UNIVERSITY OF CALIFORNIA,
IRVINE

Computational approach to characterize gene dynamics using bulk and single-nucleus RNA sequencing to study Alzheimer's disease

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology (MCSB)

by

Narges Rezaie

Dissertation Committee:
Ali Mortazavi, Chair
Kim Green
Vivek Swarup

2024

DEDICATION

To Saeed, Farideh, Niloofar, and Amirhosein
For their unconditional love and support.

TABLE OF CONTENTS

Page

LIST OF FIGURES **vi**

LIST OF TABLES **viii**

ACKNOWLEDGMENTS **ix**

VITA **x**

ABSTRACT OF THE DISSERTATION **xiii**

1 Introduction **1**
- 1.1 Abstract . . . 1
- 1.2 Introduction . . . 2
  - 1.2.1 Cell types and lineages in the brain . . . 2
  - 1.2.2 Alzheimer's disease . . . 5
  - 1.2.3 The role of glial cells during neurodegeneration . . . 7
  - 1.2.4 Mouse models of AD . . . 9
  - 1.2.5 Infer modules from bulk RNA-seq . . . 11
- 1.3 Conclusions (Theme of thesis) . . . 15

2 PyWGCNA: a Python package for weighted gene co-expression network analysis **18**
- 2.1 Abstract . . . 18
- 2.2 Introduction . . . 19
- 2.3 Materials and methods . . . 20
  - 2.3.1 Identifying co-expression modules . . . 20
  - 2.3.2 Assessing co-expression module overlap between PyWGCNA objects or to single-cell data . . . 22
- 2.4 Results . . . 22
  - 2.4.1 Analysis of bulk RNA-seq of 5xFAD and 3xTgAD mouse model . . . 23
  - 2.4.2 Analysis of bulk RNA-seq of GWAS mouse models . . . 24
- 2.5 Discussion . . . 34
- 2.6 Supplementary methods . . . 36
  - 2.6.1 co-expression network construction . . . 36
- 2.7 Materials and data availability . . . 38

3 Identification of robust cellular programs using reproducible LDA that impact sex-specific disease progression in different genotypes of a mouse model of AD **59**
3.1 Abstract . . . 59
3.2 Introduction . . . 60
3.3 Results . . . 62
3.3.1 Reproducible LDA topics using Topyfic . . . 62
3.3.2 Comparing a mouse model of AD across two genetic backgrounds . . . 64
3.3.3 Identifying topics related to cell type and cell state . . . 66
3.3.4 Recovering topics for different activation levels in microglia scRNA-seq 68
3.3.5 Topics derived from regulatory genes are sufficient to define cell types and cell states . . . 70
3.4 Discussion . . . 72
3.5 Materials . . . 75
3.5.1 Mice and tissue collection . . . 75
3.5.2 Single-nucleus isolation and fixation . . . 75
3.5.3 Microglia single-cell isolation and fixation . . . 76
3.6 Methods . . . 77
3.6.1 Datasets . . . 77
3.6.2 Preprocessing scRNA-seq and snRNA-seq data . . . 77
3.6.3 Parse Biosciences Split-seq Experiments . . . 79
3.6.4 Isolation of RNA for bulk assays . . . 80
3.6.5 Bulk RNA-seq from mouse tissues . . . 80
3.6.6 Bulk microRNA-seq from mouse tissues . . . 81
3.6.7 Bulk read mapping and quantification . . . 81
3.6.8 Bulk RNA-seq and microRNA-seq integrated analysis . . . 82
3.6.9 Selection of regulatory genes . . . 82
3.6.10 Input data to Topyfic . . . 83
3.6.11 Topic modeling using Latent Dirichlet Allocation (LDA) . . . 83
3.6.12 LDA Model Training . . . 84
3.6.13 TopModel Construction . . . 84
3.6.14 Topic object . . . 85
3.6.15 Analysis object . . . 86
3.6.16 Comparing Topics . . . 86
3.6.17 LDA parameter settings . . . 87
3.6.18 Pseudobulk calculation . . . 87
3.6.19 Principal component analysis (PCA) . . . 88
3.6.20 Topyfic analysis of single-nucleus RNA-seq data . . . 88
3.6.21 Topyfic analysis of single-cell RNA-seq data . . . 88

4 Unraveling gene expression dynamics in mouse models through PyWGCNA and Topyfic integration **96**
4.1 Abstract . . . 96
4.2 Introduction . . . 97
4.3 Results . . . 99

4.3.1 Analysis of Bulk RNA-seq 99
4.3.2 Comparison of bulk and single-nucleus RNA-seq datasets 100
4.4 Discussion 101
4.5 Methods 102

5 Future directions **117**

Bibliography **121**

LIST OF FIGURES

Page

2.1 **Overview of PyWGCNA workflow** . . . 39
2.2 **Depth-in PyWGCNA workflow** . . . 40
2.3 **Determining soft power threshold** . . . 41
2.4 **Running time of R WGCNA and PyWGCNA** . . . 42
2.5 **5xFAD modules the during progression of the AD** . . . 43
2.6 **3xTgAD module during the progression of the AD** . . . 44
2.7 **Comparison of 5xFAD modules and 3xTgAD modules** . . . 45
2.8 **Trem2 modules during the progression of the AD** . . . 46
2.9 **Trem2 neuronal module during the progression of the AD** . . . 47
2.10 **Trem2 myelination module during the progression of the AD** . . . 48
2.11 **ABCA7 modules during the progression of the AD** . . . 49
2.12 **Bin1 modules during the progression of the AD** . . . 50
2.13 **CLU modules during the progression of the AD** . . . 51
2.14 **Epha modules during the progression of the AD** . . . 52
2.15 **PicalmH465R modules during the progression of the AD** . . . 53
2.16 **Spi1 modules during the progression of the AD** . . . 54
2.17 **ABI3 modules during the progression of the AD** . . . 55
2.18 **Comparison of modules obtains from GWAS mouse models of AD.** 56

3.1 **Overview of Topyfic and datasets.** . . . 89
3.2 **Topic modeling in single nuclei from 5xFAD/BL6 and 5xFAD/CAST cortex and hippocampus.** . . . 90
3.3 **Topic modeling in scRNA-seq of microglia.** . . . 91
3.4 **Topic modeling using regulatory genes.** . . . 92
3.5 **Impact of various data and Topyfic parameters on the number of topics(K).** . . . 93
3.6 **Overview of MODEL-AD dataset.** . . . 94
3.7 **Overview of ENCODE dataset.** . . . 95

4.1 **Gene expression heatmap** . . . 105
4.2 **Matrix with the Module-Trait Relationships (MTRs) heatmap** . . . 106
4.3 **Darkred module eigengene expression** . . . 107
4.4 **Gene Ontology (GO) analysis of the Darkred module** . . . 108
4.5 **Comparison of PyWGCNA modules and gene markers of clusters** . 109
4.6 **Comparison of PyWGCNA modules and gene markers of cell types** 110

4.7 **Distribution of cosine similarities** 111
4.8 **Heatmap of cosine similarities between PywGCNA modules and Topyfic topics** 112
4.9 **Clustering dendrogram of samples based on TPM values.** 113
4.10 **Determination of soft-thresholding power** 114

LIST OF TABLES

Page

2.1 5xFAD mouse model and matching C57BL/6J mice samples. 57
2.2 3xTgAD mouse model and matching B6129SF1/J mice samples. 58

4.1 Discriotion of cuprizone cohort of bulk RNA-seq data 115
4.2 Discriotion of LPS bulk RNA-seq data 116

ACKNOWLEDGMENTS

I want to express my heartfelt appreciation to my family and friends for their constant love, encouragement, and understanding throughout this journey. My parents, Saeed and Farideh, and sister ,Niloofar, have been particularly supportive, providing unwavering belief in my abilities and celebrating my achievements with me. I am also grateful to my husband, Amirhosein, for his consistent support and encouragement. His belief in me has been a source of strength.

I extend my sincere thanks to Ali for their invaluable guidance and the opportunities they provided during my research. Their expertise, patience, and dedication significantly influenced my academic journey and the direction of this thesis.

I am grateful to my doctoral committee members, Kim Green and Vivek Swarup, for their insightful feedback and valuable suggestions, which enhanced the quality of this work.

Thanks also go to the MODEL-AD, ENCODE, and IGVF consortia for the collaboration opportunities and engaging academic discussions.

I appreciate my lab mates for their camaraderie, collaboration, and friendship. Special thanks to Dr. Gaby, Fairlie, Heidi, and our lab technicians for their support in our shared projects, creating a positive research environment.

To everyone who has been part of this journey, your support has been invaluable, and I am truly thankful for your presence in my life.

VITA

Narges Rezaie

EDUCATION

Doctor of Philosophy in Biological Sciences **2023**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Information technology in Computer engineering **2018**
Sharif University of Technology *Tehran, Iran*

PUBLICATIONS

* These authors contributed equally

Published

1. CA. Butler, AM Arvilla, G Milinkeviciute, Cd Cunha, S Kawauchi, **N Rezaie**, H Liang, D Javonillo1, A Thach, S Wang, S Collins, A Walker, KX Shi, J Neumann, A Gomez-Arboledas, LA Hohsfield, M Mapstone, AJ Tenner, FM LaFerla, A Mortazavi, GR MacGregor, KN Green. The $Abca7^{V1613M}$ variant reduces Aβ generation, plaque load, and neuronal damage. . *XXX* (2024)

2. LF Garcia-Agudo, Z Shi, I Smith, EA Kramár, K Tran, S Kawauchi, S Wang, S Collins, A Walker, K Shi, J Neumann, HY Liang, CD Cunha, G Milinkeviciute, S Morabito, E Miyoshi, **N Rezaie**, A Gomez-Arboledas, AM Arvilla, D ImanGhaemi, AJ Tenner, FM LaFerla, MA Wood, A Mortazavi, V Swarup, GR MacGregor, KN Green. $BIN1^{K358R}$ suppresses glial response to plaques in mouse model of Alzheimer's Disease. *XXX.* (2024)

3. KM Tran, S Kawauchi, E Kramár, **N Rezaie**, H Liang, J Sakr, A Gomez-Arboledas, MA Arreola, C Cunha, J Phan, S Wang, S Collins, A Walker, K Shi, J Neumann, G Filimban, Z Shi, G Milinkeviciute, DI Javonillo, K Tran, M Gantuz, S Forner, V Swarup, AJ Tenner, FM LaFerla, MA Wood, A Mortazavi, GR MacGregor, KN Green. A Trem2R47H mouse model without cryptic splicing drives age-and disease-dependent tissue damage and synaptic loss in response to plaques. *Molecular neurodegeneration.* (2023)

4. **N Rezaie**, F Reese, A Mortazavi. PyWGCNA: A Python package for weighted gene co-expression network analysis. *Bioinformatics.* (2023)

5. M Bayati*, **N Rezaie***, M Hamidi, MS Tahaei, H Rabiee. A New R Package for Categorizing Coding and Non-Coding Genes. *Preprints.* (2023)

6. H Alinejad-Rokny, R GhavamiModegh, H Rabiee, E Ramezani, **N Rezaie**, KT Tam, ARR Forrest. MaxHiC: A robust background correction model to identify biologically relevant chromatin interactions in Hi-C and capture Hi-C experiments. *PLOS Computational Biology.* (2022)

7. **N Rezaie***, M Bayati*, M Hamidi,MS Tahaei, S Khorasani, NH Lovell, J Breen, H Rabiee, H Alinejad-Rokny. Somatic point mutations are enriched in non-coding RNAs with possible regulatory function in breast cancer. *Communications Biology.* (2022)

8. DI Javonillo, KM Tran, J Phan, E Hingco, EA Kramár, Cd Cunha, S Forner, S Kawauchi, G Milinkeviciute, A Gomez-Arboledas, J Neumann, CE Banh, M Huynh, DP Matheos, **N Rezaie**, JA Alcantara, A Mortazavi, MA Wood, AJ Tenner, GR MacGregor, KN Green, FM LaFerla. Systematic phenotyping and characterization of the 3xTg-AD mouse model of Alzheimer's disease. *Frontiers in Neuroscience.* (2022)

9. S Forner, S Kawauchi, G Balderrama-Gutierrez, EA Kramár, DP Matheos, J Phan, DI Javonillo, KM Tran, E Hingco, Cd Cunha, **N Rezaie**, JA Alcantara, D Baglietto-Vargas, C Jansen, J Neumann, MA Wood, GR MacGregor, A Mortazavi, AJ Tenner, FM LaFerla, KN Green. Systematic phenotyping and characterization of the 5xFAD mouse model of Alzheimer's disease. *Scientific Data.* (2021)

In review/preparation

1. **N Rezaie**, E Rebboah, BA Williams, H Liang, F Reese, G Balderrama-Gutierrez, L Dionne, LG Reinholdt, D Trout, B Wold, A Mortazavi Identification of robust cellular programs using reproducible LDA that impact sex-specific disease progression in different genotypes of a mouse model of AD *bioRχiv*.(2024)

2. F Reese, BA Williams, G Balderrama-Gutierrez, D Wyman, MH Çelik, E Rebboah, **N Rezaie**, D Trout, M Razavi-Mohseni, Y Jiang, B Borsari, S Morabito, H Liang, C McGill, S Rahmanian, J Sakr, S Jiang, W Zeng, K Carvalho, A Weimer, LA Dionne, A McShane, K Bedi, S Elhajjajy, J Jou, I Youngworth, I Gabdank, P Sud, O Jolanki, JS Strattan, M Kagda, MP Snyder, BC Hitz, JE Moore, Z Weng, D Bennet, L Reinholdt, M Ljungman, MA Beer, MB Gerstein, L Pachter, R Guigó, BJ Wold, A Mortazavi. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *In revision.* (2023)

3. G Balderrama-Gutierrez*, H Liang*, **N Rezaie**, K Carvalho, S Forner, D Matheos, E Rebboah, KN Green, AJ Tenner, F LaFerla, A Mortazavi. Single-cell and nucleus RNA-seq in a mouse model of AD reveal activation of distinct glial subpopulations in the presence of plaques and tangles. *bioRχiv*.(2021)

SOFTWARE

PyWGCNA `https://github.com/mortazavilab/PyWGCNA`
Python library designed to do weighted correlation network analysis (WGCNA).

Topyfic `https://github.com/mortazavilab/Topyfic`
Python library designed to find reproducible topics using reproducible Latent Dirichlet Allocation (rLDA).

SELECTED PRESENTATIONS / POSTERS

IGVF Consortium Meeting **2023**
Poster

Probabilistic Modeling in Genomics conference **2023**
Poster

Network Biology conference **2023**
Poster

IGVF Consortium Meeting **2023**
Invited speaker, Poster

American Society for Human Genetics **2022**
Poster

Alzheimer's & Dementia **2022**
Poster

Alzheimer's & Dementia **2021**
Poster

Society for Neuroscience **2021**
Poster

Society for Neuroscience **2020**
Poster

ABSTRACT OF THE DISSERTATION

Computational approach to characterize gene dynamics using bulk and single-nucleus RNA sequencing to study Alzheimer's disease

By

Narges Rezaie

Doctor of Philosophy in Mathematical, Computational and Systems Biology (MCSB)

University of California, Irvine, 2024

Ali Mortazavi, Chair

The brain is a complex organ that controls thought, memory, emotion, touch, motor skills, vision, breathing, temperature, hunger, and many processes that regulate our body[1]. Alzheimer's disease (AD) is a neurodegenerative disease that is characterized by memory loss and impaired cognitive function[2]. It is associated with the accumulation of plaques and tangles in the brain[3]. The cortex and hippocampus are critical brain regions for learning because of their tasks of neural integration and memory respectively[4–6]. Therefore, these regions have been characterized exhaustively under different conditions and models to understand the cell subtypes involved[7–9]. Changes in gene expression and isoforms during development, aging, and disease are controlled by multiple, overlapping programs[10]. The gene expression profiles of distinct cell types arise reflect from complex genomic interactions among multiple simultaneous biological processes within each cell that can be altered by disease progression. Gene functionality is closely connected to its expression specificity across tissue and cell types. These functions can be inferred by the abundance and activity of co-expression networks using bulk RNA-seq[11]. Short-read single-cell RNA-seq is a widely-used method to characterize cellular heterogeneity in complex tissues based on gene expression[12]. A critical step in the analysis of large genome-wide gene expression datasets is the use of module detection methods to identify which genes vary in an informative manner and determine how these genes

organize into modules. Because of the limitations of classical clustering methods/detecting modules, numerous alternative module detection methods have been proposed, which improve upon clustering by handling co-expression in only a subset of samples, modeling the regulatory network, and/or allowing overlap between modules.

Here, I describe my work on characterizing the transcriptome of mouse cortex and hippocampus using bulk RNA-seq in conjunction with single-cell/nucleus RNA-seq to characterize changes during normal development and aging by comparing several mouse models of AD against control mice to study genes associated with neurodegeneration. First, I describe the PyWGCNA package to analyze gene expression and to infer meaningful modules of co-expressed genes that respond to different conditions such as age in different mouse models of AD using bulk RNA-seq. Then, I describe my novel reproducible grade of membership model called Topyfic, which is designed to derive topic models that correspond to cellular programs. I then apply Topyfic to distinct brain RNA-seq datasets from MODEL-AD and ENCODE and detect major changes in microglia, astrocytes, and oligodendrocytes that vary based on genotype and sex. Finally, I investigate possible ways to deconvolve modules into topics and make a connection between them. Together, these new computational methods provide novel insights into cellular programs in health and disease.

Chapter 1

Introduction

1.1 Abstract

Genetic variation, disease progression, and cell type/state-specific cellular programs impact gene functionality and expression. RNA sequencing (RNA-seq) characterizes the heterogeneity and complexity of RNA transcripts within individual samples ranging from cells to organized tissues, whole organs, and organisms. Changes in gene expression, alternative splicing, and chromatin profiles have been described as indicators of many pathologies. Alzheimer's disease (AD) is a pervasive neurodegenerative disorder that is characterized by distinctive plaques and tangles in affected brain areas. Mouse models of human Alzheimer's disease are a valuable approach for studying the underlying pathogenic mechanisms and assessing the efficacy of interventions as well as therapeutic strategies. Analysis of transcriptional changes was conducted using RNA-seq from the cortex and hippocampus of mouse models as a function of aging along with human clinical data to identify genes involved in disease progression. Genes that are expressed in the same subset of cells represent a module that may be co-regulated by shared cis-regulatory elements and a specific set of transcription

factors. Identifying such units is an important entry point to characterizing transcriptionally distinct subpopulations, including those associated with pathology or known regulators of myelination, inflammation, and neuronal survival. Here, I describe my work on characterizing the transcriptome of mouse cortex and hippocampus using bulk RNA-seq in conjunction with single-cell/nucleus RNA-seq to characterize changes during normal development and aging by comparing several mouse models of AD against control mice to study genes associated with neurodegeneration. First, I describe the PyWGCNA package to analyze gene expression and to infer meaningful modules of co-expressed genes that respond to different conditions such as age in different mouse models of AD using bulk RNA-seq. Then, I describe my novel reproducible grade of membership model called Topyfic, which is designed to derive topic models that correspond to cellular programs. I then apply Topyfic to distinct brain RNA-seq datasets from MODEL-AD and ENCODE and detect major changes in microglia, astrocytes, and oligodendrocytes that vary based on genotype and sex. Finally, I investigate possible ways to deconvolve modules into topics and make a connection between them. Together, these new computational methods provide novel insights into cellular programs in health and disease.

1.2 Introduction

1.2.1 Cell types and lineages in the brain

The brain is the most complex organ in the human body. It is the command center of the nervous system and plays a crucial role in controlling various physiological and cognitive functions, including perception, memory, emotion, and motor coordination. The brain includes at least 47 molecular distinct subclasses of cells[13], with several key cell types having distinct structures and functions. The two primary categories of cells are neurons and

glial cells[14]. Neurons are the main signaling units of the brain, communicating with each other via synapses to transmit information. Oligodendrocytes, microglia, and astrocytes are the main glial cell types. They provide structural support, insulation, and nourishment for neurons[15,16]. They also contribute to the maintenance of the brain's extracellular environment[17,18].

Neurons generate electrical impulses, known as action potentials, and release neurotransmitters to communicate with other neurons and cells. The two main subclasses of neurons are the GABAergic neurons and Glutamatergic Neurons[19]. GABAergic neuron (GABA) is a type of neuron that uses gamma-aminobutyric acid as its primary neurotransmitter[20]. GABA is an inhibitory neurotransmitter, meaning it tends to inhibit or reduce the activity of the neurons it acts upon. GABAergic neurons play a crucial role in regulating the balance between excitation and inhibition in the nervous system. Glutamatergic neuron (GLUT) is another type of neuron that predominantly uses glutamate as its primary neurotransmitter[21]. Glutamate is an excitatory neurotransmitter, which tends to enhance or promote the activity of the neurons it interacts with[22]. Glutamatergic neurons are involved in conveying excitatory signals in neural circuits and play a central role in various cognitive functions, including learning and memory. Neurons integrate the signals that they receive in their dendrites and will fire an action potential along their axons when the input signal crosses a neuron-specific threshold.

Microglia are the primary resident immune cells of the brain and spinal cord. They originate from myeloid precursor cells during embryonic development[23] and act as the primary immune defense cells in the central nervous system (CNS). Microglia become activated whenever there is an injury or infection in the CNS. Activated microglia migrate to the site of issue to engulf and to digest cellular debris, pathogens, and dead neurons, helping to clear away damaged or harmful substances[24]. Microglia can release signaling molecules that either promote or suppress inflammation, depending on the context, and are vital in maintaining a balance

between protective and potentially harmful immune responses in the CNS[25–29]. Microglia are involved in multiple neurodevelopmental processes, including the formation and refinement of neural circuits[30]. They contribute to synaptic pruning, eliminating excess or non-functional synapses during brain development. Dysregulation of microglial activity is implicated in various neurological disorders, including neurodegenerative diseases such as Alzheimer's, Parkinson's, and multiple sclerosis[31]. Overactivation of microglia may contribute to chronic inflammation, potentially exacerbating neurodegenerative processes[32].

Astrocytes are the most abundant glial cells and play a crucial role in supporting the structure and function of neurons[33]. They are essential for forming and maintaining the blood-brain barrier (BBB), providing structural support, supplying nutrients, maintaining homeostasis, promoting neuronal growth, and recycling neurotransmitters[34]. Astrocytes are integral to the overall health and function of the nervous system, and their diverse functions highlight the complexity of neural networks in the brain[35]. Dysregulation of astrocyte function is associated with various neurological disorders, including neurodegenerative diseases, epilepsy, and brain injuries. In some cases, reactive astrogliosis, which is an exaggerated response of astrocytes to injury or disease, can contribute to the progression of neurological conditions[36].

Oligodendrocytes play a crucial role in supporting and insulating neurons by producing myelin, which is a fatty substance that wraps around axons[37]. Oligodendrocytes play a role in maintaining ion homeostasis around axons by regulating the concentration of ions in the extracellular space[38]. Their role in myelination is critical for efficient communication between neurons and is essential for proper neurological function[39]. Disorders affecting oligodendrocytes and myelin, such as demyelinating diseases, can lead to impaired nerve conduction and neurological symptoms.

1.2.2 Alzheimer's disease

Alzheimer's disease is a progressive neurological disorder that primarily affects the brain, leading to cognitive decline and memory loss[40]. Over 90% of people who develop Alzheimer's dementia are age 65 or older[41]. Late-onset Alzheimer's Disease (LOAD) is the most common cause of dementia among older adults with about 1 in 9 people (10.8%) age 65 and older having Alzheimer's dementia[42]. In 2023, an estimated 7.2 million Americans aged 65 and older were living with Alzheimer's dementia and an estimated 13.8 million people will have dementia by 2060 in the USA[43]. The greatest risk factors for LOAD are older age and a family history of dementia. The latter implies the involvement of genetics, and the $\epsilon 4$ form of the apolipoprotein E (APOE) gene is known to substantially increase the risk of LOAD[44]. Over 25 genes have been implicated in LOAD on the basis of GWAS studies, such as Trem2, Clu, Abca7, Epha1, and Spi1[45]. Together, they are thought to account for 15-25% of the risk of developing LOAD[46,47]. There are also additional environmental risk factors that can be changed or modified to reduce the risk of cognitive decline and dementia such as physical activity, smoking, education, staying socially and mentally active, blood pressure, and diet[42]. There is currently no cure for Alzheimer's disease. However, there is medicine available that can temporarily reduce the symptoms and slow disease progression[48].

AD is characterized by memory loss and impaired cognitive function associated with the accumulation of plaques and tangles in the brain[49] that disrupt communication between neurons and contribute to their degeneration. This neuronal loss, along with the structural changes in the hippocampus and cortex ultimately leads to the characteristic cognitive and functional deficits observed in AD. The hippocampus is a vital region for memory and learning processes that plays a crucial role in the formation of new memories and the conversion of short-term memories into long-term memories. In the early stages of AD, the hippocampus is often one of the first areas to be affected. The accumulation of beta-Amyloid (β-Amyloid) plaques and tau tangles disrupts the normal functioning of nerve cells in the hippocampus.

As the disease progresses, the hippocampus atrophies, which contributes to memory loss and difficulties in forming and retrieving memories[50,51]. Individuals with AD often experience challenges in remembering recent events and struggle with spatial navigation. The cortex is the outer layer of the brain responsible for higher cognitive functions, including thinking, reasoning, language, and sensory perception. AD progression affects various areas of the cortex, leading to widespread cortical atrophy that contributes to cognitive decline[52]. As a result, individuals with AD may experience difficulties with language, problem-solving, decision-making, and overall cognitive function.

The specific trigger for AD is not yet fully understood however multiple theories such as abnormal protein accumulation[53], genetic, cholinergic hypothesis[54], oxidative stress[55], and hormonal variations[56] have been proposed to explain the underlying causes. Among them, abnormal protein accumulation is a key feature of several neurodegenerative diseases, including AD[57]. In AD, two primary types of abnormal protein deposits, β-Amyloid plaques, and tau tangles play a central role. β-Amyloid is a protein fragment derived from the larger amyloid precursor protein (APP). In the normal course of cellular processing, APP is cleaved into smaller fragments, including β-Amyloid. However, in AD, there is an abnormal accumulation of β-Amyloid, leading to the formation of plaques[58]. Several genes are associated with an increased risk of developing Alzheimer's disease, and some of them are directly related to the processing of β-Amyloid. Mutations in presenilin 1 (*PSEN1*) and presenilin 2 (*PSEN2*) genes are associated with early-onset familial AD (FAD)[59]. These mutations can influence the processing of APP, leading to an increase in β-Amyloid production and, indirectly, affecting tau pathology. The major risk factor for LOAD is specific alleles of the $APOE$ gene. The $APOE$ gene comes in different forms or alleles: $APOE\ \epsilon 2$, $APOE\ \epsilon 3$, and $APOE\ \epsilon 4$. The $APOE\ \epsilon 4$ allele is a well-established genetic risk factor for AD. Individuals carrying one copy of $APOE\ \epsilon 4$ have a 2-3 fold increased risk, and those with two copies have 10-15 fold higher risk of developing AD during their lifetime[60]. The $APOE\ \epsilon 4$ allele is associated with higher β-Amyloid accumulation and an increased likelihood of

developing Alzheimer's[61]. However, having the *APOE* $\epsilon 4$ allele does not guarantee that an individual will develop AD[62]. Tau is a protein encoded by the *MAPT* gene that normally stabilizes microtubules in neurons. In AD, hyperphosphorylation of tau isoforms causes it to form twisted tangles inside neurons, known as neurofibrillary tangles[63,64]. Mutations in specific genes can influence tau metabolism and contribute to the development of neurofibrillary tangles[65]. Mutations in *MAPT* can lead to abnormal tau protein and are associated with certain forms of frontotemporal dementia, which shares some pathological features with AD[66,67]. The intricate relationship between these abnormal protein accumulations, genetic factors, and the overall pathology of Alzheimer's disease is an active area of research. Other genetic and environmental factors may also contribute to the complex interplay that leads to the development and progression of AD. Identifying these factors is crucial for understanding the underlying mechanisms of the disease and developing potential therapeutic interventions.

1.2.3 The role of glial cells during neurodegeneration

In AD, glial cells undergo dynamic changes throughout the progression of the disease and play both protective and detrimental roles[68,69]. Microglia are in a surveillance state during the early stages of AD, constantly monitoring the brain for signs of damage or pathology[70]. In the intermediate stages, as β-Amyloid plaques accumulate, microglia become activated[70]. This activation involves changes in morphology, increased expression of immune-related genes, and an attempt to clear the plaques through phagocytosis[71]. Activated microglia release pro-inflammatory cytokines, contributing to neuroinflammation[72]. However, the inflammatory response may not be sufficient to completely clear the accumulating β-Amyloid, resulting in chronic inflammation and prolonged microglial activity[73,74]. Persistent activation of microglia and sustained neuroinflammation characterize advanced stages of AD[32,75]. In addition to β-Amyloid, the presence of tau tangles becomes more prominent in advanced stages and microglia start to respond to tau pathology[76,77]. In late stages, chronic activated

microglia may become dysfunctional, contributing to a toxic environment that negatively affects neuronal health[25]. The chronic inflammatory state, along with the accumulation of tau tangles, can lead to widespread neuronal damage and cognitive decline[78].

Microglia can adopt a specialized activation state known as Disease-Associated Microglia (DAM), which is associated with a specific transcriptional profile[8]. DAM is involved in the response to neurodegenerative pathology such as AD, attempting to clear abnormal protein aggregates. DAMs are associated with the expression LOAD50 GWAS candidate genes l such as apolipoprotein E (*APOE*), *TREM2*, *CLU* and *TYROBP* upregulated in DAM, whereas *CD33*, *BIN1*, *PICALM* and *PLCG2* are downregulated[8].

Astrocytes undergo reactive gliosis as AD progresses, exhibiting changes in morphology and gene expression in response to neuroinflammation and β-Amyloid accumulation[79,80]. In response to the presence of β-Amyloid plaques and other pathological changes, astrocytes become activated and release inflammatory molecules, such as cytokines and chemokines (*C5*, *CCL5*, *CCL3*, and *IL1B*)[81], as part of the brain's immune response[82]. Dysfunction in the ability of astrocytes to clear β-Amyloid may contribute to the accumulation of plaques in AD. In addition to β-Amyloid, astrocytes may contribute to the clearance of tau aggregates[83]. The activation of astrocytes, as indicated by increased *GFAP* expression, is part of the brain's attempt to respond to the presence of β-Amyloid and associated neurodegenerative changes[84]. In summary, as the disease progresses, astrocytes may become less effective in maintaining a homeostatic environment.

Several studies suggest that myelin abnormalities may be present in the brains of individuals with AD[85**?** ,86]. Changes in myelin integrity and composition, possibly related to alterations in oligodendrocyte function, have been observed in post-mortem brain tissue of individuals with Alzheimer's[86]. White matter, which consists of axons and myelin, can show structural changes in individuals with AD. Disruptions in white matter integrity, including changes in myelin, may contribute to cognitive decline in AD[87]. Other factors such as inflammation

and oxidative stress associated with AD pathology could also potentially impact the health and function of oligodendrocytes[88].

The collective impact of glial responses in neurodegenerative diseases is complex. While glial cells attempt to mitigate damage and clear pathological substances, chronic or dysregulated glial activation can contribute to a neurotoxic environment and exacerbate neuronal damage. Understanding the role of glial cells at different stages of AD is crucial for developing targeted therapeutic approaches aimed at modulating these responses to slow or to halt disease progression.

1.2.4 Mouse models of AD

The study of animal models in disease research is integral to our understanding of disease mechanisms, developing effective therapies, and exploring the influence of genetic and environmental factors[89,90]. Researchers can model various diseases, providing insights into pathophysiological processes by employing animals that share physiological and genetic similarities with humans. Mouse models are widely utilized due to their practical advantages and genetic similarities with humans for AD research[91,92]. Mice share conserved biological pathways and exhibit pathological changes akin to those seen in AD, such as the accumulation of β-Amyloid plaques and neurofibrillary tangles[93]. The relatively short lifespan of mice allows researchers to observe disease progression and test interventions within a manageable timeframe. Genetic engineering techniques such as transgenic and knockout technologies enable the manipulation of specific genes associated with AD, facilitating the study of genetic factors influencing the disease. Moreover, mice are more cost-effective and reproduce quickly compared to larger mammals. They also have a well-characterized genome which make them valuable for high-throughput experiments. The use of mouse models in Alzheimer's research provides a controlled and efficient platform for investigating disease mechanisms, testing

potential therapeutics, and advancing our understanding of the intricate interplay between genetics and environmental factors in neurodegenerative diseases.

The development and characterization of new mouse models of LOAD are critical for understanding the progression of the pathology and as a platform for the evaluation of new drugs[94]. The goal is to recapitulate three essential features: (1) plaques, (2) tangles, and (3) degeneration in the cortex as well as the hippocampus. There are at least 200 mouse models of AD (alzforum.org) that have been developed to mimic different aspects of AD[95]. Among them, the 5xFAD mouse model[96] is the commonly used transgenic mouse model developed in 2012[96]. This model contains five different familial AD mutations including three mutations in *APP* (APP KM670/671NL (Swedish), APP I716V (Florida), APP V717I (London)) and two mutations in *PSEN1* (PSEN1 M146L, PSEN1 L286V), under the control of a *THY1* mini-gene[97,98], which directs expression to forebrain neurons. These mutations are known to promote the overproduction and accumulation of β-Amyloid, so 5xFAD mice will develop robust amyloid pathologies, with plaques appearing in the brain from 2–4 months of age[96], triggering robust microgliosis and inflammatory processes as well as synaptic and neuronal loss[96,99]. Even though tau tangles are absent, the 5xFAD model is designed to accelerate the development of β-Amyloid plaques, allowing researchers to study the early and rapid onset of Alzheimer's pathology.

The 3xTgAD mouse model[100] is another transgenic mouse model frequently used in AD research developed in 2003[101]. It is unique in that it combines both β-Amyloid plaques and neurofibrillary tangles, two of the major pathological features of AD but it does not display neuronal loss[101]. It contains three key genetic mutations (APP Swedish, PSEN1 M146V, and tau P301L) associated with familial AD. APP Swedish mutation is associated with the Swedish variant of *APP*, leading to increased production of β-Amyloid plaques[102]. PSEN1 M146V mutation involves a familial Alzheimer's disease-associated mutation in the *PSEN1* gene, contributing to β-Amyloid accumulation[103]. MAPT P301L mutation involves a

mutated form of the human tau protein, leading to the formation of neurofibrillary tangles, another hallmark of AD[104]. This model has been widely used to study the interactions between β-Amyloid and tau pathology, as well as their combined effects on cognitive function and neurodegeneration[105–107].

Both of these transgenic mouse models along with over 200 other mouse models have been generated to study this progressive neurodegenerative disorder, and in many instances, these mice have yielded insights into the underlying pathogenic mechanisms[108,109]. However, many therapeutic approaches that have been demonstrated to be successful in these familial AD models have failed when evaluated in human clinical trials involving participants with late-onset AD (LOAD)[110–112]. The causative mutations in *APP*, *PSEN1*, or *PSEN2* for AD have been identified only in early-onset or familial cases, which account for $< 1\%$of all AD cases. In contrast, LOAD accounts for $> 95\%$ of all AD cases and there is a pressing need to develop new animal models that better recapitulate the underlying molecular pathways leading to LOAD[111–113]. The NIA-funded UCI MODEL-AD project is developing better mouse models to analyze the causes of LOAD. Numerous susceptibility genes have recently been identified by GWAS and genomic sequencing[114,115,115,116], albeit at much smaller hazard ratios compared to that of APOE[117]. The UCI MODEL-AD project uses CRISPR and genome replacement to model and validate eight GWAS-identified LOAD risk loci (*Abca7*, *Abi3*, *Bin1*, *Clu*, *Epha1*, *Picalm*, *Spi1*, and *Trem2*[118]) and characterized mice with each of these both on a wild-type (C57BL/6J) and 5xFAD background to determine their effects on plaque generation and damage exerted on the brain in response to pathology.

1.2.5 Infer modules from bulk RNA-seq

Bulk RNA-seq is a powerful tool for studying global gene expression patterns, identifying novel transcripts, and understanding the molecular mechanisms underlying various biologi-

cal processes, including development, disease, and responses to external stimuli[119,120]. Bulk RNA-seq allows researchers to identify differentially expressed genes associated with conditions such as Alzheimer's disease, Parkinson's disease, and autism spectrum disorders[120]. Moreover, bulk RNA-seq enables the exploration of gene expression changes during crucial stages, shedding light on the molecular mechanisms underlying neural differentiation, synaptogenesis, and maturation in neurodevelopment.

One method to infer gene function and gene–disease associations from bulk RNA-seq is weighted gene co-expression network analysis (WGCNA)[121], an approach that identifies modules of genes with similar expression patterns. It constructs networks of genes with a tendency to co-activate across a group of samples and subsequently interrogates and analyzes this network. The gene modules can be used to associate genes of unknown function with biological processes, to prioritize candidate disease genes, or to discern transcriptional regulatory programs. The network and modules can be interrogated to identify regulators, co-regulated pathways, biological processes, and hub genes.

Computational approaches to infer cell identity from single-cell and single-nucleus RNA-seq studies

Tissues consist of diverse cell types, each with highly specialized functions in multicellular organisms. Single-cell (sc) and single-nucleus (sn) RNA-seq is a relatively young technology that enables researchers to profile gene expression in individual cells/nuclei. This technological advancement facilitates the study of the heterogeneous cellular composition of complex tissues in various contexts such as development, aging, health, and disease[13,122].

In scRNA-seq workflows[123,124], whole cells are isolated, allowing for the capture and analysis of RNA from the entire cell, including both cytoplasmic and nuclear RNA. The quality of scRNA-seq data relies on the successful dissociation of tissues into viable single cells, presenting a significant challenge in highly interconnected tissues such as the brain[125]. On the

other hand, snRNA-seq[126] exclusively requires the removal of nuclei, focusing on capturing and analyzing RNA within the nucleus, while excluding cytoplasmic RNA. snRNA-seq has proven effective in analyzing diverse hard-to-dissociate tissues and cell types, including the brain[125], heart[127], adipocytes[128], and myofibers[129].

In most tissues, snRNA-seq excels in recovering attached cell types, whereas scRNA-seq exhibits a bias toward immune cell types[130]. For instance, snRNA-seq analyses of human brain samples have failed to capture a microglial activation signature in AD[131]. The integration of scRNA-seq and snRNA-seq data will enable more comprehensive transcriptome profiling and enhanced cell-type annotation in tissues.

In the analysis of sc/snRNA-seq data, cell clustering and cell type annotation are both critical steps. While these processes can be manually executed with adequate expertise, they are labor-intensive and time-consuming. Identifying genes with significant expression variability across individual cells is a common preprocessing step in the analysis of sc/snRNA-seq data. Focusing on highly variable genes (HVGs) allows researchers to prioritize genes that carry essential biological information, contributing to the observed diversity and heterogeneity in the dataset. The selection of HVGs is recommended for enhancing downstream analyses, including dimensionality reduction techniques and cell clustering algorithms. Validation and sensitivity analysis of different HVG selection methods[132–134] which may yield varying sets of genes, become essential to ensure the robustness of downstream results.

In cell clustering, cells are grouped based on their gene expression patterns, facilitating downstream tasks such as cell function recognition and cell-type annotation. Numerous cell clustering methods have been developed, with many stemming from generic clustering algorithms. For instance, pcaReduce[135] employs an iterative strategy relying on principal component analysis (PCA) and hierarchical clustering. SC3[136] was developed using k-means and PCA methods, while RaceID[137] enhances k-means by incorporating outlier detection to identify rare cell types. Beyond generic clustering algorithms, community detection-based

methods have been developed and widely employed. PhenoGraph[138], for example, utilizes shared nearest-neighbor graphs and Louvain[139] community detection to reduce clustering time costs for large-scale datasets[140]. Leiden algorithm is an improvement of the Louvain algorithm, which guarantees that communities are well connected[139].

Seurat[141] and Scanpy[142] are two popular scRNA-seq analysis packages that integrate HVGs, PCA, Louvain, Leiden, and various other methods. However, high dimensionality, inherent noise, and the rapid growth of scRNA-seq data pose challenges in cell clustering. While state-of-the-art methods have addressed issues related to high dimensionality and large cell numbers, mitigating noise associated with data sparsity remains a significant challenge.

The biological interpretation of cell clustering results, i.e. cell annotation, is crucial in scRNA-seq data analysis. Many scRNA-seq data analysis methods, due to the inclusion of PCA-based dimensionality reduction, often overlook biological significance during the clustering process. In cell annotation, genes play a pivotal role in annotating and interpreting cell clusters. Automatic cell annotation methods fall into two categories. The first category uses supervised methods such as SingleCellNet[143], SingleR[144], scmap[145], and Azimuth[146]. These methods require a labeled reference dataset, and the gene expression patterns of cell clusters are compared to this reference dataset. Clusters exhibiting similar expression patterns to a particular cell group in the reference dataset are assigned their label[144]. However, accurate annotated reference data is often not available. The second category prioritizes genes for cells or cell clusters. The gene markers are considered informative for revealing cellular diversity and indicating cell functions[147]. Methods such as SCINA[148] and CellAssign[149] assign cell types based on known marker genes but may be prone to biases associated with the markers used. Additionally, there are deep learning-based tools, such as scDHA[150] and scBalance[151], which also require reference data. In some cases, annotation tools come with a cell marker database such as scCATCH[152] built a reference database "CellMatch" combining multiple databases.

The accuracy of annotation strongly relies on the informativeness and comprehensiveness of the marker gene database. However, existing databases may not provide extensive coverage across tissue types and cell types with good specificity. These challenges can be particularly daunting for investigators new to the scRNA-seq field or those with limited background knowledge of the involved tissue and cell types.

In Chapter 3, I use a Grade of Membership model (GoM)[153] known as latent Dirichlet allocation (LDA) to annotate cells. These models generalize cluster models, allowing each sample to have membership in multiple clusters. They are employed to uncover latent and complex gene expression patterns, revealing biologically meaningful topics. LDA is a probabilistic topic model using unsupervised learning initially proposed for text mining.[154] It assumes that the observed set of words in a document is influenced by latent attributes (topics) within the document[155]. As a nonlinear method, LDA excels in handling complex, sparse, and noisy datasets[156]. Furthermore, LDA is considered interpretable as its parameters directly associate input features with latent factors or target outcomes. In bioinformatics, LDA has found application in single-cell analysis[153,157,158], novel cancer mutation signature discovery102, microbiome composition analysis[159,160], substructure exploration in metabolomics[161], and pathway–drug relationships[162].

1.3 Conclusions (Theme of thesis)

In summary, genetics and the control of gene regulation are likely heavily intertwined in AD. The MODEL-AD project is developing LOAD mouse models based on human GWAS data with each model evaluated by performing a comparative network analysis of hippocampal transcriptomes to relate changes in expression to other phenotypic changes through the combined use of bulk and single-cell/nucleus RNA-seq. By learning weighted gene co-expression networks impacted by AD-risk-enhancing genes, we will obtain insights that facilitate the

continued improvement of mouse models and will be able to suggest the genetic perturbations that would be useful for matching a particular AD subtype. As different subsets of GWAS-associated genes are known to be expressed in neurons, microglia, and astrocytes, we must then characterize gene expression at a regional, as well at the single-cell level, to identify the key factors that account for desired expression levels at the right time and place. My projects aim to characterize the transcriptome of mouse cortex and hippocampus using bulk, sn, and sc RNA-seq to describe the cell-specific changes during normal development and AD using several mouse models of AD produced by MODEL-AD.

In Chapter 2, I introduce PyWGCNA, which is a Python package that I developed for weighted correlation network analysis (WGCNA)[121]. This library offers a faster implementation compared to the R version of WGCNA and includes several additional downstream analysis modules. PyWGCNA facilitates the comparison of multiple co-expression modules to each other and external gene lists, such as marker genes from single-cell analyses. I showcase the capabilities of PyWGCNA on brain bulk RNA-seq datasets from the 5xFAD and 3xTgAD mouse models provided by the MODEL-AD consortium. We applied PyWGCNA to identify modules associated with genotypes in these datasets and compared resulting modules to find shared co-expression signatures with significant overlap across the datasets. The PyWGCNA manuscript was published in *Bioinformatics* in 2023.

In Chapter 3, I describe my work on identifying robust cellular programs using Topyfic, which is a Python package that I developed for applying reproducible Latent Dirichlet Allocation (rLDA) to single-cell/bulk RNA-seq data. This approach aims to recover meaningful topics involving key genes, such as transcription factors, associated with different cellular processes. Topyfic is applied to brain single-cell and single-nucleus datasets from 5xFAD mice crossed with either C57BL6/J (BL6) or CAST/EiJ (Cast) mice. Our goal is to identify shared and distinct cell types and states, particularly focusing on microglia. The results reveal that 8-month 5xFAD/Cast F1 males exhibit a higher level of microglial activation compared to

matching 5xFAD/BL6 F1 males, while female mice show similar levels of microglial activation. Notably, using regulatory genes such as transcription factors, microRNA host genes, and chromatin regulatory genes is sufficient to effectively capture cell types and states. This study emphasizes how Grade of Membership models with a specific vocabulary of regulatory genes can successfully identify gene expression programs in single-cell data. This approach proves valuable for quantifying both similar and divergent cell states in distinct genotypes. This work was posted as a BioRχiv preprint in Feb 2024 and is currently in revision.

In Chapter 4, I investigate the relationship between gene modules resulting from PyWGCNA and gene topics using Topyfic. I analyze bulk RNA-seq datasets developed by the MODEL-AD consortium with PyWGCNA and compare them to the gene topics identified in Chapter 3. Here, a gene module is defined as a gene module membership vector (kME) representing genes exhibiting a correlation between the expression profile of the gene and the module eigengene, while a gene topic consists of a set of genes with weights indicating their contribution to the topic. Analyzing these modules and topics simplifies the system's complexity, accelerates novel discoveries, and reveals new functionalities. To explore the relationship between gene modules and topics, I calculate the cosine similarity, quantifying the degree of similarity or dissimilarity between vectors based on the cosine of the angle between them. Visualizing these vectors in geometric space aids in discerning patterns, relationships, and potential trends that might not be apparent through numerical analysis alone. This approach facilitates the identification of upstream regulators, signaling hubs, and potential resistance pathways in diseases like Alzheimer's (AD) and cancer.

In Chapter 5, I conclude by discussing other sequencing technology assays such as long-read RNA-seq to improve the generation and evaluation of novel mouse models of AD. Additionally, I discuss other sequencing technology assays such as perturb-seq to improve our knowledge in understanding biological aspects underlying development, evolution, and disease.

Chapter 2

PyWGCNA: a Python package for weighted gene co-expression network analysis

Note: Parts of this chapter was published in *Bioinformatics* in 2023. It has been revised with additional text and data to reflect updates made to the software since then.

2.1 Abstract

Weighted Gene Co-expression Network Analysis (WGCNA) is frequently used to identify modules of genes that are co-expressed across multiple RNA-seq samples. However, the current R implementation is slow, is not designed to compare modules across different WGCNA networks, and produces results that can be challenging to interpret and visualize. We introduce the PyWGCNA Python package, which is designed to identify co-expression modules from large RNA-seq datasets. PyWGCNA has a faster implementation than the R version

of WGCNA and includes several additional downstream analysis features that can be used within/across modules. These features include functional enrichment analysis using Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), REACTOME, and inter-module analysis of protein–protein interactions, as well as comparison of multiple co-expression modules to each other and external gene lists, such as marker genes from single cells. We apply PyWGCNA to two distinct bulk RNA-seq datasets of brain tissue collected by MODEL-AD to identify modules associated with genotypes. Comparing the resulting modules enables us to discover shared co-expression signatures in the form of modules with significant overlap across the datasets.

The PyWGCNA library for Python 3 is available on PyPi and on GitHub.

2.2 Introduction

Weighted Gene Co-expression Network Analysis (WGCNA) is a widely used method for characterizing gene correlation patterns across extensive sample sets[163]. It identifies modules of highly correlated genes, summarizes these modules, correlates them with external traits, and calculates module membership. Correlation networks facilitate network-based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, such as cancer, mouse genetics, and analysis of human data. The WGCNA package[163] is implemented in the popular R language. As sequencing datasets grow larger and more complex, having a scalable implementation of WGCNA becomes increasingly important.

We introduce PyWGCNA, designed to do WGCNA and downstream analytical tasks natively in Python (Fig. 2.1.A). PyWGCNA supports co-expression network analysis for large, high-dimensional gene or transcript expression datasets, addressing the time or memory in-

efficiency that is often encountered in R. Additionally, this package can directly perform functional enrichment analysis, including GO[164], KEGG[165], and REACTOME[166], on co-expression modules to characterize the functional activity of each module. PyWGCNA also supports the addition or removal of data, allowing for iterative improvement in network construction as new samples become available or need to be excluded. Finally, PyWGCNA can compare co-expression modules from multiple PyWGCNA networks with each other to assess module reproducibility or with marker genes from scRNA-seq clusters to evaluate the functional activity or cell-type specificity of each module (Fig. 2.1.B). We demonstrate PyWGCNA's utility by identifying co-expression modules associated with genotype in bulk RNA-seq from MODEL-AD using 5xFAD and 3xTgAD mouse models of Alzheimer's disease (AD) and matching wild-type (WT) mice.

2.3 Materials and methods

2.3.1 Identifying co-expression modules

Input and initialization of the PyWGCNA object

The PyWGCNA object stores user-specified network parameters, such as the network type, and major outputs, such as the adjacency matrix. PyWGCNA can be initialized from expression data, gene metadata, and sample metadata, which can be passed to PyWGCNA all together in an AnnData[167] format or separately as a series of CSV or TSV matrices (Fig. 2.2.1). The expression data should be formatted such that the rows correspond to samples and the columns correspond to genes.

Data preprocessing

One crucial step in identifying gene co-expression modules involves calculating the correlation between genes. Excessive missing values can significantly impact the results. PyWGCNA addresses this issue by offering the capability to eliminate excessively sparse genes/transcripts or samples, as well as lowly expressed genes/transcripts. Moreover, PyWGCNA can identify and remove genes or samples with too many missing values, and it can detect outlier samples through hierarchical clustering and user-defined thresholds. While PyWGCNA can handle the removal of outlier genes or samples, we recommend preliminary preprocessing and normalization of input gene or transcript expression data, including any necessary batch correction.

Finding co-expression modules

PyWGCNA follows an identical approach to the reference WGCNA R package, differing only in default parameter choices, such as the type of network. Initially, PyWGCNA constructs a co-expression matrix by calculating the correlation between each pair of genes/transcripts from the preprocessed expression data and stores the results in an adjacency matrix. Subsequently, it builds a co-expression network through soft power thresholding of the correlation matrix, followed by the computation of the topological overlap matrix to generate the final network. Finally, PyWGCNA identifies co-expressed modules of genes/transcripts by hierarchically clustering the network and applying a dynamic tree cut.

Downstream analysis and visualization of co-expression modules

PyWGCNA offers several options for downstream analysis and visualization of co-expression modules. It can calculate module-trait correlation, compute and summarize module eigen-

gene expression across sample metadata categories, detect hub genes in each module, and perform functional enrichment analysis using databases like GO[164], KEGG[165], and REACTOME[166] via GSEApy[168] and BioMart[169]. Additionally, PyWGCNA can recover known and predicted protein–protein interactions within each module using the STRING database[170]. Each analysis option comes with user-friendly plotting tools to visualize the results, including interactive module network visualization with options to select genes for display in each module.

2.3.2 Assessing co-expression module overlap between PyWGCNA objects or to single-cell data

PyWGCNA can compare co-expression modules from multiple PyWGCNA objects by computing the Jaccard similarity coefficient and the proportion of common genes for each pair of modules between the objects. The statistical significance of the overlap is evaluated using Fisher's exact test. Employing the same approach, PyWGCNA can identify overlaps between co-expression modules and various gene lists, such as marker genes from single-cell RNA-seq. This analysis unveils the cell-type specificity of each co-expression module. In both cases, the results from these tests can be easily visualized using PyWGCNA.

2.4 Results

To assess the performance of PyWGCNA compared to the R reference of WGCNA, we utilized expression data from both gene-level (bulk short-read RNA-seq) and transcript-level (bulk long-read RNA-seq) datasets, each consisting of 100 samples from the ENCODE portal. We generated 15 subset datasets with a reduced number of genes or transcripts to analyze how runtime changes as the number of features increases. For each subset, we ran

PyWGCNA and R WGCNA three times on the same hardware configuration (32 cores, 300 GB memory). Both packages exhibited similar performance up to 16,000 genes, but PyWGCNA demonstrated twice the speed on larger datasets. Notably, PyWGCNA successfully identified modules for 96,000 transcripts, whereas we encountered memory constraints preventing the computation of a co-expression network with the same dataset using R WGCNA (Fig. 2.4).

2.4.1 Analysis of bulk RNA-seq of 5xFAD and 3xTgAD mouse model

We applied PyWGCNA to analyze 192 bulk RNA-seq samples from the cortex and hippocampus of the 5xFAD mouse model, along with matching C57BL/6J mice at four different ages (4, 8, 12, and 18 months) in both sexes (Table. 2.1)[120]. PyWGCNA successfully identified 17 gene co-expression modules associated with age, genotype, tissue, and sex (Fig. 2.5.A). The 5xFAD coral module is strongly correlated with age progression in the 5xFAD genotype ($P - value < 0.05$), as illustrated by the module eigengene expression (Fig. 2.5.B). Conversely, the 5xFAD white module showed a significant correlation with the hippocampus in both genotypes (Fig. 2.5.C). The 5xFAD coral module, comprising 1335 genes, demonstrated significant enrichment for GO terms related to immune response and neutrophil activation (Fig. 2.5.D). Notably, this module included well-known microglial activation genes such as *Cst7*, *Tyrobp*, and *Trem2*. In contrast, the 5xFAD white module (435 genes) exhibited enrichment in GO terms related to cilium movement, organization, and assembly (Fig. 2.5.E).

We further applied PyWGCNA to analyze 38 bulk RNA-seq hippocampal female samples from the 3xTgAD mouse model, along with matching WT (B6129SF1/J) mice at three different ages (4, 12, and 18 months) (Table. 2.2)[100]. This analysis yielded 17 modules correlated with age or genotype (Fig. 2.6.A). The 3xTgAD dark gray module, consisting of

380 genes, shows a strong correlation with both the 3xTgAD genotype in mice and the 18 month time point (Fig. 2.6.B). GO analysis reveals significant enrichment for genes related to neutrophil degranulation and immune response, featuring key genes such as *Csf1*, *Tyrobp*, and *Trem2* (Fig. 2.6.C).

To assess the similarity between modules found in the 5xFAD and 3xTgAD experiments, we performed module overlap tests using PyWGCNA. Our analysis revealed several modules with significant overlap, enriched for similar functions (Fig. 2.7). As expected, based on their functional enrichment, the 5xFAD coral module and the 3xTgAD dark gray module significantly overlap one another, suggesting that the co-expression network within these modules is conserved across the two familial AD mouse models (Fig. 2.7.A). Additionally, the 3xTgAD white module (434 genes), which strongly correlated with 18-month samples and enriched in cilium movement, significantly overlapped with the 5xFAD white module (Fig. 2.7.B).

2.4.2 Analysis of bulk RNA-seq of GWAS mouse models

The MODEL-AD consortium is aiming to develop mouse models that better replicate human AD. Investigating the impact of disease-associated single nucleotide polymorphisms (SNPs) on brain function and aging is crucial for building testable models to understand the mechanisms underlying AD. The UCI MODEL-AD teams build several bulk RNA-seq datasets of mouse models of AD based on eight GWAS-identified LOAD risk loci (*Abca7*, *Abi3*, *Bin1*, *Clu*, *Epha1*, *Picalm*, *Spi1*, and *Trem2*) on a wild-type (C57BL/6J) and 5xFAD background to explore changes in gene expression patterns and enhance our understanding of aging and AD.

Trem2 R47H mouse model

The TREM2 R47H variant is one of the strongest genetic risk factors for LOAD. Unraveling coding sequence changes in *TREM2* is crucial for gaining insights into both *TREM2* function and the role of microglia, which predominantly express *TREM2* in the brain[171,172]. Initial studies involving *Trem2* knockout (KO) mice revealed the essential role of *TREM2* in the microglial response to plaques and their transition to a DAM phenotype, characterized by the specific expression of genes such as *Cst7*, *Clec7a*, *Itgax*, and *Apoe*[91]. Furthermore, the absence of *TREM2*, leading to a lack of microglial reaction to plaques, paradoxically seemed to exacerbate disease progression[173–175].

To shed light on the relevance of AD, the MODEL-AD project conducted a study by crossing Trem2R47H NSS mice with the 5xFAD mouse model, which is characterized by amyloidosis. The assessment included an evaluation of gene expression at 4 and 12 months of age. Bulk RNA-seq libraries were constructed from 3 to 5 mice per genotype, sex, and tissue (hippocampus) across both time points[118]. This comprehensive approach aimed to provide a deeper understanding of the impact of the Trem2R47H variant in the context of gene expression patterns.

To investigate gene expression changes, PyWGCNA was applied to analyze a quantile-normalized TPM matrix comprising 93 bulk RNA-seq samples. This analysis identified 55 modules associated with age, genotype, and sex. Notably, the PyWGCNA recovered the inflammatory darkgrey module (Fig. 2.8.A). While there are no significant changes in eigengene values at 4 months in both 5xFAD and 5xFAD/Trem2R47H mice, an increases in the inflammation module's eigengene values are observed by 12 months, with the presence of Trem2*R47H including an even higher increase 2.8.B). The Darkgrey inflammatory module significantly enriched in GO terms related to the immune response and cytokine signaling modules 2.8.B).

The eigengene values of the Lightgrey module reveal a significant decrease in this module in 12-month-old 5xFAD/Trem2R47H compared to 5xFAD mice (Fig. 2.9.A). The Light-

grey module contains 513 genes enriched in go terms related to neuronal systems such as neurotransmitter secretion (GO:0007269) (Fig. 2.9.B). The Snow module with 183 genes reveals a clear effect of Trem2R47H in 12-month-old mice, regardless of 5xFAD genotype (Fig. 2.10.A). It is also significantly enriched in myelination and differentiation related go terms (Fig. 2.10.B). These findings suggest that the R47H variant confers age- and disease-specific effects on glial cells. Importantly, similar results have been observed in human AD tissue from TREM2 variant carriers, supporting and validating the relevance of these findings[176].

ABCA7 V1613M mouse model

GWAS of AD and control individuals have identified variants associated with the *ABCA7* locus, indicating an increased risk of developing LOAD[114,177]. Exome sequencing of *ABCA7* has further revealed nonsense and missense coding variants that are predicted to affect protein structure and function. Many of these variants are thought to result in loss-of-function mutations and have been found to be enriched in individuals with AD[116,178–180]. Expression studies have supported these findings, showing that AD brains with low levels of *ABCA7* tend to develop AD at a younger age compared to those with higher *ABCA7* expression. On the other hand, individuals expressing *ABCA7* at similar levels to healthy controls tend to develop AD at a very late age[181,182] suggesting that reduced function of *ABCA7* may play a crucial role as a risk factor in the development of AD.

The V1599M coding variant of *ABCA7* (rs117187003) was identified by exome sequencing and predicted to be deleterious, suggesting it may confer an increased risk of developing LOAD[116,179]. The MODEL-AD consortium evaluated the V1599M variant (V1613M in mice) for potential inclusion in new LOAD models and also crossed it with 5xFAD mice to give relevance to AD. This decision was based on its location within a relatively well-conserved region of homology between the mouse and human *ABCA7*.

To identify changes in gene expression associated with homozygosity for the Abca7V1613M

variant in 5xFAD mice, bulk tissue RNA-seq was conducted on hippocampi resulting 4 different mouse models (5xFAD/Abca7V1613M, Abca7V1613M, 5xFAD, and WT control mice). The analysis utilized PyWGCNA on 73 bulk RNA-seq samples at two different timepoints (4 months and 12 months old) in the hippocampus for both sexes, using genes with more than 1 TPM. After removing one sample based on hierarchical clustering at the sample level, a soft threshold of 19 was selected.

Among the 6 modules identified by PyWGCNA, the black module contains inflammation/cytokine related genes that are highly upregulated in 5xFAD mice but significantly reduced in 5xFAD/Abca7V1613M mice (Fig. 2.11.A). Eigengene values of these inflammation modules for all groups revealed a reduction in 5xFAD/Abca7V1613M mice compared to 5xFAD at both 4 and 12 months of age (Fig. 2.11.B-C). Collectively, these data indicate that 5xFAD/Abca7V1613M mice exhibit reduced microgliosis and inflammation, which is likely due to lower Aβ plaque burden compared to 5xFAD controls rather than distinct changes to microglial gene expression and function.

BIN1 K358R mouse model

GWAS has identified bridging integrator 1 (*BIN1*) as a key risk locus for LOAD[115,183,184]. The *BIN1* locus is associated with the second-highest risk for developing LOAD after *APOE*[115,183,184]. Subsequent targeted exome sequencing of LOAD patients and controls revealed that the rare *BIN1* coding variant rs138047593 (K358R) is associated with an increased risk of LOAD[116,185]. Research on the role of *BIN1* in AD has yielded conflicting results regarding its impact. In AD, BIN1 levels have been reported as both increased and decreased[186,187]. *BIN1* has been suggested to regulate the processing of Aβ, a key protein associated with AD pathology. However, reducing *BIN1* in mice did not significantly alter Aβ production or deposition[188]. Despite this, *BIN1* has been observed to accumulate around amyloid plaques in both humans and mice[189]. Notably, *BIN1* directly binds to tau and phosphorylation of tau in and around the proline-rich domain weakens its interaction

with the SH3 domain of *BIN1* [190,191].

Bin1K358R is a CRISPR/Cas9-generated mutant of Bin1 gene carrying the K426R mutation that corresponds to the human K358R SNP and has been generated by UCI MODEL-AD teams. To investigate the interaction between the BIN1 K358R variant and AD pathology, UCI MODEL-AD teams crossed BIN1 K358R mice with the 5xFAD mouse model, which is characterized by aggressive amyloidosis. The impacts of this crossbreeding were assessed on normal brain function and its interaction with amyloidosis. Therefore, 4 different mouse genotypes were generated: 5xFAD/BIN1K358R (hemizygous 5xFAD/homozygous BIN1K358R), BIN1K358R (homozygous BIN1K358R), 5xFAD (hemizygous 5xFAD), and wild-type (WT) control mice. These mice were aged until 4 and 12 months, representing stages of initial plaque production and advanced plaque load, respectively.

75 bulk RNA-seq datasets were built using hippocampus tissue from 4-5 mice per genotype and sex across both time points. PyWGCNA was employed to identify 38 co-expressed gene modules (Fig. 2.12.A). The module-trait relationship heatmap showed a high correlation with 12-month 5xFAD mice in the lightcoral module (Fig. 2.12.A). However, this correlation was significantly reduced in 12-month 5xFAD/BIN1K358R mice compared to 5xFAD mice (Fig. 2.12.A). Eigengene values of the lightcoral module for all groups confirmed a reduction in 5xFAD/BIN1K358R mice compared to 5xFAD at 12 months of age, indicating that the K358R BIN1 variant induces gene expression changes (Fig. 2.12.B). The lightcoral module contains 1,249 genes while most of them are associated with inflammation/cytokine-related terms based on GO analysis (Fig. 2.12.C). Given that BIN1 is predominantly expressed in oligodendrocytes[192], the results suggest a dampened activation of both microglia[193] and oligodendrocytes at the gene level.

Clu-h2kbKI mouse model

The clusterin (CLU) gene, also known as apolipoprotein J ($ApoJ$)[194], is a significant ge-

netic factor associated with an increased risk of LOAD in multiple GWAS[183,184,195]. *CLU* expression is known to be upregulated in degenerative conditions, including AD[196,197], due to cellular and oxidative stress or dysregulation of specific signaling pathways[198–201]. However, conflicting results exist in the literature regarding whether *CLU* expression improves or exacerbates cellular stress[202–207].

In AD, *CLU* levels are increased in the brain[208]. It has been found to bind to Aβ and play a role in Aβ deposition and clearance[209–211]. *CLU* is present in Aβ plaques, vessels of cerebral amyloid angiopathy (CAA), and associated with neurofibrillary tangles[212,213]. It also interacts with modified tau species in human AD brain tissue[214]. Importantly, different single nucleotide polymorphisms (SNPs) in the *CLU* gene may exert their effects in combination with other genetic risk factors such as *APOE4*[215], *TREM2*[216], and *BIN1*[214]. Additionally, different mRNA isoforms are produced from the *CLU* gene, and variations in the *CLU* gene may lead to alterations in the ratios of isoforms, influencing the outcome of the disease and playing a role in the development and progression of AD[217,218].

The UCI MODEL-AD team created Clu-h2kbKI mice contains a 2kb region of human DNA sequence that spans from intron 7 to exon 9, including a human LOAD *CLU* risk SNP rs2279590. To investigate the function of *CLU* in AD, we also crossed it with the 5xFAD mouse model. We then collected 12 month cortex along with the 4 and 12 month hippocampus of these mice and performed bulk RNA-seq in order to explore the changes at the gene expression level. PyWGCNA was applied to 114 bulk RNA-seq samples of cortex and hippocampus together while 4 month cortex samples were not collected. Among the 27 modules (Fig. 2.13.A), the darkgrey module contains 607 genes primarily expressed in cortex tissue (Fig. 2.13.B) while enriched in GO terms related to neuronal systems such as neurotransmitter secretion and transport (Fig. 2.13.C).

Epha1 P461L mouse model

Epha1 is another gene that was linked to LOAD through GWAS. *EphA1* is a receptor tyrosine kinase and a member of the Eph receptor family[219]. Common variants of *EphA1* have been linked to LOAD in various studies[114,177,220]. Particularly, in GWAS of Caribbean Hispanic families, a P460L coding mutation in *EphA1* loci showed a significant association with LOAD[177]. *EphA1*, along with other Eph receptors, plays a role in various cellular processes, including cell adhesion, migration, and axon guidance during developmental morphogenesis, organogenesis, pattern formation, and cell fate determination[221,222]. Eph receptors interact with ephrin ligands, participating in bidirectional signaling[219,223]. Disruption of Eph/ephrin signaling has been associated with oncogenesis[224] and immune dysregulation[225], suggesting a potential role in regulating the neuroinflammatory process and influencing the progression of AD[226,227]. Understanding the involvement of *EphA1* and related pathways in AD could provide insights into the underlying mechanisms of the disease and potential targets for therapeutic interventions.

A rare coding variant of *EphA1*, rs202178565, was identified through targeted sequencing. To investigate the impact of this variant, the UCI MODEL-AD teams made Epha1P461L mouse model which was created using CRISPR/Cas9 technology to introduce a missense mutation (P461L in mice) corresponding to the human SNP rs202178565 found in the *EPHA1* gene.

To assess the effects of the *Epha1* variant in the context of AD, the researchers crossed this Epha1P461L mouse model with the 5xFAD mouse model. Hippocampal samples were collected from four different mouse genotypes (5xFAD/Epha1P461L, Epha1P461L, 5xFAD, and wild-type) at 12 months for PyWGCNA analysis. 39 modules were identified by applying PyWGCNA to 38 bulk RNA-seq samples (Fig. 2.14.A). Eigengene values of genes in burlywood modules are highly upregulated in 5xFAD mice but significantly reduced in 5xFAD/EPHA1 mice (Fig. 2.14.B). Burlywood module contains 1,644 genes including microglia marker genes such as *Itgax*, *Trem2*, which significantly enriched in inflammation/cytokine responses (Fig. 2.14.C).

PicalmH465R mouse model

The *PICALM* (Phosphatidylinositol-binding clathrin assembly protein) gene has been identified as a genetic risk factor for LOAD through GWAS[228]. *PICALM* is involved in the internalization and trafficking of APP which is a necessary protein that forms plaques in the brains of individuals with AD[229,230]. Additionally, *PICALM* plays a role in the regulation of β-Amyloid blood-brain barrier transcytosis and transferrin receptor endocytosis in erythroblasts[231,232]. Sequencing of the *PICALM* gene revealed 16 variants[231], and one of them, the H465R variant, is a nonsynonymous missense change associated with an increased risk of AD.

To investigate the role of the PICALM H465R variant in AD, the MODEL-AD UCI teams created the PicalmH465R allele. CRISPR/Cas9 endonuclease-mediated genome editing was employed to introduce the H465R mutation. Subsequently, these PicalmH465R mice were crossed with the 5xFAD mouse model of AD, resulting in four experimental groups: WT, PICALMH458R, 5xFAD, and 5xFAD/PICALMH458R. Bulk RNA-seq analyses were conducted on hippocampal samples collected from these mice at both 4, 12, and 18 months to explore changes in gene expression, with a specific focus on *PICALM*.

17 co-expressed gene modules were identified by applying PyWGCNA to 108 bulk RNA-seq samples. Gainsboro module (587 genes) is significantly correlated with older mice, more specifically 5xFAD mice (Fig. 2.15.A). Eigengene values of genes in gainsboro modules are increasing during time in 5xFAD mice this value is decreasing in 5xFAD/PICALMH458R mice during aging (Fig. 2.15.B). There is no significant changes in eigengene values in WT and PICALMH458R mice (Fig. 2.15.B). GO terms related to microglia cell activation and cytokines are enriched in this module (Fig. 2.15.C). In summary, the results suggest a dampened activation of glial cells in mice with H465R mutation at the gene level.

Spi1 rs1377416 mouse model

SPI1, also known as PU.1 (transcription factor PU.1), is a gene that encodes a transcription factor involved in the regulation of hematopoiesis and immune system function, including microglial proliferation and activation associated with AD and other neurodegenerative conditions involving an inflammatory response[233,234]. The study confirmed that AD-associated SNP rs1377416 increases in vitro enhancer activity in murine BV-2 microglia cells[235]. To create the Spi1*rs1377416 allele, CRISPR/Cas9 endonuclease-mediated genome editing was used to introduce a C to T missense mutation in the non-coding region of *Spi1*. To investigate the effect of AD, MODEL-AD UCI teams crossed this mouse model with 5xFAD and then aged them until 4 and 12 months.

Bulk RNA-seq was performed to explore the changes in SPI1 activity and its impact on the expression of genes associated with AD pathology. PyWGCNA was applied to 77 bulk RNA-seq samples which identified 39 modules (Fig. 2.16.A). White module has a significant correlation with 12 month 5xFAD and 5xFAD/Spi1 mice. The eigengene profile of this module shows no significant changes between 5xFAD and 5xFAD/Spi1 mice (Fig. 2.16.B). GO terms related to inflammation and cytokines enriched in this module, as well (Fig. 2.16.C). Collectively, there is no significant difference 5xFAD and 5xFAD/Spi1 mice.

ABI3 S209F mouse model

ABI3 (Abelson-interactor family member 3) is a gene that has been identified as a risk gene for AD through GWAS and genetic analyses[236,237]. ABI3 is primarily expressed by microglia in the human brain. Overexpression of ABI3 has been implicated in microglial activation, leading to inflammation, which in turn may contribute to the formation of neurofibrillary tangles and neuronal death, ultimately correlating with dementia[238–241]. Among the single-nucleotide polymorphisms (SNPs) associated with AD risk, two SNPs within the ABI3 gene have been implicated[242–244]: a rare missense SNP (rs616338) within ABI3 exon 5 and a common SNP (rs28394864) located 150,000 base pairs downstream of ABI3. To investigate the function of the ABI3 gene, the Abi3S209F allele was created. CRISPR/Cas9 endonuclease-

mediated genome editing was employed to introduce an S212F mutation into the Abi3 gene. This mutant allele models an SNP (rs616338) found in human ABI3 that encodes a missense mutation associated with an increased risk of sporadic AD.

The Abi3S209F mouse model was then crossed with the 5xFAD mouse model to explore the functional implications of the ABI3 gene in the context of AD. Bulk RNA-seq samples were collected from 4 and 12 months, representing different stages of disease progression, from four different mouse models: WT, Abi3S209F, 5xFAD, and 5xFAD/Abi3S209F. Then, PyWGCNA was applied resulting 12 co-expression gene modules (Fig. 2.17.A). Calculating the correlation between each module and different traits such as age, sex, and genotype reveals the white module that contains 2,393 genes and has a significant correlation with older 5xFAD and 5xFAD/Abi3S209F mice (Fig. 2.17.A). Eigengene value of the white module in 5xFAD mice is increasing over time, especially in female mice supporting the idea that biological sex is a risk factor for developing dementia including AD (Fig. 2.17.B). GO terms related to cytokines enriched in this module suggest the role of cytokines in the inflammatory response in the body, and inflammation is increasingly recognized as a factor in the development and progression of AD (Fig. 2.17.C).

Comparison of co-expressed modules from GWAS mouse models

In assessing the concordance among modules identified in the GWAS mouse models, we employed PyWGCNA to calculate the Jaccard index for each pair of modules derived from the analysis of individual mouse models. Our analysis recovered multiple modules exhibiting substantial overlap, indicative of shared functions. We identified 4 groups of modules with Jaccard index > 0.23 and annotated them based on the biological functions they were enriched with. Notably, modules enriched for biological functions such as inflammation, cytokine signaling, and immune response, previously annotated in individual analysis, exhibited significant (P-value < 0.01) similarity to one another (denoted as M4 of Fig. 2.18). Furthermore, the eigengene profiles of these modules demonstrated a consistent pattern, with

values showing a significant increase in mice with a 5xFAD background over time, while no significant changes were observed in mice with a WT background. This convergence of module similarity and eigengene profile dynamics suggests a shared regulatory landscape underlying the genetic architecture in these mouse models, mainly in the progression of AD.

The M1 meta-module, comprising one module from each GWAS analysis, consists of genes related to tauopathy, such as Mapt consistently present across all these modules suggesting the existence of an underlying pathway associated with excitability in neuronal networks (denoted as M1 of Fig. 2.18)[245]. Furthermore, our investigation has successfully delineated distinct groups of modules, denoted as M2, associated with myelination and derived from the TREM2, ABCA7, and EPHA1 datasets. This finding implies the involvement of complex processes related to myelin in glial development, particularly emphasizing the intricate interaction between glial cells and neurons (Fig. 2.18). Additionally, modules enriched in cilium development exhibit significant overlap (M3 in Fig. 2.18), reinforcing the confirmed role of cilia in mediating Sonic Hedgehog signaling (Shh) and thereby regulating hippocampal neurogenesis[246]. These collective findings shed additional light on the molecular intricacies associated with tauopathy, myelination, and cilium development, providing valuable insights into potential therapeutic targets for neurodegenerative disorders.

2.5 Discussion

We have developed a Python package based on the original R implementation of WGCNA. PyWGCNA is capable of handling larger datasets and provides an expanded set of well-documented functions, including additional downstream analyses and visualization tools such as functional enrichment and protein–protein interactions. Notable features include multi-way comparisons between multiple PyWGCNA networks and/or other gene lists. For example, the PyWGCNA Jaccard similarity-based gene list overlap test allows for associating

specific cell types to individual modules for further interpretation of the possible functions of these modules. As the number of datasets continues to grow exponentially, we anticipate that such comparative analyses will become increasingly valuable. We hope that PyWGCNA will contribute to filling a gap in the Python bioinformatics community.

Several GWAS have found over 25 genes associated with LOAD in humans. As part of the first phase of MODEL-AD, we introduced nine GWAS-identified LOAD risk loci (Abca7, Abi3, Bin1, Clu, Epha1, Picalm, Spi1, and Trem2) into C57BL/6J mice using CRISPR/Cas9 followed by breeding with the 5xFAD mouse model, which provides a valuable platform for studying the impact of these variants on normal brain function and their interactions with aggressive amyloidosis.

The mRNA expression profiles of these mouse models as well as matching controls at different ages can then be used to identify modules of genes whose expression is impacted by ages and the specific GWAS variants in a model of amyloidosis using PyWGCNA. PyWGCNA results would help us to improve our understanding of how each variant has age- and disease-dependent effects on glial cells and neuropathology on AD progression.

Using PyWGCNA, we were able to identify gene co-expressed modules that were associated with age and/or genotypes. Variants such as Trem2R47H appear to confer age and disease-specific effects on microglia that also have been shown in human AD individuals that carry *TREM2* variant, in which microglial responses to pathology are suppressed in newly formed pathological areas but exacerbated in more advanced pathological brain areas[118]. On the other hand, some variants such as BIN1K358R and Abca7V1613M show a reduction of eigenegene values in the inflammatory module, which raises the interesting possibility that these variants might be protective instead of increasing the risk of LOAD. There is a last group of variants such as Spi1*rs1377416 where we did not detect major changes at gene expression level that have been associated with AD pathologies stating the limitations of bulk RNA-seq.

The complexity of AD pathogenesis, involving various cell types and their interactions, necessitates advanced techniques such as single-cell for a more precise characterization. Generally, it is difficult to evaluate the complete role of glial cells in AD using only bulk RNA-seq but this will help us to prioritize the list of variants and genes that need to be explored deeply.

2.6 Supplementary methods

2.6.1 co-expression network construction

Determining the soft power threshold

To construct the network, correlation values are corrected based on a soft power threshold. This threshold is crucial as it enhances the distinction between strong and weak correlations, effectively bringing weak values closer to zero. This threshold also determines the sensitivity and specificity of pairwise correlation strengths. The chosen power value must result in a network that exhibits characteristics similar to a scale-free network which is based on the idea that the likelihood of a node (gene) being connected to k other nodes (genes) diminishes following a power law ($p(k) \sim k^{-\gamma}$).

To determine this parameter, we adopted the same strategy as the original WGCNA. We try to maximize a model fit ($R^2 > 0.9$) under a scale-free topology model while minimizing the number of connections lost ($mean(connectivity) <= 100$) when fitting the model (maintaining a high mean number of connections) (Fig.2.3).

Adjacency: Pairwise gene co-expression correlation

To calculate the adjacency matrix, it is essential to determine the type of network, specifically how pairs of nodes with strong negative correlations should be treated. One option is to consider them connected, as would be the case if the correlation was positive, and creating an 'unsigned' network where the sign doesn't matter ($cor = |cor(a, b)|$). Conversely, strongly negatively correlated nodes can be considered unconnected in a 'signed' network, where the sign of a strong correlation dictates the connection status ($cor = \frac{(cor(a,b)+1)}{2}$). Another option is a 'signed hybrid' network, combining hard and soft thresholding. It involves a hard threshold at 0 ($cor = 0$ for $|cor(a, b)| \leq 0$) and soft thresholding above zero ($cor = |cor(a, b)|$ for $|cor(a, b)| > 0$).

Once a network type is chosen, the correlations are calculated, raised to the normalization power factor, and then stored in an adjacency matrix.

Topological overlap matrix (TOM)

To construct modules, meaningful relationships are extracted from the adjacency matrix. This process begins by transforming the adjacency matrix into measures of gene dissimilarity (distance of a gene from every other gene in the system) known as Topological Overlap Matrix (TOM). Subsequently, the corresponding dissimilarity is calculated.

Identify modules

Using the Dissimilarity Matrix, genes with more similar expression patterns are grouped into clusters using hierarchical/agglomerative clustering, and a dendrogram (cluster tree) of genes is constructed to identify modules. Since some modules may exhibit high similarity, they can be merged into a single cluster based on a user-defined cut-off.

2.7 Materials and data availability

RNA sequencing was performed as described[120]. Sequences were aligned to the mouse genome (mm10) and annotation was done using GENCODE v21. Reads were mapped with STAR v.2.7.3a and RSEM (v.1.3.3) was used for quantification of gene expression. Raw fastqs for each model are available in Synapse.

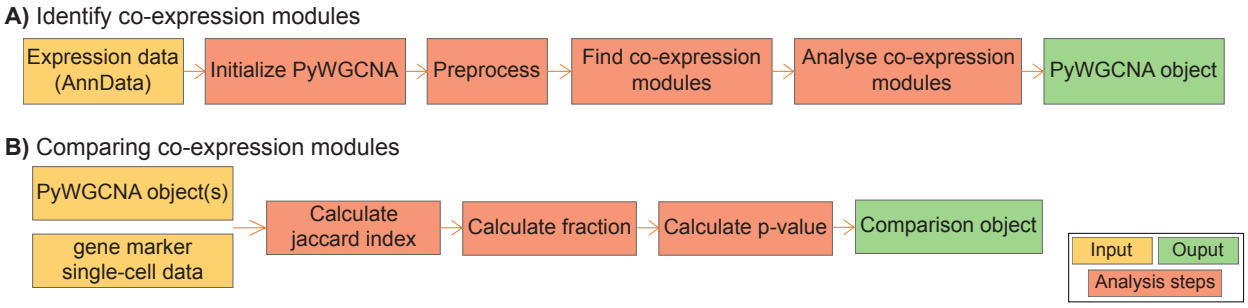

Figure 2.1: **Overview of PyWGCNA workflow. A)** Workflow of identifying co-expression modules. **B)** Workflow of comparing multiple PyWGCNA object(s) with/out gene marker list.

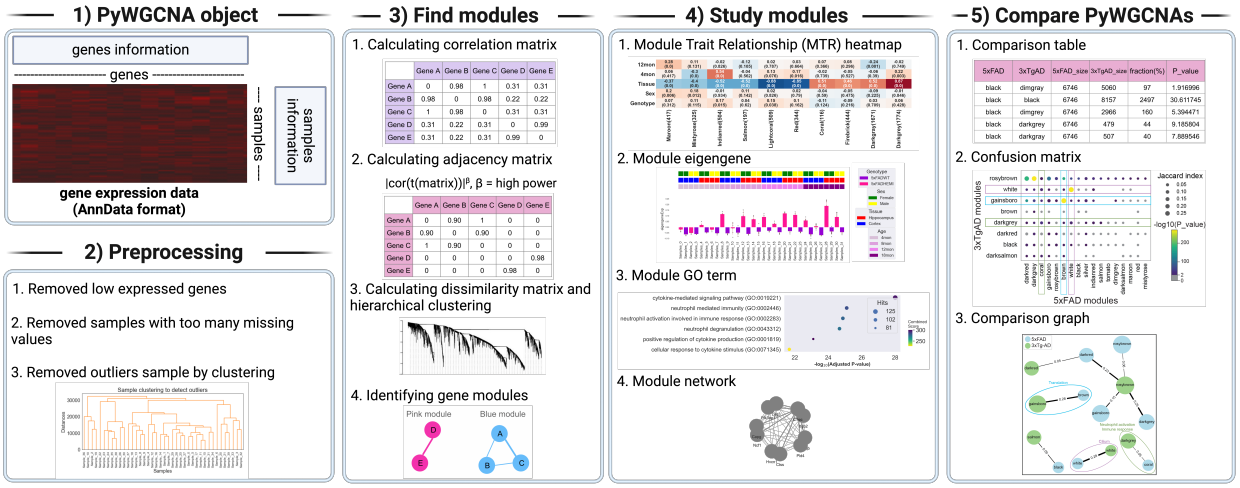

Figure 2.2: **Depth-in PyWGCNA workflow. 1)** Input of PyWGCNA (gene expression data in AnnData format) **2)** Preprocessing steps **3)** Finding co-expression modules **4)** Downstream analysis and visualization of co-expression modules **5)** Assessing co-expression module overlap between PyWGCNA objects or to single-cell data

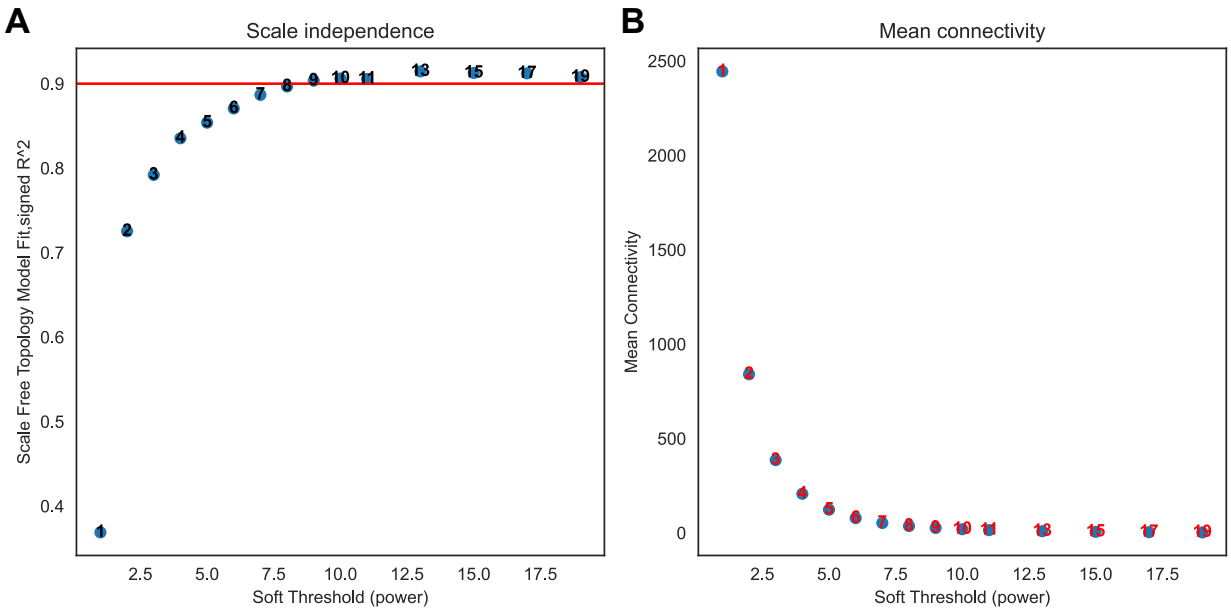

Figure 2.3: **Determining soft power threshold. A)** Scale independence value (R^2) at different soft thresholds (powers) **B)** Mean connectivity at different soft thresholds (powers)

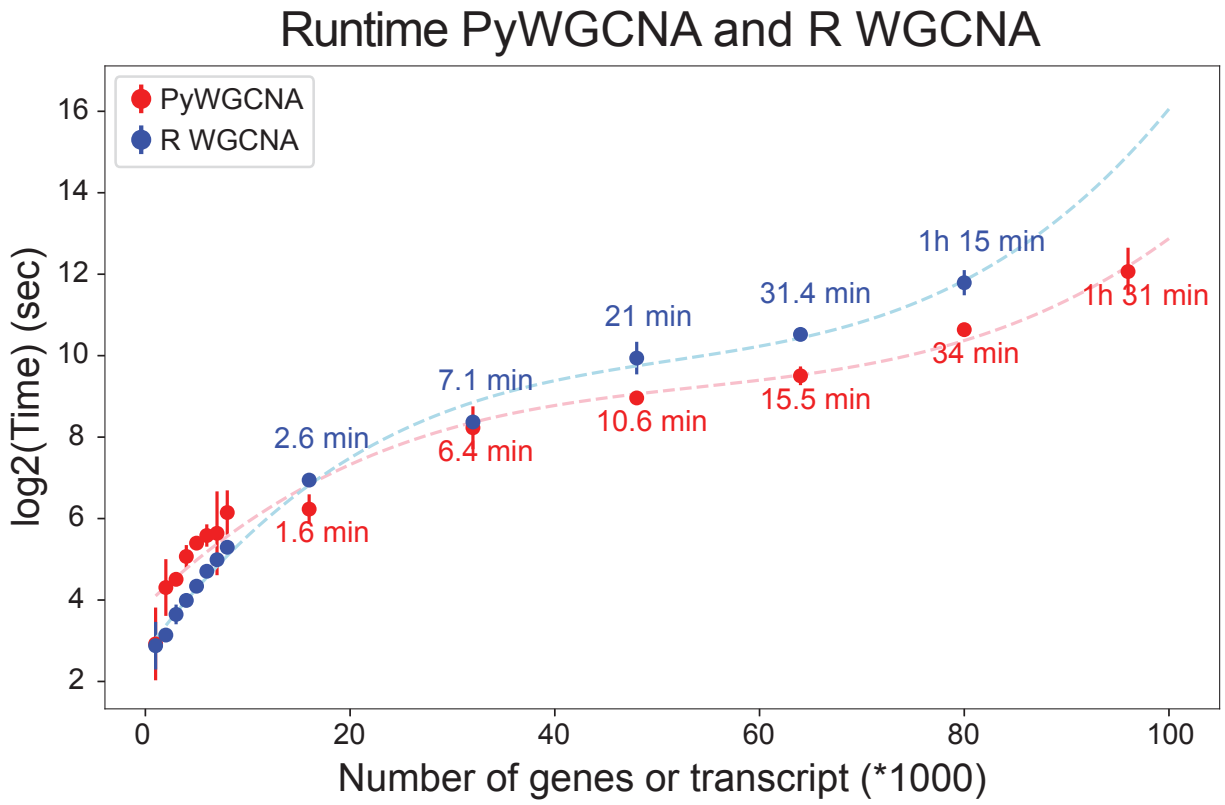

Figure 2.4: **Running time of R WGCNA and PyWGCNA.** Average runtime of PyWGCNA and R version of WGCNA versus the number of features used in downsampled datasets in triplicate

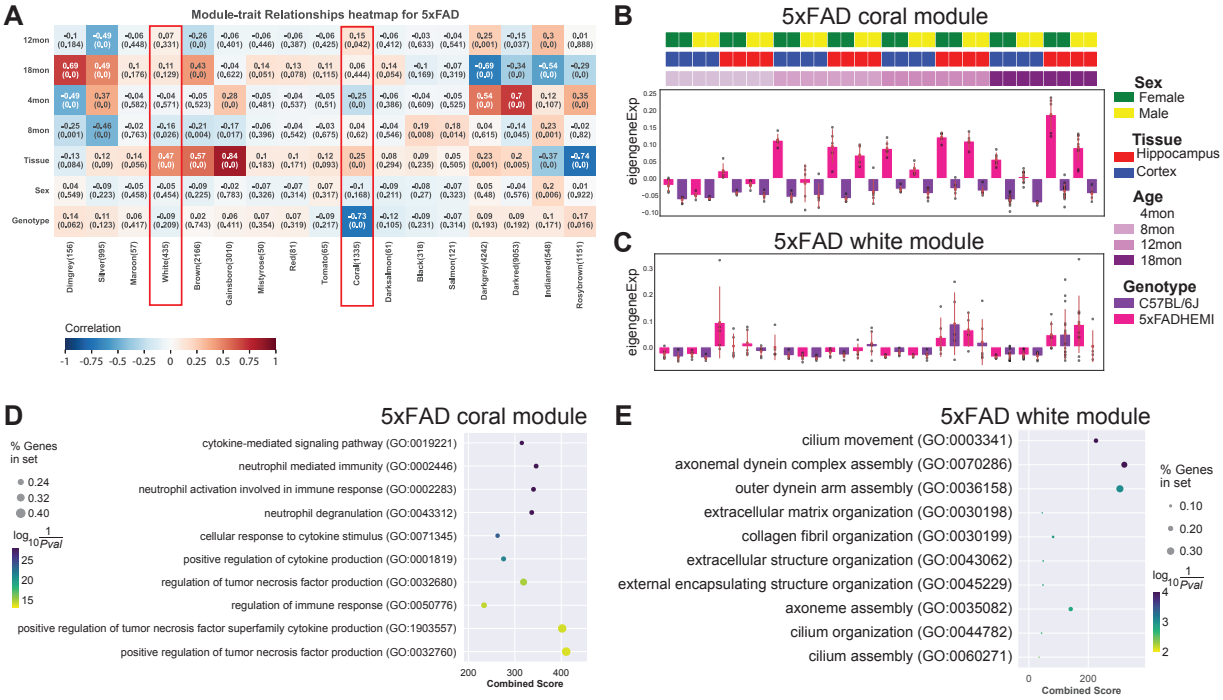

Figure 2.5: **5xFAD modules during the progression of the AD. A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** Coral and **C)** white module eigengene expression profile from 5xFAD mouse model summarized by genotype. Above, the top three rows display the metadata for each dataset including sex, tissue, and age. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. GO analysis of the genes in 5xFAD **D)** coral and **E)** white modules, respectively.

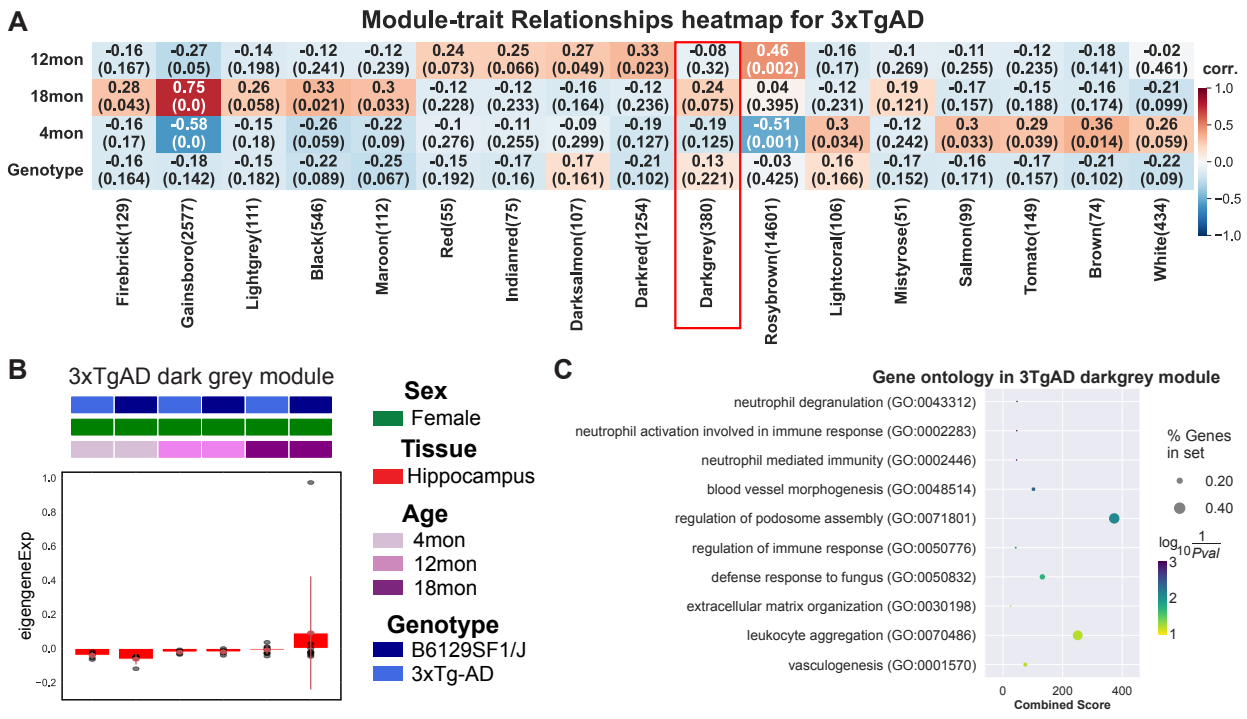

Figure 2.6: **3xTgAD module during the progression of the AD. A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** dark grey module eigengene expression profile from 3xTgAD mouse model summarized by genotype. Above, the top three rows display the metadata for each dataset including sex, tissue, and age. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in 3xTgAD dark grey module.

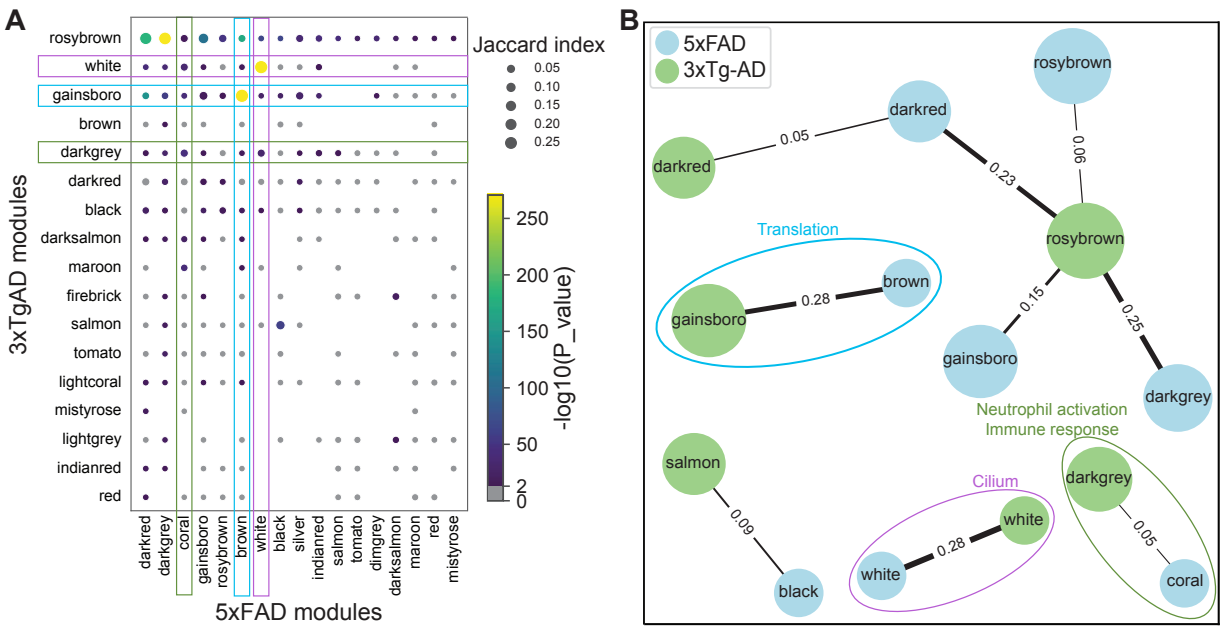

Figure 2.7: **Comparison of 5xFAD modules and 3xTgAD modules. A)** Bubble plot of module overlap test results between 5xFAD and 3xTgAD mouse models of familial AD. The dot size represents the fraction of shared genes between each pair of modules and non-gray color denotes the significance of the overlap between modules. **B)** Comparison graph of 5xFAD and 3xTgAD modules mouse models of familial AD for those with >0.05 Jaccard similarity. The thickness of the lines shows the Jaccard index value.

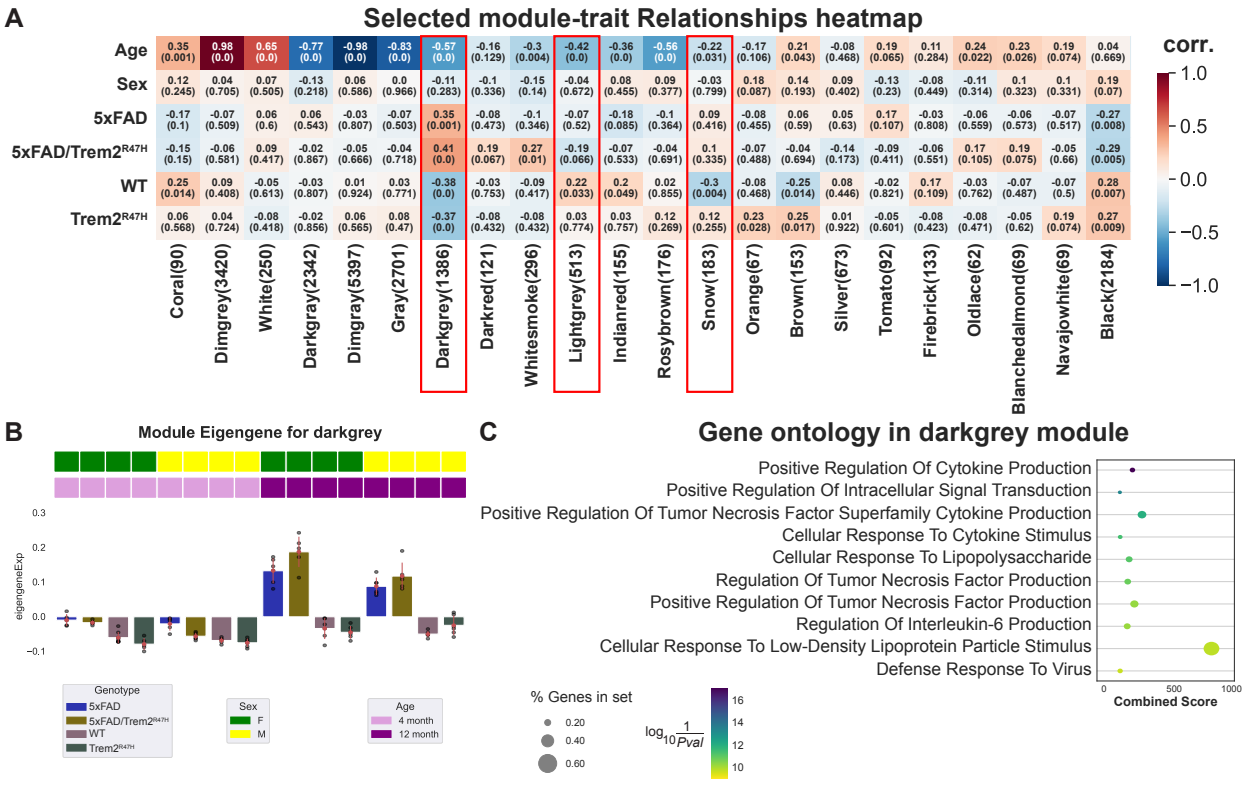

Figure 2.8: **Trem2 modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The darkgrey module eigengene expression profile from the Trem2 mouse model is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in darkgrey modules.

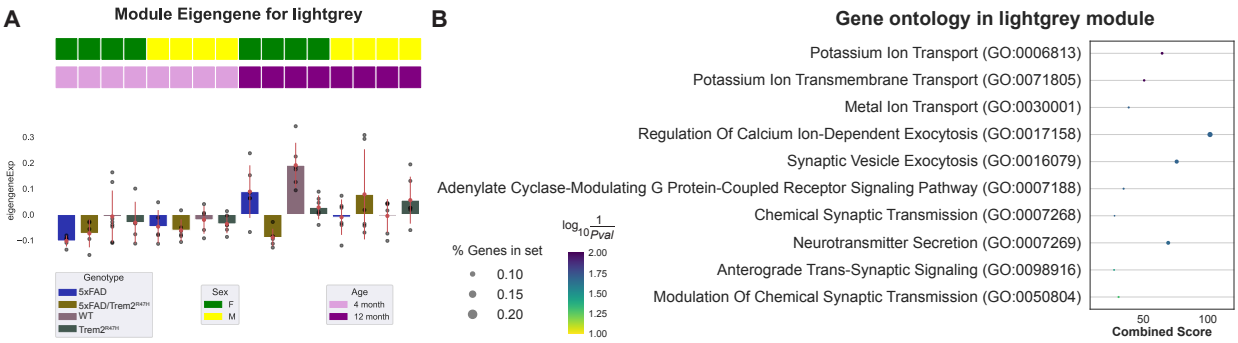

Figure 2.9: **Trem2 neuronal module during the progression of the AD A)** The lightgrey module eigengene expression profile from the Trem2 mouse model is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **B)** GO analysis of the genes in lightgrey modules.

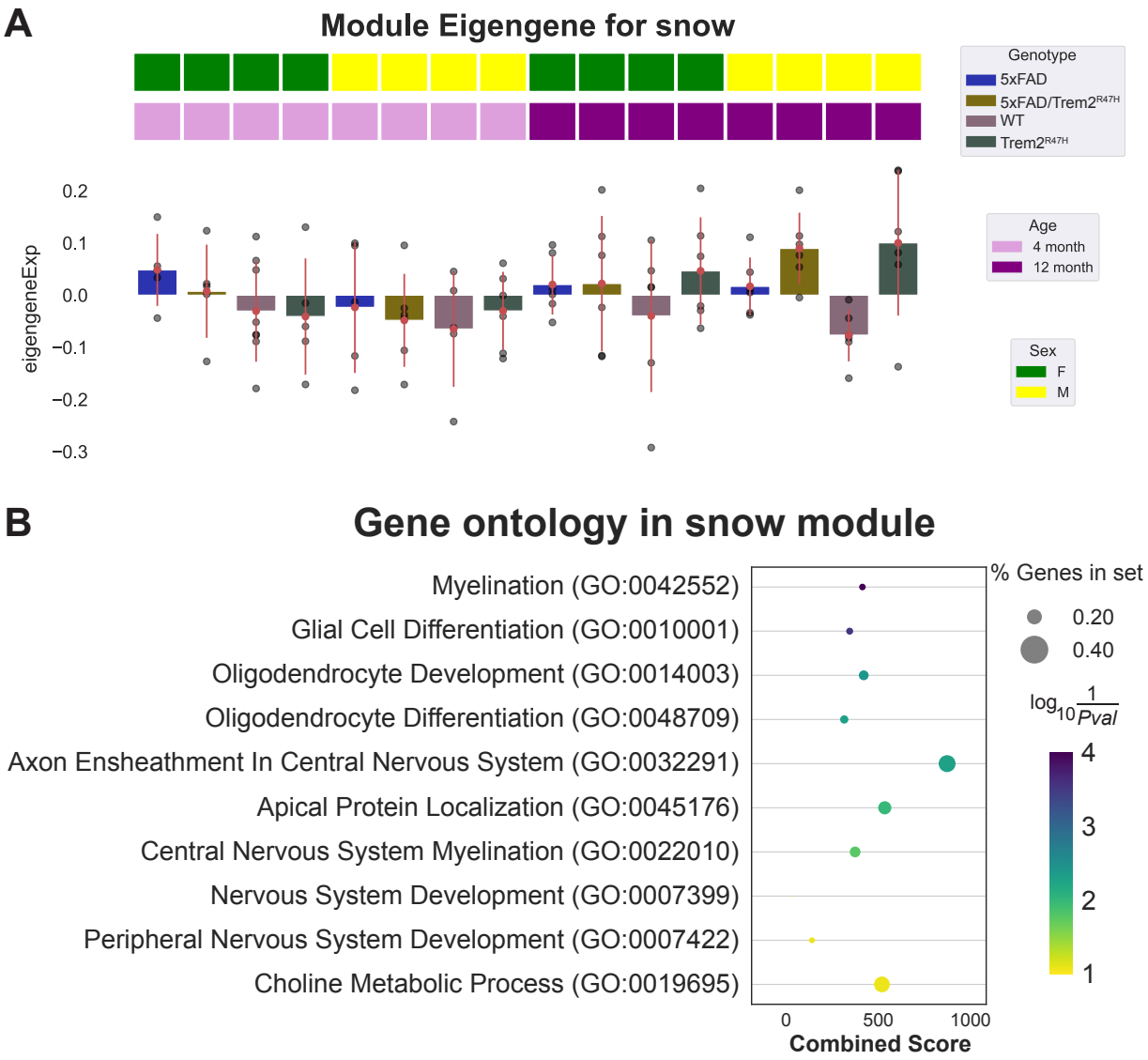

Figure 2.10: **Trem2 myelination module during the progression of the AD A)** The snow module eigengene expression profile from the Trem2 mouse model is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **B)** GO analysis of the genes in snow modules.

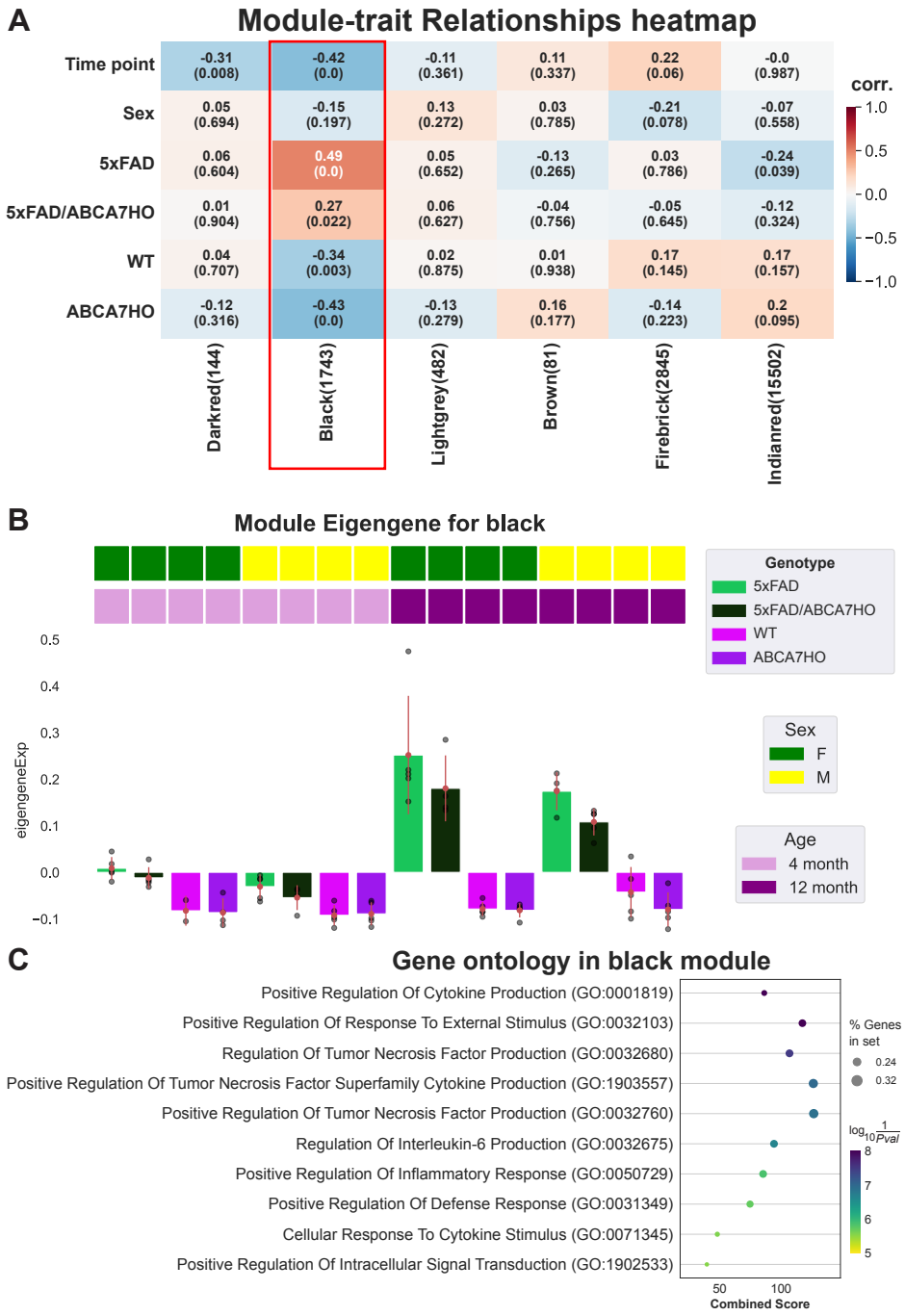

Figure 2.11: **ABCA7 modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The black module eigengene expression profile is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in black modules.

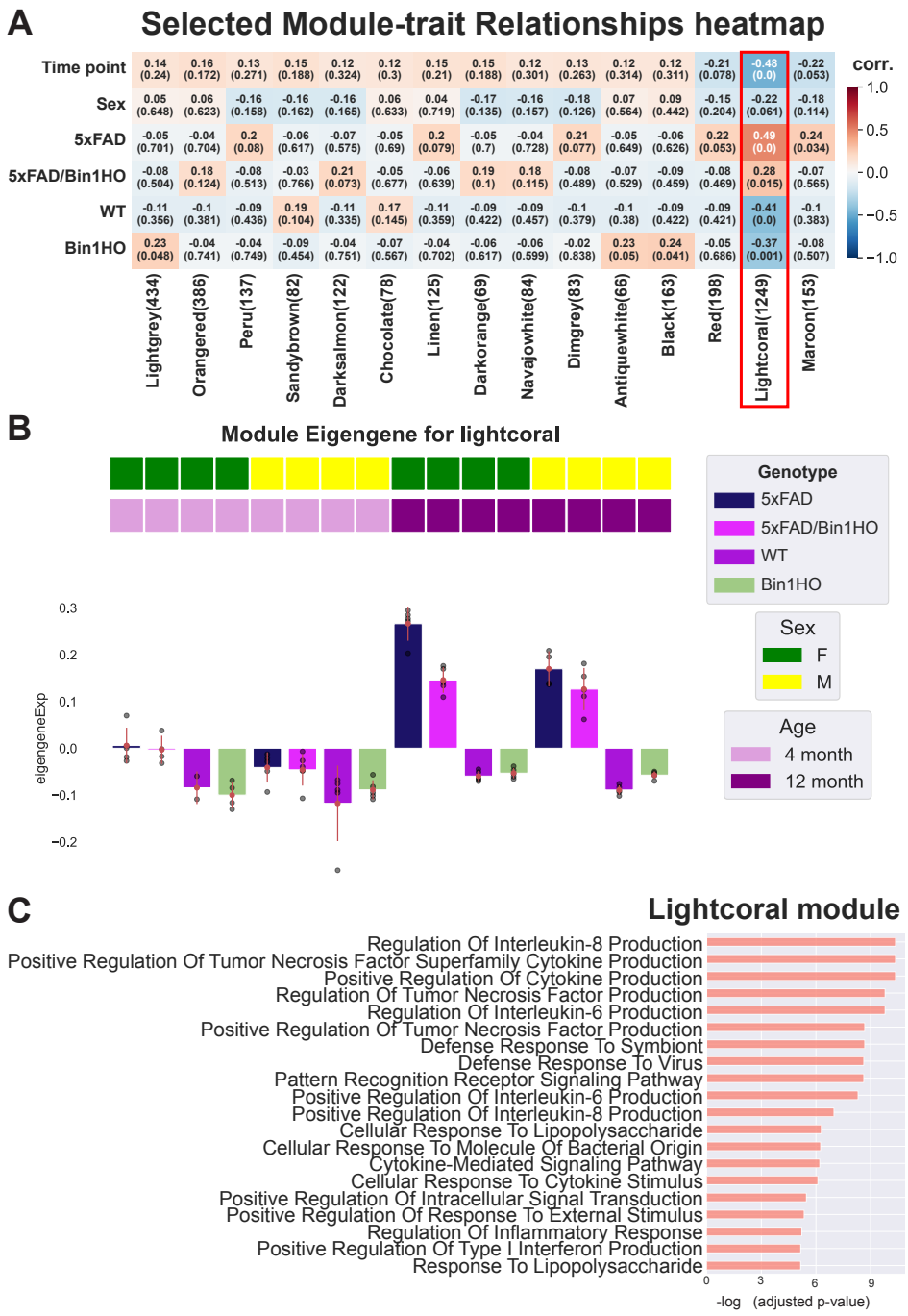

Figure 2.12: **Bin1 modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The lightcoral module eigengene expression profile is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in lightcoral modules.

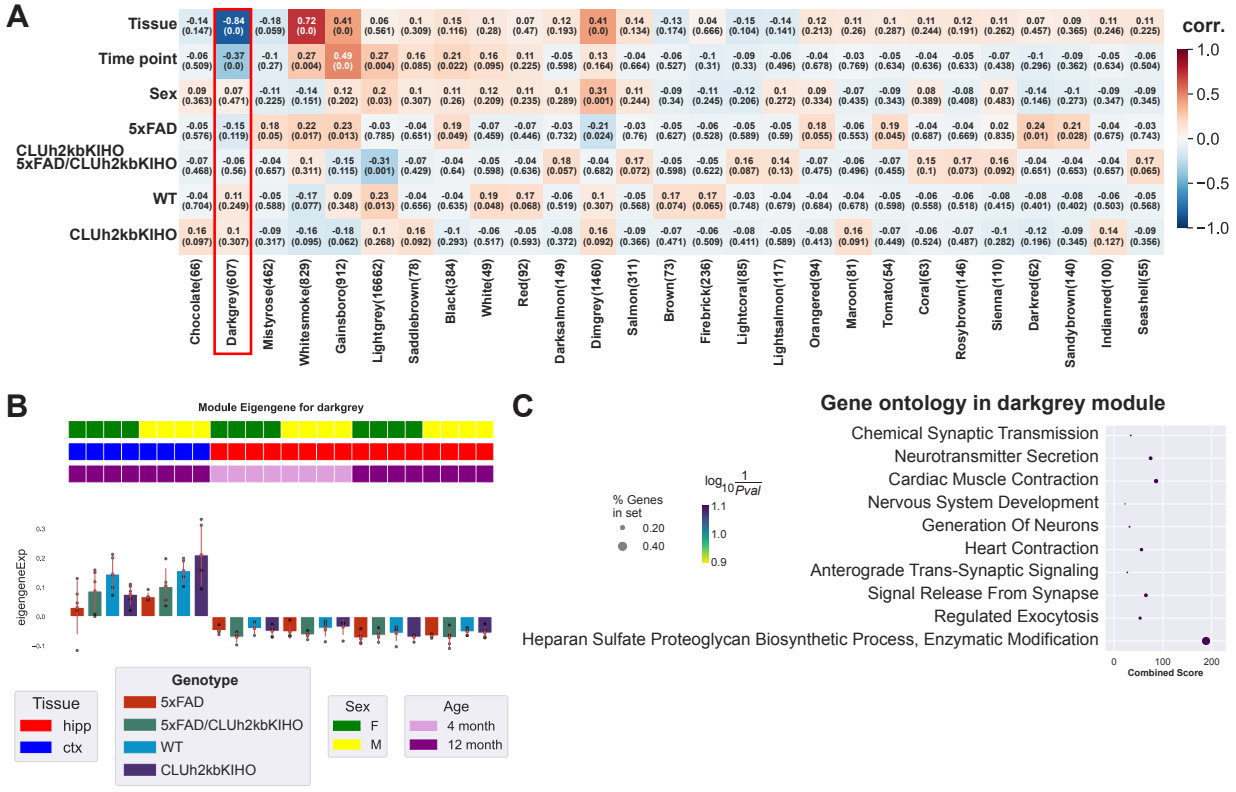

Figure 2.13: **CLU modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The darkgrey module eigengene expression profile is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in darkgrey modules.

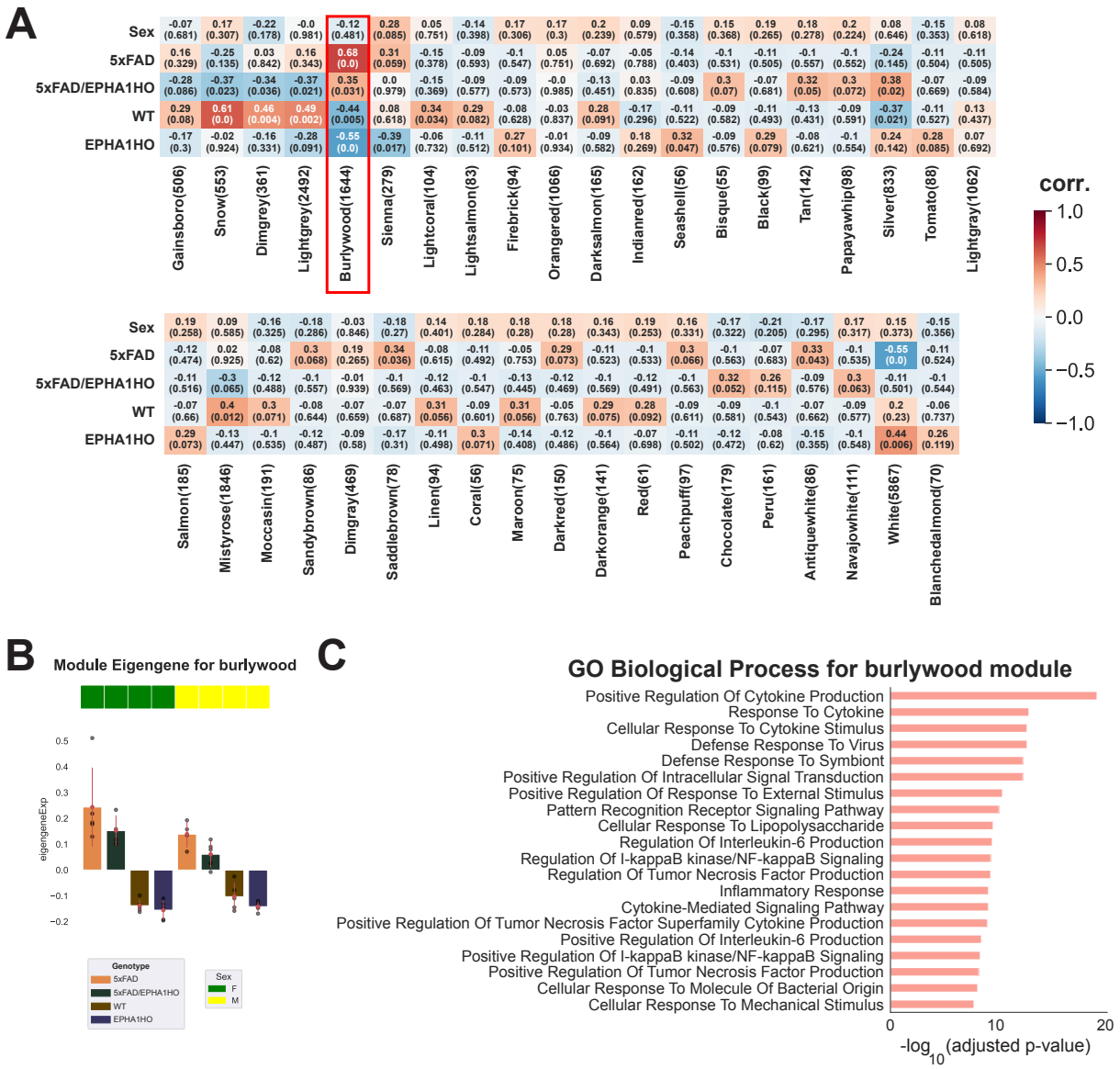

Figure 2.14: **Epha modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** the burlywood module eigengene expression profile summarized by genotype. Above, the top two rows display sex and age for each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in burlywood modules.

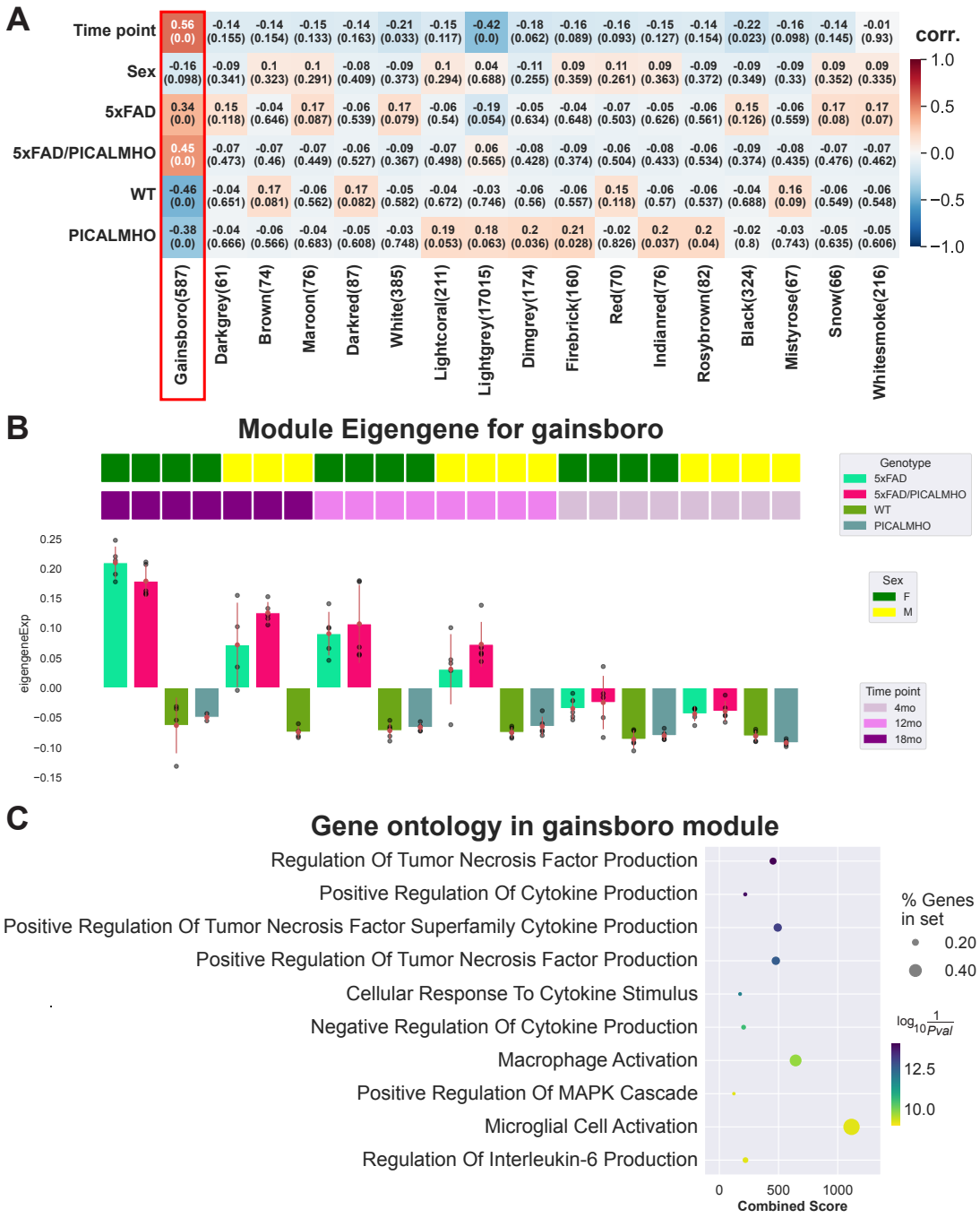

Figure 2.15: **PicalmH465R modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The gainsboro module eigengene expression profile is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in gainsboro modules.

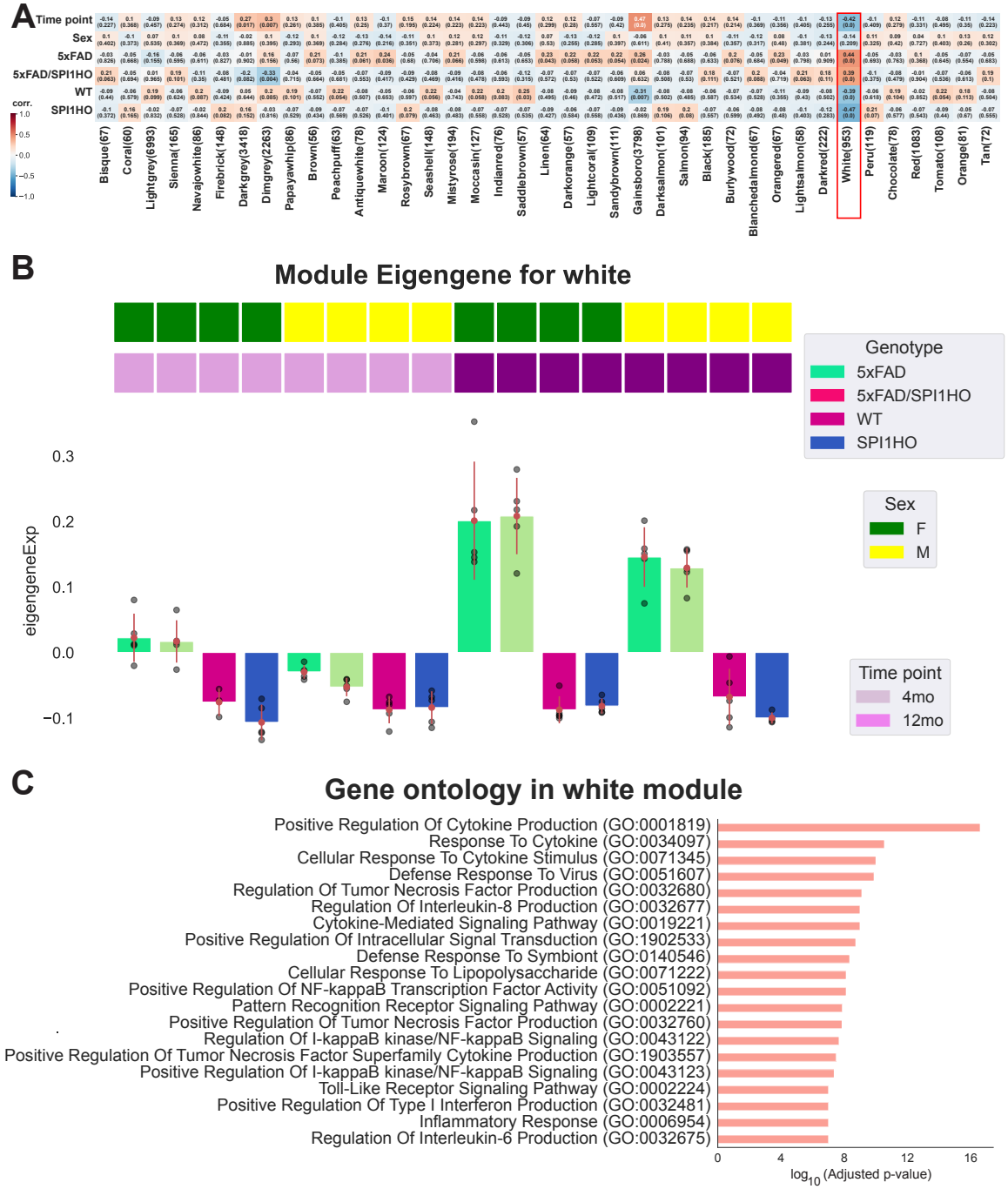

Figure 2.16: **Spi1 modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The white module eigengene expression profile is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in white modules.

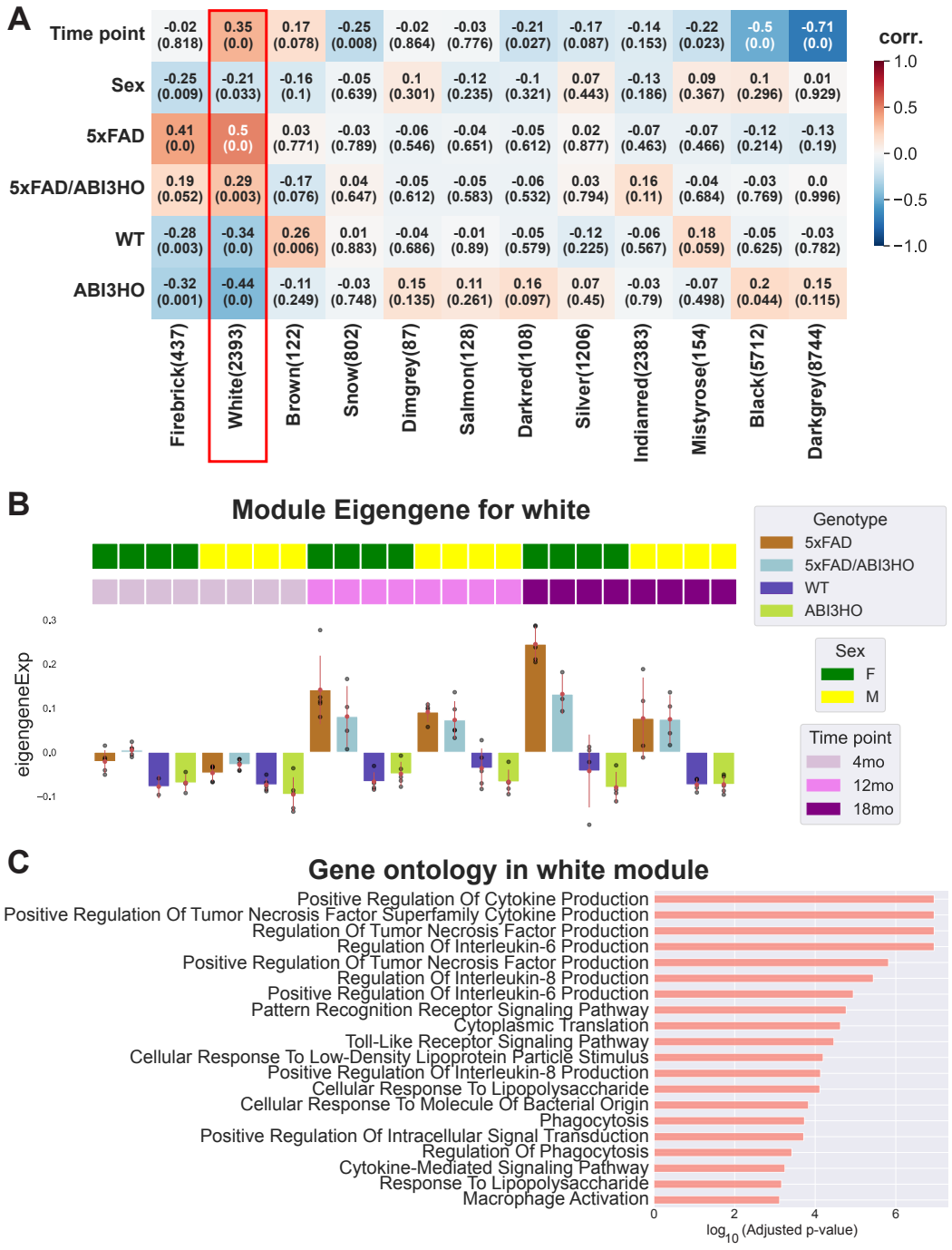

Figure 2.17: **ABI3 modules during the progression of the AD A)** Matrix with the Module-Trait Relationships (MTRs) and corresponding p-values between the detected modules on the y-axis and selected AD traits on the x-axis. The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation. **B)** The white module eigengene expression profile is summarized by genotype. Above, the top two rows display the sex and age of each sample. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points. **C)** GO analysis of the genes in white modules.

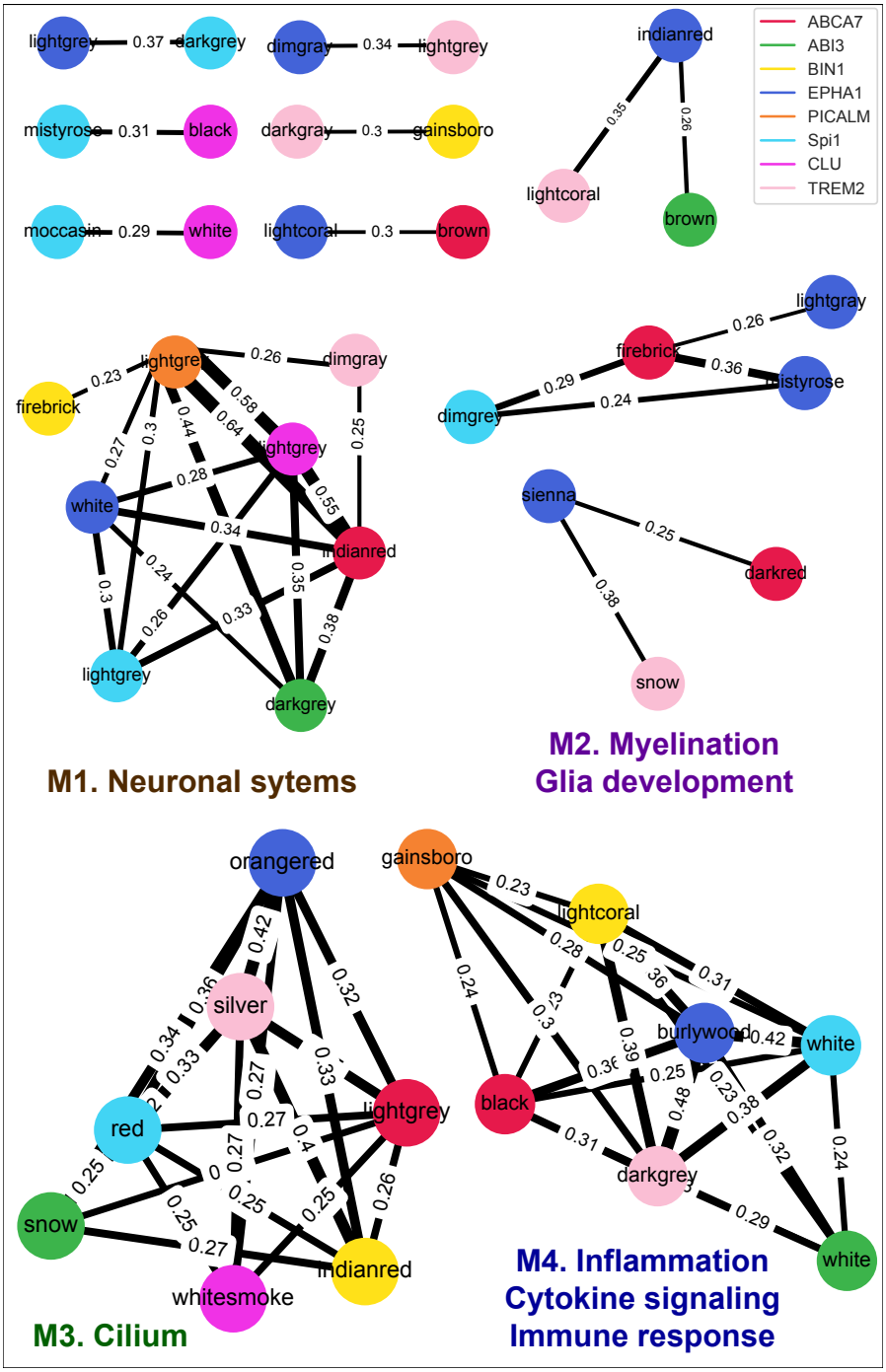

Figure 2.18: **Comparison of modules obtains from GWAS mouse models of AD.** Comparison graph of GWAS mouse model filtered by Jaccard similarity ¿ 0.15. The color of each dot shows the origin of modules and the thickness of the lines shows the Jaccard index value.

Age	Tissue	Sex	Genotype	samples
4mon	Cortex	Female	5xFADHEMI	5
4mon	Cortex	Female	C57BL/6J	5
4mon	Cortex	Male	5xFADHEMI	5
4mon	Cortex	Male	C57BL/6J	5
4mon	Hippocampus	Female	5xFADHEMI	5
4mon	Hippocampus	Female	C57BL/6J	5
4mon	Hippocampus	Male	5xFADHEMI	5
4mon	Hippocampus	Male	C57BL/6J	5
8mon	Cortex	Female	5xFADHEMI	5
8mon	Cortex	Female	C57BL/6J	5
8mon	Cortex	Male	5xFADHEMI	4
8mon	Cortex	Male	C57BL/6J	6
8mon	Hippocampus	Female	5xFADHEMI	5
8mon	Hippocampus	Female	C57BL/6J	5
8mon	Hippocampus	Male	5xFADHEMI	4
8mon	Hippocampus	Male	C57BL/6J	6
12mon	Cortex	Female	5xFADHEMI	5
12mon	Cortex	Female	C57BL/6J	5
12mon	Cortex	Male	5xFADHEMI	4
12mon	Cortex	Male	C57BL/6J	6
12mon	Hippocampus	Female	5xFADHEMI	5
12mon	Hippocampus	Female	C57BL/6J	5
12mon	Hippocampus	Male	5xFADHEMI	4
12mon	Hippocampus	Male	C57BL/6J	6
18mon	Cortex	Female	5xFADHEMI	6
18mon	Cortex	Female	C57BL/6J	15
18mon	Cortex	Male	5xFADHEMI	10
18mon	Cortex	Male	C57BL/6J	5
18mon	Hippocampus	Female	5xFADHEMI	6
18mon	Hippocampus	Female	C57BL/6J	15
18mon	Hippocampus	Male	5xFADHEMI	10
18mon	Hippocampus	Male	C57BL/6J	5

Table 2.1: 5xFAD mouse model and matching C57BL/6J mice samples.

Age	Tissue	Sex	Genotype	samples
4mon	Hippocampus	Female	3xTgAD	5
4mon	Hippocampus	Female	B6129SF1/J	5
12mon	Hippocampus	Female	3xTgAD	5
12mon	Hippocampus	Female	B6129SF1/J	5
18mon	Hippocampus	Female	3xTgAD	9
18mon	Hippocampus	Female	B6129SF1/J	9

Table 2.2: 3xTgAD mouse model and matching B6129SF1/J mice samples.

Chapter 3

Identification of robust cellular programs using reproducible LDA that impact sex-specific disease progression in different genotypes of a mouse model of AD

3.1 Abstract

The gene expression profiles of distinct cell types reflect complex genomic interactions among multiple simultaneous biological processes within each cell that can be altered by disease progression as well as genetic background. The Identification of these active cellular programs is an open challenge in the analysis of single-cell RNA-seq data. Latent Dirichlet Allocation (LDA) is a generative method used to identify recurring patterns in counts data, commonly

referred to as topics that can be used to interpret the state of each cell. However, LDA's interpretability is hindered by hyperparameter selection of the number of topics as well as the variability in topic definitions due to random initialization. We developed Topyfic, a Reproducible LDA (rLDA) package, to accurately infer the identity and activity of cellular programs in single-cell data, providing insights into the relative contributions of each program in individual cells. We apply Topyfic to brain single-cell and single-nucleus datasets of two 5xFAD mouse models of Alzheimer's disease crossed with C57BL6/J or CAST/EiJ mice to identify distinct cell types and states in cell types such as microglia. We find that 8-month 5xFAD/Cast F1 males show higher level of microglial activation than matching 5xFAD/BL6 F1 males, whereas female mice show similar levels of microglial activation. We show that regulatory genes such as TFs, microRNA host genes, and chromatin regulatory genes alone capture cell types and cell states. Our study highlights how topic modeling with a limited vocabulary of regulatory genes can identify gene expression programs in single-cell data to quantify similar and divergent cell states in distinct genotypes.

3.2 Introduction

The different cell types constituting a tissue work together to carry out the functions of that tissue in response to various developmental and environmental cues by activating specific cellular programs. Single-cell and single-nucleus RNA sequencing enable the identification of cell types, subtypes, and cell states through their single-cell transcriptomes, enhancing our understanding of cellular phenotype heterogeneity and composition within complex tissues such as the brain[124,247,248]. A common approach for cell type annotation relies primarily on unsupervised clustering methods[249], which partition cells based on the similarity of their gene expression patterns. This is followed by manual cell type assignment for each cluster based on differentially expressed markers from literature. The overall accuracy of this ap-

proach depends on both the clustering accuracy[250] and the prior knowledge of marker gene expression levels[251]. For example, marker genes could be expressed in more than one cell type, complicating the annotation process. More importantly, this cluster-based approach assumes that cells can only be part of a single cluster, thereby averaging the cell-to-cell variability within that cluster.

Here, we focus on a key challenge of inferring complex cellular states and identities that are encoded by patterns of gene expression. We assume that each cell or nucleus engages in a limited number of cellular programs, and its observed transcriptome is determined by the sum of these active programs. To represent this, we leverage grade of membership (GoM) models[153,252], allowing each cell to have partial membership in multiple cellular programs. One such model is Latent Dirichlet Allocation (LDA)[253], a probabilistic algorithm capable of inferring recurring combinations referred to as topics. LDA starts by randomly assigning topics to each word in a document, which in our case are cells[153]. Due to this random initialization, different topic assignments and, consequently, different topic representations for each document may arise in repeated runs. As a result, the topics discovered by LDA can vary across different runs of the algorithm. To address this issue of topic variability, one common approach is to use a fixed random seed before running LDA, ensuring consistent random initialization across different runs. However, there is no guarantee that this fixed seed will produce the best topics, as some randomly defined topics might be stuck in local optima that lack biological significance.

Alzheimer's disease (AD) is a progressive neurodegenerative disease characterized by memory loss[1]. Microglia are the resident macrophages of the brain that mediate brain homeostasis by regulating immune function and promoting neuronal homeostasis and neuroprotection. To maintain homeostasis, microglia damage or kill neurons with abnormal profiles[254,255]. However, not all the microglia in the brain behave identically. Single-cell RNA-seq studies have identified new microglial subtypes with unique transcriptional and functional characteristics,

termed "disease-associated microglia" (DAM) in animal models of AD[8]. DAMs are characterized by the upregulation of genes associated with late-onset AD, such as apolipoprotein E (*Apoe*), *Itgax*, *Csf1r*, and *Tyrobp*, whereas *Tmem119*, *Cd33*, and *Maf* are downregulated. However, in cluster-level analyses, these cells are typically treated either as part of distinct clusters or positioned on a pseudo time continuum between homeostasis and activation.

Here, we develop Topyfic, a Python package that (a) runs LDA multiple times with different random seeds, (b) aggregates similar topics across runs to compute reproducible topics, and (c) filters out low-participation topics. Using this strategy, we find reproducible topics and filter out noisy, irreproducible topics. We apply Topyfic to single-nucleus and single-cell datasets generated by the ENCODE and MODEL-AD consortiums from mice with and without the 5xFAD transgene in either a C57BL/6J (MODEL-AD) or a C57BL/6J x CAST/EiJ background (ENCODE) to identify topics that are (a) detected in both genotypes with and without the transgene, and (b) microglia-specific. We then train additional topics with a subset of regulatory genes such as transcription factors and show that these regulatory topics that we recover also capture cell activation using regulatory genes alone.

3.3 Results

3.3.1 Reproducible LDA topics using Topyfic

Topyfic estimates the most likely number of topics as well as maximizes the number of meaningful topics. The core idea of Topyfic is that topics that are found repeatedly across multiple LDA runs are more reliable than topics found in any single run, which could be suboptimal as the results of poor random initialization (Fig. 3.1.A). Starting from a cell-by-gene expression matrix in h5ad format, Topyfic first trains a Latent Dirichlet Allocation (LDA) model on the provided dataset using different random seeds (see methods for a detailed technical descrip-

tion of Topyfic parameter settings). Topyfic uses Leiden clustering[139] in gene-weight space to combine similar topics across individual runs to construct a consensus set of topics called the TopModel. Finally, Topyfic calculates cell-topic participation based on the TopModel and filters out any topic with user-defined low participation. Topyfic also includes several helper functions to analyze and visualize topics and topic participation in the training and testing datasets.

Determining the number of topics is a challenging step for the application of LDA. We use two diagnostic metrics and visualizations to help guide this decision. First, we train our TopModels using a different starting number of topics, followed by pruning low-participation topics. In general, applying our consensus models to repeated runs with a smaller number of topics (K) than the optimal number will lead us to discover more clusters of topics (N). As we increase K, we expect N to stay relatively stable until higher Ks result in fewer N. We therefore select our parameter K to be when K=N. Alternatively, we can also calculate the perplexity. A lower perplexity score is an indication of a better model. We observe that perplexity tends to decrease rapidly and then flattens out. In this approach, we choose the smallest value of K that is able to explain the data: i.e. the value of K at the point in which the perplexity flattens out (Fig. 3.5.A).

We evaluated the sensitivity of the resulting topics to important parameters. First, we investigated the effect of the number of cells on the number of reproducible topics. As expected, increasing the number of cells led to the identification of rarer and diverse gene expression programs with more topics (Fig. 3.5.B). In the final steps of building our TopModels, we filter topics with low cell participation. Increasing the minimum cell participation threshold enabled us to retain topics with stronger signals in each cell (Fig. 3.5.C). As a default criterion, we focused on topics that represent at least 1% of the gene expression in cells. We then use Leiden clustering to form our consensus topics, therefore all the inputs related to clustering can also be altered such as the resolution, which is a value controlling the coarse-

ness of the clustering. We can recover more topics by increasing this value, but we found the defaults adequate for our analyses below (Fig. 3.5.D).

3.3.2 Comparing a mouse model of AD across two genetic backgrounds

We first analyzed the overall characteristics of two related mouse brain datasets for the 5xFAD mouse model of AD[120] and matching controls for 8 month old mice before applying Topyfic. The genetic background of the 5xFAD mouse model of AD is C57BL/6J ("BL6") and it is normally studied as a hemizygote, i.e. with only one copy of the transgene on an otherwise regular BL6 background. The first dataset consists of single-nucleus RNA-seq from the cortex and hippocampus of 2 male and 2 female 5xFAD x CAST/EiJ F1 hybrid mice and matching BL6 x CAST/EiJ controls from the ENCODE consortium, while the second dataset consists of 2 male and 2 female cortex and hippocampus snRNA-seq as well as microglia single-cell RNA-seq of 5xFAD hemizygotes and matching BL6 controls from the MODEL-AD consortium (Fig. 3.1.B-C). As all mice from ENCODE have the BL6 x CAST/EiJ F1 background, and all MODEL-AD mice have the BL6 background, we will use the consortium names and genotypes interchangeably. All experiments were performed as previously described (refs) using the Parse Biosciences split-pool method[13,256], which we refer to as Split-seq. Separate Split-seq experiments were performed for the ENCODE and MODEL-AD mice and were deposited in their respective online repositories. For each dataset, demultiplexing and alignment were carried out using Parse Biosciences' split-pipe software and STARSolo[257]. Scrublet[258] was employed to identify doublets in each dataset, followed by quality control (QC) using Seurat[146] (Methods). The filtering successfully recovered a combined total of 110,907 nuclei and 5,546 microglia cells (Fig. 3.1.C), which were annotated using marker genes and label transfer with external reference data from the Allen Brain Institute dataset[259] (Fig. 3.6-3.7 and Methods).

Glial cells such as microglia, astrocytes, and oligodendrocytes constitute a substantial fraction of the mammalian brain, representing 27.5% of the nuclei in our snRNA-seq datasets. The proportion of glial cells is influenced by several factors, including genotype, sex, and brain region. We examined the variation of glial cells by genotype and sex in each tissue separately. As expected, we found a higher portion of microglial cells in 5xFAD mice regardless of genetic background. Despite recovering more nuclei from BL6 mice, we observed a higher proportion of glial nuclei in the BL6/CAST genotype, highlighting how genetic diversity contributes to substantial differences in glial cell abundance. Interestingly, there is more variation between sexes in mice with the BL6 background compared to mice with the BL6/CAST background. In particular, 5xFAD/CAST males have similar numbers of microglia in the hippocampus when compared to 5xFAD/CAST females, which is substantially higher than 5xFAD males (Fig. 3.1.D-E). Expression levels of marker genes for disease-associated microglia (DAM), astrocytes, and oligodendrocytes in pseudo bulk for each mouse showed higher expression in 5xFAD versus WT and more uniformity between replicates and sexes in 5xFAD/CAST than 5xFAD in both hippocampus and cortex (Fig. 3.1.F). Principal component analysis (PCA) confirmed that genotype contributes to major transcriptomic differences across the dataset, with PC2 (8.54%) corresponding to genotype and PC3 (6.14%) corresponding to brain region (Fig. 3.1.G and Methods). Comparison of the median number of UMIs across cells in each cell type in AD and WT samples reveals reproducible patterns across both genotypes, such as neurons generally having more UMIs compared to glial cell types (Fig. 3.1.H-I). Interestingly, we do not detect differences in the number of UMI between microglia whether using nuclei or whole cells (Fig. 3.1.H). In general, we observe a higher number of UMIs per nuclei in BL6/CAST genotype even though both consortia used similar sequencing depths. These results indicate that the differences are more likely associated with the genetic identity of mice, such as genotype, rather than technical procedures such as sequencing depth. In summary, 5xFAD/CAST males at 8 months show a higher proportion of DAM microglia that matches their female counterparts, unlike

regular 5xFAD males at 8 months, which have lower proportion of DAM microglia than their female counterparts.

3.3.3 Identifying topics related to cell type and cell state

We trained Topyfic using (a) 1 male replicate and 1 female replicate of WT mice from both genotypes (4 mice) and (b) 1 male replicate and 1 female replicate of 5xFAD transgene-carrying mice from both genotypes (4 mice) separately using all genes (Fig. 3.2.A) with varying numbers of topics, ranging from K = 5 to 50. This iterative process allowed us to evaluate different K values and identify the final number of topics (N) that best captured the underlying structure in our data, which was at K=15 (Fig. 3.2.B, methods). The TopModels for each genotype were aggregated to form our final TopModel with 28 topics that passed our low participation filter on the second replicates. For comparison with subsequent topics derived from different gene sets and cell types, we label these topics as asn1 through 28, where 'asn' stands for 'all genes, single-nucleus'.

We assessed the distribution of cell-topic participation, focusing on whether a topic was the predominant topic in a cell. We also performed a topic-trait relationships analysis to capture correlations between each topic and major cell types, cell states, genotype (BL6, BL6/CAST), and transgene presence (5xFAD and WT) (Fig. 3.2.C). Topics exhibiting high cell participation consistently showed enrichment for specific cell types. Conversely, topics with low participation were generally not associated with any particular cell type. We identified topics asn4, asn17, and asn26 as corresponding to microglia, each displaying varying levels of cell-topic participation. Notably, asn17 exhibited the highest participation, while asn26 showed the lowest (Fig. 3.2.C). Our snRNA-seq dataset includes 2,440 microglia, 77% of which are from mice with the 5xFAD transgene. The structure plot of microglia nuclei displays cell-topic participation as a stacked bar plot for each nucleus, grouped by

genotype, sex, and tissue (Fig 2D, methods). While both asn4 and asn17 are correlated with presence of the transgene, asn programs in microglia shows that asn17 dominates in the transgenic mice. In mice without the transgene, microglia have a predominant mixture of cellular programs consisting of 50% asn4, 20% asn17, 7% asn26, and 23% from the remaining topics. By contrast, transgenic mice show two distinct cellular program patterns, suggesting the presence of two distinct cellular states in the mice. A minority of cells show program combinations that resemble the WT samples, which indicate a population of homeostatic microglia. The majority of microglia in transgenic mice show a significantly higher ($\sim$65%) participation of asn17, which represents the heightened activation state of these microglia (Fig. 3.2.E). Thus, Topyfic recovers topics representing different cell states and cell types for minor cell types such as microglia across both genotypes.

The importance of a gene's expression to a given topic is called the gene weight. To gain insights into the differences between the two major microglia topics, we compared the weights of genes in asn4 and asn17 that have weights greater than 1. We found 120 genes specific to asn4 and 585 genes specific to asn17, as well as 1,659 genes shared between the two topics (Fig. 3.2.F). Key genes associated with homeostatic microglia, such as *Tmem119*, exhibited significantly higher weights in asn4 compared to asn17. In contrast, genes linked to disease-associated microglia (DAM) such as *Csf1r*, *Itgax*, and *Apoe*, were exclusively represented in asn17 (Fig. 3.2.F). By using an MA plot to compare topics, we found a total of 138 genes with differential weights (modified z-score > 2) in asn4, including 120 of with large absolute log ratios (M) value (> 6.5). Conversely, asn17 displayed differential weights in 835 genes (modified z-score < -2), of which 585 had absolute log ratio (M) values > 7.5. In particular, genes overexpressed in stage 2 DAMs, such as *Apoe*, *Itgax*, *Csf1r*, *Lpl*, and *Axl13* were among the differentially higher weighted genes in asn17 (Fig. 3.2.G). Genes related to microglial cell identity, such as *Tgfbr1* and *Hexb*[260], shared similar ranks in both topics, even though they had higher weights in asn17. In contrast, genes primarily expressed in homeostatic microglia such as *Tmem119* and *Slc2s5* were exclusively represented in asn4, whereas DAM

genes *Apoe*, *Itgax*, and *Csf1r* only had significant weights in asn17 (Fig. 3.2.H). Thus, genes with shared or have specific weights in a topic can be correlated to the known underlying biology, as demonstrated in this case by the microglial neuroinflammatory signatures in mice with the 5xFAD transgene.

3.3.4 Recovering topics for different activation levels in microglia scRNA-seq

Having demonstrated its performance and utility on snRNA-seq from tissues, we applied Topyfic to our complete microglia single-cell data from 5xFAD and matching BL6. After training the TopModel with multiple values of K, we selected K=5, yielding 6 topics labeled sc1-sc6 (Methods). Each topic is the top participating topic in a subset of microglia (Fig. 3.3.A). While calculating the correlation between each topic and sex did not reveal any sex-specific topics (Fig. 3.3.B), analyzing the contribution of each topic in each genotype uncovered differences in activity between the genotypes (Fig. 3.3.C). The structure plot illustrates three distinct cellular programs. The first combination of programs is high in sc6 with high levels of *Csf1*, and is more prevalent in 5xFAD than in BL6. The second combination of programs contains a relatively consistent proportion of sc2, sc3, and sc5 in both genotypes, suggesting a closer association with homeostatic microglia. A subset of these homeostatic cells also have participation of sc6 and low levels of sc1. The third combination of programs is more pronounced in 5xFAD mice and is primarily composed of sc1, with significantly lower participation of sc3 and sc5 compared to the previous program, indicating a stronger association with activated microglia (Fig. 3.3.D).

Comparison of genes with weights > 1 between sc1 and sc2 revealed differential weights in 1,794 genes, with only 24 genes in sc2, including the *miR-155* host gene (Fig. 3.3.E). MicroRNAs (miRNAs) play a role in modulating inflammatory responses in microglia, and

their profiles are altered in Alzheimer's disease (AD). Notably, the pro-inflammatory miRNA, *miR-155*, shows increased expression in the AD brain[261]. We found disease-associated microglia (DAM) genes such as *Trem2*, *Apoe*, *Itgax*, *Clec7a*, *Axl*, and *Lpl*, alongside typical microglia gene markers such as *Tmem119*, *Olfml3*, and *Cd68* with higher weights in sc1 (Fig. 3.3.E). A comparison between sc1 and sc3 reveals 562 genes with higher weights in sc1 (modified z-score $>$ 2) and 247 genes with higher weights in sc3 (modified z-score $<$ -2) (Fig. 3.3.F). Primed microglia have the potential to induce the production of amyloid β (Aβ), tau pathology, neuroinflammation, and reduce the release of neurotrophic factors. This can lead to the loss of normal neurons in both quantity and function, a phenomenon strongly associated with AD. Genes such as *Cst7* and *Slc2a5*, part of the primed microglia pathway[262], were upregulated in sc1. Comparison between sc1 and sc5 reveals 465 genes with weights higher in sc1 (modified z-score $>$ 2) and 280 genes with weights higher in sc5 (modified z-score $<$ -2) (Fig. 3.3.G). In all three comparisons, genes associated with disease associated microglia are higher in sc1 compared to the three homeostatic programs.

Weights and ranks of homeostatic and DAM genes across microglial cells and microglial nuclei are similar between topics sc1 and asn17. Genes such as *Lpl*, *Apoe*, and *Trem2* exhibit higher gene weights and lower ranks in sc1 and asn17 compared to the rest of the single-cell topics, including microglial topic asn4 (Fig 3H). Lipoprotein lipase (*Lpl*), the rate-limiting enzyme in lipoprotein hydrolysis, is predominantly expressed in microglia such as phagocytic disease-associated microglia thought to be protective in AD[263]. Increased expression of genes such as *ApoE*, *Trem2*, and *Lpl* in microglia during development, damage, and disease suggests that increased lipid metabolism is needed to fuel protective cellular functions such as phagocytosis[264]. Thus, Topyfic recovered an activated microglial state topic using either single-cell and single-nucleus RNA-seq, with the main difference corresponding to how scRNA-seq topics capture multiple subtypes of homeostatic microglial programs.

3.3.5 Topics derived from regulatory genes are sufficient to define cell types and cell states

While numerous genes are used as markers for distinct cell types and states, we hypothesized that cellular programs are fundamentally constructed from a core set of regulatory genes. Therefore, we explored identifying cellular programs using a restricted LDA vocabulary of regulatory genes. Transcription factors and other genes based on Gene Ontology (GO) term annotations were chosen based on their impact on transcriptional regulation, including known regulatory genes such as the Id family (inhibitors of DNA binding and cell differentiation, despite lacking a DNA binding domain themselves)[265] (Methods). Overall, TFs constitute approximately 50% of the 2,701 genes included in our regulatory gene list (Fig. 3.4.A). This approach aims to elucidate impactful cellular programs using a curated vocabulary.

We trained Topyfic models on 52,685 nuclei across 2,701 regulatory genes using various K values as described previously, and once again combined models with K = 15 to obtain 27 reproducible topics labeled as rsn (regulatory single nucleus). Analysis of topic-trait relationships showed that individual topics are highly correlated to specific cell types, 5xFAD transgene presence, or genotype. For instance, both rsn1 and rsn22 correspond to microglia while rsn1 also correlated to the transgene presence, rsn3 and rsn15 to astrocytes, rsn2 and rsn14 to oligodendrocytes, and a dozen topics correspond to different neuronal subtypes (Fig. 3.4.B). Similarly to the results using all genes, topics with the highest maximum cell participation are consistently enriched for specific cell types or cell states. By contrast, topics with low maximum cell participation are typically not associated with any particular cell type (Fig. 3.4.B). At our chosen resolution, all cell types with > 350 nuclei (0.32%) are associated with at least one topic. In summary, our topics demonstrate a striking alignment with our annotated cell types.

The structure plot of microglia nuclei shows that as cells activate, rsn22 is gradually replaced

by rsn1 while minor topics remain constant (Fig. 3.4.C). Comparing the gene weights of the two microglia topics reveals remarkably similar topic compositions, aligning with our expectations (Fig. 3.4.D). Only a few genes exhibit an absolute modified z-score value > 2 (93 in rsn1, 15 in rsn22) (Fig. 3.4.D). Among the 93 genes upregulated in rsn1, microglia gene markers such as *Ank*, *Hif1a*, *Arid5b*, *Creb3l2*, and *Srpk2* show the highest modified z-scores. The expression of serine/threonine-protein kinase 2 (Srpk2) is associated with the production of proinflammatory cytokines and M1 polarization of microglial cells, suggesting a potential connection to the cognitive decline observed in the AD mice model[266]. *Ank* is a membrane-phosphate transporter that has a microRNA, *Mir7117*, embedded in its intron. *Ank* exhibits the highest absolute log ratio (M) value (10.8) and is also up-regulated in laser-captured microglia in the brains of individuals with AD[267]. In summary, regulatory genes alone can differentiate between homeostatic and disease-associated microglial states.

We evaluated the similarity between topics learned using all genes (asn) and regulatory genes (rsn) using cosine similarity (methods). Using a similarity threshold of > 0.9, we identified 14 clusters of highly correlated topics that matched asn topics to rsn topics (Fig. 3.4.E). As anticipated, topics representing common cell types were found to cluster together with at least one topic from each method type. At the selected threshold, activated microglia topic rsn1 only matches topic asn17, whereas homeostatic microglia rsn22 only matches the corresponding asn4. We further clustered the nine microglia topics (2 asn, 6 sc, 2 rsn) and found higher correlation of activated microglia topic sc1 with asn17 and rsn1 than with the other asn and rsn topics (Fig. 3.4.F). Thus, the expression patterns of regulatory genes alone are adequate to define cell types in the brain and states of microglial activation.

3.4 Discussion

We developed a grade of membership (GoM) model using Latent Dirichlet allocation (LDA) called Topyfic and applied it to mouse single cell and single nucleus RNA-seq brain datasets to infer topics that capture cell types, subtypes, and cell states. Current implementations of LDA for analyzing single cell RNA-seq data do not consider the stability and consistency of the model[157,268–270]. A robust LDA model should be less sensitive to variation in the initial conditions, such as different random seeds. To achieve this, we implemented reproducible LDA (rLDA) in Topyfic, which automatically runs LDA multiple times and aggregates similar topics. In particular, our strategy identifies the optimal number of topics at the given resolution. We then score topic participation in the whole dataset, which enables us to anchor each topic to a specific cell type, guiding the subsequent inference of the global topic distributions over genes to prioritize genes differentially weighted in each topic.

We applied Topyfic to mouse brain tissues and microglia cells in control and AD mouse models such as 5xFAD to validate our strategy, considering its thorough examination in prior studies. We recovered topics that reflect different cell types, including glia with different levels of activation. Interestingly, our findings indicate increased DAM microglia in male 5xFAD/CAST F1s compared to male 5xFAD/BL6 mice. While we recover the expected sex-specific difference in the higher amount of detected DAM microglia in female 5xFAD/BL6 mice compared to male 5xFAD/BL6 mice as previously reported[120], this pattern is not seen in 5xFAD/CAST F1s, where both sexes show an equal proportion of DAM microglia. This suggests that genetic variation affects disease progression in the different sexes in mouse models of AD. While previous studies have compared APP/PS1 transgene expression in inbred wild-derived strains such as CAST also at 8 months of age and found similar levels of microglia activation in transgenic CAST and transgenic BL6, the study only analyzed females[271]. Whether it is 5xFAD/CAST F1 males that are atypical because of their increased activation or alternatively 5xFAD/BL6 inbred males that are atypical because of their lower

activation remains to be determined.

We have shown that Topyfic recovers topics related to cell types or activities, including multiple topics relating to distinct activation states in microglia. We also show that using regulatory genes is enough to identify cell types or cell states by limiting our dictionary of genes to those that are the most likely to be directly involved in transcriptional and post-transcriptional regulation. For example, the top differential gene in topic rsn1 was *Ank1*, which has a 4-fold upregulation in AD microglia[267]. While *Ank1* is a structural protein, it is also the host gene of *Mir7117*, which suggests that the microRNA could be playing a role in AD. This potential role for *Ank1* as a microRNA host gene in topic-based cellular programs would have been missed in traditional protein-coding gene marker analysis.

Cellular programs defined using topic modeling have two major benefits over traditional cluster-based approaches to single-cell analysis. First, each gene can contribute more than one topic with different weights, which is more reflective of the pleiotropic nature of gene activity. The second substantial benefit is that each nucleus can have a unique linear combination of topic membership, rather than force it to be only part of a homogeneous cluster[142,146]. This approach infers a higher and more abstract level of transcriptional activity, represented as topics. The latent topic structure is biologically interpretable, as cells carry out their functions by engaging simultaneously in multiple cellular programs related to cell identity, activation state, cell cycle, or circadian rhythm. Each pathway relies on different amounts of a gene's product, and the overall gene expression reflects the combined requirements across all topics. This GoM representation of cell-topic participation can be interpreted as a dimensionally-reduced portrayal of a cell's distinct transcriptional activities, which is what we are truly interested in elucidating when applying single-cell techniques to study already well-surveyed samples that have been previously characterized by bulk techniques. The sooner the field moves from individual gene marker-based analyses to cellular-program based approaches, the more likely we are to obtain useful biological insights to match our

wants and needs.

Acknowledgements

XXX

Contributions

A.M. and N.R. designed the package and study and wrote the manuscript. N.R. wrote Topyfic and its documentation, performed data analysis, and generated figures. E.R. performed Parse Biosciences snRNA-seq experiments for the ENCODE samples and preprocessed the data. B.A.W. processed the tissues for the ENCODE samples. H.Y.L. performed Parse Biosciences sc and snRNA-seq experiments for the MODEL-AD samples. All coauthors edited the manuscript.

Data availability

- Mouse snRNA-seq dataset from ENCODE
- Mouse sn/scRNA-seq dataset from MODEL-AD

Code availability

- Data preprocessing cell type annotation
- Data processing using Topyfic and figure generation code
- Topyfic

3.5 Materials

3.5.1 Mice and tissue collection

All mice were housed following the guidelines outlined in the Guide for Care and Use of Laboratory Animals. Approval for all experimental procedures was obtained from UCI's Institutional Animal Care and Use Committee (IACUC), adhering to both institutional and national guidelines. Model-AD samples were obtained from 5xFAD/BL6 mice (Tg(APPSwFlLon, PSEN1*M146L*L286V) 6799Vas/Mmjax, RRID: MMRRC–034840-JAX) covered under the IACUC protocol #AUP-21-100 and bred by the Transgenic Mouse Facility at UCI. Left cortex and left hippocampus tissues from 8 month old mice were snap frozen in liquid nitrogen at UCI and stored at -80°C. ENCODE samples were obtained from 5xFAD x CAST/EiJ (RRID: IMSR–JAX:000928) F1 hybrids covered by IACUC protocol #IA21-1647 and bred by Jackson Laboratories (JAX). Left cortex and left hippocampus tissues from 8 month old mice were snap frozen in liquid nitrogen at JAX and shipped to UCI on dry ice.

3.5.2 Single-nucleus isolation and fixation

All single-nucleus samples regardless of genotype or tissue were processed identically. On ice, tissues from each mouse were transferred to a chilled gentleMACS C Tube (Miltenyi Biotec cat. #130-093-237) with 2 mL Nuclei Extraction Buffer (Miltenyi Biotec cat. #130-128-024) supplemented with 0.2 $U/\mu L$ RNase Inhibitor (NEB cat. #M0314L). A gentleMACS Octo Dissociator (Miltenyi Biotec cat. #130-095-937) was used to dissociate nuclei from whole tissues. The resulting suspensions underwent rounds of filtering through mesh strainers (70 μm, Miltenyi Biotec cat. #130-110-916, then 30 μm, #130-098-458). Finally, nuclei were resuspended in PBS + 7.5% BSA (Life Technologies cat. #15260037) and 0.2 $U/\mu L$ RNase inhibitor and kept on ice. Manual counting was performed using a hemocytometer and DAPI

stain (Thermo cat. #R37606). After counting, nuclei were fixed using Parse Biosciences' Nuclei Fixation Kit v1 (Parse Biosciences cat. #WN100), following the manufacturer's protocol. Between 1 and 4 million nuclei per sample were incubated in fixation solution for 10 minutes on ice, followed by permeabilization for 3 minutes on ice. The reaction was quenched and nuclei were centrifuged and resuspended in 300 μL Nuclei Buffer (Parse Biosciences cat. #WN101) and DMSO (Parse Biosciences cat. #WN105). The fixed samples were assessed under a microscope and manually counted as previously described. Aliquots of fixed nuclei were slow-frozen in a Mr. Frosty (Thermo cat. #5100-0001) and stored at -80°C.

3.5.3 Microglia single-cell isolation and fixation

Freshly prepared tissues were used for microglia isolation. Perfused right cortex and hippocampus were dissociated together using the Adult Brain dissociation kit (Miltenyi Biotec cat. #130-107-677) and gentleMACS Octo Dissociator (Miltenyi Biotec cat. #130-095-937) with heating. The resulting suspension was filtered with a 70μm mesh strainer (Miltenyi Biotec cat. #130-110-916). Debris were removed using the debris removal solution from the dissociation kit. Myelin were removed from the single-cell suspensions using negative selection with Myelin Removal Beads II (Miltenyi Biotec cat. #130-096-733) and LS columns (Miltenyi Biotec cat. #130-042-401). The resulting cells were enriched for microglia with magnetic labeling and positive selection using CD11b MicroBeads (Miltenyi Biotec cat. #130-093-634) and LS columns. Isolated microglia eluted in 1.8mL of bead buffer (0.5% BSA in DPBS) from the LS columns were centrifuged at 550 xg for 10 min at 4°C. Cell pellet with 10-15uL bead buffer was resuspended in 730µL of Cell buffer (Parse Biosciences cat. #WF101) from the Parse Biosciences kit (V1.3.0) with 0.5% BSA (Life Technologies cat. #15260037), and counted (1:4 dilution) with TC20 Automated Cell Counter (Bio-rad cat. #1450102). Cells were fixed using Parse Biosciences' Nuclei Fixation Kit v1 (Parse

Biosciences cat. #WF102), following the manufacturer's protocol. Between 150000 - 800000 cells per sample were incubated in fixation solution for 10 minutes on ice, followed by permeabilization for 3 minutes on ice. The reaction was quenched and nuclei were centrifuged and resuspended in 150 μL Cell Buffer. The fixed samples were counted using TC20 automated cell counter. DMSO (Parse Biosciences cat. #WF105) was added to the fixed cells and cells were slow-frozen in a Mr. Frosty (Thermo cat. #5100-0001) and stored at -80°C.

3.6 Methods

3.6.1 Datasets

One microglia single-cell RNA-seq dataset of cortex and hippocampus at 8 months on 5xFAD mouse model[99] and matching wild type (C57BL/6J) were used to demonstrate Topyfic behavior on microglia cells. We also combined two single-nucleus RNA-seq datasets of cortex and hippocampus from the 5xFAD mouse model of AD in two different genetic backgrounds (B6J and B6CASTF1/J) from the Model-AD and ENCODE consortiums, respectively. The Model-AD snRNA-seq was performed in 8 month old 5xFAD and matching wild type (C57BL/6J) mice, and ENCODE snRNA-seq was performed in 8 month old 5xFAD/CAST and matching WT (B6CASTF1/J) hybrid mice.

3.6.2 Preprocessing scRNA-seq and snRNA-seq data

Raw fastq files were processed using Parse Bioscience's split-pipe software (v1.0.3p) to assign reads to single cells and nuclei. In order to provide sample-level fastqs to the ENCODE portal, all data was demultiplexed using the sample-level barcode (barcode 1) from the output of split-pipe to be aligned and quantified with the ENCODE uniform processing

pipeline (https://www.encodeproject.org/pipelines/ENCPL257SYI/). We use STARSolo with GeneFull–Ex50pAS settings and the GENCODE vM21 annotation to generate UMI count matrices, annotated using GENCODE vM21. We removed low-quality cells using a UMI cutoff of 500 based on our knee plots (Fig. S2A, Fig. 2B), then performed Scrublet[258] doublet detection. Cells and nuclei < 500 or $> 30,000$ UMIs, more than 500 genes, and a doublet score > 0.2 were removed in downstream analysis. In addition, nuclei were required to have a mitochondrial gene expression score of $< 0.5\%$, while cells had a more lenient threshold of $< 5\%$. Seurat V4[146] was used to perform normalization, UMAP dimensionality reduction, and clustering. Each dataset (Model-AD 5xFAD and B6J WT snRNA-seq, ENCODE 5xFAD/CAST snRNA-seq, and Model-AD scRNA-seq microglia) were preprocessed and clustered separately, with 50, 40, and 15 clusters, respectively, after removal of 2 doublet-high clusters from the Model-AD snRNA-seq dataset and 3 doublet-high clusters from the ENCODE dataset.

To facilitate cell type annotation, a downsampled version of the 1M whole cortex and hippocampus 10x atlas from 8 week old mice available on the Allen data portal was used to transfer subtype-level annotations using "FindTransferAnchors" in Seurat. Each cluster was then manually annotated using the resulting Allen atlas labels and marker gene expression. Overall, this process identified 13 major cell types and 34 subtypes in the snRNA-seq data, and 5 celltypes (75% of which are microglia) in the scRNA-seq data. After defining annotations for each dataset, we extracted the raw gene count matrices along with gene and sample information from Seurat objects and embedded all information into the Anndata[167] file format where gene information is 'var' and cell information is 'obs.' Then we applied depth normalization[272] individually to each dataset to prepare them for the rest of the analysis (Fig1A).

3.6.3 Parse Biosciences Split-seq Experiments

The Model-AD and ENCODE libraries were prepared in two separate experiments using Parse Biosciences' Evercode WT Kit v1 (cat. #EC-W01030), one kit per experiment, following the manufacturer's protocol. Fixed samples were thawed and added to the Round 1 barcoding plate at 15,000 nuclei/cells per well when possible across 48 wells. For microglia samples with low numbers, the entire sample was added. Each tissue sample from one individual was loaded into a single well. RNA was reverse transcribed in the fixed nuclei/cells using oligodT and random hexamer primers and the first barcode was annealed. After RT, samples were pooled and randomly distributed across 96 wells of the Round 2 ligation barcoding plate for in situ barcode ligation. After Round 2, samples were pooled and randomly redistributed into 96 wells of the Round 3 ligation barcoding plate for ligation of the third cell barcode and Illumina adapters. Finally, samples were counted with a hemocytometer and distributed into 6 subpools of 15,000 nuclei for a target of around 75,000 nominal nuclei/cells per tissue. To prepare libraries, the nuclei/cells in each subpool were lysed and the barcoded cDNA was amplified. The cDNA was purified with AMPure XP beads (Beckman Coulter cat. #A63881) and quality checked with the Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) and Bioanalyzer 2100 (Agilent cat. #G2939A) High Sensitivity DNA Kit (Agilent cat. #5067-4626). Subpool cDNA (100 ng) was fragmented and Illumina P5/P7 adapters were ligated during the last amplification, followed by size selection and quality check with the Bioanalyzer and Qubit. An Illumina NextSeq 2000 and P3 200 cycles kits (Illumina cat. #20040560) were used to sequence libraries with 5% PhiX spike-in with as paired-end, single-index reads (115/86/6/0) to an average depth of 187 M reads per Model-AD library, and 183 M reads per ENCODE library.

3.6.4 Isolation of RNA for bulk assays

For ENCODE3 and ENCODE 4 bulk RNA-seq and microRNA-seq experiments in mouse tissues, total RNA was extracted from flash-frozen mouse tissues at Caltech using the Norgen Animal Tissue RNA Purification Kit (Norgen Biotek cat. #25700). Briefly, the tissue was lysed using Buffer RL, followed by protein removal with proteinase K. DNaseI treatment on the column removed genomic DNA contamination. The resulting purified total RNA encompasses a broad spectrum of RNA sizes, including large mRNAs, lncRNAs, and microRNAs. To assess RNA concentrations, the Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) was used, while RIN values were determined using the Bioanalyzer Pico RNA kit (Agilent cat. #5067-1513).

3.6.5 Bulk RNA-seq from mouse tissues

Each cDNA library was built from 300 ng total RNA with ERCC spike-ins (Thermo cat. #4456740) using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB cat. #E7760), specifically the protocol for use with NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB cat. #E7490). Ribosomal RNA was depleted from total input RNA using the NEBNext rRNA Depletion Kit (NEB cat. #E6310). Briefly, rRNA-depleted RNA was carried through first and second strand synthesis, cDNA end prep, adapter ligation, and finally PCR amplification of the resulting libraries. The bulk RNA-seq libraries were quantified using the Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) and sequenced on an Illumina HiSeq 2500 as 100 bp single-end reads to 50 M raw read depth per library. Submission to the ENCODE portal required at least 30 M aligned reads and Spearman replicate correlation > 0.9.

3.6.6 Bulk microRNA-seq from mouse tissues

MicroRNA-seq libraries were built from 400 ng of total RNA. Adapters were ligated to the 5' and 3' ends of small RNA using its 5' phosphate and 3' hydroxyl groups, then the ligation product was reverse transcribed using SuperScript II Reverse Transcriptase (Invitrogen cat. #18064-071). The 5' adapter adds a 6-nucleotide barcode from a one of 7 sets of 4 distinct barcodes used in downstream demultiplexing. The cDNA was amplified using Phusion PCR master mix (NEB cat. #M0531S) with 58 bp reverse and 55 bp forward primers containing additional 6-nucleotide barcodes added to the 3' end. The 140 bp product containing mature microRNA (21-25 nucleotides) was size-selected using 10% TBE-Urea gel (BioRad cat. #456-6033). Libraries were isolated from the gel by agitated incubation at 70°C, 1000 RPM for 2 hours in a buffer containing 0.5 M ammonium acetate (Ambion cat. #AM9070G), 0.1% SDS (Sigma cat. #L6026-50G), and 0.1 uM EDTA (Ambion cat. #AM9261), then precipitated overnight in 50% isopropanol. Resulting microRNA-seq libraries were quantified using the Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) and sequenced on an Illumina NextSeq 2000 with P2 100 cycle kits (Illumina cat. #20046811) as 75 bp single-end reads to around 10 M raw read depth per library. Submission to the ENCODE portal required >5M aligned reads, >300 microRNAs detected at >2 CPM, and a Spearman replicate correlation > 0.85.

3.6.7 Bulk read mapping and quantification

Bulk RNA-seq data from ENCODE3 and ENCODE4 mouse tissues were processed through ENCODE uniform processing pipelines using the mm10 genome with GENCODE vM21 annotations. The bulk RNA-seq data were aligned using STAR v. 2.5.1b[273] and quantified using RSEM, which provides FPKM, TPM, and raw counts (https://www.encodeproject.org/pipelines/ENCP
After trimming adapters with cutadapt v. 3.4[274], microRNA-seq data were aligned using STAR v. 2.5.1b[273] to generate raw counts for 2,207 mature microRNAs included in GEN-

CODE vM21 (https://www.encodeproject.org/pipelines/ENCPL280YDY/).

3.6.8 Bulk RNA-seq and microRNA-seq integrated analysis

Normalized counts were concatenated using the TPM column across all bulk RNA-seq gene quantifications. Raw, unstranded counts were concatenated across all microRNA-seq quantifications, then converted to CPM. Of the bulk RNA-seq and microRNA-seq datasets, 289 were built from matching RNA biosamples across a wide variety of postnatal mouse tissues from this study and prenatal mouse tissues from ENCODE3. There were 151 datasets in common using total RNA-seq, and 138 in common using polyA plus RNA-seq. MicroRNA host genes were determined by intersecting their coordinates with gene coordinates in Gencode vM21. Of the 2,207 unique microRNA coordinates, 1,180 overlapped 980 host genes, while 1,027 did not. Spearman correlations were calculated between microRNA expression in CPM and their corresponding host gene expression in TPM across all 289 samples. Around 8%, or 170 Gencode vM21 microRNAs correlated with 174 unique host genes by a Spearman correlation >= 0.3, and another 9 are annotated microRNA host genes (e.g. *Mir133a-1hg*, *Mir124a-1hg*) and were included in the regulatory gene set used for topics modeling.

3.6.9 Selection of regulatory genes

Regulatory genes were determined by microRNA-host gene correlations, annotated transcription factors, and genes annotated with Gene Ontology (GO) terms based on their impact on transcriptional and chromatin regulation. GO term-derived genes were collapsed into 5 categories: histone modifiers, from GO terms related to histone acetyltransferases (GO: 0004402), histone deacetylases (GO: 0004407), histone methyltransferases (GO: 0042054), and histone demethylases (GO: 0032452); TBP-associated factors and members of the Mediator complex (TAF-MED, GO: 0016592 and GO: 0006352), chromatin binding (GO: 0003682), chromatin

organizers (GO: 0006325 and GO: 0030527), and transcription regulators (GO: 0140110). The final list of 2,701 expressed regulatory genes has 7 biotype categories including microRNA host genes and transcription factors.

3.6.10 Input data to Topyfic

Topyfic accepts input in the form of a preprocessed expression matrix embedded within the Anndata format[167]. This format contains gene information as 'var' and cell information as 'obs.' Users can generate this input format from the output of popular single-cell tools like Scanpy[142] or Seurat[146]. For topics modeling using the regulatory gene set, the expression matrix was subset for the genes of interest, then normalized and formatted. It's important to note that Topyfic leaves the choice of performing normalizations to the user's discretion. However, it is strongly recommended, especially when dealing with data originating from different technologies. Without normalization, there is a heightened risk of detecting topics influenced by batch differences, even when they don't represent meaningful biological signals. To mitigate this issue, we apply depth normalization[272] individually to each dataset. This approach effectively implements depth normalization and variance stabilization, enabling the accurate identification of recurring patterns while minimizing the impact of technical variations.

3.6.11 Topic modeling using Latent Dirichlet Allocation (LDA)

Topic modeling is a type of statistical model that uses unsupervised machine learning to identify groups of similar words in each document. LDA is a generative probabilistic Bayesian model that operates on the assumption that documents can be represented as random mixtures over latent topics, where each topic is characterized by a distribution over a set word vocabulary. In simpler terms, each document is a mixture of topics, and each topic is a

mixture of words, where words can be repeated in different topics with different weights. In the context of single cell/nucleus RNA-seq data, cells correspond to documents, genes to words, and counts are equivalent to word frequencies. We hypothesize that there are recurring latent patterns or "topics" in count data such as large gene expression matrices. Topics are composed of genes with distinct weights that can together recreate underlying patterns of gene expression profiles for each individual cell.

3.6.12 LDA Model Training

We employed scikit-learn's LDA implementation (v1.3)[275] with options to allow users to change default parameters including batch size and learning method based on their data (default: learning_method=online variational Bayes method, batch_size=1000, max_iter=10). Due to the random initialization of LDA algorithms, topic definitions can vary substantially each time that the algorithm is rerun, which hinders their interpretability. Therefore, we train the LDA model with several distinct random seeds (default 100 times) to capture all possible topics. After training all LDA models, we built our gene-topic weight matrix using all the obtained models. Even though learning each LDA model is not overly time-consuming, learning it 100 times can be time-intensive and will increase by increasing the number of input data(cells/nuclei). To reduce run time, we have added another feature to train each LDA model separately and then combine all of them to make the final LDA train object.

3.6.13 TopModel Construction

We assume topics that are independent of the random seeds should have a similar gene weight profile. Leveraging this hypothesis, we employ the Leiden algorithm to cluster all topics with similar gene weight profiles. In cases where batch effects may be present, we

incorporate Harmony[276] for batch correction, ensuring that the topics remain consistent across different datasets. Once the clusters are defined, Topyfic calculates the topic centroid (mean of gene weights) for each cluster to create a new gene weight matrix. Then it will trim the matrix by calculating 90% of the cumulative sum of gene weight and reassign the rest to a pseudocount (1 divided by the total number of topics), creating a new reproducible LDA model (TopModel). To enhance the quality of topics, we implemented a filtering step that discards small clusters of topics based on cell-topic participation, with the default threshold set at less than 1% of cells. If any topic is eliminated in this step, Topyfic will redo the trimming and reassign the rest to the new pseudocount. This filtering step aids in eliminating topics that may have emerged due to random seed fluctuations, focusing our analysis on the more stable and biologically meaningful topics.

3.6.14 Topic object

A topic is essentially a collection of genes, each assigned a weight denoting its contribution to that specific topic. It's important to note that a single gene can appear in multiple topics with a different weights. Topyfic offers different functional enrichment analyses for each topic, enhancing its utility and our understanding of each topic. These analyses encompass Gene Ontology (GO) analysis, Gene Set Enrichment Analysis (GSEA)[277], and pathway analysis using the REACTOME database[278]. Furthermore, Topyfic supports the comparison of the two topics. This is achieved by transforming the data onto two scales: M (log ratio) and A (mean average). These scales facilitate the calculation of a modified z-score based on the M value, allowing for meaningful comparisons between topics in terms of their gene weights.

3.6.15 Analysis object

The Analysis object is a pivotal component in the post-processing phase of Topyfic, aiding in biological interpretation of the topics and data following training of the TopModel. After successfully training of the TopModel, analysis of the TopModel itself commences. This analysis includes the calculation of 'cell-topic participation,' which quantifies the extent to which each topic contributes to each cell. In essence, it represents probability distributions for each row, ensuring that the sum of topic participation for each cell equals one. To facilitate a comprehensive understanding of the topics and their relationships, Topyfic offers several visualization tools.

Using grade-of-membership models helps us estimate membership proportions for each cell/nucleus in each topic, visualized as a structure plot[279,280]. The structure plot displays the estimated membership proportions of each cell/nucleus as a stacked bar plot, with different colors representing different reproducible topics. To enhance the visualization of inferred cellular programs from the data, cells are sorted within selected traits in a given order. Within each trait group, cells are further ordered based on their similarity in estimated membership proportions, employing Ward's linkage[281]. A pie chart summarizes the structure plot, providing a representation of the overall the contribution of each topic of all the cells/nuclei in a trait group. A topic-trait relationship heatmap visualizes the Spearman correlation coefficients between each trait and topic as a heatmap. These visualizations serve as valuable aids in interpreting the topics and their associations with other relevant traits or characteristics.

3.6.16 Comparing Topics

An advantage of employing LDA to discover topics lies in the representation of gene weights as probability distributions within each topic, ensuring that the sum of gene weights in any topic equals one. Leveraging this property, Topyfic offers a valuable feature for compar-

ing topics based on their gene membership profiles. To facilitate this comparison, Topyfic normalizes the gene weights and assesses the similarity between any pair of topics in the gene membership space. This similarity evaluation can be performed using various methods, including Pearson correlation, Spearman correlation, cosine similarity and information-theoretic metrics like the Jensen–Shannon divergence. Once these comparisons are completed, the results can be visualized as a graph or heatmap through Topyfic. This visualization allows for a clearer understanding of the relationships between different topics based on their gene memberships.

3.6.17 LDA parameter settings

In addition to determining the final number of topics, other parameters may require tuning based on the input data. We suggest using 'online' as a learning method, which uses an online variational Bayes method. This method updates the gene weight in each topic during each EM update using a mini-batch of training data. We can also tune learning_decay which controls learning rate in the online learning method. Besides these, other possible search parameters could be batch_size (number of cells to use in each EM iteration; default=1000) and max_iter (the maximum number of passes over the training data, aka epochs; default=10). Given enough computing resources, it might be worthwhile to experiment with these parameters.

3.6.18 Pseudobulk calculation

A pseudobulk sample is formed by aggregating the expression values that pass QC from groups of nuclei originating from the same individual, which represents the experimental unit of replication.

3.6.19 Principal component analysis (PCA)

Principal component analysis was performed through scikit-learn on the pseudobulk matrix, where 27 components explained the 95% of the variance in the single nucleus data.

3.6.20 Topyfic analysis of single-nucleus RNA-seq data

The normalized gene-count matrix was divided based on samples with and without the 5xFAD transgene for training the TopModel through Topyfic. Initial TopModel training was performed on the first replicate of the 5xFAD and WT samples separately, employing default parameters except minimum cell participation which was 0.5% of the total number of nuclei (120.7 for the 5xFAD dataset and 142.725 for the WT dataset) using different numbers of topics (K) ranging from 5 until 50. K=15 was chosen for further analysis. Subsequently, Topyfic was used to combine TopModels and remove topics with cell participation lower than 0.5% of the total number of nuclei in both datasets (321.11) to obtain the final reproducible topics. To demonstrate that the TopModel learned meaningful topics that could be used to analyze other related datasets, we applied the trained TopModel to the second replicate of the single-nucleus data.

3.6.21 Topyfic analysis of single-cell RNA-seq data

A processed gene-count matrix containing only microglia cells was passed as input to Topyfic. The TopModel was trained with all default parameters, except for the minimum cell participation which was set to 0.5% of the total number of nuclei (27.73) across different numbers of topics (K).

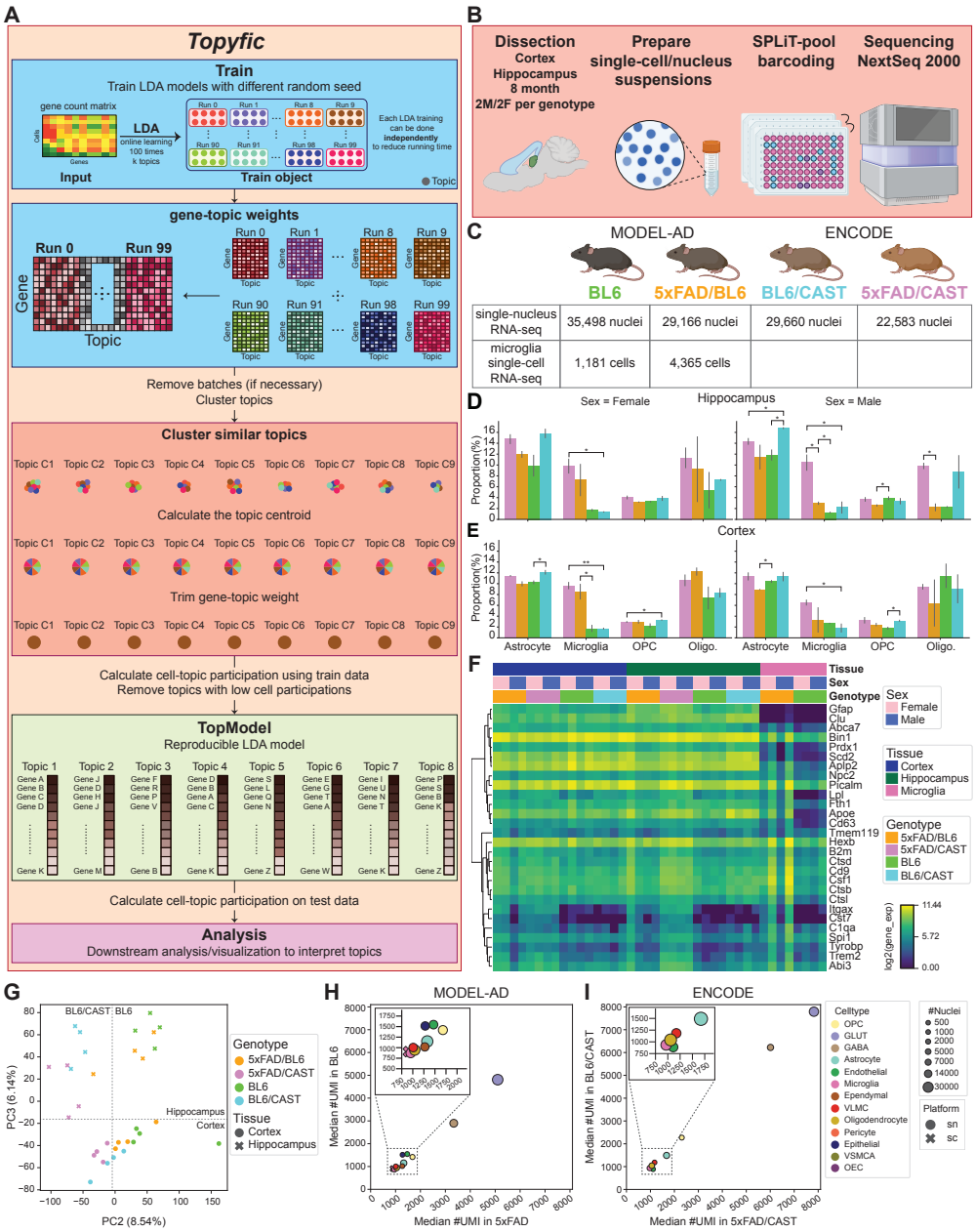

Figure 3.1: **Overview of Topyfic and datasets. A.** Overview of Topyfic workflow, as described in the text. **B.** Diagram of experimental design of single cell and single nucleus RNA-seq using Split-seq. **C.** Number of recovered nuclei/cells from each genotype after filtering. **D-E.** Proportion of glial cells recovered from a single nucleus dataset in each genotype in **D.** the hippocampus and **E.** cortex. **F.** Hierarchical clustering of gene expression markers for microglia, astrocytes, and Alzheimer's disease (AD) marker genes in each pseudobulked sample. **G.** PCA plot of pseudobulked samples. **H-I.** Comparison of the median of UMI counts in cell types in AD mice vs. WT **H.** in MODEL-AD and **I.** in ENCODE datasets. Dot size reflects the number of nuclei in each cell type.

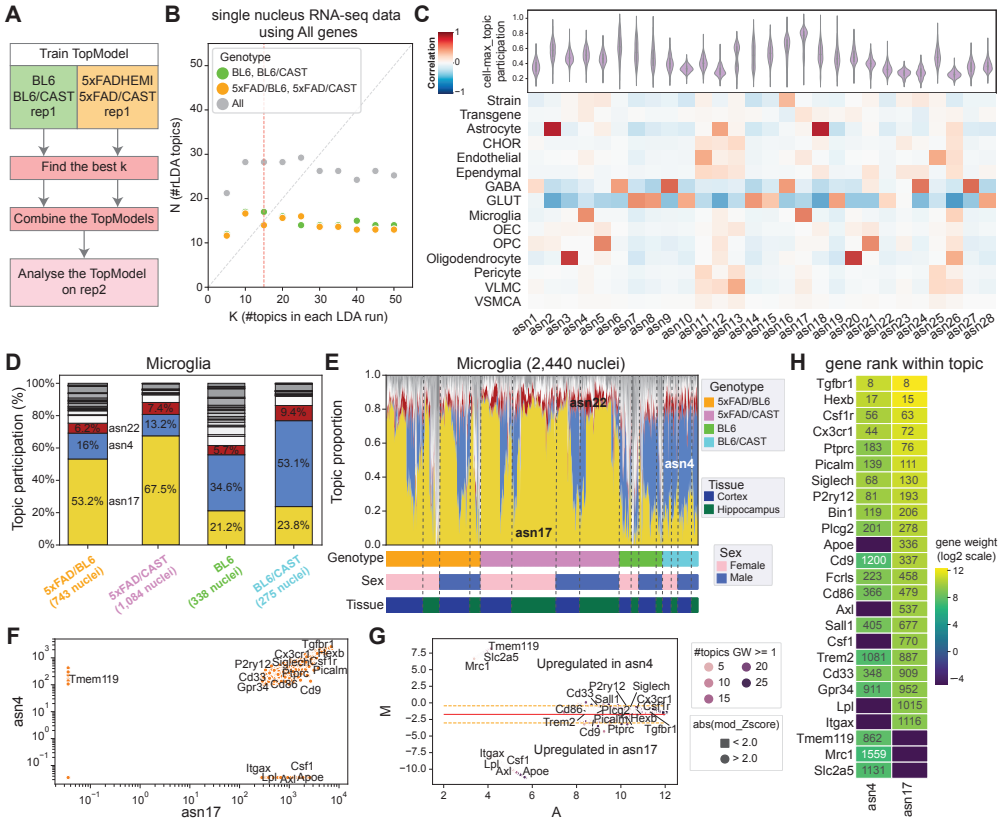

Figure 3.2: **Topic modeling in single nuclei from 5xFAD/BL6 and 5xFAD/CAST cortex and hippocampus. A.** Topics were called in 5xFAD transgenic mice and control mice separately using the first biological replicate for each mouse pair. After finding the best k to describe each training set separately, resulting topics were combined into a single TopModel, which was applied to the second technical replicate. **B.** The number of starting topics (K) versus the number of final topics (N) on the separate runs, and the combined models. Choosing K=15 for the individual runs led to a final set of 28 combined topics after filtering. **C.** Topic-trait relationship of the Spearman correlation between traits such as major cell type, strain, and transgene, across all topics. "Strain" indicates the background of mice, either BL6 (red) or BL6/CAST (blue). "Transgene" shows if the mice have the 5xFAD transgene (red) or not (blue). The violin plots show the distribution of cell topic participation whenever the topic is top ranked topic in a cell. **D.** Overall topic participation in microglia nuclei by genotype. Three major microglia topics were annotated. **E.** Structure plot of topic participation for each nucleus sorted by hierarchical clustering in each group (same genotype, sex, and tissue). **F.** Gene weights of the main two microglia topics asn4 and asn17 in log-log plot. **G.** MA plot comparing asn4 and asn17, where the X-axis (A) represents the average weight of the gene between both topics in the comparison, and Y-axis (M) represents log base 2 of the fold change of gene weight between topics. The color of each dot shows the number of topics (out of 28) where each gene has a weight above one. Modified z-score also indicated genes that were significantly differentially weighted between both topics in the comparison. **H.** Gene weights and the rank of each gene within the topic shown for asn4 and asn17. Color represents the weights of genes in the log2 scale.

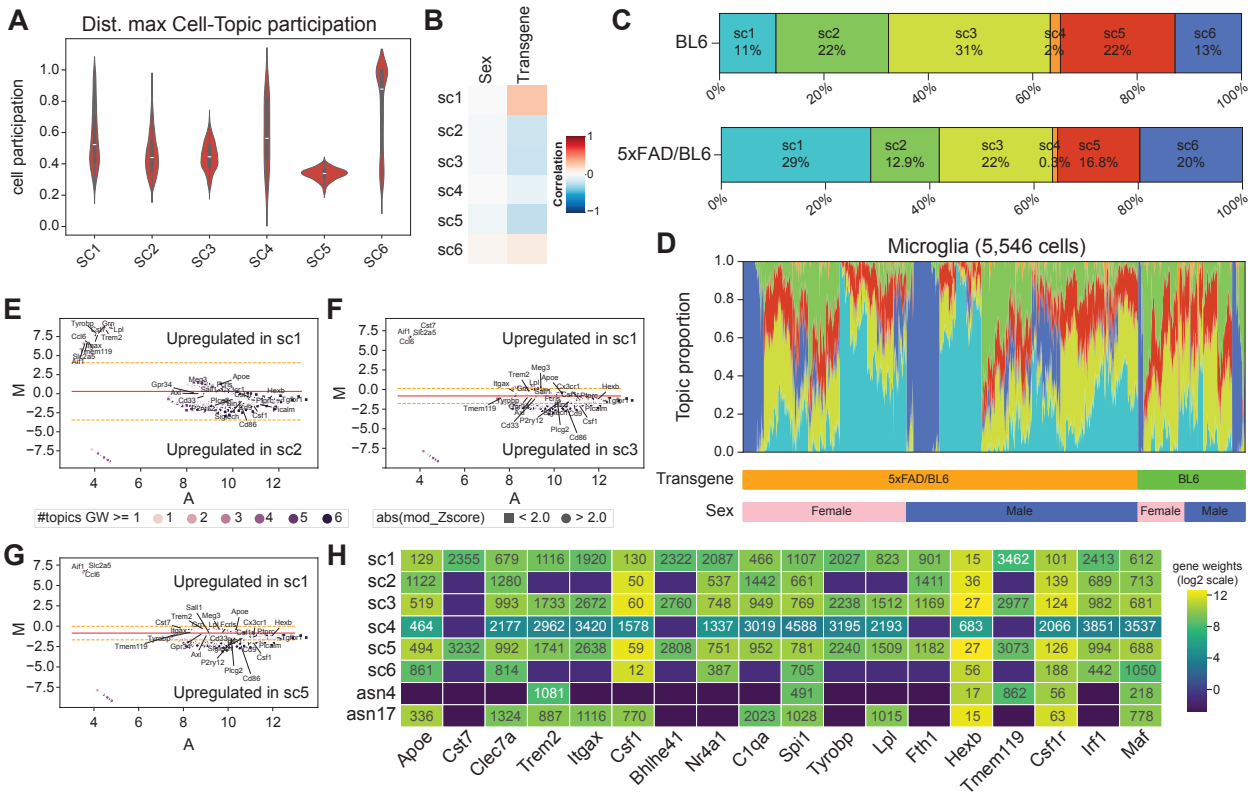

Figure 3.3: **Topic modeling in scRNA-seq of microglia. A.** Distribution of maximum cell-topic participation in each cell in each topic. **B.** Topic-trait correlation between sex or transgene and each topic. **C.** Topic participation broken down by genotype. **D.** Structure plot of microglia cells sorted by genotype and sex show topic participation in each cell. **E-G.** Comparison of gene weights for sc1 (activated microglia topic) **E.** versus sc2, **F.** versus sc3, and **G.** versus sc5. Color represents the weights of genes in the log2 scale. **H.** Gene weights of genes in interest in log2 scale with their rank.

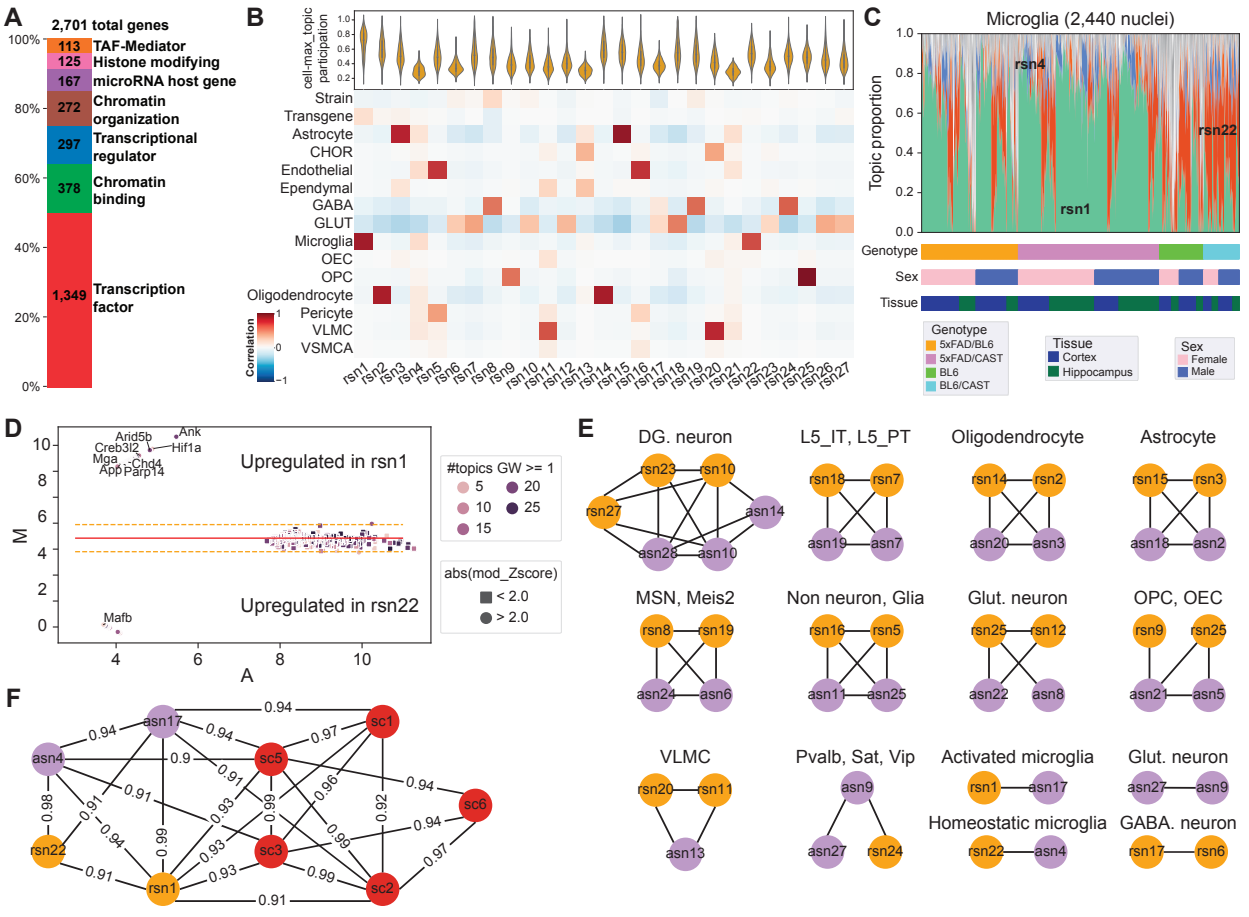

Figure 3.4: **Topic modeling using regulatory genes. A.** Breakdown of regulatory gene categories in mouse. **B.** Distribution of cell-topic participation in single-nucleus datasets. Topics rsn1 and rsn22 are enriched in microglia, with rsn1 also enriched in transgenic mice. **C.** Structure plot for microglia nuclei sorted by genotype and sex. rsn1 mostly contributes to the AD mice, whereas rsn22 is primarily found in WT mice. **D.** MA plot comparison of gene weights between rsn1 and rsn22. The color of the dots represents the number of topics with the gene weight higher than the pseudocount, and the style of each dot indicates if the difference between gene weights is significant (circle) or not (square) based on a modified z-score of M values. **E.** Topic cosine similarity > 0.9 between topic pairs using all genes and regulatory topics. **F.** Topic cosine similarity > 0.9 between microglia topics from single nucleus (asn), regulatory genes (rsn) and microglia single-cell (sc).

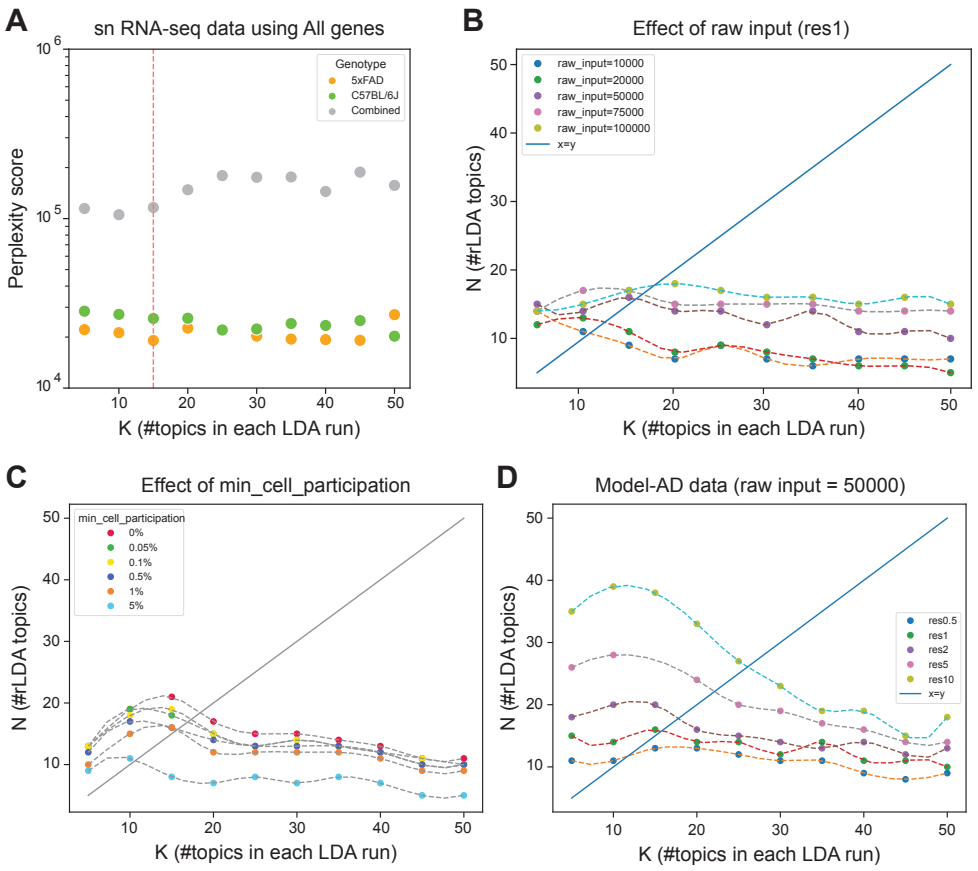

Figure 3.5: **Impact of various data and Topyfic parameters on the number of topics(K) A.** The perplexity score is based on the number of starting topics (K) on the separate runs, and the combined models. Choosing K=15 for the individual runs to have lower the perplexity score. **B.** Influence of the number of nuclei/cells on the recovered topics (K). Higher number of nuclei/cells results in a greater number of topics (K). **C.** Impact of increasing minimum cell-topic participation on the removal of smaller topics, leading to a reduction in the number of topics (K). **D.** The effect of increasing resolution on generating more clusters, subsequently increasing the number of topics (K).

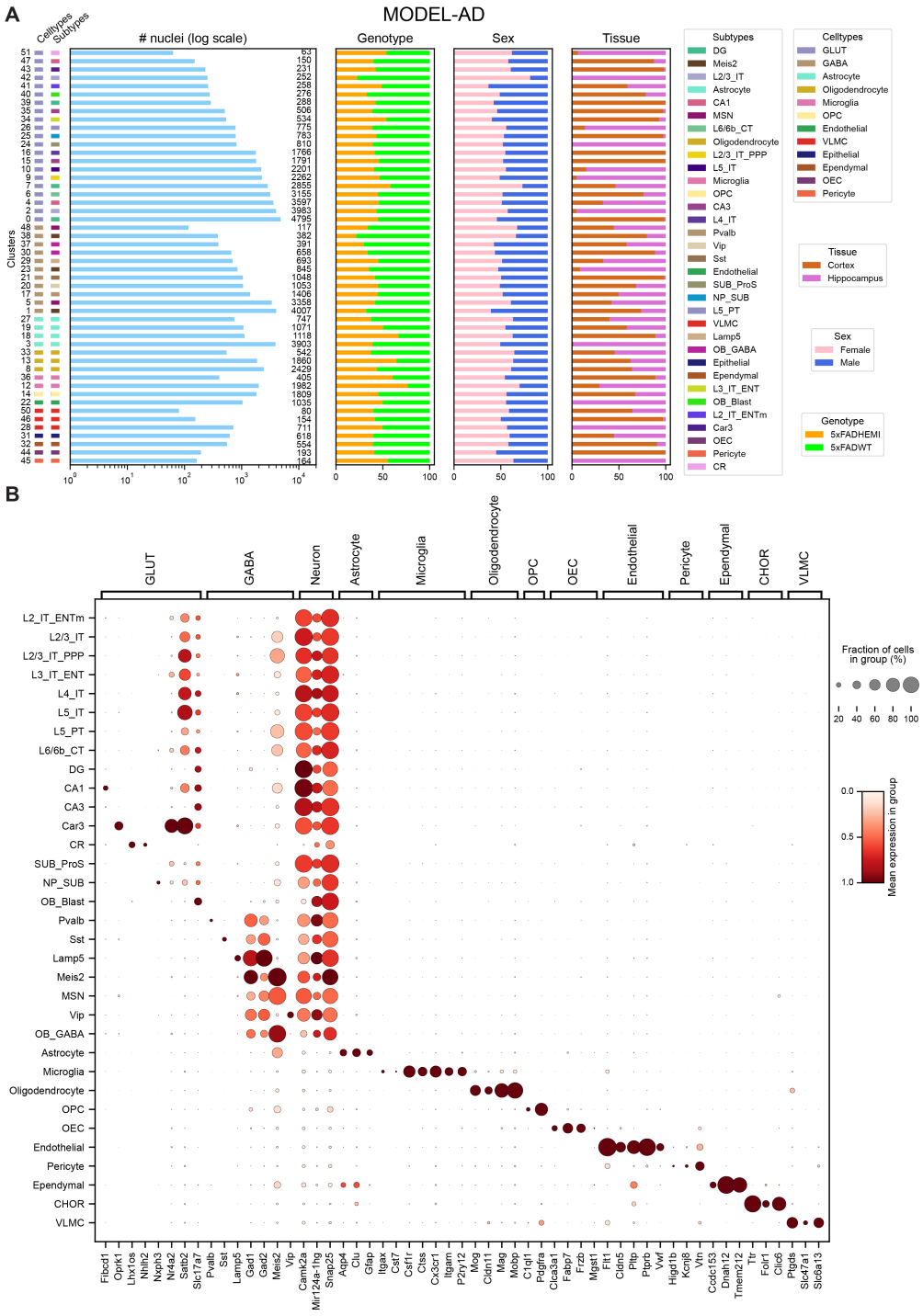

Figure 3.6: **Overview of MODEL-AD dataset. A.** Breakdown of nuclei, cell type, subtype, genotype, sex, and tissue across 52 clusters. **B.** Expression of marker genes.

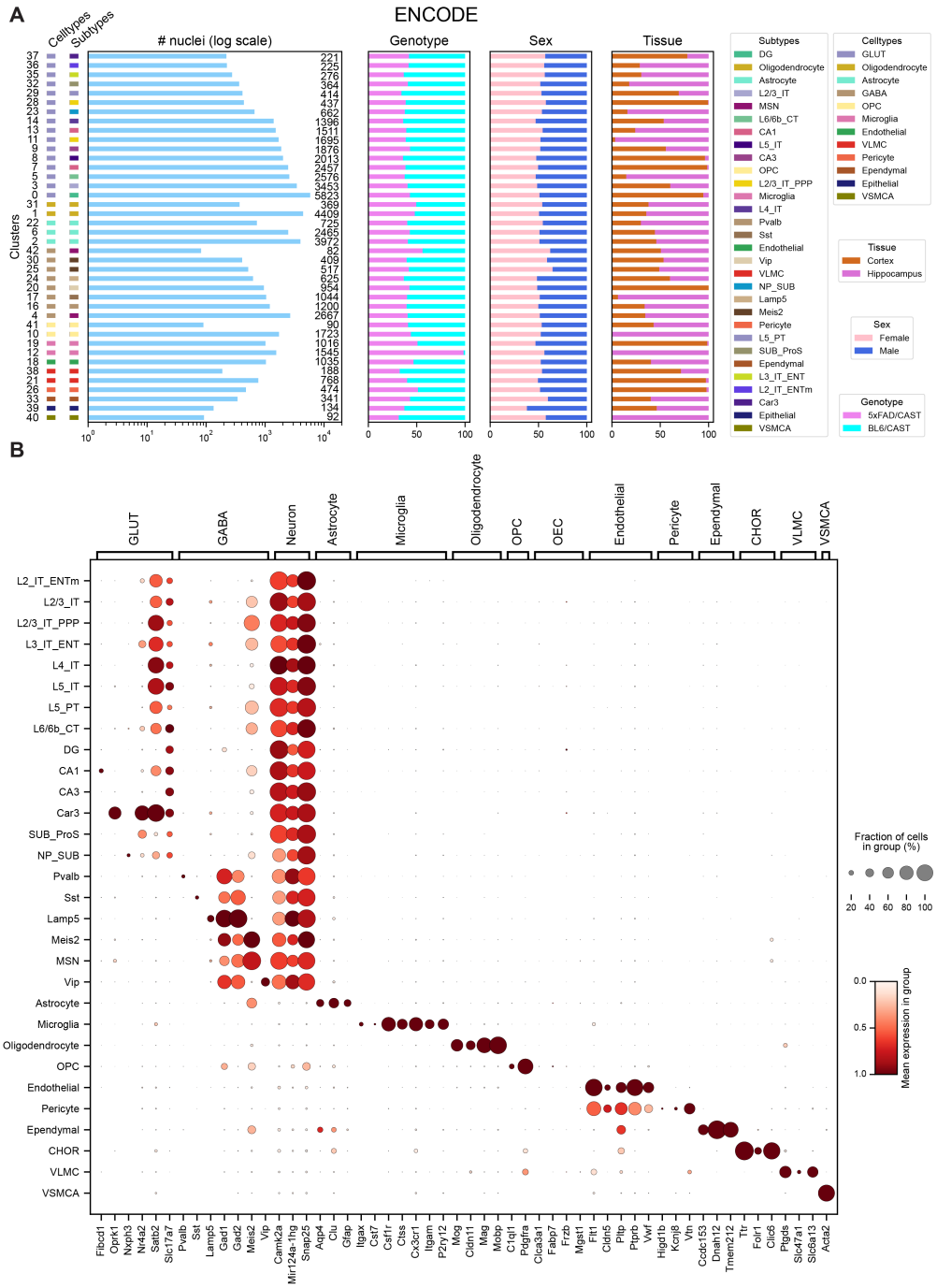

Figure 3.7: **Overview of ENCODE dataset. A.** Breakdown of nuclei, cell type, subtype, genotype, sex, and tissue across 43 clusters. **B.** Expression of marker genes.

Chapter 4

Unraveling gene expression dynamics in mouse models through PyWGCNA and Topyfic integration

4.1 Abstract

RNA sequencing (RNA-seq) has emerged as a pivotal tool for profiling transcriptomic variations in diverse conditions, including disease states. While conventional bulk RNA-seq provides an overview of average gene expression across a population of cells within a tissue, single-cell RNA sequencing (scRNA-seq) allows for a granular examination of transcriptomic profiles at the individual cell level. The Weighted Gene Co-expression Network Analysis (WGCNA) applied to bulk RNA-seq facilitates the identification of co-expressed gene modules. Conversely, the Grade of Membership (Gom) models such as LDA offer an unsupervised approach for analyzing scRNA-seq data, unveiling cellular programs. A comparative analysis between these two methodologies provides valuable insights into the underlying interactions

among genes. This chapter introduces a novel approach to compare gene modules to gene topics based on their similarities. By incorporating cellular programs as gene-topic weights to modules derived from PyWGCNA using a geometric approach, we aim to compare and contrast the two techniques. To validate our methodology, we conducted a comparative analysis using bulk RNA-seq datasets from various mouse models of Alzheimer's disease (AD) obtained from the UCI MODEL-AD consortium and explored topics identified in two 5xFAD mouse models of AD, crossed with either C57BL6/J or CAST/EiJ mice (Chapter 3). This analysis sheds light on how gene expression modules and cellular programs are complimentary in studying gene expression changes in health and disease.

4.2 Introduction

Bulk RNA-seq is the most widely used for studying the transcriptional landscape and elucidating molecular pathway alterations in human cancers. However, it presents a limitation by furnishing only the average gene expression profiles across various cell clusters, failing to capture the transcriptional heterogeneity inherent in cell populations[282]. One method to analyze bulk RNA-seq data is weighted gene co-expression network analysis which can be done through PyWGCNA[283]. Genes are systematically grouped into co-expression modules by PyWGCNA based on their similarities in expression profiles across samples. The primary objective is to pinpoint sets of genes displaying coordinated expression patterns and identify hub genes associated with specific biological functions. Particularly well-suited for studies with large sample sizes, methods such as PyWGCNA offers valuable insights into global expression patterns. It distinguishes itself from methods allowing genes to belong to multiple clusters with associated weights, as PyWGCNA uniquely assigns each gene to a single module.

In contrast to bulk RNA-seq, single-cell RNA sequencing (scRNA-seq) offers high-throughput

and high-resolution transcriptome profiling at the individual cell level. However, this advancement comes at the cost of increased data noise and variability[12]. ScRNA-seq allows for a granular understanding of cellular states and functions by capturing transcripts on a per-cell basis[284]. A popular analytical approach applied to scRNA-seq data is topic modeling, a method originally derived from natural language processing and adapted for single-cell/nucleus RNA-seq studies. Methods such as Topyfic[285] leverage topic modeling to uncover hidden latent themes (topics) within the heterogeneous landscape of individual cells. By assigning weights to genes for each topic, topic modeling can identify quantitative relationships between genes the underlying biological processes contributing to observed cellular diversity. Unlike PyWGCNA, topic modeling furthermore allows each gene to be part of multiple topics with possibly distinct weights, which matches the biological pleiotropy of genes.

While topic modeling presents a flexible and unsupervised approach, yielding valuable insights into cellular functions, it should be noted that its computational demands may pose challenges. Additionally, interpreting topics in a biological context can be intricate, thereby requiring careful consideration during analysis.

This study introduces an approach aimed at elucidating the intricate relationships between co-expression network modules derived from PyWGCNA and gene weights obtained through Topyfic using a geometric framework. The overarching objective is to characterize each topic as a linear combination of modules, where each topic is a vector of genes that can be projected into a subspace of the genes within a module. This integrated approach provides a distinctive perspective on the modular organization of gene expression, unveiling the connections between co-expression modules and latent topics.

To validate our approach, we first aggregate all bulk RNA-seq datasets generated by the MODEL-AD consortium (as detailed in Chapter 2). Subsequently, we explore gene expression changes using PyWGCNA283. These modules are then compared to topics in two

distinct 5xFAD mouse models of Alzheimer's disease (AD) crossed with either C57BL6/J or CAST/EiJ mice[285]. This comprehensive analysis aims to unravel the intricate interplay between co-expression modules and latent topics, offering valuable insights into the regulatory landscape of gene expression in the context of AD.

4.3 Results

4.3.1 Analysis of Bulk RNA-seq

Expression levels of marker genes for disease-associated microglia (DAM), astrocytes, and oligodendrocytes in pseudo-bulk for each mouse showed higher expression in older mice with a 5xFAD background versus WT in both the hippocampus and cortex (Fig. 4.1). We applied PyWGCNA to 1047 bulk RNA-seq datasets and identified 11 co-expressed modules associated with age, genotype, tissue, and sex (Fig. 4.2). The dark red module, comprising 703 genes, is significantly correlated with older mice with a 5xFAD background (Fig. 4.2). This module is enriched in GO terms related to the inflammatory response, cytokine signaling, and microglia cell activation (Fig. 4.4). Additionally, it is enriched in pathways related to the immune system, such as neutrophil degranulation. The eigengene profile of this module is significantly higher in aged mice with a 5xFAD background (Fig. 4.3). The brown module is the neuronal development module which is mainly expressed in the hippocampus in younger mice. Go terms related to neuronal systems such as generation of neurons (GO:0048699) significantly (p-value ¡ 0.05) enriched in this module as well.

4.3.2 Comparison of bulk and single-nucleus RNA-seq datasets

To evaluate the similarity between bulk and single-nucleus RNA-seq datasets, we compared PyWGCNA modules and gene markers at both the cluster and cell type levels using a module overlap test in PyWGCNA. Notably, the dark red module, which is enriched in microglia cell activation, exhibits a significant overlap with cluster 12, which is identified as a microglial cluster (Fig. 4.5, Fig. 4.6)[285]. Additionally, the brown module significantly overlaps with neuronal clusters (GABA and GLUT), which matches the biological functions enriched in this module (Fig. 4.5, Fig. 4.6).

In PyWGCNA and, more broadly, WGCNA, each gene is associated with only one module, contrasting with other clustering[139,286] and GoM methods[153,287] where genes may belong to multiple clusters with associated weights. In this context, we utilized Gene Module Membership (kME) and compared them to the normalized gene-topic weights obtained from Chapter 3 (refer to the Methods section for details). Examining the distribution of cosine similarities, we define 0.12 as a significant level based on the null hypothesis (Fig. 4.7). The inflammatory darkred module exhibits a significant cosine similarity with our microglia topics (asn4 and asn17), confirming the biological functions. While the darkred module could not tell a difference between activated and homeostatic microglia, Topyfic would provide more information related to the different states of cells suggesting the major benefit of GoM methods over WGCNA analysis.

As expected, the brown module shows greater similarity to topics associated with GLUT and GABA cell types (Fig. 4.8). Topyfic analysis recovers 6 topics associated with GABA and 7 topics were significantly enriched in GLUT, while we only are able to identify one module associated with neuron cell types through PyWGCNA, suggesting gene expression variability within a cell type. This heterogeneity in gene expression is crucial for understanding the complexity of biological systems, as it underlies the diverse responses of seemingly identical

cells to external signals, contributing to the intricacies of development, tissue homeostasis, and disease.

4.4 Discussion

In this study, we employed a framework to investigate the intricacies of gene expression dynamics. PyWGCNA, utilizing a network-based approach, allowed us to categorize genes with shared expression patterns into modules while each genes can only be present in one module. The modules identified by PyWGCNA hold potential for therapeutic target exploration, and the unraveling of regulatory networks. Simultaneously, Topyfic offered a single-cell resolution perspective, helping us uncover latent biological programs within individual cells where a gene can participate in multiple topics with different weights. This approach was crucial in revealing cellular heterogeneity, distinct cell states, and developmental trajectories within complex biological systems.

Applying PyWGCNA to the MODEL-AD consortium's bulk RNA-seq datasets led to the identification of a darkred module that is enriched in genes involved in inflammatory response with an increase over time in mice with a 5xFAD background. This finding aligns with the expected dynamic nature of inflammatory processes in AD disease progression. To gain more insight into these modules, we describe them with cellular programs defined by Topyfic. This revealed that the darkred module comprised two main cell states – homeostatic and activated microglia. Topyfic's ability to identify cellular heterogeneity and explore gene expression patterns in cellular trajectories demonstrated its superiority in capturing cell type/state-specific insights compared to the broader information provided by PyWGCNA modules.

Building on these findings, our study proposes an approach to compare PyWGCNA modules with Topyfic topics. This integration aims to leverage the strengths of both methodologies,

providing a more comprehensive understanding of gene expression dynamics across different assay types. By harmonizing the insights from these two analytical approaches, our study seeks to contribute to understanding the molecular landscape within complex biological systems. This comparative analysis lays the groundwork for future investigations into disease mechanisms and therapeutic interventions.

4.5 Methods

Bulk RNA-seq datasts

As part of the MODEL-AD consortium, we sequenced bulk RNA-seq of the cortex and hippocampus of 5xFAD mice and matching wild-type (C57BL/6J) of both sexes in four different ages (4 months, 8 months, 12 months, and 18 months)[120]. We also produced a RNA-seq time course (4 months, 12 months, and 18 months) in the 3xTg-AD mouse and matching wild-type (B6129SF1/J) using the cortex and hippocampus[100,288]. We also produced mRNA profiles of eight GWAS (TREM2R47H, ABCA7v1613M, BIN1k358R, CLUh2kbKI, EPHA1P461L, PICALMH465R, SPI1rs1377416, and ABI3S209F) mouse models of AD and crossed them with 5xFAD mice as well as bulk RNA-seq data of Trem2R47H NSS mice treated with the demyelinating agent cuprizone (Table. 4.1). They also collected bulk RNA-seq data of 8-month-old WT and homozygous Abca7V1613M mice of both sexes that were injected intraperitoneally with either 0.3mg/kg LPS or 1X PBS (saline). At 6 or 24 h post administration, mice were euthanized via CO2 inhalation and transcardially perfused with ice-cold 1X PBS (Table. 4.2). In total, we are analyzing 1047 bulk RNA-seq datasets.

single-cell and single-nucleus RNA-seq dataset

All single-nucleus RNA-seq data were meticulously produced and preprocessed as outlined in Chapter 3, ensuring a standardized and comprehensive approach.

PyWGCNA analysis

Weighted Gene Correlation Network Analysis (WGCNA) was conducted using PyWGCNA [283] on a bulk RNA-seq with default parameters. Protein-coding genes with expression greater than 2 TPM in at least 2 samples were used as input for PyWGCNA. No samples were identified and removed based on hierarchical clustering (Fig. 4.9). According to our datasets, the power that resulted in a higher similarity with a scale-free network was 11 (Fig. 4.10).

For each co-expression module identified by PyWGCNA, the module eigengene was obtained as the first principal component of the standardized expression profiles of all genes within the module. PyWGCNA also calculates, for each gene in the dataset, the Pearson correlation coefficient between the expression profile of the gene and the module eigengene of each module, referred to as Gene Module Membership (kME). This value indicates the membership strength of the gene in each module. Positive kME values suggest a positive correlation between the gene's expression profile and the module eigengene, indicating membership in the module. Conversely, negative kME values imply a negative correlation, indicating an inverse relationship with the module eigengene. A kME value close to zero suggests weak or no correlation with the module eigengene.

Topyfic analysis

Topics, derived from Chapter 3 using single-nucleus RNA-seq data all genes (asn), were used for the analysis. Gene weights were normalized to ensure comparability for further analysis, with the sum of gene weights in any topic equalling one.

Geometric space integration of PyWGCNA modules and Topyfic topics

To establish a connection between PyWGCNA modules and Topyfic topics, we computed the cosine similarity values between the kME vector of genes within the module and the normalized gene weight vector of each topic. The cosine similarity between the kME vector

of genes (u) and the normalized gene weight vector (v) is defined as $\frac{u \cdot v}{||u||_2 \cdot ||v||_2}$.

The cosine similarity is a measurement to define each gene's module membership within the geometric space as a linear model of gene weight topics. Cosine similarity values serve as slopes, reflecting the alignment of a gene's expression profile with the module eigengene obtained from PyWGCNA and the gene weight vector obtained from Topyfic. A high cosine similarity indicates a strong association, underscoring the significance of the associated topic in characterizing the gene's expression pattern within the defined module.

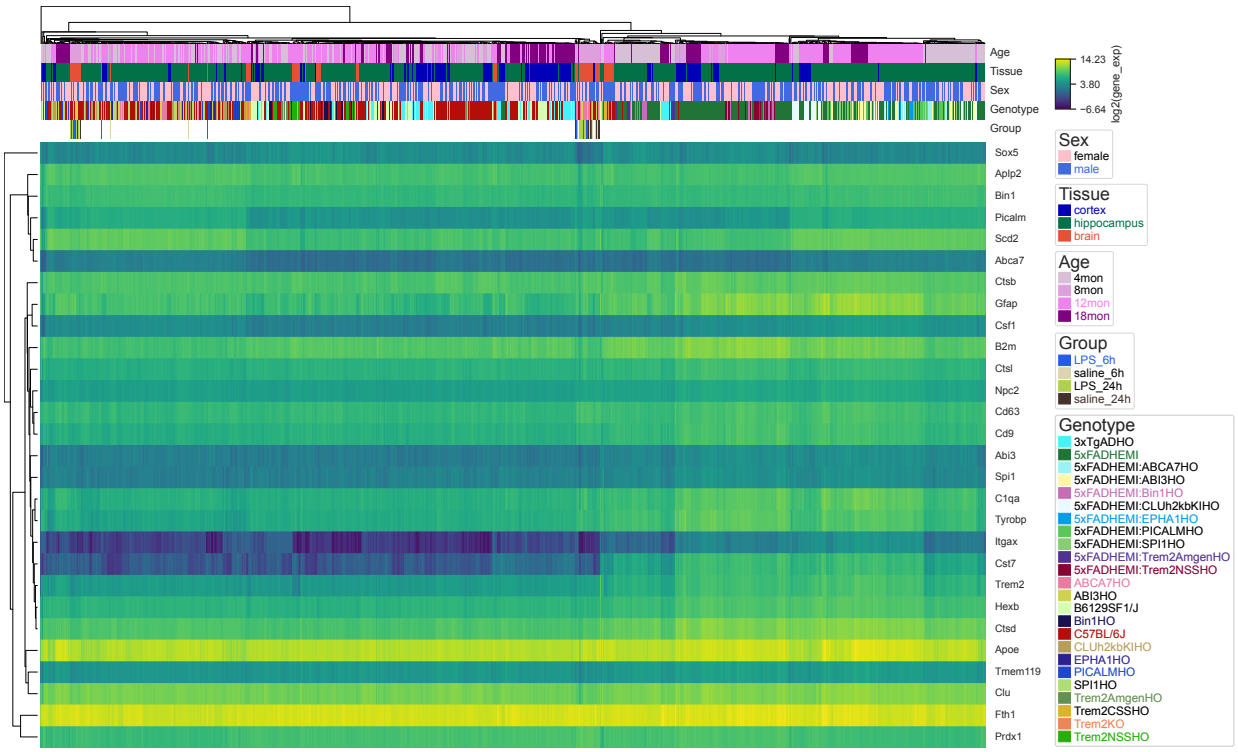

Figure 4.1: **Hierarchical clustering of gene expression markers for microglia, astrocytes, and Alzheimer's disease (AD) marker genes in each pseudobulked sample.**

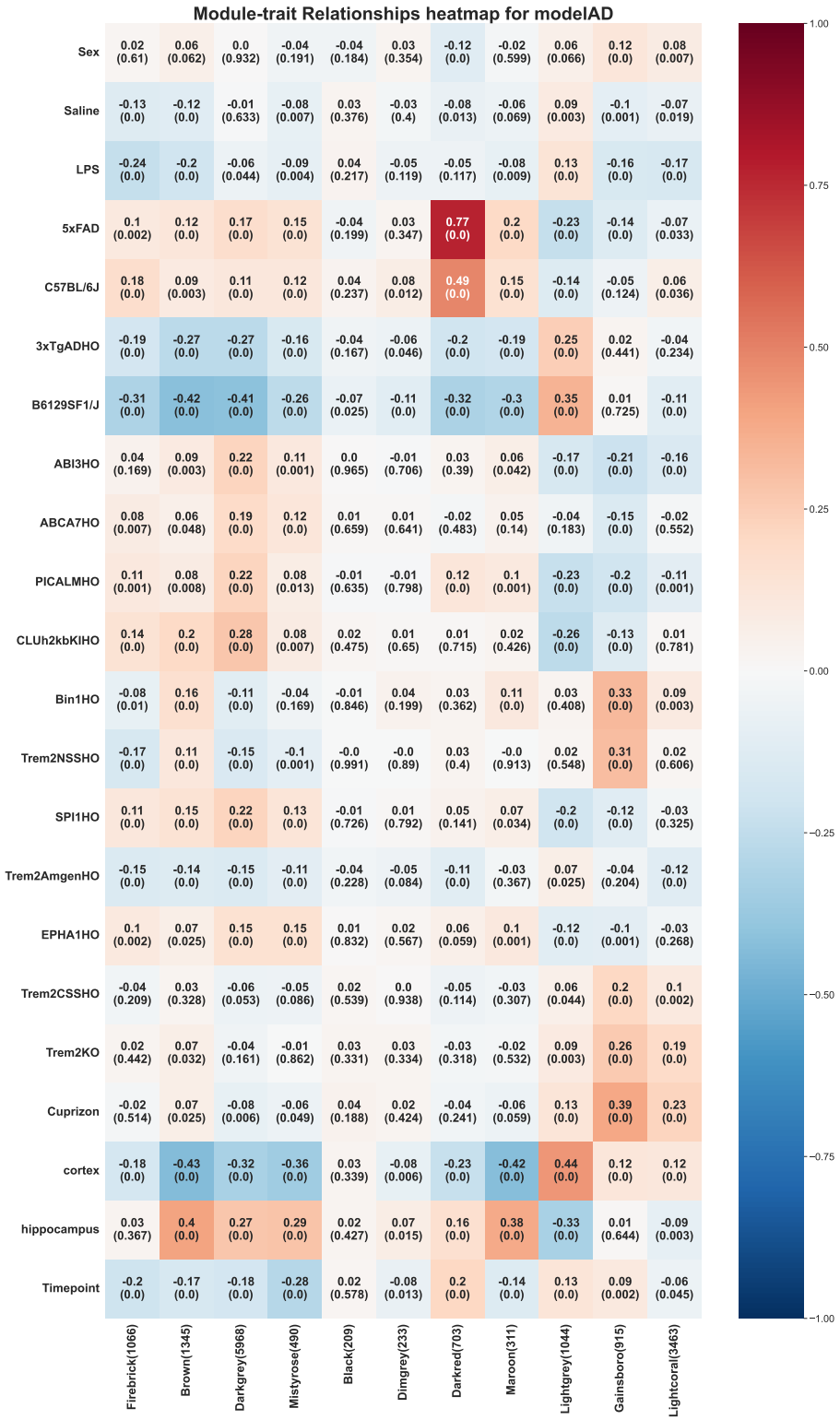

Figure 4.2: **Matrix with the Module-Trait Relationships (MTRs) heatmap** The MTRs are colored based on their correlation: red is a strong positive correlation, while blue is a strong negative correlation.

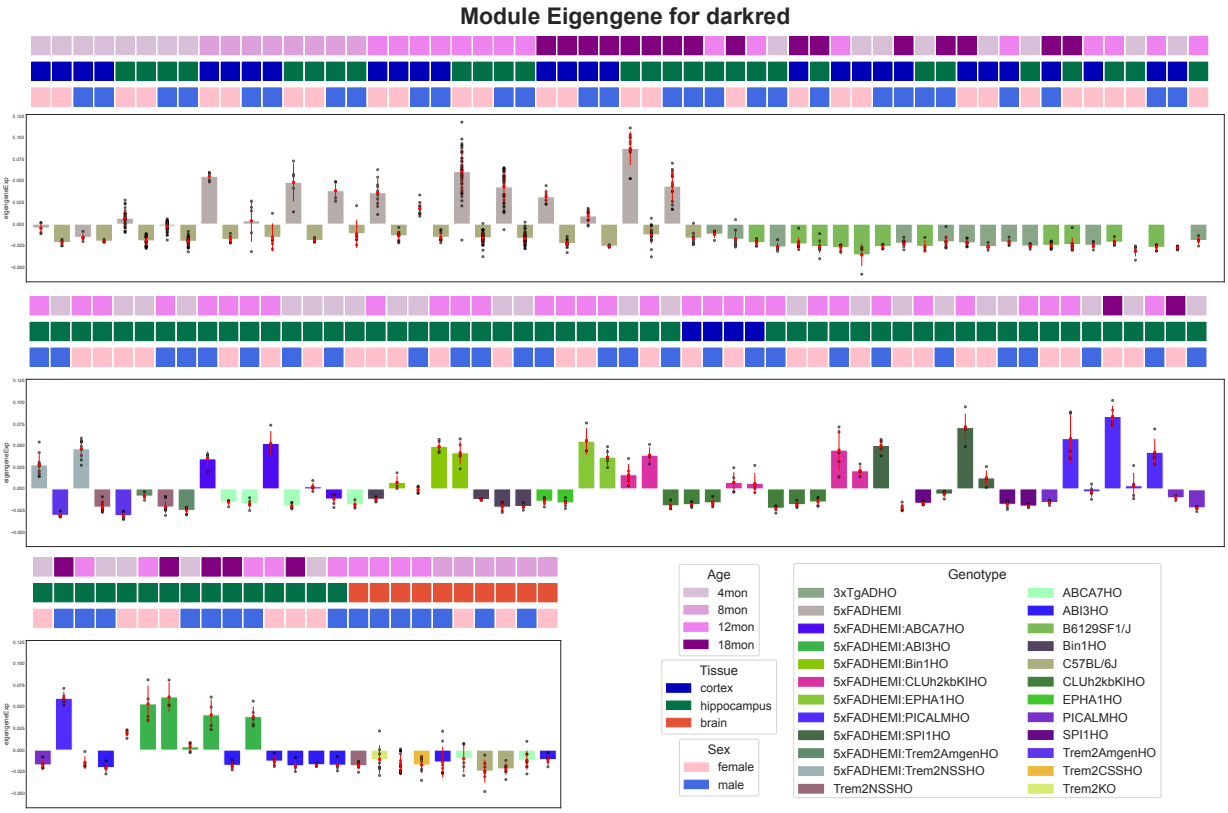

Figure 4.3: **Darkred module eigengene expression** The darkred module eigengene expression profile is summarized by genotype. Above, the top three rows display the metadata for each dataset including sex, tissue, and age. Below, the bar plot represents module eigengene expression by genotype for each dataset with individual sample module eigengene expression shown as points.

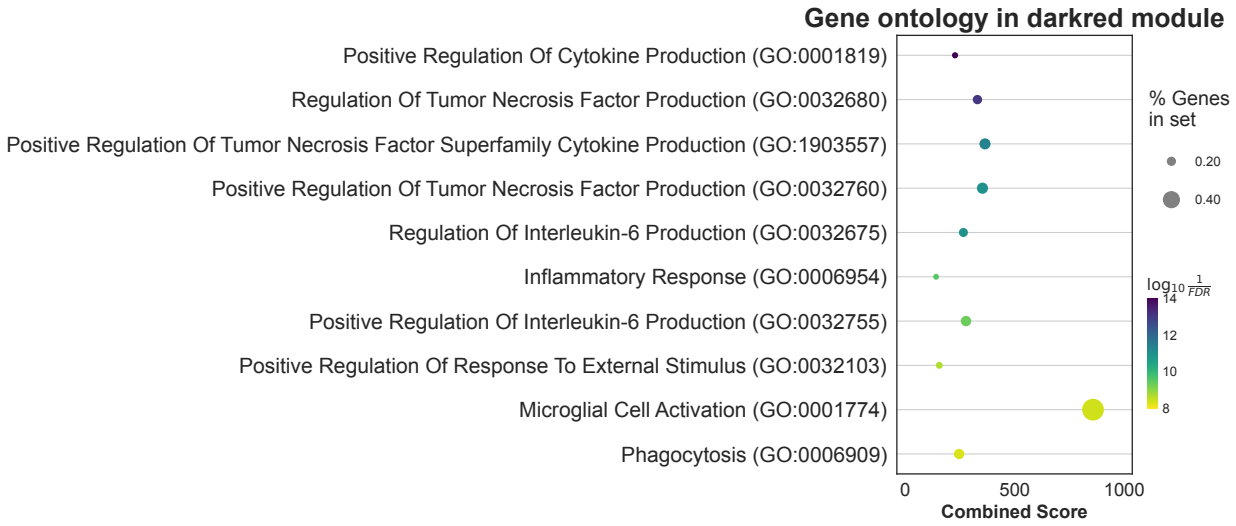

Figure 4.4: **Gene Ontology (GO) analysis of the Darkred module** GO analysis of the genes in the Darkred module.

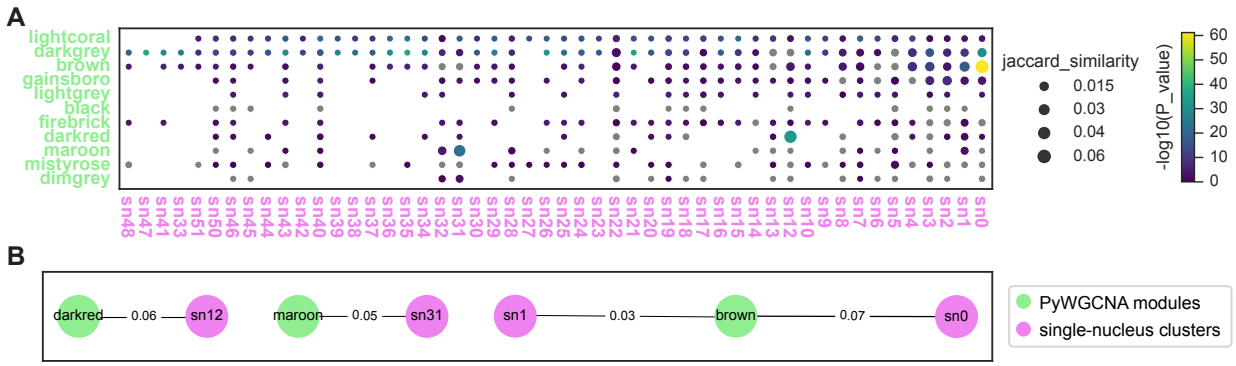

Figure 4.5: **Comparison of PyWGCNA modules and gene markers of clusters.** **A)** Bubble plot of module overlap test results between PyWGCNA modules and gene markers that were calculated for each major cell type. The dot size represents the fraction of shared genes between each pair of modules and non-gray color denotes the significance of the overlap between modules. **B)** Comparison of PyWGCNA modules and gene markers of clusters with >0.03 Jaccard similarity. The thickness of the lines shows the Jaccard index value.

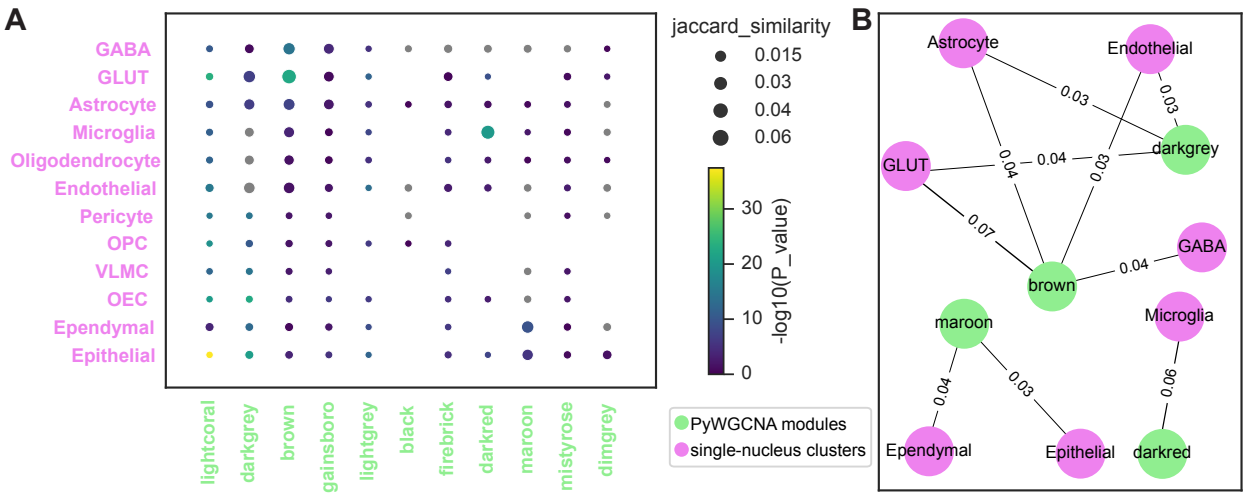

Figure 4.6: **Comparison of PyWGCNA modules and gene markers of cell types.** **A)** Bubble plot of module overlap test results between PyWGCNA modules and gene markers that were calculated for each major cell type. The dot size represents the fraction of shared genes between each pair of modules and non-gray color denotes the significance of the overlap between modules. **B)** Comparison of PyWGCNA modules and gene markers of cell types with >0.03 Jaccard similarity. The thickness of the lines shows the Jaccard index value.

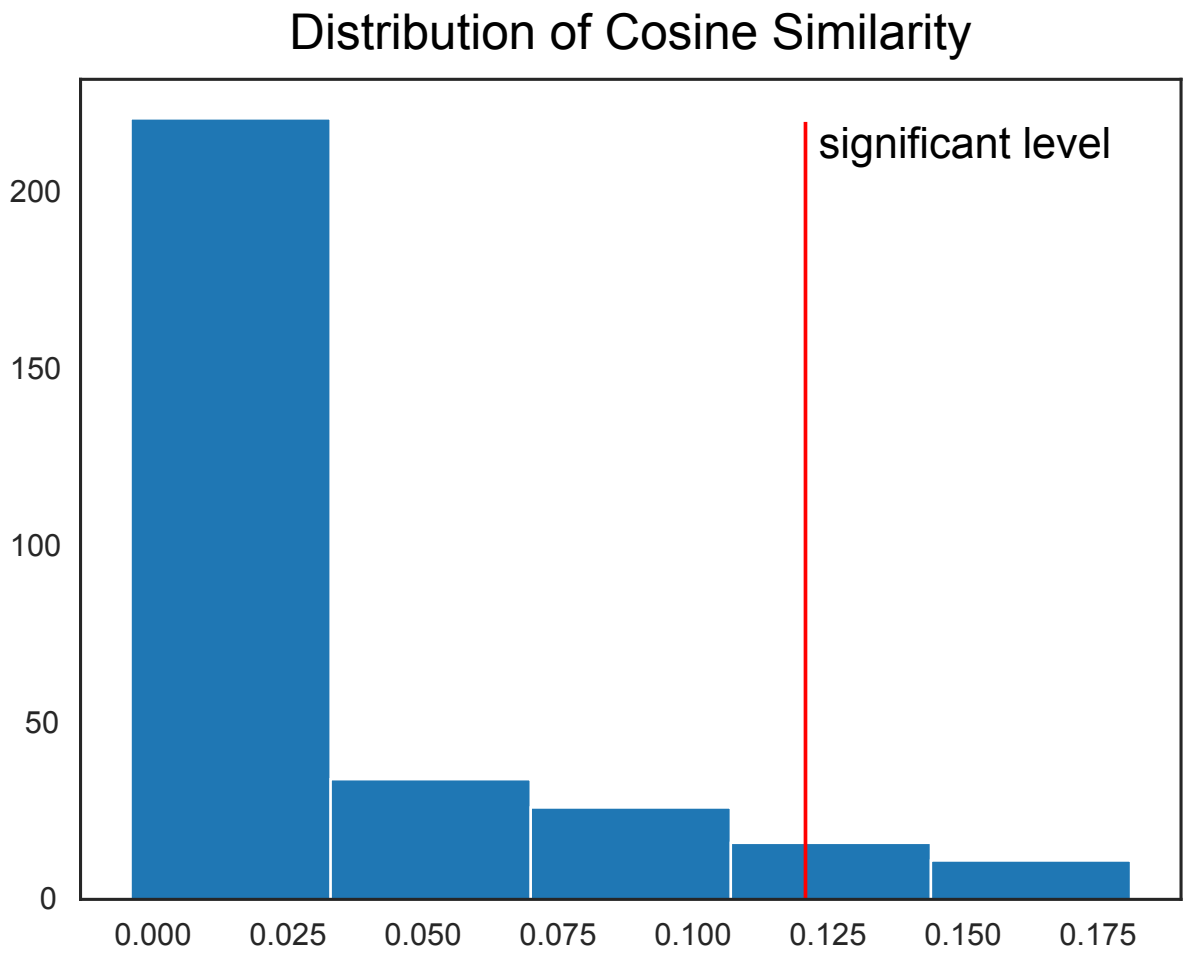

Figure 4.7: **Distribution of cosine similarities.**

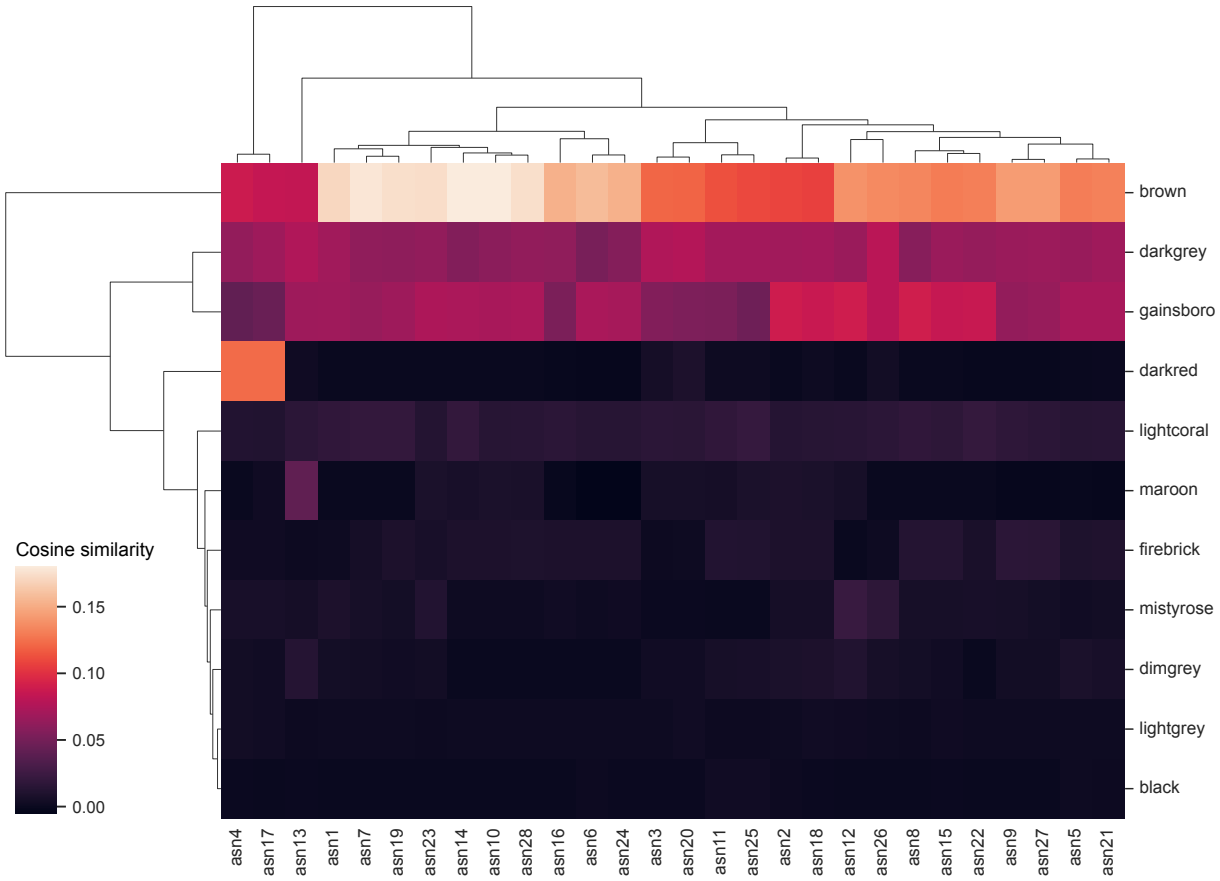

Figure 4.8: **Heatmap of cosine similarities between PywGCNA modules and Topyfic topics**

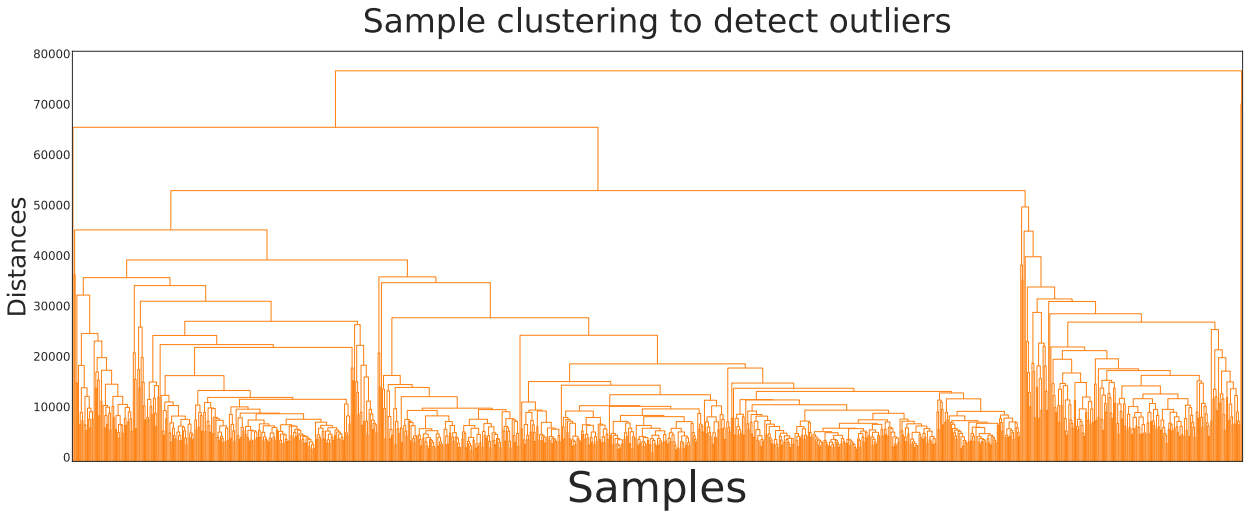

Figure 4.9: **Clustering dendrogram of samples based on TPM values.**

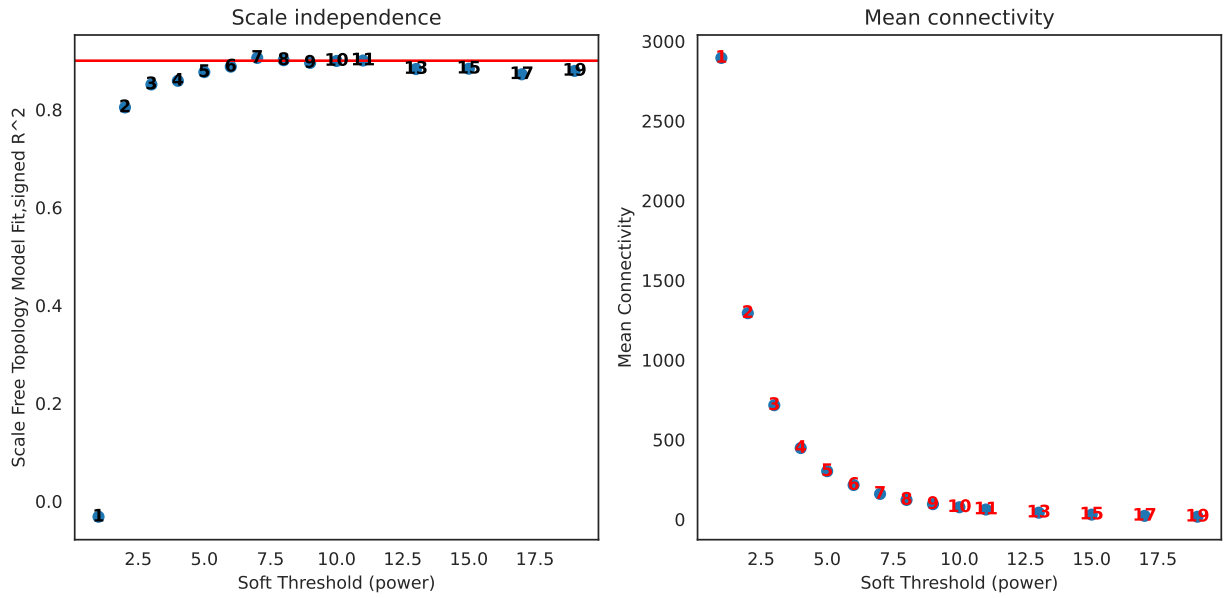

Figure 4.10: **Clustering dendrogram of samples based on TPM values. left)** The plot shows the scale-free topology fit index (y-axis) for different soft-thresholding powers (β) (x-axis). When the soft threshold is 11, R^2 is greater than 0.9. **right)** The plot shows the mean connectivity (degree, y-axis) for various soft-thresholding powers (x-axis). When the soft threshold is 11, the mean connectivity is less than 100.

mouse model	Tissue	Age	Genotype	Group	Sex	# Samples
cuprizon	brain	12mon	BL6	Cuprizon	M	8
cuprizon	brain	12mon	BL6	Control	M	8
cuprizon	brain	12mon	TREM2em1Adiuj:IUJTREM2R47H	Cuprizon	M	4
cuprizon	brain	12mon	TREM2em1Adiuj:IUJTREM2R47H	Control	M	4
cuprizon	brain	12mon	TREM2em1Aduci	Cuprizon	M	4
cuprizon	brain	12mon	TREM2em1Aduci	Control	M	4
cuprizon	brain	12mon	TREM2em2Adiuj:TREM2KO	Cuprizon	M	4
cuprizon	brain	12mon	TREM2em2Adiuj:TREM2KO	Control	M	4

Table 4.1: Discriotion of cuprizone cohort of bulk RNA-seq data

mouse model	Tissue	Age	Genotype	Group	Sex	# Samples
LPS	brain	8mon	ABI3HO	LPS(6h)	F	2
LPS	brain	8mon	ABI3HO	LPS(6h)	M	2
LPS	brain	8mon	ABCA7HO	LPS(6h)	F	2
LPS	brain	8mon	ABCA7HO	LPS(6h)	M	2
LPS	brain	8mon	BL6	LPS(6h)	F	2
LPS	brain	8mon	BL6	LPS(6h)	M	2
LPS	brain	8mon	ABI3HO	Saline(6h)	M	3
LPS	brain	8mon	ABCA7HO	Saline(6h)	M	3
LPS	brain	8mon	BL6	Saline(6h)	F	2
LPS	brain	8mon	BL6	Saline(6h)	M	1
LPS	brain	8mon	ABI3HO	LPS(24h)	F	2
LPS	brain	8mon	ABI3HO	LPS(24h)	M	2
LPS	brain	8mon	ABCA7HO	LPS(24h)	F	2
LPS	brain	8mon	ABCA7HO	LPS(24h)	M	2
LPS	brain	8mon	BL6	LPS(24h)	F	2
LPS	brain	8mon	BL6	LPS(24h)	M	2
LPS	brain	8mon	ABI3HO	Saline(24h)	F	1
LPS	brain	8mon	ABI3HO	Saline(24h)	M	2
LPS	brain	8mon	ABCA7HO	Saline(24h)	F	1
LPS	brain	8mon	ABCA7HO	Saline(24h)	M	2
LPS	brain	8mon	BL6	Saline(24h)	F	1
LPS	brain	8mon	BL6	Saline(24h)	M	2

Table 4.2: Discriotion of LPS bulk RNA-seq data

Chapter 5

Future directions

Leveraging genomics data to design better mouse models of AD

While current mouse models have played a crucial role in understanding certain aspects of Alzheimer's disease (AD) pathology, their limitations are evident. These models often oversimplify the complex genetic landscape of AD by relying on singular mutations or transgenes that fail to capture the polygenic nature observed in individuals with AD[289]. Additionally, the factors initiating AD are not yet known, and the field has primarily relied on models of the familial forms of the disease, linked to dominant genetic mutations. While these models have provided invaluable insights into disease pathogenesis, they have significant shortcomings. Current AD models are based on mutations found in familial cohorts of early-onset Alzheimer's disease (EOAD) despite EOAD only accounting for 2% of overall cases[290]. Consequently, there is a critical need for mouse models that more faithfully recapitulate the diverse genetic factors contributing to AD.

One promising path to enhance mouse models of AD involves humanizing specific genes to better replicate human AD pathology. Replacement of mouse genes with their human homologs, such as *Mapt*[291], has been attempted. However, despite successful gene replace-

ment, the subsequent cross with an APP knock-in showed no evidence of tauopathies such as neurofibrillary tangles (NFT) or neurodegeneration[291]. The integration of risk variants is a crucial aspect, but it alone does not guarantee a successful mouse model. For example, a knock-in of human variants should avoid causing aberrant splicing events not observed in humans, as was seen in several mouse models of the Trem2 R47H variant[292]. Although this particular variant is identified as an AD risk factor in humans, the mouse model exhibited a decrease in Trem2 expression due to aberrant splicing events[292,293]. Expression levels and distribution of the protein products should not change in comparison to wild-type in otherwise normal mice. Furthermore, when exploring GWAS hits, it is paramount to consider that AD shares certain pathology traits with other forms of dementia. Therefore, evaluating the specificity of identified variants in comparison to other dementias becomes crucial in refining the accuracy and relevance of mouse models in AD research.

In AD research, the use of both laboratory (lab)[294] and wild-derived mouse strains[295,296], along with Collaborative Cross (CC) lines[297], contributes to a comprehensive understanding of the genetic basis and disease mechanisms. Lab strains often inbred and genetically homogeneous, provide controlled environments for studying specific genetic factors and experimental interventions. They are valuable for assessing the effects of targeted genetic manipulations and investigating the molecular pathways involved in AD. On the other hand, wild-derived strains offer increased genetic diversity, capturing a broader range of natural genetic variations and environmental influences. These strains are instrumental in modeling the complex, polygenic nature of AD and understanding gene-environment interactions. Integrating lab, wild-derived, and CC lines allows researchers to bridge the gap between controlled experimental settings and the complexity observed in human populations. Lab strains facilitate precise genetic manipulations and mechanistic studies, while wild-derived strains and CC lines provide platforms for exploring the diverse genetic factors and phenotypic variations associated with AD. This combined approach enhances the translational relevance of preclinical studies, aiding in the identification of key genetic contributors, po-

tential therapeutic targets, and personalized treatment strategies for AD across a spectrum of genetic backgrounds.

In the past decade, long-read sequencing technologies have rapidly progressed[298,299]. They can be used to validate mouse models of AD and help us resolve complex genomic structures, detect structural variations, and capture full-length transcripts[300]. Long-read sequencing will allow researchers to inspect the precise nucleotide changes made during gene editing genome-wise, which is crucial for confirming the accuracy of the intended edits in mouse models designed to replicate specific mutations associated with AD. By analyzing the full-length transcripts of the mouse model using long RNA-seq, we have been able to assess how well the modified genes are expressed, spliced, and processed[118]. Long RNA-seq facilitates the capture of alternative splicing events in genes, providing valuable insights into mouse models where alternative splicing may contribute to disease progression[301]. By comparing splicing patterns in the model to those in wild-type mice, researchers can gain a clearer understanding of the model's accuracy. Collectively, the integration of long-read RNA-seq in these approaches allows researchers to acquire a deeper and more precise understanding of the genomic and transcriptomic landscapes of mouse models and ensures that mouse models faithfully recapitulate the genetic and molecular features associated with AD.

The successful generation of a new mouse model of AD that would mirror human pathology in terms of histology, aging, AD onset, gene expression patterns, and other relevant features would be transformative. Aligning the pathophysiological traits of this mouse model with human clinical data would facilitate the discovery of robust biomarkers as well as the pre-clinical testing of new treatments for Alzheimer's disease.

CRISPRi-Perturb-seq in Mouse Models of Alzheimer's Disease

Another promising path for exploring the function of genes involved in AD involves functional perturbations. The combination of CRISPR interference (CRISPRi) with single-cell RNA

sequencing (scRNA-seq), known as CRISPRi-Perturb-seq[302,303], has emerged as a robust method for unraveling the functional genomics of AD within mouse models. This technique enables large-scale functional genomics screening by systematically perturbing specific genes associated with AD pathology such as Apoe. Through targeted gene repression, researchers can identify key players in the disease process, contributing to a deeper understanding of AD molecular mechanisms.

Cell identity is encoded by gene regulatory networks (GRNs) that consist of directional links indicating regulatory connections, with the edge's source being the regulator gene and the sink being the target gene. GRNs are frequently employed to delineate how the expression of transcription factors influences the expression of target genes. CRISPRi-Perturb-seq also provides a unique advantage in uncovering regulatory networks and pathways affected by gene perturbations. This information is instrumental in elucidating the intricate molecular processes that contribute to AD progression. By perturbing genes linked to AD, researchers can discern how these alterations cascade through the regulatory machinery, influencing downstream signaling pathways and molecular processes. This holistic understanding of gene regulatory networks offers a comprehensive view of the interconnected cellular events that contribute to AD pathology.

Moreover, CRISPRi-Perturb-seq serves as a valuable tool for the validation of candidate genes previously identified through various approaches, such as genetics or transcriptomics. This aids in confirming the functional relevance of specific genes in the context of AD. Finally, by systematically perturbing genes associated with AD, CRISPRi-Perturb-seq assists in the identification of potential drug targets, contributing to the development of targeted therapeutic interventions for neurodegenerative diseases. In essence, CRISPRi-Perturb-seq stands as a comprehensive approach, advancing our understanding and potential treatment strategies for mouse models of AD.

Bibliography

[1] R. Carter, S. Aldridge, M. Page, and S. Parker. *The Human Brain Book*. DK Pub., 2009. ISBN 9780756654412. URL `https://books.google.com/books?id=2IfMPwAACAAJ`.

[2] Carl A Gold and Andrew E Budson. Memory loss in alzheimer's disease: implications for development of therapeutics. *Expert review of neurotherapeutics*, 8(12):1879–1891, 2008.

[3] Pietro Tiraboschi, LA Hansen, LJ Thal, and Jody Corey-Bloom. The importance of neuritic plaques and tangles to the development and evolution of ad. *Neurology*, 62 (11):1984–1989, 2004.

[4] AT Du, N Schuff, XP Zhu, WJ Jagust, BL Miller, BR Reed, JH Kramer, D Mungas, K Yaffe, HC Chui, et al. Atrophy rates of entorhinal cortex in ad and normal aging. *Neurology*, 60(3):481–486, 2003.

[5] CMA Pennartz, R Ito, PFMJ Verschure, FP Battaglia, and TW Robbins. The hippocampal–striatal axis in learning, prediction and goal-directed behavior. *Trends in neurosciences*, 34(10):548–559, 2011.

[6] Anders M Fjell, Linda McEvoy, Dominic Holland, Anders M Dale, Kristine B Walhovd, Alzheimer's Disease Neuroimaging Initiative, et al. What is normal in normal aging? effects of aging, amyloid and alzheimer's disease on the cerebral cortex and the hippocampus. *Progress in neurobiology*, 117:20–40, 2014.

[7] Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca Simmons, Sam Buckberry, Dulce Vargas Landin, Jahnvi Pflueger, et al. A single cell brain atlas in human alzheimer's disease. *Biorxiv*, page 628347, 2019.

[8] Hadas Keren-Shaul, Amit Spinrad, Assaf Weiner, Orit Matcovitch-Natan, Raz Dvir-Szternfeld, Tyler K Ulland, Eyal David, Kuti Baruch, David Lara-Astaiso, Beata Toth, et al. A unique microglia type associated with restricting development of alzheimer's disease. *Cell*, 169(7):1276–1290, 2017.

[9] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, 2018.

[10] Jason M Keil, Adel Qalieh, and Kenneth Y Kwan. Brain transcriptome databases: a user's guide. *Journal of Neuroscience*, 38(10):2399–2412, 2018.

[11] Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018.

[12] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317, 2019.

[13] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas Gray, David J Peeler, Sumit Mukherjee, Wei Chen, et al. Split-seq reveals cell types and lineages in the developing brain and spinal cord. *Science (New York, NY)*, 360(6385):176, 2018.

[14] Christopher S Von Bartheld, Jami Bahney, and Suzana Herculano-Houzel. The search

for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, 524(18):3865–3895, 2016.

[15] Joan Stiles and Terry L Jernigan. The basics of brain development. *Neuropsychology review*, 20(4):327–348, 2010.

[16] Kenneth Campbell and Magdalena Götz. Radial glia: multi-purpose cells for vertebrate brain development. *Trends in neurosciences*, 25(5):235–238, 2002.

[17] Kristjan R Jessen. Glial cells. *The international journal of biochemistry & cell biology*, 36(10):1861–1867, 2004.

[18] Leda Dimou and Magdalena Götz. Glial cells as progenitors and stem cells: new roles in the healthy and diseased brain. *Physiological reviews*, 94(3):709–737, 2014.

[19] Rafael Luján, Ryuichi Shigemoto, and Guillermina López-Bendito. Glutamate and gaba receptor signalling in the developing brain. *Neuroscience*, 130(3):567–580, 2005.

[20] Mateo Vélez-Fort, Etienne Audinat, and María Cecilia Angulo. Central role of gaba in neuron–glia interactions. *The Neuroscientist*, 18(3):237–250, 2012.

[21] Kaia Achim, Marjo Salminen, and Juha Partanen. Mechanisms regulating gabaergic neuron development. *Cellular and molecular life sciences*, 71:1395–1415, 2014.

[22] Susan J Vannucci, Fran Maher, and Ian A Simpson. Glucose transporter proteins in brain: delivery of glucose to neurons and glia. *Glia*, 21(1):2–21, 1997.

[23] Yuki Hattori. The multifaceted roles of embryonic microglia in the developing brain. *Frontiers in Cellular Neuroscience*, 17:988952, 2023.

[24] Mackenzie A Michell-Robinson, Hanane Touil, Luke M Healy, David R Owen, Bryce A Durafourt, Amit Bar-Or, Jack P Antel, and Craig S Moore. Roles of microglia in brain development, tissue maintenance and repair. *Brain*, 138(5):1138–1159, 2015.

[25] Luca Muzio, Alice Viotti, and Gianvito Martino. Microglia in neuroinflammation and neurodegeneration: from understanding to therapy. *Frontiers in neuroscience*, 15: 742065, 2021.

[26] Chao Gao, Jingwen Jiang, Yuyan Tan, and Shengdi Chen. Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets. *Signal transduction and targeted therapy*, 8(1):359, 2023.

[27] Marta Sochocka, Breno Satler Diniz, and Jerzy Leszek. Inflammatory response in the cns: friend or foe? *Molecular neurobiology*, 54:8071–8089, 2017.

[28] José A Rodríguez-Gómez, Edel Kavanagh, Pinelopi Engskog-Vlachos, Mikael KR Engskog, Antonio J Herrera, Ana M Espinosa-Oliva, Bertrand Joseph, Nabil Hajji, José L Venero, and Miguel A Burguillos. Microglia: agents of the cns pro-inflammatory response. *Cells*, 9(7):1717, 2020.

[29] Shalina S Ousman and Paul Kubes. Immune surveillance in the central nervous system. *Nature neuroscience*, 15(8):1096–1101, 2012.

[30] Lindsey C Mehl, Amritha V Manjally, Ouzéna Bouadi, Erin M Gibson, and Tuan Leng Tay. Microglia in brain development and regeneration. *Development*, 149(8): dev200425, 2022.

[31] Christopher J Bohlen, Brad A Friedman, Borislav Dejanovic, and Morgan Sheng. Microglia in brain development, homeostasis, and neurodegeneration. *Annual Review of Genetics*, 53:263–288, 2019.

[32] Cuicui Wang, Shuai Zong, Xiaolin Cui, Xueying Wang, Shuang Wu, Le Wang, Yingchao Liu, and Zhiming Lu. The effects of microglia-associated neuroinflamma-

tion on alzheimer's disease. *Frontiers in Immunology*, 14:1117172, 2023.

[33] Yonghee Kim, Jinhong Park, and Yoon Kyung Choi. The role of astrocytes in the central nervous system focused on bk channel and heme oxygenase metabolites: a review. *Antioxidants*, 8(5):121, 2019.

[34] Anna Victoria Molofsky and Benjamin Deneen. Astrocyte development: a guide for the perplexed. *Glia*, 63(8):1320–1329, 2015.

[35] Omer Ali Bayraktar, Luis C Fuentealba, Arturo Alvarez-Buylla, and David H Rowitch. Astrocyte development and heterogeneity. *Cold Spring Harbor perspectives in biology*, 7(1):a020362, 2015.

[36] Alexandra L Schober, Leigh E Wicki-Stordeur, Keith K Murai, and Leigh Anne Swayne. Foundations and implications of astrocyte heterogeneity during brain development and disease. *Trends in Neurosciences*, 45(9):692–703, 2022.

[37] Monika Bradl and Hans Lassmann. Oligodendrocytes: biology and pathology. *Acta neuropathologica*, 119:37–53, 2010.

[38] Dwight E Bergles and William D Richardson. Oligodendrocyte development and plasticity. *Cold Spring Harbor perspectives in biology*, 8(2):a020453, 2016.

[39] Sarah Kuhn, Laura Gritti, Daniel Crooks, and Yvonne Dombrowski. Oligodendrocytes in development, myelin generation and beyond. *Cells*, 8(11):1424, 2019.

[40] Martina Zvěřová. Clinical aspects of alzheimer's disease. *Clinical biochemistry*, 72:3–6, 2019.

[41] Christiane Reitz, Ekaterina Rogaeva, and Gary W Beecham. Late-onset vs non-mendelian early-onset alzheimer disease: A distinction without a difference? *Neurology Genetics*, 6(5), 2020.

[42] 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695, April 2023. ISSN 1552-5260, 1552-5279. doi: 10.1002/alz.13016. URL `https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13016`.

[43] Kumar B Rajan, Jennifer Weuve, Lisa L Barnes, Elizabeth A McAninch, Robert S Wilson, and Denis A Evans. Population estimate of people with clinical alzheimer's disease and mild cognitive impairment in the united states (2020–2060). *Alzheimer's & dementia*, 17(12):1966–1975, 2021.

[44] Scott C Neu, Judy Pa, Walter Kukull, Duane Beekly, Amanda Kuzma, Prabhakaran Gangadharan, Li-San Wang, Klaus Romero, Stephen P Arneric, Alberto Redolfi, et al. Apolipoprotein e genotype and sex risk factors for alzheimer disease: a meta-analysis. *JAMA neurology*, 74(10):1178–1189, 2017.

[45] Brian W Kunkle, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Vincent Damotte, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J Van Der Lee, Alexandre Amlie-Wolf, et al. Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing. *Nature genetics*, 51(3):414–430, 2019.

[46] Joyce Vrijsen, Ameen Abu-Hanna, Sophia E de Rooij, and Nynke Smidt. Association between dementia parental family history and mid-life modifiable risk factors for dementia: A cross-sectional study using propensity score matching within the lifelines cohort. *BMJ open*, 11(12):e049918, 2021.

[47] Jill S Goldman, Susan E Hahn, Jennifer Williamson Catania, Susan Larusse-Eckert, Melissa Barber Butson, Malia Rumbaugh, Michelle N Strecker, J Scott Roberts, Wylie

Burke, Richard Mayeux, et al. Genetic counseling and testing for alzheimer disease: joint practice guidelines of the american college of medical genetics and the national society of genetic counselors. *Genetics in medicine*, 13(6):597–605, 2011.

[48] Kelsey Roberts Noam Grysman Jiaxi Lu Rita Khoury, Amy Gallop and George T. Grossberg. Pharmacotherapy for alzheimer's disease: what's new on the horizon? *Expert Opinion on Pharmacotherapy*, 23(11):1305–1323, 2022. doi: 10.1080/14656566.2022.2097868.

[49] Angela L Guillozet, Sandra Weintraub, Deborah C Mash, and M Marsel Mesulam. Neurofibrillary tangles, amyloid, and memory in aging and mild cognitive impairment. *Archives of neurology*, 60(5):729–736, 2003.

[50] Peng Chen, ZhiLei Guo, and Benhong Zhou. Insight into the role of adult hippocampal neurogenesis in aging and alzheimer's disease. *Ageing Research Reviews*, 84:101828, 2023.

[51] Evgenia Salta, Orly Lazarov, Carlos P Fitzsimons, Rudolph Tanzi, Paul J Lucassen, and Se Hoon Choi. Adult hippocampal neurogenesis in alzheimer's disease: A roadmap to clinical relevance. *Cell Stem Cell*, 30(2):120–136, 2023.

[52] Bang-Sheng Wu, Ya-Ru Zhang, Hong-Qi Li, Kevin Kuo, Shi-Dong Chen, Qiang Dong, Yong Liu, and Jin-Tai Yu. Cortical structure and the risk for alzheimer's disease: a bidirectional mendelian randomization study. *Translational psychiatry*, 11(1):476, 2021.

[53] Chen Ma, Fenfang Hong, and Shulong Yang. Amyloidosis in alzheimer's disease: Pathogeny, etiology, and related therapeutic directions. *Molecules*, 27(4):1210, 2022.

[54] Elaine K Perry. The cholinergic hypothesis—ten years on. *British Medical Bulletin*, 42(1):63–69, 1986.

[55] Clemence Cheignon, M Tomas, D Bonnefont-Rousselot, Peter Faller, Christelle Hureau, and Fabrice Collin. Oxidative stress and the amyloid beta peptide in alzheimer's disease. *Redox biology*, 14:450–464, 2018.

[56] Sara Merlo, Simona Federica Spampinato, and Maria Angela Sortino. Estrogen and alzheimer's disease: Still an attractive topic despite disappointment from early clinical results. *European Journal of Pharmacology*, 817:51–58, 2017.

[57] Eric Karran and Bart De Strooper. The amyloid cascade hypothesis: are we poised for success or failure? *Journal of neurochemistry*, 139:237–252, 2016.

[58] Harald Hampel, John Hardy, Kaj Blennow, Christopher Chen, George Perry, Seung Hyun Kim, Victor L Villemagne, Paul Aisen, Michele Vendruscolo, Takeshi Iwatsubo, et al. The amyloid-β pathway in alzheimer's disease. *Molecular psychiatry*, 26 (10):5481–5503, 2021.

[59] Malamati Kourti and Athanasios Metaxas. A systematic review and meta-analysis of tau phosphorylation in mouse models of familial alzheimer's disease. *Neurobiology of Disease*, page 106427, 2024.

[60] Benjamin R Troutwine, Laylan Hamid, Colton R Lysaker, Taylor A Strope, and Heather M Wilkins. Apolipoprotein e and alzheimer's disease. *Acta Pharmaceutica Sinica B*, 12(2):496–510, 2022.

[61] Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2): 106–118, 2013.

[62] Michael E Belloy, Shea J Andrews, Yann Le Guen, Michael Cuccaro, Lindsay A Farrer, Valerio Napolioni, and Michael D Greicius. Apoe genotype and alzheimer disease risk across age, sex, and population ancestry. *JAMA neurology*, 80(12):1284–1294, 2023.

[63] Heiko Braak and Kelly Del Tredici. Neurofibrillary tangles. *Encyclopedia of movement disorders, vol 2: HP*, pages 265–269, 2010.

[64] Prakash N Kendre, Ajinkya Pote, Rasika Bhalke, Bhupendra Gopalbhai Prajapati, Shirish P Jain, and Devesh Kapoor. Lipid nanoparticles in targeting alzheimer's disease. In *Alzheimer's Disease and Advanced Drug Delivery Strategies*, pages 283–295. Elsevier, 2024.

[65] Michel Goedert, R Anthony Crowther, Sjors HW Scheres, and Maria Grazia Spillantini. Tau and neurodegeneration. *Cytoskeleton*, 81(1):95–102, 2024.

[66] Lukai Zheng, Anna Rubinski, Jannis Denecke, Ying Luan, Ruben Smith, Olof Strandberg, Erik Stomrud, Rik Ossenkoppele, Diana Otero Svaldi, Ixavier Alonzo Higgins, et al. Combined connectomics, mapt gene expression, and amyloid deposition to explain regional tau deposition in alzheimer disease. *Annals of neurology*, 2023.

[67] Matthew J Benskey, Spencer Panoushek, Takashi Saito, Takaomi C Saido, Tessa Grabinski, and Nicholas M Kanaan. Behavioral and neuropathological characterization over the adult lifespan of the human tau knock-in mouse. *Frontiers in Aging Neuroscience*, 15, 2023.

[68] Md Sahab Uddin and Lee Wei Lim. Glial cells in alzheimer's disease: From neuropathological changes to therapeutic implications. *Ageing Research Reviews*, 78:101622, 2022.

[69] I Lopategui Cabezas, A Herrera Batista, and G Pentón Rol. The role of glial cells in alzheimer disease: potential therapeutic implications. *Neurología (English Edition)*, 29(5):305–309, 2014.

[70] Jifei Miao, Haixia Ma, Yang Yang, Yuanpin Liao, Cui Lin, Juanxia Zheng, Muli Yu, and Jiao Lan. Microglia in alzheimer's disease: pathogenesis, mechanisms, and therapeutic potentials. *Frontiers in Aging Neuroscience*, 15:1201982, 2023.

[71] David V Hansen, Jesse E Hanson, and Morgan Sheng. Microglia in alzheimer's disease. *Journal of Cell Biology*, 217(2):459–472, 2018.

[72] Joshua A Smith, Arabinda Das, Swapan K Ray, and Naren L Banik. Role of pro-inflammatory cytokines released from microglia in neurodegenerative diseases. *Brain research bulletin*, 87(1):10–20, 2012.

[73] Jenny U Johansson, Nathaniel S Woodling, Ju Shi, and Katrin I Andreasson. Inflammatory cyclooxygenase activity and pge2 signaling in models of alzheimer's disease. *Current immunology reviews*, 11(2):125–131, 2015.

[74] Dawling A Dionisio-Santos, John A Olschowka, and M Kerry O'Banion. Exploiting microglial and peripheral immune cell crosstalk to treat alzheimer's disease. *Journal of neuroinflammation*, 16(1):1–13, 2019.

[75] Yongle Cai, Jingliu Liu, Bin Wang, Miao Sun, and Hao Yang. Microglia in the neuroinflammatory pathogenesis of alzheimer's disease and related therapeutic targets. *Frontiers in immunology*, 13:856376, 2022.

[76] Thomas Vogels, Adriana-Natalia Murgoci, and Tomáš Hromádka. Intersection of pathological tau and microglia at the synapse. *Acta neuropathologica communications*, 7:1–25, 2019.

[77] Yijun Chen and Yang Yu. Tau and neuroinflammation in alzheimer's disease: Interplay

mechanisms and clinical translation. *Journal of Neuroinflammation*, 20(1):165, 2023.
[78] Michael A DeTure and Dennis W Dickson. The neuropathological diagnosis of alzheimer's disease. *Molecular neurodegeneration*, 14(1):1–18, 2019.
[79] Agnieszka M Jurga, Martyna Paleczna, Justyna Kadluczka, and Katarzyna Z Kuter. Beyond the gfap-astrocyte protein markers in the brain. *Biomolecules*, 11(9):1361, 2021.
[80] Georgia R Frost and Yue-Ming Li. The role of astrocytes in amyloid production and alzheimer's disease. *Open biology*, 7(12):170228, 2017.
[81] Sung S Choi, Hong J Lee, Inja Lim, Jun-ichi Satoh, and Seung U Kim. Human astrocytes: secretome profiles of cytokines and chemokines. *PloS one*, 9(4):e92325, 2014.
[82] Egle Cekanaviciute and Marion S Buckwalter. Astrocytes: integrative regulators of neuroinflammation in stroke and other neurological diseases. *Neurotherapeutics*, 13: 685–701, 2016.
[83] Sofia Söllvander, Elisabeth Nikitidou, Robin Brolin, Linda Söderberg, Dag Sehlin, Lars Lannfelt, and Anna Erlandsson. Accumulation of amyloid-β by astrocytes result in enlarged endosomes and microvesicle-induced apoptosis of neurons. *Molecular neurodegeneration*, 11(1):1–19, 2016.
[84] Elly M Hol and Milos Pekny. Glial fibrillary acidic protein (gfap) and the astrocyte intermediate filament system in diseases of the central nervous system. *Current opinion in cell biology*, 32:121–130, 2015.
[85] Michel Maitre, Hélène Jeltsch-David, Nwife Getrude Okechukwu, Christian Klein, Christine Patte-Mensah, and Ayikoe-Guy Mensah-Nyagan. Myelin in alzheimer's disease: culprit or bystander? *Acta Neuropathologica Communications*, 11(1):1–18, 2023.
[86] Ewa Papuć and Konrad Rejdak. The role of myelin damage in alzheimer's disease pathology. *Archives of Medical Science*, 16(2):345–341, 2018.
[87] Sara E Nasrabady, Batool Rizvi, James E Goldman, and Adam M Brickman. White matter changes in alzheimer's disease: a focus on myelin and oligodendrocytes. *Acta neuropathologica communications*, 6(1):1–10, 2018.
[88] Peibin Zou, Chongyun Wu, Timon Cheng-Yi Liu, Rui Duan, and Luodan Yang. Oligodendrocyte progenitor cells in alzheimer's disease: from physiology to pathology. *Translational Neurodegeneration*, 12(1):52, 2023.
[89] Natalia Ledo Husby Phillips and Tania L Roth. Animal models and their contribution to our understanding of the relationship between environments, epigenetic modifications, and behavior. *Genes*, 10(1):47, 2019.
[90] Adriana Domínguez-Oliva, Ismael Hernández-Ávalos, Julio Martínez-Burnes, Adriana Olmos-Hernández, Antonio Verduzco-Mendoza, and Daniel Mota-Rojas. The importance of animal models in biomedical research: Current insights and applications. *Animals*, 13(7):1223, 2023.
[91] Raquel Sanchez-Varo, Marina Mejias-Ortega, Juan Jose Fernandez-Valenzuela, Cristina Nuñez-Diaz, Laura Caceres-Palomo, Laura Vegas-Gomez, Elisabeth Sanchez-Mejias, Laura Trujillo-Estrada, Juan Antonio Garcia-Leon, Ines Moreno-Gonzalez, et al. Transgenic mouse models of alzheimer's disease: An integrative analysis. *International Journal of Molecular Sciences*, 23(10):5404, 2022.
[92] Alicia M Hall and Erik D Roberson. Mouse models of alzheimer's disease. *Brain*

research bulletin, 88(1):3–12, 2012.

[93] Leon M Tai, Juan Maldonado Weng, Mary Jo LaDu, and Scott T Brady. Relevance of transgenic mouse models for alzheimer's disease. *Progress in molecular biology and translational science*, 177:1–48, 2021.

[94] Edward Rockenstein, Leslie Crews, and Eliezer Masliah. Transgenic animal models of neurodegenerative diseases and their application to treatment development. *Advanced drug delivery reviews*, 59(11):1093–1102, 2007.

[95] URL `https://www.alzforum.org/research-models/alzheimers-disease`.

[96] Sadim Jawhar, Anna Trawicka, Carolin Jenneckens, Thomas A Bayer, and Oliver Wirths. Motor deficits, neuron loss, and reduced anxiety coinciding with axonal degeneration and intraneuronal aβ aggregation in the 5xfad mouse model of alzheimer's disease. *Neurobiology of aging*, 33(1):196–e29, 2012.

[97] Diederik Moechars, K Lorent, Bart De Strooper, Ilse Dewachter, and Freddy Van Leuven. Expression in brain of amyloid precursor protein mutated in the alpha-secretase site causes disturbed behavior, neuronal degeneration and premature death in transgenic mice. *The EMBO Journal*, 15(6):1265–1274, 1996.

[98] Miguel Vidal, Roger Morris, Frank Grosveld, and Eugenia Spanopoulou. Tissue-specific control elements of the thy-1 gene. *The EMBO journal*, 9(3):833–840, 1990.

[99] Holly Oakley, Sarah L Cole, Sreemathi Logan, Erika Maus, Pei Shao, Jeffery Craft, Angela Guillozet-Bongaarts, Masuo Ohno, John Disterhoft, Linda Van Eldik, et al. Intraneuronal β-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial alzheimer's disease mutations: potential factors in amyloid plaque formation. *Journal of Neuroscience*, 26(40):10129–10140, 2006.

[100] Dominic I Javonillo, Kristine M Tran, Jimmy Phan, Edna Hingco, Enikö A Kramár, Celia da Cunha, Stefania Forner, Shimako Kawauchi, Giedre Milinkeviciute, Angela Gomez-Arboledas, et al. Systematic phenotyping and characterization of the 3xtg-ad mouse model of alzheimer's disease. *Frontiers in Neuroscience*, 15:785276, 2022.

[101] Salvatore Oddo, Antonella Caccamo, Jason D Shepherd, M Paul Murphy, Todd E Golde, Rakez Kayed, Raju Metherate, Mark P Mattson, Yama Akbari, and Frank M LaFerla. Triple-transgenic model of alzheimer's disease with plaques and tangles: intracellular aβ and synaptic dysfunction. *Neuron*, 39(3):409–421, 2003.

[102] Bo Zhou, Jacqueline G Lu, Alberto Siddu, Marius Wernig, and Thomas C Südhof. Synaptogenic effect of app-swedish mutation in familial alzheimer's disease. *Science translational medicine*, 14(667):eabn9380, 2022.

[103] Shinya Tokuhiro, Taisuke Tomita, Hiroshi Iwata, Takuo Kosaka, Takaomi C Saido, Kei Maruyama, and Takeshi Iwatsubo. The presenilin 1 mutation (m146v) linked to familial alzheimer's disease attenuates the neuronal differentiation of ntera 2 cells. *Biochemical and biophysical research communications*, 244(3):751–755, 1998.

[104] Bernardino Ghetti, Adrian L Oblak, Bradley F Boeve, Keith A Johnson, Bradford C Dickerson, and Michel Goedert. Invited review: frontotemporal dementia caused by microtubule-associated protein tau gene (mapt) mutations: a chameleon for neuropathology and neuroimaging. *Neuropathology and applied neurobiology*, 41(1):24–46, 2015.

[105] Kurt R Stover, Mackenzie A Campbell, Christine M Van Winssen, and Richard E Brown. Early detection of cognitive deficits in the 3xtg-ad mouse model of alzheimer's

disease. *Behavioural brain research*, 289:29–38, 2015.

[106] Alejandro R Roda, Gisela Esquerda-Canals, Joaquim Martí-Clúa, and Sandra Villegas. Cognitive impairment in the 3xtg-ad mouse model of alzheimer's disease is affected by aβ-immunotherapy and cognitive stimulation. *Pharmaceutics*, 12(10):944, 2020.

[107] Anna Escrig, Carla Canal, Paula Sanchis, Olaya Fernández-Gayol, Alejandro Montilla, Gemma Comes, Amalia Molinero, Mercedes Giralt, Lydia Giménez-Llort, Christoph Becker-Pauly, et al. Il-6 trans-signaling in the brain influences the behavioral and physio-pathological phenotype of the tg2576 and 3xtgad mouse models of alzheimer's disease. *Brain, Behavior, and Immunity*, 82:145–159, 2019.

[108] Ted M Dawson, Todd E Golde, and Clotilde Lagier-Tourenne. Animal models of neurodegenerative diseases. *Nature neuroscience*, 21(10):1370–1379, 2018.

[109] Alessandra C Martini, Stefania Forner, Laura Trujillo-Estrada, David Baglietto-Vargas, and Frank M LaFerla. Past to future: what animal models have taught us about alzheimer's disease. *Journal of Alzheimer's Disease*, 64(s1):S365–S378, 2018.

[110] Jeffrey Cummings, Yadi Zhou, Garam Lee, Kate Zhong, Jorge Fonseca, and Feixiong Cheng. Alzheimer's disease drug development pipeline: 2023. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 9(2):e12385, 2023.

[111] Anthony King. The search for better animal models of alzheimer's disease. *Nature*, 559(7715):S13–S13, 2018.

[112] Masashi Kitazawa, Rodrigo Medeiros, and Frank M LaFerla. Transgenic mouse models of alzheimer disease: developing a better model as a tool for therapeutic interventions. *Current pharmaceutical design*, 18(8):1131–1147, 2012.

[113] Hiroki Sasaguri, Per Nilsson, Shoko Hashimoto, Kenichi Nagata, Takashi Saito, Bart De Strooper, John Hardy, Robert Vassar, Bengt Winblad, and Takaomi C Saido. App mouse models for alzheimer's disease preclinical studies. *The EMBO journal*, 36(17): 2473–2487, 2017.

[114] Paul Hollingworth, Denise Harold, Rebecca Sims, Amy Gerrish, Jean-Charles Lambert, Minerva M Carrasquillo, Richard Abraham, Marian L Hamshere, Jaspreet Singh Pahwa, Valentina Moskvina, et al. Common variants at abca7, ms4a6a/ms4a4e, epha1, cd33 and cd2ap are associated with alzheimer's disease. *Nature genetics*, 43(5):429–435, 2011.

[115] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458, 2013.

[116] Badri N Vardarajan, Mahdi Ghani, Amanda Kahn, Stephanie Sheikh, Christine Sato, Sandra Barral, Joseph H Lee, Rong Cheng, Christiane Reitz, Rafael Lantigua, et al. Rare coding mutations identified by sequencing of a lzheimer disease genome-wide association studies loci. *Annals of neurology*, 78(3):487–498, 2015.

[117] Morgan Robinson, Brenda Y Lee, and Francis T Hane. Recent progress in alzheimer's disease research, part 2: genetics and epidemiology. *Journal of Alzheimer's Disease*, 57(2):317–330, 2017.

[118] Kristine M Tran, Shimako Kawauchi, Enikö A Kramár, Narges Rezaie, Heidi Yahan Liang, Jasmine S Sakr, Angela Gomez-Arboledas, Miguel A Arreola, Celia da Cunha, Jimmy Phan, et al. A trem2r47h mouse model without cryptic splicing drives age-and

disease-dependent tissue damage and synaptic loss in response to plaques. *Molecular neurodegeneration*, 18(1):1–26, 2023.

[119] David Baglietto-Vargas, Stefania Forner, Lena Cai, Alessandra C Martini, Laura Trujillo-Estrada, Vivek Swarup, Marie Minh Thu Nguyen, Kelly Do Huynh, Dominic I Javonillo, Kristine Minh Tran, et al. Generation of a humanized aβ expressing mouse demonstrating aspects of alzheimer's disease-like pathology. *Nature communications*, 12(1):2421, 2021.

[120] Stefania Forner, Shimako Kawauchi, Gabriela Balderrama-Gutierrez, Enikö A Kramár, Dina P Matheos, Jimmy Phan, Dominic I Javonillo, Kristine M Tran, Edna Hingco, Celia da Cunha, et al. Systematic phenotyping and characterization of the 5xfad mouse model of alzheimer's disease. *Scientific Data*, 8(1):270, 2021.

[121] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

[122] Xinyi Guo, Yuanyuan Zhang, Liangtao Zheng, Chunhong Zheng, Jintao Song, Qiming Zhang, Boxi Kang, Zhouzerui Liu, Liang Jin, Rui Xing, et al. Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing. *Nature medicine*, 24 (7):978–985, 2018.

[123] Jeanette Baran-Gale, Tamir Chandra, and Kristina Kirschner. Experimental design for single-cell rna sequencing. *Briefings in functional genomics*, 17(4):233–239, 2018.

[124] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[125] Trygve E Bakken, Rebecca D Hodge, Jeremy A Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PloS one*, 13(12):e0209648, 2018.

[126] Rebecca Waag and Johannes Bohacek. Single-nucleus rna-sequencing in brain tissue. *Current Protocols*, 3(11):e919, 2023.

[127] Nathan R Tucker, Mark Chaffin, Stephen J Fleming, Amelia W Hall, Victoria A Parsons, Kenneth C Bedi Jr, Amer-Denis Akkad, Caroline N Herndon, Alessandro Arduini, Irinna Papangeli, et al. Transcriptional and cellular diversity of the human heart. *Circulation*, 142(5):466–482, 2020.

[128] Wenfei Sun, Hua Dong, Miroslav Balaz, Michal Slyper, Eugene Drokhlyansky, Georgia Colleluori, Antonio Giordano, Zuzana Kovanicova, Patrik Stefanicka, Lucia Balazova, et al. snrna-seq reveals a subpopulation of adipocytes that regulates thermogenesis. *Nature*, 587(7832):98–102, 2020.

[129] Michael J Petrany, Casey O Swoboda, Chengyi Sun, Kashish Chetal, Xiaoting Chen, Matthew T Weirauch, Nathan Salomonis, and Douglas P Millay. Single-nucleus rna-seq identifies transcriptional heterogeneity in multinucleated skeletal myofibers. *Nature communications*, 11(1):6374, 2020.

[130] Michal Slyper, Caroline BM Porter, Orr Ashenberg, Julia Waldman, Eugene Drokhlyansky, Isaac Wakiro, Christopher Smillie, Gabriela Smith-Rosario, Jingyi Wu, Danielle Dionne, et al. A single-cell and single-nucleus rna-seq toolbox for fresh and frozen human tumors. *Nature medicine*, 26(5):792–802, 2020.

[131] Nicola Thrupp, Carlo Sala Frigerio, Leen Wolfs, Nathan G Skene, Nicola Fattorelli,

Suresh Poovathingal, Yannick Fourne, Paul M Matthews, Tom Theys, Renzo Mancuso, et al. Single-nucleus rna-seq is not suitable for detection of microglial activation genes in humans. *Cell reports*, 32(13), 2020.

[132] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[133] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11):1093–1095, 2013.

[134] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS computational biology*, 11(6):e1004333, 2015.

[135] Justina Žurauskienė and Christopher Yau. pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC bioinformatics*, 17:1–11, 2016.

[136] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[137] Josip S Herman, null Sagar, and Dominic Gruen. Fateid infers cell fate bias in multipotent progenitors from single-cell rna-seq data. *Nature methods*, 15(5):379–386, 2018.

[138] Joseph A DiGiuseppe, Jolene L Cardinali, William N Rezuke, and Dana Pe'er. Phenograph and visne facilitate the identification of abnormal t-cell populations in routine clinical flow cytometric data. *Cytometry Part B: Clinical Cytometry*, 94(5):744–757, 2018.

[139] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

[140] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[141] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, pages 1–12, 2023.

[142] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

[143] Yuqi Tan and Patrick Cahan. Singlecellnet: a computational tool to classify single cell rna-seq data across platforms and across species. *Cell systems*, 9(2):207–213, 2019.

[144] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

[145] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362, 2018.

[146] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[147] Akira Cortal, Loredana Martignetti, Emmanuelle Six, and Antonio Rausell. Gene signature extraction and cell identity recognition at the single-cell level with cell-id. *Nature biotechnology*, 39(9):1095–1102, 2021.

[148] Ze Zhang, Danni Luo, Xue Zhong, Jin Huk Choi, Yuanqing Ma, Stacy Wang, Elena Mahrt, Wei Guo, Eric W Stawiski, Zora Modrusan, et al. Scina: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes*, 10(7):531, 2019.

[149] Allen W Zhang, Ciara O'Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, et al. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling. *Nature methods*, 16(10):1007–1015, 2019.

[150] Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N Luu, and Tin Nguyen. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications*, 12(1):1029, 2021.

[151] Yuqi Cheng, Xingyu Fan, Jianing Zhang, and Yu Li. A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data. *Communications Biology*, 6(1):545, 2023.

[152] Xin Shao, Jie Liao, Xiaoyan Lu, Rui Xue, Ni Ai, and Xiaohui Fan. sccatch: automatic annotation on cell types of clusters from single-cell rna sequencing data. *Iscience*, 23 (3), 2020.

[153] Kushal K Dey, Chiaowen Joyce Hsiao, and Matthew Stephens. Visualizing the structure of rna-seq expression data using grade of membership models. *PLoS genetics*, 13 (3):e1006599, 2017.

[154] David M Blei and John D Lafferty. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC, 2009.

[155] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[156] Daniel Spakowicz, Shaoke Lou, Brian Barron, Jose L Gomez, Tianxiao Li, Qing Liu, Nicole Grant, Xiting Yan, Rebecca Hoyd, George Weinstock, et al. Approaches for integrating heterogeneous rna-seq data reveal cross-talk between microbes and genes in asthmatic patients. *Genome biology*, 21:1–22, 2020.

[157] Hyeon-Jin Kim, Galip Gürkan Yardımcı, Giancarlo Bonora, Vijay Ramani, Jie Liu, Ruolan Qiu, Choli Lee, Jennifer Hesson, Carol B Ware, Jay Shendure, et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data. *PLoS computational biology*, 16(9):e1008173, 2020.

[158] Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nature methods*, 16(5):397–400, 2019.

[159] Ko Abe, Masaaki Hirayama, Kinji Ohno, and Teppei Shimamura. A latent allocation model for the analysis of microbial composition and disease. *BMC bioinformatics*, 19: 171–177, 2018.

[160] Jifang Yan, Guohui Chuai, Tao Qi, Fangyang Shao, Chi Zhou, Chenyu Zhu, Jing Yang,

Yifei Yu, Cong Shi, Ning Kang, et al. Metatopics: an integration tool to analyze microbial community profile by topic model. *BMC genomics*, 18(1):1–5, 2017.

[161] Justin Johan Jozias van Der Hooft, Joe Wandy, Michael P Barrett, Karl EV Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743, 2016.

[162] Naruemon Pratanwanich and Pietro Lio. Exploring the complexity of pathway–drug relationships using latent dirichlet allocation. *Computational biology and chemistry*, 53:144–152, 2014.

[163] P Langfelder. Hs, 2008. wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559.

[164] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

[165] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2017.

[166] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

[167] Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *BioRxiv*, pages 2021–12, 2021.

[168] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757, 2023.

[169] Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, et al. Ensembl 2022. *Nucleic acids research*, 50(D1):D988–D995, 2022.

[170] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49 (D1):D605–D612, 2021.

[171] Thorlakur Jonsson, Hreinn Stefansson, Stacy Steinberg, Ingileif Jonsdottir, Palmi V Jonsson, Jon Snaedal, Sigurbjorn Bjornsson, Johanna Huttenlocher, Allan I Levey, James J Lah, et al. Variant of trem2 associated with the risk of alzheimer's disease. *New England Journal of Medicine*, 368(2):107–116, 2013.

[172] Rita Guerreiro, Aleksandra Wojtas, Jose Bras, Minerva Carrasquillo, Ekaterina Rogaeva, Elisa Majounie, Carlos Cruchaga, Celeste Sassi, John SK Kauwe, Steven Younkin, et al. Trem2 variants in alzheimer's disease. *New England Journal of Medicine*, 368(2):117–127, 2013.

[173] William J Meilandt, Hai Ngu, Alvin Gogineni, Guita Lalehzadeh, Seung-Hye Lee, Karpagam Srinivasan, Jose Imperio, Tiffany Wu, Martin Weber, Agatha J Kruse, et al. Trem2 deletion reduces late-stage amyloid plaque accumulation, elevates the aβ42: Aβ40 ratio, and exacerbates axonal dystrophy and dendritic spine loss in the

ps2app alzheimer's mouse model. *Journal of Neuroscience*, 40(9):1956–1974, 2020.

[174] Yaming Wang, Tyler K Ulland, Jason D Ulrich, Wilbur Song, John A Tzaferis, Justin T Hole, Peng Yuan, Thomas E Mahan, Yang Shi, Susan Gilfillan, et al. Trem2-mediated early microglial response limits diffusion and toxicity of amyloid plaques. *Journal of Experimental Medicine*, 213(5):667–675, 2016.

[175] Peng Yuan, Carlo Condello, C Dirk Keene, Yaming Wang, Thomas D Bird, Steven M Paul, Wenjie Luo, Marco Colonna, David Baddeley, and Jaime Grutzendler. Erratum: Trem2 haplodeficiency in mice and humans impairs the microglia barrier function leading to decreased amyloid compaction and severe axonal dystrophy (neuron (**)**(*–*)(10.1016/j. neuron. 2016.05. 003)(s0896627316301635)). *Neuron*, 92(1): 252–264, 2016.

[176] Stefan Prokop, Kelly R Miller, Sergio R Labra, Rose M Pitkin, Kevt'her Hoxha, Sneha Narasimhan, Lakshmi Changolkar, Alyssa Rosenbloom, Virginia M-Y Lee, and John Q Trojanowski. Impact of trem2 risk variants on brain region-specific immune activation and plaque microenvironment in alzheimer's disease patient brain samples. *Acta neuropathologica*, 138:613–630, 2019.

[177] Adam C Naj, Gyungah Jun, Gary W Beecham, Li-San Wang, Badri Narayan Vardarajan, Jacqueline Buros, Paul J Gallins, Joseph D Buxbaum, Gail P Jarvik, Paul K Crane, et al. Common variants at ms4a4/ms4a6e, cd2ap, cd33 and epha1 are associated with late-onset alzheimer's disease. *Nature genetics*, 43(5):436–441, 2011.

[178] Arne De Roeck, Tobi Van den Bossche, Julie van der Zee, Jan Verheijen, Wouter De Coster, Jasper Van Dongen, Lubina Dillen, Yalda Baradaran-Heravi, Bavo Heeman, Raquel Sanchez-Valle, et al. Deleterious abca7 mutations and transcript rescue mechanisms in early onset alzheimer's disease. *Acta neuropathologica*, 134:475–487, 2017.

[179] Celeste Sassi, Michael A Nalls, Perry G Ridge, Jesse R Gibbs, Jinhui Ding, Michelle K Lupton, Claire Troakes, Katie Lunnon, Safa Al-Sarraj, Kristelle S Brown, et al. Abca7 p. g215s as potential protective factor for alzheimer's disease. *Neurobiology of aging*, 46:235–e1, 2016.

[180] Liene Bossaerts, Elisabeth Hendrickx Van de Craen, Rita Cacace, Bob Asselbergh, and Christine Van Broeckhoven. Rare missense mutations in abca7 might increase alzheimer's disease risk by plasma membrane exclusion. *Acta neuropathologica communications*, 10(1):43, 2022.

[181] Nicholas N Lyssenko and Domenico Praticò. Abca7 and the altered lipidostasis hypothesis of alzheimer's disease. *Alzheimer's & Dementia*, 17(2):164–174, 2021.

[182] Nicholas N Lyssenko, Xinghua Shi, and Domenico Praticò. The alzheimer's disease gwas risk alleles in the abca7 promoter and 5'region reduce abca7 expression. *Acta Neuropathologica*, 144(3):585–587, 2022.

[183] Sudha Seshadri, Annette L Fitzpatrick, M Arfan Ikram, Anita L DeStefano, Vilmundur Gudnason, Merce Boada, Joshua C Bis, Albert V Smith, Minerva M Carrasquillo, Jean Charles Lambert, et al. Genome-wide analysis of genetic loci associated with alzheimer disease. *Jama*, 303(18):1832–1840, 2010.

[184] Denise Harold, Richard Abraham, Paul Hollingworth, Rebecca Sims, Amy Gerrish, Marian L Hamshere, Jaspreet Singh Pahwa, Valentina Moskvina, Kimberley Dowzell, Amy Williams, et al. Genome-wide association study identifies variants at clu and

picalm associated with alzheimer's disease. *Nature genetics*, 41(10):1088–1093, 2009.

[185] Laura Luukkainen, Seppo Helisalmi, Laura Kytövuori, Riitta Ahmasalo, Eino Solje, Annakaisa Haapasalo, Mikko Hiltunen, Anne M Remes, and Johanna Krüger. Mutation analysis of the genes linked to early onset alzheimer's disease and frontotemporal lobar degeneration. *Journal of Alzheimer's Disease*, 69(3):775–782, 2019.

[186] Diego Marques-Coelho, Lukas da Cruz Carvalho Iohan, Ana Raquel Melo de Farias, Amandine Flaig, Jean-Charles Lambert, and Marcos Romualdo Costa. Differential transcript usage unravels gene expression alterations in alzheimer's disease human brains. *npj Aging and Mechanisms of Disease*, 7(1):2, 2021.

[187] Elizabeth BC Glennon, Isobel J Whitehouse, J Scott Miners, Patrick G Kehoe, Seth Love, Katherine AB Kellett, and Nigel M Hooper. Bin1 is decreased in sporadic but not familial alzheimer's disease or in aging. *PloS one*, 8(10):e78806, 2013.

[188] Robert J Andrew, Pierre De Rossi, Phuong Nguyen, Haley R Kowalski, Aleksandra J Recupero, Thomas Guerbette, Sofia V Krause, Richard C Rice, Lisa Laury-Kleintop, Steven L Wagner, et al. Reduction of the expression of the late-onset alzheimer's disease (ad) risk-factor bin1 does not affect amyloid pathology in an ad mouse model. *Journal of Biological Chemistry*, 294(12):4477–4487, 2019.

[189] Pierre De Rossi, Robert J Andrew, Timothy F Musial, Virginie Buggia-Prevot, Guilian Xu, Moorthi Ponnusamy, Han Ly, Sofia V Krause, Richard C Rice, Valentine de l'Estoile, et al. Aberrant accrual of bin1 near alzheimer's disease amyloid deposits in transgenic models. *Brain Pathology*, 29(4):485–501, 2019.

[190] J Chapuis, F Hansmannel, Marc Gistelinck, A Mounier, C Van Cauwenberghe, KV Kolen, F Geller, Y Sottejeau, D Harold, P Dourlen, et al. Increased expression of bin1 mediates alzheimer genetic risk by modulating tau pathology. *Molecular psychiatry*, 18(11):1225–1234, 2013.

[191] Yoann Sottejeau, Alexis Bretteville, François-Xavier Cantrelle, Nicolas Malmanche, Florie Demiaute, Tiago Mendes, Charlotte Delay, Harmony Alves Dos Alves, Amandine Flaig, Peter Davies, et al. Tau phosphorylation regulates the interaction between bin1's sh3 domain and tau's proline-rich domain. *Acta neuropathologica communications*, 3:1–12, 2015.

[192] Pierre De Rossi, Virginie Buggia-Prévot, Benjamin LL Clayton, Jared B Vasquez, Carson Van Sanford, Robert J Andrew, Ruben Lesnick, Alexandra Botté, Carole Deyts, Someya Salem, et al. Predominant expression of alzheimer's disease-associated bin1 in mature oligodendrocytes and localization to white matter tracts. *Molecular neurodegeneration*, 11:1–21, 2016.

[193] Ari Sudwarts, Supriya Ramesha, Tianwen Gao, Moorthi Ponnusamy, Shuai Wang, Mitchell Hansen, Alena Kozlova, Sara Bitarafan, Prateek Kumar, David Beaulieu-Abdelahad, et al. Bin1 is a key regulator of proinflammatory and neurodegeneration-related activation in microglia. *Molecular Neurodegeneration*, 17(1):1–27, 2022.

[194] Douglas P Wightman, Iris E Jansen, Jeanne E Savage, Alexey A Shadrin, Shahram Bahrami, Dominic Holland, Arvid Rongve, Sigrid Børte, Bendik S Winsvold, Ole Kristian Drange, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for alzheimer's disease. *Nature genetics*, 53(9):1276–1282, 2021.

[195] Jean-Charles Lambert, Simon Heath, Gael Even, Dominique Campion, Kristel Sleegers, Mikko Hiltunen, Onofre Combarros, Diana Zelenika, Maria J Bullido,

Béatrice Tavernier, et al. Genome-wide association study identifies variants at clu and cr1 associated with alzheimer's disease. *Nature genetics*, 41(10):1094–1099, 2009.

[196] Tapio Nuutinen, Tiina Suuronen, Anu Kauppinen, and Antero Salminen. Clusterin: a forgotten player in alzheimer's disease. *Brain research reviews*, 61(2):89–104, 2009.

[197] Miguel Calero, Agueda Rostagno, Blas Frangione, and Jorge Ghiso. Clusterin and alzheimer's disease. *Alzheimer's Disease: Cellular and Molecular Aspects of Amyloid* β, pages 273–298, 2005.

[198] Ioannis P Trougakos and Efstathios S Gonos. Oxidative stress in malignant progression: the role of clusterin, a sensitive cellular biosensor of free radicals. *Advances in cancer research*, 104:171–210, 2009.

[199] Troels Schepeler, Francisco Mansilla, Lise L Christensen, Torben F Ørntoft, and Claus L Andersen. Clusterin expression can be modulated by changes in tcf1-mediated wnt signaling. 2007.

[200] Claudia GUTACKER, Gerd KLOCK, Patrick DIEL, and Claudia KOCH-BRANDT. Nerve growth factor and epidermal growth factor stimulate clusterin gene expression in pc12 cells. *Biochemical Journal*, 339(3):759–766, 1999.

[201] Paul Wong, Daniel Taillefer, Jonathon LAKINS, Jean Pineault, Gerald Chader, and Martin Tenniswood. Molecular characterization of human trpm-2/clusterin, a gene associated with sperm maturation, apoptosis and neurodegeneration. *European Journal of Biochemistry*, 221(3):917–925, 1994.

[202] Claire Troakes, Rachel Smyth, Farzana Noor, Satomi Maekawa, Richard Killick, Andrew King, and Safa Al-Sarraj. Clusterin expression is upregulated following acute head injury and localizes to astrocytes in old head injury. *Neuropathology*, 37(1):12–24, 2017.

[203] Pablo Trindade, Brian Hampton, Alex C Manhães, and Alexandre E Medina. Developmental alcohol exposure leads to a persistent change on astrocyte secretome. *Journal of neurochemistry*, 137(5):730–743, 2016.

[204] Nayoung Kim, Jae Y Han, Gu S Roh, Hyun J Kim, Sang S Kang, Gyeong J Cho, Jae Y Park, and Wan S Choi. Nuclear clusterin is associated with neuronal apoptosis in the developing rat brain upon ethanol exposure. *Alcoholism: Clinical and Experimental Research*, 36(1):72–82, 2012.

[205] Anouk Imhof, Yves Charnay, Philippe G Vallet, Bruce Aronow, Eniko Kovari, Lars E French, Constantin Bouras, and Panteleimon Giannakopoulos. Sustained astrocytic clusterin expression improves remodeling after brain ischemia. *Neurobiology of disease*, 22(2):274–283, 2006.

[206] Byung Hee Han, Ronald B DeMattos, Laura L Dugan, Jeong Sook Kim-Han, Robert P Brendza, John D Fryer, Malca Kierson, John Cirrito, Kevin Quick, Judith AK Harmony, et al. Clusterin contributes to caspase-3–independent brain injury following neonatal hypoxia-ischemia. *Nature medicine*, 7(3):338–343, 2001.

[207] Steven S Schreiber, Georges Tocco, Imad Najm, and Michel Baudry. Seizure activity causes a rapid increase in sulfated glycoprotein-2 messenger rna in the adult but not the neonatal rat brain. *Neuroscience letters*, 153(1):17–20, 1993.

[208] J Scott Miners, Polly Clarke, and Seth Love. Clusterin levels are increased in a lzheimer's disease and influence the regional distribution of aβ. *Brain Pathology*, 27 (3):305–313, 2017.

[209] Priyanka Narayan, Angel Orte, Richard W Clarke, Benedetta Bolognesi, Sharon Hook, Kristina A Ganzinger, Sarah Meehan, Mark R Wilson, Christopher M Dobson, and David Klenerman. The extracellular chaperone clusterin sequesters oligomeric forms of the amyloid-β1- 40 peptide. *Nature structural & molecular biology*, 19(1):79–83, 2012.

[210] Nazhakaiti Palihati, Yuanhong Tang, Yajuan Yin, Ding Yu, Gang Liu, Zhenzhen Quan, Junjun Ni, Yan Yan, and Hong Qing. Clusterin is a potential therapeutic target in alzheimer's disease. *Molecular Neurobiology*, pages 1–15, 2023.

[211] Md Sahab Uddin, Md Tanvir Kabir, Mst Begum, Md Siddiqul Islam, Tapan Behl, Ghulam Md Ashraf, et al. Exploring the role of clu in the pathogenesis of alzheimer's disease. *Neurotoxicity Research*, 39(6), 2021.

[212] Aleksandra M Wojtas, Silvia S Kang, Benjamin M Olley, Maureen Gatherer, Mitsuru Shinohara, Patricia A Lozano, Chia-Chen Liu, Aishe Kurti, Kelsey E Baker, Dennis W Dickson, et al. Loss of clusterin shifts amyloid deposition to the cerebrovasculature via disruption of perivascular drainage pathways. *Proceedings of the National Academy of Sciences*, 114(33):E6962–E6971, 2017.

[213] PL McGeer, T Kawamata, and DG Walker. Distribution of clusterin in alzheimer brain tissue. *Brain research*, 579(2):337–341, 1992.

[214] Yuan Zhou, Ikuo Hayashi, Jacky Wong, Katherine Tugusheva, John J Renger, and Celina Zerbinatti. Intracellular clusterin interacts with brain isoforms of the bridging integrator 1 and with the microtubule-associated protein tau in alzheimer's disease. *PloS one*, 9(7):e103187, 2014.

[215] Lígia Ramos dos Santos, Jucimara Ferreira Figueiredo Almeida, Lúcia Helena Sagrillo Pimassoni, Renato Lírio Morelato, and Flavia de Paula. The combined risk effect among bin1, clu, and apoe genes in alzheimer's disease. *Genetics and Molecular Biology*, 43, 2020.

[216] Felix L Yeh, Yuanyuan Wang, Irene Tom, Lino C Gonzalez, and Morgan Sheng. Trem2 binds to apolipoproteins, including apoe and clu/apoj, and thereby facilitates uptake of amyloid-beta by microglia. *Neuron*, 91(2):328–340, 2016.

[217] Seonggyun Han, Kwangsik Nho, and Younghee Lee. Alternative splicing regulation of an alzheimer's risk variant in clu. *International Journal of Molecular Sciences*, 21(19): 7079, 2020.

[218] I-Fang Ling, Jiraganya Bhongsatiern, James F Simpson, David W Fardo, and Steven Estus. Genetics of clusterin isoform expression and alzheimer's disease risk. *PloS one*, 7(4):e33923, 2012.

[219] Hisamaru Hirai, Yoshiro Maru, Koichi Hagiwara, Junji Nishida, and Fumimaro Takaku. A novel putative tyrosine kinase receptor encoded by the eph gene. *Science*, 238(4834):1717–1720, 1987.

[220] Minerva M Carrasquillo, Olivia Belbin, Talisha A Hunter, Li Ma, Gina D Bisceglio, Fanggeng Zou, Julia E Crook, V Shane Pankratz, Sigrid B Sando, Jan O Aasly, et al. Replication of epha1 and cd33 associations with late-onset alzheimer's disease: a multi-centre case-control study. *Molecular neurodegeneration*, 6:1–9, 2011.

[221] Rüdiger Klein. Eph/ephrin signalling during development. *Development*, 139(22): 4105–4109, 2012.

[222] Kwok-On Lai and Nancy Y Ip. Synapse development and plasticity: roles of ephrin/eph receptor signaling. *Current opinion in neurobiology*, 19(3):275–283, 2009.

[223] Elena B Pasquale. Eph-ephrin bidirectional signaling in physiology and disease. *Cell*, 133(1):38–52, 2008.

[224] Elena B Pasquale. Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nature Reviews Cancer*, 10(3):165–180, 2010.

[225] Jae Min Shin, Moon Soo Han, Jae Hyung Park, Seung Hyeok Lee, Tae Hoon Kim, and Sang Hag Lee. The epha1 and epha2 signaling modulates the epithelial permeability in human sinonasal epithelial cells and the rhinovirus infection induces epithelial barrier dysfunction via epha2 receptor signaling. *International Journal of Molecular Sciences*, 24(4):3629, 2023.

[226] Katsuaki Ieguchi. Eph as a target in inflammation. *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, 15(2):119–128, 2015.

[227] Jianjun Ma, Zhidong Wang, Siyuan Chen, Wenhua Sun, Qi Gu, Dongsheng Li, Jinhua Zheng, Hongqi Yang, and Xue Li. Epha1 activation induces neuropathological changes in a mouse model of parkinson's disease through the cxcl12/cxcr4 signaling pathway. *Molecular neurobiology*, 58:913–925, 2021.

[228] Anamika Misra, Sankha Shubhra Chakrabarti, and Indrajeet Singh Gambhir. New genetic players in late-onset alzheimer's disease: Findings of genome-wide association studies. *The Indian journal of medical research*, 148(2):135, 2018.

[229] Wei Xu, Lan Tan, and Jin-Tai Yu. The role of picalm in alzheimer's disease. *Molecular neurobiology*, 52:399–413, 2015.

[230] Kunie Ando, Siranjeevi Nagaraj, Fahri Küçükali, Marie-Ange De Fisenne, Andreea-Claudia Kosa, Emilie Doeraene, Lidia Lopez Gutierrez, Jean-Pierre Brion, and Karelle Leroy. Picalm and alzheimer's disease: an update and perspectives. *Cells*, 11(24):3994, 2022.

[231] Domenico Azarnia Tehran, Gaga Kochlamazashvili, Niccolò P Pampaloni, Silvia Sposini, Jasmeet Kaur Shergill, Martin Lehmann, Natalya Pashkova, Claudia Schmidt, Delia Löwe, Hanna Napieczynska, et al. Selective endocytosis of ca2+-permeable ampars by the alzheimer's disease risk factor calm bidirectionally controls synaptic plasticity. *Science Advances*, 8(21):eabl5032, 2022.

[232] Zhen Zhao, Abhay P Sagare, Qingyi Ma, Matthew R Halliday, Pan Kong, Kassandra Kisler, Ethan A Winkler, Anita Ramanathan, Takahisa Kanekiyo, Guojun Bu, et al. Central role for picalm in amyloid-β blood-brain barrier transcytosis and clearance. *Nature neuroscience*, 18(7):978–987, 2015.

[233] Bilal Cakir, Yoshiaki Tanaka, Ferdi Ridvan Kiral, Yangfei Xiang, Onur Dagliyan, Juan Wang, Maria Lee, Allison M Greaney, Woo Sub Yang, Catherine DuBoulay, et al. Expression of the transcription factor pu. 1 induces the generation of microglia-like cells in human cortical organoids. *Nature Communications*, 13(1):430, 2022.

[234] Anna A Pimenova, Manon Herbinet, Ishaan Gupta, Saima I Machlovi, Kathryn R Bowles, Edoardo Marcora, and Alison M Goate. Alzheimer's-associated pu. 1 expression levels regulate microglial inflammatory response. *Neurobiology of disease*, 148: 105217, 2021.

[235] Elizabeta Gjoneska, Andreas R Pfenning, Hansruedi Mathys, Gerald Quon, Anshul Kundaje, Li-Huei Tsai, and Manolis Kellis. Conserved epigenomic signals in mice and humans reveal immune basis of alzheimer's disease. *Nature*, 518(7539):365–369, 2015.

[236] Olivia J Conway, Minerva M Carrasquillo, Xue Wang, Jenny M Bredenberg, Joseph S Reddy, Samantha L Strickland, Curtis S Younkin, Jeremy D Burgess, Mariet Allen, Sarah J Lincoln, et al. Abi3 and plcg2 missense variants as risk factors for neurodegenerative diseases in caucasians and african americans. *Molecular neurodegeneration*, 13(1):1–12, 2018.

[237] Maartje A Nieuwenhuis, Matteusz Siedlinski, Maarten van den Berge, Raquel Granell, Xingnan Li, Marijke Niens, Pieter van der Vlies, Janine Altmüller, Peter Nürnberg, Marjan Kerkhof, et al. Combining genomewide association study and lung eqtl analysis provides evidence for novel genes associated with asthma. *Allergy*, 71(12):1712–1720, 2016.

[238] Yi-Jun Xu, Ngan Pan Bennett Au, and Chi Him Eddie Ma. Functional and phenotypic diversity of microglia: implication for microglia-based therapies for alzheimer's disease. *Frontiers in aging neuroscience*, 14:896852, 2022.

[239] Alessandra Webers, Michael T Heneka, and Paul A Gleeson. The role of innate immune responses and neuroinflammation in amyloid accumulation and progression of alzheimer's disease. *Immunology and cell biology*, 98(1):28–41, 2020.

[240] Sho Takatori, Wenbo Wang, Akihiro Iguchi, and Taisuke Tomita. Genetic risk factors for alzheimer disease: emerging roles of microglia in disease pathomechanisms. *Reviews on Biomarker Studies in Psychiatric and Neurodegenerative Disorders*, pages 83–116, 2019.

[241] Angela K Hodges, Thomas M Piers, David Collier, Oliver Cousins, and Jennifer M Pocock. Pathways linking alzheimer's disease risk genes expressed highly in microglia. *Neuroimmunology and Neuroinflammation*, 8:245–268, 2021.

[242] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, 51(3):404–413, 2019.

[243] Maria Carolina Dalmasso, Luis Ignacio Brusco, Natividad Olivar, Carolina Muchnik, Claudia Hanses, Esther Milz, Julian Becker, Stefanie Heilmann-Heimbach, Per Hoffmann, Federico A Prestia, et al. Transethnic meta-analysis of rare coding variants in plcg2, abi3, and trem2 supports their general contribution to alzheimer's disease. *Translational psychiatry*, 9(1):55, 2019.

[244] Rebecca Sims, Sven J Van Der Lee, Adam C Naj, Céline Bellenguez, Nandini Badarinarayan, Johanna Jakobsdottir, Brian W Kunkle, Anne Boland, Rachel Raybould, Joshua C Bis, et al. Rare coding variants in plcg2, abi3, and trem2 implicate microglial-mediated innate immunity in alzheimer's disease. *Nature genetics*, 49(9):1373–1384, 2017.

[245] Helena Targa Dias Anastacio, Natalie Matosin, and Lezanne Ooi. Neuronal hyperexcitability in alzheimer's disease: what are the drivers behind this aberrant phenotype? *Translational Psychiatry*, 12(1):257, 2022.

[246] Kalyani B Karunakaran, Srilakshmi Chaparala, Cecilia W Lo, and Madhavi K Ganapathiraju. Cilia interactome with predicted protein–protein interactions reveals connections to alzheimer's disease, aging and other neuropsychiatric processes. *Scientific reports*, 10(1):15629, 2020.

[247] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee,

Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[248] Suijuan Zhong, Shu Zhang, Xiaoying Fan, Qian Wu, Liying Yan, Ji Dong, Haofeng Zhang, Long Li, Le Sun, Na Pan, et al. A single-cell rna-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*, 555(7697):524–528, 2018.

[249] Farzana Anowar, Samira Sadaoui, and Bassant Selim. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378, 2021.

[250] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[251] Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728, 2019.

[252] G Seibel. Study of the diffusion of solid-state radiotracers by the gruzin method. *The International Journal of Applied Radiation and Isotopes*, 15:679–693, 1964.

[253] Donghwa Kim, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information sciences*, 477:15–29, 2019.

[254] Zhihong Chen and Bruce D Trapp. Microglia and neuroprotection. *Journal of neurochemistry*, 136:10–17, 2016.

[255] Aleksandra Deczkowska, Hadas Keren-Shaul, Assaf Weiner, Marco Colonna, Michal Schwartz, and Ido Amit. Disease-associated microglia: a universal immune sensor of neurodegeneration. *Cell*, 173(5):1073–1081, 2018.

[256] Elisabeth Rebboah, Fairlie Reese, Katherine Williams, Gabriela Balderrama-Gutierrez, Cassandra McGill, Diane Trout, Isaryhia Rodriguez, Heidi Liang, Barbara J Wold, and Ali Mortazavi. Mapping and modeling the genomic basis of differential rna isoform expression at single-cell resolution with lr-split-seq. *Genome biology*, 22(1):1–28, 2021.

[257] Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. Starsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus rna-seq data. *Biorxiv*, pages 2021–05, 2021.

[258] Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 8(4):281–291, 2019.

[259] Zizhen Yao, Cindy TJ van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

[260] Oleg Butovsky, Mark P Jedrychowski, Craig S Moore, Ron Cialic, Amanda J Lanser, Galina Gabriely, Thomas Koeglsperger, Ben Dake, Pauline M Wu, Camille E Doykan, et al. Identification of a unique tgf-β–dependent molecular and functional signature in microglia. *Nature neuroscience*, 17(1):131–143, 2014.

[261] Macarena S Aloi, Katherine E Prater, Raymond EA Sánchez, Asad Beck, Jasmine L Pathan, Stephanie Davidson, Angela Wilson, C Dirk Keene, Horacio de la Iglesia, Suman Jayadev, et al. Microglia specific deletion of mir-155 in alzheimer's disease

mouse models reduces amyloid-β pathology but causes hyperexcitability and seizures. *Journal of Neuroinflammation*, 20(1):60, 2023.

[262] Delphine Boche and Marcia N Gordon. Diversity of transcriptomic microglial phenotypes in aging and alzheimer's disease. *Alzheimer's & Dementia*, 18(2):360–376, 2022.

[263] Bailey A Loving, Maoping Tang, Mikaela C Neal, Sachi Gorkhali, Robert Murphy, Robert H Eckel, and Kimberley D Bruce. Lipoprotein lipase regulates microglial lipid droplet accumulation. *Cells*, 10(2):198, 2021.

[264] Bailey A Loving and Kimberley D Bruce. Lipid and lipoprotein metabolism in microglia. *Frontiers in physiology*, 11:393, 2020.

[265] Cornelia Roschger and Chiara Cabrele. The id-protein family in developmental and cancer-associated pathways. *Cell Communication and Signaling*, 15(1):1–26, 2017.

[266] Ziqi Tian, Wenfang Zeng, Cuihuan Yan, Qiang Li, Nan Li, Lin Ruan, Jie Li, Xiaoguang Yao, and Si Li. Srpk2 expression and beta-amyloid accumulation are associated with bv2 microglia activation. *Frontiers in Integrative Neuroscience*, 15:742377, 2022.

[267] Diego Mastroeni, Shobana Sekar, Jennifer Nolz, Elaine Delvaux, Katie Lunnon, Jonathan Mill, Winnie S Liang, and Paul D Coleman. Ank1 is up-regulated in laser captured microglia in alzheimer's brain; the importance of addressing cellular heterogeneity. *PloS one*, 12(7):e0177814, 2017.

[268] Alexandrina Pancheva, Helen Wheadon, Simon Rogers, and Thomas D Otto. Using topic modeling to detect cellular crosstalk in scrna-seq. *PLOS Computational Biology*, 18(4):e1009975, 2022.

[269] Xiaotian Wu, Hao Wu, and Zhijin Wu. Penalized latent dirichlet allocation model in single-cell rna sequencing. *Statistics in Biosciences*, pages 1–20, 2021.

[270] Qi Yang, Zhaochun Xu, Wenyang Zhou, Pingping Wang, Qinghua Jiang, and Liran Juan. An interpretable single-cell rna sequencing data clustering method based on latent dirichlet allocation. *Briefings in Bioinformatics*, page bbad199, 2023.

[271] Hongtian Stanley Yang, Kristen D Onos, Kwangbom Choi, Kelly J Keezer, Daniel A Skelly, Gregory W Carter, and Gareth R Howell. Natural genetic variation determines microglia heterogeneity in wild-derived mouse models of alzheimer's disease. *Cell reports*, 34(6), 2021.

[272] A Sina Booeshaghi, Ingileif B Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. *bioRxiv*, pages 2022–05, 2022.

[273] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[274] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[275] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.

[276] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16 (12):1289–1296, 2019.

[277] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.

[278] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 11 2021. ISSN 0305-1048.

[279] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

[280] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[281] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[282] Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial rna sequencing. *International Journal of Oral Science*, 13(1):36, 2021.

[283] Narges Rezaie, Farilie Reese, and Ali Mortazavi. Pywgcna: A python package for weighted gene co-expression network analysis. *Bioinformatics*, 39(7):btad415, 2023.

[284] Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9):1428–1442, 2020.

[285] Narges Rezaie, Elisabeth Rebboah, Brian A Williams, Heidi Yahan Liang, Fairlie Reese, Gabriela Balderrama-Gutierrez, Louise Dionne, Laura G Reinholdt, Diane Trout, Barbara Wold, et al. Identification of robust cellular programs using reproducible lda that impact sex-specific disease progression in different genotypes of a mouse model of ad. *bioRxiv*, pages 2024–02, 2024.

[286] David Combe, Christine Largeron, Mathias Géry, and Előd Egyed-Zsigmond. I-louvain: An attributed graph clustering method. In *Advances in Intelligent Data Analysis XIV: 14th International Symposium, IDA 2015, Saint Etienne. France, October 22-24, 2015. Proceedings 14*, pages 181–192. Springer, 2015.

[287] Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8:e43803, 2019.

[288] Gabriela Balderrama-Gutierrez, Heidi Liang, Narges Rezaie, Klebea Carvalho, Stefania Forner, Dina Matheos, Elisabeth Rebboah, Kim N Green, Andrea J Tenner, Frank LaFerla, et al. Single-cell and nucleus rna-seq in a mouse model of ad reveal activation of distinct glial subpopulations in the presence of plaques and tangles. *bioRxiv*, pages 2021–09, 2021.

[289] Luciano D'Adamio. Transfixed by transgenics: how pathology assumptions are slowing progress in alzheimer's disease and related dementia research. *EMBO Molecular*

Medicine, 15(11):e18479, 2023.

[290] Temitope Ayodele, Ekaterina Rogaeva, Jiji T Kurup, Gary Beecham, and Christiane Reitz. Early-onset alzheimer's disease: what is missing in research? *Current neurology and neuroscience reports*, 21:1–10, 2021.

[291] Takashi Saito, Naomi Mihira, Yukio Matsuba, Hiroki Sasaguri, Shoko Hashimoto, Sneha Narasimhan, Bin Zhang, Shigeo Murayama, Makoto Higuchi, Virginia MY Lee, et al. Humanization of the entire murine mapt gene provides a murine model of pathological human tau propagation. *Journal of Biological Chemistry*, 294(34):12754–12765, 2019.

[292] Xianyuan Xiang, Thomas M Piers, Benedikt Wefers, Kaichuan Zhu, Anna Mallach, Bettina Brunner, Gernot Kleinberger, Wilbur Song, Marco Colonna, Jochen Herms, et al. The trem2 r47h alzheimer's risk variant impairs splicing and reduces trem2 mrna and protein in mice but not in humans. *Molecular neurodegeneration*, 13:1–14, 2018.

[293] Seonggyun Han, Yirang Na, Insong Koh, Kwangsik Nho, and Younghee Lee. Alternative splicing regulation of low-frequency genetic variants in exon 2 of trem2 in alzheimer's disease by splicing-based aggregation. *International Journal of Molecular Sciences*, 22(18):9865, 2021.

[294] Takato Suzuki, Kyoko Nishiyama, Koji Kawata, Kotaro Sugimoto, Masato Isome, Shigeo Suzuki, Ruriko Nozawa, Yoko Ichikawa, Yoshihisa Watanabe, and Tatsuo Suzutani. Effect of the lactococcus lactis 11/19-b1 strain on atopic dermatitis in a clinical test and mouse model. *Nutrients*, 12(3):763, 2020.

[295] Richard A Miller, James M Harper, Robert C Dysko, Stephen J Durkee, and Steven N Austad. Longer life spans and delayed maturation in wild-derived mice. *Experimental biology and medicine*, 227(7):500–508, 2002.

[296] Keiji Mochida, Ayumi Hasegawa, Naoki Otaka, Daiki Hama, Takashi Furuya, Masaki Yamaguchi, Eri Ichikawa, Maiko Ijuin, Kyuichi Taguma, Michiko Hashimoto, et al. Devising assisted reproductive technologies for wild-derived strains of mice: 37 strains from five subspecies of mus musculus. *PloS one*, 9(12):e114305, 2014.

[297] D Kuriakose and Z Xiao. The collaborative cross mouse–a powerful tool for studying complex genetic conditions of neurogenesis and brain development. *Stem Cell Res Int, 5 (2), 54*, 63, 2022.

[298] Darren J Burgess. Genomics: Next regeneration sequencing for reference genomes. *Nature Reviews Genetics*, 19(3):125–126, 2018.

[299] Martin O Pollard, Deepti Gurdasani, Alexander J Mentzer, Tarryn Porter, and Manjinder S Sandhu. Long reads: their purpose and place. *Human molecular genetics*, 27 (R2):R234–R241, 2018.

[300] Vivien Marx. Method of the year: long-read sequencing. *Nature Methods*, 20(1):6–11, 2023.

[301] M Brandon Titus, Adeline W Chang, and Eugenia C Olesnicky. Exploring the diverse functional and regulatory consequences of alternative splicing in development and disease. *Frontiers in Genetics*, page 2380, 2021.

[302] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[303] Joseph M Replogle, Thomas M Norman, Albert Xu, Jeffrey A Hussmann, Jin Chen, J Zachery Cogan, Elliott J Meer, Jessica M Terry, Daniel P Riordan, Niranjan Srinivas, et al. Combinatorial single-cell crispr screens by direct guide rna capture and targeted sequencing. *Nature biotechnology*, 38(8):954–961, 2020.