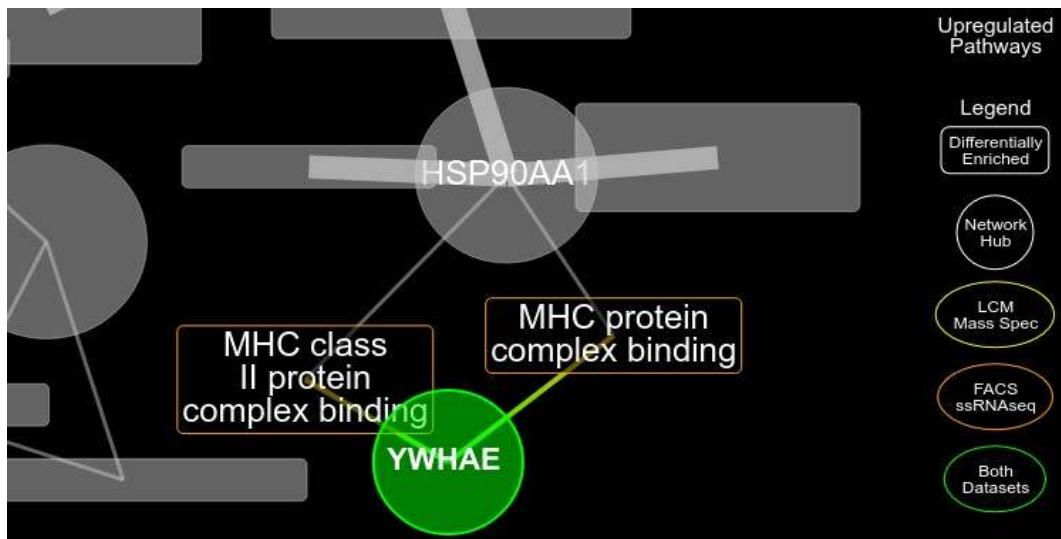


Figure 176. Pathway network of YWHAE. Figure format previous described in Fig. 142.

**Figure 177.**

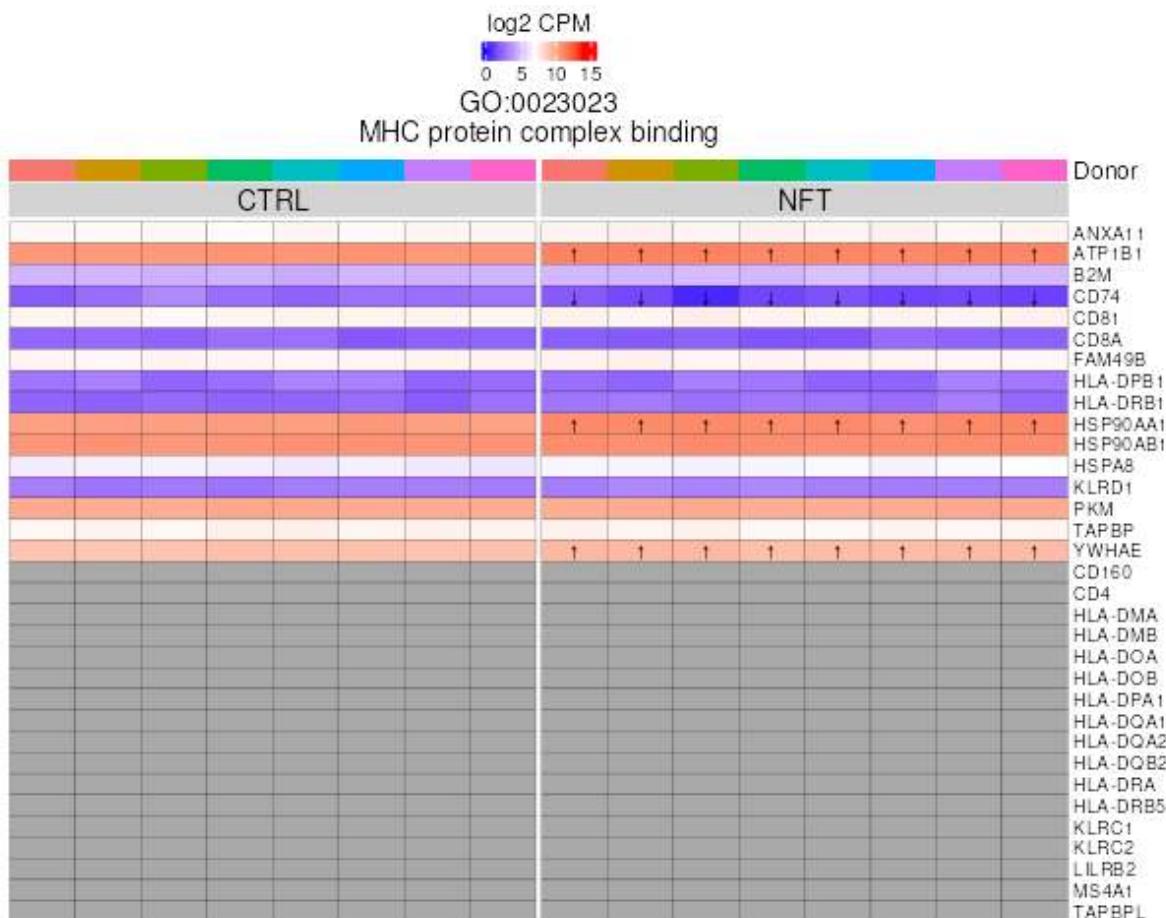


Figure 177: Heatmap showing the expression of genes within the “MHC protein complex binding” pathway in the FACS ssRNAseq dataset. Format previously described in Fig. 135.

Other features in Figure 175 have been previously mentioned. Like with the network for HSP90AA1, the Na+/K+-ATPase complex, which includes ATP1B1 (Na+/K+-ATPase beta subunit 1) and ATP1A1 (Na+/K+-ATPase alpha subunit 1), critical regulators of neuronal excitability and ion homeostasis (Sahoo et al., 2016). In AD, A $\beta$  and tau pathology have been linked to disruptions in Na+/K+-ATPase activity, potentially leading to calcium overload and excitotoxicity (Petrushanko et al., 2016). And as in with SQSTM1, FTH1, which is a key iron-storage protein that maintains cellular iron homeostasis and protects neurons from oxidative stress (Di Sanzo et al., 2022). Finally, PRNP, which as discussed in relation to APP, is implicated in protein misfolding diseases like CJD and AD (Calero et al., 2011). It is less clear how YWHAE interacts with these features due to a lack of research interest in YWHAE compared to the previously mentioned hubs, all of which are big names in disease research.

It is potentially more insightful to look into other pathways containing YWHAE that was shown differentially enriched. In the LCM Mass Spec, the only gene set was “phosphoserine residue binding”, which are the specific motifs targeted by 14-3-3 proteins when mediating protein localisation. This is a potential path to direct interaction with AD as phosphoserine has been reported to be elevated in AD post-mortem brain tissue (Klunk et al., 1991). The proteins involved in this process is shown in Figure 178. Figure 179 shows all differentially enriched pathways containing YWHAE in the LCM Mass Spec dataset. Out of them, it is interesting to see pathways related to sequestration. In the context of proteins, sequestering is the process by which specific proteins are isolated or bound by other molecules, preventing them from interacting with their usual cellular targets or participating in biological processes; it is often as a regulatory mechanism in signalling pathways, stress responses, or protein aggregation disorders (H. Yang & Hu, 2016). The genes involved in the gene set “protein sequestering activity” is shown in Figure 180. In both this figure, and Figure 178, it can be observed that YHWAE often functions alongside related components of the 14-3-3 class, such as YHWAE and YHWAZ.

**Figure 178.**

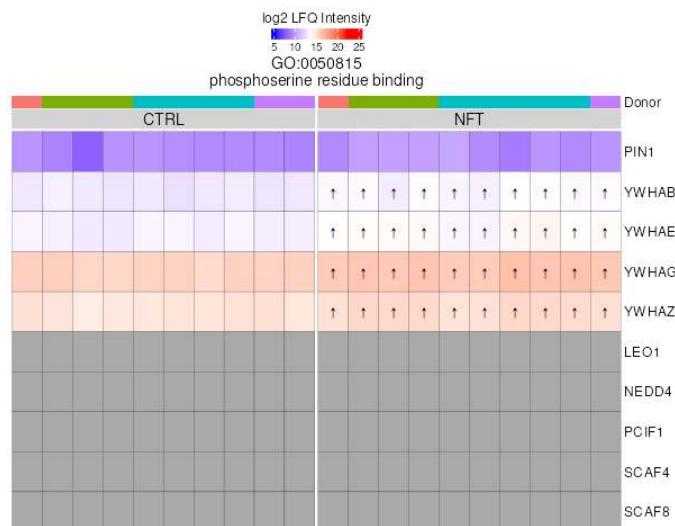


Figure 178: Heatmap showing the unscaled abundance of proteins within the “phosphoserine residue binding” pathway in the LCM Mass Spec dataset. Figure format previously described in Fig. 135.

**Figure 179.**

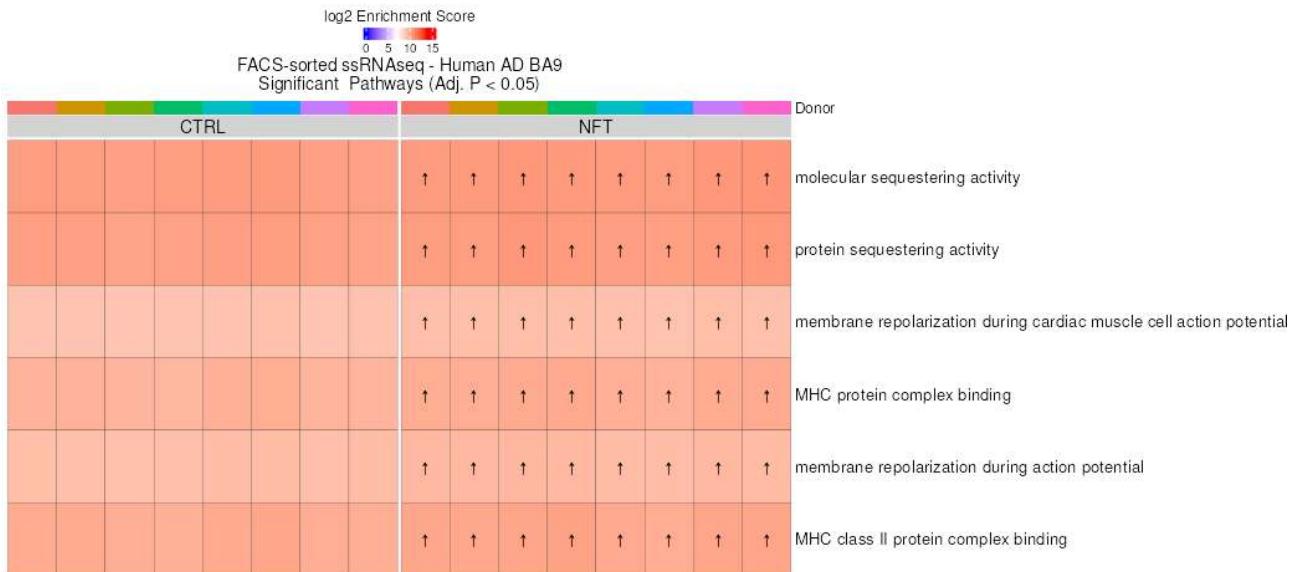


Figure 179: Heatmap showing enrichment of all differentially enriched pathways containing YWHAE in the FACS ssRNAseq dataset. Figure format previously described in Fig. 146.

**Figure 180.**

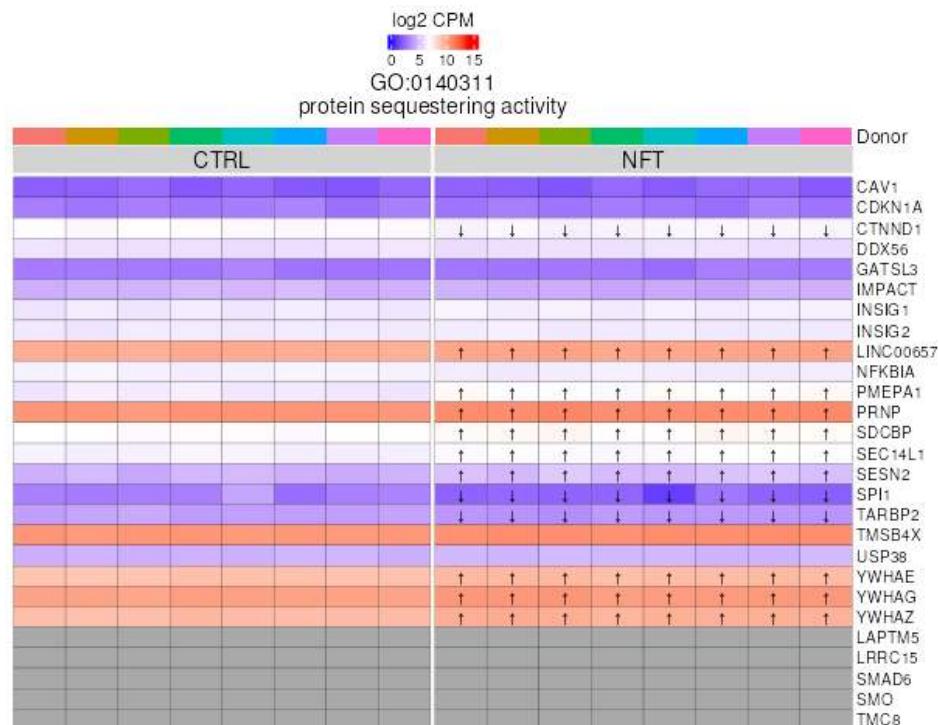


Figure 180: Heatmap showing the unscaled expression of genes within the “protein sequestering activity” pathway in the FACS ssRNAseq dataset. Figure format previously described in Fig. 135.

#### 6.4.6 WASF1

WASF1 (WAVE1, WASP-family verprolin homologous protein 1) is a critical regulator of actin cytoskeleton remodeling, playing an essential role in synaptic plasticity, dendritic spine formation, and neuronal connectivity (Han & Ko, 2023b). It is a core component of the WAVE regulatory complex (WRC), which activates the Arp2/3 complex to drive actin polymerization and dendritic spine maintenance. WASF1 has been shown to colocalise directly with tau in the 3xTg AD mouse model (Watamura et al., 2016). The present analysis highlights NCKAP1 and MAPT as top interacting partners in terms of shared differentially expressed pathways in the LCM Mass Spec and FACS ssRNAseq datasets.

**Figure 181.**

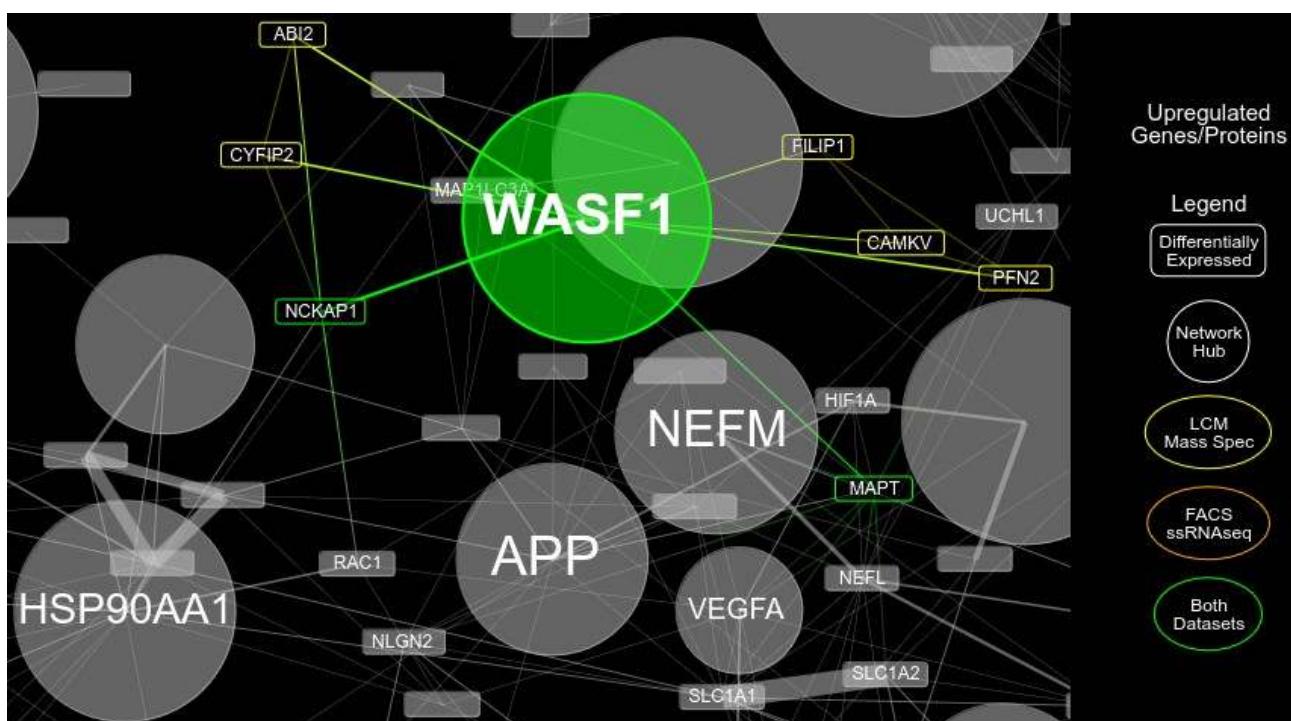


Figure 181. Feature network of WASF1. Figure format previously described in Fig. 140.

NCKAP1 (Nck-associated protein 1) is previously described, as it was also associated with the network hub HSP90AA1. NCKAP1, one of five key components of the WRC, regulates actin cytoskeleton dynamics, contributing to synaptic plasticity and the structural integrity of neurons (Han & Ko, 2023b). While I could not find a report of NCKAP1 being upregulated in AD, its gene has been reported to be downregulated in microarray data from AD prefrontal cortex tissue (Y. Zhu et al., 2023). Another group revealed that overexpression of NCKAP1 led to an upregulation of actin polymerization-associated proteins, including CYFIP1, ABI2, WAVE1, and WAVE2, while NCKAP1 knockdown resulted in their decreased expression (Noh et al., 2023). They further showed the NCKAP1 ameliorated defects in phagocytic function in an amyotrophic lateral sclerosis (ALS) model of microglia-like cells. Another study compared *WASF1* between human AD and healthy control neocortical tissue and reported a reduction in gene expression (Ceglia et

al., 2015). That study also showed that significantly lowering *Wasf1* led to a marked decrease in A $\beta$  levels and reversed memory impairments in the Tg-APPswe AD mouse model. This may suggest that the WRC as a whole may be a protective factor in neurodegeneration, and the upregulation observed in the datasets of this thesis is representative of a compensatory mechanism.

(Watamura et al., 2016) reports that WAVE1 (WASF1) and tau directly colocalise in the 3xTg AD mouse model. Interestingly, (Takata et al., 2009) reports that both tau and amyloid pathologies are required for WAVE1 accumulation, as accumulation was not detected in the JNPL3 or Tg2576 mice, which respectively feature each pathology separately. However, robust WAVE1 accumulation was observed in the 3xTg model, which harbour both pathologies. In AD, the kinases Cdk5 and GSK3- $\beta$  are heavily implicated in the regulation of cytoskeletal dynamics and axonal transport. Their activity influences these processes by phosphorylating critical molecules including WAVE1, as well as tau (Ceglia et al., 2010). It is possible that kinase dysregulation is in fact the causal factor in the observed upregulation of both WAVE1 and tau, leading to their aggregation, loss of cytoskeletal stability, and dysfunction of the hypothesised protective effects of the WAVE regulatory complex.

**Figure 182.**

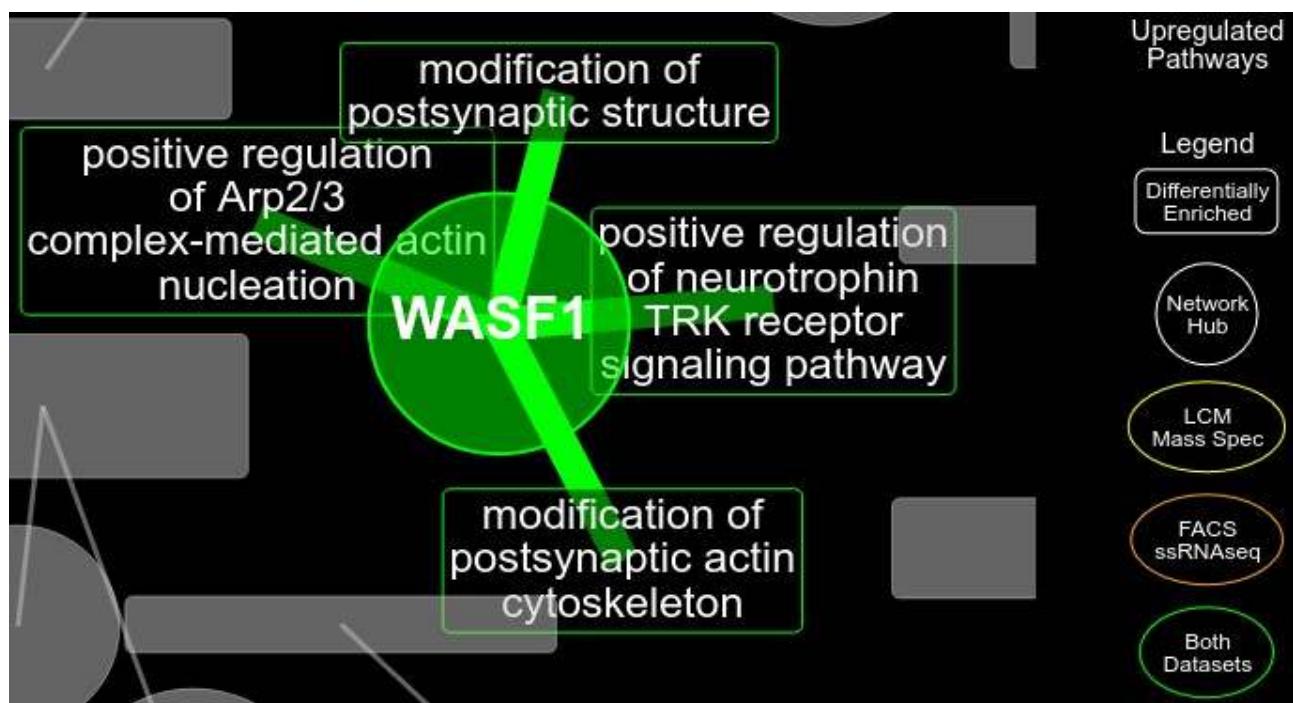


Figure 182. Pathway network of WASF1. Figure format previous described in Fig. 142.

The discussion presented above is supportive of the pruned pathways associated with WASF1 shown in Figure 182. An additional pathway not discussed is “positive regulation of neurotrophin TRK receptor signaling pathway”. The Trk receptor signaling pathway involves a family of receptor tyrosine kinases, TrkA, TrkB, and TrkC, each activated by specific neurotrophins such as nerve growth factor (NGF), brain-derived neurotrophic

factor (BDNF), and neurotrophins 3 and 4 (NT3 and NT4). Upon neurotrophin binding, Trk receptors initiate intracellular signaling cascades that regulate diverse domains such as neuronal survival, axonal and dendritic growth, and synaptic plasticity (E. J. Huang & Reichardt, 2003). Neurotrophin related pathways appear to be the link between WASF1 and MAPT in this dataset, as they both appear in the pathway “cellular response to brain-derived neurotrophic factor stimulus”. This pathway was only differentially enriched in the FACS ssRNAseq dataset and the genes within the pathway are shown in Figure 183.

**Figure 183.**

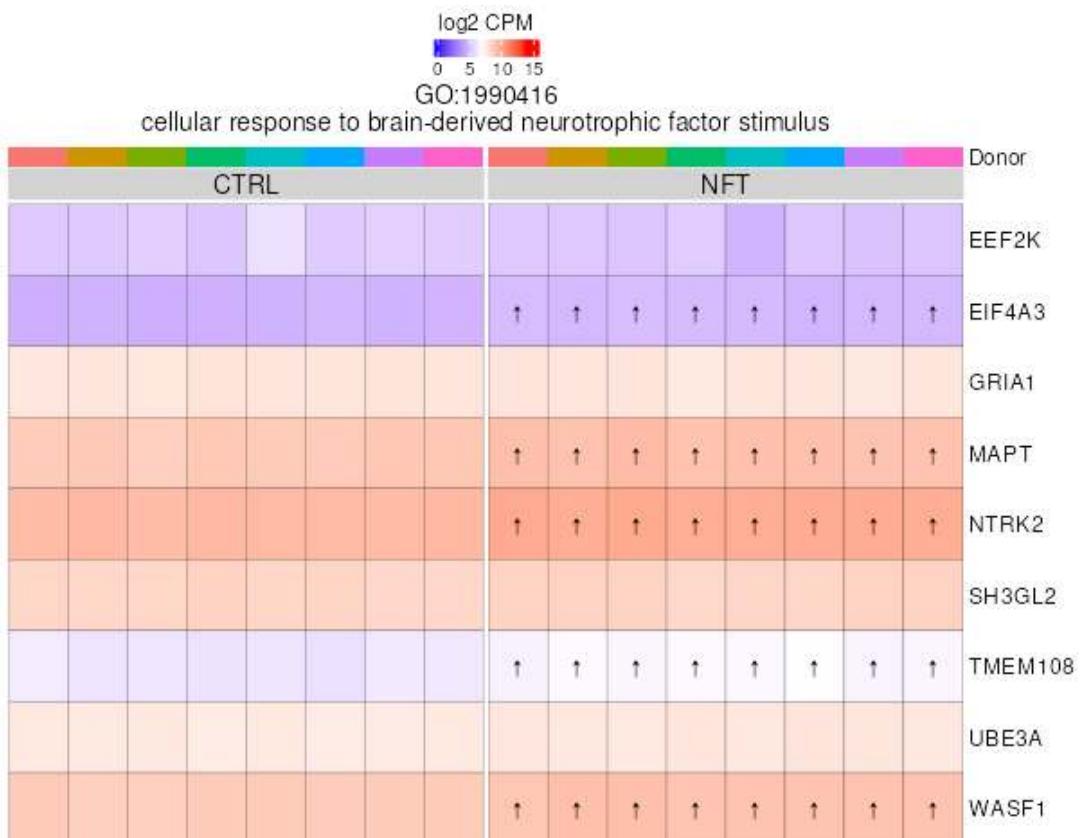


Figure 183: Heatmap showing the unscaled expression of genes within the “cellular response to brain-derived neurotrophic factor stimulus” pathway in the FACS ssRNAseq dataset. Figure format previously described in Fig. 135.

#### 6.4.7 CNTNAP1

CNTNAP1 (Contactin-associated protein 1 or Caspr) is a neuronal adhesion molecule essential for axonal organization, synaptic function, and neuronal excitability. It plays a crucial role in maintaining the integrity of paranodal junctions in myelinated neurons, facilitating axon-glial interactions that are critical for proper neuronal signaling (W. Li et al., 2020b). In this analysis, CNTNAP1 interacts with several genes and proteins involved in cytoskeletal organization, endosomal trafficking, and synaptic stability, including ANK2, VPS35, PALM, and NEFL.

**Figure 184.**

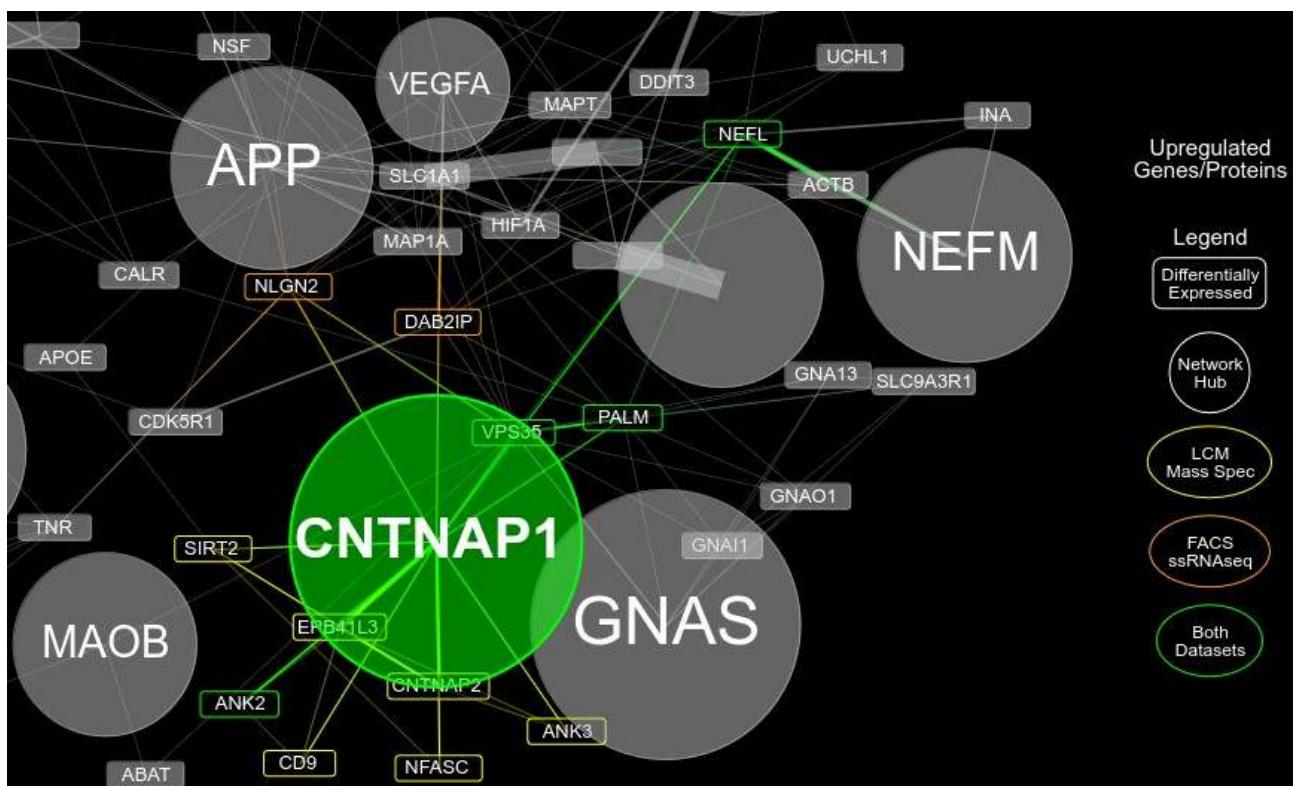


Figure 184. Feature network of CNTNAP1. Figure format previously described in Fig. 140.

ANK2 (Ankyrin-2, also known as Ankyrin-B) is a scaffolding protein involved in linking membrane proteins to the actin cytoskeleton, particularly at the nodes of Ranvier and paranodal junctions in axons (Kawano et al., 2022). VPS35 (Vacuolar Protein Sorting 35) is a core component of the retromer complex, responsible for endosomal sorting and protein recycling (A. Wu et al., 2024). PALM (Paralemmin-1) is a scaffolding protein involved in neuronal membrane dynamics, synaptic plasticity, and dendritic spine remodeling (Macarrón-Palacios et al., 2025). NEFL, as previously discussed, is a major component of the neuronal cytoskeleton, critical for axon stability and intracellular transport (Campos-Melo et al., 2018). The interaction between CNTNAP1 and many of these largely structural components may largely contribute to the instability induced by tau pathology in AD.

The pathway network analysis performed in these datasets support the role of CNTNAP1 in paranodal junctions on both a transcriptomic and proteomic level (Figure 185).

Inspection of additional pathways differentially enriched in the FACS ssRNASeq dataset also shows a term called “myelin assembly”, which may provide a link between this so-far neuron focused analysis to glia (Figure 186). Of the above mentioned genes, this pathway also incorporates ANK2, which itself is implicated in myelin specific processes. Gene expression of CNTNAP1 has been reported as upregulated human AD post-mortem tissue, though interpretation of such data has predominantly been in the context of regulating APP function (Bamford et al., 2020). However, extensive work outside of the AD research field have reported that mutations in CNTNAP1 is highly involved in myelination disorders (W. Li et al., 2020c), and related mechanisms may be at play when CNTNAP1 is dysregulated alongside other key genes in tangle-bearing neurons.

**Figure 185.**

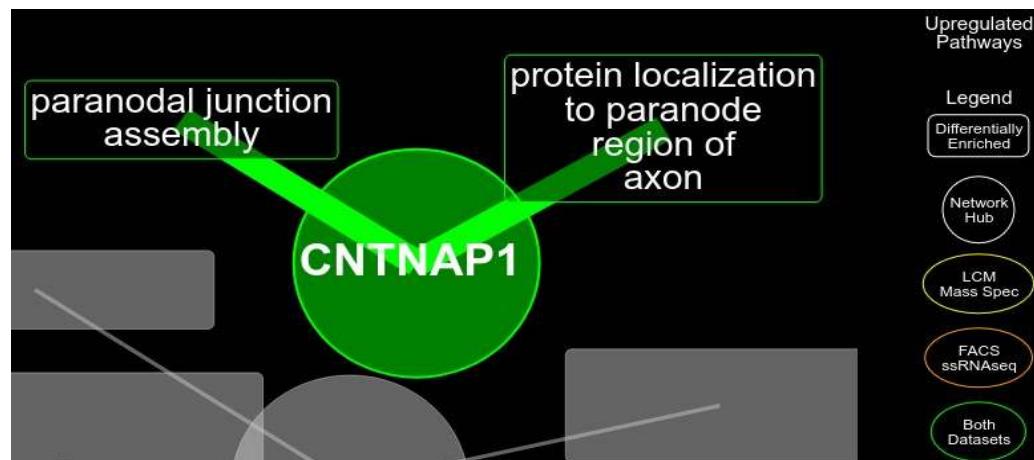


Figure 185. Pathway network of CNTNAP1. Figure format previous described in Fig. 142.

**Figure 186.**

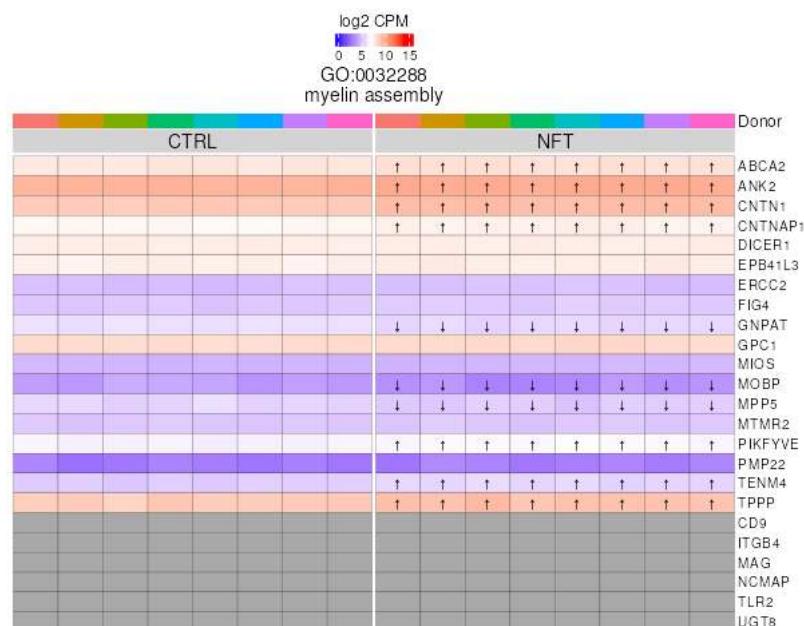


Figure 186: Heatmap showing the unscaled expression of genes within the “myelin assembly” pathway in the FACS ssRNASeq dataset. Figure format previously described in Fig. 135.

#### 6.4.8 GOT2

Glutamate oxaloacetate transaminase 2 (GOT2), also known as aspartate aminotransferase, is a component of the malate-aspartate shuttle (MAS), a cytosolic-mitochondrial pathway responsible for transferring reducing equivalents (donation of an electron into an electron recipient) into mitochondria to sustain oxidative phosphorylation (Borst, 2020). The reaction is a rapid process that bypasses the tricarboxylic acid (TCA) cycle, and is thought to be required for neuronal activity (Yudkoff et al., 1994). Studies have shown that GOT2 is decreased in both AD model mice (H. Li et al., 2023) and human AD cases (Choe et al., 2024; Mahajan et al., 2020). However, in this thesis work, GOT2 was identified as upregulated in both the transcriptomic and proteomic datasets. Its interacting partners include DLST and GLUD1, which were also upregulated in both datasets (Figure 187).

**Figure 187.**

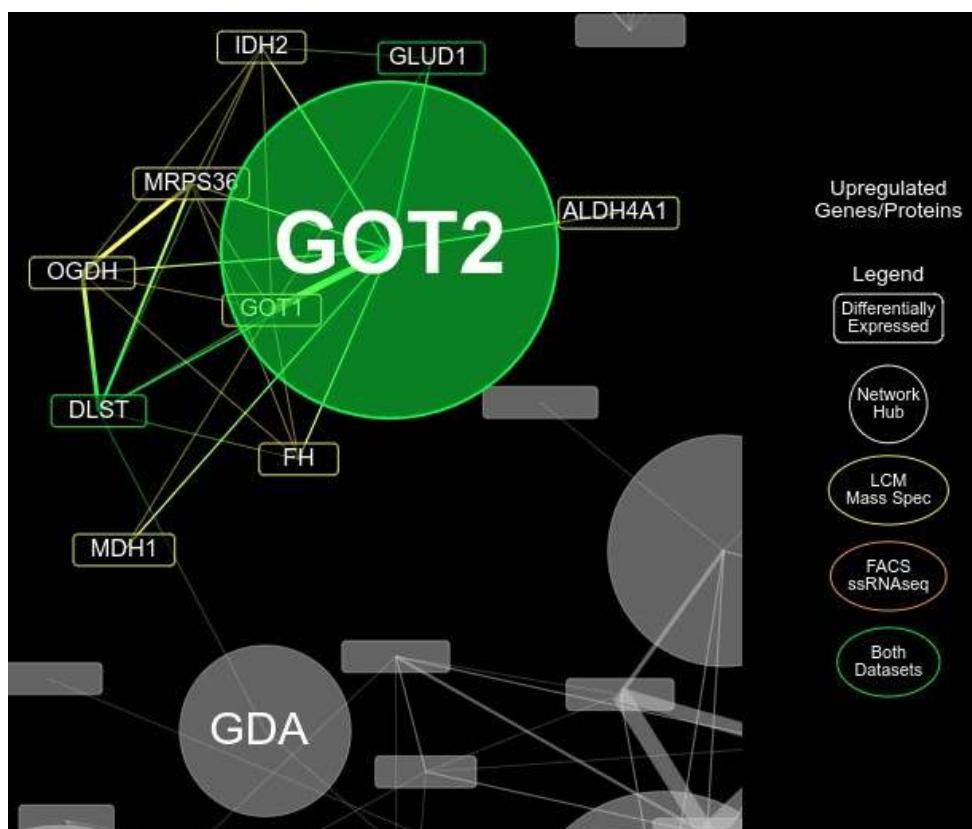


Figure 187: Feature network of GOT2. Figure format previously described in Fig. 140.

Dihydrolipoamide S-succinyltransferase (DLST) is a core component of the 2-oxoglutarate dehydrogenase complex (OGDHc), which catalyzes the conversion of 2-oxoglutarate to succinyl-CoA and CO<sub>2</sub> in the tricarboxylic acid (TCA) cycle (Mellid et al., 2023). Meanwhile, Glutamate dehydrogenase 1 (GLUD1) is a mitochondrial enzyme that catalyses the reversible conversion of glutamate to α-ketoglutarate and ammonia, playing a pivotal role in amino acid metabolism and energy production (Plaitakis et al., 2017).

Since *GLUD1* is integral to neurotransmitter regulation, modulating the primary excitatory neurotransmitter glutamate, and *DLST* contributes to energy metabolism, they both play important roles in synaptic transmission and overall neuronal function. The contribution of *DLST* gene to genetic risk of AD has had inconsistent results in the literature (Matsushita et al., 2001; Sheu et al., 1999), while *GLUD1* has shown more consistency as reviewed in (Mathioudakis et al., 2023).

After pruning, the pathway network showed no data for GOT2. Therefore Figures 188 and 189 instead visualise the all differentially enriched pathways containing GOT2 in each dataset as heatmaps. It is reassuring to observe that all of the pathways converge towards basic metabolic processes and include those related to glutamate, oxaloacetate, and 2-oxoglutarate (Figure 190). Also interesting is that these processes are more associated with the proteomics dataset; a noticeable difference in enrichment levels is observed between the two datasets. This runs counter to the data thus far, where pathway enrichment always had more coverage in the transcriptomics dataset. An NGS dataset will always provide more coverage than mass spectrometry with current technology, which makes that case unsurprising. On the other hand, recall that in the methodology for this analysis, due to the sparsity of the proteomics dataset, gene sets were only required to have half of its features present in the input data. Figure 190 is an example of this, and likely results in the term suggesting catabolism despite the absence of GAD1 and GAD2, which are probably the key features driving the catabolic aspect of the term. While it may indeed be the case that GOT2-related processes take place more on the protein than transcript-level, one should be cautious that this was not driven by methodology decisions that unintentionally inflated the enrichment of some gene sets on the proteomics side. Nevertheless, in theory this should not significantly impact differential enrichment, just the relative expression of some gene sets when comparing the two datasets against each other.

**Figure 188.**

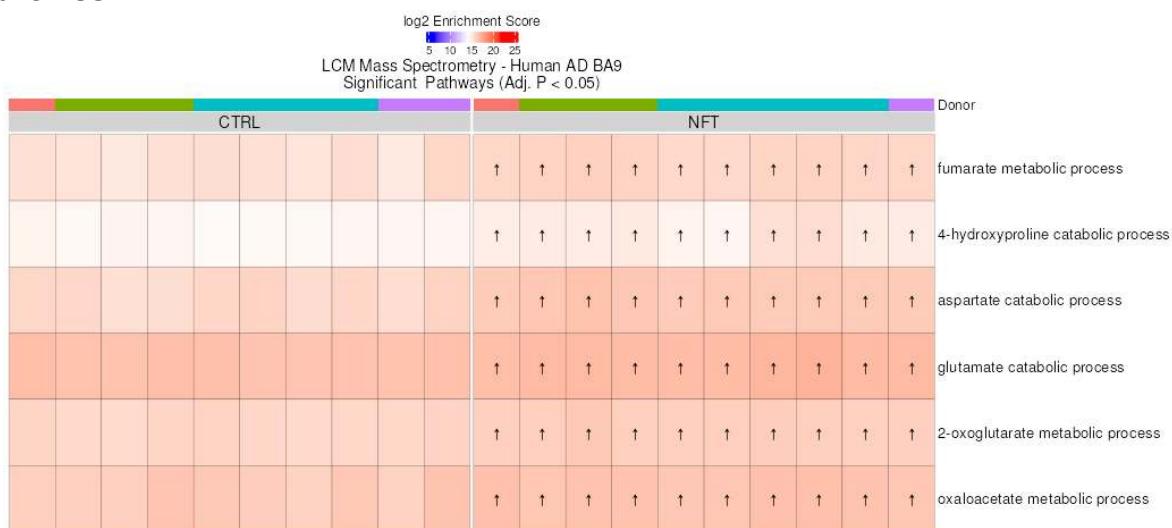


Figure 188: Heatmap showing enrichment of all differentially enriched pathways containing GOT2 in the LCM Mass Spec dataset. Figure format previously described in Fig. 146.

**Figure 189.**

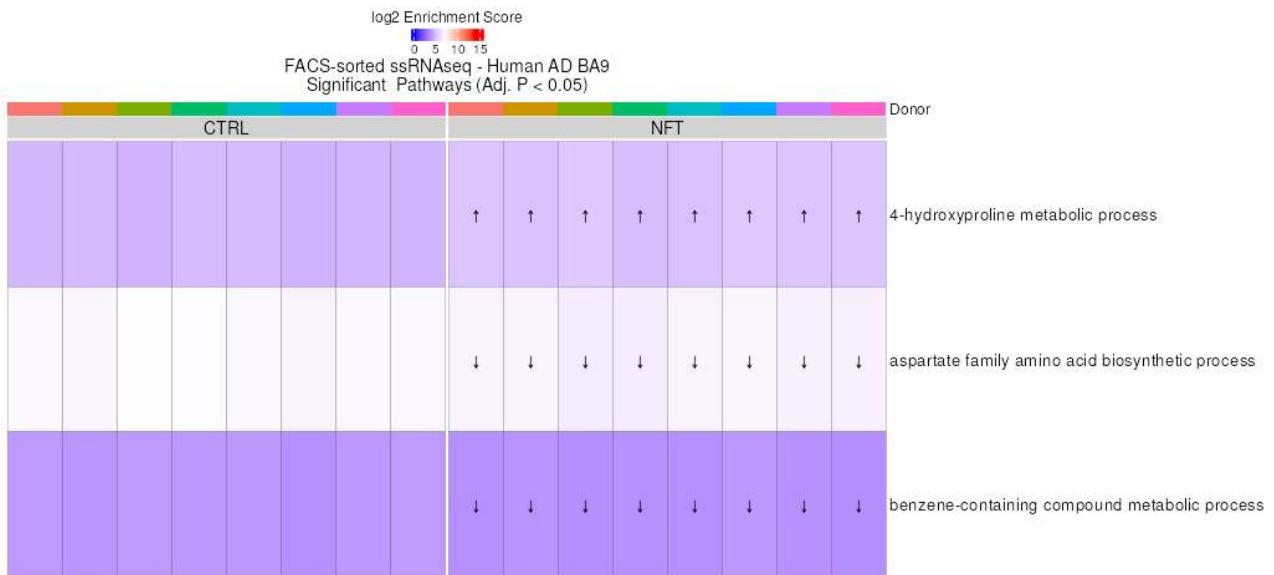


Figure 189: Heatmap showing enrichment of all differentially enriched pathways containing GOT2 in the FACS ssRNAseq dataset. Figure format previously described in Fig. 146.

**Figure 190.**

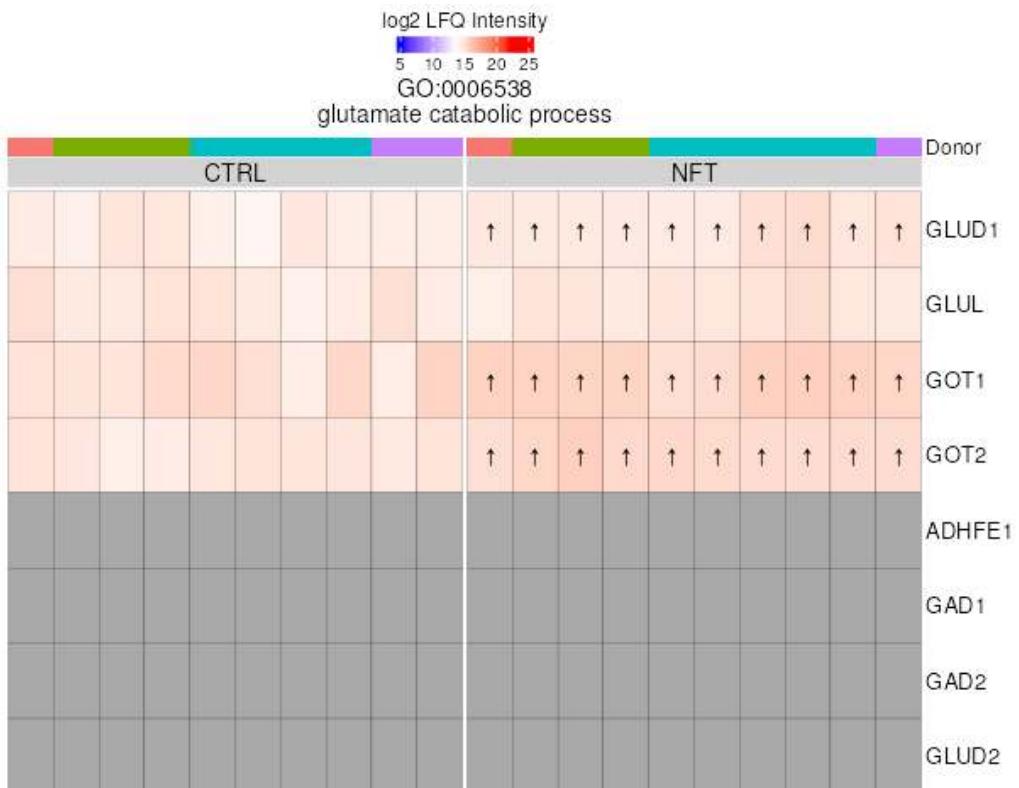


Figure 190: Heatmap showing the unscaled abundance of proteins within the "glutamate catabolic process" pathway in the LCM Mass Spec dataset. Figure format previously described in Figure 135.

## 7. Discussion

### 7.1 ImputeFinder and GeneFunnel Resolve Obstacles in Real-world Analysis

In this work, I develop novel methods for tackling common problems in transcriptomic and proteomic analysis. I explain their methodologies and explore their properties extensively, as well as perform benchmarks in synthetic data, real data, and against other contemporary methods. I show that the new methods are intuitive to reason with and address issues that other methods fail to satisfactorily solve. The methods are performant and were leaders in all metrics defined in the benchmarking experiments. Finally, I utilised the methods for real-world analysis to investigate the molecular changes underlying tangle-bearing neurons in Alzheimer's Disease.

While the transcriptomics dataset (FACS ssRNAseq) was of exemplary quality for re-analysis, the in-house proteomics dataset (LCM Mass Spec) could be considered pilot quality at present. It is in fact the earliest of several similarly designed experiments performed by our group, and recent improvements in the protocol has improved the most recent iterations to a much higher standard than before. With the current dataset, I was forced to remove half of the data due to poor capture of proteins. The remaining samples still suffered from issues related to missing values. This made the dataset a prime candidate for more sophisticated missing value handling using ImputeFinder, previously developed for the analysis of another complex proteomics dataset (Fowler et al., 2025).

ImputeFinder performed exceptionally well in this dataset, resulting in the clean separation of samples by treatment group (tangle-bearing vs. non-tangle-bearing) on a PCA plot (Figure 74). This helped maximise retention of analysable proteins in the already sparse label-free mass spectrometry dataset, allowing for the testing of differential abundance in 665 proteins from the original 1,547. Through the comprehensive exploration and benchmarking process of the method, I could be confident that these proteins were valid to keep and any missing values were imputed in a sensible manner.

GeneFunnel was primarily employed to help manage and interpret the mass output of data that result from omics research. Many genes/proteins (features) participate in a multitude of functional domains and it can be of value to infer as to which of these domains are relevant in an analysis. GeneFunnel helped identify the functionalities of features based on their membership in GO annotations, but more importantly, pinpointed which of those annotations are actually active in the datasets or show differences between groups. This information was further enhanced by coupling it with network analysis, facilitating unique approaches to feature prioritisation. In this work, traditional statistics such as p-value were not the sole criteria, it was combined with metrics such as if a feature had many of its possible GO annotations differentially enriched. This, for example, inferred that not just the feature itself was differentially expressed, but also the totality of its functionalities.

The end result of usage of these methods was the identification of a small set of hub features, with final focus on those upregulated in both datasets, totalling just 8, out of the 1,771 genes and 262 proteins upregulated in total. The methods themselves then allowed in-depth characterisation of these hubs in the context of the dataset, as detailed in Section 6.4. As will be further deduced from the coming sections, the mechanisms exposed by just these 8 hubs cover a wide variety of functional domains, demonstrating that such analysis is not only useful for deciding on individual genes/proteins of focus, but that they can also be used to distil overall themes from the analysis.

Beyond interpretational benefit, both ImputeFinder and GeneFunnel aim to improve the sensitivity of the analysis. In regards to imputation, this is an obvious goal, however, ImputeFinder additionally has specific criteria for when to discard proteins entirely from the analysis. Quality control aside, making such decisions has non-negligible impact on the severity of multiple testing correction, which often plague naive omics analyses. Likewise, by performing a gene set enrichment focused analysis, one can choose to not test features which are not contained in the gene sets, a valid but often overlooked approach for only analysing decently characterised features. Finally, by statistically testing gene sets themselves, it is possible to capture more subtle changes across many features, where few to none of those features would be detected as statistically different individually.

## 7.2 Evidence of Successful Isolation of NFT-bearing Neurons on the Gene and Protein-level

As shown in Figure 133, tangle-bearing neurons compared to within donor non-tangle-bearing neuron controls, exhibit robust changes on the proteomic and transcriptomic-level. In both modalities, nearly half of all tested features were identified as differentially expressed/abundant. This was the case when testing pathways in the LCM Mass Spec dataset as well, though less so in the FACS ssRNAseq dataset (37% differentially enriched). Moreover, 134 out of the 289 differentially abundant proteins were also differentially expressed on the gene-level, showing considerable agreement between the two datasets. In the FACS ssRNAseq dataset, adjusted p-values as low as  $1.08 \times 10^{-12}$  were found (*NNAT*), while in the LCM Mass Spec dataset, adjusted p-values were as low as  $2.57 \times 10^{-9}$  (*MAPT*). This information also suggests that variance between donors was low and that the methods are consistent and reproducible between samples. Note though, that the LCM Mass Spec experiment will benefit from further optimisation, as nearly half of samples had to be discarded due to low or no protein capture (Section 2.5.3). Of the samples kept however, the data was of sufficient quality to achieve strong significance.

It is reassuring in the context of this analysis that *MAPT* is not only differentially expressed in both datasets, but also had the lowest adjusted p-value in the LCM Mass Spec dataset. *MAPT* encodes the tau protein, which is the primary protein that becomes aggregated in neurofibrillary tangles. To find it differentially abundant on the protein-level provides strong evidence that the experiment achieves its goals, that is the separation of tangle-bearing and non-tangle-bearing neurons. Similarly, neurofilaments were also strongly differentially expressed in both datasets, with *NEFM* being selected as a shared hub. In recent years, neurofilaments show great promise for being a biomarker of Alzheimer's Disease (Giuffrè et al., 2023), and in advanced AD, they often co-localise with aggregated tau (Schmidt & Trojanowski, 1990).

Finally, I observed strong differentially enrichment of the pathway "neurofibrillary tangle" in both datasets, which was unbiasedly selected as a pathway hub for the analysis. This pathway is defined by GO as comprised of *MAPT* and several neurofilaments, as well as *CLU* (Clusterin) and *PICALM* (Phosphatidylinositol Binding Clathrin Assembly Protein). The later two features are also well-characterised classical AD risk genes (Carrasquillo et al., 2010). Together, these finds support not only robust differential changes between tangle-bearing and non-tangle-bearing neurons on the protein and gene-level, but also evidence that the cell sorting methods succeeded in isolating the two cell populations.

### 7.3 Similarly Designed Studies Support Findings in This Analysis

The work performed here directly re-analyses an existing transcriptomics dataset (Otero-Garcia et al., 2022) as well as initial analysis on an in-house proteomics dataset. For the in-house dataset, LCM Mass Spec, there is no perfect study for comparison. Other studies besides the one for re-analysis that performed within donor isolation of tangle-bearing neurons all analysed transcripts (Dunckley et al., 2006; Ginsberg et al., 2000), though they did utilise laser-capture microdissection (LCM) like the in-house dataset. However, the work of (Hondius et al., 2021) does use LCM with mass spectrometry to analyse the proteome of tangle-bearing neurons in AD, with the big difference being that it is not a within-donor comparison; non-tangle-bearing neurons were all isolated from non-demented control donors. This not only introduces additional donor variability to the analysis but also now confounds the effect with general processes associated with AD beyond that of neurofibrillary tangles. Moreover, their study was not focused strictly on tangle-bearing neurons, but also granulovacuolar degeneration (GVD), and cells with GVD were collected within the same patient donors as tangles, unlike cells without either feature. Nonetheless, it appears to be the best dataset for comparison with the LCM mass spectrometry dataset in this thesis work.

One of the strongest points of concordance between this analysis and the referenced datasets is the top feature sorted by adjusted p-value. Comparing (Otero-Garcia et al., 2022) is made complicated by the fact that they focused on differential expression in each cell-type separately. While they do present figures of shared differentially expressed genes, it is a collation of cell-types beyond the cell-types chosen in my re-analysis. Moreover, a CSV or Excel file of these cell-types could not be found in the supplemental data. Since the authors do release comprehensive Excel files of differentially expressed genes for each cell-type on an individual basis (Table S6 from their publication), this file was used to compare their analysis with the one in this work. I specifically selected their analysis of their EX2 cell subtype, corresponding to Layer 2-4 CUX+ excitatory neurons in the prefrontal cortex. This cell-type corresponds to the largest group of cell subtypes used in my analysis (Figure 59), so is likely a good compromise.

The top DE gene in EX2 from the analysis by (Otero-Garcia et al., 2022) is *NNAT*, which is the same top DE gene found in my analysis (Figure 138). They report an adjusted p-value of  $1.37 \times 10^{-156}$ , which is substantially inflated from the  $1.08 \times 10^{-12}$  in this thesis work. This results from the use of pseudobulking in my analysis, which facilitates the use of true biological replication (8 donors compared to thousands of cells), a practical solution for avoiding pseudoreplication bias (Zimmerman et al., 2021). The authors did not use pseudobulking, so there is insufficient evidence to suggest that their DE genes, particularly lower ranks ones, will replicate across different patient donors. Nonetheless, it is reassuring that their top DE gene aligns with the top DE genes in the pseudobulked analysis performed here.

Interestingly, despite pseudoreplication bias, the authors report far fewer DE genes than I found, 743 vs. 3,369, both using an adjusted p-value cutoff of 0.05. It is unclear what contributed to this difference, though it can be hypothesised that the pipeline in this thesis work takes substantial efforts to reduce multiple testing burden. The most salient source of this is the subsetting of tested genes to those contained within the Gene Ontology, resulting in the testing of just 8,950 genes compared to the original matrix containing tens of thousands of genes. I could not find data indicating the universe of genes tested by (Otero-Garcia et al., 2022), but it is possible that it was much larger. The approach elected in this thesis work comes with the caveat of removing genes that may have been DE, even strongly DE, just because they are not in the Gene Ontology. Therefore less characterised genes, which may indeed be great candidates for further work, are excluded in favour of greater statistical power for better characterised genes. These differences may reflect in differences in the overlap of DE genes in the analysis by (Otero-Garcia et al., 2022) and my analysis. I report that 556 genes overlap, which means that 187 genes in the (Otero-Garcia et al., 2022) dataset were either not analysed in my analysis or did not reach significance, perhaps due to reduced power from pseudobulking. Nonetheless, this overlap shows that the majority of their DE genes align with mine, regardless of the substantial differences in methodology.

Regarding the analysis by (Hondius et al., 2021) and my analysis on the in-house LCM Mass Spec dataset, the top protein also matches. Crucially, this top protein is MAPT in both analyses, providing confidence that both datasets successfully isolated tangle-bearing neurons and that protein expression of MAPT itself is a good marker for evidence of this. They report an adjusted p-value of  $5.05 \times 10^{-60}$  for MAPT vs.  $2.57 \times 10^{-9}$  in the in-house dataset. There are several clear reasons for this. Their dataset had much larger sample size, having an N of 12, while the in-house dataset only had an N of 4 after removal of low-quality samples. As reported in Section 2.5, the in-house dataset had substantial quality issues, which while addressed to the best of my ability, may have still manifested in the remaining samples. One readily observed way is that by the end of pre-processing, samples had a variable number of technical replicates. While the function *duplicateCorrelation* provides handling of this, it is nonetheless a less-than-ideal situation. In any case, the differential abundance pipeline may have salvaged the data substantially, as my analysis discovered 262 differentially abundant proteins (DAPs AKA DEPs) with an adjusted p-value cutoff of 0.05. With that same cutoff, the authors of (Hondius et al., 2021) found 197 DAPs. It is unclear what their universe of proteins for testing were, but they report the quantification of 2,596 proteins. In my analysis, I tested 665 proteins and like the transcriptomics analysis, reduction of multiple testing burden may have played a key role in improving power of the analysis. The overlap between the two analyses is 151 DAPs, which like the transcriptomics analysis, shows that the majority of their DAPs align with those in the in-house dataset and associated analysis. These analyses in the transcriptomic and proteomic datasets therefore strongly suggest that the work performed in this thesis align well with existing research in a similar area, despite large differences in methodology.

## 7.4 Caveats and Interpretational Considerations

It is important to consider some caveats and interpretation considerations associated with the datasets and analysis at hand. Tangle-bearing neurons were isolated on the basis of Anti-Phospho-Tau AT8 immunohistochemical staining. The AT8 antibody is a widely used monoclonal antibody with high specificity (D. Li & Cho, 2020) that specifically recognises tau protein phosphorylated at residues Ser202 and Thr205 (Goedert et al., 1995). AT8 detects early-stage tau phosphorylation events, which may or may not progress to form insoluble aggregates characteristic of advanced tauopathies. Therefore, AT8 positivity alone does not confirm the presence of mature neurofibrillary tangles. Follow-up work with other antibodies may be beneficial to confirm the validity of this work, as well as provide further precision as to what is being captured during the neuronal isolation process. It may also be insightful to investigate other varieties of tau-related pathology in AD, such as dystrophic neurites, known for co-localising with amyloid plaques as well as hyperphosphorylated tau (Moloney et al., 2021). Related to this is the notable under-representation of differential enrichment of explicit cell death pathways, both in the analysis of this work and the original analysis of re-analysed data (Otero-Garcia et al., 2022). It is not out of the question that indeed the neurons captured are those that are more resilient to cell death or pathology as a whole, and thus able to be isolated and profiled. It is a significant technical challenge, but would be of great benefit if future methods can somehow facilitate profiling of neurons that are known to die at a later time point but before their death. A future technology one might envision is a minimally invasive live imaging protocol that can sample transcript expression and protein abundance. There are reports of proof-of-concepts of such technology (W. Chen et al., 2022), but further review is needed to ascertain the eventual viability of this direction.

The use of gene set enrichment in this analysis provides various benefits in terms of sensitivity to subtly altered pathways, multiple testing correction burden, and prioritisation and interpretation, as discussed extensively throughout Section 4. Here, I wish to highlight some potential pitfalls as well that pertain to interpretation of the results on real-world data. First and foremost, while I attempt to explore the properties and benchmark GeneFunnel extensively, it is a part of the novel work of this thesis, and has not yet undergone peer-review. Indeed, in a real-world situations, where compromises are necessary, GeneFunnel can produce unintended results. In Figure 188 and 190, one can observe that the gene set “glutamate catabolic process” is highly enriched but half the proteins were undetected, or very lowly abundant. These proteins include GAD1 and GAD2, crucial towards the “catabolic” aspect of the gene set. When possible, I prefer to only analyse gene sets for which data of all its features can be found in the input data, like the FACS ssRNAseq dataset. But the sparsity of the LCM Mass Spec dataset led to the decision to only require that half the proteins are present in a gene set, leading to some edge cases such as this. A related issue is the dependency of these analyses on the contents of the gene sets. I used the Gene Ontology, which is both non-disease specific and generally curated by non-experts on the huge swathes of biology it covers. Even if a gene set enrichment method is accurate, the sets used in the analysis may not always be.

## 7.5 Emergent Themes of Tangle-bearing Neuron Pathophysiology

This study presents a within-donor comparison of transcriptome and proteome in tangle-bearing (AT8+) neurons and adjacent AT8- neurons from Alzheimer's disease brains. The novelty lies in the joint analysis of RNA-protein concordance within the same cellular and pathological context, rather than across unmatched donors or bulk tissue. Pairing AT8+ and AT8- neurons from the same case reduces inter-individual and cellular composition confounding and allows the observed signal to be attributed to the presence of neurofibrillary tangles.

A further contribution is methodological. A single analysis framework integrated both modalities so that gene and pathway level inferences are directly comparable. Quality control, statistical testing, and gene set scoring used the same procedures in RNA and protein data, and quantitative criteria was established (i.e. network analysis) to measure agreement and divergence between the two data types. Rigour was enforced by addressing issues like pseudoreplication through pseudobulking, and multiple testing was controlled by taking into account both assays in tandem.

To organise the breadth of differential expression and gene-set results, an interactive network representation was used to summarise relationships between features and associated pathways that move in tandem with disease state, with fine-tuned filtering procedures to move between broader and more selective views. The integration of these networks, as well as interactive heatmaps, into publicly accessible web viewers helps encourage the dissemination and reuse of data in easily explorable formats.

Within this experiment, a compact set of hubs, NEFM, APP, SQSTM1, HSP90AA1, YWHAE, WASF1, CNTNAP1 and GOT2, showed consistent upregulation in AT8+ neurons relative to nearby AT8- neurons in both RNA and protein. The recurrence of these features across donors and analyses supports them as robust characteristics of the tangle-bearing state rather than technical or study-specific effects. The focus on a small, reproducible set is deliberate and intended to aid interpretation and downstream validation.

Practically, the comparison returns two products for use beyond this study. First, a set of features observed in both datasets within the same donors and cellular context. These can be taken forward for confirmation in adjacent material by immunolabelling, targeted proteomics, and other orthogonal assays, and serve as a foundation for pathways that were also observed enriched. Second are findings confined to one dataset or another. These frame more nuanced questions regarding the relationship between transcriptional programmes and protein accumulation in the context of disease.

In the sections that follow, I organise the results into discrete themes and relate them to existing literature. I do not impose a definite causal sequence, as the present data does not have the capability to resolve directionality, and hypotheses which remain speculative are made clearly distinguishable from the descriptive results.

### 7.5.1 Co-aggregation of Neurofilaments and Microtubule Destabilisation

NEFM (neurofilament medium polypeptide) was shown to have increased gene expression and protein abundance in the comparison of tangle-bearing neurons with non-tangle-bearing neurons. Neurofilaments are a class of intermediate filament proteins that contribute to the support of microtubules, a key component of the structure of axons (A. Yuan & Nixon, 2021). Along with tau, neurofilaments bind to the outer perimeter of microtubules to help maintain axonal caliber. In disorders that affect microtubule integrity such as Alzheimer's Disease, neurofilaments and tau become mislocalised. Tau commonly aggregates while neurofilaments have a tendency to fragment. A high-level schematic of these roles can be seen in Figure 191.

**Figure 191.**

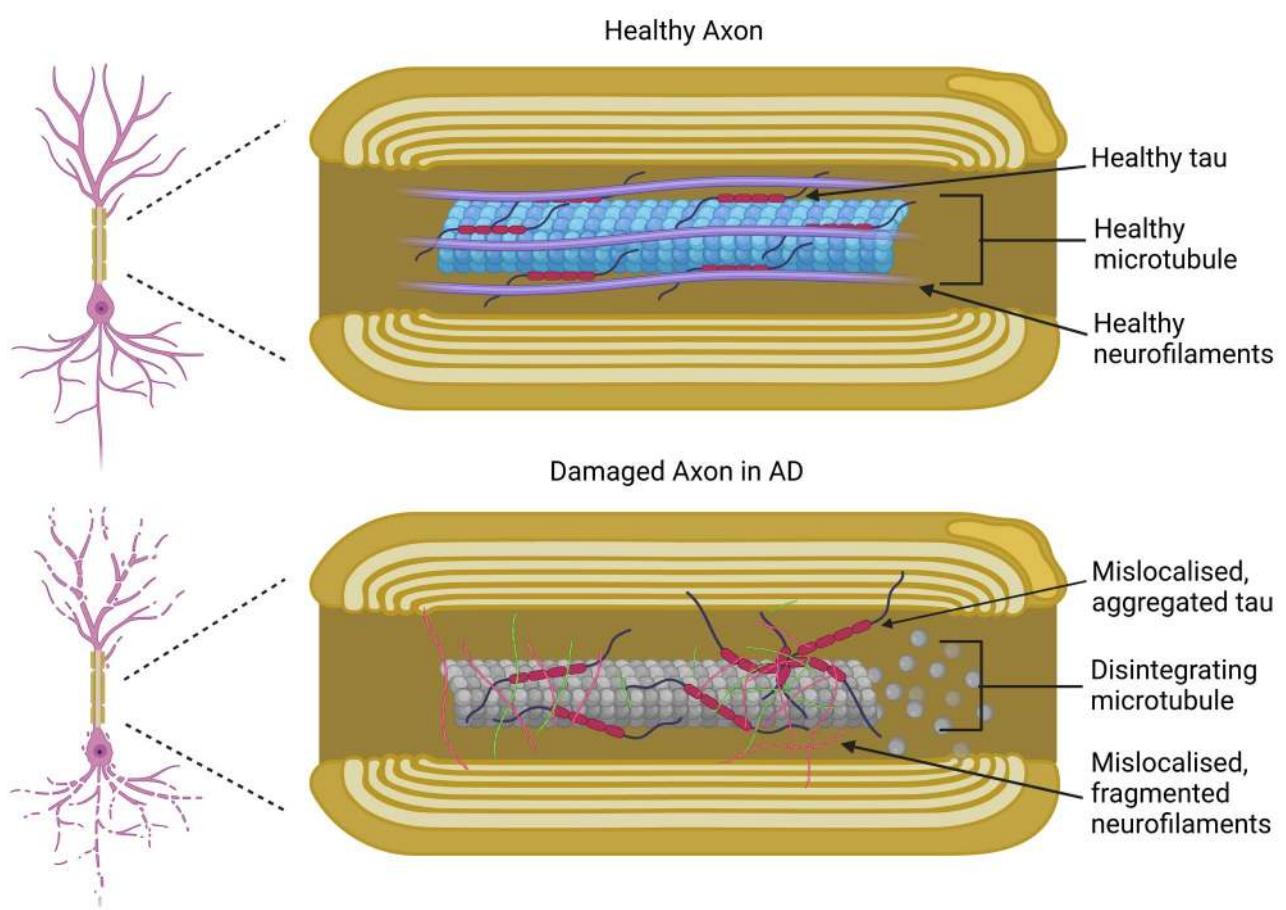


Figure 191: Proposed disease mechanism of neurofilament overabundance in tangle-bearing neurons based on the results of the analysis. Neurofilaments mislocalise alongside tau, leading to microtubule instability and disintegration. The mislocalisation and fragmentation of neurofilaments may also promote their escape from the axonal space, resulting in increased detection in CSF and blood plasma. Furthermore, fragmentation may result in neurofilament dyshomeostasis, an adverse condition where shorter chains become the predominant species. Created with BioRender.com.

The detection of neurofilaments in neurofibrillary tangles date as far back as 1979, when (Ishii et al., 1979) showed immunofluorescence co-localisation in the AD hippocampus. This finding has faced scepticism over the years; in 1987 it was suggested that neurofilament antibodies are cross-reactive with PHF-tau (Nukina et al., 1987), and in 2003, a study could not replicate direct co-localisation with tangles and furthermore reported that neurofilament aggregates were only found in a subset of PHF-tau-positive neurons in the EC (Porchet et al., 2003). Nonetheless, other groups have replicated the original finding by (Ishii et al., 1979), and have further revealed that neurofilaments are phosphorylated in AD (Haugh & Probst, 1986), and that they may mark selectively vulnerable neurons (J. H. Morrison et al., 1987). Additionally, monoclonal antibodies against NEFM and NFH were shown by (Rasool et al., 1984) to strongly label nearly all neurofibrillary tangles in AD cortical tissue. They also performed a biochemical extraction to remove neurofilament, finding that most isolated NFTs still retain tau immunoreactivity while largely losing neurofilament antibody reactivity. They concluded that neurofilament content in tangles appear to represent partial fragments or cross-linked epitopes rather than a core structural component. Consistent with this, a quantitative analysis found that NFTs are composed predominantly of tau protein and that only restricted segments contained NEFM and NFH. Furthermore, tau epitopes far outnumbered neurofilament epitopes in tangles (Schmidt & Trojanowski, 1990). The authors showed as well that neurofilament markers tend to co-localise with tau lesions especially in advanced stages of tangle accumulation, suggesting that as cytoskeletal degeneration progresses, neurofilament proteins increasingly become entrapped in or co-aggregate with tau filaments. Finally, though it is inferential evidence, in this thesis work, neurofilaments were detected as differentially enriched alongside associated members of the GO term “neurofibrillary tangle” (Figure 155).

Regardless of controversies regarding the direct co-localisation of neurofilaments with neurofibrillary tangles or other pathological features, its accumulation is undisputed in Alzheimer’s Disease as well as a range of other neurodegenerative diseases including ALS and Parkinson’s Disease (Q. Liu et al., 2011). (Vickers et al., 2016) provides a great review of the mechanistic contributions of neurofilaments to AD, particularly within the scope of cytoskeleton dysfunction in axons, as depicted in Figure 191. Neurofilaments are described by the triplet proteins NEFL, NEFM, and NEFH (light, medium, and heavy) based on the length of extension of the C-terminal tail domain. They are specific to neurons and closely integrate with axons and myelinating glia to structure the cytoskeleton and facilitate axonal transport. Phosphorylation state is tightly regulated to enable timely and precise control of axon calibre (Brown, 1998). The high reliance of axonal integrity on neurofilaments mean that neurofilament disruption results in an observable loss of microtubule structure (King et al., 2001). This is alongside a collapse in filamentation and formation of filamentous aggregates (Siedler et al., 2014). The loss of microtubule support is particularly detrimental because microtubules serve as tracks for axonal transport, and without them, the movement of proteins, vesicles, and organelles are severely impaired. This underlie observations that tangle-bearing neurons show accumulations of cargo (e.g.,

amyloid precursor protein, mitochondria, and neurofilaments themselves) in the soma (Stamer et al., 2002).

Neurofilaments have also seen rising popularity as a biomarker for neural injury and in AD they are elevated in the CSF and blood plasma (Giuffrè et al., 2023). It has been proposed that during axonal injury, neurofilaments escape into the extracellular space, resulting in its detection in locations peripheral to the CNS (Khalil et al., 2024). Through autophagy and proteolysis, neurofilaments are known fragment into a variety of degradation products (A. Yuan et al., 2017), and this may contribute to neurofilament translocation. It is possible that this may have contributed to low detection of NEFH species on the protein-level in this thesis work, though this does not explain the low expression of the *NFH* gene as well. Regarding interpretation of the analysis in this work, it is insightful that in NEFM (as well as NEFL) upregulation is taking place on both the protein and gene-level. An increase in differential gene expression suggests that neurofilament gene expression itself may be a contributing factor to neurofilament aggregation. On the other hand, a lack of differential expression would suggest aggregation occurs as a direct consequence of external factors, such as tau or amyloid pathology. Upregulation of neurofilament gene expression may also be a compensatory response due to axonal injury, as has been suggested (H. Wang et al., 2012). This is supported by separate studies in AD model mice that knocked down Nfl, showing an increase in the AD-like phenotype (Fernandez-Martos et al., 2015; Weston et al., 2017). Little direct research on neurofilament gene expression in humans could be found, though work from 1994 reports large reductions in NEFM and NEFL (Kittur et al., 1994). Though the evidence supports the thesis that elevated neurofilament protein abundance may signal co-aggregation with tau and microtubule/cytoskeletal dysfunction, more work is needed to elucidate the role of neurofilament gene expression in disease and physiology.

### 7.5.2 Potentially Protective Role of the Non-Amyloigenic Pathway

AD is characterized by the accumulation of extracellular amyloid-beta (A $\beta$ ) fibrils in plaques in addition to the intracellular tau neurofibrillary tangles focused on in this research. Rather than the proliferation of plaques themselves however, and besides tau pathological staging, multiple studies have identified the loss of synapses and associated synaptic dysfunction as pathological changes that closely correlate with cognitive decline in AD (Rajmohan & Reddy, 2017; Sirisi et al., 2024). The role of amyloid plaques on synaptic integrity therefore represents an indirect link with disease severity (H. Zhang et al., 2022). In this thesis research, I observed an upregulation of amyloid precursor protein (APP) on the transcript and protein-level. While much has been publicised regarding APP's contribution to synaptic dysfunction, I wish to also discuss underappreciated roles of APP in neuroprotection, including synapse support. This is largely dependent on whether APP is processed through amyloigenic or non-amyloigenic pathways, respectively (Figure 192). Although the datasets in this analysis lack sufficient information for determining the predominant pathway, I speculate that non-amyloigenic may be more active on the basis

that amyloid plaques are not conventionally found co-localised with neurons bearing neurofibrillary tangles.

**Figure 192.**

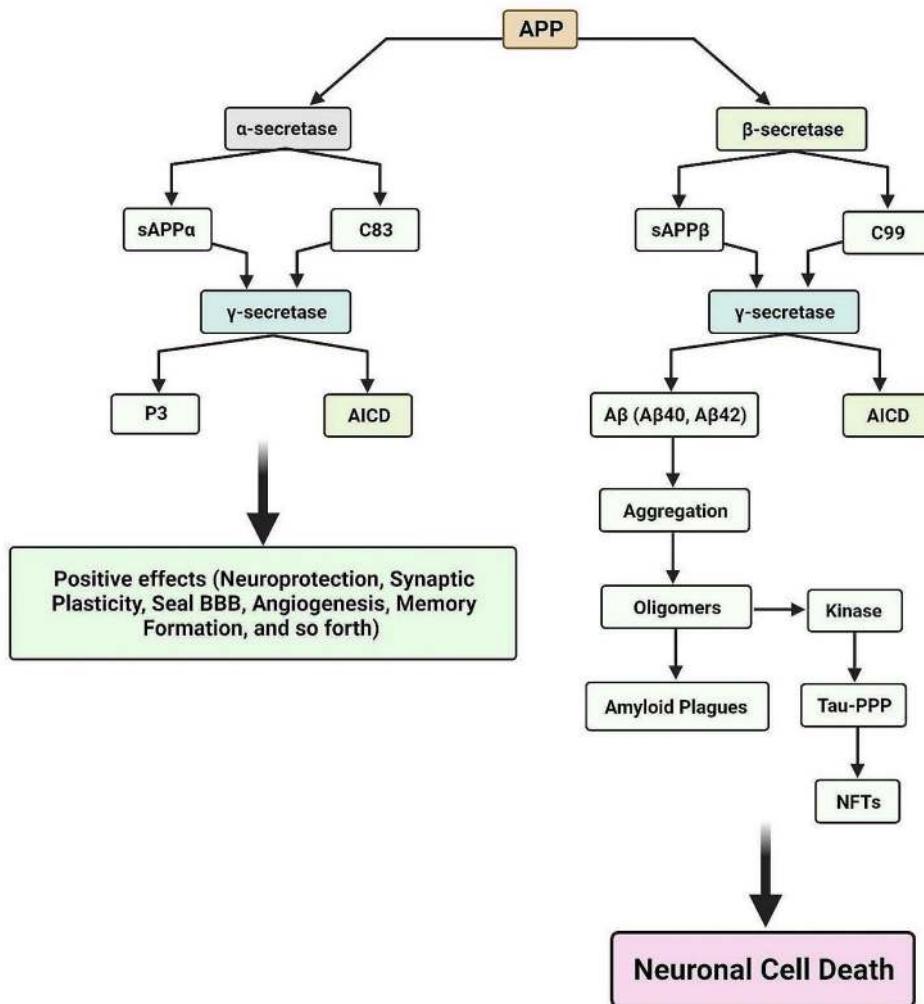


Figure 192: APP can be processed through two main pathways, non-amyloigenic (left) and amyloigenic (right). Often underappreciated, there is substantial research indicating that the non-amyloigenic pathway confers significant positive effects for neural health. APP was found upregulated on both the gene and protein-level in the present analysis of neurofibrillary tangle-bearing neurons, but amyloid plaques are not conventionally found co-localised to NFTs, suggesting that the non-amyloigenic pathway may be the predominant pathway at play. Figure reproduced from (Azargoonjahromi, 2024).

The amyloidogenic pathway involves sequential processing of APP by β-secretase and γ-secretase. Initially, β-secretase cleaves APP to produce soluble APP beta (sAPP $\beta$ ) and a membrane-bound C-terminal fragment known as C99 (CTF $\beta$ ). Subsequently, γ-secretase cleaves C99 within the cell membrane, resulting in the release of the APP intracellular domain (AICD) and amyloid-beta (A $\beta$ ) peptides, primarily A $\beta$ 40 and A $\beta$ 42 (Rodríguez-Manotas et al., 2012). The accumulation of A $\beta$  in AD can activate kinases such as

GSK-3 $\beta$ , CDK5, and MAPKs, leading to abnormal phosphorylation of tau protein and its subsequent aggregation into NFTs. Additionally, disruption of phosphatases, enzymes that remove phosphate groups, can further contribute to tau hyperphosphorylation. (H. Zhang et al., 2021).

The non-amyloidogenic pathway prevents the formation of A $\beta$ . Here,  $\alpha$ -secretase cleaves APP within the A $\beta$  region, producing soluble  $\alpha$ -APP fragments (sAPP $\alpha$ ) and a membrane-bound fragment known as C-terminal fragment alpha (CTF $\alpha$  or C83). This C83 fragment is further cleaved by  $\gamma$ -secretase, generating non-toxic P3 peptides and the APP intracellular domain (AICD). Through this pathway, APP is processing without producing harmful products, reducing the potential for amyloid pathology (Nhan et al., 2015).

As reviewed in (Azargoonjahromi, 2024), growing evidence suggests that under certain conditions, A $\beta$  can exert beneficial physiological functions, including neuroprotection, antioxidation, and trophic support. Human A $\beta$  peptides have been shown to decrease apoptosis when introduced to neuronal cell cultures (Chan et al., 1999). And in cultured cortical neurons, inhibition of  $\beta$ - and  $\gamma$ -secretases or treatment with antibodies that aggregate A $\beta$  leads to decreased cell survival, an effect that is fully reversed upon supplementation with A $\beta$ 1–40 (Plant et al., 2003). Furthermore, In a study using neural stem cells (NSCs), oligomeric A $\beta$ 1–42 was observed to enhance the survival and differentiation of NSCs from the striatum and hippocampus (Lopez-Toledano, 2004). Interestingly, this outcome did not occur upon exposure to either A $\beta$ 1–40, A $\beta$ 25–35, or their fibrillar peptide forms.

In the work in this thesis, the top enriched pathway associated with APP, aside from those directly related to A $\beta$ , was "positive regulation of long-term synaptic potentiation" (Figure 162). A $\beta$  peptides have been shown experimentally to induce long-term potentiation (LTP), for example after application of low concentrations of A $\beta$  1–42 into hippocampal slice preparations in mice (Puzzo et al., 2008) and rats (J. Wu et al., 1995). Potential mechanisms underlying these effects include an increase in acetylcholine release into synapses and heightened probability of postsynaptic neuron depolarisation, thereby promoting synaptic strengthening (Q. Huang et al., 2022). Evidence points towards these neuroprotective mechanisms as being mediated via NMDA receptors rather than AMPA receptors (J. Wu et al., 1995). Additionally, A $\beta$  1–40 has been implicated in promoting synaptic plasticity through modulation of cholesterol dynamics within neuronal membranes (Koudinov & Koudinova, 2003). These findings suggest that there is a potential for APP upregulation in tangle-bearing neurons to manifest as a protective or compensatory response, dependent on the activity of the non-amyloigenic pathway over the amyloigenic pathway. Though this cannot be confirmed with the data at hand, the improbability of co-localisation of amyloid plaques suggests that this is not improbable. Well designed experiments to gauge the activity of the competing pathways will be informative to better understand the interplay of APP within NFT environments. It is likely also valuable to further investigate genes/proteins co-upregulated with APP in the enriched synaptic potential-related gene sets (Figure 163).

### 7.5.3 p62 Accumulation and Dysregulation of Autophagy

In my analysis I report an increase in gene expression and protein abundance of SQSTM1/p62 in neurofibrillary tangle-bearing neurons compared to tangle-free neurons. Mechanistically, this elevation of p62 likely reflects a cellular response to proteostasis stress. p62 is a selective autophagy receptor that normally binds polyubiquitinated proteins and organelles and delivers them to autophagosomes via its LC3-interacting region (Kraft et al., 2016). In healthy conditions, p62 is continually turned over by autophagy; thus an accumulation of p62 often indicates impaired autophagic flux (Blaudin De Thé et al., 2021). The high p62 levels in tangle-bearing neurons could therefore signify that the autophagy-lysosomal pathway is overwhelmed or stalled, leading to p62 protein being stabilised and accumulated rather than degraded. Recent mechanistic work has illuminated how an overload of p62 on tau fibrils can actually impede degradation. Tau fibrils heavily coated by p62 fail to recruit other crucial autophagy adapters like TAX1BP1, thereby stalling autophagosome formation and cargo turnover (Ferrari et al., 2024). In other words, while p62 docks onto the tangles, the downstream steps of autophagy may not fully engage. A schematic of how this situation may pertain to neurofibrillary tangles is shown in Figure 178.

**Figure 193.**

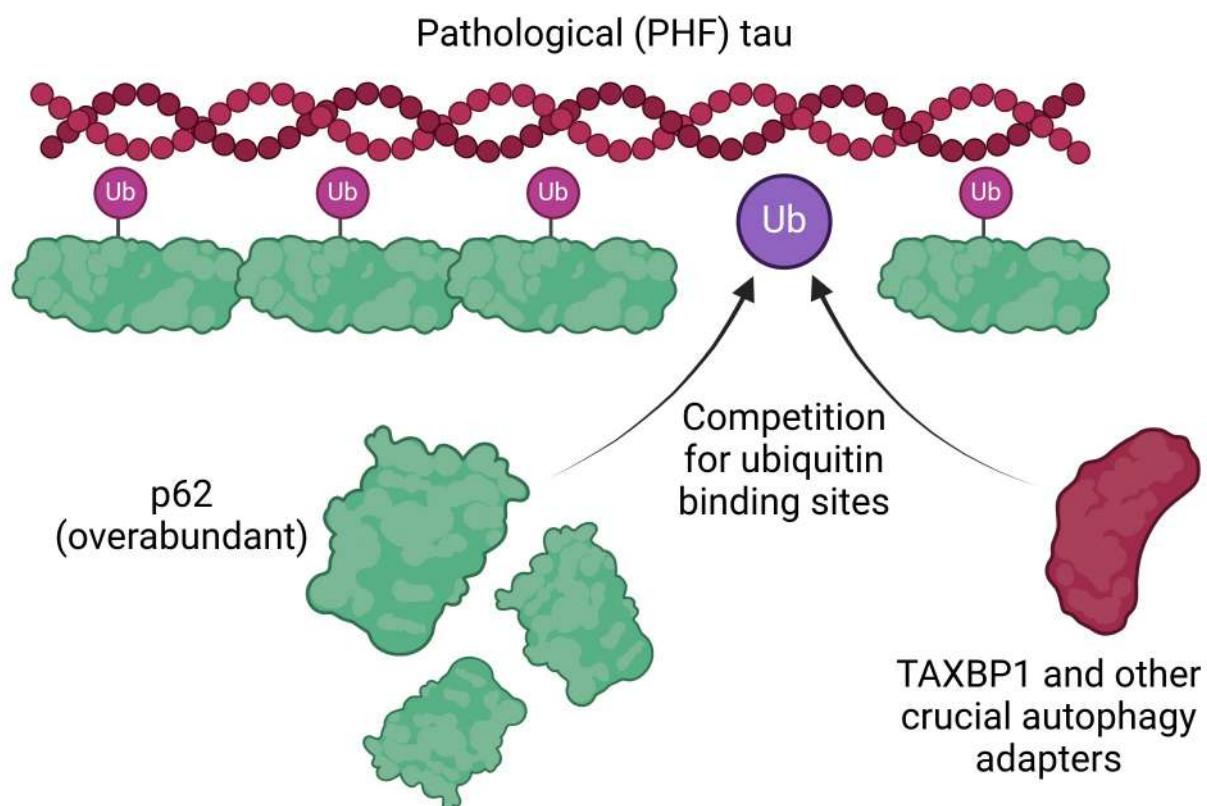


Figure 193: Proposed disease mechanism of p62 upregulation in tangle-bearing neurons based on the results of the analysis. p62 accumulation may lead to its overabundance

relative to other key factors, like TAXBP1, that are essential to forming a functional autophagosome for autophagic breakdown of toxic species such as pathological tau forms. p62 competes with these factors for ubiquitin binding sites on the autophagy target and p62 overabundance may saturate and over-compete for these sites. Created with BioRender.com.

TAXBP1 (AKA TAX1BP1), was not within either dataset for statistical testing, indicating that it was too lowly present for analysis. It should not have been removed as a result of filtering on Gene Ontology members, as it had been annotated by the group. Nevertheless, the related gene *TAX1BP3* was detected as lowly upregulated in tangle-bearing neurons in the FACS ssRNASeq dataset (adjusted p-value < 0.0231). Considering the greater level of differential expression of *SQSTM1* (adjusted p-value < 0.00675) and differential protein abundance (adjusted p-value < 0.00117), this may highlight a severe imbalance between p62 and other co-factors in NFT disease response. It would be interesting to explore other co-factors in this dataset to see if any have comparable effect sizes as p62. Experimentally, one may overexpress or introduce co-factors exogenously to an NFT disease model to see if it can rescue the phenotype by better balancing the ratio of p62 to other autophagy adapters.

p62's role in tangle-bearing neurons appears to be directly tied to its ability to recognise and bind tau aggregates. In AD brain tissue, p62 has been found to strongly bind to NFTs composed of hyperphosphorylated tau (Kuusisto et al., 2002). Through its ubiquitin-binding UBA domain, p62 can attach to ubiquitinated tau species and tether them to the autophagy machinery via LC3 binding (Babu et al., 2005), suggesting that p62 is actively attempting to target tau aggregates for degradation. In the tangle-bearing neurons of this analysis, the co-localisation of protein p62 with tau pathology likely represents this effort to clear tau. At the transcript level, increased *SQSTM1* mRNA might be a compensatory upregulation, potentially driven by stress-responsive transcription factors (e.g. via NRF2), as the neuron attempts to boost its clearance capacity. However, the persistence of NFTs despite p62 enrichment indicates a breakdown in the clearance process. One other possibility besides the imbalance of co-factors is that tau aggregates trap p62 in an autophagy-incompetent state. Prior studies have shown that in AD, p62 becomes sequestered within NFTs, which may reduce the pool of functional p62 available in the cytosol (Du et al., 2009). In summary, elevated presence and expression of p62 may signal a genuinely functional autophagy response but may also be a sign of over-accumulation that ultimately impedes rather than rescues neurons.

#### 7.5.4 Chaperone Co-factors and the Dual Roles of the HSP90 Complex

Heat shock protein 90 alpha (HSP90AA1) is a ubiquitous ATP-dependent chaperone that plays a central role in the cellular misfolded protein response (Ou et al., 2014a). HSP90AA1 directly binds tau and influences its folding state and stability. It interacts across broad regions of the tau molecule, including aggregation-prone domains (Shelton et al., 2017), forming a complex that can either refold tau or hold it in a soluble state. The

analysis in this thesis reports an upregulation of HSP90AA1 on both the transcriptomic and proteomic-level in tangle-bearing neurons. In experimental models, elevating Hsp90 (and co-chaperone Hsp70) levels promotes tau solubility and enhances tau's binding to microtubules, leading to a reduction in insoluble, aggregated tau (Dou et al., 2003). This chaperoning activity also correlates with lower tau hyperphosphorylation. Conversely, pharmacological inhibition of HSP90 (e.g. geldanamycin) has been shown to promote the reduction of disease-associated tau (Opattova et al., 2012). Another study likewise showed that tau protein interaction with Hsp90 promotes its assembly into filamentous aggregates (Tortosa et al., 2009). The elevation and reduction of HSP90 appears to have contradictory effects on tau pathology in the literature, but this may in fact be due to variation in the activity of interacting partners with HSP90. As shown in Figure 194, HSP90 can adopt different conformation depending on these partners, which can have widely different effects on HSP90 function.

**Figure 194.**

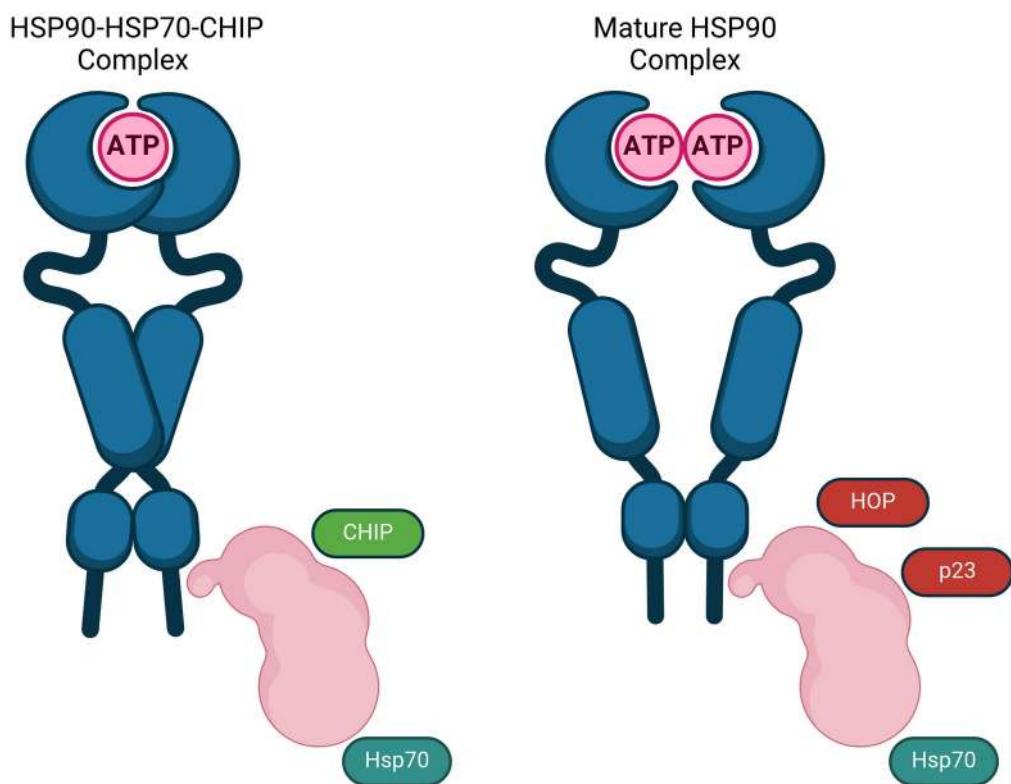


Figure 194: Proposed disease mechanism of p62 upregulation in tangle-bearing neurons based on the results of the analysis. Dependent on if HSP90 is bound to the CHIP complex or the HOP-p23 complex (forming what is known as the Mature HSP90 Complex), HSP90 carries out different functions on its protein target, either breaking them down or refolding them. This has been shown experimentally to have significant impacts on the interaction between HSP90 and tau. Created with BioRender.com.

An important function of HSP90AA1 is the triaging of misfolded proteins for degradation via the ubiquitin-proteasome system. HSP90 serves as a scaffold that, together with Hsp70 and the E3 ubiquitin ligase CHIP, can target aberrant tau for proteasomal destruction (Dickey et al., 2007). In fact, the HSP90-HSP70-CHIP complex specifically recognises phosphorylated, pathogenic tau species and tags them with ubiquitin for clearance (Nadel et al., 2023). Essentially, co-chaperones may determine whether tau is refolded or degraded when targeted by HSP90. A review by (Ou et al., 2014b) postulates that recruitment of CHIP and related factors favours tau disposal, whereas assembly of the so-called mature HSP90 folding complex (which includes adaptors like HOP and p23) can prevent tau degradation. In AD, an imbalance in these co-chaperone interactions could impair the efficient proteasomal clearance of tau. For instance, if HSP90 remains engaged in a refolding mode with tau (possibly due to an excess of mature co-chaperones), it may inadvertently shield tau from ubiquitination, allowing pathological forms of tau to persist. The intricacies of HSP90 triage highlights HSP90AA1's dual role, it can either rescue tau pathology or potentiate it, and simply assessing its gene or protein levels in isolation may be insufficient for understanding its role in disease. Deeper analysis of the pathway-level data produced by GeneFunnel may guide directions for follow-up work, such as investigation of factors co-regulated with HSP90AA1 like those seen in Figure 172.

### 7.5.5 Sequestration of Tau Dephosphorylating Phosphatases

YWHAE (14-3-3 $\epsilon$ ), shown as upregulated in tangle-bearing neurons in my analysis, plays a significant role in modulating the subcellular localisation of a diverse range of binding partners, thereby influencing their function and activity (Foote & Zhou, 2012). 14-3-3 $\epsilon$  exerts its effects on protein localisation primarily through phosphorylation-dependent binding. By recognising specific phosphoserine or phosphothreonine motifs on target proteins, 14-3-3 $\epsilon$  can induce conformational changes that inhibit localisation signals, such as nuclear localisation signals (NLS) or nuclear export signals (NES), effectively sequestering these proteins in-place or into particular cellular compartments. This mechanism ensures that proteins are localised appropriately in response to various cellular signals, maintaining cellular homeostasis. Interestingly, it has been proposed that this sequestration activity may be active against tau phosphoryl residues, effectively preventing potentially protective dephosphorylating phosphatases from carrying out their functions (Sluchanko & Gusev, 2011). A visualisation of this scenario is shown in Figure 195.

**Figure 195.**

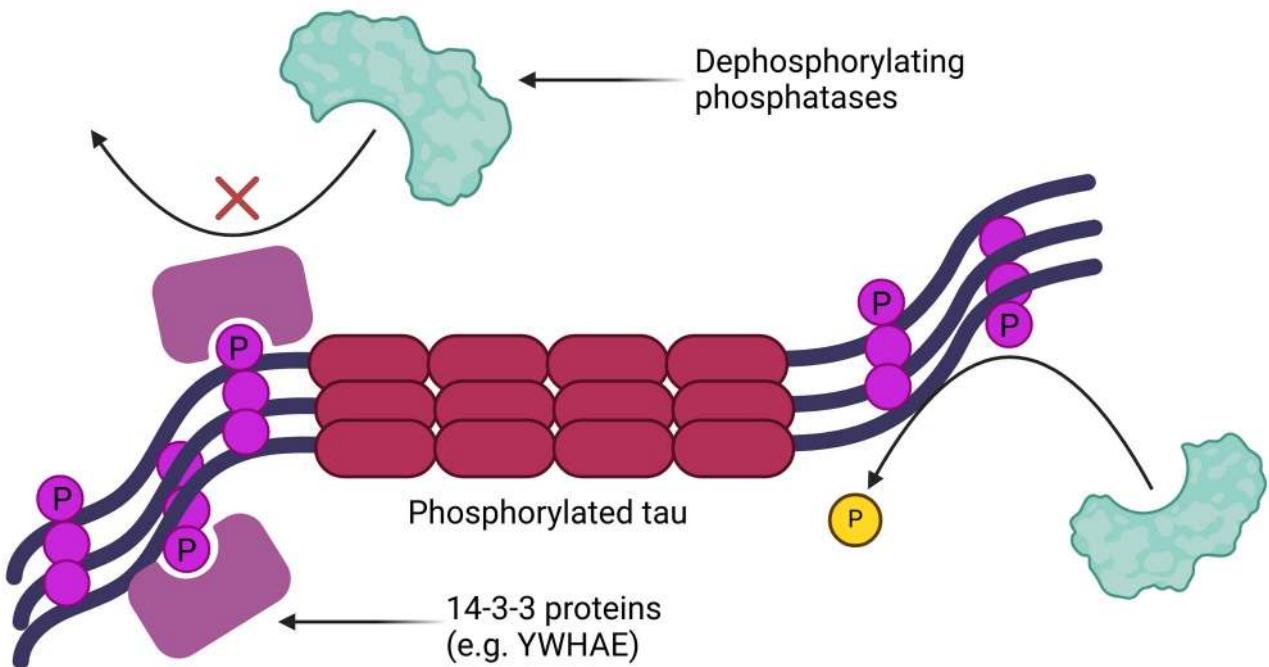


Figure 195: Proposed disease mechanism of 14-3-3 protein upregulation in tangle-bearing neurons based on the results of the analysis. Dephosphorylating phosphatases are generally considered to be protective in AD by removing phosphoryl groups from hyperphosphorylated tau. These phosphoryl groups also bind 14-3-3 proteins, effectively sequestering them from phosphatase activity and contributing to the continuation of AD disease progression. Created with BioRender.com.

In further detail, (Sluchanko & Gusev, 2011) describe that when tau detaches from microtubules, it becomes accessible for phosphorylation by multiple protein kinases, creating potential binding sites for 14-3-3 proteins. While high-affinity binding of 14-3-3 proteins to these phosphorylated sites may inhibit the aggregation of hyperphosphorylated tau, this binding may also sequester phosphorylated residues from protein phosphatases, thereby preventing tau dephosphorylation. Furthermore, in the event of sequestration of phosphorylated tau within aggregates directly, that structure may become more highly stabilised, making degradation increasingly difficult. These ideas remain theories, though inhibition of dephosphorylation by 14-3-3 proteins has been observed in different biological contexts (Kacirova et al., 2017).

Aside from this putative mechanism, early work has identified 14-3-3 proteins as present NFTs in postmortem hippocampal tissue obtained from AD patients (Layfield et al., 1996). Subsequent immunolocalisation studies have demonstrated that these proteins accumulate both within and surrounding NFTs in AD-affected brains (Umahara et al., 2004). Tau was demonstrated to be a binding partner of 14-3-3 proteins in (Hashiguchi et al., 2000), where it was also shown to promote its hyperphosphorylation. Additionally, the

14-3-3 $\zeta$  form is proposed to function as an adaptor protein, mediating the interaction between GSK3 $\beta$  and tau, thus promoting GSK3 $\beta$ -dependent tau phosphorylation (Z. Yuan et al., 2004). On the other hand, 14-3-3 proteins potentially confer neuroprotection by supporting aggresome formation, thus aiding in the sequestration and subsequent degradation of toxic misfolded proteins (Kopito, 2000).

### 7.5.6 Recruitment of Neurotrophic Factors Through the WRC

WASF1 is a component of the WAVE regulatory complex (WRC), which promotes actin polymerisation and is essential for neuronal morphology and function (Dahl et al., 2003). The WRC is recruited to the plasma membrane upon BDNF-TrkB activation and this recruitment facilitates actin-dependent endocytosis of the BDNF-TrkB complex, highlighting the role of WASF1 in actin cytoskeletal remodeling in response to neurotrophic signaling (C. Xu et al., 2016). Additionally, BDNF-TrkB signalling has been shown to regulate actin cytoskeleton dynamics through pathways involving Rac1 and other intermediates, further supporting the role of neurotrophic factors in modulating actin remodelling (Gonzalez et al., 2016). However, neurotrophic signalling is frequently impaired in AD, leading to synaptic instability and progressive loss of neuronal connectivity (Zuccato & Cattaneo, 2009). The upregulation of WASF1 in NFT-bearing neurons, as well as the prioritisation of BDNF-TrkB signalling pathways in the GeneFunnel analysis, may reflect an attempt to compensate for deficits in neurotrophic support, that in turn should reinforce synaptic and cytoskeletal structure and function.

**Figure 196.**

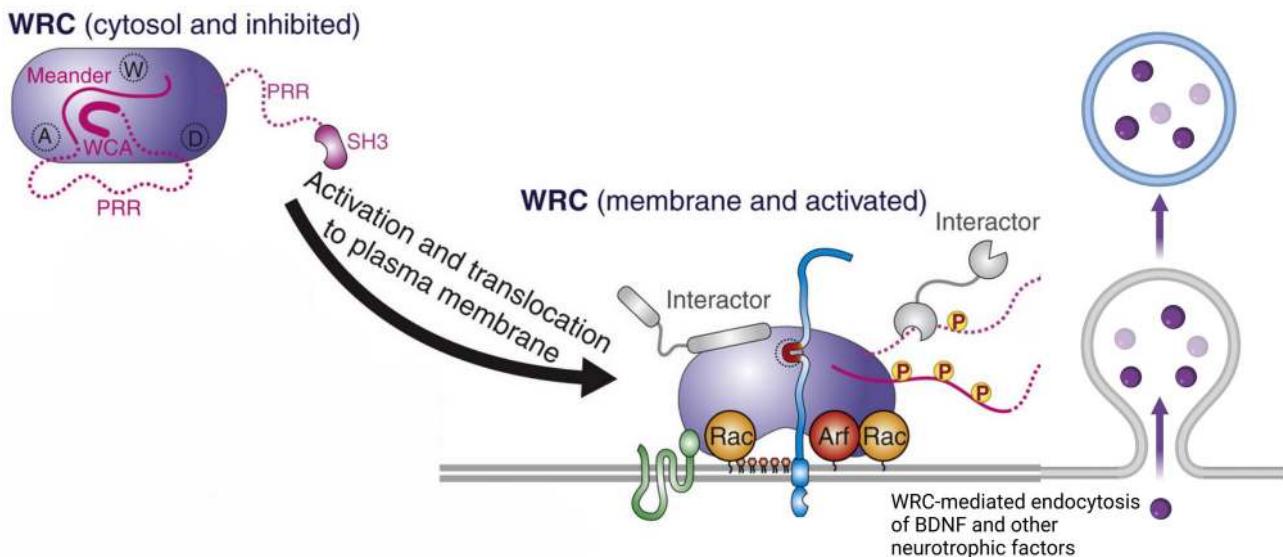


Figure 196: Proposed protective mechanism of activation of the WRC in tangle-bearing neurons based on the results of the analysis. The WRC, of which WASF1 is a major component, becomes active upon translocation to the plasma membrane. This mediates the endocytosis of BDNF and other neurotropic factors through BDNF-TrkB signalling. Created with BioRender.com using elements from (Rottner et al., 2021).

Beyond its role in cytoskeletal regulation, WASF1 is also implicated in intracellular trafficking pathways relevant to neurotrophic signalling. It has been shown to regulate endosomal dynamics and receptor trafficking, suggesting that increased WASF1 expression may influence the localisation and availability of neurotrophic factor receptors (Yokota et al., 2007). This could have profound effects on neuronal survival and synaptic function, particularly in AD where neurotrophic signalling is already compromised. In my analysis, neurotrophin related pathways appear to link WASF1 and MAPT, as they both appear in the differentially enriched pathway “cellular response to brain-derived neurotrophic factor stimulus”.

Regarding direct interactions with tau, work with the 3xTg AD mouse model by (Watamura et al., 2016) demonstrated that tau and WASF1 directly co-localise. This was also observed in 3xTg mice by (Takata et al., 2009) who additionally showed that co-localisation requires the presence of amyloid and tau pathologies. They included JNPL3 and Tg2576 mice in their experiment, which respectively feature tau and amyloid pathology separately, finding that the co-localisation is lost in those models but not the 3xTg model. Lastly, the key kinases Cdk5 and GSK3- $\beta$  have also been shown to phosphorylate WAVE1, in addition to tau (Ceglia et al., 2010), suggesting their close correlation downstream of signalling pathways, particularly in disease-promoting conditions.

### 7.5.7 Glial Involvement Through Paranodal Junctions

The paranodal junction is a specialised site of neuron-glia interaction where the myelin sheath is tightly anchored to the axon. CNTNAP1, shown upregulated in tangle-bearing neurons in both the proteomics and transcriptomics datasets, encodes contactin-associated protein 1 (Caspr), the key axonal adhesion molecule at this junction, linking the neuronal membrane to the flanking myelin loops (Ishibashi & Baba, 2022). This adhesion not only ensures physical attachment of oligodendrocyte processes to the axon, but also creates a barrier that compartmentalises the axonal membrane into node, paranode, and juxtaparanode regions. Thus, under physiological conditions, CNTNAP1 is an essential site of neuron-glia interaction at the paranodes, it not only physically tethers myelin to axons, but also helps coordination with oligodendrocytes for proper synaptic transmission. A schematic of the role of Caspr in these structures can be seen in Figure 197.

**Figure 197.**

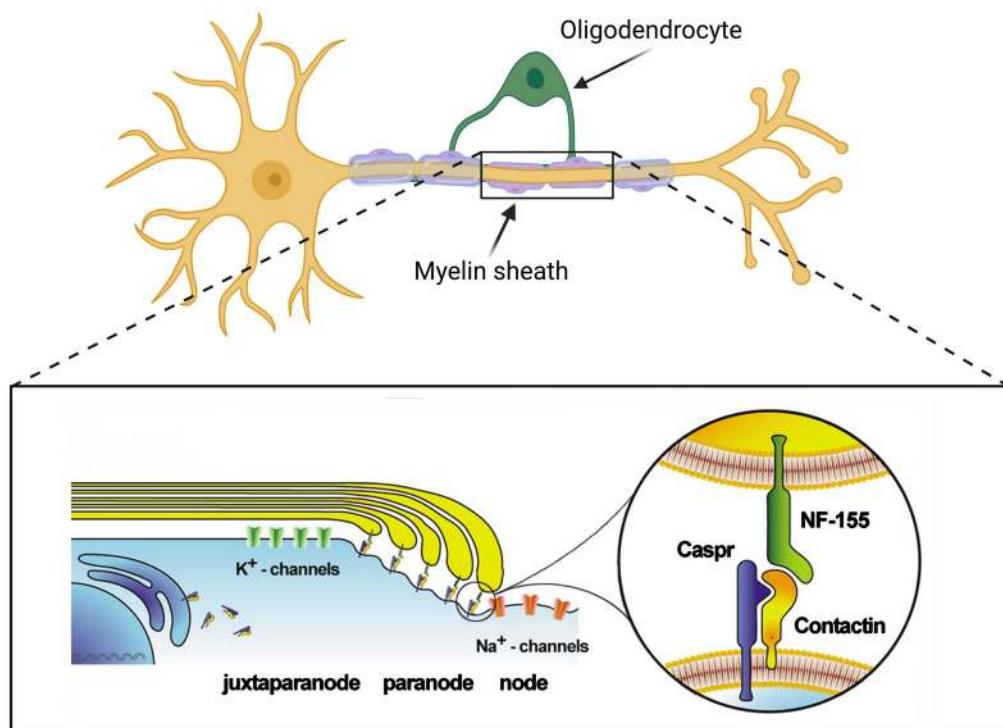


Figure 197: Physiological function of Caspr at the paranodal junction. Contactin forms an intracellular complex with Caspr, essential for the transport of this complex to the paranodal axolemma, where it interacts with NF155 present on the terminal loops of myelin. Created with BioRender.com using elements from (Boyle et al., 2001).

In the context of AD, emerging evidence points to the significant involvement of myelinating glia and their interactions with neurons. White matter changes are now recognised as part of AD pathology, as the accumulation of A $\beta$  and NFTs not only lead to synaptic and neuronal loss, but can also induce oligodendrocyte injury and myelin degeneration (Papuć & Rejdak, 2020). Oligodendroglial cell death and compromised myelin sheaths are observed in AD brains, and myelin impairment may even precede classical amyloid and tau lesions (Couttas et al., 2016). This context raises the possibility that neuron-glia junctions, such as paranodes, become destabilised during AD progression, thereby contributing to neurodegenerative mechanisms.

Tau pathology within neurons might directly or indirectly perturb paranodal junctions. In healthy axons, tau assists in maintaining the cytoskeleton that help position and transport paranodal proteins, however, the mislocalisation of tau forms NFTs that disrupt axonal transport and cytoskeletal organization. Such disruption could impair the delivery or anchoring of Caspr at paranodes, leading to junctional instability. Moreover, dying or

dysfunctional neurons might fail to sustain normal expression levels of axonal adhesion molecules. In line with this, pathological studies in demyelinating disease show that when myelin is lost, affected axons downregulate Caspr, and paranodal junctions disintegrate (Wolswijk, 2003). A similar phenomenon in AD could mean that tangle-bearing neurons undergoing degenerative changes might also exhibit altered Caspr distribution or loss as myelin support begins to fail. While it is unclear how this might directly translate to upregulation of Caspr and its associated gene *CNTNAP1*, it could be a compensatory response or general dyshomeostasis of its regulation and production.

#### 7.5.8 Compensation of Impaired Glutamate Recycling

Finally, the observed upregulation of aspartate aminotransferase (GOT2) in NFT-bearing neurons, alongside its enriched pathways (Figures 188 to 190), suggests significant alterations in neuronal metabolic processing and neurotransmitter homeostasis associated with AD. GOT2 is an enzyme found in high levels in the liver but also brain, central to the malate-aspartate shuttle (MAS), playing a crucial role in maintaining cellular redox balance and energy metabolism through facilitating the reversible conversion of glutamate and oxaloacetate into  $\alpha$ -ketoglutarate and aspartate (Borst, 2020). It is found in both a mitochondrial (GOT2) and cytoplasmic form (GOT1), fulfilling similar roles as either species. Relating GOT1/2 to a disease context, A $\beta$  oligomers have been shown to increase the abundance of glutamate in the extracellular space, leading to excitotoxicity (Hu et al., 2014; S. Li & Selkoe, 2020). As a protective response, or coincidentally beneficial side-effect, upregulation of GOT2 in tangle-bearing neurons adjacent to A $\beta$  oligomers or fibrils may provide an alternative pathway for turnover of excess glutamate through its conversion into  $\alpha$ -ketoglutarate and aspartate. A diagram of this process is shown in Figure 198.

**Figure 198.**

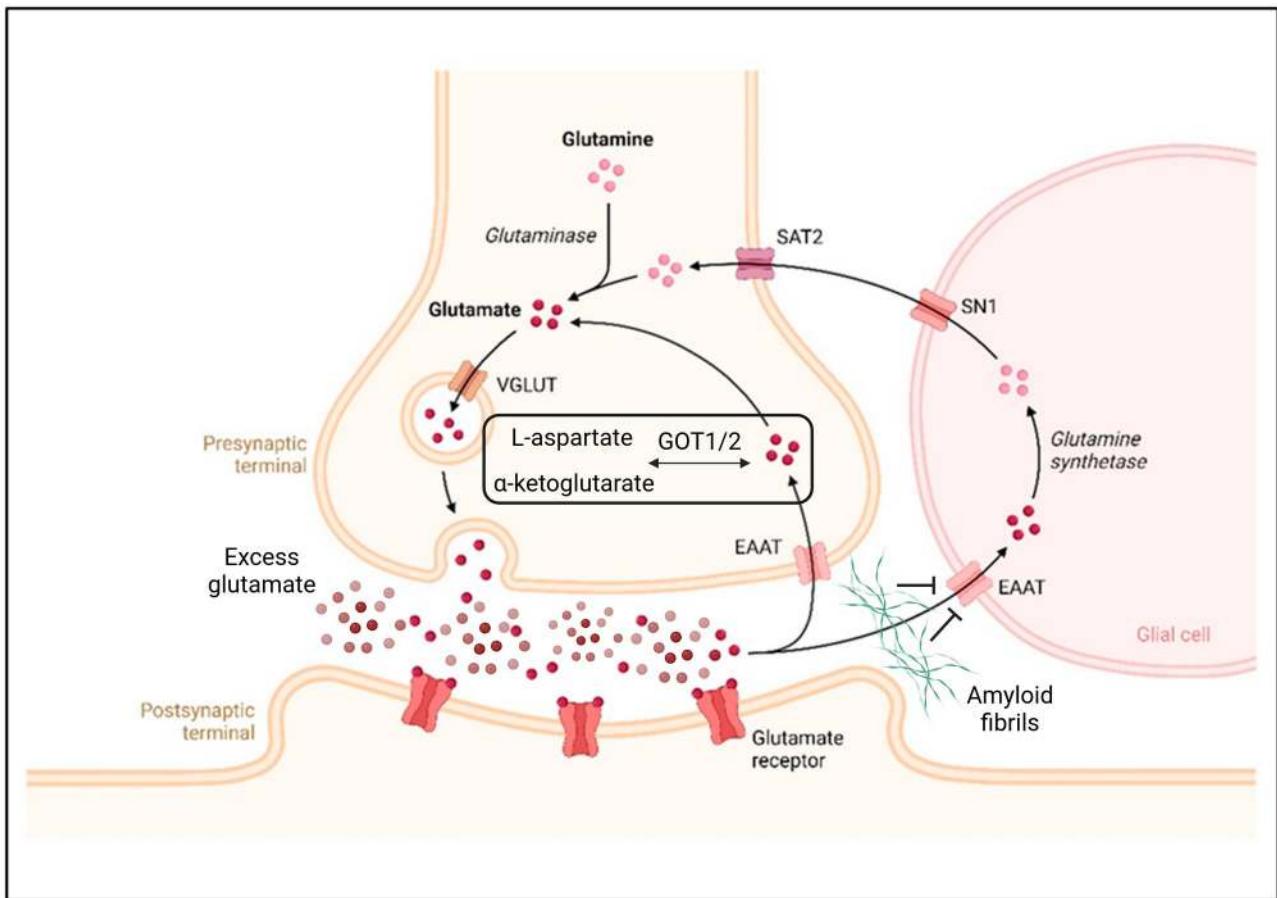


Figure 198: Proposed protective mechanism of GOT2 upregulation in tangle-bearing neurons based on the results of the analysis. A $\beta$  oligomers or fibrils block the functioning of EAAT receptors on glial cells (typically astrocytes) nearby neurons, halting the recycling of glutamine through glutamine synthetase. This manifests as an excess of glutamate in the synaptic cleft, leading to impaired synaptic function and excitotoxicity. GOT1/2 facilitates the reversible conversion of glutamate into  $\alpha$ -ketoglutarate and aspartate, and may compensate for impaired recycling by offering an alternative pathway for breakdown of excess glutamate. Created with BioRender.com using elements from (Puranik & Song, 2024).

In AD, oligomeric forms of A $\beta$  are known to impair synaptic plasticity. This disruption occurs partly due to A $\beta$ -induced downregulation of glutamate transporters, leading to glutamate spillover and subsequent overactivation of NMDA receptors, specifically the NMDA receptor 2B subunit (NMDA-R2B) (Hu et al., 2014; S. Li & Selkoe, 2020). Consequently, interventions that enhance glutamate clearance or block specific glutamate receptor subtypes have shown potential in mitigating A $\beta$ -mediated synaptic dysfunction and associated memory impairments (Puranik & Song, 2024). Seemingly unrelated, studies have revealed correlations between altered liver enzyme activities, particularly elevated aspartate aminotransferase (GOT1/2), and an increased risk of AD (Nho et al., 2019). Additionally, post-mortem analyses identified a substantial increase (approximately

1.5-fold) in aspartate aminotransferase activity in multiple cortical regions of individuals with AD compared to cognitively normal subjects (D'Aniello et al., 2005).

GOT1/2, in addition with oxaloacetate, catalyses the conversion of glutamate into  $\alpha$ -ketoglutarate, which effectively scavenges excess glutamate, facilitating its clearance from the brain (D. Zhang et al., 2019). In animal models, administration of aspartate aminotransferase in rats resulted in the rapid reduction of brain glutamate levels, rescuing synaptic plasticity impairments induced by A $\beta$  and inflammatory cytokines such as TNF $\alpha$  (D. Zhang et al., 2016). And a more recent study using APP/PS1 transgenic mice showed a significantly greater density of hippocampal CA1 synapses, accompanied by improved mitochondrial structural integrity, compared to untreated control mice (H. Li et al., 2023).

More generally on the topic of metabolic dysfunction, Alzheimer's Disease is associated with significant metabolic impairments. In fact, a study utilising LCM paired with microarray in AD post-mortem tissue (though in non-tangle-bearing neurons), highlighted the reduced expression of energy metabolism genes (W. S. Liang et al., 2008). Furthermore, a comprehensive metabolomic analysis revealed widespread metabolic dysregulation in the AD brain, affecting pathways related to bioenergetics, cholesterol metabolism, and neurotransmitter balance (Batra et al., 2023). More relevant to GOT2 in particular, glutamatergic neurotransmission decreases with age, and progression of AD also correlates with dysfunctions in the glutamatergic system (D. Huang et al., 2017). GOT2 gene expression has been reported as downregulated in AD bulk cortical tissue (Choe et al., 2024; Mahajan et al., 2020), though its expression/abundance has yet to be reported in NFT-bearing neurons. Given the therapeutic potential of GOT1/2, recovery of its physiological levels in AD may provide rescue of overall neural dysfunction, while it remains speculative as to whether its elevated levels in NFT-bearing neurons is truly protective or ultimately harmful.

## 8. Conclusion

This thesis provides a comprehensive review of the history and evolution of transcriptomics and proteomics, tracing their development from early methodologies to modern high-throughput techniques. These fields have undergone decades of technological advancement, continuously refining analytical methods and expanding their applications in biomedical research. Within this broader context, this work focuses specifically on two major computational challenges: gene set enrichment analysis and missing value imputation, both of which remain active areas of methodological development.

A key contribution of this thesis is the identification and systematic investigation of critical problem areas within these two computational domains. To address these challenges, I develop and introduce novel computational methods, designed to enhance the accuracy and interpretability of enrichment analysis and imputation strategies. These methods are implemented as open-source software, available to the research community as standalone R packages for integration into existing computational workflows. Web-based interactive viewers also comprise a significant portion of this work that enables researchers to explore the data freely.

Beyond computational development, this thesis applies these methods to an important and underexplored problem in Alzheimer's Disease research: the molecular characteristics of neurofibrillary tangle (NFT)-bearing neurons. NFTs, composed of hyperphosphorylated tau protein, are a defining feature of Alzheimer's pathology, yet the specific molecular determinants that govern their presence in affected neurons remain poorly understood. To advance this field, our lab introduces a new proteomics dataset, providing the research community with a valuable resource for investigating tau pathology at the protein level. Additionally, I perform a reanalysis of an existing single-cell transcriptomics dataset, integrating these two datasets to obtain a multi-omics perspective on NFTs in AD.

Applying my computational methods to these datasets yields a restricted set of 8 key molecular features that are consistently upregulated in both the transcriptomic and proteomic data: NEFM, APP, SQSTM1, HSP90AA1, YWHAE, WASF1, CNTNAP1, and GOT2. Each of these features plays a distinct but interconnected role in the pathophysiology of tangle-bearing neurons in AD, revealing the complex interplay of distinct domains. In detail, I discuss the potential impact of neurofilament co-aggregation coupled with microtubule destabilisation, a potentially protective role of the non-amyloigenic pathway, p62 accumulation and dysregulation of autophagy, the influence of HSP90 co-factors on protein folding and degradation, the sequestering of tau dephosphorylating phosphatases, the recruitment of neurotrophic factors such as BDNF, glial involvement through paranodal junctions, and compensatory responses to impaired glutamate recycling. This analysis lays the groundwork for future validation and exploration into these domains that may someday coalesce into a unified model of neurofibrillary tangle pathology in Alzheimer's Disease.

## 8.1 Overall Limitations and Future Directions

While this analysis provides valuable insights into the molecular factors underlying NFT-bearing neurons in Alzheimer's Disease, it is inherently constrained by the availability and quality of datasets. At the time of writing, no proteomic datasets existed that specifically quantified tangle-bearing neurons at single-cell resolution, which prompted our lab to generate the first such dataset. However, as with any novel dataset, this first-of-its-kind proteomic dataset contains technical challenges that could be improved with future optimization of sample preparation, mass spectrometry acquisition, and data processing pipelines. Enhancing these aspects in subsequent iterations of the dataset could significantly improve signal clarity, coverage, and reproducibility, allowing for more precise insights into protein-level alterations in NFT-bearing neurons.

The availability of transcriptomic datasets were also highly limited, with only one high-quality FACS-based single-cell RNA-seq dataset available for this study. While this dataset underwent rigorous quality control and was determined to be suitable for differential expression analysis, it lacks a definitive ground truth for comparison. Older LCM-based microarray datasets exist, but these were determined to be of insufficient quality for integration into the present analysis due to poor resolution, lower dynamic range, and technical artifacts. The lack of an independent transcriptomic dataset for validation remains a notable constraint, underscoring the need for replication studies and additional single-cell sequencing efforts.

An interpretive limitation is that the RNA and protein datasets were not derived from the same donors or tissue sections. They are matched by region and disease stage, but they are not paired at the individual level. Consequently, agreement between RNA and protein should be read as convergent evidence in similar contexts rather than as evidence of coupling within the same cells or donors. Relatedly, Alzheimer's disease contains mixed 3-repeat and 4-repeat tau, and the relative proportions and phosphorylation states vary by region and disease stage. Selection by AT8 enriches for specific phospho-epitopes and may not capture all tau species equally. Conclusions therefore generalise to AT8-positive tangles in the sampled region, not to all tau pathology.

The precision of single-cell isolation by LCM is another constraint. Sections were cut at 8 µm thickness, and material from adjacent cells or processes in the z-plane cannot be completely excluded. Profiles should be regarded as soma-enriched rather than perfectly isolated single cells. In the FACS material, although the strategy targeted neuronal somas, a small minority of events showed astrocytic or OPC markers. Non-specific staining, incomplete doublet exclusion and neuron-glia aggregates are plausible technical contributors, though a biological contribution cannot be excluded.

This study also introduces novel computational methodologies, both of which were benchmarked against existing approaches. However, while these methods demonstrated strong performance in controlled benchmarking scenarios, they have not yet been

validated in widespread real-world applications. Their effectiveness depends on a number of underlying assumptions that may not hold universally, particularly as new ground truth datasets become available for comparison.

For example, GeneFunnel's functional class scoring depends on curated resources (e.g. Gene Ontology). These resources are uneven across the genome and favour well-studied genes, which can bias hub weighting and enrichment toward highly curated features (for example, APP or HSP90AA1) while down-weighting less characterised genes. Alternative weighting schemes that account for annotation density, or incorporate citation-independent priors, would mitigate this bias. Similarly, ImputeFinder relies on distinguishing MAR (Missing At Random) and MNAR (Missing Not At Random) values based on an observed relationship between protein intensity and missingness. While this assumption was empirically tested within the available datasets, it remains unclear whether this relationship will hold under future ground truth datasets or across other proteomic workflows and instrumentation.

To address dataset, current work in our lab is focused on replicating the FACS single-soma RNAseq dataset and generating higher quality LCM mass spectrometry datasets. The LCM dataset used in this study exhibited substantial donor and technical variability, which necessitated the removal of a large number of samples due to inconsistencies in quality and coverage. While the remaining dataset was determined to be suitable for analysis, the loss of samples was not ideal and introduced additional uncertainty into the final results. Future iterations of LCM mass spec datasets will prioritise enhancing sample consistency, improving technical reproducibility, and optimising protein extraction from LCM-captured material to reduce inter-sample variability.

Despite these limitations, this study provides a strong foundation for future investigations into the molecular determinants of NFT-bearing neurons in Alzheimer's Disease. By integrating multi-omics data, novel computational tools, and network-based analyses, this work establishes a framework for systematically identifying key molecular players in NFT pathology. However, functional validation remains essential to determine whether the identified differentially expressed genes, proteins, and enriched pathways play causal roles in driving neurodegeneration or are simply correlates of tangle formation. The findings presented here can serve as a roadmap for targeted experimental studies, such as gene perturbations, proteomic interaction mapping, or simply additional confirmation using orthogonal techniques such as RNAscope. Future research can build upon this work to refine therapeutic targets, uncover novel biomarkers, and deepen our understanding of the complex landscape of neurofibrillary tangles in Alzheimer's Disease progression.

## 9. References

- Aceves, M., Granados, J., Leandro, A. C., Peralta, J., Glahn, D. C., Williams-Blangero, S., Curran, J. E., Blangero, J., & Kumar, S. (2024). Role of Neurocellular Endoplasmic Reticulum Stress Response in Alzheimer's Disease and Related Dementias Risk. *Genes*, 15(5), 569.  
<https://doi.org/10.3390/genes15050569>
- Acosta, D., Powell, F., Zhao, Y., & Raj, A. (2018). Regional vulnerability in Alzheimer's disease: The role of cell-autonomous and transneuronal processes. *Alzheimer's & Dementia*, 14(6), 797–810.  
<https://doi.org/10.1016/j.jalz.2017.11.014>
- Adzhubei, A. A., Tolstova, A. P., Strelkova, M. A., Mitkevich, V. A., Petrushanko, I. Yu., & Makarov, A. A. (2022). Interaction Interface of A $\beta$ 42 with Human Na,K-ATPase Studied by MD and ITC and Inhibitor Screening by MD. *Biomedicines*, 10(7), 1663.  
<https://doi.org/10.3390/biomedicines10071663>
- Aghili-Ashtiani, A. (2021). Upper bounds on deviations from the mean and the mean absolute deviation. *Italian Journal of Pure and Applied Mathematics*.
- Aldridge, S., & Teichmann, S. A. (2020). Single cell transcriptomics comes of age. *Nature Communications*, 11(1), 4307.  
<https://doi.org/10.1038/s41467-020-18158-5>
- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2022). *rmarkdown: Dynamic documents for r* [Manual]. <https://github.com/rstudio/rmarkdown>
- Alonso, A. D., Cohen, L. S., Corbo, C., Morozova, V., Elldrissi, A., Phillips, G., & Kleiman, F. E. (2018). Hyperphosphorylation of Tau Associates With Changes in Its Function Beyond Microtubule Stability. *Frontiers in Cellular Neuroscience*, 12, 338. <https://doi.org/10.3389/fncel.2018.00338>

- Alquezar, C., Arya, S., & Kao, A. W. (2021). Tau Post-translational Modifications: Dynamic Transformers of Tau Function, Degradation, and Aggregation. *Frontiers in Neurology*, 11, 595532. <https://doi.org/10.3389/fneur.2020.595532>
- Amaral, D. G., & Witter, M. P. (1989). The three-dimensional organization of the hippocampal formation: A review of anatomical data. *Neuroscience*, 31(3), 571–591. [https://doi.org/10.1016/0306-4522\(89\)90424-7](https://doi.org/10.1016/0306-4522(89)90424-7)
- Arendt, T., Stieler, J. T., & Holzer, M. (2016). Tau and tauopathies. *Brain Research Bulletin*, 126, 238–292. <https://doi.org/10.1016/j.brainresbull.2016.08.018>
- Arneberg, R., Rajalahti, T., Flikka, K., Berven, F. S., Kroksveen, A. C., Berle, M., Myhr, K.-M., Vedeler, C. A., Ulvik, R. J., & Kvalheim, O. M. (2007). Pretreatment of Mass Spectral Profiles: Application to Proteomic Data. *Analytical Chemistry*, 79(18), 7014–7026. <https://doi.org/10.1021/ac070946s>
- Arnsten, A. F. T., Datta, D., Tredici, K. D., & Braak, H. (2020). Hypothesis: Tau pathology is an initiating factor in sporadic Alzheimer's disease. *Alzheimer's & Dementia*, alz.12192. <https://doi.org/10.1002/alz.12192>
- Astillero-Lopez, V., Villar-Conde, S., Gonzalez-Rodriguez, M., Flores-Cuadrado, A., Ubeda-Banon, I., Saiz-Sanchez, D., & Martinez-Marcos, A. (2024). Proteomic analysis identifies HSP90AA1, PTK2B, and ANXA2 in the human entorhinal cortex in Alzheimer's disease: Potential role in synaptic homeostasis and A $\beta$  pathology through microglial and astroglial cells. *Brain Pathology*, 34(4), e13235. <https://doi.org/10.1111/bpa.13235>
- Azargoonjahromi, A. (2024). The duality of amyloid- $\beta$ : Its role in normal and Alzheimer's disease states. *Molecular Brain*, 17(1), 44. <https://doi.org/10.1186/s13041-024-01118-1>
- Babu, J. R., Geetha, T., & Wooten, M. W. (2005). Sequestosome 1/p62 shuttles polyubiquitinated tau for proteasomal degradation. *Journal of*

- Neurochemistry*, 94(1), 192–203. <https://doi.org/10.1111/j.1471-4159.2005.03181.x>
- Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N., & Patrick, E. (2019). Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Frontiers in Immunology*, 10, 2657. <https://doi.org/10.3389/fimmu.2019.02657>
- Bamford, R. A., Widagdo, J., Takamura, N., Eve, M., Anggono, V., & Oguro-Ando, A. (2020). The Interaction Between Contactin and Amyloid Precursor Protein and Its Role in Alzheimer's Disease. *Neuroscience*, 424, 184–202. <https://doi.org/10.1016/j.neuroscience.2019.10.006>
- Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., ... Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269), 108–112. <https://doi.org/10.1038/nature08460>
- Barbier, P., Zejneli, O., Martinho, M., Lasorsa, A., Belle, V., Smet-Nocca, C., Tsvetkov, P. O., Devred, F., & Landrieu, I. (2019). Role of Tau as a Microtubule-Associated Protein: Structural and Functional Aspects. *Frontiers in Aging Neuroscience*, 11, 204. <https://doi.org/10.3389/fnagi.2019.00204>
- Barker, S. J., Raju, R. M., Milman, N. E. P., Wang, J., Davila-Velderrain, J., Gunter-Rahman, F., Parro, C. C., Bozzelli, P. L., Abdurrob, F., Abdelaal, K., Bennett, D. A., Kellis, M., & Tsai, L.-H. (2021). MEF2 is a key regulator of cognitive potential and confers resilience to neurodegeneration. *Science Translational Medicine*, 13(618), eabd7695. <https://doi.org/10.1126/scitranslmed.abd7695>
- Barkovits, K., Pacharra, S., Pfeiffer, K., Steinbach, S., Eisenacher, M., Marcus, K., & Uszkoreit, J. (2020). Reproducibility, Specificity and Accuracy of

- Relative Quantification Using Spectral Library-based Data-independent Acquisition. *Molecular & Cellular Proteomics*, 19(1), 181–197.  
<https://doi.org/10.1074/mcp.RA119.001714>
- Barykin, E. P., Mitkevich, V. A., Kozin, S. A., & Makarov, A. A. (2017). Amyloid β Modification: A Key to the Sporadic Alzheimer's Disease? *Frontiers in Genetics*, 8, 58. <https://doi.org/10.3389/fgene.2017.00058>
- Bateman, N. W., Goulding, S. P., Shulman, N. J., Gadok, A. K., Szumlinski, K. K., MacCoss, M. J., & Wu, C. C. (2014). Maximizing Peptide Identification Events in Proteomic Workflows Using Data-Dependent Acquisition (DDA). *Molecular & Cellular Proteomics*, 13(1), 329–338.  
<https://doi.org/10.1074/mcp.M112.026500>
- Batra, R., Arnold, M., Wörheide, M. A., Allen, M., Wang, X., Blach, C., Levey, A. I., Seyfried, N. T., Ertekin-Taner, N., Bennett, D. A., Kastenmüller, G., Kaddurah-Daouk, R. F., Krumsiek, J., & for the Alzheimer's Disease Metabolomics Consortium (ADMC). (2023). The landscape of metabolic brain alterations in Alzheimer's disease. *Alzheimer's & Dementia*, 19(3), 980–998. <https://doi.org/10.1002/alz.12714>
- Bayerlová, M. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics*, 15.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1), 38–44. <https://doi.org/10.1038/nbt.4314>
- Blaudin De Thé, F.-X., Lassus, B., Schaler, A. W., Fowler, S. L., Goulbourne, C. N., Jeggo, R., Mannoury La Cour, C., Millan, M. J., & Duff, K. E. (2021). P62 accumulates through neuroanatomical circuits in response to tauopathy propagation. *Acta Neuropathologica Communications*, 9(1), 177.  
<https://doi.org/10.1186/s40478-021-01280-w>

- Bohush, A., Bieganowski, P., & Filipek, A. (2019). Hsp90 and Its Co-Chaperones in Neurodegenerative Diseases. *International Journal of Molecular Sciences*, 20(20), 4976. <https://doi.org/10.3390/ijms20204976>
- Bonam, S. R., Bayry, J., Tschan, M. P., & Muller, S. (2020). Progress and Challenges in the Use of MAP1LC3 as a Legitimate Marker for Measuring Dynamic Autophagy In Vivo. *Cells*, 9(5), 1321. <https://doi.org/10.3390/cells9051321>
- Bordi, M., Berg, M. J., Mohan, P. S., Peterhoff, C. M., Alldred, M. J., Che, S., Ginsberg, S. D., & Nixon, R. A. (2016). Autophagy flux in CA1 neurons of Alzheimer hippocampus: Increased induction overburdens failing lysosomes to propel neuritic dystrophy. *Autophagy*, 12(12), 2467-2483. <https://doi.org/10.1080/15548627.2016.1239003>
- Borràs, E., & Sabidó, E. (2017). What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. *Proteomics*, 17(17-18), 1700180. <https://doi.org/10.1002/pmic.201700180>
- Borst, P. (2020). The malate-aspartate shuttle (Borst cycle): How it started and developed into a major metabolic pathway. *IUBMB Life*, 72(11), 2241-2259. <https://doi.org/10.1002/iub.2367>
- Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21), 9546-9551. <https://doi.org/10.1073/pnas.0914005107>
- Boyarko, B., & Hook, V. (2021). Human Tau Isoforms and Proteolysis for Production of Toxic Tau Fragments in Neurodegeneration. *Frontiers in Neuroscience*, 15, 702788. <https://doi.org/10.3389/fnins.2021.702788>
- Boyle, M. E. T., Berglund, E. O., Murai, K. K., Weber, L., Peles, E., & Ranscht, B. (2001). Contactin Orchestrates Assembly of the Septate-like Junctions at

- the Paranode in Myelinated Peripheral Nerve. *Neuron*, 30(2), 385–397.  
[https://doi.org/10.1016/S0896-6273\(01\)00296-3](https://doi.org/10.1016/S0896-6273(01)00296-3)
- Braak, H., & Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*, 82(4), 239–259.  
<https://doi.org/10.1007/BF00308809>
- Braak, H., Rüb, U., Schultz, C., & Tredici, K. D. (2006). Vulnerability of cortical neurons to Alzheimer's and Parkinson's diseases. *Journal of Alzheimer's Disease*, 9(s3), 35–44. <https://doi.org/10.3233/JAD-2006-9S305>
- Bramer, L. M., Irvahn, J., Piehowski, P. D., Rodland, K. D., & Webb-Robertson, B.-J. M. (2021). A Review of Imputation Strategies for Isobaric Labeling-Based Shotgun Proteomics. *Journal of Proteome Research*, 20(1), 1–13.  
<https://doi.org/10.1021/acs.jproteome.0c00123>
- Brown, A. (1998). Contiguous phosphorylated and non-phosphorylated domains along axonal neurofilaments. *Journal of Cell Science*, 111(4), 455–467.  
<https://doi.org/10.1242/jcs.111.4.455>
- Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10), 630–644. <https://doi.org/10.1038/s41576-020-0258-4>
- Buchholz, S., & Zempel, H. (2024). The six brain-specific TAU isoforms and their role in Alzheimer's disease and related neurodegenerative dementia syndromes. *Alzheimer's & Dementia*, 20(5), 3606–3628.  
<https://doi.org/10.1002/alz.13784>
- Budimlija, Z. M., Lechpammer, M., Popolek, D., Fogt, F., Prinz, M., & Bieber, F. R. (2005). Forensic Applications of Laser Capture Microdissection: Use in DNA-Based Parentage Testing and Platform Validation. *Croat Med J*.
- Bull, C., Byrne, R. M., Fisher, N. C., Corry, S. M., Amirkhah, R., Edwards, J., Hillson, L., Lawler, M., Ryan, A., Lamrock, F., Dunne, P. D., & Malla, S. B. (2024). Evaluation of Gene Set Enrichment Analysis (GSEA) tools

- highlights the value of single sample approaches over pairwise for robust biological discovery. *bioRxiv*. <https://doi.org/10.1101/2024.03.15.585228>
- Busche, M. A., & Hyman, B. T. (2020). Synergy between amyloid- $\beta$  and tau in Alzheimer's disease. *Nature Neuroscience*, 23(10), 1183–1193.  
<https://doi.org/10.1038/s41593-020-0687-6>
- Buschow, S. I., Van Balkom, B. W. M., Aalberts, M., Heck, A. J. R., Wauben, M., & Stoorvogel, W. (2010). MHC class II-associated proteins in B-cell exosomes and potential functional implications for exosome biogenesis. *Immunology & Cell Biology*, 88(8), 851–856.  
<https://doi.org/10.1038/icb.2010.64>
- Bustin, S. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, 25(2), 169–193. <https://doi.org/10.1677/jme.0.0250169>
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5), 411–420.  
<https://doi.org/10.1038/nbt.4096>
- Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3). <https://doi.org/10.18637/jss.v045.i03>
- Calero, O., Bullido, M. J., Clarimón, J., Frank-García, A., Martínez-Martín, P., Lleó, A., Rey, M. J., Rábano, A., Blesa, R., Gómez-Isla, T., Valdivieso, F., De Pedro-Cuesta, J., Ferrer, I., & Calero, M. (2011). Genetic Cross-Interaction between APOE and PRNP in Sporadic Alzheimer's and Creutzfeldt-Jakob Diseases. *PLoS ONE*, 6(7), e22090.  
<https://doi.org/10.1371/journal.pone.0022090>
- Campos-Melo, D., Hawley, Z. C. E., & Strong, M. J. (2018). Dysregulation of human NEFM and NEFH mRNA stability by ALS-linked miRNAs. *Molecular Brain*, 11(1), 43. <https://doi.org/10.1186/s13041-018-0386-3>

- Candia, J., & Ferrucci, L. (2024). Assessment of Gene Set Enrichment Analysis using curated RNA-seq-based benchmarks. *PLoS ONE*, 19(5), e0302696. <https://doi.org/10.1371/journal.pone.0302696>
- Canovas, B., & Nebreda, A. R. (2021). Diversity and versatility of p38 kinase signalling in health and disease. *Nature Reviews Molecular Cell Biology*, 22(5), 346–366. <https://doi.org/10.1038/s41580-020-00322-w>
- Capano, C. P., Pernas-Alonso, R., & Di Porzio, U. (2000). Neurofilament homeostasis and motoneurone degeneration. *BioEssays*, 23(1), 24–33. [https://doi.org/10.1002/1521-1878\(200101\)23:1%253C24::AID-BIES1004%253E3.0.CO;2-H](https://doi.org/10.1002/1521-1878(200101)23:1%253C24::AID-BIES1004%253E3.0.CO;2-H)
- Carrasquillo, M. M., Belbin, O., Hunter, T. A., Ma, L., Bisceglie, G. D., Zou, F., Crook, J. E., Pankratz, V. S., Dickson, D. W., Graff-Radford, N. R., Petersen, R. C., Morgan, K., & Younkin, S. G. (2010). Replication of CLU, CR1, and PICALM Associations With Alzheimer Disease. *Archives of Neurology*, 67(8). <https://doi.org/10.1001/archneurol.2010.147>
- Castle, A. R., & Gill, A. C. (2017). Physiological Functions of the Cellular Prion Protein. *Frontiers in Molecular Biosciences*, 4. <https://doi.org/10.3389/fmolb.2017.00019>
- Ceglia, I., Kim, Y., Nairn, A. C., & Greengard, P. (2010). Signaling pathways controlling the phosphorylation state of WAVE1, a regulator of actin polymerization. *Journal of Neurochemistry*, 114(1), 182–190. <https://doi.org/10.1111/j.1471-4159.2010.06743.x>
- Ceglia, I., Reitz, C., Gresack, J., Ahn, J.-H., Bustos, V., Bleck, M., Zhang, X., Martin, G., Simon, S. M., Nairn, A. C., Greengard, P., & Kim, Y. (2015). APP intracellular domain-WAVE1 pathway reduces amyloid- $\beta$  production. *Nature Medicine*, 21(9), 1054–1059. <https://doi.org/10.1038/nm.3924>
- Chan, A. D. C., Dharmarajan, A. A., Atwood, C. S., Huang, X., Tanzi, R. E., Bush, A. I., & Martins, R. N. (1999). Anti-apoptotic action of Alzheimer A $\beta$ . *Alzheimer's Reports*.

- Chen, G., Xu, T., Yan, Y., Zhou, Y., Jiang, Y., Melcher, K., & Xu, H. E. (2017). Amyloid beta: Structure, biology and structure-based therapeutic development. *Acta Pharmacologica Sinica*, 38(9), 1205–1235.  
<https://doi.org/10.1038/aps.2017.28>
- Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., & Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608(7924), 733–740.  
<https://doi.org/10.1038/s41586-022-05046-9>
- Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., & Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biology*, 19(1), 70. <https://doi.org/10.1186/s13059-018-1438-9>
- Chen, W., Zhang, S., Williams, J., Ju, B., Shaner, B., Easton, J., Wu, G., & Chen, X. (2020). A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing. *Computational and Structural Biotechnology Journal*, 18, 861–873. <https://doi.org/10.1016/j.csbj.2020.03.026>
- Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5, 1438.  
<https://doi.org/10.12688/f1000research.8987.2>
- Chin, J., Massaro, C. M., Palop, J. J., Thwin, M. T., Yu, G.-Q., Bien-Ly, N., Bender, A., & Mucke, L. (2007). Reelin Depletion in the Entorhinal Cortex of Human Amyloid Precursor Protein Transgenic Mice and Humans with Alzheimer's Disease. *Journal of Neuroscience*, 27(11), 2727–2733.  
<https://doi.org/10.1523/JNEUROSCI.3758-06.2007>
- Choe, K., Ali, M., Lardenoije, R., Riemens, R. J. M., Pishva, E., Bickel, H., Weyerer, S., Hoffmann, P., Pentzek, M., Riedel-Heller, S., Wiese, B.,

- Scherer, M., Wagner, M., Mastroeni, D., Coleman, P. D., Ramirez, A., Ramakers, I. H. G. B., Verhey, F. R. J., Rutten, B. P. F., ... Van Den Hove, D. L. A. (2024). Alzheimer's disease-specific transcriptomic and epigenomic changes in the tryptophan catabolic pathway. *Alzheimer's Research & Therapy*, 16(1), 259. <https://doi.org/10.1186/s13195-024-01623-4>
- Chong, C.-M., Ke, M., Tan, Y., Huang, Z., Zhang, K., Ai, N., Ge, W., Qin, D., Lu, J.-H., & Su, H. (2018). Presenilin 1 deficiency suppresses autophagy in human neural stem cells through reducing γ-secretase-independent ERK/CREB signaling. *Cell Death & Disease*, 9(9), 879. <https://doi.org/10.1038/s41419-018-0945-7>
- Choudhary, S., & Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology*, 23(1), 27. <https://doi.org/10.1186/s13059-021-02584-9>
- Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., & Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, 8(4), 315-328.e8. <https://doi.org/10.1016/j.cels.2019.03.010>
- Constans, A. (2002). Arraying the Genome. *TheScientist*.
- Couttas, T. A., Kain, N., Suchowerska, A. K., Quek, L.-E., Turner, N., Fath, T., Garner, B., & Don, A. S. (2016). Loss of ceramide synthase 2 activity, necessary for myelin biosynthesis, precedes tau pathology in the cortical pathogenesis of Alzheimer's disease. *Neurobiology of Aging*, 43, 89–100. <https://doi.org/10.1016/j.neurobiolaging.2016.03.027>
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367–1372. <https://doi.org/10.1038/nbt.1511>
- Csardi, G., & Nepusz, T. (2006). *The igraph software package for complex network research*. 9.

- Cummings, J., Zhou, Y., Lee, G., Zhong, K., Fonseca, J., & Cheng, F. (2024). Alzheimer's disease drug development pipeline: 2024. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 10(2), e12465. <https://doi.org/10.1002/trc2.12465>
- Dahl, J. P., Wang-Dunlop, J., Gonzales, C., Goad, M. E. P., Mark, R. J., & Kwak, S. P. (2003). Characterization of the WAVE1 Knock-Out Mouse: Implications for CNS Development. *The Journal of Neuroscience*, 23(8), 3343–3352. <https://doi.org/10.1523/JNEUROSCI.23-08-03343.2003>
- D'Aniello, A., Fisher, G., Migliaccio, N., Cammisa, G., D'Aniello, E., & Spinelli, P. (2005). Amino acids and transaminases activity in ventricular CSF and in brain of normal and Alzheimer patients. *Neuroscience Letters*, 388(1), 49–53. <https://doi.org/10.1016/j.neulet.2005.06.030>
- Das, S., McClain, C. J., & Rai, S. N. (2020). Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy*, 22(4), 427. <https://doi.org/10.3390/e22040427>
- De Strooper, B. (2010). Proteases and Proteolysis in Alzheimer Disease: A Multifactorial View on the Disease Process. *Physiological Reviews*, 90(2), 465–494. <https://doi.org/10.1152/physrev.00023.2009>
- Del Tredici, K., & Braak, H. (2020). To stage, or not to stage. *Current Opinion in Neurobiology*, 61, 10–22. <https://doi.org/10.1016/j.conb.2019.11.008>
- Di Sanzo, M., Cozzolino, F., Battaglia, A. M., Aversa, I., Monaco, V., Sacco, A., Biamonte, F., Palmieri, C., Procopio, F., Santamaria, G., Ortuso, F., Pucci, P., Monti, M., & Faniello, M. C. (2022). Ferritin Heavy Chain Binds Peroxiredoxin 6 and Inhibits Cell Proliferation and Migration. *International Journal of Molecular Sciences*, 23(21), 12987. <https://doi.org/10.3390/ijms232112987>
- Dickey, C. A., Kamal, A., Lundgren, K., Klosak, N., Bailey, R. M., Dunmore, J., Ash, P., Shoraka, S., Zlatkovic, J., Eckman, C. B., Patterson, C., Dickson, D.

- W., Nahman, N. S., Hutton, M., Burrows, F., & Petrucelli, L. (2007). The high-affinity HSP90-CHIP complex recognizes and selectively degrades phosphorylated tau client proteins. *Journal of Clinical Investigation*, 117(3), 648-658. <https://doi.org/10.1172/JCI29715>
- Dickson, T. C., King, C. E., McCormack, G. H., & Vickers, J. C. (1999). Neurochemical Diversity of Dystrophic Neurites in the Early and Late Stages of Alzheimer's Disease. *Experimental Neurology*, 156(1), 100-110. <https://doi.org/10.1006/exnr.1998.7010>
- Dorostkar, M. M., Zou, C., Blazquez-Llorca, L., & Herms, J. (2015). Analyzing dendritic spine pathology in Alzheimer's disease: Problems and opportunities. *Acta Neuropathologica*, 130(1), 1-19. <https://doi.org/10.1007/s00401-015-1449-5>
- Dou, F., Netzer, W. J., Tanemura, K., Li, F., Hartl, F. U., Takashima, A., Gouras, G. K., Greengard, P., & Xu, H. (2003). Chaperones increase association of tau protein with microtubules. *Proceedings of the National Academy of Sciences*, 100(2), 721-726. <https://doi.org/10.1073/pnas.242720499>
- Drummond, E., Pires, G., MacMurray, C., Askenazi, M., Nayak, S., Bourdon, M., Safar, J., Ueberheide, B., & Wisniewski, T. (2020). Phosphorylated tau interactome in the human Alzheimer's disease brain. *Brain*, 15.
- Du, Y., Wooten, M. C., Gearing, M., & Wooten, M. W. (2009). Age-associated oxidative damage to the p62 promoter: Implications for Alzheimer disease. *Free Radical Biology and Medicine*, 46(4), 492-501. <https://doi.org/10.1016/j.freeradbiomed.2008.11.003>
- Dudoit, S., Gentleman, R. C., & Quackenbush, J. (2003). Open Source Software for the Analysis of Microarray Data. *BioTechniques*, 34(sup3), 7. <https://doi.org/10.2144/mar03dudoit>
- Dunckley, T., Beach, T. G., Ramsey, K. E., Grover, A., Mastroeni, D., Walker, D. G., LaFleur, B. J., Coon, K. D., Brown, K. M., Caselli, R., Kukull, W., Higdon, R., McKeel, D., Morris, J. C., Hulette, C., Schmeichel, D., Reiman, E. M.,

- Rogers, J., & Stephan, D. A. (2006). Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. *Neurobiology of Aging*, 27(10), 1359–1371. <https://doi.org/10.1016/j.neurobiolaging.2005.08.013>
- Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, 72(1), 28–36. <https://doi.org/10.1080/00031305.2017.1375990>
- Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*.
- Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, 71, 1054–1063. <https://doi.org/10.1016/j.csda.2013.02.005>
- Engmann, O. (2009). Crosstalk between Cdk5 and GSK3 $\beta$ : Implications for Alzheimer's Disease. *Frontiers in Molecular Neuroscience*, 2. <https://doi.org/10.3389/neuro.02.002.2009>
- Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software*, 48(4). <https://doi.org/10.18637/jss.v048.i04>
- Fang, X., Zhuang, X., Zheng, L., Lv, Y., Gao, F., Mo, C., & Zheng, X. (2025). SQSTM1 upregulation-induced iron overload triggers endothelial ferroptosis in nicotine-exacerbated atherosclerosis. *Life Sciences*, 361, 123330. <https://doi.org/10.1016/j.lfs.2024.123330>
- Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246(4926), 64–71. <https://doi.org/10.1126/science.2675315>
- Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D. B., Saviola, A. J., Yu, N.-K., & Yates, J. R. (2020). Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *Journal of Proteome*

- Research*, 19(8), 3153–3161.  
<https://doi.org/10.1021/acs.jproteome.0c00153>
- Fernandez-Martos, C. M., King, A. E., Atkinson, R. A. K., Woodhouse, A., & Vickers, J. C. (2015). Neurofilament light gene deletion exacerbates amyloid, dystrophic neurite, and synaptic pathology in the APP/PS1 transgenic model of Alzheimer's disease. *Neurobiology of Aging*, 36(10), 2757–2767. <https://doi.org/10.1016/j.neurobiolaging.2015.07.003>
- Ferré, C. A., Thouard, A., Bétourné, A., Le Dorze, A.-L., Belenguer, P., Miquel, M.-C., Peyrin, J.-M., Gonzalez-Dunia, D., & Szelechowski, M. (2021). HSPA9/Mortalin mediates axo-protection and modulates mitochondrial dynamics in neurons. *Scientific Reports*, 11(1), 17705.  
<https://doi.org/10.1038/s41598-021-97162-1>
- Foote, M., & Zhou, Y. (2012). 14-3-3 proteins in neurological disorders. *International Journal of Biochemistry and Molecular Biology*.
- Foroutan, M., Bhuva, D. D., Lyu, R., Horan, K., Cursons, J., & Davis, M. J. (2018). Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, 19(1), 404. <https://doi.org/10.1186/s12859-018-2435-4>
- Fowler, S. L., Behr, T. S., Turkes, E., O'Brien, D. P., Cauhy, P. M., Rawlinson, I., Edmonds, M., Foiani, M. S., Schaler, A., Crowley, G., Bez, S., Ficulle, E., Tsefou, E., Fischer, R., Geary, B., Gaur, P., Miller, C., D'Acunzo, P., Levy, E., ... Ryskeldi-Falcon, B. (2025). Tau filaments are tethered within brain extracellular vesicles in Alzheimer's disease. *Nature Neuroscience*, 28(1), 40–48. <https://doi.org/10.1038/s41593-024-01801-5>
- Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11), 1129–1164.  
<https://doi.org/10.1002/spe.4380211102>
- Fu, H., Hardy, J., & Duff, K. E. (2018). Selective vulnerability in neurodegenerative diseases. *Nature Neuroscience*, 21(10), 1350–1358.  
<https://doi.org/10.1038/s41593-018-0221-2>

- Fu, H., Possenti, A., Freer, R., Nakano, Y., Villegas, N. C. H., Tang, M., Cauhy, P. V. M., Lassus, B. A., Chen, S., Fowler, S. L., Figueroa, H. Y., Huey, E. D., Johnson, G. V. W., Vendruscolo, M., & Duff, K. E. (2019). A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. *Nature Neuroscience*, 22(1), 47–56. <https://doi.org/10.1038/s41593-018-0298-7>
- Furcila, D., Domínguez-Álvaro, M., DeFelipe, J., & Alonso-Nanclares, L. (2019). Subregional Density of Neurons, Neurofibrillary Tangles and Amyloid Plaques in the Hippocampus of Patients With Alzheimer's Disease. *Frontiers in Neuroanatomy*, 13, 99. <https://doi.org/10.3389/fnana.2019.00099>
- Gan, L., Cookson, M. R., Petrucelli, L., & La Spada, A. R. (2018). Converging pathways in neurodegeneration, from genetics to mechanisms. *Nature Neuroscience*, 21(10), 1300–1309. <https://doi.org/10.1038/s41593-018-0237-7>
- Gardner, M. L., & Freitas, M. A. (2021). Multiple Imputation Approaches Applied to the Missing Value Problem in Bottom-Up Proteomics. *International Journal of Molecular Sciences*, 22(17), 9650. <https://doi.org/10.3390/ijms22179650>
- Gatto, L., Gibb, S., & Rainer, J. (2021). MSnbase, Efficient and Elegant R-Based Processing and Visualization of Raw Mass Spectrometry Data. *Journal of Proteome Research*, 20(1), 1063–1069. <https://doi.org/10.1021/acs.jproteome.0c00313>
- Gatto, L., & Lilley, K. S. (2012). MSnbase—an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2), 288–289. <https://doi.org/10.1093/bioinformatics/btr645>
- Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., Zimmer, R., & Waldron, L. (2020).

- Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*, bbz158. <https://doi.org/10.1093/bib/bbz158>
- Germain, P.-L., Lun, A., Meixide, C. G., Macnair, W., & Robinson, M. D. (2022). Doublet identification in single-cell sequencing data. *F1000Research*.
- Gerschutz, A., Heinsen, H., Grunblatt, E., Wagner, A. K., Bartl, J., Meissner, C., Fallgatter, A. J., Al-Sarraj, S., Troakes, C., Ferrer, I., Arzberger, T., Deckert, J., Riederer, P., Fischer, M., Tatschner, T., & Monoranu, C. M. (2014). Neuron-Specific Alterations in Signal Transduction Pathways associated with Alzheimer's Disease. *Journal of Alzheimer's Disease*, 8.
- Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., Varga, Z., Wild, P. J., Günther, D., & Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4), 417–422. <https://doi.org/10.1038/nmeth.2869>
- Ginsberg, S. D., Hemby, S. E., & Trojanowski, J. Q. (2000). Expression profile of transcripts in Alzheimer's disease tangle-bearing CA1 neurons. *Annals of Neurology*, 11.
- Giuffrè, G. M., Quaranta, D., Costantini, E. M., Citro, S., Martellacci, N., De Ninno, G., Vita, M. G., Guglielmi, V., Rossini, P. M., Calabresi, P., & Marra, C. (2023). Cerebrospinal fluid neurofilament light chain and total-tau as biomarkers of neurodegeneration in Alzheimer's disease and frontotemporal dementia. *Neurobiology of Disease*, 186, 106267. <https://doi.org/10.1016/j.nbd.2023.106267>
- Goedert, M., Spillantini, M. G., & Crowther, R. A. (1991). Tau Proteins and Neurofibrillary Degeneration. *Brain Pathology*, 1(4), 279–286. <https://doi.org/10.1111/j.1750-3639.1991.tb00671.x>
- Gonzalez, A., Moya-Alvarado, G., Gonzalez-Billaut, C., & Bronfman, F. C. (2016). Cellular and molecular mechanisms regulating neuronal growth by brain-

- derived neurotrophic factor. *Cytoskeleton*, 73(10), 612-628.  
<https://doi.org/10.1002/cm.21312>
- Götz, J., Halliday, G., & Nisbet, R. M. (2019). Molecular Pathogenesis of the Tauopathies. *Annual Review of Pathology: Mechanisms of Disease*, 14(1), 239-261. <https://doi.org/10.1146/annurev-pathmechdis-012418-012936>
- Grantham, J. (2020). The Molecular Chaperone CCT/TRiC: An Essential Component of Proteostasis and a Potential Modulator of Protein Aggregation. *Frontiers in Genetics*, 11, 172.  
<https://doi.org/10.3389/fgene.2020.00172>
- Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L., & Marioni, J. C. (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature Communications*, 9(1), 2667. <https://doi.org/10.1038/s41467-018-05083-x>
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), 2847-2849. <https://doi.org/10.1093/bioinformatics/btw313>
- Guan, S., Taylor, P. P., Han, Z., Moran, M. F., & Ma, B. (2020). Data Dependent-Independent Acquisition (DDIA) Proteomics. *Journal of Proteome Research*, 19(8), 3230-3237.  
<https://doi.org/10.1021/acs.jproteome.0c00186>
- Guo, Y., Yu, D., Cupp-Sutton, K. A., Liu, X., & Wu, S. (2022). Optimization of protein-level tandem mass tag (TMT) labeling conditions in complex samples with top-down proteomics. *Analytica Chimica Acta*, 1221, 340037. <https://doi.org/10.1016/j.aca.2022.340037>
- Haberman, R. P., Branch, A., & Gallagher, M. (2017). Targeting Neural Hyperactivity as a Treatment to Stem Progression of Late-Onset Alzheimer's Disease. *Neurotherapeutics*, 14(3), 662-676.  
<https://doi.org/10.1007/s13311-017-0541-z>

- Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1), 296. <https://doi.org/10.1186/s13059-019-1874-1>
- Han, K. A., & Ko, J. (2023a). Orchestration of synaptic functions by WAVE regulatory complex-mediated actin reorganization. *Experimental & Molecular Medicine*, 55(6), 1065–1075. <https://doi.org/10.1038/s12276-023-01004-1>
- Han, K. A., & Ko, J. (2023b). Orchestration of synaptic functions by WAVE regulatory complex-mediated actin reorganization. *Experimental & Molecular Medicine*, 55(6), 1065–1075. <https://doi.org/10.1038/s12276-023-01004-1>
- Hänelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14(1), 7. <https://doi.org/10.1186/1471-2105-14-7>
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., ... Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13), 3573-3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048>
- Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2), 293–304. <https://doi.org/10.1038/s41587-023-01767-y>
- Hara, M., Hirokawa, K., Kamei, S., & Uchihara, T. (2013). Isoform transition from four-repeat to three-repeat tau underlies dendrosomatic and regional progression of neurofibrillary pathology. *Acta Neuropathologica*, 125(4), 565–579. <https://doi.org/10.1007/s00401-013-1097-6>

- Hardy, J. A., & Higgins, G. A. (1992). Alzheimer's Disease: The Amyloid Cascade Hypothesis. *Science*, 256(5054), 184–185.  
<https://doi.org/10.1126/science.1566067>
- Hashiguchi, M., Sobue, K., & Paudel, H. K. (2000). 14-3-3 $\zeta$  Is an Effector of Tau Protein Phosphorylation. *Journal of Biological Chemistry*, 275(33), 25247–25254. <https://doi.org/10.1074/jbc.M003738200>
- Haugh, M. C., & Probst, A. (1986). Alzheimer neurofibrillary tangles contain phosphorylated and hidden neurofilament epitopes. *Journal of Neurology, Neurosurgery, and Psychiatry*.
- Haynes, W. A., Tomczak, A., & Khatri, P. (2018). Gene annotation bias impedes biomedical research. *Scientific Reports*, 8(1), 1362.  
<https://doi.org/10.1038/s41598-018-19333-x>
- He, Z., Guo, J. L., McBride, J. D., Narasimhan, S., Kim, H., Changolkar, L., Zhang, B., Gathagan, R. J., Yue, C., Dengler, C., Stieber, A., Nitla, M., Coulter, D. A., Abel, T., Brunden, K. R., Trojanowski, J. Q., & Lee, V. M.-Y. (2018). Amyloid- $\beta$  plaques enhance Alzheimer's brain tau-seeded pathologies by facilitating neuritic plaque tau aggregation. *Nature Medicine*, 24(1), 29–38. <https://doi.org/10.1038/nm.4443>
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8.  
<https://doi.org/10.1016/j.ygeno.2015.11.003>
- Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real Time Quantitative PCR. *Genome Methods*.
- Henstridge, C. M., Hyman, B. T., & Spires-Jones, T. L. (2019). Beyond the neuron-cellular interactions early in Alzheimer disease pathogenesis. *Nature Reviews Neuroscience*, 20(2), 94–108.  
<https://doi.org/10.1038/s41583-018-0113-1>

- Hillenkamp, F., Karas, M., Beavis, R. C., & Chait, B. T. (1991). Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. *Analytical Chemistry*, 63(24), 1193A-1203A. <https://doi.org/10.1021/ac00024a716>
- Hippel, P. T. von, & Bartlett, J. (2019). Maximum likelihood multiple imputation: Faster imputations and consistent standard errors without posterior draws. *arXiv*. <https://doi.org/10.48550/arXiv.1210.0870>
- Hoffman, G. E., & Roussos, P. (2021). Dream: Powerful differential expression analysis for repeated measures designs. *Bioinformatics*, 37(2), 192–201. <https://doi.org/10.1093/bioinformatics/btaa687>
- Hondius, D. C., Koopmans, F., Leistner, C., Pita-Illobre, D., Peferoen-Baert, R. M., Marbus, F., Paliukhovich, I., Li, K. W., Rozemuller, A. J. M., Hoozemans, J. J. M., & Smit, A. B. (2021). The proteome of granulovacuolar degeneration and neurofibrillary tangles in Alzheimer's disease. *Acta Neuropathologica*, 141(3), 341–358. <https://doi.org/10.1007/s00401-020-02261-4>
- Hu, N.-W., Nicoll, A. J., Zhang, D., Mably, A. J., O'Malley, T., Purro, S. A., Terry, C., Collinge, J., Walsh, D. M., & Rowan, M. J. (2014). mGlu5 receptors and cellular prion protein mediate amyloid- $\beta$ -facilitated synaptic long-term depression in vivo. *Nature Communications*, 5(1), 3374. <https://doi.org/10.1038/ncomms4374>
- Huang, D., Liu, D., Yin, J., Qian, T., Shrestha, S., & Ni, H. (2017). Glutamate-glutamine and GABA in brain of normal aged and patients with cognitive impairment. *European Radiology*, 27(7), 2698–2705. <https://doi.org/10.1007/s00330-016-4669-8>
- Huang, E. J., & Reichardt, L. F. (2003). Trk Receptors: Roles in Neuronal Signal Transduction. *Annual Review of Biochemistry*, 72(1), 609–642. <https://doi.org/10.1146/annurev.biochem.72.121801.161629>
- Huang, Q., Liao, C., Ge, F., Ao, J., & Liu, T. (2022). Acetylcholine bidirectionally regulates learning and memory. *Journal of Neurorestoratology*, 10(2), 100002. <https://doi.org/10.1016/j.jnrt.2022.100002>

- Huang, T., Bruderer, R., Muntel, J., Xuan, Y., Vitek, O., & Reiter, L. (2020). Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition. *Molecular & Cellular Proteomics*, 19(2), 421–430.  
<https://doi.org/10.1074/mcp.RA119.001705>
- Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl\_1), S96–S104. [https://doi.org/10.1093/bioinformatics/18.suppl\\_1.S96](https://doi.org/10.1093/bioinformatics/18.suppl_1.S96)
- Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., & Krijgsveld, J. (2019). Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols*, 14(1), 68–85.  
<https://doi.org/10.1038/s41596-018-0082-x>
- Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E., & Engelhard, V. H. (1992). Characterization of Peptides Bound to the Class I MHC Molecule HLA-A2.1 by Mass Spectrometry. *Science*, 255(5049), 1261–1263.  
<https://doi.org/10.1126/science.1546328>
- International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., ... Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. <https://doi.org/10.1038/35057062>
- Iseki, E., Marui, W., Kosaka, K., & Uéda, K. (1999). Frequent coexistence of Lewy bodies and neurofibrillary tangles in the same neurons of patients with diffuse Lewy body disease. *Neuroscience Letters*, 265(1), 9–12.  
[https://doi.org/10.1016/S0304-3940\(99\)00178-0](https://doi.org/10.1016/S0304-3940(99)00178-0)

- Iseki, E., Yamamoto, R., Murayama, N., Minegishi, M., Togo, T., Katsuse, O., Kosaka, K., Akiyama, H., Tsuchiya, K., Rohan, de S., Andrew, L., & Arai, H. (2006). Immunohistochemical investigation of neurofibrillary tangles and their tau isoforms in brains of limbic neurofibrillary tangle dementia. *Neuroscience Letters*, 405(1-2), 29-33.  
<https://doi.org/10.1016/j.neulet.2006.06.036>
- Ishibashi, T., & Baba, H. (2022). Paranodal Axoglial Junctions, an Essential Component in Axonal Homeostasis. *Frontiers in Cell and Developmental Biology*, 10, 951809. <https://doi.org/10.3389/fcell.2022.951809>
- Ishii, T., Haga, S., & Tokutake, S. (1979). Presence of neurofilament protein in Alzheimer's neurofibrillary tangles (ANT): An immunofluorescent study. *Acta Neuropathologica*, 48(2), 105-112.  
<https://doi.org/10.1007/BF00691151>
- Janesick, A., Shelansky, R., Gottscho, A. D., Wagner, F., Williams, S. R., Rouault, M., Beliakoff, G., Morrison, C. A., Oliveira, M. F., Sicherman, J. T., Kohlway, A., Abousoud, J., Drennon, T. Y., Mohabbat, S. H., 10x Development Teams, & Taylor, S. E. B. (2023). High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1), 8353.  
<https://doi.org/10.1038/s41467-023-43458-x>
- Jiang, J., Parameshwaran, K., Seibenhener, M. L., Kang, M., Suppiramaniam, V., Huganir, R. L., Diaz-Meco, M. T., & Wooten, M. W. (2009). AMPA receptor trafficking and synaptic plasticity require SQSTM1/p62. *Hippocampus*, 19(4), 392-406. <https://doi.org/10.1002/hipo.20528>
- Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., & Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, 11(1), 1760.  
<https://doi.org/10.1038/s41598-021-81279-4>

- Josephs, K. A. (2017). Current Understanding of Neurodegenerative Diseases Associated With the Protein Tau. *Mayo Clinic Proceedings*, 92(8), 1291–1303. <https://doi.org/10.1016/j.mayocp.2017.04.016>
- Jouanne, M., Rault, S., & Voisin-Chiret, A.-S. (2017). Tau protein aggregation in Alzheimer's disease: An attractive target for the development of novel therapeutic agents. *European Journal of Medicinal Chemistry*, 139, 153–167. <https://doi.org/10.1016/j.ejmech.2017.07.070>
- Kacirova, M., Novacek, J., Man, P., Obsilova, V., & Obsil, T. (2017). Structural Basis for the 14-3-3 Protein-Dependent Inhibition of Phosducin Function. *Biophysical Journal*, 112(7), 1339–1349. <https://doi.org/10.1016/j.bpj.2017.02.036>
- Kanaan, N. M. (2024). Tau here, tau there, tau almost everywhere: Clarifying the distribution of tau in the adult CNS. *Cytoskeleton*, 81(1), 107–115. <https://doi.org/10.1002/cm.21820>
- Kanak, D. J., Rose, G. M., Zaveri, H. P., & Patrylo, P. R. (2013). Altered Network Timing in the CA3-CA1 Circuit of Hippocampal Slices from Aged Mice. *PLoS ONE*, 8(4), e61364. <https://doi.org/10.1371/journal.pone.0061364>
- Kang, Y., Burton, L., Lau, A., & Tate, S. (2017). SWATH-ID: An instrument method which combines identification and quantification in a single analysis. *Proteomics*, 17(10), 1500522. <https://doi.org/10.1002/pmic.201500522>
- Karpievitch, Y. V., Nikolic, S. B., Wilson, R., Sharman, J. E., & Edwards, L. M. (2014). Metabolomics Data Normalization with EigenMS. *PLoS ONE*, 9(12), e116221. <https://doi.org/10.1371/journal.pone.0116221>
- Kawano, S., Baba, M., Fukushima, H., Miura, D., Hashimoto, H., & Nakazawa, T. (2022). Autism-associated ANK2 regulates embryonic neurodevelopment. *Biochemical and Biophysical Research Communications*, 605, 45–50. <https://doi.org/10.1016/j.bbrc.2022.03.058>
- Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., & Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved

- tissue and cells. *Nature Methods*, 10(9), 857–860.  
<https://doi.org/10.1038/nmeth.2563>
- Ke, Y. D., Chan, G., Stefanoska, K., Au, C., Bi, M., Müller, J., Przybyla, M., Feiten, A., Prikas, E., Halliday, G. M., Piguet, O., Kiernan, M. C., Kassiou, M., Hodges, J. R., Loy, C. T., Mattick, J. S., Ittner, A., Kril, J. J., Sutherland, G. T., & Ittner, L. M. (2019). CNS cell type-specific gene profiling of P301S tau transgenic mice identifies genes dysregulated by progressive tau accumulation. *Journal of Biological Chemistry*, 294(38), 14149–14162.  
<https://doi.org/10.1074/jbc.RA118.005263>
- Khalil, M., Teunissen, C. E., Lehmann, S., Otto, M., Piehl, F., Ziemssen, T., Bittner, S., Sormani, M. P., Gattringer, T., Abu-Rumeileh, S., Thebault, S., Abdelhak, A., Green, A., Benkert, P., Kappos, L., Comabella, M., Tumani, H., Freedman, M. S., Petzold, A., ... Kuhle, J. (2024). Neurofilaments as biomarkers in neurological disorders—Towards clinical application. *Nature Reviews Neurology*, 20(5), 269–287. <https://doi.org/10.1038/s41582-024-00955-x>
- Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, 8(2), e1002375. <https://doi.org/10.1371/journal.pcbi.1002375>
- Kim, H., Golub, G. H., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation. *Bioinformatics*, 21(2), 187–198.  
<https://doi.org/10.1093/bioinformatics/bth499>
- King, C. E., Canty, A. J., & Vickers, J. C. (2001). Alterations in neurofilaments associated with reactive brain changes and axonal sprouting following acute physical injury to the rat neocortex. *Neuropathology and Applied Neurobiology*, 27(2), 115–126. <https://doi.org/10.1046/j.1365-2990.2001.00317.x>

- Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *BioEssays*, 32(6), 524–536.  
<https://doi.org/10.1002/bies.200900181>
- Kittur, S., Hoh, J., Endo, H., Tourtellotte, W., Weeks, B. S., Markesberry, W., & Adler, W. (1994). Cytoskeletal Neurofilament Gene Expression in Brain Tissue from Alzheimer's Disease Patients. I. Decrease in NF-L and NF-M Message. *Journal of Geriatric Psychiatry and Neurology*, 7(3), 153–158.  
<https://doi.org/10.1177/089198879400700305>
- Klunk, W. E., McClure, R. J., & Pettegrew, J. W. (1991). L-Phosphoserine, a Metabolite Elevated in Alzheimer's Disease, Interacts with Specific L-Glutamate Receptor Subtypes. *Journal of Neurochemistry*, 56(6), 1997–2003. <https://doi.org/10.1111/j.1471-4159.1991.tb03458.x>
- Kobro-Flatmoen, A., & Witter, M. P. (2019). Neuronal chemo-architecture of the entorhinal cortex: A comparative review. *European Journal of Neuroscience*, 50(10), 3627–3662. <https://doi.org/10.1111/ejn.14511>
- Kong, W., Hui, H. W. H., Peng, H., & Goh, W. W. B. (2022). Dealing with missing values in proteomics data. *Proteomics*, 22(23–24), 2200092.  
<https://doi.org/10.1002/pmic.202200092>
- Kopito, R. R. (2000). Aggresomes, inclusion bodies and protein aggregation. *Trends in Cell Biology*, 10(12), 524–530. [https://doi.org/10.1016/S0962-8924\(00\)01852-3](https://doi.org/10.1016/S0962-8924(00)01852-3)
- Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2016). Fast gene set enrichment analysis. *bioRxiv*.  
<https://doi.org/10.1101/060012>
- Koudinov, A. R., & Koudinova, N. V. (2003). Amyloid beta protein restores hippocampal long term potentiation: A central role for cholesterol? *Neurobiology of Lipids*.

- Kovacs, G. (2016). Molecular Pathological Classification of Neurodegenerative Diseases: Turning towards Precision Medicine. *International Journal of Molecular Sciences*, 17(2), 189. <https://doi.org/10.3390/ijms17020189>
- Kraft, L. J., Dowler, J., Manral, P., & Kenworthy, A. K. (2016). Size, organization, and dynamics of soluble SQSTM1 and LC3-SQSTM1 complexes in living cells. *Autophagy*, 12(9), 1660–1674.  
<https://doi.org/10.1080/15548627.2016.1199299>
- Kretzschmar, H. (2009). Brain banking: Opportunities, challenges and meaning for the future. *Nature Reviews Neuroscience*, 10(1), 70–78.  
<https://doi.org/10.1038/nrn2535>
- Kumar, A. V., Mills, J., & Lapierre, L. R. (2022). Selective Autophagy Receptor p62/SQSTM1, a Pivotal Player in Stress and Aging. *Frontiers in Cell and Developmental Biology*, 10, 793328.  
<https://doi.org/10.3389/fcell.2022.793328>
- Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS ONE*, 12(5), e0177459.  
<https://doi.org/10.1371/journal.pone.0177459>
- Kuusisto, E., Salminen, A., & Alafuzoff, I. (2002). Early accumulation of p62 in neurofibrillary tangles in Alzheimer's disease: Possible role in tangle formation. *Neuropathology and Applied Neurobiology*, 28(3), 228–237.  
<https://doi.org/10.1046/j.1365-2990.2002.00394.x>
- L. Lun, A. T., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1), 75. <https://doi.org/10.1186/s13059-016-0947-7>
- Layfield, R., Fergusson, J., Aitken, A., Lowe, J., Landon, M., & Mayer, R. J. (1996). Neurofibrillary tangles of Alzheimer's disease brains contain 14-3-3 proteins. *Neuroscience Letters*, 209(1), 57–60.  
[https://doi.org/10.1016/0304-3940\(96\)12598-2](https://doi.org/10.1016/0304-3940(96)12598-2)

Lazar, C., Gatto, L., Ferro, M., Bruley, C., & Burger, T. (2016). Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of Proteome Research*, 15(4), 1116–1125.

<https://doi.org/10.1021/acs.jproteome.5b00981>

Leng, K., Li, E., Eser, R., Piergies, A., Sit, R., Tan, M., Neff, N., Li, S. H., Rodriguez, R. D., Suemoto, C. K., Leite, R. E. P., Ehrenberg, A. J., Pasqualucci, C. A., Seeley, W. W., Spina, S., Heinsen, H., Grinberg, L. T., & Kampmann, M. (2021). Molecular characterization of selectively vulnerable neurons in Alzheimer's disease. *Nature Neuroscience*.

<https://doi.org/10.1038/s41593-020-00764-7>

Lenoir, T., & Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. *Journal of Biomedical Discovery and Collaboration*, 1(1), 11. <https://doi.org/10.1186/1747-5333-1-11>

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., ... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17), 4325–4333.

<https://doi.org/10.1073/pnas.1720115115>

Li, H., Zhang, D., Wang, X., Wang, S., & Xiao, M. (2023). Protective effect of glutamic-oxaloacetic transaminase on hippocampal neurons in Alzheimer's disease using model mice. *Neuroscience Letters*, 803, 137194. <https://doi.org/10.1016/j.neulet.2023.137194>

Li, M., & Smyth, G. K. (2023). Neither random nor censored: Estimating intensity-dependent probabilities for missing values in label-free proteomics. *Bioinformatics*, 39(5), btad200.

<https://doi.org/10.1093/bioinformatics/btad200>

- Li, S., & Selkoe, D. J. (2020). A mechanistic hypothesis for the impairment of synaptic plasticity by soluble A $\beta$  oligomers from Alzheimer's brain. *Journal of Neurochemistry*, 154(6), 583–597.  
<https://doi.org/10.1111/jnc.15007>
- Li, W., Yang, L., Tang, C., Liu, K., Lu, Y., Wang, H., Yan, K., Qiu, Z., & Zhou, W. (2020a). Mutations of CNTNAP1 led to defects in neuronal development. *JCI Insight*, 5(21), e135697. <https://doi.org/10.1172/jci.insight.135697>
- Li, W., Yang, L., Tang, C., Liu, K., Lu, Y., Wang, H., Yan, K., Qiu, Z., & Zhou, W. (2020b). Mutations of CNTNAP1 led to defects in neuronal development. *JCI Insight*, 5(21), e135697. <https://doi.org/10.1172/jci.insight.135697>
- Li, W., Yang, L., Tang, C., Liu, K., Lu, Y., Wang, H., Yan, K., Qiu, Z., & Zhou, W. (2020c). Mutations of CNTNAP1 led to defects in neuronal development. *JCI Insight*, 5(21), e135697. <https://doi.org/10.1172/jci.insight.135697>
- Liang, J.-H., & Jia, J.-P. (2014). Dysfunctional autophagy in Alzheimer's disease: Pathogenic roles and therapeutic implications. *Neuroscience Bulletin*, 30(2), 308–316. <https://doi.org/10.1007/s12264-013-1418-8>
- Liang, W. S., Reiman, E. M., Valla, J., Dunckley, T., Beach, T. G., Grover, A., Niedzielko, T. L., Schneider, L. E., Mastroeni, D., Caselli, R., Kukull, W., Morris, J. C., Hulette, C. M., Schmeichel, D., Rogers, J., & Stephan, D. A. (2008). Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences*, 105(11), 4441–4446.  
<https://doi.org/10.1073/pnas.0709259105>
- Lin, C.-H., Chin, Y., Zhou, M., Sobol, R. W., Hung, M.-C., & Tan, M. (2024). Protein lipoylation: Mitochondria, cuproptosis, and beyond. *Trends in Biochemical Sciences*, 49(8), 729–744. <https://doi.org/10.1016/j.tibs.2024.04.002>
- Lin, J., Fallahi-Sichani, M., Chen, J., & Sorger, P. K. (2016). Cyclic Immunofluorescence (CyclIF), A Highly Multiplexed Method for Single-cell

- Imaging. *Current Protocols in Chemical Biology*, 8(4), 251–264.  
<https://doi.org/10.1002/cpch.14>
- Lin, T., Tjernberg, L. O., & Schedin-Weiss, S. (2021). Neuronal Trafficking of the Amyloid Precursor Protein—What Do We Really Know? *Biomedicines*, 9(7), 801. <https://doi.org/10.3390/biomedicines9070801>
- Liu, L., Drouet, V., Wu, J. W., Witter, M. P., Small, S. A., Clelland, C., & Duff, K. (2012). Trans-Synaptic Spread of Tau Pathology In Vivo. *PLoS ONE*, 7(2), e31302. <https://doi.org/10.1371/journal.pone.0031302>
- Liu, M., & Dongre, A. (2021). Proper imputation of missing values in proteomics datasets for differential expression analysis. *Briefings in Bioinformatics*, 22(3), bbaa112. <https://doi.org/10.1093/bib/bbaa112>
- Liu, Q., Xie, F., Alvarado-Diaz, A., Smith, M. A., Moreira, P. I., Zhu, X., & Perry, G. (2011). Neurofilamentopathy in Neurodegenerative Diseases. *The Open Neurology Journal*.
- Liu, Z., Lv, C., Zhao, W., Song, Y., Pei, D., & Xu, T. (2012). NR2B-Containing NMDA Receptors Expression and Their Relationship to Apoptosis in Hippocampus of Alzheimer's Disease-Like Rats. *Neurochemical Research*, 37(7), 1420–1427. <https://doi.org/10.1007/s11064-012-0726-0>
- Lopez-Toledano, M. A. (2004). Neurogenic Effect of -Amyloid Peptide in the Development of Neural Stem Cells. *Journal of Neuroscience*, 24(23), 5439–5444. <https://doi.org/10.1523/JNEUROSCI.0974-04.2004>
- Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, 15(6), e8746. <https://doi.org/10.15252/msb.20188746>
- Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., & Marioni, J. C. (2019). EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, 20(1), 63. <https://doi.org/10.1186/s13059-019-1662-y>

- Luo, H.-B., Xia, Y.-Y., Shu, X.-J., Liu, Z.-C., Feng, Y., Liu, X.-H., Yu, G., Yin, G., Xiong, Y.-S., Zeng, K., Jiang, J., Ye, K., Wang, X.-C., & Wang, J.-Z. (2014). SUMOylation at K340 inhibits tau degradation through deregulating its phosphorylation and ubiquitination. *Proceedings of the National Academy of Sciences*, 111(46), 16586–16591.  
<https://doi.org/10.1073/pnas.1417548111>
- Luo, R., Colangelo, C. M., Sessa, W. C., & Zhao, H. (2009). Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins. *Statistics in Biosciences*, 1(2), 228–245.  
<https://doi.org/10.1007/s12561-009-9013-2>
- Ma, S., Attarwala, I. Y., & Xie, X.-Q. (2019). SQSTM1/p62: A Potential Target for Neurodegenerative Disease. *ACS Chemical Neuroscience*, 10(5), 2094–2114. <https://doi.org/10.1021/acschemneuro.8b00516>
- Macarrón-Palacios, V., Hubrich, J., Metzendorf, N. G., Kneilmann, S., Trapp, M., Acuna, C., Patrizi, A., D'Este, E., & Kilimann, M. W. (2025). Paralemmin-1 controls the nanoarchitecture of the neuronal submembrane cytoskeleton. *Science AdvAnceS*.
- Mahajan, U. V., Varma, V. R., Griswold, M. E., Blackshear, C. T., An, Y., Oommen, A. M., Varma, S., Troncoso, J. C., Pletnikova, O., O'Brien, R., Hohman, T. J., Legido-Quigley, C., & Thambisetty, M. (2020). Dysregulation of multiple metabolic networks related to brain transmethylation and polyamine pathways in Alzheimer disease: A targeted metabolomic and transcriptomic study. *PLOS Medicine*, 17(1), e1003012.  
<https://doi.org/10.1371/journal.pmed.1003012>
- Majtán, T., Bukovská, G., & Timko, J. (2004). DNA microarrays—Techniques and applications in microbial systems. *Folia Microbiologica*, 49(6), 635.  
<https://doi.org/10.1007/BF02931546>

- Maleki, F., Ovens, K., Hogan, D. J., & Kusalik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. *Frontiers in Genetics*, 11, 654. <https://doi.org/10.3389/fgene.2020.00654>
- Malone, J. H., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1), 34. <https://doi.org/10.1186/1741-7007-9-34>
- Malpas, C. B., Sharmin, S., & Kalincik, T. (2020). The histopathological staging of tau, but not amyloid, corresponds to antemortem cognitive status, dementia stage, functional abilities and neuropsychiatric symptoms. *International Journal of Neuroscience*, 1-10. <https://doi.org/10.1080/00207454.2020.1758087>
- Manczak, M., & Reddy, P. H. (2013). Abnormal Interaction of Oligomeric Amyloid- $\beta$  with Phosphorylated Tau: Implications to Synaptic Dysfunction and Neuronal Damage. *Journal of Alzheimer's Disease*, 36(2), 285-295. <https://doi.org/10.3233/JAD-130275>
- Mango, D., Saidi, A., Cisale, G. Y., Feligioni, M., Corbo, M., & Nisticò, R. (2019). Targeting Synaptic Plasticity in Experimental Models of Alzheimer's Disease. *Frontiers in Pharmacology*, 10, 778. <https://doi.org/10.3389/fphar.2019.00778>
- Mann, M. (2006). Functional and quantitative proteomics using SILAC. *Nature Reviews Molecular Cell Biology*, 7(12), 952-958. <https://doi.org/10.1038/nrm2067>
- Mann, M. (2016). Origins of mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology*.
- Marx, V. (2021). Method of the Year: Spatially resolved transcriptomics. *Nature Methods*, 18(1), 9-14. <https://doi.org/10.1038/s41592-020-01033-y>
- Marx, V. (2023). Method of the year: Long-read sequencing. *Nature Methods*, 20(1), 6-11. <https://doi.org/10.1038/s41592-022-01730-w>

- Masters, C. L., Bateman, R., Blennow, K., Rowe, C. C., Sperling, R. A., & Cummings, J. L. (2015). Alzheimer's disease. *Nature Reviews Disease Primers*, 1(1), 15056. <https://doi.org/10.1038/nrdp.2015.56>
- Mathioudakis, L., Dimovasili, C., Bourbouli, M., Latsoudis, H., Kokosali, E., Gouna, G., Vogiatzi, E., Basta, M., Kapetanaki, S., Panagiotakis, S., Kanterakis, A., Boumpas, D., Lionis, C., Plaitakis, A., Simos, P., Vgontzas, A., Kafetzopoulos, D., & Zaganas, I. (2023). Study of Alzheimer's disease-and frontotemporal dementia-associated genes in the Cretan Aging Cohort. *Neurobiology of Aging*, 123, 111-128.  
<https://doi.org/10.1016/j.neurobiolaging.2022.07.002>
- Matsushima, Y., Takahashi, K., Yue, S., Fujiyoshi, Y., Yoshioka, H., Aihara, M., Setoyama, D., Uchiumi, T., Fukuchi, S., & Kang, D. (2021). Mitochondrial Lon protease is a gatekeeper for proteins newly imported into the matrix. *Communications Biology*, 4(1), 974. <https://doi.org/10.1038/s42003-021-02498-z>
- Matsushita, S., Arai, H., Yuzuriha, T., Kato, M., Matsui, T., Urakami, K., & Higuchi, S. (2001). No association between DLST gene and Alzheimer's disease or Wernicke-Korsakoff syndrome. *Neurobiology of Aging*, 22(4), 569-574.  
[https://doi.org/10.1016/S0197-4580\(01\)00225-1](https://doi.org/10.1016/S0197-4580(01)00225-1)
- Mattiasson, G., Friberg, H., Hansson, M., Elmér, E., & Wieloch, T. (2003). Flow cytometric analysis of mitochondria from CA1 and CA3 regions of rat hippocampus reveals differences in permeability transition pore activation: Flow cytometry and hippocampal mitochondria. *Journal of Neurochemistry*, 87(2), 532-544. <https://doi.org/10.1046/j.1471-4159.2003.02026.x>
- McCarthy, D. J., Campbell, K. R., Lun, A. T. L., & Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, btw777.  
<https://doi.org/10.1093/bioinformatics/btw777>

- McGurk, K. A., Dagliati, A., Chiasseroni, D., Lee, D., Plant, D., Baricevic-Jones, I., Kelsall, J., Eineman, R., Reed, R., Geary, B., Unwin, R. D., Nicolaou, A., Keavney, B. D., Barton, A., Whetton, A. D., & Geifman, N. (2020). The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination. *Bioinformatics*, 36(7), 2217–2223. <https://doi.org/10.1093/bioinformatics/btz898>
- McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. <http://arxiv.org/abs/1802.03426>
- McKetney, J., Runde, R. M., Hebert, A. S., Salamat, S., Roy, S., & Coon, J. J. (2019). Proteomic Atlas of the Human Brain in Alzheimer's Disease. *Journal of Proteome Research*, 18(3), 1380–1391. <https://doi.org/10.1021/acs.jproteome.9b00004>
- McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 263–269. <https://doi.org/10.1016/j.jalz.2011.03.005>
- Meier, F., Brunner, A.-D., Koch, S., Koch, H., Lubeck, M., Krause, M., Goedecke, N., Decker, J., Kosinski, T., Park, M. A., Bache, N., Hoerning, O., Cox, J., Räther, O., & Mann, M. (2018). Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Molecular & Cellular Proteomics*, 17(12), 2534–2545. <https://doi.org/10.1074/mcp.TIR118.000900>
- Mellid, S., García, F., Leandro-García, L. J., Díaz-Talavera, A., Martínez-Montes, Á. M., Gil, E., Calsina, B., Monteagudo, M., Letón, R., Roldán-Romero, J. M.,

- Santos, M., Lanillos, J., Valdivia, C., Martínez-Puente, N., De Nicolás-Hernández, J., Jiménez, S., Pérez-Martínez, M., Honrado, E., Coloma, J., ... Cascón, A. (2023). DLST mutations in pheochromocytoma and paraganglioma cause proteome hyposuccinylation and metabolic remodeling. *Cancer Communications*, 43(7), 838-843.  
<https://doi.org/10.1002/cac2.12427>
- Mendez, M. F. (2012). Early-onset Alzheimer's Disease: Nonamnestic Subtypes and Type 2 AD. *Archives of Medical Research*, 9.
- Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*.
- Method of the Year 2013. (2014). *Nature Methods*, 11(1), 1-1.  
<https://doi.org/10.1038/nmeth.2801>
- Method of the Year 2024: Spatial proteomics. (2024). *Nature Methods*, 21(12), 2195-2196. <https://doi.org/10.1038/s41592-024-02565-3>
- Mizuseki, K., Royer, S., & Diba, K. (2012). Activity dynamics and behavioral correlates of CA3 and CA1 hippocampal pyramidal neurons. *Hippocampus*, 22.
- Moloney, C. M., Lowe, V. J., & Murray, M. E. (2021). Visualization of neurofibrillary tangle maturity in Alzheimer's disease: A clinicopathologic perspective for biomarker research. *Alzheimer's & Dementia*, 17(9), 1554-1574. <https://doi.org/10.1002/alz.12321>
- Montagne, A., Nation, D. A., Pa, J., Sweeney, M. D., Toga, A. W., & Zlokovic, B. V. (2016). Brain imaging of neurovascular dysfunction in Alzheimer's disease. *Acta Neuropathologica*, 131(5), 687-707.  
<https://doi.org/10.1007/s00401-016-1570-0>
- Montine, T. J., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Dickson, D. W., Duyckaerts, C., Frosch, M. P., Masliah, E., Mirra, S. S., Nelson, P. T., Schneider, J. A., Thal, D. R., Trojanowski, J. Q., Vinters, H. V., & Hyman, B. T. (2012). National Institute on Aging-Alzheimer's Association guidelines

- for the neuropathologic assessment of Alzheimer's disease: A practical approach. *Acta Neuropathologica*, 123(1), 1-11.  
<https://doi.org/10.1007/s00401-011-0910-3>
- Morrison, B. M., Hof, P. R., & Morrison, J. H. (1998). Determinants of neuronal vulnerability in neurodegenerative diseases: Neuronal Vulnerability in Neurodegenerative Diseases. *Annals of Neurology*, 44(S1), S32-S44.  
<https://doi.org/10.1002/ana.410440706>
- Morrison, J. H., Lewis, D. A., Campbell, M. J., Huntley, G. W., Benson, D. L., & Bouras, C. (1987). A monoclonal antibody to non-phosphorylated neurofilament protein marks the vulnerable cortical neurons in Alzheimer's disease. *Brain Research*, 416(2), 331-336.  
[https://doi.org/10.1016/0006-8993\(87\)90914-0](https://doi.org/10.1016/0006-8993(87)90914-0)
- Mrdjen, D., Fox, E. J., Bukhari, S. A., Montine, K. S., Bendall, S. C., & Montine, T. J. (2019). The basis of cellular and regional vulnerability in Alzheimer's disease. *Acta Neuropathologica*, 138(5), 729-749.  
<https://doi.org/10.1007/s00401-019-02054-4>
- Muddapu, V. R., Dharshini, S. A. P., Chakravarthy, V. S., & Gromiha, M. M. (2020). Neurodegenerative Diseases - Is Metabolic Deficiency the Root Cause? *Frontiers in Neuroscience*, 14, 213.  
<https://doi.org/10.3389/fnins.2020.00213>
- Muir, R., Diot, A., & Poulton, J. (2016). Mitochondrial content is central to nuclear gene expression: Profound implications for human health. *BioEssays*, 38(2), 150-156. <https://doi.org/10.1002/bies.201500105>
- Nadel, C. M., Thwin, A. C., Callahan, M., Lee, K., Connelly, E., Craik, C. S., Southworth, D. R., & Gestwicki, J. E. (2023). The E3 Ubiquitin Ligase, CHIP/STUB1, Inhibits Aggregation of Phosphorylated Proteoforms of Microtubule-associated Protein Tau (MAPT). *Journal of Molecular Biology*, 435(11), 168026. <https://doi.org/10.1016/j.jmb.2023.168026>

Nhan, H. S., Chiang, K., & Koo, E. H. (2015). The multifaceted nature of amyloid precursor protein and its proteolytic fragments: Friends and foes. *Acta Neuropathologica*, 129(1), 1–19. <https://doi.org/10.1007/s00401-014-1347-2>

Nho, K., Kueider-Paisley, A., Ahmad, S., MahmoudianDehkordi, S., Arnold, M., Risacher, S. L., Louie, G., Blach, C., Baillie, R., Han, X., Kastenmüller, G., Trojanowski, J. Q., Shaw, L. M., Weiner, M. W., Doraiswamy, P. M., Van Duijn, C., Saykin, A. J., Kaddurah-Daouk, R., & for the Alzheimer's Disease Neuroimaging Initiative and the Alzheimer Disease Metabolomics Consortium. (2019). Association of Altered Liver Enzymes With Alzheimer Disease Diagnosis, Cognition, Neuroimaging Measures, and Cerebrospinal Fluid Biomarkers. *JAMA Network Open*, 2(7), e197978. <https://doi.org/10.1001/jamanetworkopen.2019.7978>

Nichols, E., Szoek, C. E. I., Vollset, S. E., Abbasi, N., Abd-Allah, F., Abdela, J., Aichour, M. T. E., Akinyemi, R. O., Alahdab, F., Asgedom, S. W., Awasthi, A., Barker-Collo, S. L., Baune, B. T., Béjot, Y., Belachew, A. B., Bennett, D. A., Biadgo, B., Bijani, A., Bin Sayeed, M. S., ... Murray, C. J. L. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, 18(1), 88–106. [https://doi.org/10.1016/S1474-4422\(18\)30403-4](https://doi.org/10.1016/S1474-4422(18)30403-4)

Nilssen, E. S., Jacobsen, B., Fjeld, G., Nair, R. R., Blankvoort, S., Kentros, C., & Witter, M. P. (2018). Inhibitory Connectivity Dominates the Fan Cell Network in Layer II of Lateral Entorhinal Cortex. *The Journal of Neuroscience*, 38(45), 9712–9727. <https://doi.org/10.1523/JNEUROSCI.1290-18.2018>

Noh, M.-Y., Kwon, M.-S., Oh, K.-W., Nahm, M., Park, J., Kim, Y.-E., Ki, C.-S., Jin, H. K., Bae, J., & Kim, S. H. (2023). Role of NCKAP1 in the Defective Phagocytic Function of Microglia-Like Cells Derived from Rapidly

- Progressing Sporadic ALS. *Molecular Neurobiology*, 60(8), 4761–4777.  
<https://doi.org/10.1007/s12035-023-03339-2>
- Nolan, A., De Paula Franca Resende, E., Petersen, C., Neylan, K., Spina, S., Huang, E., Seeley, W., Miller, Z., & Grinberg, L. T. (2019). Astrocytic Tau Deposition Is Frequent in Typical and Atypical Alzheimer Disease Presentations. *Journal of Neuropathology & Experimental Neurology*, 78(12), 1112–1123. <https://doi.org/10.1093/jnen/nlz094>
- Nukina, N., Kosik, K. S., & Selkoe, D. J. (1987). Recognition of Alzheimer paired helical filaments by monoclonal neurofilament antibodies is due to crossreaction with tau protein. *Proc. Natl. Acad. Sci. USA*.
- Nygaard, H. B., & Strittmatter, S. M. (2009). Cellular Prion Protein Mediates the Toxicity of -Amyloid Oligomers. *ARCH NEUROL*, 66(11).
- Nyrén, P., & Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry*, 151(2), 504–509. [https://doi.org/10.1016/0003-2697\(85\)90211-8](https://doi.org/10.1016/0003-2697(85)90211-8)
- Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088–2096.  
<https://doi.org/10.1093/bioinformatics/btg287>
- O'Brien, J. J., Gunawardena, H. P., Paulo, J. A., Chen, X., Ibrahim, J. G., Gygi, S. P., & Qaqish, B. F. (2018). The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The Annals of Applied Statistics*, 12(4). <https://doi.org/10.1214/18-AOAS1144>
- Oh, M. M., Simkin, D., & Disterhoft, J. F. (2016). Intrinsic Hippocampal Excitability Changes of Opposite Signs and Different Origins in CA1 and CA3 Pyramidal Neurons Underlie Aging-Related Cognitive Deficits. *Frontiers in Systems Neuroscience*, 10.  
<https://doi.org/10.3389/fnsys.2016.00052>

- Oomens, J. E., Jansen, W. J., Janssen, O., Verhey, F. R., Vos, S. J., Visser, P. J., & Amyloid Biomarker Study Group. (2021). Associations of lifestyle factors with amyloid positivity in cognitively normal individuals at different levels of genetic risk: The amyloid biomarker study. *Alzheimer's & Dementia*, 17(S10), e054090. <https://doi.org/10.1002/alz.054090>
- Opattova, A., Filipcik, P., Cente, M., & Novak, M. (2012). Intracellular Degradation of Misfolded Tau Protein Induced by Geldanamycin is Associated with Activation of Proteasome. *Journal of Alzheimer's Disease*, 33(2), 339–348. <https://doi.org/10.3233/JAD-2012-121072>
- Osorio, D., & Cai, J. J. (2021). Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics*, 37(7), 963–967.  
<https://doi.org/10.1093/bioinformatics/btaa751>
- Otero-Garcia, M., Mahajani, S. U., Wakhloo, D., Tang, W., Xue, Y.-Q., Morabito, S., Pan, J., Oberhauser, J., Madira, A. E., Shakouri, T., Deng, Y., Allison, T., He, Z., Lowry, W. E., Kawaguchi, R., Swarup, V., & Cobos, I. (2022). Molecular signatures underlying neurofibrillary tangle susceptibility in Alzheimer's disease. *Neuron*, 110(18), 2929-2948.e8.  
<https://doi.org/10.1016/j.neuron.2022.06.021>
- Ou, J.-R., Tan, M.-S., Xie, A.-M., Yu, J.-T., & Tan, L. (2014a). Heat Shock Protein 90 in Alzheimer's Disease. *BioMed Research International*, 2014, 1–7.  
<https://doi.org/10.1155/2014/796869>
- Ou, J.-R., Tan, M.-S., Xie, A.-M., Yu, J.-T., & Tan, L. (2014b). Heat Shock Protein 90 in Alzheimer's Disease. *BioMed Research International*, 2014, 1–7.  
<https://doi.org/10.1155/2014/796869>
- Panza, F., Lozupone, M., Logroscino, G., & Imbimbo, B. P. (2019). A critical appraisal of amyloid- $\beta$ -targeting therapies for Alzheimer disease. *Nature Reviews Neurology*, 15(2), 73–88. <https://doi.org/10.1038/s41582-018-0116-6>

Papuć, E., & Rejdak, K. (2020). The role of myelin damage in Alzheimer's disease pathology. *Archives of Medical Science*, 16(2), 345–341.

<https://doi.org/10.5114/aoms.2018.76863>

Peng, Y., Yao, S., Chen, Q., Jin, H., Du, M., Xue, Y., & Liu, S. (2024). True or false? Alzheimer's disease is type 3 diabetes: Evidences from bench to bedside. *Ageing Research Reviews*, 99, 102383.

<https://doi.org/10.1016/j.arr.2024.102383>

Petrushanko, I. Yu., Mitkevich, V. A., Anashkina, A. A., Adzhubei, A. A., Burnysheva, K. M., Lakunina, V. A., Kamanina, Y. V., Dergousova, E. A., Lopina, O. D., Ogunshola, O. O., Bogdanova, A. Yu., & Makarov, A. A. (2016). Direct interaction of beta-amyloid with Na,K-ATPase as a putative regulator of the enzyme function. *Scientific Reports*, 6(1), 27738.

<https://doi.org/10.1038/srep27738>

Pham, P. V. (2018). Medical Biotechnology. In *Omics Technologies and Bio-Engineering* (pp. 449–469). Elsevier. <https://doi.org/10.1016/B978-0-12-804659-3.00019-1>

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., & Smyth, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics*, 10(2), 946–963. <https://doi.org/10.1214/16-AOAS920>

Pickett, E. K., Herrmann, A. G., McQueen, J., Abt, K., Dando, O., Tulloch, J., Jain, P., Dunnett, S., Sohrabi, S., Fjeldstad, M. P., Calkin, W., Murison, L., Jackson, R. J., Tzioras, M., Stevenson, A., d'Orange, M., Hooley, M., Davies, C., Colom-Cadena, M., ... Spires-Jones, T. L. (2019). Amyloid Beta and Tau Cooperate to Cause Reversible Behavioral and Transcriptional Deficits in a Model of Alzheimer's Disease. *Cell Reports*, 29(11), 3592–3604.e5. <https://doi.org/10.1016/j.celrep.2019.11.044>

Pietz, T., Gupta, S., Schlaffner, C. N., Ahmed, S., Steen, H., Renard, B. Y., & Baum, K. (2024). PEPPERMINT: Peptide abundance imputation in mass

- spectrometry-based proteomics using graph neural networks. *Bioinformatics*, 40(Supplement\_2), ii70–ii78.  
<https://doi.org/10.1093/bioinformatics/btae389>
- Plaitakis, A., Kalf-Ezra, E., Kotzamani, D., Zaganas, I., & Spanaki, C. (2017). The Glutamate Dehydrogenase Pathway and Its Roles in Cell and Tissue Biology in Health and Disease. *Biology*, 6(1), 11.  
<https://doi.org/10.3390/biology6010011>
- Plant, L. D., Boyle, J. P., Smith, I. F., Peers, C., & Pearson, H. A. (2003). The Production of Amyloid Peptide Is a Critical Requirement for the Viability of Central Neurons. *The Journal of Neuroscience*.
- Porchet, R., Probst, A., Dráberová, E., Dráber, P., Riederer, I. M., & Riederer, B. M. (2003). Differential subcellular localization of phosphorylated neurofilament and tau proteins in degenerating neurons of the human entorhinal cortex. *NeuroReport*, 14(7), 929–933.  
<https://doi.org/10.1097/01.wnr.0000072844.93264.31>
- Potier, B., Krzywkowski, P., Lamour, Y., & Dutar, P. (1994). Loss of calbindin-immunoreactivity in CA1 hippocampal stratum radiatum and stratum lacunosum-moleculare interneurons in the aged rat. *Brain Research*, 661(1-2), 181–188. [https://doi.org/10.1016/0006-8993\(94\)91195-9](https://doi.org/10.1016/0006-8993(94)91195-9)
- Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia*, 9(1), 63-75.e2.  
<https://doi.org/10.1016/j.jalz.2012.11.007>
- Puranik, N., & Song, M. (2024). Glutamate: Molecular Mechanisms and Signaling Pathway in Alzheimer's Disease, a Potential Therapeutic Target. *Molecules*, 29(23), 5744. <https://doi.org/10.3390/molecules29235744>
- Puzzo, D., Privitera, L., Leznik, E., Fà, M., Staniszewski, A., Palmeri, A., & Arancio, O. (2008). Picomolar Amyloid- $\beta$  Positively Modulates Synaptic

- Plasticity and Memory in Hippocampus. *The Journal of Neuroscience*, 28(53), 14537–14545. <https://doi.org/10.1523/JNEUROSCI.2692-08.2008>
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing*.
- Radenovic, L., Korenic, A., Maleeva, G., Osadchenko, I., Kovalenko, T., & Skibo, G. (2011). Comparative Ultrastructural Analysis of Mitochondria in the CA1 and CA3 Hippocampal Pyramidal Cells Following Global Ischemia in Mongolian Gerbils. *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, 294(6), 1057–1065.  
<https://doi.org/10.1002/ar.21390>
- Rajmohan, R., & Reddy, P. H. (2017). Amyloid-Beta and Phosphorylated Tau Accumulations Cause Abnormalities at Synapses of Alzheimer's disease Neurons. *Journal of Alzheimer's Disease*, 57(4), 975–999.  
<https://doi.org/10.3233/JAD-160612>
- Rasool, C. G., Abraham, C., Anderton, B. H., Haugh-, M., Kahn, J., & Selkoe, D. J. (1984). Alzheimer's Disease: Immunoreactivity of Neurofibrillary Tangles With Anti-Neurofilament and Anti-Paired Helical Filament Antibodies. *Brain Research*.
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, gkz369.  
<https://doi.org/10.1093/nar/gkz369>
- Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, 12(1), 480.  
<https://doi.org/10.1186/1471-2105-12-480>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. <https://doi.org/10.1093/nar/gkv007>

Roberts, D. S., Loo, J. A., Tsybin, Y. O., Liu, X., Wu, S., Chamot-Rooke, J., Agar, J. N., Paša-Tolić, L., Smith, L. M., & Ge, Y. (2024). Top-down proteomics. *Nature Reviews Methods Primers*, 4(1), 38.

<https://doi.org/10.1038/s43586-024-00318-2>

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.

<https://doi.org/10.1093/bioinformatics/btp616>

Rodríguez, J. J., Noristani, H. N., Hilditch, T., Olabarria, M., Yeh, C. Y., Witton, J., & Verkhratsky, A. (2013). Increased densities of resting and activated microglia in the dentate gyrus follow senile plaque formation in the CA1 subfield of the hippocampus in the triple transgenic model of Alzheimer's disease. *Neuroscience Letters*, 552, 129–134.

<https://doi.org/10.1016/j.neulet.2013.06.036>

Rodríguez-Manotas, M., Amorín-Díaz, M., Cabezas-Herrera, J., Acedo-Martínez, A., & Llorca-Escuín, I. (2012). Are γ-secretase and its associated Alzheimer's disease γ problems? *Medical Hypotheses*, 78(2), 299–304.

<https://doi.org/10.1016/j.mehy.2011.11.007>

Rottner, K., Stradal, T. E. B., & Chen, B. (2021). WAVE regulatory complex.

*Current Biology*, 31(10), R512–R517.

<https://doi.org/10.1016/j.cub.2021.01.086>

Ru, Q., Li, Y., Chen, L., Wu, Y., Min, J., & Wang, F. (2024). Iron homeostasis and ferroptosis in human diseases: Mechanisms and therapeutic prospects. *Signal Transduction and Targeted Therapy*, 9(1), 271.

<https://doi.org/10.1038/s41392-024-01969-z>

Rubenstein, K. (2002). The Once and Future Microarray Market. *Drug Discovery World*. <https://www.ddw-online.com/the-once-and-future-microarray-market-1253-200210/>

- Rybicka, M., Stalke, P., & Bielawski, K. P. (2016). Current molecular methods for the detection of hepatitis B virus quasispecies. *Reviews in Medical Virology*, 26(5), 369–381. <https://doi.org/10.1002/rmv.1897>
- Sahoo, S. S., Mishra, C., Rout, M., Nayak, G., Mohanty, S. T., & Panigrahy, K. K. (2016). Comparative in silico and protein-protein interaction network analysis of ATP1A1 gene. *Gene Reports*, 5, 134–139. <https://doi.org/10.1016/j.genrep.2016.10.004>
- Salomon, R., Kaczorowski, D., Valdes-Mora, F., Nordon, R. E., Neild, A., Farbehi, N., Bartonicek, N., & Gallego-Ortega, D. (2019). Droplet-based single cell RNAseq tools: A practical guide. *Lab on a Chip*, 19(10), 1706–1727. <https://doi.org/10.1039/c8lc01239c>
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>
- Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5), 495–502. <https://doi.org/10.1038/nbt.3192>
- Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N., Mumbach, M. R., Pierce, S. E., Corces, M. R., Shah, P., Bell, J. C., Jhutty, D., Nemec, C. M., Wang, J., Wang, L., Yin, Y., Giresi, P. G., Chang, A. L. S., ... Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37(8), 925–936. <https://doi.org/10.1038/s41587-019-0206-z>
- Saxena, S., & Caroni, P. (2011). Selective Neuronal Vulnerability in Neurodegenerative Diseases: From Stressor Thresholds to Degeneration. *Neuron*, 71(1), 35–48. <https://doi.org/10.1016/j.neuron.2011.06.031>
- Scheepbouwer, C., Hackenberg, M., van Eijndhoven, M. A. J., Gerber, A., Pegtel, M., & Gómez-Martín, C. (2023). NORMSEQ: A tool for evaluation, selection

- and visualization of RNA-Seq normalization methods. *Nucleic Acids Research*, 51(W1), W372–W378. <https://doi.org/10.1093/nar/gkad429>
- Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & Van Der Flier, W. M. (2021). Alzheimer's disease. *The Lancet*, 397(10284), 1577–1590. [https://doi.org/10.1016/S0140-6736\(20\)32205-4](https://doi.org/10.1016/S0140-6736(20)32205-4)
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235), 467–470. <https://doi.org/10.1126/science.270.5235.467>
- Schmidt, M. L., & Trojanowski, J. Q. (1990). Relative Abundance of Tau and Neurofilament Epitopes in Hippocampal Neurofibrillary Tangles. *American Journal of Pathology*, 136(5).
- Schneider, A., Biernat, J., Von Bergen, M., Mandelkow, E., & Mandelkow, E.-M. (1999). Phosphorylation that Detaches Tau Protein from Microtubules (Ser262, Ser214) Also Protects It against Aggregation into Alzheimer Paired Helical Filaments. *Biochemistry*, 38(12), 3549–3558. <https://doi.org/10.1021/bi981874p>
- Selkoe, D. J., & Hardy, J. (2016). The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine*, 8(6), 595–608. <https://doi.org/10.15252/emmm.201606210>
- Shelton, L. B., Koren, J., & Blair, L. J. (2017). Imbalances in the Hsp90 Chaperone Machinery: Implications for Tauopathies. *Frontiers in Neuroscience*, 11, 724. <https://doi.org/10.3389/fnins.2017.00724>
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), 1135–1145. <https://doi.org/10.1038/nbt1486>
- Sheu, K.-F. R., Brown, A. M., Haroutunian, V., Kristal, B. S., Thaler, H., Lesser, M., Kalaria, R. N., Relkin, N. R., Mohs, R. C., Lilius, L., Lannfelt, L., & Blass, J. P. (1999). Modulation by DLST of the genetic risk of Alzheimer's disease in a

- very elderly population. *Annals of Neurology*, 45(1), 48–53.  
[https://doi.org/10.1002/1531-8249\(199901\)45:1%253C48::AID-ART9%253E3.0.CO;2-V](https://doi.org/10.1002/1531-8249(199901)45:1%253C48::AID-ART9%253E3.0.CO;2-V)
- Shireman, J. M., Cheng, L., Goel, A., Garcia, D. M., Partha, S., Quiñones-Hinojosa, A., Kendzierski, C., & Dey, M. (2023). Spatial transcriptomics in glioblastoma: Is knowing the right zip code the key to the next therapeutic breakthrough? *Frontiers in Oncology*, 13, 1266397.  
<https://doi.org/10.3389/fonc.2023.1266397>
- Shuken, S. R. (2023). An Introduction to Mass Spectrometry-Based Proteomics. *Journal of Proteome Research*, 22(7), 2151–2171.  
<https://doi.org/10.1021/acs.jproteome.2c00838>
- Siedler, D. G., Chuah, M. I., Kirkcaldie, M. T. K., Vickers, J. C., & King, A. E. (2014). Diffuse axonal injury in brain trauma: Insights from alterations in neurofilaments. *Frontiers in Cellular Neuroscience*, 8.  
<https://doi.org/10.3389/fncel.2014.00429>
- Simillion, C., Liechti, R., Lischer, H. E. L., Ioannidis, V., & Bruggmann, R. (2017). Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, 18(1), 151. <https://doi.org/10.1186/s12859-017-1571-6>
- Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry-based proteomics. *The Biochemist*, 42(5), 64–69.  
<https://doi.org/10.1042/BIO20200057>
- Sirisi, S., Sánchez-Aced, É., Belbin, O., & Lleó, A. (2024). APP dyshomeostasis in the pathogenesis of Alzheimer's disease: Implications for current drug targets. *Alzheimer's Research & Therapy*, 16(1), 144.  
<https://doi.org/10.1186/s13195-024-01504-w>
- Sivanich, M. K., Gu, T., Tabang, D. N., & Li, L. (2022). Recent advances in isobaric labeling and applications in quantitative proteomics. *Proteomics*, 22(19–20), 2100256. <https://doi.org/10.1002/pmic.202100256>

- Sluchanko, N. N., & Gusev, N. B. (2011). Probable Participation of 14-3-3 in Tau Protein Oligomerization and Aggregation. *Journal of Alzheimer's Disease*, 27(3), 467–476. <https://doi.org/10.3233/JAD-2011-110692>
- Sok, J., Wang, X.-Z., Batchvarova, N., Kuroda, M., Harding, H., & Ron, D. (1999). CHOP-Dependent Stress-Inducible Expression of a Novel Form of Carbonic Anhydrase VI. *Molecular and Cellular Biology*, 19(1), 495–504. <https://doi.org/10.1128/MCB.19.1.495>
- Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4), 255–261. <https://doi.org/10.1038/nmeth.4612>
- Spillantini, M. G., Goedert, M., Crowther, R. A., Murrell, J. R., Farlow, M. R., & Ghetti, B. (1997). Familial multiple system tauopathy with presenile dementia: A disease with abundant neuronal and glial tau filaments. *Proceedings of the National Academy of Sciences*, 94(8), 4113–4118. <https://doi.org/10.1073/pnas.94.8.4113>
- Staal, Y. C., Van Herwijnen, M. H., Van Schooten, F. J., & Van Delft, J. H. (2005). Application of four dyes in gene expression analyses by microarrays. *BMC Genomics*, 6(1), 101. <https://doi.org/10.1186/1471-2164-6-101>
- Stamer, K., Vogel, R., Thies, E., Mandelkow, E., & Mandelkow, E.-M. (2002). Tau blocks traffic of organelles, neurofilaments, and APP vesicles in neurons and enhances oxidative stress. *The Journal of Cell Biology*, 156(6), 1051–1063. <https://doi.org/10.1083/jcb.200108057>
- Stanika, R. I., Winters, C. A., Pivovarova, N. B., & Andrews, S. B. (2010). Differential NMDA receptor-dependent calcium loading and mitochondrial dysfunction in CA1 vs. CA3 hippocampal neurons. *Neurobiology of Disease*, 37(2), 403–411. <https://doi.org/10.1016/j.nbd.2009.10.020>
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>

- Stelzl, L. S., Pietrek, L. M., Holla, A., Oroz, J., Sikora, M., Köfinger, J., Schuler, B., Zweckstetter, M., & Hummer, G. (2022). Global Structure of the Intrinsically Disordered Protein Tau Emerges from Its Local Structure. *JACS Au*, 2(3), 673–686. <https://doi.org/10.1021/jacsau.1c00536>
- Stranahan, A. M., & Mattson, M. P. (2010). Selective Vulnerability of Neurons in Layer II of the Entorhinal Cortex during Aging and Alzheimer's Disease. *Neural Plasticity*, 2010, 1–8. <https://doi.org/10.1155/2010/108190>
- Strang, K. H., Golde, T. E., & Giasson, B. I. (2019). MAPT mutations, tauopathy, and mechanisms of neurodegeneration. *Laboratory Investigation*, 99(7), 912–928. <https://doi.org/10.1038/s41374-019-0197-x>
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7), 1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Takata, K., Kitamura, Y., Nakata, Y., Matsuoka, Y., Tomimoto, H., Taniguchi, T., & Shimohama, S. (2009). Involvement of WAVE Accumulation in A $\beta$ /APP Pathology-Dependent Tangle Modification in Alzheimer's Disease. *The American Journal of Pathology*, 175(1), 17–24. <https://doi.org/10.2353/ajpath.2009.080908>
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), 377–382. <https://doi.org/10.1038/nmeth.1315>

- Thal, D. R., Rüb, U., Orantes, M., & Braak, H. (2002). Phases of A $\beta$ -deposition in the human brain and its relevance for the development of AD. *Neurology*, 58(12), 1791–1800. <https://doi.org/10.1212/WNL.58.12.1791>
- Tomfohr, J., Lu, J., & Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1), 225. <https://doi.org/10.1186/1471-2105-6-225>
- Tortosa, E., Santa-Maria, I., Moreno, F., Lim, F., Perez, M., & Avila, J. (2009). Binding of Hsp90 to Tau Promotes a Conformational Change and Aggregation of Tau Protein. *Journal of Alzheimer's Disease*, 17(2), 319–325. <https://doi.org/10.3233/JAD-2009-1049>
- Trojanowski, J. Q. (1998). Aggregation of Neurofilament and  $\alpha$ -Synuclein Proteins in Lewy Bodies. *ARCH NEUROL*, 55.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520–525. <https://doi.org/10.1093/bioinformatics/17.6.520>
- Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9), 731–740. <https://doi.org/10.1038/nmeth.3901>
- Uchida, Y., Tachikawa, M., Obuchi, W., Hoshi, Y., Tomioka, Y., Ohtsuki, S., & Terasaki, T. (2013). A study protocol for quantitative targeted absolute proteomics (QTAP) by LC-MS/MS: Application for inter-strain differences in protein expression levels of transporters, receptors, claudin-5, and marker proteins at the blood-brain barrier in ddY, FVB, and C57BL/6J mice. *Fluids and Barriers of the CNS*, 10(1), 21. <https://doi.org/10.1186/2045-8118-10-21>
- Uemura, N., Uemura, M. T., Luk, K. C., Lee, V. M.-Y., & Trojanowski, J. Q. (2020). Cell-to-Cell Transmission of Tau and  $\alpha$ -Synuclein. *Trends in Molecular*

- Medicine*, 26(10), 936–952.  
<https://doi.org/10.1016/j.molmed.2020.03.012>
- Uhlen, M., & Quake, S. R. (2023). Sequential sequencing by synthesis and the next-generation sequencing revolution. *Trends in Biotechnology*, 41(12), 1565–1572. <https://doi.org/10.1016/j.tibtech.2023.06.007>
- Umahara, T., Uchihara, T., Tsuchiya, K., Nakamura, A., Iwamoto, T., Ikeda, K., & Takasaki, M. (2004). 14-3-3 proteins and zeta isoform containing neurofibrillary tangles in patients with Alzheimer's disease. *Acta Neuropathologica*, 108(4), 279–286. <https://doi.org/10.1007/s00401-004-0885-4>
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., ... Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291.
- Vezyroglou, A., Akilapa, R., Barwick, K., Koene, S., Brownstein, C. A., Holder-Espinasse, M., Fry, A. E., Németh, A. H., Tofaris, G. K., Hay, E., Hughes, I., Mansour, S., Mordekar, S. R., Splitt, M., Turnpenny, P. D., Demetriou, D., Koopmann, T. T., Ruivenkamp, C. A. L., Agrawal, P. B., ... Balasubramanian, M. (2022). The Phenotypic Continuum of ATP1A3-Related Disorders. *Neurology*, 99(14).  
<https://doi.org/10.1212/WNL.0000000000200927>
- Vickers, J. C., Kirkcaldie, M. T., Phipps, A., & King, A. E. (2016). Alterations in neurofilaments and the transformation of the cytoskeleton in axons may provide insight into the aberrant neuronal changes of Alzheimer's disease. *Brain Research Bulletin*, 126, 324–333.  
<https://doi.org/10.1016/j.brainresbull.2016.07.012>
- Vogels, T., Leuzy, A., Cicognola, C., Ashton, N. J., Smolek, T., Novak, M., Blennow, K., Zetterberg, H., Hromadka, T., Zilka, N., & Schöll, M. (2019).

- Propagation of Tau Pathology: Integrating Insights From Postmortem and In Vivo Studies. *Biological Psychiatry*, S0006322319317445.  
<https://doi.org/10.1016/j.biopsych.2019.09.019>
- Wang, H., Wu, M., Zhan, C., Ma, E., Yang, M., Yang, X., & Li, Y. (2012). Neurofilament proteins in axonal regeneration and neurodegenerative diseases. *Neural Regeneration Research*.
- Wang, J., Tung, Y. C., Wang, Y., Li, X. T., Iqbal, K., & Grundke-Iqbal, I. (2001). Hyperphosphorylation and accumulation of neurofilament proteins in Alzheimer disease brain and in okadaic acid-treated SY5Y cells. *FEBS Letters*, 507(1), 81–87. [https://doi.org/10.1016/S0014-5793\(01\)02944-1](https://doi.org/10.1016/S0014-5793(01)02944-1)
- Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., & Von Mering, C. (2012). PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. *Molecular & Cellular Proteomics*, 11(8), 492–500. <https://doi.org/10.1074/mcp.O111.014704>
- Wang, X., Lian, Q., Dong, H., Xu, S., Su, Y., & Wu, X. (2024). Benchmarking Algorithms for Gene Set Scoring of Single-cell ATAC-seq Data. *Genomics, Proteomics & Bioinformatics*, 22(2), qzae014.  
<https://doi.org/10.1093/gpbjnl/qzae014>
- Wang, X., Pal, R., Chen, X., Limpeanchob, N., Kumar, K. N., & Michaelis, E. K. (2005). High intrinsic oxidative stress may underlie selective vulnerability of the hippocampal CA1 region. *Molecular Brain Research*, 140(1-2), 120–126. <https://doi.org/10.1016/j.molbrainres.2005.07.018>
- Wang, X.-L., & Li, L. (2021). Cell type-specific potential pathogenic genes and functional pathways in Alzheimer's Disease. *BMC Neurology*, 21(1), 381. <https://doi.org/10.1186/s12883-021-02407-1>
- Wang, Y., & Mattson, M. P. (2014). L-type Ca<sub>2+</sub> currents at CA1 synapses, but not CA3 or dentate granule neuron synapses, are increased in 3xTgAD mice in an age-dependent manner. *Neurobiology of Aging*, 8.

- Wang, Z., Karkossa, I., Großkopf, H., Rolle-Kampczyk, U., Hackermüller, J., Von Bergen, M., & Schubert, K. (2021). Comparison of quantitation methods in proteomics to define relevant toxicological information on AhR activation of HepG2 cells by BaP. *Toxicology*, 448, 152652.  
<https://doi.org/10.1016/j.tox.2020.152652>
- Wang, Z.-T., Zhang, C., Wang, Y.-J., Dong, Q., Tan, L., & Yu, J.-T. (2020). Selective neuronal vulnerability in Alzheimer's disease. *Ageing Research Reviews*, 62, 101114. <https://doi.org/10.1016/j.arr.2020.101114>
- Ward, B., Pyr Dit Ruys, S., Balligand, J.-L., Belkhir, L., Cani, P. D., Collet, J.-F., De Greef, J., Dewulf, J. P., Gatto, L., Haufroid, V., Jodogne, S., Kabamba, B., Lingurski, M., Yombi, J. C., Vertommen, D., & Elens, L. (2024). Deep plasma proteomics with data-independent acquisition: A fastlane towards biomarkers identification. *bioRxiv*.  
<https://doi.org/10.1101/2024.02.23.581160>
- Watamura, N., Toba, J., Yoshii, A., Nikkuni, M., & Ohshima, T. (2016). Colocalization of phosphorylated forms of WAVE1, CRMP2, and tau in Alzheimer's disease model mice: Involvement of Cdk5 phosphorylation and the effect of ATRA treatment. *Journal of Neuroscience Research*, 94(1), 15–26. <https://doi.org/10.1002/jnr.23674>
- Webel, H., Niu, L., Nielsen, A. B., Locard-Paulet, M., Mann, M., Jensen, L. J., & Rasmussen, S. (2024). Imputation of label-free quantitative mass spectrometry-based proteomics data using self-supervised deep learning. *Nature Communications*, 15(1), 5405. <https://doi.org/10.1038/s41467-024-48711-5>
- Weggen, S., & Beher, D. (2012). Molecular consequences of amyloid precursor protein and presenilin mutations causing autosomal-dominant Alzheimer's disease. *Alzheimer's Research & Therapy*, 4(2), 9.  
<https://doi.org/10.1186/alzrt107>

- Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, 8(1), 663. <https://doi.org/10.1038/s41598-017-19120-0>
- Weston, P. S. J., Poole, T., Ryan, N. S., Nair, A., Liang, Y., Macpherson, K., Druyeh, R., Malone, I. B., Ahsan, R. L., Pemberton, H., Klimova, J., Mead, S., Blennow, K., Rossor, M. N., Schott, J. M., Zetterberg, H., & Fox, N. C. (2017). Serum neurofilament light in familial Alzheimer disease: A marker of early neurodegeneration. *Neurology*, 89(21), 2167-2175. <https://doi.org/10.1212/WNL.0000000000004667>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wijesooriya, K., Jadaan, S. A., Perera, K. L., Kaur, T., & Ziemann, M. (2022). Urgent need for consistent standards in functional enrichment analysis. *PLoS Computational Biology*, 18(3), e1009935. <https://doi.org/10.1371/journal.pcbi.1009935>
- Wilde, G. J. C., Pringle, A. K., Wright, P., & Iannotti, F. (2002). Differential Vulnerability of the CA1 and CA3 Subfields of the Hippocampus to Superoxide and Hydroxyl Radicals In Vitro. *Journal of Neurochemistry*, 69(2), 883-886. <https://doi.org/10.1046/j.1471-4159.1997.69020883.x>
- Wilkinson, J. (2021). High-performance liquid chromatography-tandem mass spectrometry for analysis of aquatic contaminants: A high-level introduction to the technique. In *Monitoring Environmental Contaminants* (pp. 1-17). Elsevier. <https://doi.org/10.1016/B978-0-444-64335-3.00004-9>
- Willems, P., Fels, U., Staes, A., Gevaert, K., & Van Damme, P. (2021). Use of Hybrid Data-Dependent and -Independent Acquisition Spectral Libraries Empowers Dual-Proteome Profiling. *Journal of Proteome Research*, 20(2), 1165-1177. <https://doi.org/10.1021/acs.jproteome.0c00350>
- Witter, M. P., Doan, T. P., Jacobsen, B., Nilssen, E. S., & Ohara, S. (2017). Architecture of the Entorhinal Cortex A Review of Entorhinal Anatomy in

- Rodents with Some Comparative Notes. *Frontiers in Systems Neuroscience*, 11, 46. <https://doi.org/10.3389/fnsys.2017.00046>
- Witter, M. P., & Moser, E. I. (2006). Spatial representation and the architecture of the entorhinal cortex. *Trends in Neurosciences*, 29(12), 671–678. <https://doi.org/10.1016/j.tins.2006.10.003>
- Wöhrle, J., Krämer, S. D., Meyer, P. A., Rath, C., Hügle, M., Urban, G. A., & Roth, G. (2020). Digital DNA microarray generation on glass substrates. *Scientific Reports*, 10(1), 5770. <https://doi.org/10.1038/s41598-020-62404-1>
- Wolswijk, G. (2003). Changes in the expression and localization of the paranodal protein Caspr on axons in chronic multiple sclerosis. *Brain*, 126(7), 1638–1649. <https://doi.org/10.1093/brain/awg151>
- Wu, A., Lee, D., & Xiong, W.-C. (2024). VPS35 or retromer as a potential target for neurodegenerative disorders: Barriers to progress. *Expert Opinion on Therapeutic Targets*, 28(8), 701–712. <https://doi.org/10.1080/14728222.2024.2392700>
- Wu, J., Anwyl, R., & Rowan, M. J. (1995).  $\beta$ -Amyloid selectively augments NMDA receptor-mediated synaptic transmission in rat hippocampus. *NeuroReport*.
- Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman and Hall/CRC. <https://bookdown.org/yihui/rmarkdown>
- Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman and Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>
- Xu, C., Fu, X., Zhu, S., & Liu, J.-J. (2016). Retrolinkin recruits the WAVE1 protein complex to facilitate BDNF-induced TrkB endocytosis and dendrite outgrowth. *Molecular Biology of the Cell*, 27(21), 3342–3356. <https://doi.org/10.1091/mbc.E16-05-0326>
- Xu, Y., Zhang, S., & Zheng, H. (2019). The cargo receptor SQSTM1 ameliorates neurofibrillary tangle pathology and spreading through selective

- targeting of pathological MAPT (microtubule associated protein tau). *Autophagy*, 15(4), 583–598.  
<https://doi.org/10.1080/15548627.2018.1532258>
- Xu, Y., Zhao, M., Han, Y., & Zhang, H. (2020). GABAergic Inhibitory Interneuron Deficits in Alzheimer's Disease: Implications for Treatment. *Frontiers in Neuroscience*, 14, 660. <https://doi.org/10.3389/fnins.2020.00660>
- Yadav, P., Selvaraj, B. T., Bender, F. L. P., Behringer, M., Moradi, M., Sivadasan, R., Dombert, B., Blum, R., Asan, E., Sauer, M., Julien, J.-P., & Sendtner, M. (2016). Neurofilament depletion improves microtubule dynamics via modulation of Stat3/stathmin signaling. *Acta Neuropathologica*, 132(1), 93–110. <https://doi.org/10.1007/s00401-016-1564-y>
- Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C., & Bu, G. (2019). Apolipoprotein E and Alzheimer disease: Pathobiology and targeting strategies. *Nature Reviews Neurology*, 15(9), 501–518. <https://doi.org/10.1038/s41582-019-0228-7>
- Yang, H., & Hu, H. (2016). Sequestration of cellular interacting partners by protein aggregates: Implication in a loss-of-function pathology. *The FEBS Journal*, 283(20), 3705–3717. <https://doi.org/10.1111/febs.13722>
- Yang, Y., Kong, B., Jung, Y., Park, J.-B., Oh, J.-M., Hwang, J., Cho, J. Y., & Kweon, D.-H. (2018). Soluble N-Ethylmaleimide-Sensitive Factor Attachment Protein Receptor-Derived Peptides for Regulation of Mast Cell Degranulation. *Frontiers in Immunology*, 9, 725.  
<https://doi.org/10.3389/fimmu.2018.00725>
- Yokota, Y., Ghashghaei, H. T., Han, C., Watson, H., Campbell, K. J., & Anton, E. S. (2007). Radial Glial Dependent and Independent Dynamics of Interneuronal Migration in the Developing Cerebral Cortex. *PLoS ONE*.
- Yuan, A., & Nixon, R. A. (2021). Neurofilament Proteins as Biomarkers to Monitor Neurological Diseases and the Efficacy of Therapies. *Frontiers in Neuroscience*, 15, 689938. <https://doi.org/10.3389/fnins.2021.689938>

- Yuan, A., Rao, M. V., Veeranna, & Nixon, R. A. (2017). Neurofilaments and Neurofilament Proteins in Health and Disease. *Cold Spring Harbor Perspectives in Biology*, 9(4), a018309.  
<https://doi.org/10.1101/cshperspect.a018309>
- Yuan, A., Sasaki, T., Kumar, A., Peterhoff, C. M., Rao, M. V., Liem, R. K., Julien, J.-P., & Nixon, R. A. (2012). Peripherin Is a Subunit of Peripheral Nerve Neurofilaments: Implications for Differential Vulnerability of CNS and Peripheral Nervous System Axons. *Journal of Neuroscience*, 32(25), 8501-8508. <https://doi.org/10.1523/JNEUROSCI.1081-12.2012>
- Yuan, M., Wang, Y., Huang, Z., Jing, F., Qiao, P., Zou, Q., Li, J., & Cai, Z. (2023). Impaired autophagy in amyloid-beta pathology: A traditional review of recent Alzheimer's research. *The Journal of Biomedical Research*, 37(1), 30. <https://doi.org/10.7555/JBR.36.20220145>
- Yuan, Z., Agarwal-Mawal, A., & Paudel, H. K. (2004). 14-3-3 Binds to and Mediates Phosphorylation of Microtubule-associated Tau Protein by Ser9-phosphorylated Glycogen Synthase Kinase 3 $\beta$  in the Brain. *Journal of Biological Chemistry*, 279(25), 26105-26114.  
<https://doi.org/10.1074/jbc.M308298200>
- Yudkoff, M., Nelson, D., Daikhin, Y., & Erecińska, M. (1994). Tricarboxylic acid cycle in rat brain synaptosomes. Fluxes and interactions with aspartate aminotransferase and malate/aspartate shuttle. *Journal of Biological Chemistry*, 269(44), 27414-27420. [https://doi.org/10.1016/S0021-9258\(18\)47001-9](https://doi.org/10.1016/S0021-9258(18)47001-9)
- Zámbó, B., Várady, G., Padányi, R., Szabó, E., Németh, A., Langó, T., Enyedi, Á., & Sarkadi, B. (2017). Decreased calcium pump expression in human erythrocytes is connected to a minor haplotype in the ATP2B4 gene. *Cell Calcium*, 65, 73-79. <https://doi.org/10.1016/j.ceca.2017.02.001>

- Zappia, L., & Oshlack, A. (2018). Clustering trees: A visualization for evaluating clusterings at multiple resolutions. *GigaScience*, 7(7), giy083.  
<https://doi.org/10.1093/gigascience/giy083>
- Zhang, D., Mably, A. J., Walsh, D. M., & Rowan, M. J. (2016). Peripheral Interventions Enhancing Brain Glutamate Homeostasis Relieve Amyloid  $\beta$ - and TNF $\alpha$ - Mediated Synaptic Plasticity Disruption in the Rat Hippocampus. *Cerebral Cortex*, cercor; bhw193v1.  
<https://doi.org/10.1093/cercor/bhw193>
- Zhang, D., Xiao, M., Wang, L., & Jia, W. (2019). Blood-Based Glutamate Scavengers Reverse Traumatic Brain Injury-Induced Synaptic Plasticity Disruption by Decreasing Glutamate Level in Hippocampus Interstitial Fluid, but Not Cerebral Spinal Fluid, In Vivo. *Neurotoxicity Research*, 35(2), 360–372. <https://doi.org/10.1007/s12640-018-9961-8>
- Zhang, H., Delafield, D. G., & Li, L. (2023). Mass spectrometry imaging: The rise of spatially resolved single-cell omics. *Nature Methods*, 20(3), 327–330.  
<https://doi.org/10.1038/s41592-023-01774-6>
- Zhang, H., Jiang, X., Ma, L., Wei, W., Li, Z., Chang, S., Wen, J., Sun, J., & Li, H. (2022). Role of A $\beta$  in Alzheimer's-related synaptic dysfunction. *Frontiers in Cell and Developmental Biology*, 10, 964075.  
<https://doi.org/10.3389/fcell.2022.964075>
- Zhang, H., Wei, W., Zhao, M., Ma, L., Jiang, X., Pei, H., Cao, Y., & Li, H. (2021). Interaction between A $\beta$  and Tau in the Pathogenesis of Alzheimer's Disease. *International Journal of Biological Sciences*, 17(9), 2181–2192.  
<https://doi.org/10.7150/ijbs.57078>
- Zhang, J., Zhang, Y., Wang, J., Xia, Y., Zhang, J., & Chen, L. (2024). Recent advances in Alzheimer's disease: Mechanisms, clinical trials and new drug development strategies. *Signal Transduction and Targeted Therapy*, 9(1), 211. <https://doi.org/10.1038/s41392-024-01911-3>

- Zhang, W., Jiao, B., Xiao, T., Liu, X., Liao, X., Xiao, X., Guo, L., Yuan, Z., Yan, X., Tang, B., & Shen, L. (2020). Association of rare variants in neurodegenerative genes with familial Alzheimer's disease. *Annals of Clinical and Translational Neurology*, 7(10), 1985–1995.  
<https://doi.org/10.1002/acn3.51197>
- Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., & Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, 73(1), 130-142.e5.  
<https://doi.org/10.1016/j.molcel.2018.10.020>
- Zhang, X., Smits, A. H., van Tilburg, G. B., Ovaa, H., Huber, W., & Vermeulen, M. (2018). Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nature Protocols*, 13(3), 530–550.  
<https://doi.org/10.1038/nprot.2017.147>
- Zhu, K., Wang, X., Sun, B., Wu, J., Lu, H., Zhang, X., Liang, H., Zhang, D., & Liu, C. (2019). Primary Age-Related Tauopathy in Human Subcortical Nuclei. *Frontiers in Neuroscience*, 13, 529.  
<https://doi.org/10.3389/fnins.2019.00529>
- Zhu, Y., Kong, L., Han, T., Yan, Q., & Liu, J. (2023). Machine learning identification and immune infiltration of disulfidoptosis-related Alzheimer's disease molecular subtypes. *Immunity, Inflammation and Disease*, 11(10), e1037. <https://doi.org/10.1002/iid3.1037>
- Zhuang, H., Wang, H., & Ji, Z. (2022). findPC: An R package to automatically select the number of principal components in single-cell analysis. *Bioinformatics*, 38(10), 2949–2951.  
<https://doi.org/10.1093/bioinformatics/btac235>
- Zimmerman, K. D., Espeland, M. A., & Langefeld, C. D. (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*, 12(1), 738. <https://doi.org/10.1038/s41467-021-21038-1>

- Zuccato, C., & Cattaneo, E. (2009). Brain-derived neurotrophic factor in neurodegenerative diseases. *Nature Reviews Neurology*, 5(6), 311-322.  
<https://doi.org/10.1038/nrneurol.2009.54>
- Zuehlke, A. D., Beebe, K., Neckers, L., & Prince, T. (2015). Regulation and function of the human HSP90AA1 gene. *Gene*, 570(1), 8-16.  
<https://doi.org/10.1016/j.gene.2015.06.018>
- Zyla, J., Marczyk, M., Weiner, J., & Polanska, J. (2017). Ranking metrics in gene set enrichment analysis: Do they matter? *BMC Bioinformatics*, 18(1).  
<https://doi.org/10.1186/s12859-017-1674-0>

## 10. Appendix

### Derivation A.

Derivation of simplified MAD (Mean Absolute Deviation) equation in sets with only one non-zero value.

Starting with the general MAD equation (Aghili-Ashtiani, 2021):

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Assume a vector  $x = [a, 0, 0, \dots, 0]$  of length  $n$ , where only the first value is non-zero.

For the non-zero entry:

$$\left|a - \frac{a}{n}\right| = a \left(1 - \frac{1}{n}\right) = a \cdot \frac{n-1}{n}$$

For each of the zero entries:

$$\left|0 - \frac{a}{n}\right| = \frac{a}{n}$$

Upon summing deviations:

$$\sum |x_i - \bar{x}| = a \cdot \frac{n-1}{n} + (n-1) \cdot \frac{a}{n}$$

Simplified:

$$\frac{a(n-1)}{n} + \frac{a(n-1)}{n} = \frac{2a(n-1)}{n}$$

After dividing  $n$  to arrive at the MAD:

$$\text{MAD} = \frac{1}{n} \cdot \frac{2a(n-1)}{n} = \frac{2a(n-1)}{n^2}$$