Spring 5-15-2019

Integrative Analysis to Investigate Complex Interaction in Alzheimer’s Disease

Zeran Li
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Neurosciences

Dissertation Examination Committee:
Carlos Cruchaga, Chair
Sharlee Climer
Joseph Corbo
Christina Gurnett
Oscar Harari
Celeste Karch
John Rice

Integrative Analysis to Investigate Complex Interaction in Alzheimer's Disease

by
Zeran Li

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2019
St. Louis, Missouri

Table of Contents

List of Figures v

List of Tables viii

List of Abbreviations x

Acknowledgments xiii

Abstract xvi

Chapter 1: Introduction and Overview 1

1.1 The Alzheimer’s Disease discovery and its impact nowadays 2

1.2 Cognitive deficits as the primary clinical symptoms and related measurements 3

1.3 Diagnostic evidence from genetic, imaging, blood and CSF biomarkers 4

1.3.1 Genetics 4

1.3.2 CSF and plasma biomarkers 8

1.3.3 Imaging 10

1.4 Neuropathological verification for postmortem assessment of AD 13

1.5 Relation to other neurodegenerative disorders 15

1.6 Dissertation Overview 16

Chapter 2: Genetic variants associated with Alzheimer’s disease confer different cerebral cortex cell-type population structure 19

2.1 Abstract 20

2.2 Introduction 21

2.2.1 Altered cellular composition confounds downstream transcriptomic analysis 21

2.2.2 Digital deconvolution approach 21

2.3 Methods 23

2.3.1 Subjects and Samples 23

2.3.2 RNA-Seq QC and Alignment 30

2.3.3 Expression quantification 55

2.3.4 Reference panel 55

2.3.5 Inference of the cellular population structure 59

2.3.6 Accuracy and Robustness Evaluation 62

2.3.7 Statistical Analysis 65

2.4 Results 69

2.4.1 Study design 69

2.4.2 Development of a reference panel to estimate brain cellular population structure 73

2.4.3 Optimization, validation and accuracy estimation of the reference panel and digital deconvolution method 78

2.4.4 Deconvolution of bulk RNA-Seq of non-demented and AD brains shows a characteristic signature for neurodegeneration 81

2.4.5 The cellular population structure differs between ADAD vs LOAD........ 89

2.4.6 Specific genetic variants confer a distinctive cell composition profile........ 94

2.5 Discussion 101

2.6 Conclusions 103

Chapter 3: The *TMEM106B* rs1990621 protective variant is associated with increased neuronal proportion........ 105

3.1 Abstract 106

3.2 Introduction 107

3.2.1 Alzheimer's disease in the context of multi-cell type interactions 107

3.2.2 Cell type composition inferred from bulk RNA-Seq deconvolution 109

3.2.3 Use cell type composition in cell type QTL (cQTL) to identify novel loci for AD risk 110

3.3 Methods 111

3.3.1 Study participants 111

3.3.2 Standard protocol approvals, registrations and patient consents........ 112

3.3.3 Data collection and generation 114

3.3.4 Data QC and preprocessing 114

3.3.5 Data analysis 119

3.3.6 Data availability 120

3.4 Results 120

3.4.1 Study design 120

3.4.2 *TMEM106B* variants associated with neuronal proportion 123

3.4.3 Neuronal protective effect of *TMEM106B* variants observed in neurodegenerative disorders and normal aging participants........ 128

3.4.4 Functional annotation of rs1990621........ 129

3.4.5 The impact of other neurodegenerative risk loci on neuronal proportion 131

3.5 Discussions 134

Chapter 4: System biology approaches revealed transcriptomic profiles of *TREM2* and *PSEN1* 138

4.1 Abstract 139

4.2 Introduction 141

4.2.1 From polygenic to omnigenic - a network interpretation 141

4.2.2 Network analysis as a powerful tool 142

4.2.3 Basic concepts in gene co-expression networks 144

4.2.4 Network analysis in Alzheimer's disease 147

4.3 Methods 149

4.3.1 Samples 149

4.3.2 Data processing and quality control 150

4.3.3 Repeated measures correlation 152

4.3.4 Network construction 152

4.3.5 Network robustness evaluation 154

4.3.6 Network functional annotation 155

4.4 Results 155

4.4.1 Study design 155

4.4.2 *MS4A* gene clusters are associated with *TREM2* in AD sporadic and control network analysis 158

4.4.3 Disrupted and acquired genes identified in network module of *PSEN1* mutation carriers 168

4.5 Discussion 175

Chapter 5: Conclusions and future directions 178

5.1 Dissertation work contributed to AD research 179

5.2 Aging, proteinopathy, and neurodegeneration 180

5.3 Future directions in developing therapies for AD 183

References 185

Curriculum Vitae 205

List of Figures

Chapter 1: Introduction and Overview

Figure 1.1: Amyloid cascade hypothesis……………………………………………………….5

Figure 1.2: AD gene risk allele frequency and risk effect……. ………………..…………….8

Figure 1.3: APP protein cleavage and amyloid β proteins …………………………………….9

Figure 1.4: Tau isoforms ………………………………….. ………………...………….10

Figure 1.5: Amyloid β PET radiotracer imaging ……………….. ………………...…………11

Figure 1.6: Tau radiotracer imaging ...……………………….. ………………...…………13

Chapter 2: Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure

Figure 2.1: Sequence quality check.………………………….. ………………...…………37

Figure 2.2: Per tile sequence quality ……………...………….. ………………...…………38

Figure 2.3: Per sequence quality scores……………………………………………………….39

Figure 2.4: Per base sequence content ……………………….. ………………...…………41

Figure 2.5: Per base N content ……………………………….. ………………...………42

Figure 2.6: Sequence length distribution ……………………………….. ………………...43

Figure 2.7: Adapter content ………………………………….. ………………...…………45

Figure 2.8: Sequence duplication levels ……………………………….. …………………47

Figure 2.9: Per sequence GC content …………………………….. …………………....49

Figure 2.10: Overrepresented sequences ……………………………….. ………………...50

Figure 2.11: Kmer Content ……………………………………………………………...51

Figure 2.12: *PSEN1* delE9 Sashimi Plot using IGV ………………………………...………..53

Figure 2.13: GWAS PCA of ethnicity check …………………………………...………….54

Figure 2.14: PCA of samples included in the reference panel …………………………..58

Figure 2.15: Leave-one-out evaluation of reference panel ……………………………………...61

Figure 2.16: Chimeric library deconvolution simulation …………………………………….. ...63

Figure 2.17: Covariate correlation ………………………………….. …………………...…..66

Figure 2.18: Transcriptome-wise principle component analysis ………………………………..67

Figure 2.19: Top 60 Genes for PC1 and PC2 heatmap …………………………...…………..68

Figure 2.20: Study design ……………………………..………………...……………71

Figure 2.21: Comparison of cell proportions estimated from full-depth and down-sampled RNA-Seq data …………………………………...……………………………76

Figure 2.22: Cellular population structure of cell-type specific samples …………………...…..79

Figure 2.23: Microglia and oligodendrocyte proportions inferred from RNA-Seq of Mayo Clinic and MSBB studies …………………...………….……………………………...83

Figure 2.24: Cellular population structure for AD and Pathological Aging (PA) ………………84

Figure 2.25: Cell-type distributions of the samples included in the Mayo Clinic and Mount Sinai Brain Bank ……………………………………...……………………………86

Figure 2.26: Neurons and astrocytes distributions for the brains included in the Mount Sinai Brain Bank stratified by CDR and Braak staging ......…………………...…………87

Figure 2.27: Neuron and astrocyte distributions from the DIAN and Knight-ADRC brains …...93

Figure 2.28: Effect of the *APOE ε4* allele and *TREM2* coding variants on the cellular population structure ……………………………………………………………...……………..97

Figure 2.29: Neurons and astrocytes distributions for samples included in the Mayo Clinic and Mount Sinai Brain Bank (MSBB) stratified by *APOE ε4* allele……………………98

Chapter 3: The *TMEM106B* rs1990621 protective variant is associated with increased neuronal proportion

Figure 3.1: Genomic PCA analysis ……………………………………………………….115

Figure 3.2: Genomic IBD analysis ………………………………….. …………………...…116

Figure 3.3: Major CNS cell type proportions derived from RNA-Seq datasets ……………....118

Figure 3.4: Study design ……………………………….. ………………...…………...122

Figure 3.5: Discovery and replication phases Manhattan and QQ plots ………………………124

Figure 3.6: SNP-based and gene-based meta-analysis …………………………………..….125

Figure 3.7: Meta-Tissue analysis results of rs1990621……………………………………...126

Figure 3.8: Variant rs1990621 functional annotation and local plot ...……………...…………130

Figure 3.9: *TMEM106B* and *TOMM40*/*APOE* regions local plot ……………………………...133

Chapter 4: System biology approaches revealed transcriptomic profiles of *TREM2* and *PSEN1*

Figure 4.1: Network medicine from obesity to the "diseasome"………………………………143

Figure 4.2: Random, small-world, and scale-free network properties …………………………146

Figure 4.3: Small-world network generated from Watts-Strogatz model ……………………...146

Figure 4.4: TREM2 module identified in LOAD cohorts ………………………...…………147

Figure 4.5: Study design ……………………………………………………………………….157

Figure 4.6: TREM2 network soft thresholding …………………………………………...…161

Figure 4.7: TREM2 network TOM clustering and gene module assignment ………………….162

Figure 4.8: Robust WGCNA of anterior prefrontal cortex ……………………...…………..164

Figure 4.9: *TREM2-MS4A* Gene Network.……………………………………………..…….167

Figure 4.10: PSEN1 seeded network gene selection……………………………...…………168

Figure 4.11: PSEN1 network TOM clustering and gene module assignment ……………...…170

Figure 4.12: *PSEN1* is mostly expressed in oligodendrocyte …….……………...…………170

Figure 4.13: PSEN1 module genes annotation ...……………………………...………….172

Chapter 5: Conclusions and future directions

Figure 5.1: Proteinopathy model in neurodegenerative disorders ……………………………..182

List of Tables

Chapter 1: Introduction and Overview

Table 1.1: CSF biomarkers…………………………………………………………………..9

Table 1.2: Neurodegenerative disease comparison……..…………………...…………………...16

Chapter 2: Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure

Table 2.1: Demographics and disease status of cohorts from four brain bank resources……….24

Table 2.2: Demographics and AD mutation carriers of DIAN and Knight-ADRC cohorts……..26

Table 2.3: Summarized quality check results integrated with Multi-QC………………………..32

Table 2.4: Illumina TruSeq Stranded RNA HT adapter and index ……………………...……44

Table 2.5: DIAN and Knight ADRC FastQC summary…………………………………………52

Table 2.6: Reference samples for each cell type…………………………………………………56

Table 2.7: Gene markers for principal brain cell types………… ………………...…………..57

Table 2.8: Simulated chimeric tissue cell-type composition……………………………………64

Table 2.9: Evaluation of deconvolution accuracy……………………… …………………......77

Table 2.10: Comparison of the cellular population structure (AD vs. neuropath-free controls) from the brains in the Mayo Clinic and Mount Sinai Brain Bank…………………..82

Table 2.11: Comparison of the cellular proportions estimated using the method PSEA in AD and control brains from the Mayo and Mount Sinai Brain Bank………………………..88

Table 2.12: Cell-type proportions comparison of subjects diagnosed with Pathological Aging...88

Table 2.13: Effect of AD in the neuronal and astrocytic proportions in distinct cerebral cortex areas……………………………………………………………………………89

Table 2.14: Cellular population structure altered in the parietal lobe from AD brains in the DIAN study and Knight-ADRC brain bank……………………………………………...91

Table 2.15: Comparison of the cellular proportions estimated using the method PSEA in AD and control brains from the DIAN and Knight-ADRC ……………………………….....92

Table 2.16: Gene specific cellular proportion analysis for Knight-ADRC and Mount Sinai Brain Bank studies ……………………………………………………………………..99

Chapter 3: The *TMEM106B* rs1990621 protective variant is associated with increased neuronal proportion

Table 3.1: Demographic information for cohorts included in the study………………………..113

Table 3.2: General information of seven studies evolved in the analysis………………………113

Table 3.3: rs1990621 (chr7:12283873) major allele C is significantly associated with decreased neuronal proportions………………………………………………………......123

Table 3.4: rs1990621 (chr7:12283873) major allele C is significantly associated with decreased neuronal proportions in AD, Control, and other non-AD neurodegenerative disorders……………………………….. ……………………...…………....129

Table 3.5: Neuronal proportion cQTL p-values were reported for variants previously identified in AD risk (by Lambert et al.), FTD risk (by Ferrari et al.), and FTD-TDP risk (by Van Deerlin et al.) studies………………………..…………………………………132

Chapter 4: System biology approaches revealed transcriptomic profiles of *TREM2* and *PSEN1*

Table 4.1: TREM2-MS4A Study Demographic………………………………………………..149

Table 4.2: PSEN1 Study Demographic……………………………………………………...150

Table 4.3: Robust WGCNA simulation of TREM2 module…………………………………....165

Table 4.4: PSEN1 module disrupted and emerged genes pathway analysis………...…………174

List of Abbreviations

Aβ	Amyloid β
AD	Alzheimer’s Disease
ADAD	Autosomal Dominant Alzheimer’s Disease
ALS	Amyotrophic lateral sclerosis
AMP-AD	Advanced Medicines Partnership - Alzheimer’s Disease
APC	Anterior prefrontal cortex
BM9	Dorsal lateral prefrontal cortex
BM10	Anterior prefrontal cortex
BM22	Superior temporal gyrus
BM24	Ventral anterior cingulate cortex
BM36	Parahippocampal gyrus
BM44	Inferior frontal gyrus
CB	Cerebellum
CDR	Clinical Dementia Rating
CERAD	The Consortium to Establish a Registry for Alzheimer’s Disease
cQTL	cell type QTL
CSF	Cerebrospinal fluid
CTX	Cortex
DIAN	Dominantly Inherited Alzheimer Network
DLB	Dementia with Lewy bodies
DLPFC	Dorsal lateral prefrontal cortex
DSA	Digital Sorting Algorithm
EOAD	Early-onset Alzheimer’s disease

FCX	Frontal cortex
FTD	Frontotemporal dementia
FTLD	Frontotemporal lobar degeneration
GO term	Gene Ontology Enrichment Analysis
IFG	Inferior frontal gyrus
IGAP	The International Genomics of Alzheimer's Project
iPSC	Induced pluripotent stem cell
Knight ADRC	Charles F. and Joanne Knight Alzheimer's Disease Research Center
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOAD	Late-onset Alzheimer's disease
meanProfile	Implementation of method Population-Specific Expression Analysis
MSBB	Mount Sinai Brain Bank
MSSM	Mount Sinai School of Medicine
NIA	National institute of aging
NPCs	Neural progenitor cells
PA	Pathological aging
PAR	Parietal cortex
PC	Principal component
PCA	Principal component analyses
PD	Parkinson's disease
PHG	Parahippocampal gyrus
PMI	Post-mortem index
PSEA	Population-Specific Expression Analysis
PSP	Progressive Supranuclear Palsy
ptau	Phosphorylated tau

QC	Quality control
RIN	RNA integrity number
rmcorr	Repeated measures correlation
RMSE	Root-mean-square error
SCZ	Schizophrenia
SNP	Single nucleotide polymorphism
ssNMF	Semi-supervised non-negative matrix factorization
STG	Superior temporal gyrus
TC	Temporal cortex
TIN	Transcript integrity number
TOM	Topological overlap matrix
TRAP-seq	Translating ribosome affinity purification sequencing
WGCNA	Weighted correlation network analysis

Acknowledgments

Firstly, I would like to express my gratitude to my advisor Dr. Carlos Cruchaga for his support of my PhD study and related research. I thank him for his patience, encouragement, insights, and high standard that lead me through my PhD training. I appreciate all the learning and training opportunities he has provided to me. His guidance and mentorship helped me in all the time of research and writing of this thesis.

I would also like to thank Dr. Oscar Harari and Dr. Sharlee Climer, who have spent great amount of time and efforts teaching and guiding me through my thesis work, especially Dr. Harari who mentored me in all three research projects documented in this thesis. I would also like to thank the rest of my thesis committee: Dr. Joseph Corbo, Dr. Christina Gurnett, Dr. Celeste Karch, and Dr. John Rice for their insightful comments and encouragement and also for the hard questions which incented me to think through my research from various perspectives. I thank all my committee members for helping me clarify my research goals and resolve scientific questions.

I thank all my fellow lab mates for their companion on this journey. I thank them for rigorous discussions in science, technical supports in work, daily small talks in life that I have had during my PhD training at Washington University School of Medicine. I thank them for being with me and supporting me through the high and low times throughout my years here.

I would also like to thank the Neuroscience Program of the Division of Biology and Biomedical Science and the Department of Psychiatry, especially Dr. Lawrence Snyder, Dr. Erik Herzog, Sally Vogt, Dr. Deanna Barch, Sridhar Kandala, and Dr. John Pruett for supporting me in going through hard times and making the transition.

Finally, I would like to thank my family and friends for their continuous supports, especially my parents. They are the backbone of my life. Without their love and supports I would not be able to complete this PhD training.

Zeran Li

Washington University in St. Louis

May 2019

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

Integrative Analysis to Investigate Complex Interaction in Alzheimer's Disease

by

Zeran Li

Doctor of Philosophy in Biology and Biomedical Sciences

Neurosciences

Washington University in St. Louis, 2019

Carlos Cruchaga, Chair

Alzheimer's disease (AD) is a neurodegenerative disorder featuring progressive cognitive and functional deficits. Pathologically, AD is characterized by tau and amyloid β protein deposition in the brain. As the sixth leading cause of death in the U.S., the disease course usually last from 7 to 10 years on average before the consequential death. In 2019 there are estimated 5.8 million Americans living with AD affecting 16 million family members. At certain stage of the disease course, patients with inability of maintaining their daily functioning highly depend on caregivers, primarily family caregivers, that incur estimated 18.4 billion unpaid hours of cares, which is equivalent to 232 billion dollars. These huge economic burdens and inevitable emotional distress on the family and the society would also increase as the number of AD affected population could triple by 2050.

Altered cellular composition is associated with AD progression and decline in cognition, such as neuronal loss and astrocytosis, which is a key feature in neurodegeneration but has often been overlooked in transcriptome research. To explore the cellular composition changes in AD, I developed a deconvolution pipeline for bulk RNA-Seq to account for cell type specific effects in brain tissues. I found that neuronal and astrocyte relative proportions differ between healthy and

diseased brains and also among AD cases that carry specific genetic risk variants. Brain carriers of pathogenic mutations in *APP*, *PSEN1*, or *PSEN2* presented lower neuron and higher astrocyte relative proportions compared to sporadic AD. Similarly, the *APOE ε4* allele also showed decreased neuronal and increased astrocyte relative proportions compared to AD non-carriers. In contrast, carriers of variants in *TREM2* risk showed a lower degree of neuronal loss compared to matched AD cases in multiple independent studies. These findings suggest that genetic risk factors associated with AD etiology have a specific effect on the cellular composition of AD brains. The digital deconvolution approach provides an enhanced understanding of the fundamental molecular mechanisms underlying neurodegeneration, enabling the analysis of large bulk RNA-sequencing studies for cell composition. It also suggests that correcting for the cellular structure when performing transcriptomic analysis will lead to novel insights of AD.

With deconvolution methods to delineate cell population changes in disease condition, it would help interpret transcriptomics results and reveal transcriptional changes in a cell type specific manner. One application demonstrated in this dissertation work is to use cell type proportion as quantitative trait to identify genetic factors associated with cellular composition changes. I performed cell type QTL analysis and identified a common pathway associated with neuronal protection underlying aging brains in the presence or absence of neurodegenerative disease symptoms. A protective variant of *TMEM106B*, which was previously identified with a protective effect in FTD, was identified to be associated with neuronal proportion in aging brains, suggesting a common pathway underlying neuronal protection and cognitive reservation in elderly. This extended analysis yield from deconvolution results demonstrated one promising direction of using deconvolution followed by cell type QTL analysis in identifying new genes or pathways underlying neurodegenerative or aging brains.

To understand the complexity of the brain under disease condition, network analysis as a large-scale system-level approach provides unbiased and data-driven view to identify gene-gene interactions altered by disease status. Using network analysis, I replicated and reconfirmed the co-expression pattern between *MS4A* gene cluster and *TREM2* in sporadic AD, from which further evidence was inferred from Bayesian network analysis to show that *MS4A4A* might be a potential regulator of *TREM2* that is validated by *in-vitro* experiments. In Autosomal Dominant AD (ADAD) cohort, disrupted and acquired genes were identified from *PSEN1* mutation carriers. Among these genes, previously identified AD risk genes and pathways were revealed along with novel findings. These results demonstrated the great potential of applying network approach in identifying disease associated genes and the interactions among them.

To conclude the dissertation work from methodological, empirical, and theoretical levels, deconvolution pipeline for bulk RNA-Seq, cell type QTL analysis, and network analysis approaches were applied to understand transcriptome changes underlying disease etiology. From which previous AD related findings were replicated that validated the methods, and novel genes and pathways were identified as potential new therapeutic targets. Based on prior knowledge and empirical evidence observed from this dissertation work, a model is proposed to explain how genetic factors are assembled as a highly interconnected interactome network to affect proteinopathy observed in neurodegenerative disorders, that cause cellular composition changes in the brain, which ultimately leads to cognitive and functional deficits observed in AD patients.

Chapter 1: Introduction and Overview

1.1 The Alzheimer’s Disease discovery and its impact nowadays

In 1901, Alois Alzheimer, a German psychiatrist and a lecturer at the Munich University Hospital received a patient case of a 51 years female named Auguste D[129]. She was sent by her husband describing her symptoms as paranoid, progressive sleep and memory disturbance, aggressiveness, crying, and confusion. This lady was admitted to the Community Psychiatric Hospital at Frankfurt am Main, and remained impatient until her death in 1906. The brain material of her autopsy was sent to Alzheimer for examination, from which he observed distinctive plaques and neurofibrillary tangles in the histology, which were later identified as pathological hallmarks of Alzheimer’s disease. In 1906, Alzheimer presented his finding of this “peculiar” dementia case in the 37th Meeting of South-West German Psychiatrists in Tubingen. Although at the meeting it did not spur much interests from the audience, Alzheimer’s finding was included as “Alzheimer's disease” in the 3th edition of his coworker Emil Kraepelin’s text ‘Psychiatrie’ in 1910[129].

For the past one hundred years since its first diagnosis, Alzheimer’s Disease (AD) is like a shadow that never leaves, and it also grows larger as human life expectancy increases as age being its most important risk factor. In 2019, there are estimated 5.8 million Americans living with AD affecting 16 million family members. At certain stage of the disease course, patients with inability of maintaining their daily functioning highly depend on caregivers, primarily family caregivers, that incur estimated 18.4 billion unpaid hours of cares, which is equivalent to 232 billion dollars[15, 69]. These huge economic burdens and inevitable emotional distress on the family and the society would also increase as the number of AD affected population could triple by 2050[121].

As the sixth leading cause of death in the U.S., the disease course usually last from 7 to 10 years on average before the consequential death[133]. Due to progressive neuronal death in the affected brain regions, apart from cognitive functions it will also disable movement functions of the patients with the results being long term bed-bound and later having swallowing problems that ultimately lead to organ failure or lethal aspiration pneumonia[15].

1.2 Cognitive deficits as the primary clinical symptoms and related measurements

Among all the neurodegenerative disorder that would result in dementia, Alzheimer's Disease is the most common form[23]. Impaired declarative memory is usually the first noticeable sign but sometimes it could also be other executive functions such as planning or problem-solving skills. More detailed cognitive deficit rating scales, such as the clinical dementia rating[189] (CDR), have been developed to further categorize cognitive performance for both clinical and research purposes. CDR = 0 is considered as normal without dementia; CDR = 0.5 is very mild dementia; CDR = 1 is mild dementia; CDR = 2 is moderate dementia; CDR = 3 is severe dementia.

As our knowledge about neurodegenerative disorders accumulates, neuropsychological tests also evolve to optimize their diagnostic and prognostic utilities. Efforts have been spent on ensuring the test construct validity and stability with appropriate norms[88]. Test construct validity hinges heavily on an accurate and unambiguous design that projects the clinical batteries to their designated cognitive domains to ensure they are measuring what they are designed to measure. The assessment criteria of AD developed by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) integrates evidence from multiple modalities including clinical, behavioral, demographical, neuropsychological, neuropathological, neuroimaging, family history, and postmortem materials to standardize and ensure accurate diagnosis[91].

1.3 Diagnostic evidence from genetic, imaging, blood and CSF biomarkers

Symptoms, such as memory loss or difficulties with executive functions, are usually what AD patients first complained to the physicians, however, pathologic changes have developed decades (10-20 years) before cognitive symptoms onsets[138, 211]. To capture preclinical stages and the disease development trajectory and dynamics, a variety of biomarkers or diagnostic evidence based on genetic, CSF, blood, and imaging biomarkers have been developed to facilitate early detection and differentiation among different dementia sub forms.

1.3.1 Genetics

Amyloid β is one of the two pathological hallmarks of AD, which was first isolated from a late onset AD (LOAD) patient[105]. Later, the same authors isolated cerebrovascular amyloid protein from Down's syndrome, a disease caused by the presence of all or part of a third copy of chromosome 21(trisomy 21). Because of the close resemblance of the two proteins and the cerebrovascular amyloid protein discovery in Down's syndrome, the authors accurately predicted that the amyloid β gene might be located on chromosome 21[104]. Three amyloid β related genes *amyloid-beta precursor protein* (*APP), presenilin 1* (*PSEN1*), *presenilin 2* (*PSEN2*) were identified associated with familial AD with in a Mendelian dominance pattern and high penetrance[148]. Then an amyloid cascade was hypothesized (**Figure 1.1**) suggesting over accumulation or failed clearance of amyloid β is the central event in the pathogenesis of AD, which led to neuronal and synaptic dysfunction, and ultimately to cognitive deficits[106]. Many statements surround the amyloid centered theory have been fulfilled but one issue raised regarding whether amyloid β being the cause or the consequence of AD. If amyloid accumulation is the leading cause, then drugs targeting clearance of amyloid should ameliorate the symptom. However, so far, none of the drug developments targeting amyloid pathway is successful, which may suggest the alternative hypothesis that amyloid accumulation might be the consequence or by-product of AD.

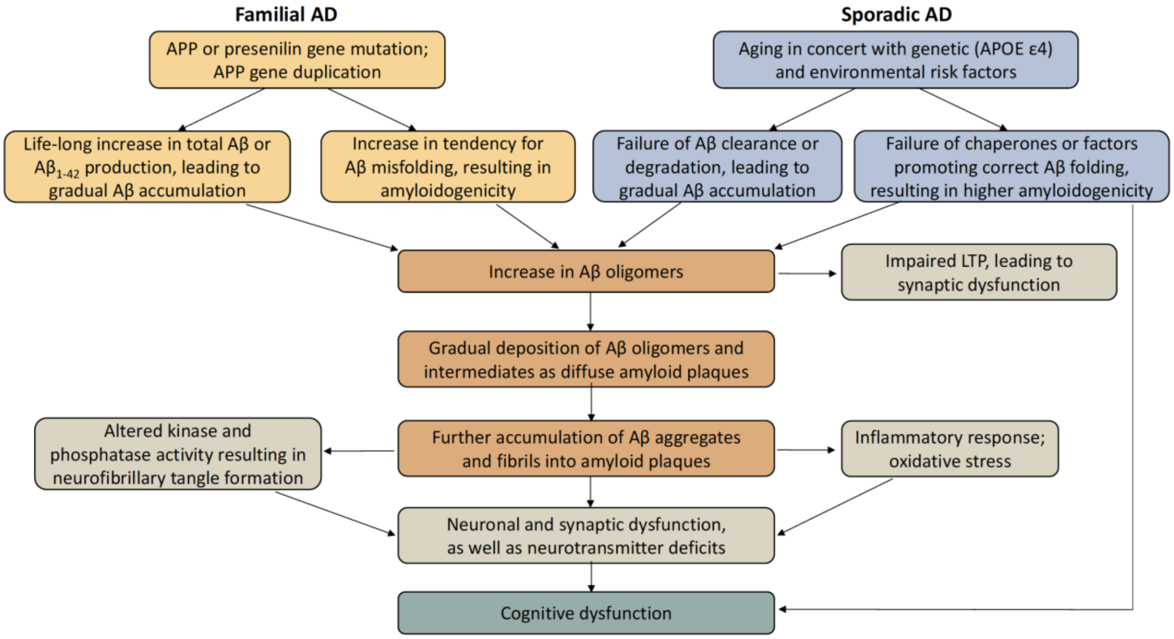

Familial AD
APP or presenilin gene mutation; APP gene duplication
Sporadic AD
Aging in concert with genetic (APOE ε4) and environmental risk factors
Life-long increase in total Aβ or Aβ1-42 production, leading to gradual Aβ accumulation
Increase in tendency for Aβ misfolding, resulting in amyloidogenicity
Failure of Aβ clearance or degradation, leading to gradual Aβ accumulation
Failure of chaperones or factors promoting correct Aβ folding, resulting in higher amyloidogenicity
Increase in Aβ oligomers
Impaired LTP, leading to synaptic dysfunction
Gradual deposition of Aβ oligomers and intermediates as diffuse amyloid plaques
Altered kinase and phosphatase activity resulting in neurofibrillary tangle formation
Further accumulation of Aβ aggregates and fibrils into amyloid plaques
Inflammatory response; oxidative stress
Neuronal and synaptic dysfunction, as well as neurotransmitter deficits
Cognitive dysfunction

Figure 1.1 Amyloid cascade hypothesis. Image reproduced from Blennow et al.[29] with permission.

A second doubt surround the amyloid pathway comes from the differences in inheritance mechanism between sporadic and familial AD. Unlike familial AD which has a clear Mendelian inheritance pattern with three major gene players and an early onset of disease manifestation, sporadic AD is late onset and attributed to complex traits with multiple risk genes located throughout the genome[162]. It seems that genes with rare variants such as *APP*, *PSEN1*, *PSEN2* in familial AD exert high risk effects to familial AD, whereas many genes with common variants exert medium or low risk effects to sporadic AD (**Figure 1.2**). For sporadic AD, *apolipoprotein E (APOE)* is the gene with the largest and dosage dependent effect[84]. *APOE* has three major alleles, protective allele ε2, common allele ε3, and risk allele ε4. In Caucasian cohorts, carriers of two ε4 alleles have increased risk of 14.9 relative to two ε3 alleles. Walking down to the risk ladder, one ε4 allele renders increased risk of 2.6 for ε2/ε4, and increased risk of 3.2 for ε3/ε4. Carriers with protective allele ε2 have reduced risk of 0.6 in both ε2/ε2 and ε2/ε3[84]. *APOE* influence LOAD risk in an amyloid dependent manner[235, 252]. The rare *TREM2* variant p.R47H (rs75932628) carriers exhibit increased AD risk by a range from 1.7-fold to 3.4-fold[112, 212]. This gene is related to microglia and immune system through neuroinflammation[79]. The involvement of immune system leads to another hypothesis for AD surround inflammation and infection with microbial triggers, for example herpes infection[219] and oral *P. gingivalis* infection[75, 203]. Instead of thinking genetic variants as disease causing factors, is it possible that the genetic vulnerability that failed to protect the brain from insult is the cause of sporadic AD? For example, the vulnerability of blood brain barrier[254] and neuroinflammation triggered by microbial, stress, or even lack of sleep[68] initiate amyloid protein accumulation, then leads to neuronal death and subsequent cognitive deficits. The shift of thinking paradigm may drive therapeutic strategy and drug design switching from amyloid clearance to anti-inflammation or boost immune resilience to insults. Apart from researches focusing on protein coding genes, investigation of non-coding RNA in neurodegeneration[232] and 3D spatial structure of genome[218] may also shed light on figuring out the mechanism underlying AD.

As shown in **Figure 1.2**, AD has a substantial but heterogeneous genetic component. The rare mutations in the *APP*, *PSEN1* and *PSEN2*[64, 238] that cause autosomal dominant AD (ADAD) only account for 1-2% of overall AD cases. There are also early-onset AD cases with unknown genetic risk factors that remain elusive. Apart from early-onset AD, the most common manifestation of AD presents late-onset (LOAD) and accounts for the majority of the cases (90-95%). Despite appearing sporadic in nature, a complex genetic architecture underlies LOAD risk. *APOE* ε4 as discussed above is the most common genetic risk factor. In addition, recent whole genome and whole exome analysis have identified rare coding variants in *TREM2*[26, 113], *PLD3*[58], *ABCA7*[65, 249] and *SORL1*[86, 226] that are associated with AD and confer risk comparable to that of carrying one *APOE* ε4 allele. Besides age at onset, the clinical presentations of LOAD and ADAD are remarkably similar with an amnestic and cognitive impairment phenotype[230, 258]. A minor fraction of cases of ADAD have additional neurological findings, sometimes also seen in LOAD[230, 258]. Twin studies have estimated that the heritability of Alzheimer's disease is about 0.74 and argued that unexplained variance is due to environmental factors[101]. So far, genetic studies have identified around 30 common and rare genetic variants that contribute to the AD phenotypes; however, these genes with disease susceptibility only explain a small proportion of the genetic heritability of the AD population. The remaining unexplained heritability has been named as missing heritability[80, 177, 178]. In **Chapter 4**, an omnigenic model will be discussed from a network perspective to explain a potential cause of the missing heritability problem.

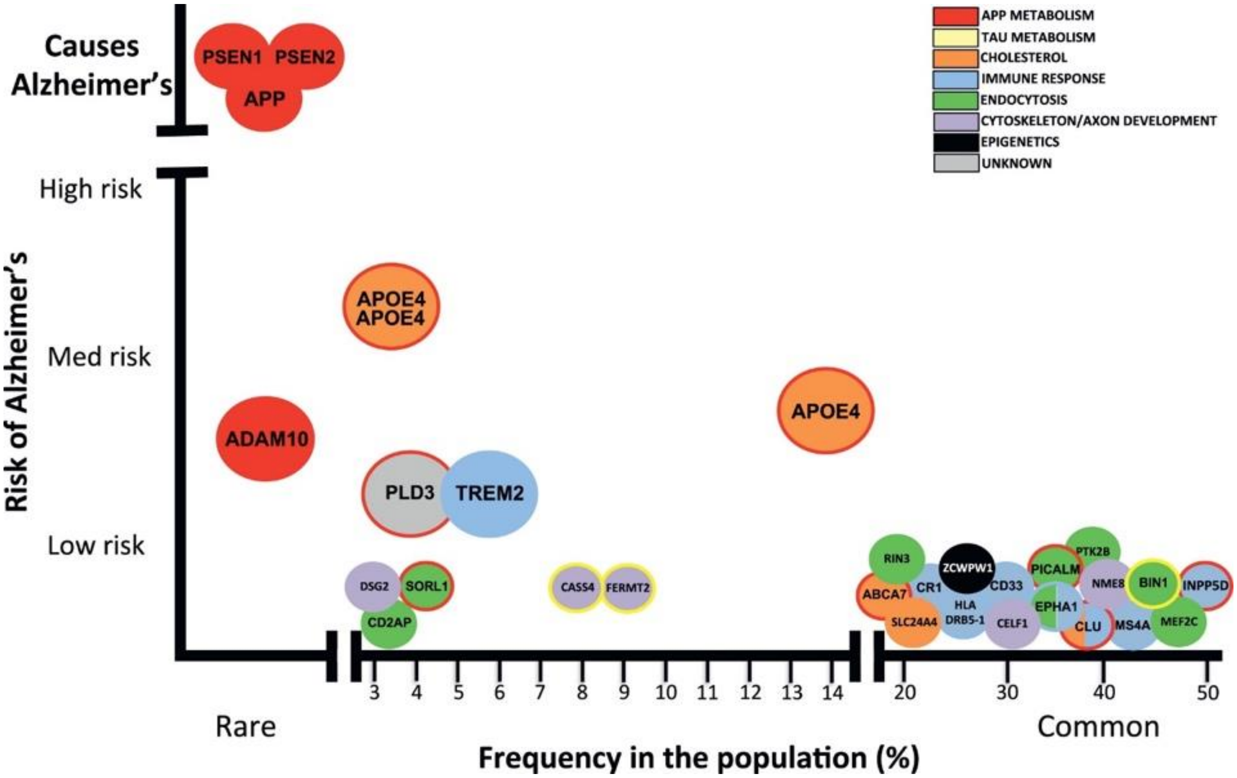


Figure 1.2 AD gene risk allele frequency and risk effect. Image from Karch and Goate[148] with permission.

1.3.2 CSF and plasma biomarkers

Biomarkers are objective measures of biological or pathogenic changes that can be used as diagnostic, prognostic, or disease progression measurement tools. Cerebrospinal fluid (CSF) is a clear and colorless fluid that the brain and spinal cord are immersed in. Due to its direct contact with the extracellular space of the brain, it is optimal as a biomarker to capture biological or pathogenic changes in the brain. CSF biomarkers can be divided to basic and core biomarkers. Basic biomarkers measure basic brain function that might be changed in AD condition but not specific to AD, which include measurements of blood brain barrier and immune system response to chronic inflammation (**Table 1.1** basic biomarker section). The core biomarkers measure AD specific molecular pathology that is specific to AD, including APP metabolism and amyloid deposition, tau phosphorylation and axonal degeneration (**Table 1.1** core biomarker section).

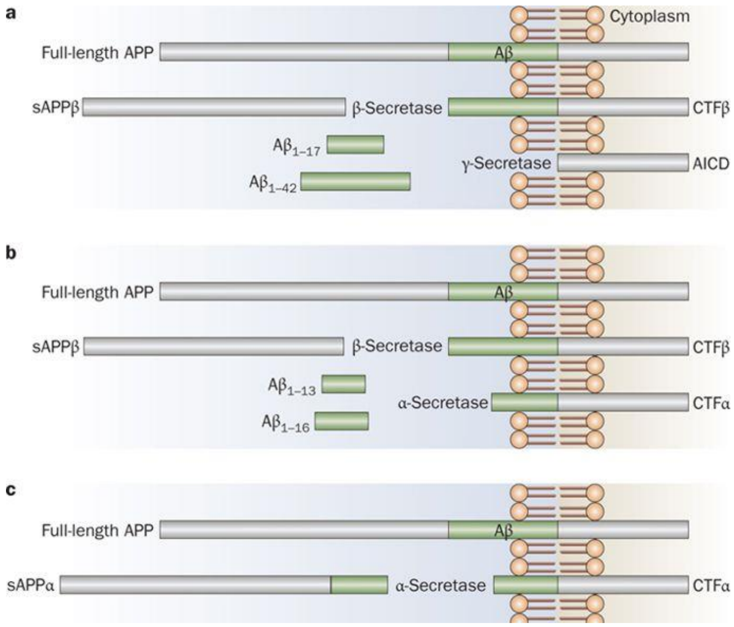


Figure 1.3 APP protein cleavage and amyloid β proteins. Image from Blennow et al.[29] with permission.

Table 1.1 CSF biomarkers

Biomarker	**Pathogenic process**	**Biomarker level change in AD**
	Basic biomarker	
CSF cell count	inflammation	unchanged
CSF: serum albumin ratio	BBB function	Pure AD: unchanged; AD with cerebrovascular pathology: increase
IgG or IgM index or oligoclonal bands	Intrathecal immunoglobin production	unchanged
	Core biomarker	
Aβ 1-42	APP metabolism and plaque formation	AD and prodromal AD: reduction
p-tau181 and p-tau231	Tau phosphorylation	AD and prodromal AD: increase
total tau	Axonal degeneration	AD and prodromal AD: increase

Amyloid β proteins of different length are cleaved from APP protein by beta-secretase and gamma-secretase (**Figure 1.3**), and $A\beta_{40}$, $A\beta_{42}$, and $A\beta_{40:42}$ ratios are primarily measured as biomarkers. Six different tau isoforms can be divided through alternative splicing from exon 2, 3, and 10 of *MAPT* gene (**Figure 1.4a**). There are a number of phosphorylation sites of threonine and serine across the tau isoforms but the commonly referred as phosphorylated tau levels are measured from Thr181 or Thr231 phosphorylation sites (**Figure 1.4b**). Other CSF biomarkers related to neuronal and synaptic proteins, for example, visinin-like protein 1 (VLP-1) and synaptosomal-associated protein 25 (SNAP-25), and oxidative stress markers such as F2-isoprostanes were also be able to differentiate AD from controls.

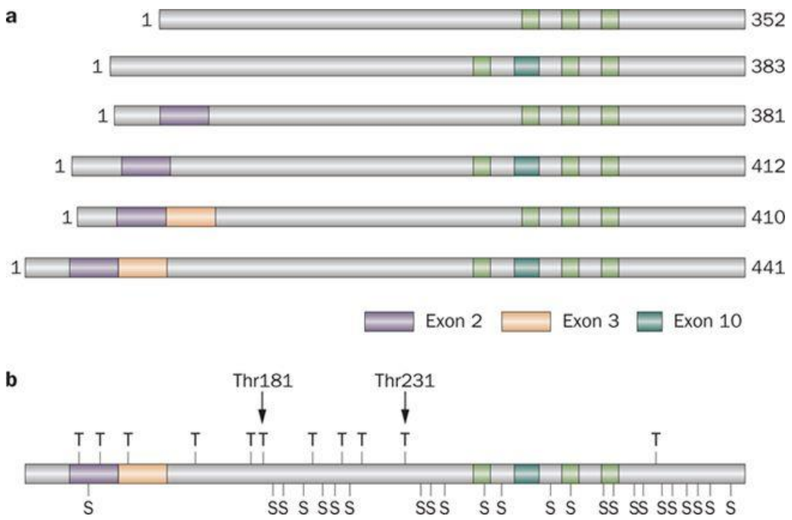


Figure 1.4 Tau isoforms. a) Tau protein isoforms and b) phosphorylation sites. Image from Blennow et al.[29] with permission.

CSF biomarkers have shown great potential in measuring pre-symptomatic changes before the plaque becomes too widespread or the proceeding of irreversible neurodegeneration. However, collection of CSF through lumbar puncture is invasive with potential risks of post-lumbar puncture headache, back discomfort or pain, bleeding, and brainstem herniation. Therefore, developments of non-invasive biomarkers, such as plasma derived biomarker, have been pursuit to look for alternative sources other than CSF, for example plasma $A\beta_{1-42}$[180], $A\beta_{40}$ to $A\beta_{42}$ ratio[109], $APP_{669–711}$ to $A\beta_{42}$ ratio[195], α 2-Macroglobulin (α2M) and complement factor H (CFH)[135], neurofilament light protein (NFL)[179].

1.3.3 Imaging

As amyloid β aggregation being directly related to AD and potentially a predictor of AD decades before cognitive deficits, Aβ PET imaging has been used clinically as a diagnosis tool and prognosis measurement[270]. Although the Aβ's role in AD is still under debate, it has been shown that amyloid β deposition proceeds neuronal loss or cerebral atrophy observed on MRI. Five different amyloid β tracers showed that the frontal, temporal and posterior cingulate cortices showed the highest retention rate for Aβ (**Figure 1.5**) that correlates with regional amyloid β plaque density and the sequence of amyloid β deposition found in post-mortem brains in sporadic AD[32]. Noticeably, the pattern of amyloid β tracer retention in familial AD mainly located in striatal region which is different from the patterns of sporadic AD. Based on the retention patterns observed, amyloid β imaging could also be used to differentiate sporadic AD from dementia with Lewy bodies (DLB) and early-onset AD from frontotemporal lobar degeneration (FTLD), because DLB exhibits a posterior retention pattern that sporadic AD does not have, and FTLD should not have C-PIB retention. However, it is worth noticing that about 25% cognitive normal elderly also have fibrillary Aβ deposition[185], which had been observed long before the amyloid β PET imaging era[216].

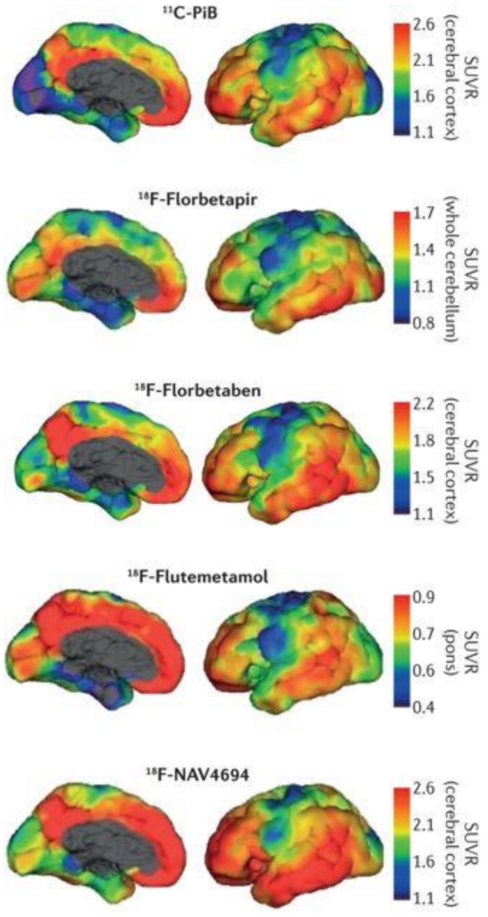


Figure 1.5 Amyloid β PET radiotracer imaging. Surface projection images obtained from five AD patients with different amyloid beta PET radiotracer. Five images showed consistent pattern with highest retention in frontal, temporal and posterior cingulate cortices representing amyloid β deposition in the brain. Image from Villemagne et al.[270] with permission.

Similar to amyloid β imaging, tau imaging also showed consistent tau tracer retention pattern (**Figure 1.6**) with post-mortem studies, besides it is more correlated with neuronal injury markers. As opposed to amyloid β imaging which focusing on total amyloid β load, regional tau distribution in terms of density and topological distribution of tau provide more information in disease progression than total tau load.

Apart from tracers based PET that requires tracer injection into the patients, gradient recalled echo MRI based approach post-processing method, Gradient Echo Contrast Imaging (GEPCI), has been developed utilizing transverse relaxation rate constant to avoid tracer injection. The GEPCI metrics showed strong correlation with both amyloid β accumulation measured from PET and cognitive performance[292]. Another non-tracer imaging technique based on functional connectivity MRI also showed success in differentiating APOE4+ from APOE4- carriers in the absence of amyloid deposition[240] suggesting early genetic effect can be measured with functional connectivity MRI. Strong evidences suggested the default mode network is strongly associated with AD[17, 111, 116, 241].

Since 2011, the diagnosis guideline for AD in the U.S. had been revised from the 1984 diagnostic criteria, which was mainly based on the clinical judgement of the patient's symptoms, to incorporate biomarker tests[15]. An A/T/N system have been proposed to integrate multiple biomarker modalities: "A" refers to β-amyloid related biomarker including amyloid PET or CSF Aβ42; "T" refers to tau related biomarker including CSF phosphor tau or tau PET; and "N" refers to non-specific neurodegeneration or neuronal injury biomarkers including ^{18}F-FDG PET, structural MRI or CSF total tau[137]. With either positive or negative binary traits defined by respective cutoffs within each category, a biomarker profile integrating multimodal measurements can be established for the subject to inform diagnosis[136]. As mentioned above, tau imaging result is highly correlated with neuronal function biomarker ^{18}F-FDG PET, thus one potential problem with ATN is that the integrative approach includes highly correlated metrics may incur a cost of redundant tests or repetitive information. Besides, the binary traits may over simplify the complexity of AD manifestation as opposed to a more quantitative approach.

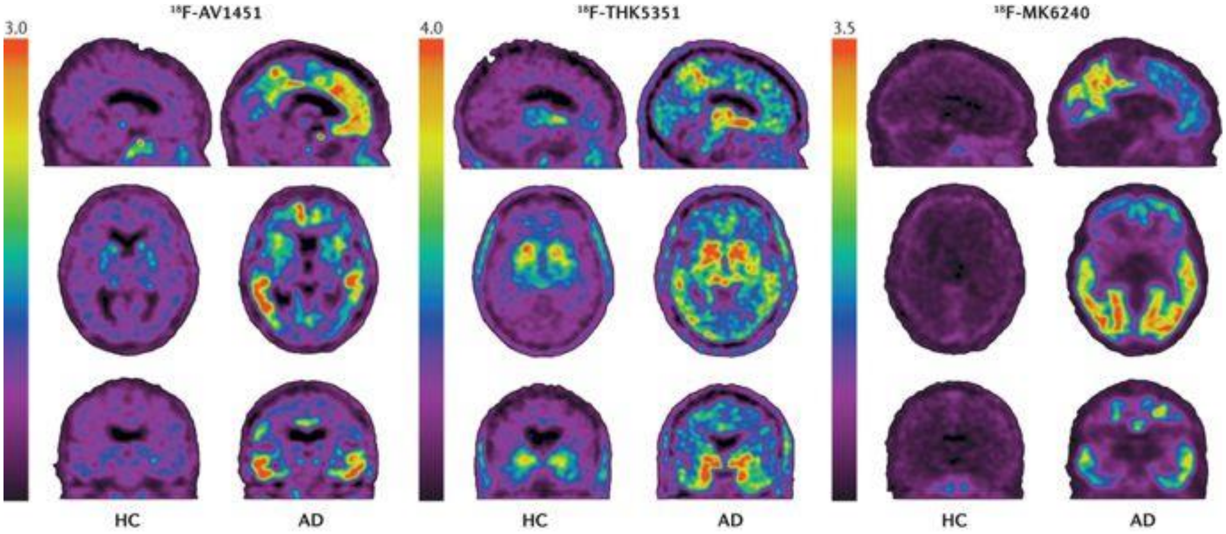


Figure 1.6 Tau radiotracer imaging. Representative PET images with three different tau radiotracers. Top row is sagittal view; center row is transverse view; bottom row is coronal view. Compared to healthy elderly controls (HC), AD patients showed tracer retention in mesial temporal, temporoparietal and posterior cingulate cortical regions. ^{18}F-THK5351-PET and ^{18}F-flortaucipir-PET in HC show 'off-target' retention in the stratum. ^{18}F-THK5351-PET also show tracer retention in the striatum. Image from Villemagne et al.[270] with permission.

1.4 Neuropathological verification for postmortem assessment of AD

Mostly for research purposes, neuropathological assessments are performed during autopsy on post mortem materials to verify the clinical diagnosis of AD. The CERAD neuropathology criteria contains gross and microscopic findings focusing on hippocampus, amygdala and various cortical regions[91, 186]. They use a semi-quantitative approach to assess frequency of senile plaques, including both neuritic plaques relative to the patent's age and diffuse plaques, neurofibrillary tangles, and others such as cerebrovascular changes. From those together with clinical history, a categorical assessment result will be derived to report the certainty of AD diagnosis, and they are definite, probable, possible, or no evidence of AD.

Other commonly used neuropathological assessment with slightly different focuses are Braak and Braak[34], Khachaturian[150], NIA-Reagan Institute[5], and the Tierney A3[259] criteria. Braak and Braak staging focuses on the distribution patterns of neurofibrillary tangles and neuropil threads[34], which is divided into six stages – stage I and II are characterized by either mild or severe alteration of the transentorhinal region; stage III and IV are marked by conspicuous changes in both transentorhinal region and entorhinal cortex; stage V and VI include destruction of all isocortical associated regions. Khachaturian[150] documented consensus criteria of AD diagnosis reached upon by the neuropathology panel at the "research workshop on the diagnosis of Alzheimer's Disease" organized by National Institute of Aging (NIA) in 1983. These criteria focus on microscopic findings in frontal, temporal, and parietal lobes, the amygdala, the hippocampal formation, the basal ganglia, the substantia nigra, the cerebellar cortex, and the spinal cord. The number of amyloid plaque and neurofibrillary tangles per field for different age ranges are specified for AD diagnosis. Another more recent consensus recommendation of postmortem diagnostic criteria for AD are proposed to reassesses the previous criteria documented in Khachaturian[150]. This meeting was led by both NIA and Nancy Reagan Institute of the Alzheimer's Association. This NIA-Reagan Institute[5] criteria emphasize on the heterogeneous clinicopathological characteristics of AD, thus the diagnosis are only probabilistic rather than deterministic statements in any given patient based on both CERAD and Braak and Braak staging criteria. Besides, there might be other pathological process involved along with AD that contribute to dementia, for example stroke, Parkinson's disease, progressive supranuclear palsy, and etc. A study that compared different pathological criteria found the CERAD category of definite AD closely resemble the cases that fulfil the Tierney A3[259] AD criteria[194].

1.5 Relation to other neurodegenerative disorders

Under the umbrella term of neurodegenerative disease resulted from neuronal loss, patients suffering from various cognitive or motor deficits are categorized into different arbitrarily defined diseases based on their clinical manifestations. Despite distinct symptoms and affected brain regions (**Table 1.2**), different neurodegenerative diseases share some common features that may suggest potential shared mechanisms underlying disease etiology[99]. For example, the two major clinical manifestations, cognitive deficit and motor deficit, divide the realm into two halves. Age is the most important risk factor for all of the neurodegenerative diseases. Aggregation and progression of misfolded proteins are also involved in all of them, although being the cause or the result of the disease is still under debate. The common features suggest common pathways being altered in neurodegenerative diseases, including protein quality control, the autophagy-lysosome pathway, mitochondria homeostasis, protein seeding and propagation of stress granules, and synaptic toxicity and network dysfunction[99]. Genetically, *MAPT* gene (microtubule associated protein tau) only plays a modest role in sporadic AD but a substantial role in Frontotemporal Dementia (FTD) and Progressive Supranuclear Palsy (PSP). The most important susceptibility region for late-onset AD surrounding the *APOE* gene is involved in non-AD neurodegenerative disorders and conditions[62], such as Lewy body dementias(LBD)[25, 284], Parkinson's disease (PD)[40], amyloid angiopathy[25, 110, 285], TDP-43 proteinopathy[283], hippocampal sclerosis[25, 83, 283].

Table 1.2 Neurodegenerative diseases comparison

	Major Symptoms	Cerebral Cortex	Basal Ganglia	Thalamus	Hippocampus	Cerebellum	Brain Stem	Protein Aggregation
AD	Cognitive	Affected	Affected	Affected	Affected	-	-	Aβ, tau
FTD	Cognitive	Affected	Affected	Affected	-	-	-	TDP-43, tau, FUS
LBD	Cognitive	Affected	-	-	Affected	-	Affected	Aβ, tau, α-synuclein
ALS	Motor	-	-	-	-	-	Affected	TDP-43, FUS, UPS
HD	Motor	Affected	Affected	-	-	-	-	polyglutamine protein
PD	Motor	-	Affected	Affected	-	-	-	α-synuclein
MSA	Motor	-	Affected	-	-	Affected	Affected	α-synuclein

1.6 Dissertation Overview

As discussed above, unlike the rare familial Mendelian dominant AD, late onset AD inherited as complex traits are more common in the population and resulted from dozens of variants involving genes distributed across the whole genome. To tackle the complex interaction in AD, the primary purpose of this dissertation is to apply integrative analysis approaches to demystify and obtain a more accurate and comprehensible picture of AD etiology.

Alzheimer’s disease is characterized by neuronal loss and astrocytosis in the cerebral cortex. However, the specific effects that pathological mutations and coding variants associated with AD have on the cellular composition of the brain are often ignored. In **chapter 2**, to investigate cerebral cortex cell-type population structure I developed an *in-silico* deconvolution method to infer cellular composition from RNA-Seq. I firstly assembled a reference panel to model the transcriptomic signature of neurons, astrocytes, oligodendrocytes and microglia. The panel was created by analyzing expression data from purified cell lines. I evaluated alternative digital deconvolution methods and selected the best performing ones for my primary analyses. I tested the digital deconvolution accuracy on induced pluripotent stem cell (iPSC) derived neurons/microglia cells and neuronal Translating Ribosome Affinity Purification followed by RNA-Seq. Finally, I verified its accuracy by creating artificial admixture with pre-defined cellular proportions. Once the deconvolution approach was optimized, I calculated the cell proportion in AD cases and controls from the different brain regions of LOAD and ADAD datasets. I found that neuronal and astrocyte relative proportions differ between healthy and diseased brains and also among AD cases that carry specific genetic risk variants. Brain carriers of pathogenic mutations in *APP*, *PSEN1* or *PSEN2* presented lower neurons and higher astrocytes relative proportions compared to sporadic AD. Similarly, the *APOE ε*4 allele also showed decreased neuronal and increased astrocyte relative proportions compared to AD non-carriers. On the contrary, carriers of variants in *TREM2* risk showed a lower degree of neuronal loss than matched AD cases in multiple independent studies. These findings suggest that genetic risk factors associated with AD etiology have a specific imprinting in the cellular composition of AD brains.

In **chapter 3**, I utilized cell-type proportions inferred from deconvolution procedure as disease status proxy to identify new genetic variants related to AD. Instead of using disease phenotype, studies which used endo-phenotypes, such as CSF APOE levels[59], CSF amyloid-β, tau, and phosphorylated tau (ptau$_{181}$)[71], and AD proxy[174] have successfully uncovered other variants associated with AD. Using cell type composition inferred from RNA-Seq data as a disease status proxy, I performed cell type association analysis to identify potential new locus that are related to cellular population changes in disease cohort. We imputed and merged genotyping data from seven studies (five centered on neurodegeneration; two focused on schizophrenia and multiple tissue controls), and derived major CNS cell type proportions as described in chapter 2 from cortical RNA-Seq data. Neuronal proportion were normalized by subtracting the mean from each tissue deconvolution results after removing outliers. I identified a variant rs1990621 located in the *TMEM106B* gene region significantly associated with neuronal proportion in cortical RNA-Seq dataset. This variant is in high LD with rs1990622 ($r^2 = 0.98$) which was previously identified as a protective variant in FTD cohorts[266]. In conclusion, I have identified a variant associated with neuronal proportion with potential protective effect in neurodegeneration disorders.

In **Chapter 4**, using network analysis I replicated and reconfirmed the co-expression pattern between *MS4A* gene cluster and *TREM2* in sporadic AD, from which further evidence was inferred from Bayesian network analysis to show that *MS4A4A* might be a potential regulator of *TREM2* that is validated by *in-vitro* experiments. In Autosomal Dominant AD (ADAD) cohort, disrupted and acquired genes were identified from *PSEN1* mutation carriers. Among the genes, previous identified AD related gene and pathways were revealed together with novel findings. These results demonstrated the great potential of applying network approach in identifying disease associated genes and the interactions among them.

In conclusion, contribution from this dissertation work to AD research is summarized in **Chapter 5**, and future directions in research to facilitate diagnosis, intervention, and disease-modifying therapies are also discussed in the context of this dissertation work.

Chapter 2: Genetic variants associated with Alzheimer’s disease confer different cerebral cortex cell-type population structure

This chapter was adapted from:

Li Z, Del-aguila JL, Dube U, et al. Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure. Genome Med. 2018;10(1):43

2.1 Abstract

Background: Alzheimer's disease (AD) is characterized by neuronal loss and astrocytosis in the cerebral cortex. However, the specific effects that pathological mutations and coding variants associated with AD have on the cellular composition of the brain are often ignored.

Methods: I developed and optimized a cell-type-specific expression reference panel and employed digital deconvolution methods to determine brain cellular distribution in three independent transcriptomic studies.

Results: I found that neuronal and astrocyte relative proportions differ between healthy and diseased brains and also among AD cases that carry specific genetic risk variants. Brain carriers of pathogenic mutations in *APP, PSEN1*, or *PSEN2* presented lower neuron and higher astrocyte relative proportions compared to sporadic AD. Similarly, the *APOE ε4* allele also showed decreased neuronal and increased astrocyte relative proportions compared to AD non-carriers. In contrast, carriers of variants in *TREM2* risk showed a lower degree of neuronal loss compared to matched AD cases in multiple independent studies.

Conclusions: These findings suggest that genetic risk factors associated with AD etiology have a specific imprinting in the cellular composition of AD brains. My digital deconvolution method provides an enhanced understanding of the fundamental molecular mechanisms underlying neurodegeneration, enabling the analysis of large bulk RNA-sequencing studies for cell composition and suggests that correcting for the cellular structure when performing transcriptomic analysis will lead to novel insights of AD.

2.2 Introduction

2.2.1 Altered cellular composition confounds downstream transcriptomic analysis

Altered cellular composition is associated with AD progression and decline in cognition. Neuronal loss in the hippocampus is characteristic in the initial stages of AD, which could explain early memory disturbances[205, 282]. As the disease progresses, neuronal death is observed throughout the cerebral cortex. Furthermore, ~25% of individuals who die by ~75 years of age who were cognitively normal also presented substantial cerebral lesions that resemble AD pathology, including amyloid plaque, NFTs, and neuronal loss[132]. Thus, the identification of the brain cellular population structure is essential for understanding neurodegenerative disease progression[107]. However, stereology protocols for counting neurons can be tedious, require extensive training and are susceptible to technical artifacts which may lead to biased quantification of cell-type distributions[107].

Recently there has been a growing interest in understanding the transcriptomic changes attributed to AD[9, 46, 98, 184, 197, 206, 247, 287], as these may point to underlying molecular mechanisms of disease. These studies are typically designed to analyze the expression profiles of large cohorts ascertained from homogenized regions of the brain (e.g. bulk RNA-Seq) of affected and control donors. However, bulk RNA-Seq captures the gene expression of all of the constituent cells in the sampled tissue, and the altered cellular composition associated with AD has been reported to confound downstream analyses[247].

2.2.2 Digital deconvolution approach

Digital deconvolution approaches enhance the interrogation of expression profiles to identify the cellular population structure of individual samples, alleviating the requirement of additional neurostereology procedures. These approaches have been developed, tested and applied to ascertain cellular composition altered in many traits[157, 199, 242, 293]. However, digital deconvolution has not been applied to identify the cellular population structure from RNA-Seq from human brain of AD cases and controls. Technical constraints restrict the dissociation of cells from the brains for very specific conditions[38, 290, 291]. Nevertheless, a limited number of RNA-Seq from isolated cell populations from the brain have been generated[38, 290, 291]. Using these resources, I am now able to generate a reference panel for digital deconvolution of human brain bulk RNA-Seq data.

I sought to investigate the cellular population structure in AD by analyzing RNA-Seq from multiple brain regions of LOAD participants. To do so, I assembled a novel brain reference panel and evaluated the accuracy of digital deconvolution methods by analyzing additional cell-type specific RNA-Seq samples and by creating synthetic admixtures with defined cellular distributions. Then I analyzed large cohorts of pathologically confirmed AD cases and controls (N = 613) and verified that it predicts cellular distribution patterns consistent with neurodegeneration. Finally, I generated RNA-Seq from the parietal lobe of participants from the Charles F. and Joanne Knight Alzheimer's Disease Research Center (Knight-ADRC)[153], including non-demented controls, LOAD cases, with enriched proportions of carriers of high-risk coding variants associated with AD, and also ADAD from The Dominantly Inherited Alzheimer Network[72] (DIAN). I compared the cell composition in ADAD and LOAD; and also evaluated differences among carriers of coding high-risk variants in *PLD3*, *TREM2* and *APOE* ε4. My findings indicated that cell-type composition differs among carriers of specific genetic risk factors, which might be revealing distinct pathogenic mechanisms contributing to disease etiology.

2.3 Methods

2.3.1 Subjects and Samples

DIAN and Knight-ADRC

Parietal lobe tissue of post-mortem brain was obtained with informed consent for research use and were approved by Washington University in St. Louis review board. RNA was extracted from frozen brain using Tissue Lyser LT and RNeasy Mini Kit (Qiagen, Hilden, Germany). RNA-Seq Paired end reads with read length of 2×150bp were generated using Illumina HiSeq 4000 with a mean coverage of 80 million reads per sample (**Table 2.1**; **Table 2.2**). RNA-Seq was generated for 19 brains from The Dominantly Inherited Alzheimer Network (DIAN), 84 brains with late-onset AD and 16 non-demented controls from The Charles F. and Joanne Knight Alzheimer's Disease Research Center (Knight ADRC)[153]. The AD brains selected from the Knight-ADRC are enriched for carrier of variants in *TREM2* (N=20; **Table 2.2**) and *PLD3* (N=33; **Table 2.2**). The clinical status of participants was neuropathologically confirmed[187]. We identified three additional participants from the Knight-ADRC study with *PSEN1* (p.A79V, p.I143T and p.S170F) mutations. CDR scores were obtained during regular visits throughout the study prior to the subject’s decease[190]. A range of other pathological measurement were collected during autopsy including Braak staging, as previously described[35].

Table 2.1 Demographics and disease status of cohorts from four brain bank resources.

	Mayo[a]	**MSBB**[b]	**DIAN**	**Knight-ADRC**
Sample Size	191	300	19	103
Age	83 ± 7.77	83.3 ± 7.55	50.6 ± 7.06	85.1 ± 9.78
% Male	45.5	36	68.4	38.8
% ***APOE*** **ε4+**	33.2	31.7	14.3	45.6
Brain weight	-	-	1187.7 ± 184.5	1138.1 ± 142.5
AD[c]	82	135	19	87
PA[d]	29	0	0	0
Control	80	85	0	16
CDR[e] **= 0**	-	40	0	13
CDR = 0.5	-	40	0	9
CDR = 1	-	30	2	11
CDR = 2	-	44	4	14
CDR = 3	-	146	1	56

[a] Mayo stands for Mayo Clinic.

[b] MSBB stands for Mount Sinai Brain Bank.

[c] AD stands for Alzheimer's Disease.

[d] PA stands for pathological aging (amyloid plaques but no tau tangles).

[e] CDR stands for clinical dementia rating for available samples.

RNA was extracted from frozen brain tissues using Tissue Lyser LT and RNeasy Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instruction. RIN (RNA integrity) and DV200 were measured with RNA 6000 Pico Assay using Bioanalyzer 2100 (Agilent Technologies). The RIN is determined by the software on the Bioanalyzer taking into account the entire electrophoretic trace of the RNA including the presence or absence of degradation products. The DV200 value is defined as the percentage of nucleotides greater than 200nt. RIN and DV200 for all the samples can be found on **Table 2.2**. The yield of each sample is determined by the Quant-iT RNA Assay (Life Technologies) on the Qubit Fluorometer (Fisher Scientific). The cDNA library was prepared with the TruSeq Stranded Total RNA Sample Prep with Ribo-Zero Gold kit (Illumina) and then sequenced by HiSeq 4000 (Illumina) using 2×150 paired end reads at McDonnell Genome Institute, Washington University in St. Louis with a mean of 58.14 ± 8.62 million reads. Number of reads and other QC metrics can be found in **Table 2.2**.

Mayo Clinic Brain Bank

Mayo Clinic Brain Bank RNA-Seq was accessed from the AMP-AD portal (synapse ID = 5550404; accessed January 2017) (**Table 2.1**). Paired end reads of 2×101bp were generated by Illumina HiSeq 2000 sequencers for an average of 134.9 million reads per sample. Neuropathology criteria, quality control procedures, RNA extraction and sequencing details are explained elsewhere[9].

Table 2.2 Demographics and AD mutation carriers of DIAN and Knight-ADRC cohorts.

	DIAN & Knight-ADRC			Total DIAN &	ADAD vs LOAD
	ADAD	**LOAD**	**Control**	**Knight-ADRC**	**t-tests p-value**
RIN	5.69 ± 1.13	6.44 ± 1.16	6.71 ± 1.18	6.34 ± 1.19	9.04×10^{-03}
DV200	86.59 ± 4.12	89.48 ± 3.85	91.19 ± 2.54	89.18 ± 3.97	5.82×10^{-03}
PMI	14.57 ± 10.29	13.05 ± 6.66	10.52 ± 6.09	12.99 ± 7.4	5.16×10^{-01}
Age	51.27 ± 11.13	85.72 ± 6.83	87.08 ± 10.2	79.69 ± 15.67	2.61×10^{-13}
Male %	0.64	0.39	0.38	0.43	4.68×10^{-02}
APOE4+ %	0.3	0.52	0.06	0.44	1.94×10^{-01}
CDR	2.2 ± 0.79	2.37 ± 0.93	0.22 ± 0.31	2.04 ± 1.14	5.42×10^{-01}
Braak	5.94 ± 0.24	4.84 ± 1.29	1.93 ± 0.88	4.61 ± 1.62	8.81×10^{-10}
Number of Total Reads (Million)	60.92 ± 5.6	57.7 ± 9.28	56.6 ± 7.98	58.14 ± 8.62	4.47×10^{-02}
Uniquely Mapped Reads %	79.72 ± 4.28	80.74 ± 4.49	81.06 ± 5.96	80.6 ± 4.65	3.32×10^{-01}
Mapped to Multiple Loci Reads %	16.39 ± 2.1	15.56 ± 2.2	15.08 ± 3.3	15.64 ± 2.36	1.07×10^{-01}
Disease Status	22	84	16	122	-
APP	3	0	0	3	-
PSEN1	18	0	0	18	-
PSEN2	1	0	0	1	-
TREM2	0	20	0	20	-
PLD3 [a]	0	33	0	33	-
UNC5C [a]	0	4	0	4	-
Sporadic AD	0	29	0	29	-

[a] There are two Knight-ADRC subjects that carry both *PLD3* and *UNC5C* variants.

RNA-Seq based transcriptome data was generated from post-mortem brain tissue collected from cerebellum (189 samples) and temporal cortex (191 samples) of Caucasian subjects[2, 9]. RNA was extracted using Trizol® reagent and cleaned with Qiagen RNeasy. RIN measurement was performed with Agilent Technologies 2100 Bioanalyzer. Samples with RIN greater than 5 were included. Library was prepared by Mayo Clinic Medical Genome Facility Gene Expression and Sequencing Cores with TruSeq RNA Sample Prep Kit (Illumina).

Mount Sinai Brain Bank

Mount Sinai Brain Bank RNA-Seq study was downloaded from the AMP-AD portal (synapse ID = 3157743; accessed January 2017) (**Table 2.1**). Single end reads of 100 nucleotides was generated by Illumina HiSeq 2500 System (Illumina, San Diego, CA) for an average of 38.7 million reads per sample[3].

This dataset contains 1030 samples collected from four post-mortem brain regions of 300 subjects: anterior prefrontal cortex (BA10), superior temporal gyrus (BA22), parahippocampal gyrus (BA36), and inferior frontal gyrus (BA44). RNA-Seq was generated using the TruSeq RNA Sample Preparation Kit v2 and Ribo-Zero rRNA removal kit (Illumina, San Diego, CA)[3].

iPSC-derived neurons

Dermal fibroblasts were obtained from skin biopsies from research participants in the Knight-ADRC (Fibroblast lines: F11362, F12455, and F13504). Human fibroblasts were reprogrammed into iPSC using non-integrating Sendai virus carrying OCT3/4, SOX2, KLF4, and cMYC[255, 265]. iPSCs were manually selected and expanded on Matrigel in mTesR1 (StemCell Techologies). iPSCs were characterized for expression of pluripotency markers by immunocytochemistry and quantitative PCR (qPCR). qPCR with probes specific to Sendai virus were used to confirm the absence of virus in the isolated clones. All cell lines were confirmed to have a normal karyotype based on G-band karyotyping. To generate cortical neurons, iPSCs were plated in a v-bottom plate in neural induction media (StemCell Technologies; 65,000 per well) to form highly uniform neural aggregates. After 5 days, neural aggregates were transferred onto PLO/laminin-coated tissue culture plates. Neural rosettes formed over 5-7 days. The resulting neural rosettes were then isolated by enzymatic selection (StemCell Technologies) and cultured as neural progenitor cells (NPCs). NPCs were then differentiated by culturing in neural maturation medium (neurobasal medium supplemented with B27, GDNF, BDNF, cAMP). RNA was collected from the cells and sequenced following the same protocol and processing pipeline as the DIAN and Knight-ADRC dataset.

In addition, I accessed RNA-Seq data generated for iPSC-derived neurons from the Broad iPSC study[7] (Synapse ID: syn3607401). Forebrain neurons from wild-type background were generated using an embryoid body-based protocol to produce neural progenitor cells (day 17) and mature neurons (day 57 and day 100). RNA was purified using a PureLink RNA mini-kit (Life Technologies). RNA-Seq libraries were prepared using Illumina Strand Specific TruSeq protocol, and sequenced to obtain an average of 75M reads in pairs reads per sample.

TRAP-seq mice

All animal procedures were performed in accordance with the guidelines of Washington University's Institutional Animal Care and Use Committee. The Rosa26^{fsTRAP} mice (Gt(ROSA)26Sor$^{tm1(CAG-EGFP/Rpl10a,-birA)Wtp}$)[294] (The Jackson Laboratory) were crossed with PV^{Cre} mice (Pvalb$^{tm1(cre)Arbr}$)[128] (The Jackson Laboratory) to produce PV-TRAP mice directing expression of EGFP-L10a ribosomal fusion protein in parvalbumin (PV) expressing cells.

Purification of cell-type specific mRNA by translating ribosome affinity purification (TRAP) was described previously[122] with modifications. Briefly, PV-TRAP mouse brain was removed and quickly washed in ice-cold dissection buffer (1× HBSS, 2.5 mM HEPES-KOH (pH 7.3), 35 mM glucose, and 4 mM $NaHCO_3$ in RNase-free water). Barrel cortex was rapidly dissected and flash-frozen in liquid nitrogen, and then stored at -80 °C until use. Affinity matrix was prepared with 150 µl of Streptavidin MyOne T1 Dynabeads, 60 µg of Biotinylated Protein L, and 25 µg of each of GFP antibodies 19C8 and 19F7. The tissue was homogenized on ice in 1 ml of tissue-lysis buffer (20 mM HEPES KOH (pH 7.4), 150 mM KCl, 10 mM $MgCl_2$, EDTA-free protease inhibitors, 0.5 mM DTT, 100 µg/ml cycloheximide, and 10 µl/ml rRNasin and Superasin). Homogenates were centrifuged for 10 min at 2,000 × *g*, 4 °C, and 1/9 sample volume of 10% NP-40 and 300 mM DHPC were added to the supernatant at final concentration of 1% (vol/vol). After incubation on ice for 5 min, the lysate was centrifuged for 10 min at 20,000 × *g* to pellet insolubilized material. Then 200 µl of freshly resuspended affinity matrix was added to the supernatant and incubated at 4 °C for 16–18 hours with gentle end-over-end mixing in a tube rotator. After incubation, the beads were collected with a magnet and resuspended in 1000 µl of high-salt buffer (20 mM HEPES KOH (pH 7.3), 350 mM KCl, 10 mM $MgCl_2$, 1% NP-40, 0.5 mM DTT and 100 µg/ml cycloheximide), and collected with magnet as above. After 4 times of washing with high-salt buffer, RNA was extracted using Absolutely RNA Nanoprep Kit (Agilent Technologies) following manufacturer's instruction. RNA quantification was measured using Qubit RNA HS Assay Kit (Life Technologies) and the integrity was determined by Bioanalyzer 2100 using an RNA Pico chip (Agilent Technologies). The cDNA library was prepared with Clontech SMARTer and then sequenced by HiSeq3000. Single end reads of 50 base pairs were generated for an average of 29.2 million reads per sample (24 samples).

iPSC-derived microglia

The data was accessed from the AMP-AD portal (Synapse ID: syn7203233). This dataset is comprised of iPSC-derived microglia (N = 10) from human primitive streak-like cells[77]. Within 30 days of differentiation, myeloid progenitors coexpressing CD14 and CX3CR1 were generated. These iPSC-derived microglia were able to perform phagocytosis and elicit ADP-induced intracellular Ca^{2+} transients that asserted their microglia identity as opposed to macrophage. Single-ended RNA-Seq data was generated with the Illumina HiSeq 2500 platform following the Illumina protocol.

2.3.2 RNA-Seq QC and Alignment

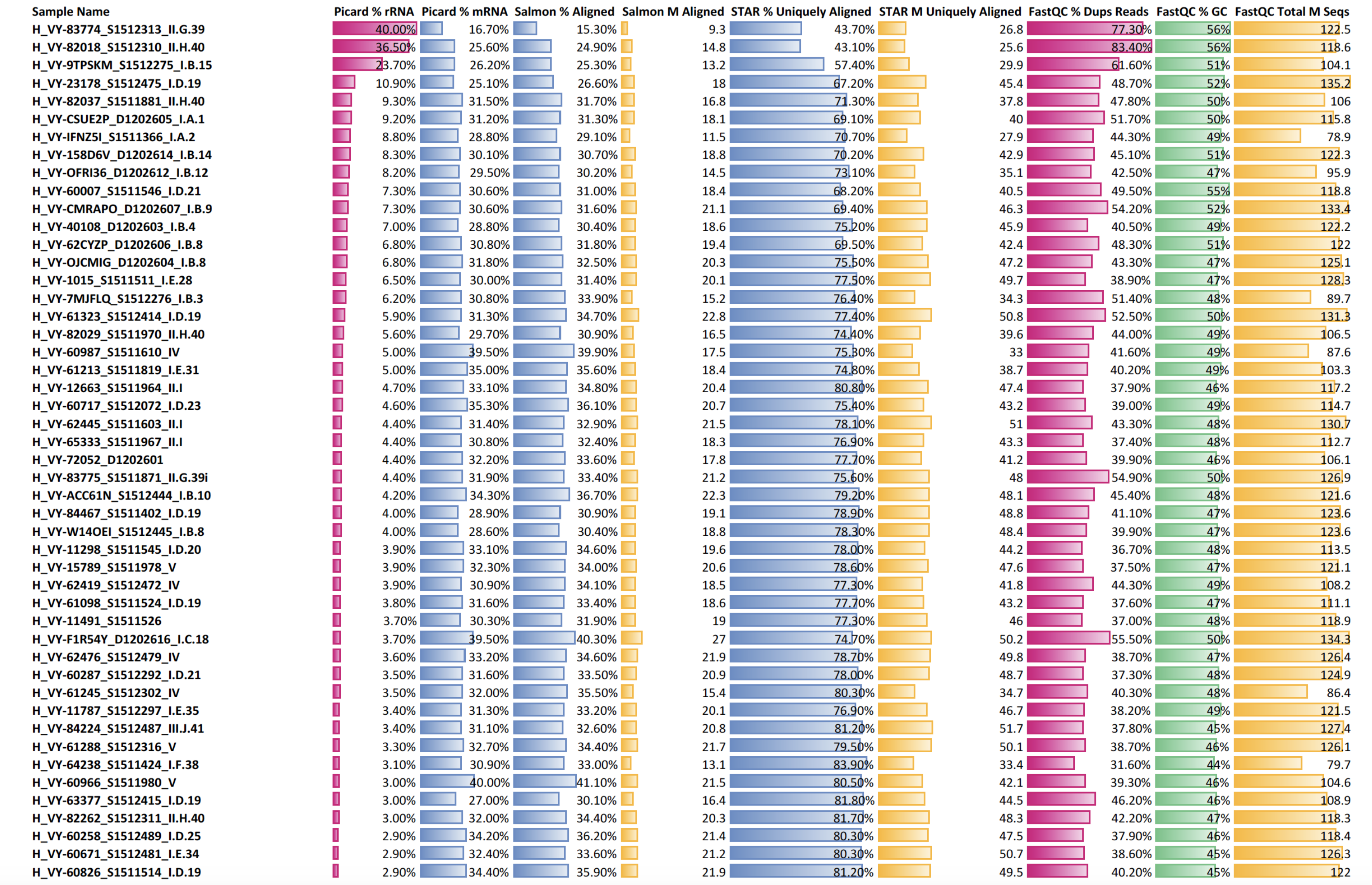
FastQC was applied to DIAN and Knight-ADRC RNA-Seq data to perform quality check on various aspects of sequencing quality[231]. Each category of FastQC will be explained with pass or fail examples together with summary results ascertained from the DIAN and Knight-ADRC combined dataset. QC result explanations were obtained from the developer's website[11].

The DIAN and Knight-ADRC dataset was aligned to human GRCh37 primary assembly using Star (ver 2.5.2b)[74]. I used the primary assembly and aligned reads to the assembled chromosomes, un-localized and unplaced scaffolds, and discarded alternative haploid sequences. Sequencing metrics, including coverage, distribution of reads in the genome[4], ribosomal and mitochondrial contents and alignment quality, were further obtained by applying Picard CollectRnaSeqMetrics (ver 2.8.2) to detect sample deviation. QC results from FastQC, Star, Picard, and Salmon are merged with multi-QC software to generated integrated summary reports (**Table 2.3**).

Problematic samples summary:

- RIN < 5 & DV200 < 75
    - H_VY-82018_S1512310_II.H.40
    - H_VY-83774_S1512313_II.G.39
- Low Yield
    - H_VY-60410_S1511525_I.D.19
    - H_VY-62240_S1511620_IV
- High Ribosomal RNA
    - H_VY-83774_S1512313_II.G.39
    - H_VY-82018_S1512310_II.H.40
    - H_VY-9TPSKM_S1512275_I.B.15
- High Median 5' to 3' bias
    - H_VY-60007_S1511546_I.D.21
    - H_VY-83774_S1512313_II.G.39
- Ethnicity non-Europeans
    - H_VY-83665_S1511508_I.E.31
    - H_VY-61256_S1511537_I.E.31
    - H_VY-62275_S1511542_I.D.19
    - H_VY-11964_S1512298_I.E.29
    - H_VY-23178_S1512475_I.D.19
    - H_VY-62464_S1512484_VI.N
- Transcriptome-wise outliers
    - H_VY-F1R54Y_D1202616_I.C.18
    - H_VY-1XYTL9_D1202619_I.B.10
    - H_VY-61377_S1511495_I.D.19
    - H_VY-61245_S1512302_IV
    - H_VY-61684_S1512304_IV
    - H_VY-60007_S1511546_I.D.21
    - H_VY-83774_S1512313_II.G.39

Table 2.3 Summarized quality check results integrated with Multi-QC



Sample Name	Picard % rRNA	Picard % mRNA	Salmon % Aligned	Salmon M Aligned	STAR % Uniquely Aligned	STAR M Uniquely Aligned	FastQC % Dups Reads	FastQC % GC	FastQC Total M Seqs
H_VY-83774_S1512313_II.G.39	40.00%	16.70%	15.30%	9.3	43.70%	26.8	77.30%	56%	122.5
H_VY-82018_S1512310_II.H.40	36.50%	25.60%	24.90%	14.8	43.10%	25.6	83.40%	56%	118.6
H_VY-9TPSKM_S1512275_I.B.15	23.70%	26.20%	25.30%	13.2	57.40%	29.9	61.60%	51%	104.1
H_VY-23178_S1512475_I.D.19	10.90%	25.10%	26.60%	18	67.20%	45.4	48.70%	52%	135.2
H_VY-82037_S1511881_II.H.40	9.30%	31.50%	31.70%	16.8	71.30%	37.8	47.80%	50%	106
H_VY-CSUE2P_D1202605_I.A.1	9.20%	31.20%	31.30%	18.1	69.10%	40	51.70%	50%	115.8
H_VY-IFNZ5I_S1511366_I.A.2	8.80%	28.80%	29.10%	11.5	70.70%	27.9	44.30%	49%	78.9
H_VY-158D6V_D1202614_I.B.14	8.30%	30.10%	30.70%	18.8	70.20%	42.9	45.10%	51%	122.3
H_VY-OFRI36_D1202612_I.B.12	8.20%	29.50%	30.20%	14.5	73.10%	35.1	42.50%	47%	95.9
H_VY-60007_S1511546_I.D.21	7.30%	30.60%	31.00%	18.4	68.20%	40.5	49.50%	55%	118.8
H_VY-CMRAPO_D1202607_I.B.9	7.30%	30.60%	31.60%	21.1	69.40%	46.3	54.20%	52%	133.4
H_VY-40108_D1202603_I.B.4	7.00%	28.80%	30.40%	18.6	75.20%	45.9	40.50%	49%	122.2
H_VY-62CYZP_D1202606_I.B.8	6.80%	30.80%	31.80%	19.4	69.50%	42.4	48.30%	51%	122
H_VY-OJCMIG_D1202604_I.B.8	6.80%	31.80%	32.50%	20.3	75.50%	47.2	43.30%	47%	125.1
H_VY-1015_S1511511_I.E.28	6.50%	30.00%	31.40%	20.1	77.50%	49.7	38.90%	47%	128.3
H_VY-7MJFLQ_S1512276_I.B.3	6.20%	30.80%	33.90%	15.2	76.40%	34.3	51.40%	48%	89.7
H_VY-61323_S1512414_I.D.19	5.90%	31.30%	34.70%	22.8	77.40%	50.8	52.50%	50%	131.3
H_VY-82029_S1511970_II.H.40	5.60%	29.70%	30.90%	16.5	74.40%	39.6	44.00%	49%	106.5
H_VY-60987_S1511610_IV	5.00%	39.50%	39.90%	17.5	75.30%	33	41.60%	49%	87.6
H_VY-61213_S1511819_I.E.31	5.00%	35.00%	35.60%	18.4	74.80%	38.7	40.20%	49%	103.3
H_VY-12663_S1511964_II.I	4.70%	33.10%	34.80%	20.4	80.80%	47.4	37.90%	46%	117.2
H_VY-60717_S1512072_I.D.23	4.60%	35.30%	36.10%	20.7	75.40%	43.2	39.00%	49%	114.7
H_VY-62445_S1511603_II.I	4.40%	31.40%	32.90%	21.5	78.10%	51	43.30%	48%	130.7
H_VY-65333_S1511967_II.I	4.40%	30.80%	32.40%	18.3	76.90%	43.3	37.40%	48%	112.7
H_VY-72052_D1202601	4.40%	32.20%	33.60%	17.8	77.70%	41.2	39.90%	46%	106.1
H_VY-83775_S1511871_II.G.39i	4.40%	31.90%	33.40%	21.2	75.60%	48	54.90%	50%	126.9
H_VY-ACC61N_S1512444_I.B.10	4.20%	34.30%	36.70%	22.3	79.20%	48.1	45.40%	48%	121.6
H_VY-84467_S1511402_I.D.19	4.00%	28.90%	30.90%	19.1	78.90%	48.8	41.10%	47%	123.6
H_VY-W14OEI_S1512445_I.B.8	4.00%	28.60%	30.40%	18.8	78.30%	48.4	39.90%	47%	123.6
H_VY-11298_S1511545_I.D.20	3.90%	33.10%	34.60%	19.6	78.00%	44.2	36.70%	48%	113.5
H_VY-15789_S1511978_V	3.90%	32.30%	34.00%	20.6	78.60%	47.6	37.50%	47%	121.1
H_VY-62419_S1512472_IV	3.90%	30.90%	34.10%	18.5	77.30%	41.8	44.30%	49%	108.2
H_VY-61098_S1511524_I.D.19	3.80%	31.60%	33.40%	18.6	77.70%	43.2	37.60%	47%	111.1
H_VY-11491_S1511526	3.70%	30.30%	31.90%	19	77.30%	46	37.00%	48%	118.9
H_VY-F1R54Y_D1202616_I.C.18	3.70%	39.50%	40.30%	27	74.70%	50.2	55.50%	50%	134.3
H_VY-62476_S1512479_IV	3.60%	33.20%	34.60%	21.9	78.70%	49.8	38.70%	47%	126.4
H_VY-60287_S1512292_I.D.21	3.50%	31.60%	33.50%	20.9	78.00%	48.7	37.30%	48%	124.9
H_VY-61245_S1512302_IV	3.50%	32.00%	35.50%	15.4	80.30%	34.7	40.30%	48%	86.4
H_VY-11787_S1512297_I.E.35	3.40%	31.30%	33.20%	20.1	76.90%	46.7	38.20%	49%	121.5
H_VY-84224_S1512487_III.J.41	3.40%	31.10%	32.60%	20.8	81.20%	51.7	37.80%	45%	127.4
H_VY-61288_S1512316_V	3.30%	32.70%	34.40%	21.7	79.50%	50.1	38.70%	46%	126.1
H_VY-64238_S1511424_I.F.38	3.10%	30.90%	33.00%	13.1	83.90%	33.4	31.60%	44%	79.7
H_VY-60966_S1511980_V	3.00%	40.00%	41.10%	21.5	80.50%	42.1	39.30%	46%	104.6
H_VY-63377_S1512415_I.D.19	3.00%	27.00%	30.10%	16.4	81.80%	44.5	46.20%	46%	108.9
H_VY-82262_S1512311_II.H.40	3.00%	32.00%	34.40%	20.3	81.70%	48.3	42.20%	47%	118.3
H_VY-60258_S1512489_I.D.25	2.90%	34.20%	36.20%	21.4	80.30%	47.5	37.90%	46%	118.4
H_VY-60671_S1512481_I.E.34	2.90%	32.40%	33.60%	21.2	80.30%	50.7	38.60%	45%	126.3
H_VY-60826_S1511514_I.D.19	2.90%	34.40%	35.90%	21.9	81.20%	49.5	40.20%	45%	122

H_VY-83665_S1511508_I.E.31	2.90%	34.40%	36.30%	22.4	83.10%	51.2	37.40%	46%	123.2
H_VY-F11362.1d1F10_Neuron_3_S1512505_VII.Q.47	2.90%	46.30%	46.80%	30	83.30%	53.5	52.00%	47%	128.4
H_VY-60324_S1512496	2.80%	35.60%	37.00%	23.1	79.30%	49.4	41.90%	47%	124.6
H_VY-60173_S1511775_VI.N	2.70%	29.10%	31.60%	19.2	79.30%	48.3	40.00%	48%	121.8
H_VY-62481_S1511392_I.D.19	2.70%	35.80%	38.20%	19.5	82.40%	42.1	37.20%	47%	102.2
H_VY-1XYTL9_D1202619_I.B.10	2.60%	38.10%	38.50%	24.6	76.40%	48.7	49.80%	47%	127.5
H_VY-62069_S1512474_IV	2.60%	29.10%	31.60%	19.7	79.20%	49.4	42.80%	47%	124.7
H_VY-62099_S1511885_IV	2.60%	30.30%	32.50%	18.3	80.90%	45.7	35.40%	47%	112.9
H_VY-84619_S1512321_I.D.25	2.60%	40.80%	42.50%	22.7	82.70%	44.2	41.50%	45%	106.9
H_VY-62464_S1512484_VI.N	2.50%	32.60%	34.50%	20.4	82.20%	48.6	38.50%	45%	118.3
H_VY-1319_S1512296_I.E.35	2.40%	29.10%	30.80%	16.2	80.10%	42	39.10%	45%	104.8
H_VY-60591_S1511501_I.D.25	2.40%	31.40%	33.50%	19.5	79.80%	46.5	37.20%	47%	116.5
H_VY-60942_S1511504_I.D.19	2.40%	28.20%	30.60%	18.8	80.50%	49.6	35.70%	46%	123.2
H_VY-61608_S1512323_I.D.19	2.40%	31.60%	35.30%	21.6	82.60%	50.5	40.00%	47%	122.4
H_VY-82655_S1511882_I.F.38	2.40%	31.20%	33.10%	19.3	80.20%	46.8	35.20%	46%	116.7
H_VY-F12455.8_Astro_IGF_2_S1512510_VII.O.43	2.40%	63.80%	64.20%	41.9	83.20%	54.3	68.20%	49%	130.5
H_VY-1307_D1202600_VI.K	2.30%	33.80%	34.90%	22.6	77.90%	50.4	42.20%	46%	129.4
H_VY-60975_S1511609_IV	2.30%	32.90%	34.60%	20.9	79.90%	48.2	38.90%	46%	120.7
H_VY-62077_S1511498_I.E.35	2.30%	35.10%	36.90%	22.6	81.40%	49.8	38.10%	46%	122.5
H_VY-62894_S1511520_I.D.25	2.30%	35.90%	37.20%	22.5	78.10%	47.3	38.70%	48%	121.2
H_VY-BVG3QG_D1202617_I.A.2	2.30%	38.90%	40.50%	25	82.20%	50.7	41.90%	47%	123.3
H_VY-12184_S1512488_III.J.41	2.20%	31.30%	33.40%	17	79.90%	40.7	37.30%	47%	101.8
H_VY-61084_S1511436_I.F.38	2.20%	35.70%	37.00%	23.7	80.10%	51.4	39.30%	47%	128.3
H_VY-61256_S1511537_I.E.31	2.20%	32.90%	34.50%	21.4	81.00%	50.2	37.20%	46%	123.9
H_VY-62460_S1511984_V	2.20%	35.50%	37.30%	22.3	82.00%	49	37.30%	46%	119.5
H_VY-63491_S1511398_I.E.33	2.20%	40.80%	42.70%	18.7	84.00%	36.8	38.90%	46%	87.7
H_VY-82669_S1511519_I.D.25	2.20%	31.10%	33.30%	19	82.30%	47	34.60%	46%	114.2
H_VY-F11362.1d1F10_Neuron_1_S1512503_VII.Q.47	2.20%	49.80%	51.10%	33.1	84.20%	54.5	52.90%	47%	129.5
H_VY-62435_S1512307_IV	2.10%	30.00%	31.80%	19.8	81.00%	50.3	37.80%	47%	124.4
H_VY-2968OM_S1512443_I.B.13	2.00%	32.70%	34.60%	21.5	80.30%	49.9	49.60%	46%	124.2
H_VY-60427_S1512314_V	2.00%	32.60%	34.50%	20.9	80.30%	48.7	33.70%	47%	121.3
H_VY-62661_S1511768_IV	2.00%	33.10%	35.30%	21.8	79.30%	48.8	43.40%	47%	123.2
H_VY-82034_S1511971_II.H.40	2.00%	31.70%	34.20%	19.9	83.20%	48.4	35.10%	45%	116.4
H_VY-12114_S1512416_I.D.24	1.90%	28.30%	30.10%	20.6	80.70%	55.3	37.90%	45%	137
H_VY-1DKYRE_D1202618_I.B.11	1.90%	35.80%	37.00%	23.4	78.00%	49.3	43.60%	47%	126.2
H_VY-60324_S1511770_VI.N	1.90%	33.90%	36.20%	18.2	81.10%	40.9	37.00%	47%	100.8
H_VY-60738_S1511465_IV	1.90%	38.20%	40.00%	25.3	83.10%	52.6	40.10%	45%	126.6
H_VY-61074_S1512315_V	1.90%	29.40%	31.90%	19	81.50%	48.6	36.10%	45%	119.3
H_VY-61464_S1511421_I.D.19	1.90%	35.40%	36.70%	16.6	80.20%	36.3	48.60%	45%	90.5
H_VY-61736_S1511884_IV	1.90%	29.60%	32.60%	20.8	81.70%	52	42.90%	46%	127.3
H_VY-83775_S1511973_II.G.39	1.90%	33.20%	35.60%	21.4	82.40%	49.5	41.40%	45%	120.2
H_VY-F12455.8_Astro_3_S1512508_VII.O.43	1.90%	60.80%	61.70%	40.3	85.70%	55.9	58.40%	48%	130.5
H_VY-60919_S1511505_I.D.25	1.80%	36.10%	37.90%	23.1	81.90%	49.8	40.70%	46%	121.6
H_VY-22528_S1511420_I.E.33	1.70%	27.90%	30.10%	19.6	81.80%	53.3	40.00%	45%	130.2
H_VY-60564_S1511510_I.E.35	1.70%	29.90%	31.90%	18.5	80.70%	46.9	32.50%	46%	116.2
H_VY-82397_S1511396_I.B.17	1.70%	41.80%	43.40%	23.5	80.10%	43.5	44.30%	48%	108.6
H_VY-82654_S1511974_I.D.25	1.70%	31.40%	33.90%	19.2	83.80%	47.3	38.80%	46%	113
H_VY-13188_S1511812_VI.N	1.50%	30.20%	32.30%	16.7	81.80%	42.3	34.60%	45%	103.4
H_VY-60752_D1202602_VI.K	1.50%	41.20%	42.70%	26.2	82.50%	50.7	42.90%	46%	122.8

H_VY-62558_S1512493	1.50%	29.30%	31.50%	19.6	82.80%	51.5	39.00%	45%	124.4
H_VY-82019_S1511985_VI.L	1.50%	30.00%	32.50%	16.5	81.60%	41.3	33.30%	46%	101.2
H_VY-Y1XG2W_D1202622_I.B.5	1.50%	36.40%	38.10%	22.6	79.60%	47.2	40.70%	46%	118.6
H_VY-40116_S1511499_I.D.25	1.40%	31.30%	33.20%	20.9	81.40%	51.2	37.80%	45%	125.8
H_VY-61377_S1511495_I.D.19	1.40%	43.80%	45.60%	26.7	83.00%	48.6	47.10%	48%	117
H_VY-62319_S1511703_IV	1.40%	32.70%	35.10%	21.2	83.10%	50.3	36.90%	46%	121
H_VY-62373_S1511621_IV	1.40%	33.10%	35.10%	22.5	81.40%	52.2	37.20%	45%	128.3
H_VY-13012_S1511977_V	1.30%	33.00%	35.30%	21	83.90%	49.9	34.10%	45%	118.9
H_VY-62216_S1511886_IV	1.30%	29.30%	32.30%	19.3	84.10%	50.3	32.90%	46%	119.6
H_VY-62240_S1511620_IV	1.30%	36.30%	37.80%	3.6	81.50%	7.8	24.80%	46%	19.2
H_VY-64370_S1511427_I.D.25	1.30%	31.70%	34.10%	19.9	84.80%	49.5	37.60%	45%	116.6
H_VY-F11362.1d1B6_Neuron_2_S1512501_VII.O.45	1.30%	54.40%	55.20%	33.9	84.00%	51.6	57.90%	48%	123
H_VY-479_S1511606_VI.N	1.20%	33.40%	35.20%	19.9	82.20%	46.5	33.20%	45%	113.3
H_VY-60186_S1512499	1.20%	33.30%	35.10%	24	80.90%	55.3	40.60%	45%	136.7
H_VY-60410_S1511525_I.D.19	1.20%	37.70%	39.10%	3.2	80.00%	6.6	23.60%	45%	16.4
H_VY-60951_S1512291_II.I	1.20%	34.20%	36.20%	21.9	82.70%	50.2	37.40%	45%	121.3
H_VY-61589_S1512295_I.D.24	1.20%	33.30%	35.60%	22.7	82.90%	52.8	40.00%	45%	127.3
H_VY-61873_S1511544_I.D.19	1.20%	42.60%	44.80%	24.7	84.90%	46.8	39.70%	46%	110.3
H_VY-62275_S1511542_I.D.19	1.20%	37.60%	40.20%	23.5	85.00%	49.7	36.20%	47%	117
H_VY-62433_S1512294_I.D.22	1.20%	35.50%	38.30%	17.4	83.50%	38	36.10%	46%	91
H_VY-62766_S1511965_II.I	1.20%	37.00%	40.10%	22.3	85.70%	47.7	37.00%	47%	111.3
H_VY-65229_S1511872_VI.M.42i	1.20%	31.20%	33.90%	17.5	82.50%	42.5	33.20%	46%	103.1
H_VY-83832_S1511496_I.D.25	1.20%	32.60%	34.70%	21.2	82.20%	50.2	42.20%	46%	122.3
H_VY-11231_S1512486_III.J.41	1.10%	30.40%	32.00%	17.6	82.50%	45.4	34.00%	45%	110.1
H_VY-167_S1511608_IV	1.10%	37.10%	38.60%	24.4	81.80%	51.7	40.70%	45%	126.4
H_VY-61391_D1202306	1.10%	32.50%	34.90%	22.4	83.80%	53.9	35.90%	45%	128.6
H_VY-61872_S1511615_IV	1.10%	33.70%	36.30%	18.4	86.20%	43.5	33.20%	46%	101
H_VY-62157_S1511618_IV	1.10%	35.70%	38.00%	23.9	83.20%	52.3	39.70%	45%	125.9
H_VY-62215_S1511607_VI.N	1.10%	34.00%	36.20%	20.7	82.90%	47.3	35.50%	46%	114.2
H_VY-65229_S1511986_VI.N	1.10%	31.40%	34.10%	19.7	82.80%	47.9	34.60%	45%	115.8
H_VY-84228_S1511534_I.E.33	1.10%	36.70%	38.70%	22.4	83.40%	48.4	35.30%	46%	116
H_VY-UCCKJF_D1202615_I.B.7	1.10%	31.40%	33.50%	18.6	83.40%	46.4	35.70%	45%	111.3
H_VY-11964_S1512298_I.E.29	1.00%	40.70%	42.60%	27.8	83.80%	54.6	42.10%	45%	130.4
H_VY-61681_S1511401_I.E.35	1.00%	38.40%	40.80%	26.2	86.80%	55.7	40.40%	45%	128.2
H_VY-61819_S1511528_I.E.33	1.00%	33.00%	35.90%	20.1	86.50%	48.4	42.10%	46%	111.9
H_VY-63976_S1511873_VI.M.42	1.00%	30.20%	35.40%	22.5	85.00%	54	39.90%	46%	127
H_VY-12152_S1511976_V	0.90%	33.90%	35.70%	22.8	81.50%	52	38.80%	45%	127.6
H_VY-12152_S1512495	0.90%	29.20%	31.70%	16	84.10%	42.4	29.70%	44%	100.8
H_VY-12608_S1512320_I.D.19	0.90%	32.50%	34.60%	22.2	81.60%	52.3	37.00%	45%	128.1
H_VY-60772_S1511696_VI.N	0.90%	31.60%	33.90%	21.3	84.40%	53.1	34.90%	45%	125.9
H_VY-61394_S1511988_VI.N	0.90%	36.00%	38.30%	22.9	85.90%	51.3	36.30%	44%	119.5
H_VY-61402_S1511604_V	0.90%	38.50%	40.40%	26.3	83.50%	54.3	40.90%	45%	130.1
H_VY-61632_S1512322_I.D.19	0.90%	31.70%	34.30%	21.8	82.50%	52.5	37.20%	45%	127.3
H_VY-QNVRDM_D1202613_I.B.8	0.90%	34.30%	36.60%	23.6	84.70%	54.7	39.10%	45%	129.3
H_VY-12355_S1512490	0.80%	35.20%	37.20%	23.1	83.60%	52	38.70%	45%	124.5
H_VY-1442_S1512485_VI.N	0.80%	38.70%	40.30%	21.3	83.20%	43.9	37.10%	46%	105.6
H_VY-61472_S1511611_IV	0.80%	36.00%	38.40%	22.2	86.20%	50	41.70%	45%	115.9
H_VY-61565_S1511975_I.E.35	0.80%	35.90%	38.30%	24.1	84.50%	53.1	38.70%	45%	125.8
H_VY-61575_S1512418_I.E.30	0.80%	41.30%	42.80%	23.7	84.00%	46.4	39.70%	46%	110.6

H_VY-61649_D1202287_I.E.32	0.80%	39.00%	41.20%	25	85.70%	52.1	38.70%	45%	121.6
H_VY-61684_S1512304_IV	0.80%	33.40%	36.20%	21.4	83.20%	49	42.00%	45%	117.8
H_VY-61741_S1512305_IV	0.80%	32.40%	35.10%	23.3	83.10%	55.3	39.80%	45%	133
H_VY-64269_S1512482_II.I	0.80%	32.20%	35.00%	12.7	86.30%	31.4	30.10%	44%	72.7
H_VY-72221_S1511778_VI.N	0.80%	33.20%	35.60%	21.7	84.50%	51.5	35.90%	45%	121.9
H_VY-85194_S1511391_I.B.9	0.80%	36.40%	38.80%	23.6	83.60%	50.9	39.20%	46%	121.8
H_VY-60290_S1511547_I.D.21	0.70%	34.10%	36.00%	22.3	83.10%	51.5	36.60%	45%	124.1
Pooled_RNA_2896115730	0.70%	46.10%	48.50%	32.3	85.50%	57	52.90%	47%	133.4
H_VY-40510_S1512480_I.E.35	0.60%	34.40%	36.40%	22.8	84.70%	53	40.70%	44%	125
H_VY-60180_S1511776_VI.N	0.60%	45.50%	47.00%	30.4	84.30%	54.4	46.20%	45%	129.1
H_VY-60974_S1512301_VI.N	0.60%	35.50%	37.70%	21.8	83.50%	48.3	34.70%	45%	115.7
H_VY-61609_S1512483_VI.N	0.60%	36.40%	38.80%	23.9	87.50%	53.9	37.00%	43%	123.2
H_VY-61744_S1511530_I.E.35	0.50%	34.10%	36.60%	18.5	86.40%	43.7	34.90%	45%	101.1
H_VY-62658_S1512478_VI.N	0.50%	36.30%	38.70%	22.7	86.20%	50.7	33.40%	44%	117.6
H_VY-F11362.1d1B6_Neuron_3_S1512502_VII.O.45	0.50%	49.70%	50.00%	22.6	82.60%	37.3	46.70%	47%	90.4
H_VY-F11362.1d1F10_Neuron_2_S1512504_VII.Q.47	0.50%	51.10%	52.20%	20.8	79.20%	31.5	48.30%	49%	79.5
H_VY-F12455.8_Astro_2_S1512507_VII.O.43	0.50%	65.60%	66.60%	38.9	87.50%	51.1	56.60%	48%	116.8
H_VY-60524_S1511987_VI.N	0.40%	36.30%	38.50%	24.6	83.80%	53.5	36.60%	45%	127.7
H_VY-61280_S1511982_V	0.40%	36.40%	39.20%	25.2	86.80%	55.6	46.90%	45%	128.2
H_VY-F11362.1d1B6_Neuron_1_S1512500_VII.O.45	0.40%	53.40%	54.70%	28.8	84.20%	44.4	52.60%	47%	105.4
H_VY-61587_S1512303_IV	0.30%	33.60%	36.80%	22.9	88.60%	55.2	36.20%	44%	124.8
H_VY-F12455.8_Astro_IGF_3_S1512511_VII.O.43	0.30%	67.10%	68.60%	39.6	88.10%	50.9	58.00%	47%	115.6
Pooled_RNA_2896115783	0.30%	43.30%	47.70%	29.2	87.90%	53.8	46.00%	45%	122.5
H_VY-F12455.8_Astro_IGF_1_S1512509_VII.O.43	0.20%	66.30%	68.10%	30.3	87.80%	39	57.20%	47%	88.8
Averge	2.96%	34.98%	36.85%	21.4	80.57%	46.8	41.26%	47%	116.2

Per base sequence quality

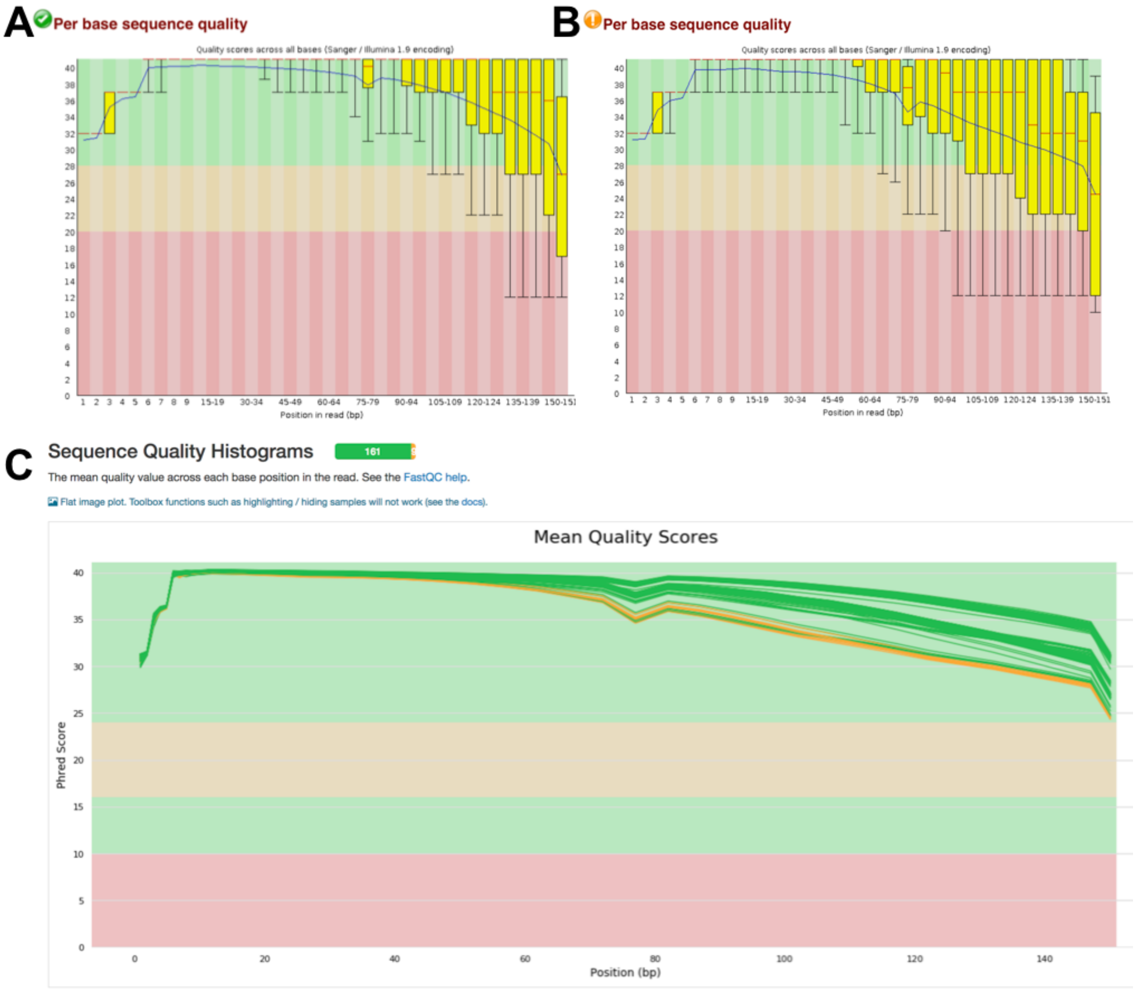
Figure 2.1A and **Figure 2.1B** are passed and failed example for the per base sequence quality check. As its name suggested, this analysis summarizes over all sequence quality of each sample for each read base. In my case, the read length is 151 base pairs represented in the x-axis. The y-axis on the graph showed the quality scores. The sequence quality is calculated as $Q = -10\times\log_{10}(e)$ where 'e' is the estimated probability of the base call being wrong. Thus, higher score indicates higher quality, and it ranges from 0 to 40. A quality score of 20 represents an error rate of 1 in 100, and a quality score of 40 represents an error rate of 1 in 10,000 and a call accuracy of 99.99%. Green region represents good quality calls from 28 to 40; orange region represents calls of reasonable quality from 20 to 28; red represents poor quality calls with quality less than 20, with a call accuracy of 99%. The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.

Each column of box and whisker plot is the summarized quality score of all the reads for that particular base position. "The central red line is the median value; the yellow box represents the inter-quartile range (25-75%); the upper and lower whiskers represent the 10% and 90% points. The blue solid line represents the mean quality"[11]. **Figure 2.1C** is the overall per base sequence quality for all sample, which showed 161 passed and 9 samples with warning.

The 9 warning samples are:

- H_VY-61609_S1512483_VI.N.bam

- H_VY-63976_S1511873_VI.M.42.bam
- H_VY-72221_S1511778_VI.N.bam
- H_VY-82037_S1511881_II.H.40.bam
- H_VY-82397_S1511396_I.B.17.bam
- H_VY-83665_S1511508_I.E.31.bam
- H_VY-83775_S1511871_II.G.39i.bam
- H_VY-F11362.1d1B6_Neuron_2_S1512501_VII.O.45.bam
- H_VY-F11362.1d1F10_Neuron_1_S1512503_VII.Q.47.bam



The mean quality value across each base position in the read. See the FastQC help.

Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).

Figure 2.1 Sequence quality check. A) Passed sample from H_VY-1DKYRE_D1202618_I.B.11 and **B)** Warning sample from H_VY-83775_S1511871_II.G.39i and **C)** Summary sequence quality score showed 161 passed (green lines) and 9 warning (yellow lines) for DIAN and Knight ADRC dataset.

Per tile sequence quality

Because we used an Illumina library that retains its original sequence identifiers, the sequencing output also documents each read’s flowcell tile information. Thus, for Illumina sequencing data FastQC reports quality scores from each tile across all the base positions to see if there was a loss in quality associated with any particular part of the flowcell. The plot shows the deviation from the average quality for each tile. Cold color indicates good quality and warm color indicates bad quality that a tile had worse qualities than other tiles for that base. A passed sample plot should be blue all over. FastQC user manual explains that “reasons for seeing warnings or errors on this plot could be transient problems such as bubbles going through the flowcell, or they could be more permanent problems such as smudges on the flowcell or debris inside the flowcell lane”[11]. My samples all passed the test except one showed warning (**Figure 2.2B**). I observed that there was a quality drop on the end of the reads in some tiles of the flowcells.

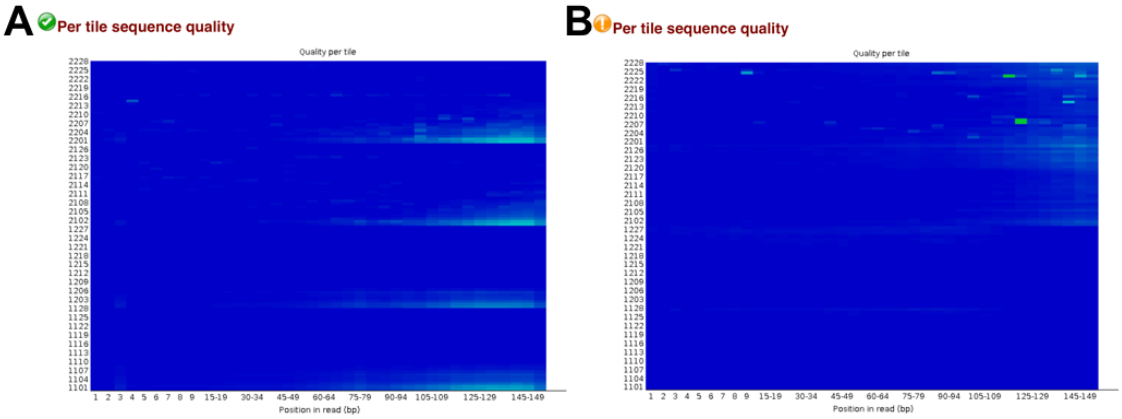


Figure 2.2 Per tile sequence quality. A) passed sample from H_VY-1DKYRE_D1202618_I.B.11 **B)** Warning sample from H_VY-62240_S1511620_IV

Per sequence quality scores

The per sequence quality score report allows me to see the overall quality distribution of my reads and to detect if I have reads with universally low-quality scores. If some reads are poorly imaged, for example, when they are on the edge of the field of view, they will have universally poor quality[11]. My samples showed that the majority of reads are good quality ranging from 28 to 40 (**Figure 2.3**).

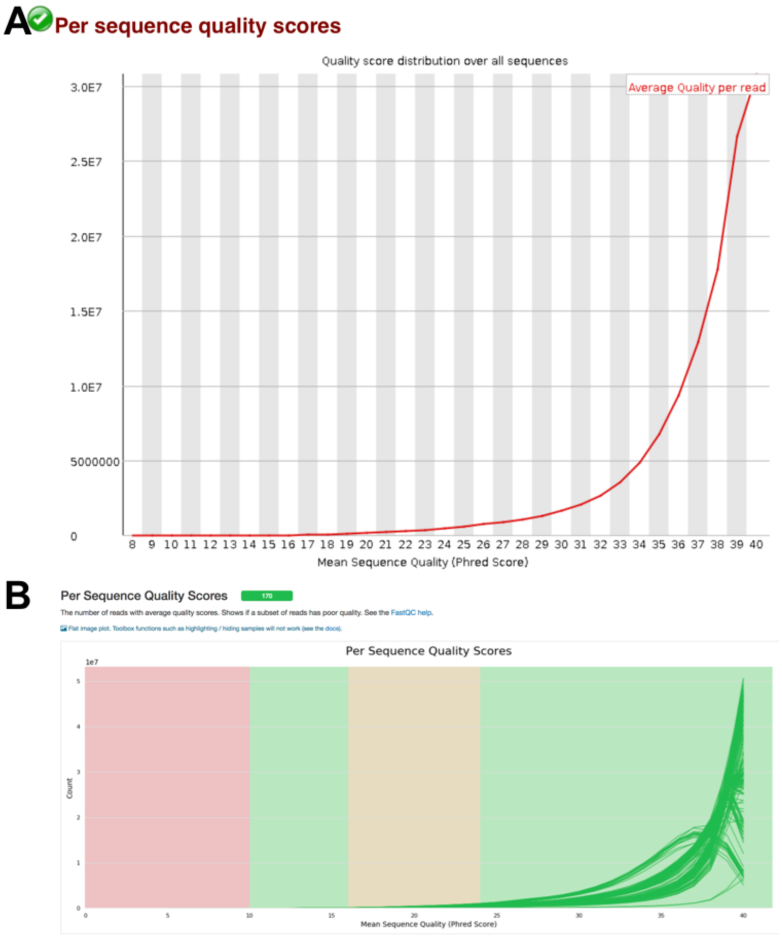


Figure 2.3 Per sequence quality scores. A) passed example from H_VY-1DKYRE_D1202618_I.B.11 **B)** summary results for all samples showed they all passed for this test.

Per Base Sequence Content

Per Base Sequence Content plots out the proportion of DNA bases for each base position, which are the percentage of A, T, C, G. In a random library, I would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base may not be equally 25% for each type of nucleic acid, but should reflect the overall amount of these bases in the sequenced genome and should not be hugely imbalanced from each other in any case[11].

It is worth noting that some types of library will always produce biased sequence composition, normally at the start of the read, which is what is happening in my samples with Illumina. “Libraries produced by Illumina priming use random hexamers (including nearly all RNA-Seq libraries). Those hexamers were fragmented using transposases, which inherit an intrinsic bias in the positions at which reads start. This bias does not concern an absolute sequence, but instead provides enrichment of a number of different K-mers at the 5' end of the reads. Therefore, it is common for Illumina sequencers to have the first 7-9 bases with unbalanced base sequence contents. While this is a true technical bias, it can be corrected by trimming, but in most cases, it does not seem to adversely affect the downstream analysis”[11]. It will however produce a warning or error in this module when the difference between A and T, or G and C is greater than 10% (warning) and 20% (failed) in any position. After the first nine positions, the lines run flat and in parallel with each other indicating balanced and unbiased nucleic acid contents (**Figure 2.4**). Due to the technical bias of Illumina library, this module produced 38 warning and 132 failed test results for the dataset.

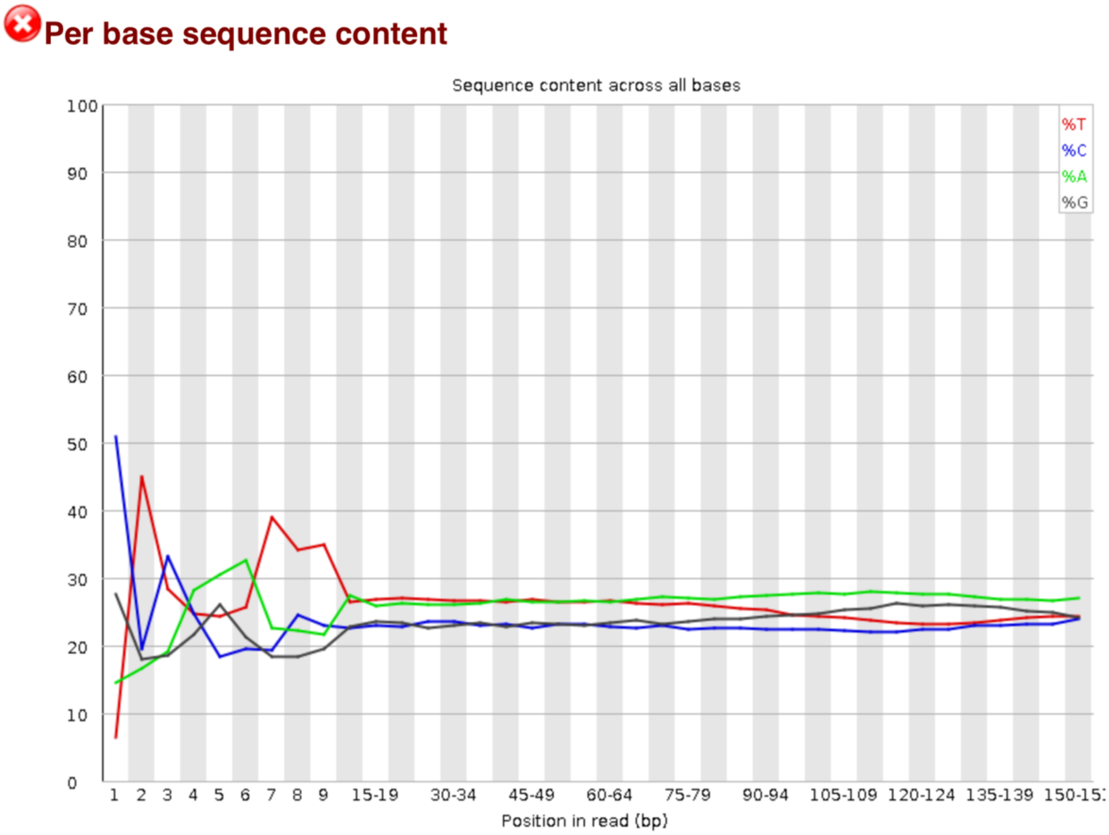
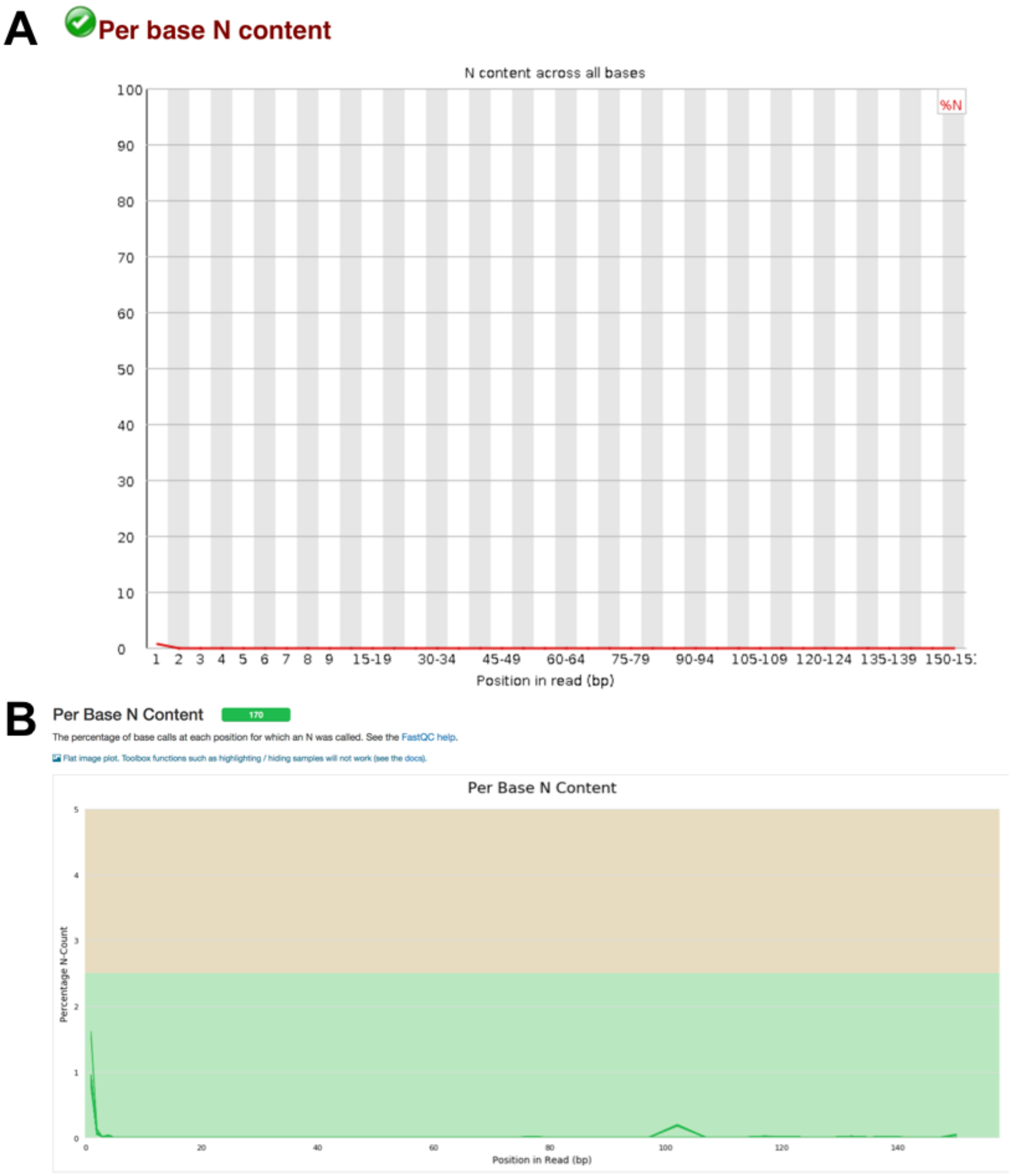


Figure 2.4 Per base sequence content. The percentage of each DNA nucleic acid type for each base position was labeled and color coded respectively.

Per base N content

“If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called”[11]. My samples only showed a very low N content at the beginning of the reads due to unavoidable library technical bias, and at 100bp for some samples (**Figure 2.5**). All samples passed this test.



The percentage of base calls at each position for which an N was called. See the FastQC help.

Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).

Figure 2.5 Per base N content. A) Passed sample from H_VY-1DKYRE_D1202618_I.B.11 and B) summary for all samples.

<u>Sequence length distribution</u>

“Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end. This module generates a graph showing the distribution of fragment sizes in the file which was analyzed”[11]. In many cases this will produce a simple graph showing a peak only at one size, which is the case for my sample shown here (**Figure 2.6**), but for variable length FastQ files this will show the relative amounts of each different size of sequence fragment[11].

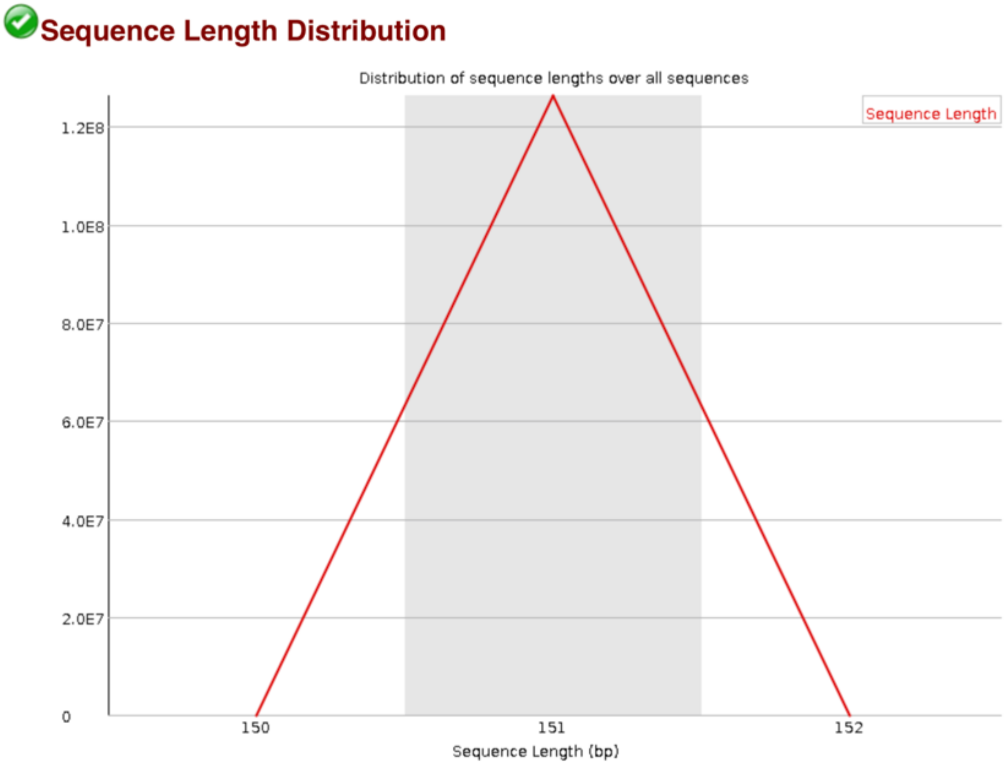


Figure 2.6 Sequence length distribution. Passed sample from H_VY-1DKYRE_D1202618_I.B.11

Adapter content

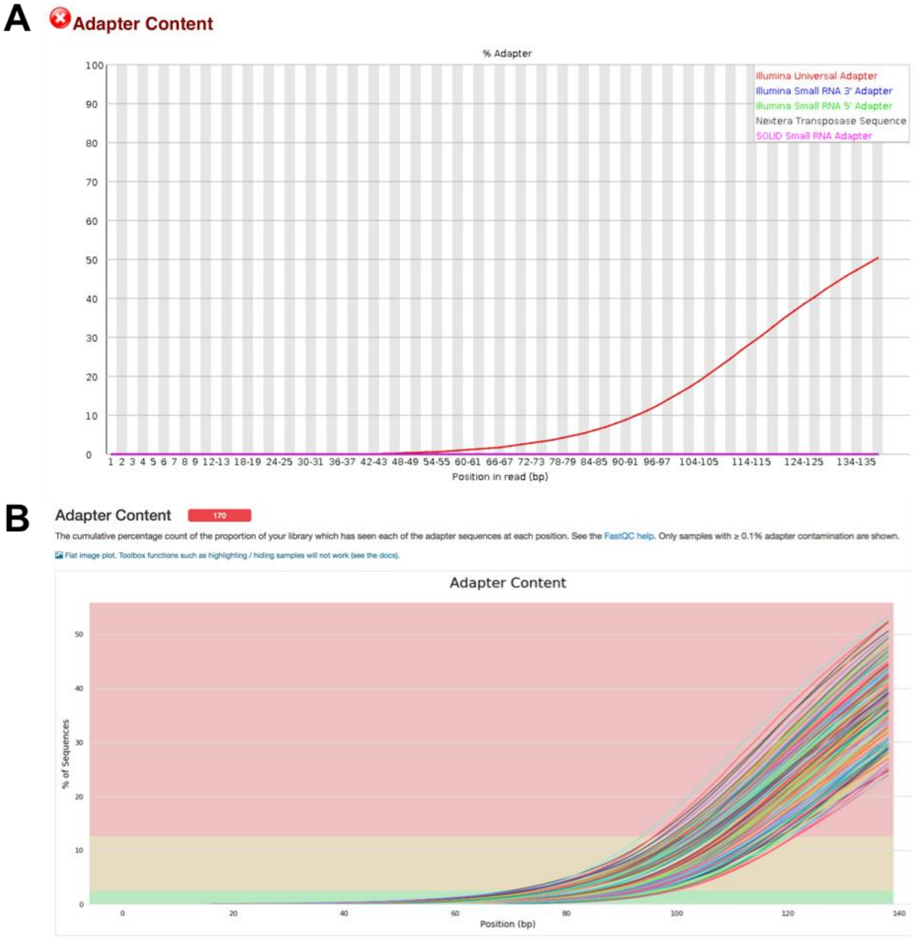
I have adapter contamination due to adapter read-through problem associated with fragmented short reads. Adapter source was predicted and my sample showed potential adapter contamination from Illumina universal adapter (**Table 2.4; Figure 2.7A**). All samples failed this test (**Figure 2.7B**).

Expected observations with adapter dimer contamination[11]:

- Drop in per base sequence quality after base 60
- Possible bi-modal distribution of per sequence quality scores
- Distinct pattern observed in per bases sequence content up to base 60
- Spike in per sequence GC content
- Overrepresented sequence matching adapter
- Adapter content > 0% starting at base 1

Table 2.4 Illumina TruSeq Stranded RNA HT adapter and index

Adapter	
D501-D508 adapter	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
D701-D712 adapter	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG
Index	
i5 index name	i5 index bases
D501	TATAGCCT
D502	ATAGAGGC
D503	CCTATCCT
D504	GGCTCTGA
D505	AGGCGAAG
D506	TAATCTTA
D507	CAGGACGT
D508	GTACTGAC
i7 index name	i7 index bases
D701	ATTACTCG
D702	TCCGGAGA
D703	CGCTCATT
D704	GAGATTCC
D705	ATTCAGAA
D706	GAATTCGT
D707	CTGAAGCT



The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the FastQC help. Only samples with ≥ 0.1% adapter contamination are shown.

Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).

Figure 2.7 Adapter content. A) Failed sample from H_VY-1DKYRE_D1202618_I.B.11 and **B)** summarized for all samples

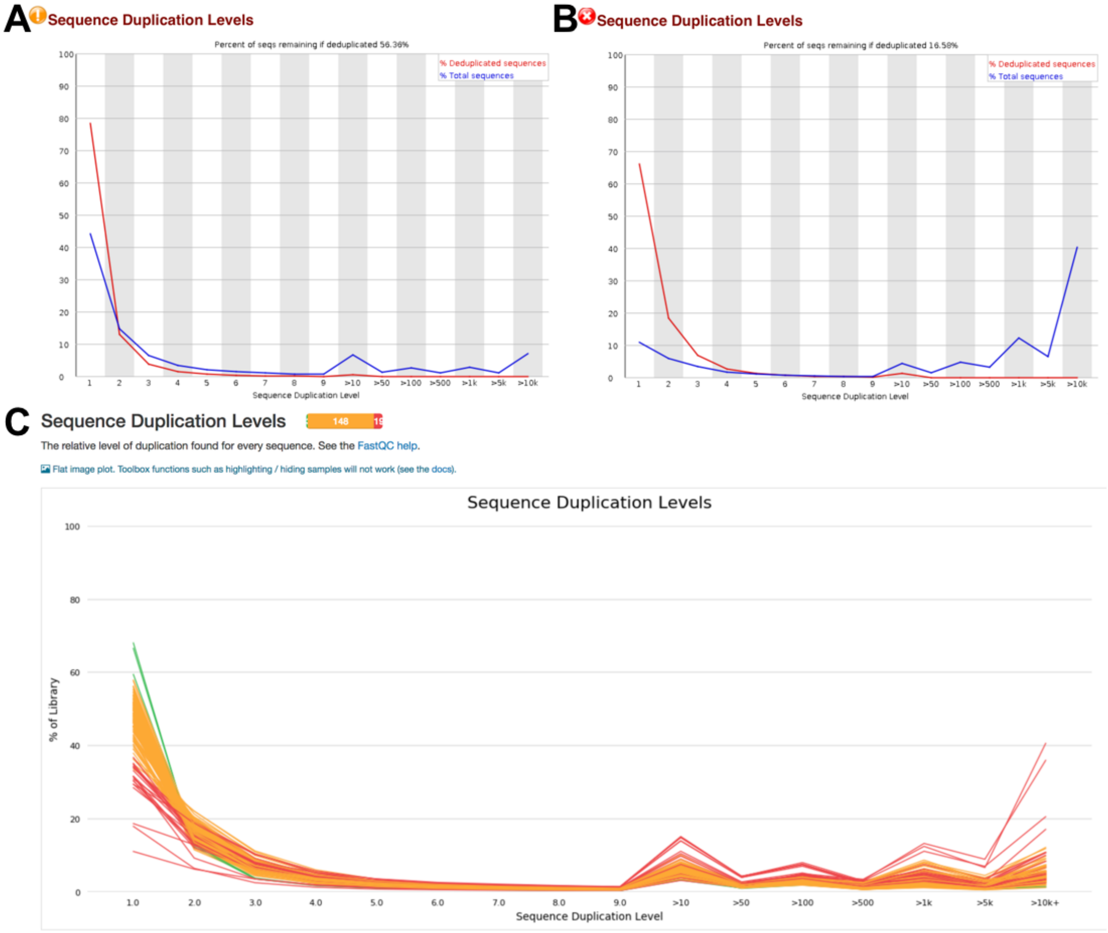
Sequence duplication level

"The plot shows the proportion of the library which is made up of the duplicated sequences in each of the different duplication level bins. There are two lines on the plot. The blue line takes the full sequence set and shows how its duplication levels are distributed. In the red plot the sequences are de-duplicated and the proportions shown are the proportions of the remained different duplication levels in the original data after removing the duplicated sequences. In a properly diverse library most sequences should fall into the far left of the plot in both the red and blue lines. A general level of enrichment, indicating broad over sequencing in the library will tend to flatten the lines”[11], lowering the low end and generally raising other categories (**Figure 2.8B** blue line). More specific enrichments of subsets, or the presence of low complexity contaminants will tend to produce spikes towards the right of the plot (**Figure 2.8A** blue line). “These high duplication peaks will most often appear in the blue trace as they make up a high proportion of the original library, but usually disappear in the red trace as they make up an insignificant proportion of the deduplicated set. If peaks persist in the red trace then this suggests that there are a large number of different highly duplicated sequences which might indicate either a contaminant set or a very severe technical duplication. The module also calculates an expected overall loss of sequence were the library to be deduplicated shown in the figure headline at the top of the plot, which gives a reasonable impression of the potential overall level of loss”[11].

Note that both biological duplication and technical duplication were not differentiated in this analysis and the way to differentiate these two categories is to examine if the duplicated reads are mostly from physically connected genome regions after alignment[11]. High coverage data has more reads so it is not surprised to see higher duplication level. Notice that among the three passed samples, two are low yield samples. I would also expect the majority of duplications are from rRNA so the samples with high rRNA will have high duplication levels, which is proved by comparing the two figures showing one normal rRNA sample (**Figure 2.8A**) and one high rRNA sample (**Figure 2.8B**). In STAR alignment, the default setting for reads aligned to multiple location is 10, and when it’s above 20 it will not map to the reference and those will go to unmapped category, so from the duplication level plot >10 bin percentage I would be able to have a rough estimation of the percentage of reads that would map to multiple location and unmapped percentage of STAR alignment results. For project that look at unmapped section of the samples this duplication level information might be extremely important, for example circular RNA or microbial RNA focused projects.

Notice that among the three passed samples, two are low yield samples:

- H_VY-12152_S1512495.bam
- H_VY-60410_S1511525_I.D.19.bam (low yield)
- H_VY-62240_S1511620_IV.bam (low yield)



The relative level of duplication found for every sequence. See the FastQC help.

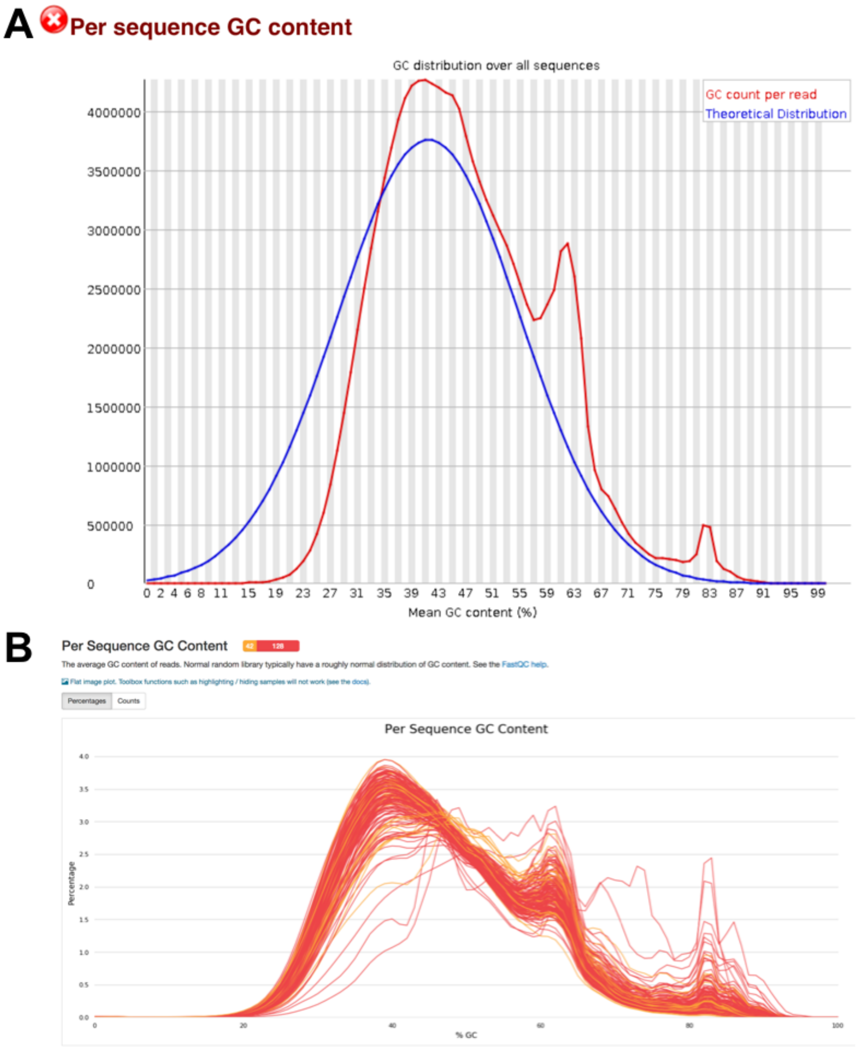
Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).

Figure 2.8 Sequence duplication levels. A) Warning sample from H_VY-1DKYRE_D1202618_I.B.11 and B) failed sample from H_VY-82018_S1512310_II.H.40 and C) summary sample for all samples

Per sequence GC content

“This module measures the GC content across the whole sequence length and compares it to a modelled normal distribution of GC content. In a normal random library, I would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the sequenced genome. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since software doesn't know what the sequenced genome's GC content should be”[11].

This module will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads. Warnings in this module usually indicate a problem with the library. Sharp peaks on an otherwise smooth distribution are normally the result of a specific contaminant (adapter dimers for example, **Figure 2.9A**), which may well be picked up by the overrepresented sequences module. Broader peaks may represent contamination with a different species. In my samples, the distribution showed sharp peaks on an otherwise smooth distribution or severely deviated from normal distribution may indicate adapter dimers contamination, which might be picked by other failed modules, such as overrepresented sequences, adapter content, and kmer content. Overall, I have 42 samples with warning and 128 failed samples (**Figure 2.9B**).



The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the FastQC help.

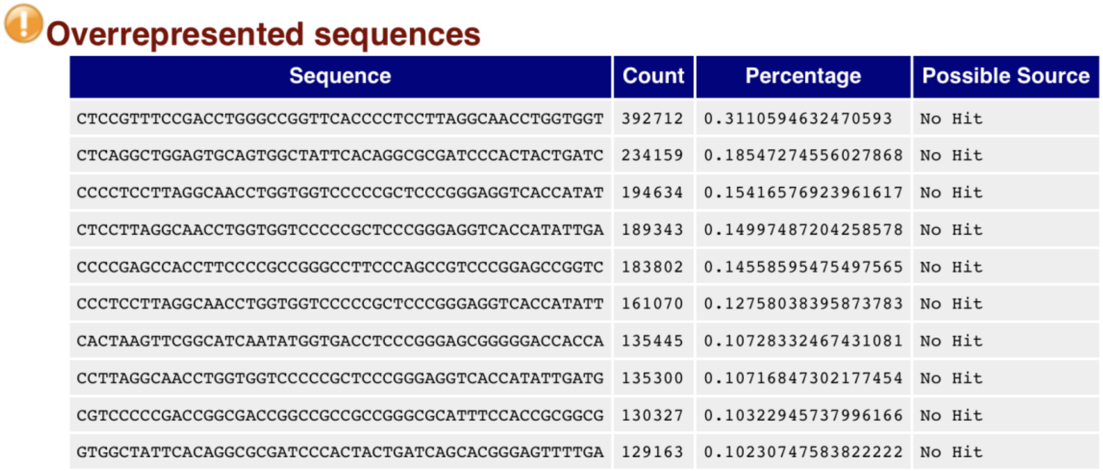
Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).

Figure 2.9 Per sequence GC content. A) Failed sample from H_VY-1DKYRE_D1202618_I.B.11 and B) GC content curves for all sample.

Overrepresented sequences

“A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as expected”[11]. This module lists all of the sequence which make up more than 0.1% of the total (**Figure 2.10**). “To conserve memory only sequences which appear in the first

100,000 sequences are tracked to the end of the file. For each overrepresented sequence the program will look for matches in a database of common contaminants and will report the best hit it finds. Hits must be at least 20bp in length and have no more than 1 mismatch. Finding a hit does not necessarily mean that this is the source of the contamination, but may point me in the right direction. Because the duplication detection requires an exact sequence match over the whole length of the sequence any reads over 75bp in length are truncated to 50bp for the purposes of this analysis. Even so, longer reads are more likely to contain sequencing errors which will artificially increase the observed diversity and will tend to underrepresent highly duplicated sequences”[11].



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGTGGT	392712	0.3110594632470593	No Hit
CTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	234159	0.18547274556027868	No Hit
CCCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATAT	194634	0.15416576923961617	No Hit
CTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTGA	189343	0.14997487204258578	No Hit
CCCCGAGCCACCTTCCCCGCCGGGCCTTCCCAGCCGTCCCGGAGCCGGTC	183802	0.14558595475497565	No Hit
CCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATT	161070	0.12758038395873783	No Hit
CACTAAGTTCGGCATCAATATGGTGACCTCCCGGGAGCGGGGGACCACCA	135445	0.10728332467431081	No Hit
CCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTGATG	135300	0.10716847302177454	No Hit
CGTCCCCCGACCGGCGACCGGCCGCCGCCGGGCGCATTTCCACCGCGGCG	130327	0.10322945737996166	No Hit
GTGGCTATTCACAGGCGCGATCCCACTACTGATCAGCACGGGAGTTTTGA	129163	0.10230747583822222	No Hit

Figure 2.10 Overrepresented sequences. Warning sample from H_VY-1DKYRE_D1202618_I.B.11.

<u>Kmer Content</u>

The Kmer Content module will do a generic analysis of all of the Kmers in my sample library to find those which do not have even coverage through the length of my reads (**Figure 2.11**). The top six kmers are plotted in the graph. This can find a number of different sources of bias in the library which can include the presence of read-through adapter sequences building up on the end of the sequences. The presence of any overrepresented sequences in my library (such as adapter dimers) will cause the Kmer plot to be dominated by the Kmers these sequences contain. What I have in my samples are two folds: 1. Unbalanced sequence content for the first 9 bases which is intrinsic technical bias associated with Illumina library. 2. Adapter read through problem at the end of the reads which can be corrected with STAR alignment using adapter soft clipping option.

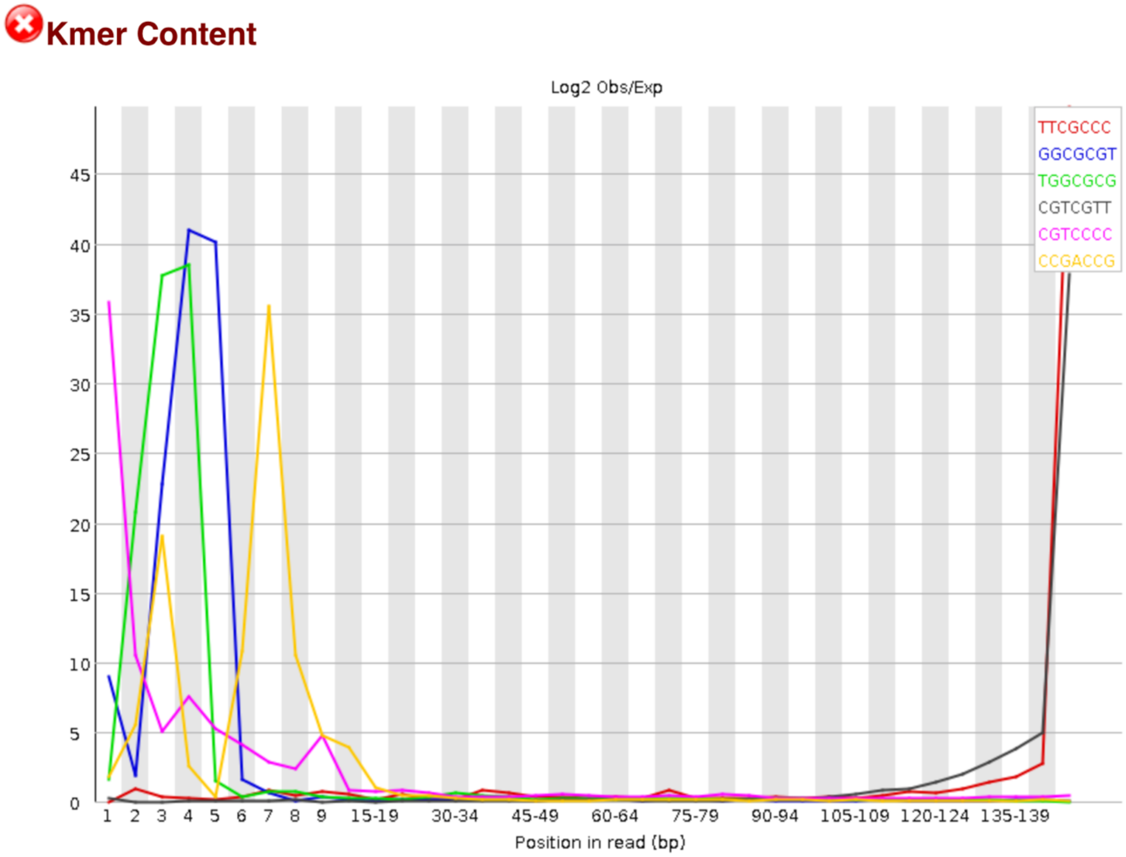


Figure 2.11 Kmer Content. Failed sample from H_VY-1DKYRE_D1202618_I.B.11.

In summary, my DIAN and Knight ADRC samples have good sequence quality in general, reflected in categories such as per base, per tile, per sequence quality scores, and per base N content (**Table 2.5**). Because we used Illumina sequencing, the first nine bases of each read have technical bias, therefore, none of the samples passed per base sequence content test. However, it does not impact downstream analysis. Due to potential rRNA and adapter contamination, several test metrics captured these observations, for example, per sequence GC content, sequence duplication levels, overrepresented sequences, adapter content, and kmer content. Samples with high rRNA contamination were not different from the other samples in terms of deconvolution results, but will be excluded from other downstream analysis. Adapter sequences could be soft clipped during alignment or trimmed ahead of alignment.

Table 2.5 DIAN and Knight ADRC FastQC summary

	Pass	Warning	Fail
Basic Statistics	170	0	0
Per base sequence quality	161	9	0
Per tile sequence quality	169	1	0
Per sequence quality scores	170	0	0
Per base sequence content	0	38	132
Per sequence GC content	0	42	128
Per base N content	170	0	0
Sequence Length Distribution	170	0	0
Sequence Duplication Levels	3	148	19
Overrepresented sequences	11	159	0
Adapter Content	0	0	170
Kmer Content	0	0	170

IGV visualization and IBD to verify sample identity

Aligned and sorted bam files were loaded into IGV[224] to perform visual inspection of target variants. For example, visualization of a *PSEN1*_S290C delE9 carrier in the top red track is compared to a non-carrier in the bottom blue track shown in **Figure**

2.12 using IGV sashimi plot. DelE9 is a mutation (Chr14: 73673093 G>A) in *PSEN1* gene that result in exon 9 exclusion. In this carrier subject, in whom harbors heterozygous exon 9 deletion, the peak in the middle represents gene expression level of *PSEN1* exon 9 flanked by exon 8 and exon 10. Because this carrier is heterozygous for delE9, the exon 9 peak is half in height compared to its neighbors on either side, and there are more reads (79 reads) that skipped exon 9 and linked exon 8 and exon 10 compared to the non-carrier in the bottom track (0 read). Samples carrying unexpected variants or missing expected variants were labeled as potential swapped samples. In addition, variants were called from RNA-Seq following BWA/GATK pipeline[171, 181]. The identity of the samples was later verified by performing IBD analysis against genomic typing from GWAS chipsets.

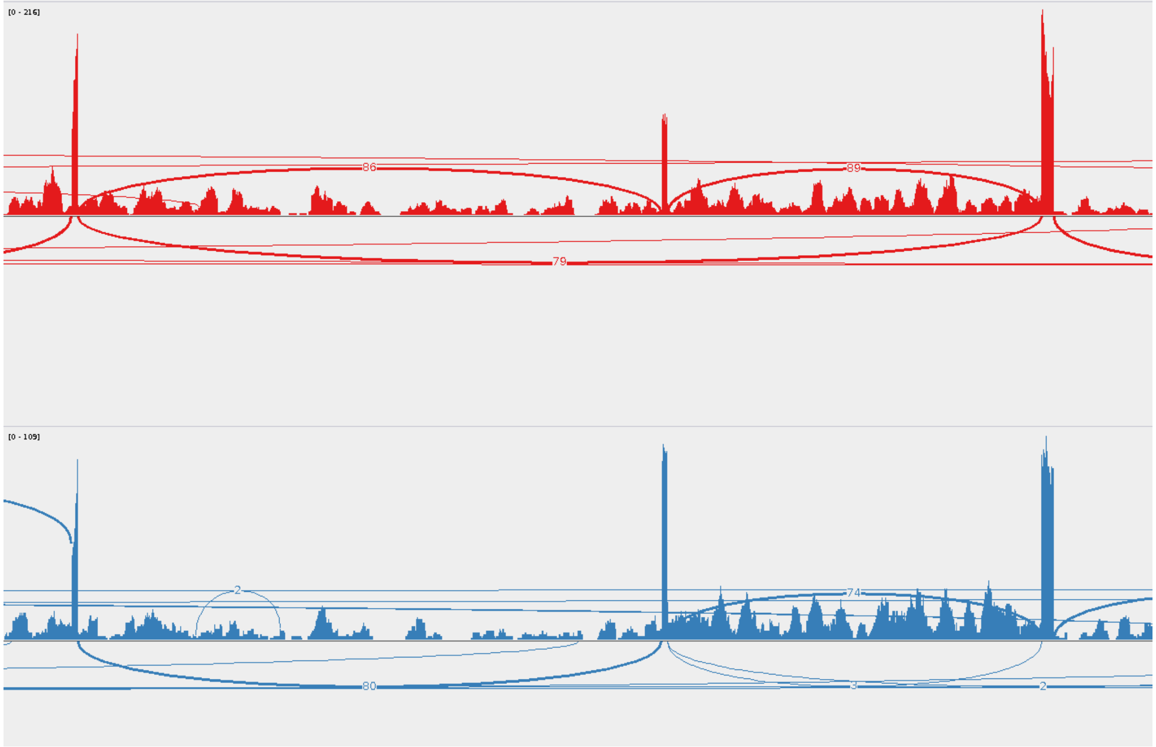


Figure 2.12 ***PSEN1*** **delE9 Sashimi Plot using IGV.** Top red track is from sample H_VY-2968OM_S1512443_I.B.13, a PSEN1_S290C delE9 carrier, in whom harbors heterozygous exon 9 deletion. Bottom blue track is from a non-carrier subject.

GWAS PCA and Ethnicity Check

GWAS principal component analysis (PCA) components were extracted for matched RNA-Seq subjects to check ethnic identity. Among RNA-Seq subjects there are 6 subjects are African Americans, while the rest are European Americans (**Figure 2.13**).

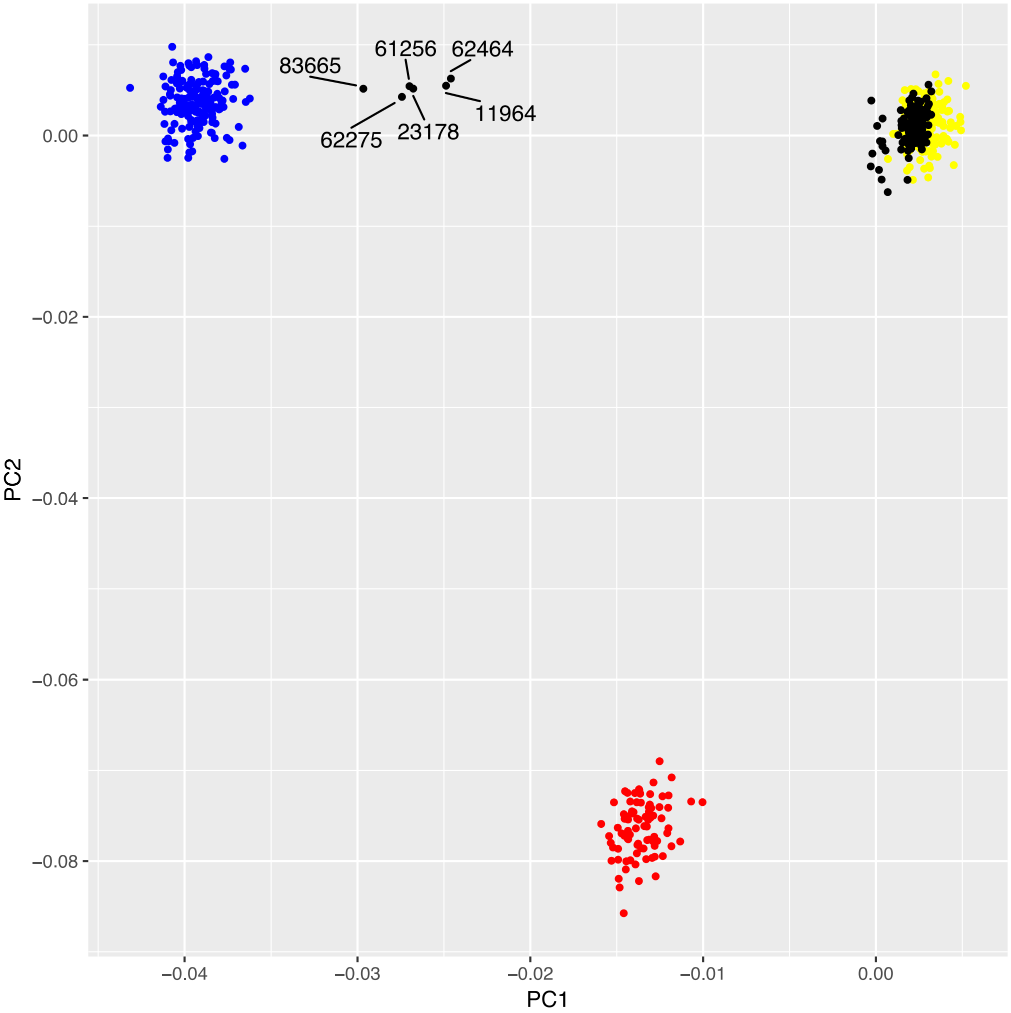


Figure 2.13 GWAS PCA of ethnicity check. HapMap Europeans are color coded as yellow; HapMap Africans are color coded as blue; HapMap Asians are color coded as red. My samples are color coded as black, which mostly fall into the European ethnic group except six subjects labeled with their IDs.

2.3.3 Expression quantification

I applied Salmon transcript expression quantification (ver 0.7.2)[208] to infer the gene expression for all samples included in the reference panel and participants in the Mayo, MSBB, DIAN and Knight-ADRC. I quantified the coding transcripts of *Homo Sapiens* included in the GENCODE reference genome (GRCh37.75). Similarly, I quantified the expression of the mice samples included in the reference panel using the *Mus Musculus* reference genome (mm10).

2.3.4 Reference panel

Reference Samples

I assembled a cell-type specific reference panel from publicly available RNA-Seq datasets comprised of both immunopanning collected or iPSC derived neurons, astrocytes, oligodendrocytes, and microglial cells from human and murine samples. For immunopanning collected cells, antibodies for cell-type specific antigens were utilized to bind and immobilize their targeted cell types in order to immunoprecipitate and purify each cell type from the suspensions[290]. cDNA synthesis was accomplished using Ovation RNA-Seq system V2 (Nugen 7102) and library prepared with Next Ultra RNA-Seq library prep kit from Illumina (NEB E7530) and NEBNext® multiplex oligos from Illumina (NEB E7335 E7500). TruSeq RNA Sample Prep Kit (Illumina) was used to prepare library for paired-end sequence on 100ng of total RNA extracted from each sample. Illumina HiSeq 2000 Sequencer was used to sequence all libraries[290].

Both human adult temporal cortex tissue, collected from patients receiving neurological surgeries, and mice cells were disassociated, sorted and sequenced as described elsewhere[291], and deposited in the Gene Expression Omnibus GSE73721 and GSE52564. I also accessed neural progenitor cells (day 17) and mature human neurons (day 57 and 100) from Broad iPSC deposited in the AMP-AD portal[7] and neural progenitor cells and iPSC-derived neurons from[37]. Broad iPSC derived neurons accessed from AMP-AD portal were generated using an embryoid body-based protocol to differentiate into forebrain neurons[1]. Wild-type cells used in the protocol were obtained from UConn StemCell Core. RNA was purified using PureLink RNA mini-kit (Life Technologies) and libraries were prepared by Broad Institute's Genomics Platform using TruSeq protocol. Please refer to **Table 2.6** for additional information.

Table 2.6 Reference samples for each cell type. GEO accession numbers for cell-type specific samples.

	Reference Sample			
Type	**Human**		**Mouse**	**Human iPSC**[a]
Neuron	GSM1901333		GSM1269905	YZ2-100day
			GSM1269906	YZ3-100day
				YZ4-100day
				YZ5-100day
Astrocyte	GSM1901309	GSM1901317	GSM1269903	Astrocyte1
	GSM1901310	GSM1901318	GSM1269904	Astrocyte2
	GSM1901311	GSM1901319		
	GSM1901312	GSM1901320		
	GSM1901313	GSM1901321		
	GSM1901314	GSM1901322		
	GSM1901315	GSM1901323		
	GSM1901316	GSM1901324		
Oligodendrocyte	GSM1901335		GSM1269911	
	GSM1901336		GSM1269912	
	GSM1901338			
Microglia	GSM1901339		GSM1269913	
	GSM1901340		GSM1269914	
	GSM1901341			

[a] Samples accessed from the Broad iPSC cell-lines deposited in the AMP-AD.

Marker Genes

The reference panel was assembled with samples from four distinct cell types. A redundant set of well-known cell-type markers was selected from the literature[41, 131, 291] (**Table 2.7**). Principal component analysis was performed on the reference panel using R function *prcomp* (version 3.3.3) to verify that the expressions of these gene were clustering samples by their cell types (**Figure 2.14b; Figure 2.15a**).

Table 2.7 Gene markers for principal brain cell types.

	Cell Marker	
Type	**Human**	**Mouse**[a]
Neuron	STMN2	Stmn2
	SYN1	Syn1
	SYT1	Syt1
	GAD1	Gad1
	CCK	Cck
Astrocyte	GFAP	Gfap
	ALDH1L1	Aldh1l1
	AQP4	Aqp4
	GJA1	Gja1
	SOX9	Sox9
Oligodendrocyte	MOG	Mog
	MOBP	Mobp
	SOX10	Sox10
	GPR37	Gpr37
Microglia	TLR2	Tlr2
	CX3CR1	Cx3cr1
	IL1A	Il1a

[a] Mouse homologous genes were identified from Mouse Genome Database.

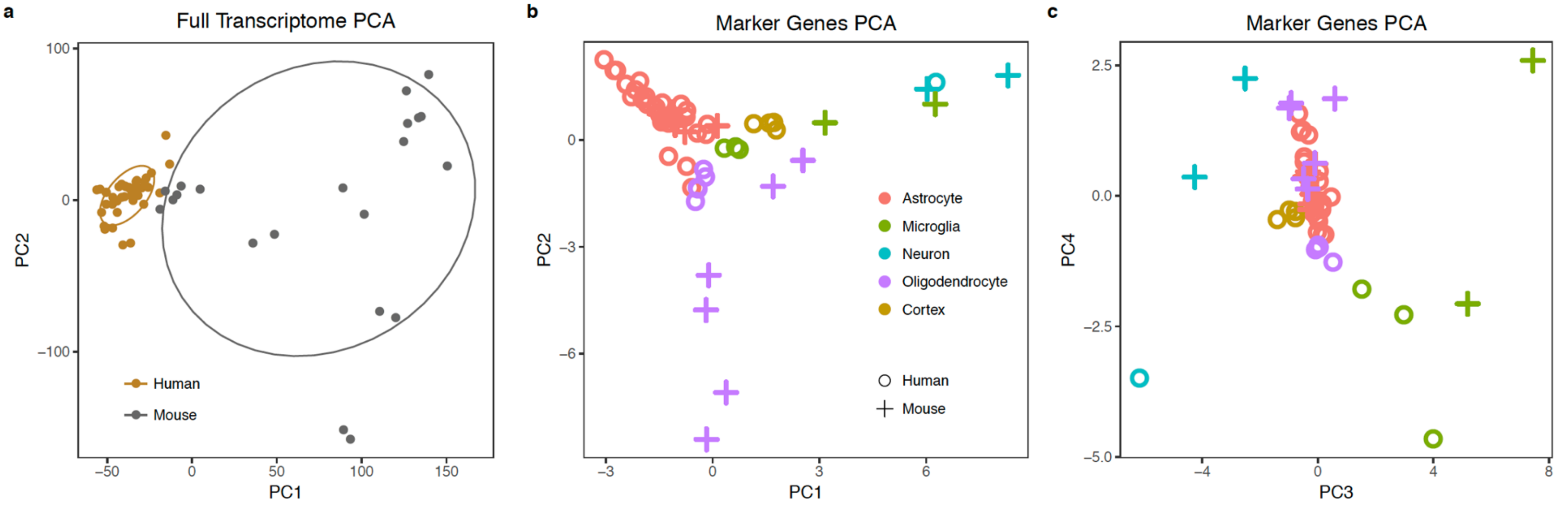


Figure 2.14 PCA of samples included in the reference panel. a) Transcriptome-wide. Genes included in the reference panel b) PC1 vs PC2 and c) PC3 vs PC4.

2.3.5 Inference of the cellular population structure

I ascertained alternative computation deconvolution algorithms implemented in the CellMix package (ver 1.6). Based on accuracy and robustness evaluation results I compared and reported the following three algorithms that outperformed the others: Digital Sorting Algorithm (named “DSA”)[293], which employs linear modeling to infer cell distributions; the method population-specific expression analysis (PSEA, also named meanProfile in CellMix implementation)[157] that calculates estimated expression profiles relative to the average of the marker gene list for each cell type[157]; and a semi-supervised learning method that employs non-negative matrix factorization (ssNMF in CellMix implementation)[103]. I employed a leave-one-out cross-validation procedure to evaluate the accuracy provided by each method. The best performing algorithm ssNMF integrates cell-type marker genes to resolve the drawbacks of completely unsupervised standard non-negative matrix factorization. I followed the standard procedure described in the CellMix package, that included the extraction of marker genes from the reference samples (function extractMarkers from the CellMix package), and the posterior invocation of the function *ged* to infer cellular population from the gene expression of bulk RNA-Seq data. Besides, I tested additional methods which provided considerably lower accuracy (least-squares fit[8], quadratic programing[108]) or no significant difference (support vector regression[199] or latent variable analysis[51]) to the methods presented.

I selected the reference samples that provide the most faithful transcriptomic profile for their respective cell types by following a leave-one-out cross validation approach. I trained iteratively deconvolution models using all but one of the samples that was tested. Only samples predicted with a composition higher than 80% were kept for the reference panel (**Table 2.6; Figure 2.15b**).

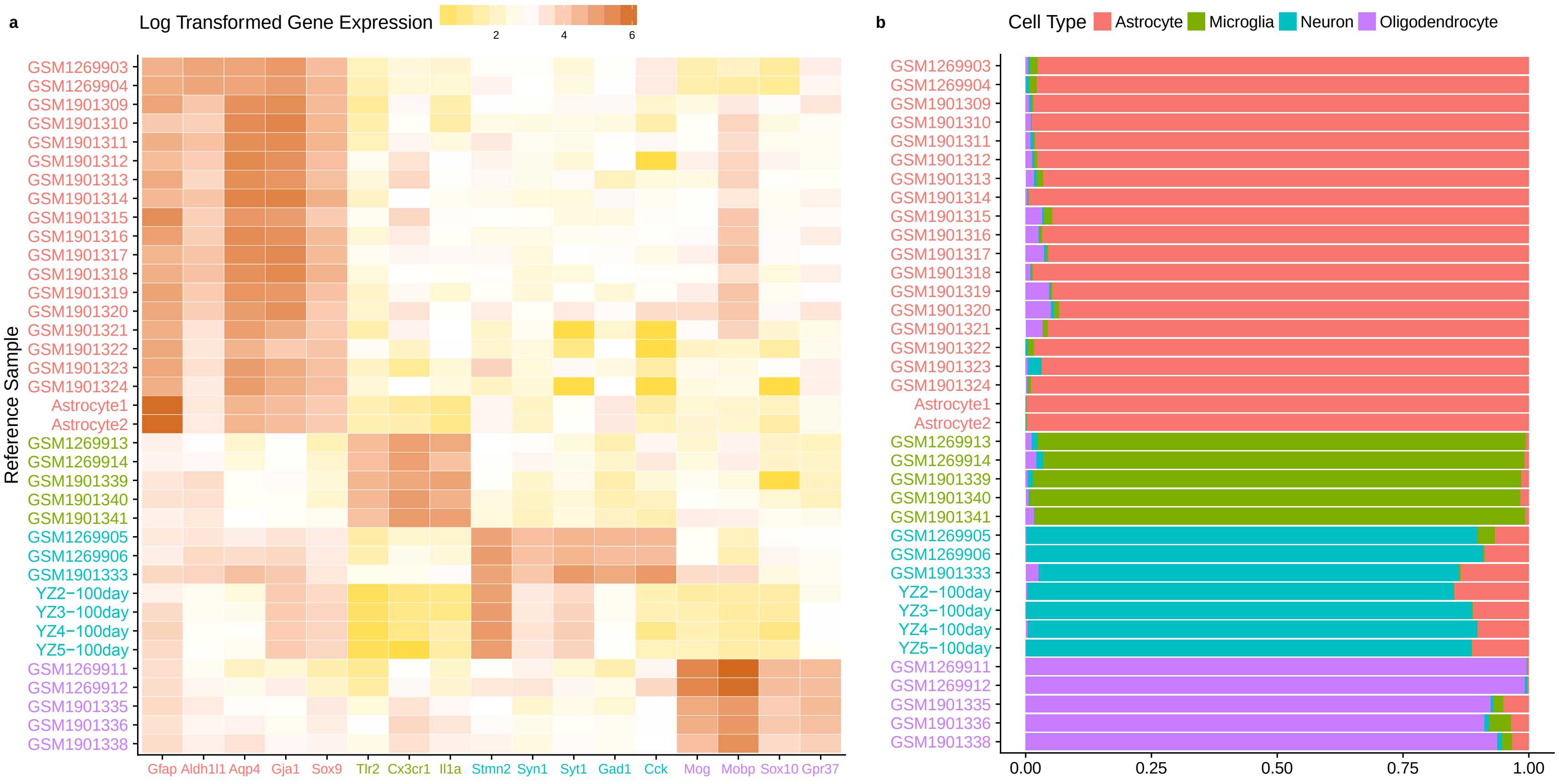


Figure 2.15 Leave-one-out evaluation of reference panel. a) Gene expression levels (log-transformed) for reference panel. The cell types of the isolated/iPSC-derived samples are color-coded and labeled on the y-axis. b) A leave-one-out procedure to obtain the cellular proportion for each of the samples of the reference panel was used. Cell-type proportions are shown as stacked percentage (red: astrocytes; green: microglia; blue: neuron; purple: oligodendrocyte).

2.3.6 Accuracy and Robustness Evaluation

Chimeric validation

To emulate heterogeneous tissue with known and controlled cellular composition, I generated chimeric libraries pooling reads (to a total of 400,000) contributed from the human reference samples (See **Table 2.6**). This process was repeated 720 times, using alternative reference samples to model each cell type. The proportion of reads that the libraries of neurons, astrocytes, oligodendrocytes and microglia provided to the chimeric libraries varied in predefined ranges (**Figure 2.16**). As a result, each of the chimeric libraries contained reads that followed 32 different distributions (neuronal reads contributed between 2 to 36% of reads, astrocytes between 22 to 76%, oligodendrocytes between 6 to 62% and microglia between 1 to 5%). Refer to **Table 2.8** for detailed description of the 32 different distributions. I quantified the chimeric reads using Salmon (v0.7.2)[208], and employed the reference samples that did not contribute reads to the chimeric library as reference panel for the deconvolution methods.

Overall, I applied my digital deconvolution analyses to 23,040 (720 × 32) chimeric libraries. I evaluated the accuracy using the root-mean-square error (RMSE, **Equation 2.1** to compare the digital deconvolution cellular proportion estimates (method ssNMF) versus the defined proportion of reads specific to each of the chimeric libraries:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}i - yi)^2}{n}} \qquad \textbf{(Equation 2.1)}$$

$$\hat{y}i - estimated\ value, yi - observed\ value$$

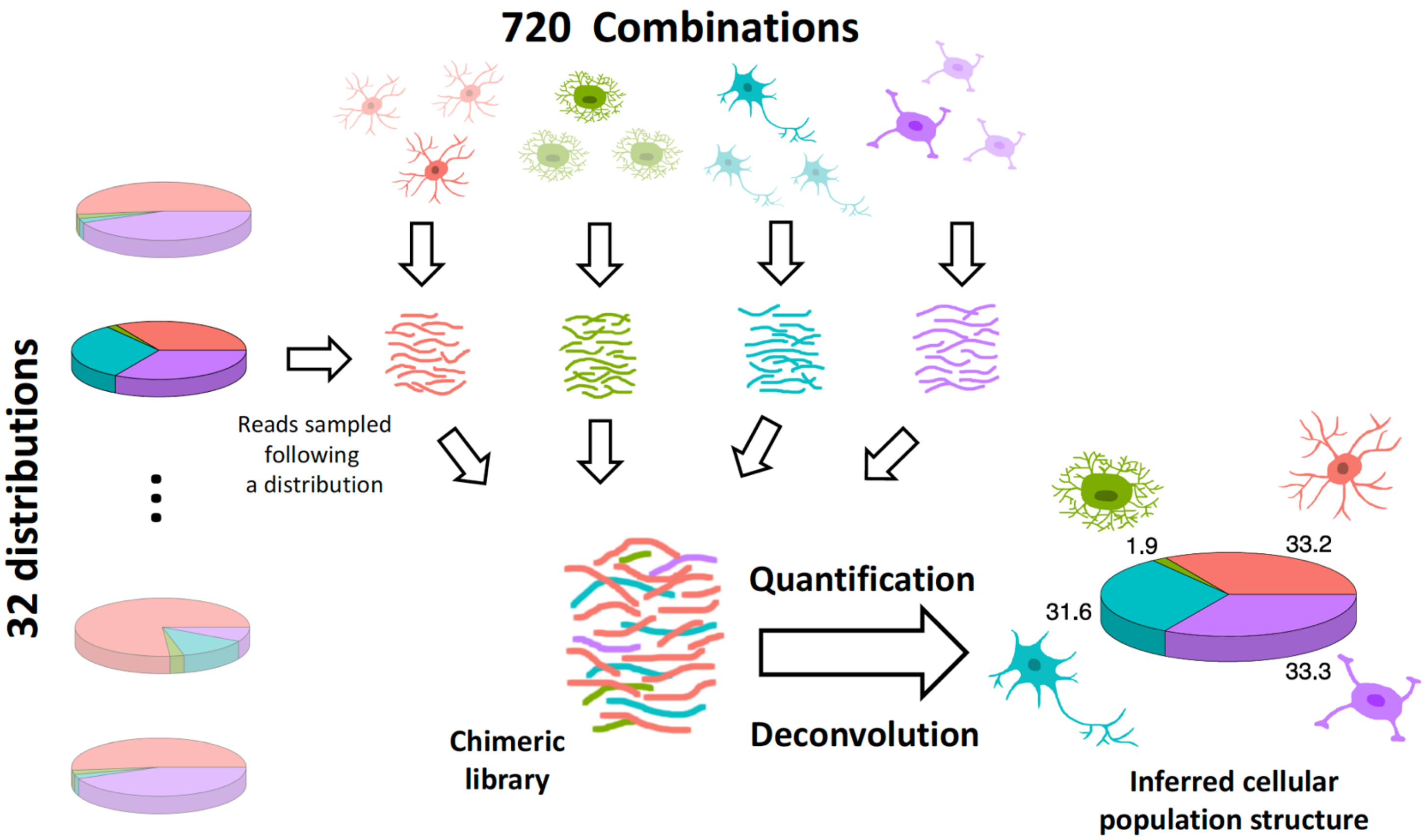


Figure 2.16 Chimeric library deconvolution simulation. Human cell-type specific reference samples contributed 720 different combinations to generate chimeric libraries. Reads were randomly sampled following 32 pre-specified distributions (Neuronal reads contributed between 2 to 36% of reads, astrocytes between 22 to 76%, oligodendrocytes between 6 to 62% and microglia between 1 to 5%). Each chimeric library was quantified and the cellular distribution estimated using digital deconvolution. These estimates were compare to prior distribution.

Table 2.8 Simulated chimeric tissue cell-type composition. Percentages of reads contributed to the synthetic chimeric libraries.

		Percentage of reads			
		Neuron	**Astrocyte**	**Oligodendrocyte**	**Microglia**
Configurations	**Neuron**	**2.31**	51.7	43.5	2.49
		10	72	15.2	2.81
		10.1	44.7	42.9	2.31
		18.9	65.9	12.4	2.85
		18.9	41.6	37	2.53
		25.4	57.6	14.5	2.49
		31.6	33.2	33.3	1.9
		35.9	46	15.6	2.51
	Astrocyte	20.3	**22.2**	55.8	1.75
		17.2	**34.6**	46.2	2.01
		33.6	**43.1**	20.7	2.52
		15.6	**47.8**	34.1	2.53
		24.2	**56.7**	16.5	2.59
		19.2	**65.5**	12.3	2.99
		9.81	**66.7**	20.6	2.95
		12.9	**76.4**	8.1	2.59
	Oligodendrocyte	18.2	72.2	**6.86**	2.8
		23.3	62.7	**11.2**	2.87
		25	57.3	**15.1**	2.56
		18.2	61.6	**17.3**	2.94
		23.7	49.2	**24.6**	2.44
		20.3	45.5	**31.8**	2.5
		14.7	37.8	**45.4**	2.13
		9.77	26.3	**62.2**	1.67
	Microglia	17.7	29.9	51.4	**1.01**
		23.8	55.4	19.4	**1.38**
		17.5	41.9	39	**1.67**
		24.7	60.9	12.3	**2.01**
		15	51.2	31.4	**2.47**
		20.1	64.8	12.3	**2.71**
		12.7	48.3	34.9	**4.09**
		21.3	60.6	13.5	**4.57**

For each target cell type the distribution was pre-defined to cover a broad range of biological viable proportions (highlighted in bold).

I also tested whether the deconvolution results were dominated by the expression of any specific marker gene and ascertained the robustness of the inferred cellular population structure to any possibly altered expression of marker genes. To do so, I performed the deconvolution analysis discarding each of the marker genes one at a time and evaluated how these distributions differed in comparison to the full gene reference panel.

2.3.7 Statistical Analysis

Transcriptome PCA and covariate analysis

Number of reads gene quantification results from Salmon were normalized with DeSeq2's VarianceStabilizingTransformation function after removing genes with total reads count less than 10,000. First 10 principle components were extracted from PCA, and single and stepwise covariate analysis using linear model were performed to investigate what covariates would affect data quality and downstream analysis. RNA quality (RIN and DV200), post-mortem index (PMI), sequencing pooling, sex, age at death, and brain tissue origin were included in either single covariate and stepwise covariate analysis. The results showed RNA quality measurements RIN and DV200, and age at death are the most important confounding factors that impact the analysis. Covariate correlation was performed to examine any correlation among the factors to avoid including highly correlated covariates that lead to over correction of the model, for example, RIN and DV200 are associated with quality but they are highly correlated as well. Since RIN and DV200 are highly correlated, using RIN only is recommended because other publicly ascertained datasets only have RIN measurements. (**Figure 2.17**). Later in **Chapter 3**, I will replace RIN for transcript integrity number (TIN), which will be inferred directly from RNA-Seq data. Notice that ribosomal RNA contents are negatively correlated with uniquely mapped reads, because those rRNA will be mapped to multiple locations due to its highly conserved and similar sequences. Besides, uniquely mapped reads also anti-correlated with incorrect strand reads, percentage of reads that mapped to multiple loci, and median 5' to 3' bias. The neuronal and astrocyte proportion I inferred from the RNA-Seq data are also highly negatively correlated.

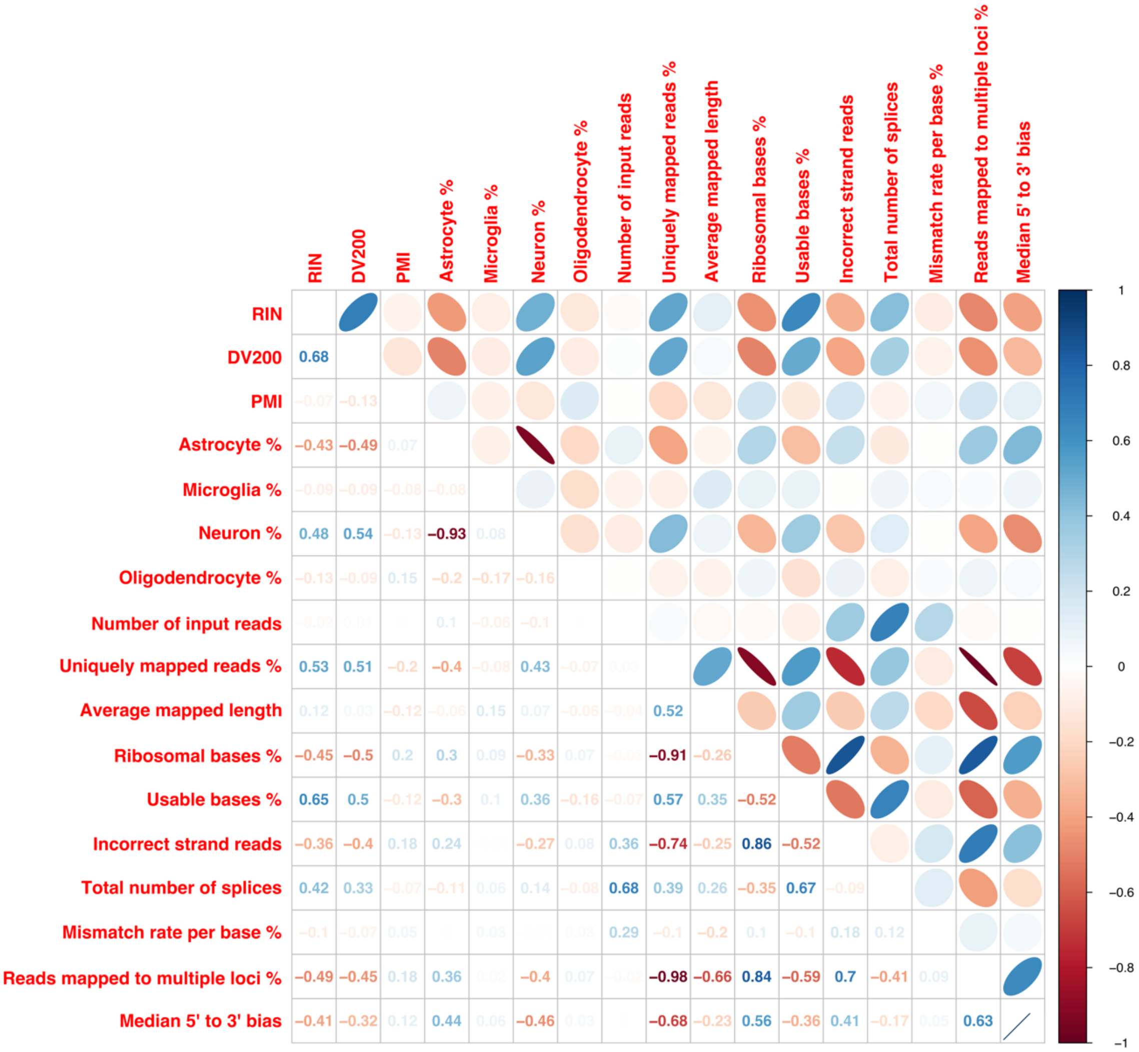


Figure 2.17 Covariate correlation. Major covariates of RNA-Seq and inferred cell type compositions from deconvolution were correlated to avoid including highly correlated covariates in downstream analysis.

Transcriptome-wise PCA showed seven subjects are very different from the rest and they appeared as outlier on PC1 vs PC2 plot (**Figure 2.18**). iPSC samples were excluded from this analysis and analyzed separately in another study[143]. Among the seven outliers, 60007 and 83774 also have high median 5' to 3' bias. The top 30 genes that contributed the most to PC1 and top 30 to PC2 were extracted and plotted as a heatmap to show the dramatic difference between these outlier subjects and the rest of the samples (**Figure 2.19**). Notice that the left most 5 subjects are the outliers on transcriptome-wise PCA, but the other two outliers 83774 and 60007 do not have an obvious clustering pattern in the heatmap, suggesting that high median 5' to 3' bias also contribute variances observed in PC1 and PC2.

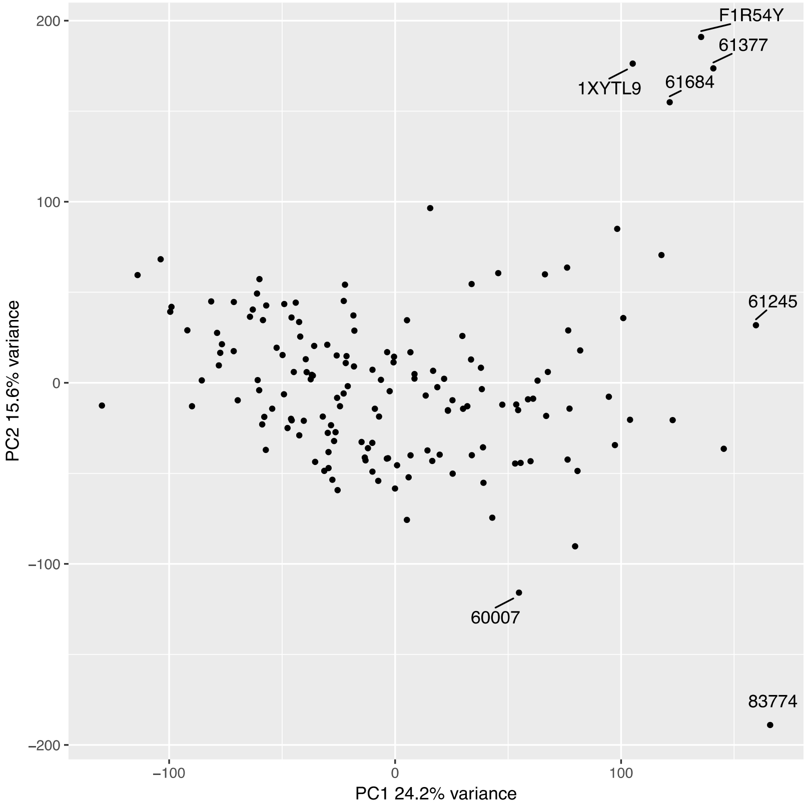


Figure 2.18 Transcriptome-wise principle component analysis. Subject transcriptome-wise PCA were plotted using PC1 and PC2 with outliers labeled with subject ID.

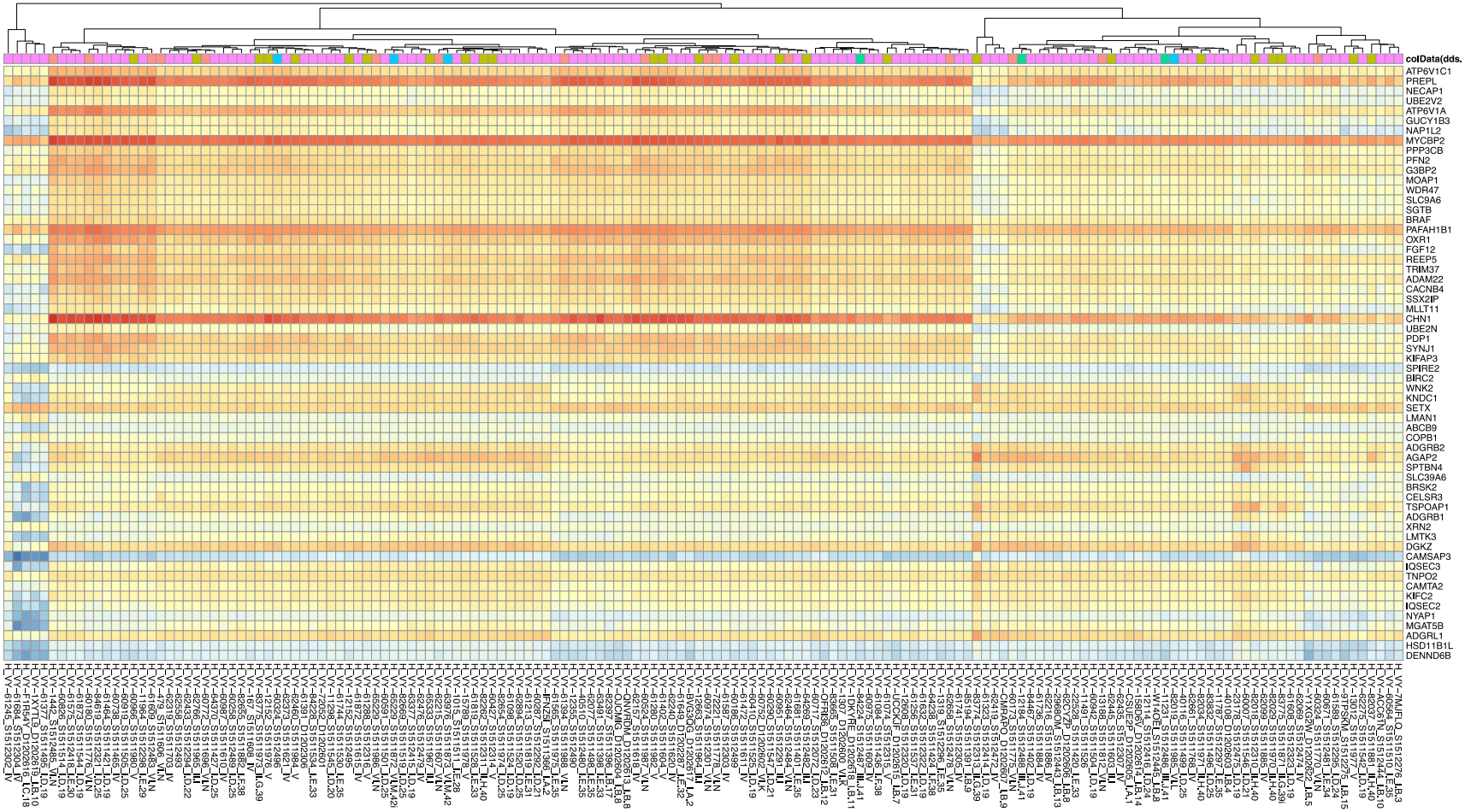


Figure 2.19 Top 60 Genes for PC1 and PC2 heatmap. The left most 5 subjects are the outliers on transcriptome-wise PCA, and they are clustered as a separate group in the top 30 genes contributing the most variance to PC1 and PC2. The other two outliers 83774 and 60007 also have high median 5’ to 3’ bias that do not have an obvious clustering pattern in the graph.

Cell type proportions and disease status association analysis

I employed linear regression models to test the association between cell-type proportions and disease status (R Foundation for Statistical Computing, ver.3.3.3). Stepwise discriminant analysis (stepAIC function of R package MASS, version 7.3-45) was used to determine significant covariates, and correct for confounding effects. I included RNA integrity number (RIN), batch, age at death and post-mortem interval (PMI) as covariates for the Mayo Clinic analyses. For Mount Sinai Brain Bank analyses, I corrected for RIN, PMI, race, batch and age at death. I also used linear-mixed models to perform multiple-region association analysis, employing random slopes and random intercepts grouping by observations and by donors[253], and correcting for the same covariates previously described.

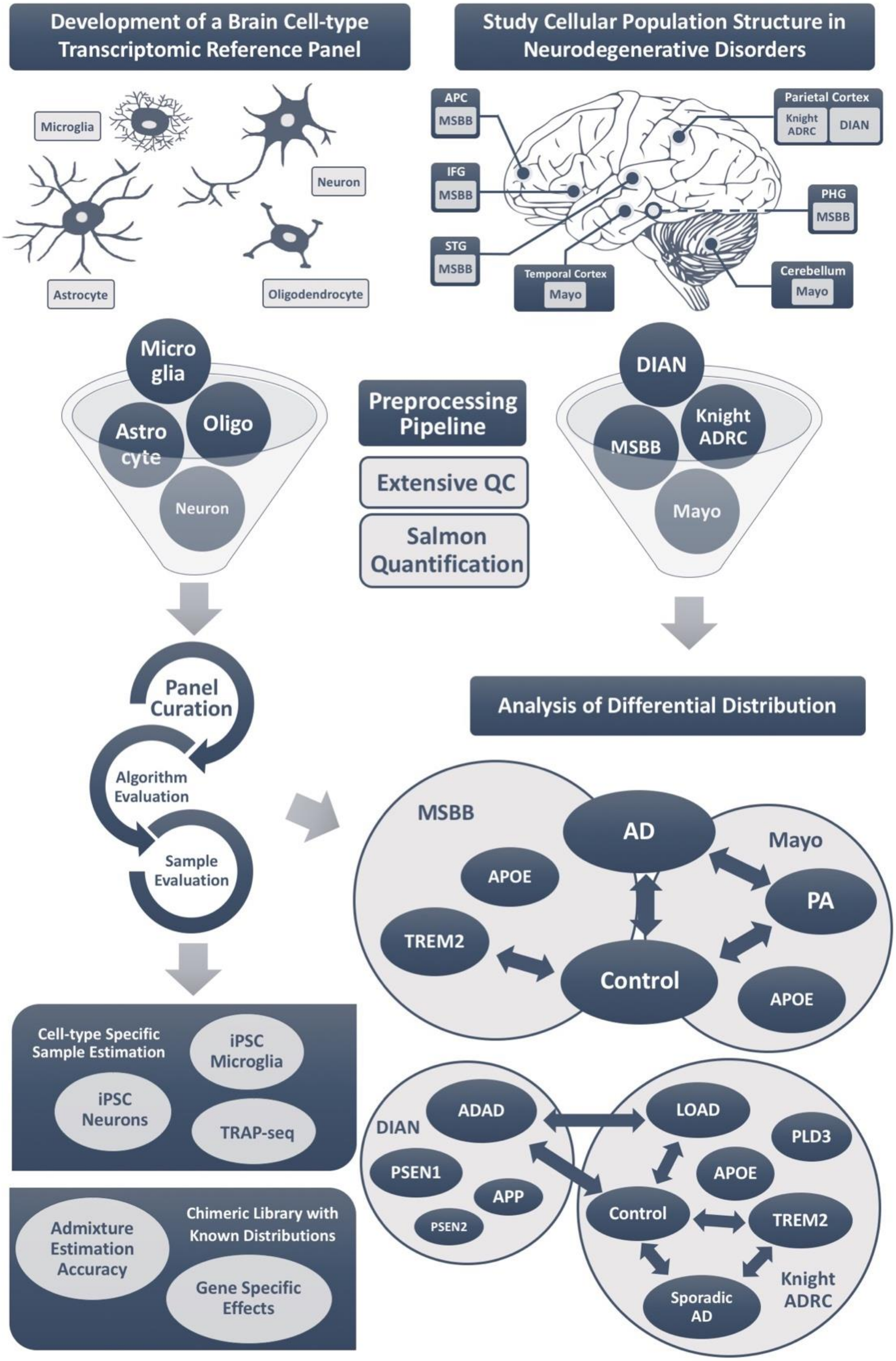
To analyze the DIAN and Knight-ADRC studies I applied linear-mixed models (function lmer and Anova, R packages lme4 ver.1.1 and car ver.2.1, respectively), clustering at family level to ascertain the effect of the neuropathological status in the cell proportion, and corrected for RIN and PMI. For late-onset specific analyses I also corrected for age at death. Cellular composition shown as proportions were plotted using R package ggplot2 (ver 2.2.1)

2.4 Results

2.4.1 Study design

To infer cellular composition from RNA-Seq, I firstly assembled a reference panel to model the transcriptomic signature of neurons, astrocytes, oligodendrocytes and microglia. The panel was created by analyzing expression data from purified cell lines. I evaluated alternative digital deconvolution methods and selected the best performing for my primary analyses. I tested the digital deconvolution accuracy on induced pluripotent stem cell (iPSC) derived neurons/microglia cells and neuronal Translating Ribosome Affinity Purification followed by RNA-Seq (TRAP-seq; **Figure 2.20**). Finally, I verified its accuracy by creating artificial admixture with pre-defined cellular proportions.

Figure 2.20 Study design. Development of the brain cell-type transcriptomic reference panel (**left column**): the expression signatures of key cell types of the brain were curated by compiling publicly available RNA-Seq data from neurons, astrocytes, oligodendrocytes and microglia. The panel was curated iteratively to retain only those samples that showed the most faithful expression signature, while evaluating alternative digital deconvolution methods. The accuracy of digital deconvolution to estimate brain cellular proportion was validated using additional cell-type specific samples, and also by generating chimeric libraries. To study cellular population structure in AD (**right column**), I accessed publicly available datasets from the Advanced Medicines Partnership-AD knowledge portal (AMP-AD), including Mayo Clinic and Mount Sinai Brain Bank datasets. In addition, we generated RNA-Seq from participants of the Knight-ADRC and The Dominantly Inherited Alzheimer (DIAN) studies. These three studies generated RNA-Seq data from pathological aging brains, Alzheimer's disease cases, and neuropath-free controls for a total of six cerebral cortex regions and cerebellum. I quantified the gene expression for all of the samples included in these studies using the same RNA-Seq processing pipeline. Using digital deconvolution methods, I estimated the brain cellular proportions of the samples and compared the proportion between AD cases and controls. I study the cell structure of brains carriers of Mendelian pathological mutations and variants that confer high-risk to AD. Anterior prefrontal cortex – APC; superior temporal gyrus – STG; parahippocampal gyrus – PHG; inferior frontal gyrus – IFG; Mount Sinai Brain Bank – MSBB; Alzheimer's disease – AD; pathological aging – PA.


Development of a Brain Cell-type Transcriptomic Reference Panel
Study Cellular Population Structure in Neurodegenerative Disorders
Microglia
Neuron
Astrocyte
Oligodendrocyte
APC
MSBB
IFG
MSBB
STG
MSBB
Temporal Cortex
Mayo
Parietal Cortex
Knight ADRC
DIAN
PHG
MSBB
Cerebellum
Mayo
Micro glia
Astro cyte
Oligo
Neuron
Preprocessing Pipeline
Extensive QC
Salmon Quantification
DIAN
MSBB
Knight ADRC
Mayo
Panel Curation
Algorithm Evaluation
Sample Evaluation
Analysis of Differential Distribution
MSBB
AD
Mayo
APOE
PA
TREM2
Control
APOE
Cell-type Specific Sample Estimation
iPSC Microglia
iPSC Neurons
TRAP-seq
Admixture Estimation Accuracy
Chimeric Library with Known Distributions
Gene Specific Effects
DIAN
ADAD
LOAD
PLD3
PSEN1
APOE
APP
PSEN2
Control
TREM2
Sporadic AD
Knight ADRC

Once the deconvolution approach was optimized, I calculated the cell proportion in AD cases and controls from the different brain regions of Mayo and MSBB datasets. The RNA-Seq data for the Mayo Clinic study (N = 191)[9] and Mount Sinai (MSSM) Brain Bank (MSBB; N = 300)[3] are deposited in the Advanced Medicines Partnership-AD (AMP-AD) knowledge portal (Synapse ID: syn5550404 and syn3157743; **Table 2.1**). The Mayo study includes RNA-Seq from the temporal cortex and cerebellum for AD affected and non-demented controls, in addition to pathological aging participants (**Figure 2.20**). The MSBB also profiled four additional cerebral cortex areas: anterior prefrontal cortex - APC, superior temporal gyrus - STG, parahippocampal gyrus – PHG, and inferior frontal gyrus – IFG; **Table 2.1; Figure 2.20**). I restricted the case-control analysis to subjects with definite AD and autopsy confirmed controls. In addition, we generated RNA-Seq from parietal lobe for participants of the Knight-ADRC (84 late-onset cases, carriers of genetic risk factors and 16 controls; **Table 2.2**) and The Dominantly Inherited Alzheimer Network (DIAN; 19 carriers of mutations in *APP*, *PSEN1*, *PSEN2*) (**Table 2.1; Figure 2.20).** I employed the same pipeline to process all of the samples in order to avoid any bias. Furthermore, RNA-Seq from the Knight-ADRC and DIAN studies allowed us to compare the cell composition from ADAD vs LOAD brains, and similarly to test for differences in brain of controls, sporadic AD who do not carry any known high-risk variant, carriers of high-risk variants in *TREM2* (N = 20), *PLD3* (N = 33), and *APOE* ε4 allele.

2.4.2 Development of a reference panel to estimate brain cellular population structure

Due to limited availability of brain cell-type specific transcriptomic data, I compiled reference samples from different sources, including single-population RNA-Seq from mice and human (immunopan-purified oligodendrocytes, neurons, astrocytes and microglia and iPSC-derived neurons and astrocytes).

I selected 17 well accepted genes that tag brain cell types based on literature reviews[41, 131, 291] . A visual inspection of the expression of these marker genes in the samples I compiled suggested a divergent transcriptomic profile among the cell types (**Figure 2.15a**). The PCA showed that their expression was sufficient to cluster samples of neurons, astrocytes, oligodendrocytes and microglia with their respective cell types, regardless of the species of the reference samples (**Figure 2.14b; Table 2.6**). I observed that first principal component (PC) captured the expression profile of astrocytes; as shown by the significant association of the expression of astrocyte marker genes with ($p < 8.05\times10^{-03}$). The second PC captured the expression of genes whose expression is characteristic to oligodendrocytes ($p < 2.52\times10^{-02}$). The third PC was negatively associated with neuronal genes ($p < 1.11\times10^{-05}$) and positively with microglia ($p < 1.42\times10^{-02}$). Overall, the principal component analysis (**Figure 2.14b**) indicated that these genes can effectively cluster samples by their cell-type.

Given the technical and biological heterogeneity of the samples I compiled for reference panel, I carried out an optimization phase to identify those samples that showed the most faithful expression profile to represent their respective cell types (See **Methods**). From the leave-one-out cross-validation results, I noticed that not all of the cell-type specific samples were predicted as expected (defined with a correct prediction proportion higher than 80%). Samples failed this criteria were due to various reasons, for example, the expression profile of immunopan-purified astrocytes collected from mice[291], human fetal[291] or human sclerotic hippocampal[291] brains were reported with altered expression[38, 291] that differed to an extent that could not be accurately ascertained by deconvolution methods. Similarly, neuronal proportion inferred from iPSC-derived neurons from schizophrenic donors[38] and iPSC-derived neurons collected at early stages of differentiation (< 100 days; Synapse ID: syn3607401) were also lower than 80%. These samples did not cluster with their expected cell types in the marker gene PCA either, and coincidently the leave-one-out cross-validation indicated that these samples had an expression signature that differed from the other samples of the same cell type.

To evaluate the performance of reference panel performance and test out different deconvolution algorithms, I employed and compared six digital deconvolution methods implemented in the CellMix package (ver 1.6) to infer cellular composition from reference samples RNA-Seq data, including qprog[108], cs-qprog, DSA[293], ssFrobenius[103], meanProfile[157], deconf[221]. The deconvolution performance of reference panel was evaluated by following a leave-one-out cross-validation procedure to compare the predicted cellular composition with its expected cellular identity of each cell-type specific sample. The accuracy of this comparison was quantified using the root-mean-squared error (RMSE) calculation. A semi-supervised method adapted from non-negative matrix factorization[103] (ssNMF – named ssFrobenius in CellMix) generated the most accurate predictions; and I verified that similar results were obtained by the method population-specific expression analysis[157] (PSEA – named meanProfile in CellMix).

I ascertained the effect that sequencing depth has in the accuracy of deconvolution. I generated low-coverage versions (Picard DownsampleSam ver 2.8.2) of the samples that included a reduced number of randomly sampled reads (400,000 reads per sample), quantified the gene expression, inferred their cellular population proportions, and compared the distribution estimates with their full-depth libraries sequencing (more than 30 million reads per sample). I observed that the deconvolution was robust to the sequencing coverage, as shown by a correlation r^2=0.98 ($p < 2.2\times10^{-16}$; **Figure 2.21**). My final reference panel (**Table 2.6; Table 2.7**) had a very high confidence to predict cell types with a mean predicted accuracy = 95.2%; s.d. = 4.3, and a root-mean-square error (RMSE) = 0.06 (**Table 2.9**).

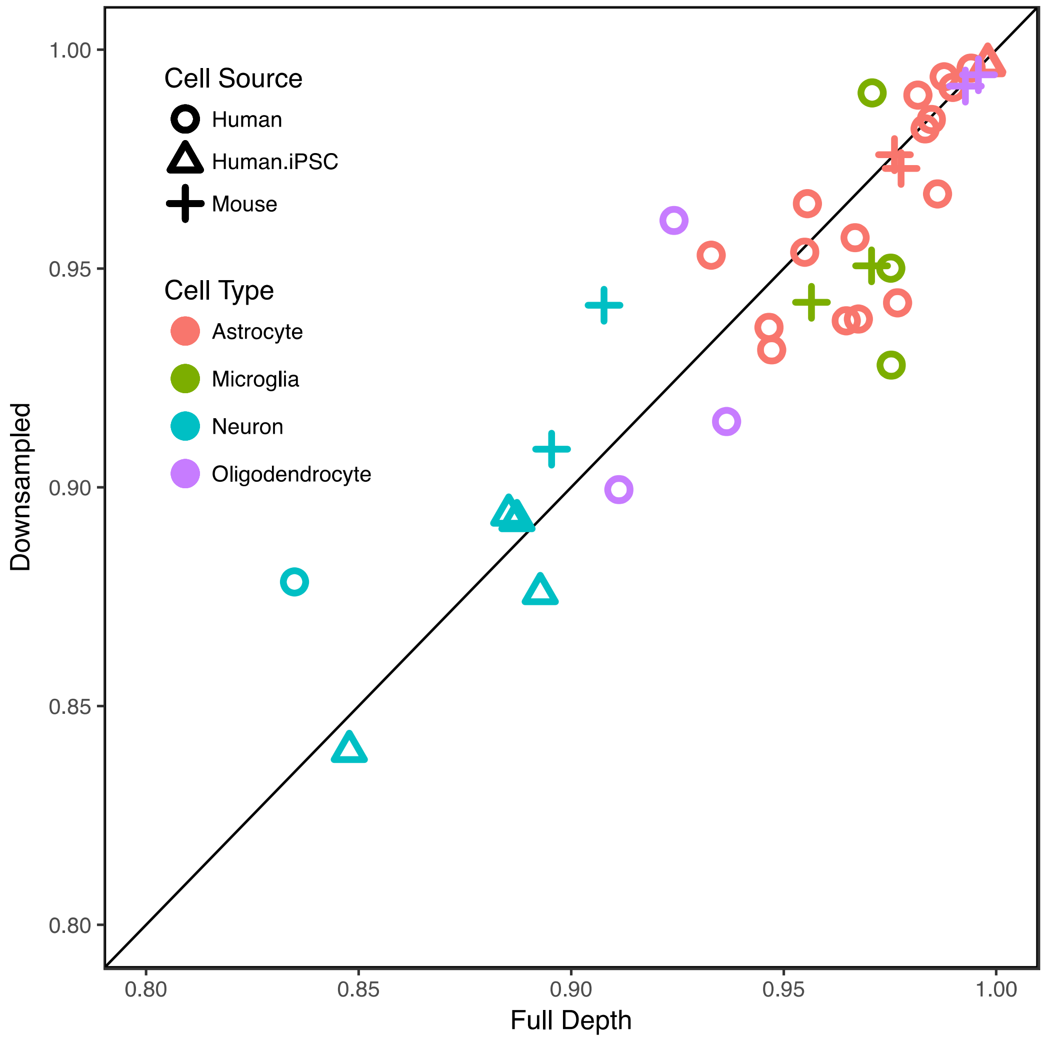


Figure 2.21 Comparison of cell proportions estimated from full-depth and down-sampled RNA-Seq data. Each sample of the reference panel sample was down-sampled (400,000 reads) and cellular population structure inferred following leave-one-out procedure. Cell-type proportions of the samples inferred using the full-depth RNA-Seq data are presented along the X-axis, and along the Y-axis the counterparts inferred using shallow RNA-Seq.

Table 2.9 Evaluation of deconvolution accuracy. Overall and cell-type specific root-mean-squared error (RMSE) for reference panel, calculated using the leave-one-out approach for three deconvolution algorithms implemented in CellMix package.

Algorithm	Overall	Neuron	Astrocyte	Oligodendrocyte	Microglia
ssNMF[a]	0.064	0.054	0.055	0.028	0.017
PSEA[b]	0.089	0.08	0.052	0.058	0.025
DSA[c]	0.465	0.32	0.328	0.291	0.295

[a] ssNMF: semi-supervised learning non-negative matrix factorization.

[b] PSEA: population-specific expression analysis (also named meanProfile in CellMix implementation).

[c] DSA: Digital sorting algorithm.

2.4.3 Optimization, validation and accuracy estimation of the reference panel and digital deconvolution method

Once I identified the optimal approach to perform digital deconvolution from brain RNA-Seq, I benchmarked it by using three sets of independent pure cell populations and simulated chimeric libraries.

I firstly validated the accuracy to predict neuronal composition by generating RNA-Seq for eight iPSC-derived cortical neurons (see **Methods**). I observed an accurate prediction in these independent cell lines (mean neuronal proportion = 94.8% and s.d. = 1.1%; **Figure 2.22a**). I also ascertained the cellular composition of mRNA extracted from the barrel cortex neurons isolated by Translating Ribosome Affinity Purification (TRAP) in 24 mice. TRAP is a method that captures cell-type specific mRNA translation by purifying tagged ribosomal subunit and capturing the mRNA it bound to[122]. I observed an average of neuronal proportion = 96.7% and s.d. = 1.2% (**Figure 2.22b**). Similarly, I assessed the RNA-Seq data generated for iPSC-derived microglia (N = 10) deposited in the AMP-AD portal (Synapse ID: syn7203233) and inferred their cellular population structure and observed a mean microglia proportion = 86.6% and s.d. = 7.1% (**Figure 2.22c**).

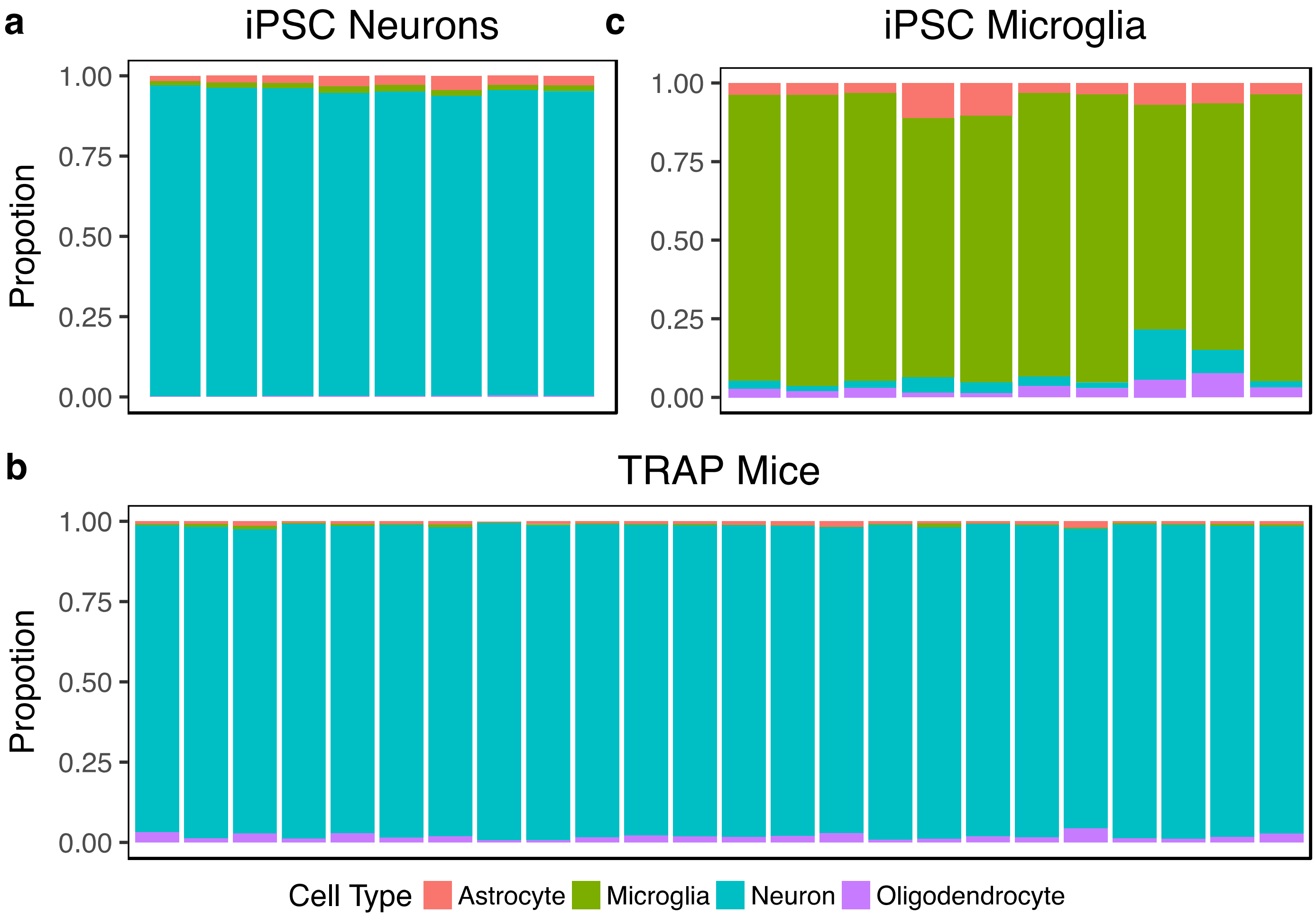


Figure 2.22 Cellular population structure of cell-type specific samples. Cell-type proportions shown as stacked percentage (red: astrocytes; green: microglia; blue: neuron; purple: oligodendrocyte). a) iPSC derived cortical neurons (N = 8). b) mouse barrel cortex neurons isolated by Translating Ribosome Affinity Purification (TRAP) procedure (N = 24). c) iPSC derived microglia (N = 10).

To evaluate the accuracy of digital deconvolution for measuring cell-type proportion from cell-type admixtures, I simulated RNA-Seq libraries by pooling reads from individual cell types into well-defined proportions. I combined randomly sampled reads from neurons, astrocytes, oligodendrocytes and microglia to create chimeric libraries that mimic bulk RNA-Seq from brain, but with a range of pre-defined cell-type distributions (**Figure 2.16**). I then quantified the gene expression for the chimeric libraries and inferred the cell-type distribution (employing for the reference panel samples that did not contribute reads to the chimeric libraries). This process was repeated 23,040 times, choosing distinct human samples to represent each cell type and varying the proportions in 32 alternative distributions (See methods and **Table 2.8**). The overall error (RMSE) compared to known proportions = 0.08.

Finally, I evaluated whether any gene included in the reference panel was dominating the inference of cell proportions. I re-calculated the cell-type distributions of the chimeric libraries, but dropping each of the genes from the reference panel one at a time. I observed a negligible difference between the cellular population structure inferred using the full reference and the gene-dropped panels (average RMSE = 0.022, s.d. < 0.01). In this way, I verified that the proportions inferred using the reference panel are not driven by the expression of a single gene. This reassured us the inference should be robust to any bias introduced by the potential association of a single gene included in the reference panel with a particular trait.

2.4.4 Deconvolution of bulk RNA-Seq of non-demented and AD brains shows a characteristic signature for neurodegeneration

Pathologically, AD is associated with neuronal death and gliosis specifically in the cerebral cortex. I evaluated whether I could exploit deconvolution methods using my reference panel to detect altered cellular population structure from the bulk RNA-Seq, and whether this corresponded to known pathological alterations.

I initially analyzed the RNA-Seq from the Mayo Clinic Brain Bank that includes bulk RNA-Seq from the temporal cortex (TC) and cerebellum (CB) for 191 participants[9] (**Table 2.1**). In the TC, I observed a significant higher astrocyte relative proportion ($\beta = 0.23$; $p = 5.01\times10^{-09}$; **Table 2.10; Figure 2.25; Table 2.11**) in AD brains compared to controls brains. I also found a significant lower relative proportion of neurons ($\beta = -0.17$; $p = 1.58\times10^{-07}$; **Table 2.10; Figure 2.25; Table 2.11)** and oligodendrocytes ($\beta = -0.07$; $p = 1.8\times10^{-02}$; **Table 2.10; Figure 2.23; Table 2.11**). As expected, given the absence of pathology, I did not observe a significant difference in the cell-type composition in the CB (**Table 2.10)**.

Table 2.10 Comparison of the cellular population structure (AD vs. neuropath-free controls) from the brains in the Mayo Clinic and Mount Sinai Brain Bank.

	Brain Regions	Sample Size	Neuron		Astrocyte		Oligodendrocyte		Microglia	
Mayo	**AD vs Control**	N	Effect	P-value	Effect	P-value	Effect	P-value	Effect	P-value
	Cerebellum	119	-0.03	2.74×10^{-01}	0.05	8.65×10^{-02}	-0.02	1.07×10^{-01}	-3.19×10^{-04}	9.19×10^{-01}
	Temporal Cortex	119	-0.17	1.58×10^{-07}	0.23	5.01×10^{-09}	-0.07	1.8×10^{-02}	-2.03×10^{-03}	5.48×10^{-01}
Mount Sinai Brain Bank	**AD vs Control**									
	Anterior Prefrontal Cortex	184	-0.04	8.14×10^{-04}	0.06	8.11×10^{-05}	-0.01	3.36×10^{-02}	-3.18×10^{-03}	1.12×10^{-02}
	Superior Temporal Gyrus	167	-0.08	3.49×10^{-07}	0.1	1.45×10^{-07}	-0.01	5.8×10^{-02}	-3.17×10^{-03}	5.78×10^{-02}
	Parahippocampal Gyrus	160	-0.11	1.35×10^{-08}	0.13	5.48×10^{-10}	-0.02	1.79×10^{-03}	-3.18×10^{-03}	1.35×10^{-01}
	Inferior Frontal Gyrus	159	-0.04	3.12×10^{-03}	0.06	3.58×10^{-04}	-0.01	4.39×10^{-02}	-3.98×10^{-03}	1.64×10^{-02}
	Clinical Dementia Rating									
	Anterior Prefrontal Cortex	184	-0.02	9.38×10^{-04}	0.02	2.07×10^{-04}	-3.43×10^{-03}	1.25×10^{-01}	-1.46×10^{-03}	4.95×10^{-03}
	Superior Temporal Gyrus	167	-0.03	1.87×10^{-06}	0.04	3.33×10^{-07}	-0.01	2.1×10^{-02}	-1.02×10^{-03}	1.49×10^{-01}
	Parahippocampal Gyrus	160	-0.04	8.56×10^{-06}	0.04	2.85×10^{-06}	-0.01	8.7×10^{-02}	-1.94×10^{-03}	2.53×10^{-02}
	Inferior Frontal Gyrus	159	-0.02	8.29×10^{-05}	0.03	1.4×10^{-05}	-4.64×10^{-03}	6.7×10^{-02}	-1.46×10^{-03}	3.11×10^{-02}
	Braak Staging									
	Anterior Prefrontal Cortex	173	-0.01	1.21×10^{-02}	0.01	1.27×10^{-03}	-3.09×10^{-03}	2.77×10^{-02}	-7.04×10^{-04}	3.12×10^{-02}
	Superior Temporal Gyrus	158	-0.02	2.22×10^{-07}	0.02	2.77×10^{-07}	-2.91×10^{-03}	1.17×10^{-01}	-5.47×10^{-04}	1.97×10^{-01}
	Parahippocampal Gyrus	147	-0.02	1.83×10^{-06}	0.03	9.6×10^{-08}	-0.01	1.49×10^{-03}	-3.71×10^{-04}	4.97×10^{-01}
	Inferior Frontal Gyrus	152	-0.01	1.01×10^{-02}	0.01	8.56×10^{-04}	-3.55×10^{-03}	2.37×10^{-02}	-1.01×10^{-03}	1.74×10^{-02}
	Mean Amyloid Plaques									
	Anterior Prefrontal Cortex	184	-1.88×10^{-03}	3.6×10^{-03}	2.82×10^{-03}	1.03×10^{-04}	-7.99×10^{-04}	2.13×10^{-03}	-1.46×10^{-04}	1.72×10^{-02}
	Superior Temporal Gyrus	167	-4.2×10^{-03}	7.73×10^{-08}	0.01	4.63×10^{-08}	-6.08×10^{-04}	9.01×10^{-02}	-2.04×10^{-04}	1.5×10^{-02}
	Parahippocampal Gyrus	160	-4.96×10^{-03}	5.05×10^{-09}	0.01	1.26×10^{-10}	-9.99×10^{-04}	1.85×10^{-03}	-2.1×10^{-04}	2.58×10^{-02}
	Inferior Frontal Gyrus	159	-2.58×10^{-03}	3.82×10^{-04}	3.53×10^{-03}	1.96×10^{-05}	-7.41×10^{-04}	1.51×10^{-02}	-2.04×10^{-04}	1.26×10^{-02}

The cell-type proportions from AD cases and control were inferred from bulk RNA-seq using the ssNMF method. Effects of AD and associations with additional clinical and pathological phenotypes in cell-type distributions were estimated using linear regression model.

The distribution of microglia was similar in the TC and CB from AD and control brains (**Table 2.10; Figure 2.23)**. The proportion of microglia was lower than any other cell types. The Mayo dataset also includes brains from individuals with pathological aging (PA; **Table 2.1**); which is neuropathologically defined by amyloid β (Aβ) senile plaque deposits but little or no neurofibrillary tau pathology[9, 192]. I observed a significant lower relative proportion of microglia in PA brains compared to AD in both TC and CB (**Table 2.12; Figure 2.24**)[169]. Therefore, I speculated that the lack of changes in the AD microglial population was neither due to low statistical power nor the inability of my method to estimate the microglial proportions, but reflected unaltered neuropathological observations in AD brains.

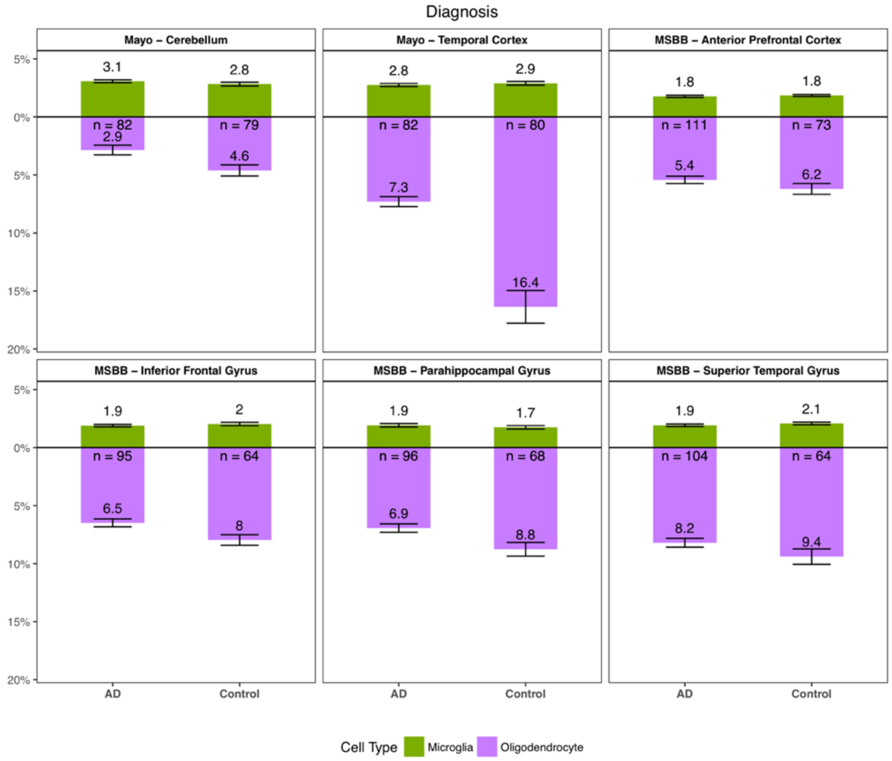


Figure 2.23 Microglia and oligodendrocyte proportions inferred from RNA-Seq of Mayo Clinic and Mount Sinai Brain Bank (MSBB) studies. Mean microglial (green) and oligodendrocyte (purple) proportion for AD cases and neuropath-free controls (bars indicate the standard deviation). The numbers of subjects are indicated below x-axis.

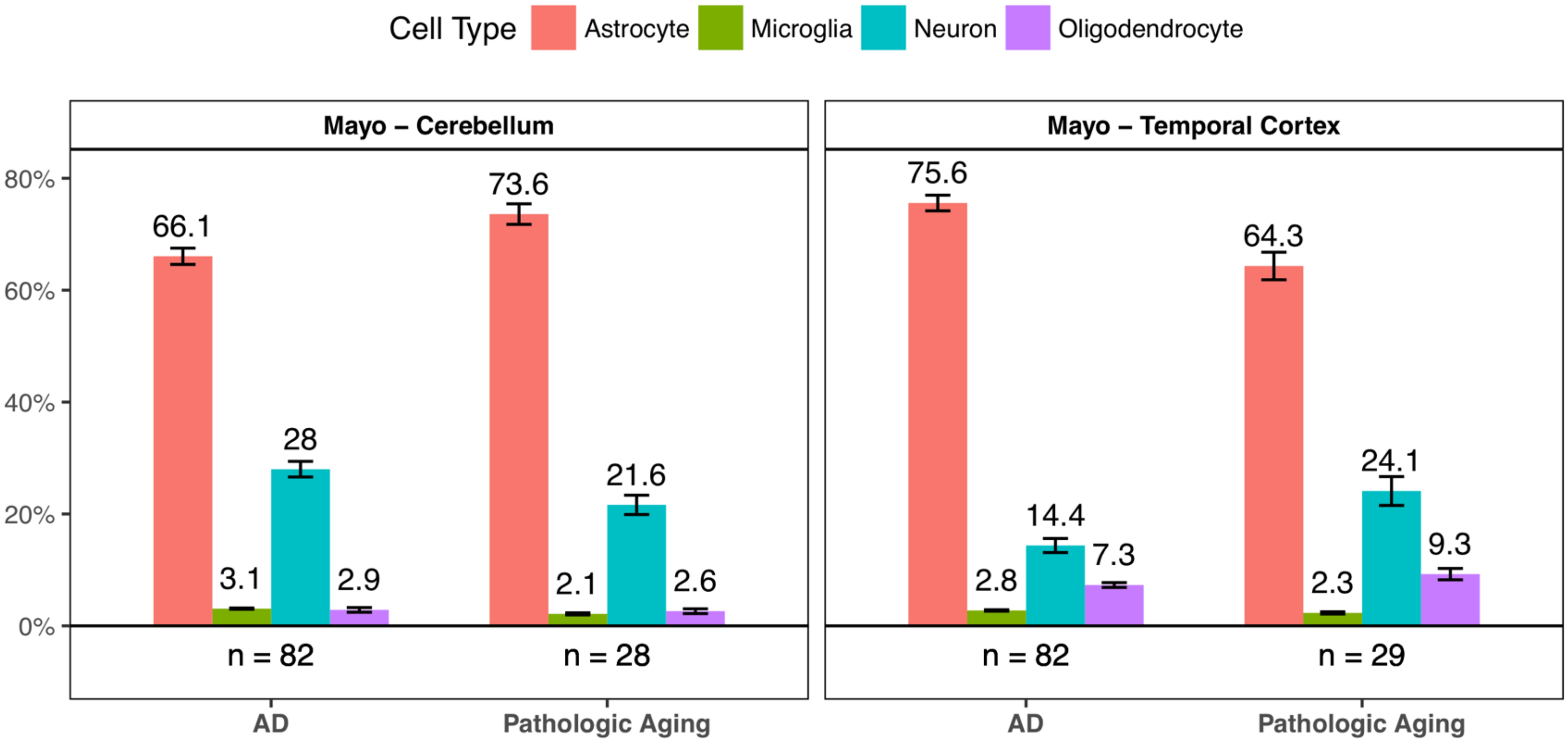


Figure 2.24 Cellular population structure for Alzheimer’s disease (AD) and Pathological Aging (PA) subjects included in the Mayo Clinic study. Columns height represent the mean proportions. The numbers of subjects for each group is reported below x-axis.

I also analyzed data from the MSBB, which contains bulk RNA-Seq for four additional cerebral cortex areas (APC, STG, PHG, IFG). Replicating my findings from the Mayo dataset I observed a significant lower relative proportion in neurons and increase in astrocytes in all four areas (**Table 2.10; Figure 2.25; and Table 2.11**). The strongest effect size was detected in the parahippocampal gyrus and superior temporal gyrus ($p < 3.49\times10^{-07}$) (**Table 2.10; Table 2.13**). Neuropathological studies have described that the parahippocampal gyrus in one of the first brain areas in which AD pathology occurs[33, 78, 267]. I also observed a significant and strong correlation between neuronal and astrocyte relative proportions and last ascertained clinical status (Clinical Dementia Rating - CDR), and number of amyloid plaques and Braak staging (**Table 2.10; Figure 2.25; Figure 2.26**).

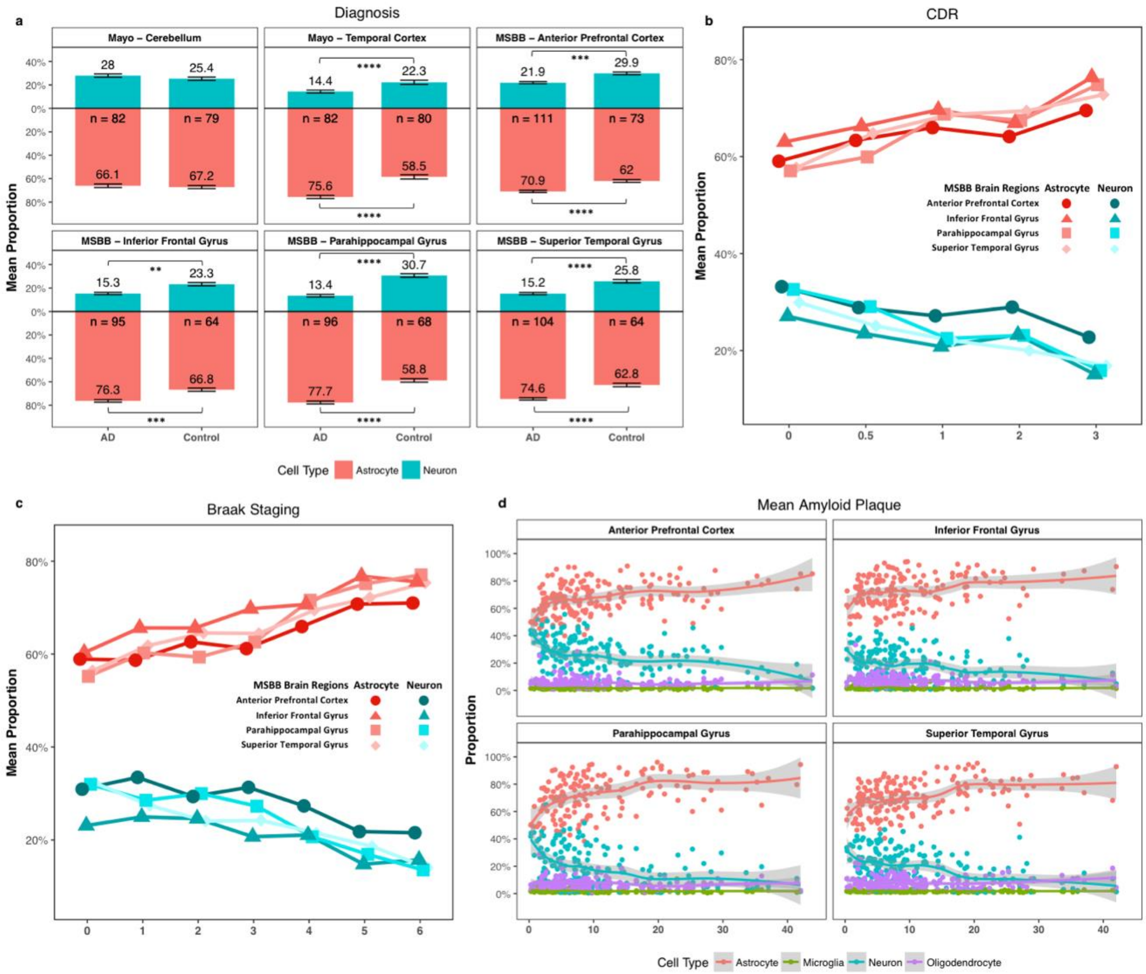


Figure 2.25 Cell-type distributions of the samples included in the Mayo Clinic and Mount Sinai Brain Bank. Mean neuronal (blue) and astrocytic proportion (red) for **a)** Alzheimer's disease affected brains (AD) and controls (bars indicate standard deviations). The numbers of subjects for each group are shown below the x-axis. Distribution for additional clinical and pathological phenotypes reported for the Mount Sinai Brain Bank (MSBB): **b)** clinical dementia rating scores (CDR) and **c)** Braak and Braak staging. **d)** Brain cell-type proportions (x-axis) plotted against the mean number of amyloid plaque (values greater than 0; y-axis). Standard errors were depicted in shaded area with LOESS smooth curve fitted to cell-type proportions derived from deconvolution. (** P< 0.01; *** P< 1.0×10^{-3}; and **** P< 1.0×10^{-4}).

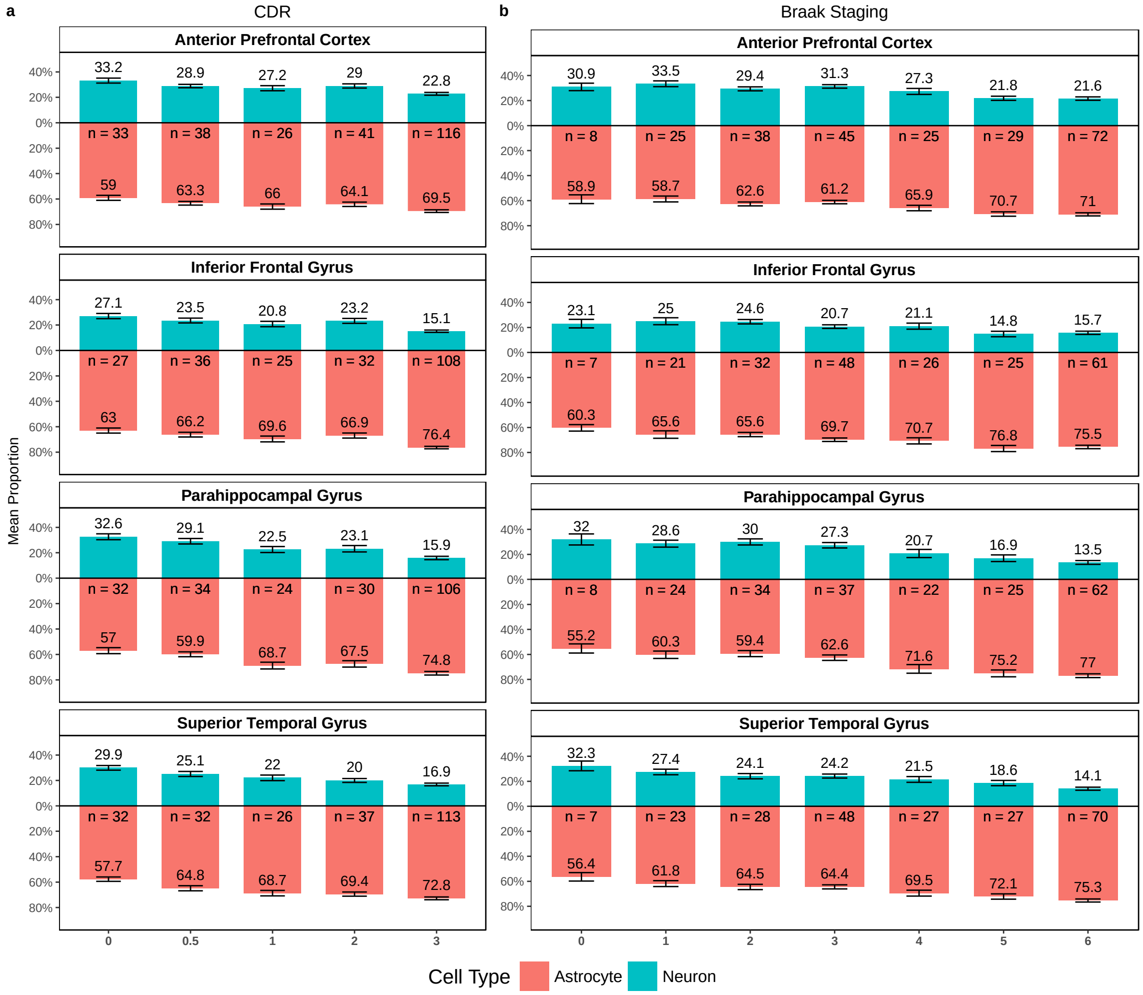


Figure 2.26 Neurons and astrocytes distributions for the brains included in the Mount Sinai Brain Bank stratified by CDR and Braak staging. Neuron (blue) and astrocyte (red) proportions for the plotted against a) CDR. b) Braak Staging.

Table 2.11 Comparison of the cellular proportions estimated using the method PSEA in AD and control brains from the Mayo and Mount Sinai Brain Bank.

Brain Regions	**Sample Size**	**Neuron**		**Astrocyte**		**Oligodendrocyte**		**Microglia**	
Mayo	N	Effect	P-value	Effect	P-value	Effect	P-value	Effect	P-value
Cerebellum	119	-0.04	1.05×10^{-01}	0.06	2.09×10^{-02}	-0.02	1.53×10^{-01}	8.79×10^{-04}	6.53×10^{-01}
Temporal Cortex	119	-0.17	7.1×10^{-08}	0.23	8.43×10^{-09}	-0.07	3.1×10^{-02}	3.91×10^{-04}	8.67×10^{-01}
Mount Sinai Brain Bank									
Anterior Prefrontal Cortex	184	-0.04	1.75×10^{-03}	0.06	1.02×10^{-04}	-0.01	2.68×10^{-02}	-1.81×10^{-03}	2.76×10^{-02}
Superior Temporal Gyrus	167	-0.07	2.57×10^{-06}	0.09	1.06×10^{-06}	-0.01	8.55×10^{-02}	-1.77×10^{-03}	1.32×10^{-01}
Parahippocampal Gyrus	160	-0.1	2.37×10^{-08}	0.13	2.66×10^{-10}	-0.02	2.74×10^{-03}	-1.72×10^{-03}	2.44×10^{-01}
Inferior Frontal Gyrus	159	-0.04	4.07×10^{-03}	0.06	5.96×10^{-04}	-0.01	1.15×10^{-01}	-2.81×10^{-03}	1.84×10^{-02}

Table 2.12 Cell-type proportions comparison of subjects diagnosed with Pathological Aging. The cell-type proportions inferred from RNA-seq data using the ssNMF method. Distribution in Pathological Aging (PA) brains, AD cases and neuropath-free controls are compared using linear regression model.

	Sample Size	**Neuron**		**Astrocyte**		**Oligodendrocyte**		**Microglia**	
PA vs AD	N	Effect	P-value	Effect	P-value	Effect	P-value	Effect	P-value
Cerebellum	75	0.05	6.66×10^{-02}	-0.06	3.83×10^{-02}	0.02	1.11×10^{-01}	-0.01	2.29×10^{-04}
Temporal Cor	76	0.19	2.51×10^{-06}	-0.21	1.02×10^{-06}	0.03	5.57×10^{-02}	-0.01	5.05×10^{-04}
PA vs Control									
Cerebellum	94	0.02	5.38×10^{-01}	-0.04	1.77×10^{-01}	0.02	1.7×10^{-01}	4.65×10^{-03}	1.73×10^{-01}
Temporal Cor	91	-0.05	1.91×10^{-01}	-0.03	4.58×10^{-01}	0.07	2.75×10^{-02}	5.39×10^{-03}	1.53×10^{-01}

Table 2.13 Effect of AD in the neuronal and astrocytic proportions in distinct cerebral cortex areas. Comparison of the effect that AD has in the cell-type distribution in the four cerebral cortex areas ascertained in the Mount Sinai Study (ANCOVA). We report p-values for the pairwise comparison (upper triangle for neuron and lower triangle for astrocyte).

		Upper Half - Neuron		
MSBB regions	**APC**	**STG**	**PHG**	**IFG**
APC[a]	-	7.79×10^{-04}	9.02×10^{-03}	$<2.22\times10^{-16}$
STG[b]	4.89×10^{-01}	-	9.35×10^{-01}	2.12×10^{-11}
PHG[c]	9.90×10^{-01}	7.06×10^{-01}	-	8.53×10^{-13}
IFG[d]	1.28×10^{-10}	2.23×10^{-06}	8.86×10^{-08}	-
		Lower Half - Astrocyte		

[a] **APC** - Anterior Prefrontal Cortex.

[b] **STG** - Superior Temporal Gyrus.

[c] **PHG** - Parahippocampal Gyrus.

[d] **IFG** - Inferior Frontal Gyrus.

2.4.5 The cellular population structure differs between ADAD vs LOAD

While the loss of neurons is a common feature of AD, it is not clear whether the mechanism holds true across different forms of AD or AD cases carrying different genetic risk variants. Therefore, I investigated whether AD with distinct etiologies showed different cellular compositions. We generated RNA-Seq data from the parietal lobe of participants enrolled in Knight-ADRC (84 LOAD, 3 ADAD, and 16 neuropath-free controls) and DIAN (19 ADAD) studies (**Table 2.1; Table 2.2**). I selected the LOAD and ADAD participants to match for CDR at death, brain weight and sex distributions (See **Table 2.2**).

Using digital deconvolution, I determined the cellular composition for these brains. I observed a significant lower relative proportion of neurons ($\beta = -0.02$, p = 2.66×10^{-02}) and significant higher relative proportion of astrocyte in AD ($\beta = 0.03$, p =

5.48×10^{-03}) for the combined LOAD and ADAD brains compared to controls (**Table 2.14; Figure 2.27; Table 2.15**), consistent with my findings in the Mayo and MSBB datasets. Similarly, the joint analysis of the brains from Knight-ADRC and DIAN showed a significant association between the neuronal and astrocyte relative proportions and neuropathological measures (Braak staging: $\beta = -0.03$, $p = 8.51\times10^{-06}$ for neurons and $\beta = 0.03$, $p = 3.83\times10^{-06}$ for astrocytes; **Table 2.14**; **Figure 2.27b**) as well as for clinical measures (CDR: $\beta = -0.02$, $p = 2.66\times10^{-02}$ for neurons and $\beta = 0.03$ and $p = 5.48\times10^{-03}$ for astrocytes; **Table 2.14**; **Figure 2.27c**). I did not observe a significant difference in the compositions of microglia or oligodendrocytes (**Table 2.14**; **Fig S8**).

Table 2.14 Cellular population structure altered in the parietal lobe from AD brains in the DIAN study and Knight-ADRC brain bank.

Disease Status	**Sample Size**	**Neuron**		**Astrocyte**		**Oligodendrocyte**		**Microglia**	
AD Status	N	Effect	P-value	Effect	P-value	Effect	P-value	Effect	P-value
AD[a] vs Control	122	-0.11	5.52×10^{-04}	0.14	2.48×10^{-05}	-0.03	6.5×10^{-02}	-2.64×10^{-03}	2.49×10^{-01}
ADAD vs Control	38	-0.19	3.94×10^{-07}	0.24	1.57×10^{-10}	-0.04	8.5×10^{-03}	-0.01	7.77×10^{-05}
LOAD vs Control	100	-0.09	5.67×10^{-03}	0.12	3.34×10^{-04}	-0.02	1.06×10^{-01}	-1.70×10^{-03}	4.57×10^{-01}
ADAD vs LOAD									
Braak matched	42	-0.08	1.03×10^{-02}	0.11	9.26×10^{-04}	-0.03	7.1×10^{-02}	-1.46×10^{-03}	7.01×10^{-01}
Braak corrected	91	-0.09	4.71×10^{-03}	0.11	5.24×10^{-04}	-0.02	1.77×10^{-01}	-2.41×10^{-03}	4.25×10^{-01}
CDR corrected	94	-0.12	2.11×10^{-03}	0.13	6.29×10^{-04}	-0.02	3.8×10^{-01}	-3.11×10^{-03}	2.41×10^{-01}
Clinical Dementia Rating									
AD[a] and Controls	110	-0.02	2.66×10^{-02}	0.03	5.48×10^{-03}	-0.01	2×10^{-01}	-4.63×10^{-04}	4.77×10^{-01}
ADAD and Controls	26	-0.08	4.12×10^{-04}	0.11	1.78×10^{-07}	0.01	4.03×10^{-03}	-1.55×10^{-03}	1.75×10^{-08}
LOAD and Controls	100	-0.02	3.22×10^{-02}	0.03	7.01×10^{-03}	-0.01	1.81×10^{-01}	-4.64×10^{-04}	5.11×10^{-01}
Braak Staging									
AD[a] and Controls	106	-0.03	8.51×10^{-06}	0.03	3.83×10^{-06}	-4.24×10^{-03}	2.04×10^{-01}	-2.52×10^{-04}	6.81×10^{-01}
ADAD and Controls	33	-0.05	2.37×10^{-05}	0.06	2.45×10^{-05}	-0.01	2.29×10^{-01}	-7.2×10^{-04}	4.89×10^{-01}
LOAD and Controls	88	-0.03	7.41×10^{-04}	0.03	4.63×10^{-04}	-3.72×10^{-03}	3.29×10^{-01}	-1.66×10^{-04}	7.86×10^{-01}

[a] AD includes both autosomal dominant AD (ADAD) and late-onset AD (LOAD).

The cellular population structure was inferred using the ssNMF method. Effects and p-values for the association with disease status, clinical dementia rating and Braak staging using generalized mixed models. We identified similar trends with approximately the same significance levels.

Table 2.15 Comparison of the cellular proportions estimated using the method PSEA in AD and control brains from the DIAN and Knight-ADRC

Disease Status	**Sample Size**	**Neuron**		**Astrocyte**		**Oligodendrocyte**		**Microglia**	
	N	Effect	P-value	Effect	P-value	Effect	P-value	Effect	P-value
AD[a] vs Control	122	-0.1	1.41×10^{-03}	0.12	7×10^{-05}	-0.02	1.97×10^{-01}	-1.84×10^{-03}	2.4×10^{-01}
ADAD[b] vs Control	38	-0.18	2.35×10^{-06}	0.22	3.03×10^{-10}	-0.05	3.47×10^{-02}	-3.77×10^{-03}	1.56×10^{-04}
LOAD[c] vs Control	100	-0.08	1.16×10^{-02}	0.1	6.82×10^{-04}	-0.02	2.5×10^{-01}	-1.12×10^{-03}	4.91×10^{-01}
ADAD vs LOAD	106	-0.09	9.47×10^{-04}	0.12	1.1×10^{-05}	-0.03	6.23×10^{-02}	-1.14×10^{-04}	9.38×10^{-01}

[a] AD Includes all of the AD affected subjects from the Knight-ADRC and DIAN studies.

[b] ADAD: autosomal dominant AD, carriers of pathogenic mutation in *APP* , *PSEN1* or *PSEN2* .

[c] LOAD: late-onset AD patients not carrying any autosomal dominant mutation.

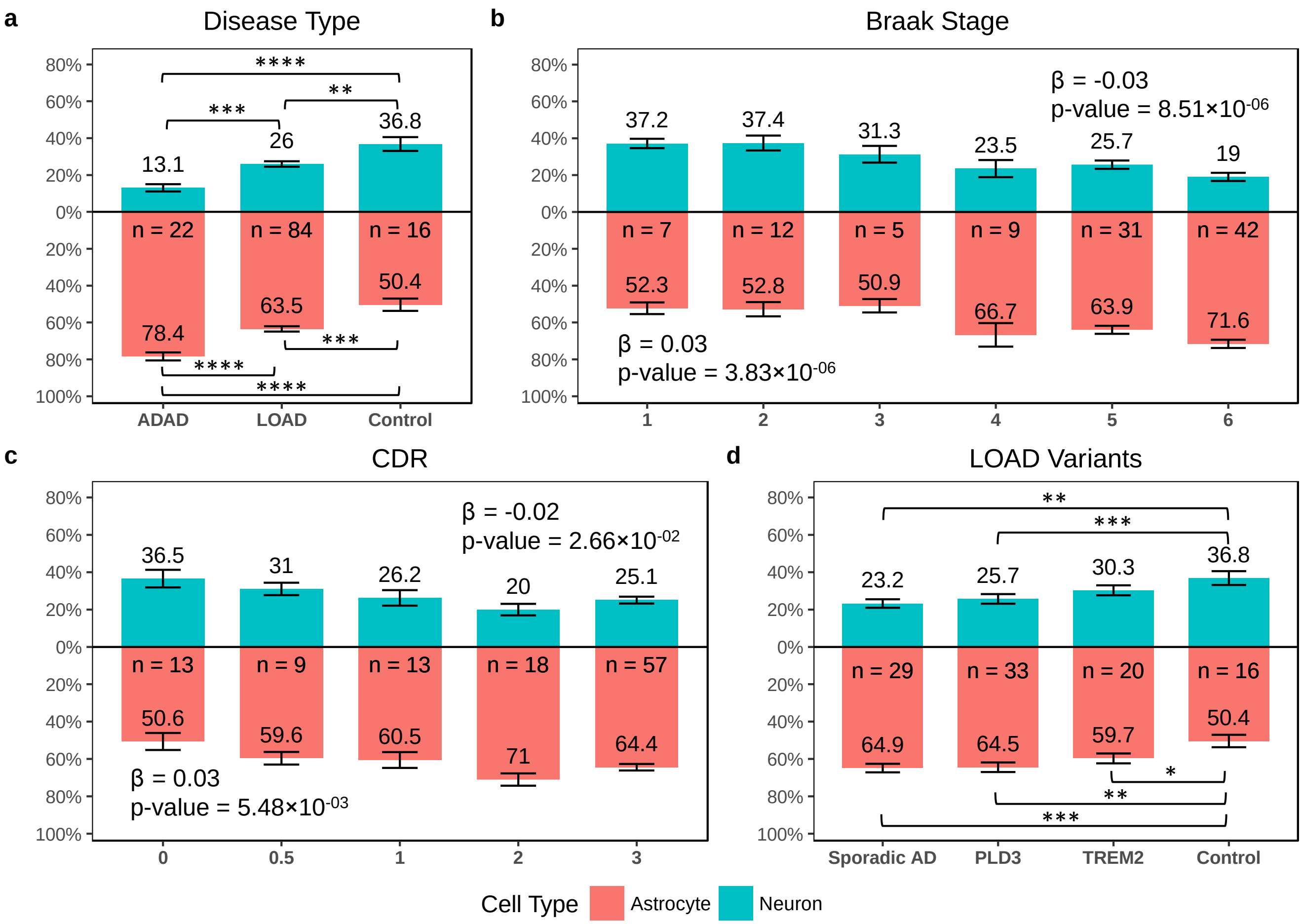


Figure 2.27 Neuron and astrocyte distributions from the DIAN and Knight-ADRC brains. a) Mean neuronal (blue) and astrocytic (red) proportions for carriers of pathogenic mutations in APP, PSEN1 or PSEN2 (ADAD), late-onset AD (LOAD) and neuropath-free controls (bars indicate standard deviations). Neuronal and astrocytic proportions plotted against b) Braak Staging; c) by Clinical Dementia Rating. d) Cell-type distributions for carriers of AD genetic risk factors. Lines indicate significance levels (*P< 0.05; ** P< 0.01; *** P< 1.0×10^{-3}; **** P< 1.0×10^{-4}).

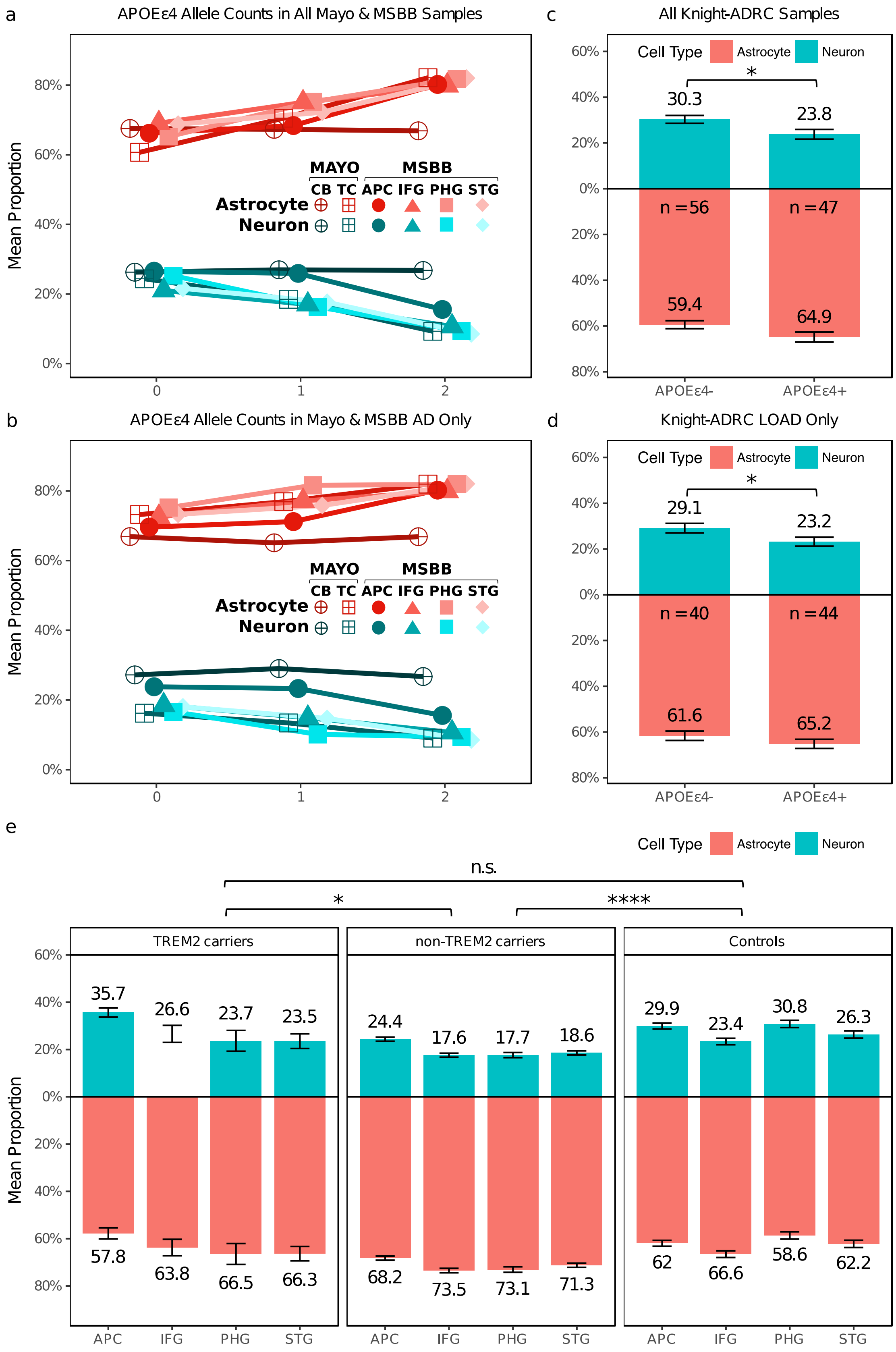
Next, I compared the cell proportion of LOAD vs ADAD and found that the cell composition differs between them. I firstly selected the LOAD brains (N = 25) to match the Braak staging distribution of ADAD brains (N = 17). The ADAD brains showed a significant lower neuronal proportion compared to LOAD brains (β = -0.08; p = 1.03×10^{-02}; **Table 2.14**), and increased relative astrocyte proportion (β = 0.11; p = 9.26×10^{-04}; **Table 2.14**). Then, I analyzed the entire Knight-ADRC LOAD brains, by extending the model to correct for Braak stages. I also observed significant lower relative neuronal proportion (β = -0.09; p = 4.71×10^{-03}; **Table 2.14; Figure 2.27a; Table 2.15)** and increased relative astrocyte proportion (β = 0.11; p = 5.24×10^{-04}; **Table 2.14; Figure 2.27a; Table 2.15** in ADAD brains compared to LOAD. I observed the same cellular differences when I corrected for CDR at death (β = -0.12; p = 2.11×10^{-03} for neurons and β = 0.13; p = 6.29×10^{-04} for astrocytes; **Table 2.14; Figure 2.27bc**). In summary, my results indicate that ADAD individuals present a higher neuronal death even in the same stage of the disease, suggesting that in ADAD neuronal death play a more important role in pathogenesis than sporadic AD, in which other factors such as inflammation or immune response may be involved.

2.4.6 Specific genetic variants confer a distinctive cell composition profile

A variety of genetic variants increase risk of LOAD; however, it is unclear if the cellular mechanisms are the same across these distinct risk factors. Therefore, I tested the hypothesis that distinct genetic causes of LOAD have characteristic cellular population signatures.

I initially ascertained the effect of *APOE* ε4 on the cell-type composition. I observed a significant lower relative proportion of neurons (β = -0.06 for each of the ε4 alleles; p = 9.91×10^{-03}) and increase of relative proportion of astrocytes (β = 0.10; p = 4.15×10^{-02}) from the TC included in the Mayo Clinic dataset (**Table 2.20; Figure 2.28a; Figure 2.29a**). This finding was replicated when I performed a multi-region analysis of the MSBB dataset (β = -0.04; p = 2.60×10^{-03} and β = 0.05; p = 1.31×10^{-03} for neurons and astrocytes respectively; **Table 2.16; Figure 2.28a; Table 2.20; Figure 2.29a**). Given the strong risk conferred by the *APOE* ε4 allele[56], I studied its effects on the cell-type composition by restricting my analysis to AD brains. I observed a significant association in the multi-area analysis of the MSBB dataset (β = -0.03 p = 4.01×10^{-02}; **Table 2.16; Figure 2.28b; Table 2.21; Figure 2.29b**) and also a significant increase in relative proportion of astrocytes (β = 0.03; p = 1.23×10^{-02}; **Table 2.16; Figure 2.28b; Table 2.21; Figure 2.29b**). I also observed a significant decrease in relative proportion of neurons (β = -0.06; p = 2.11×10^{-02}; **Table 2.16**; **Figure 2.28c**) when I analyzed the LOAD and control brains from the Knight-ADRC. When I restricted the analysis to AD brains from the Knight-ADRC and compared the *APOE* ε4 carriers (N = 46) to non-carriers (N = 41) I also observed decreased relative neuronal proportion (β = -0.06; p = 2.69×10^{-02}; **Table 2.16**; **Figure 2.28d**). I extended the models to correct for the Braak stages, and observed a significant association for the relative proportion of neurons with the *APOE* ε4 allele in the Knight-ADRC dataset (β = -0.06; p = 3.66×10^{-02}; **Table 2.16**), and a significant association for the relative proportion of astrocytes in the MSBB (β = 0.04; p = 4.89×10^{-02}; **Table 2.16**). Furthermore, I performed a meta-analysis to combine the evidence of both studies and observed a significant association of the relative neuronal proportion with *APOE* ε4 allele ($p=1.86\times10^{-02}$) and marginally significant association for the relative astrocytic relative proportion (p=0.09).

Figure 2.28 Effect of the *APOE ε4* allele and *TREM2* coding variants on the cellular population structure. Mean neuronal (blue) and astrocytic (red) proportions for **a)** AD cases and controls in the Knight-ADRC brains categorized by *APOE* ε4 carriers vs. non-carriers and **b)** AD cases of Knight-ADRC brain bank (bars indicate standard deviations). **c)** AD cases and controls in the Mayo Clinic and MSBB **d)** AD cases in the Mayo Clinic and MSBB. **e)** Neuronal (blue) and astrocyte (red) distributions for samples included in the Mount Sinai brain bank stratified by *TREM2* genetic status. APC: Anterior Prefrontal Cortex; STG: Superior Temporal Gyrus; PHG: Parahippocampal Gyrus; IFG: Inferior Frontal Gyrus; (n.s. $P > 0.05$; * $P < 0.05$; **** $P < 1.0\times10^{-4}$)


a
APOEε4 Allele Counts in All Mayo & MSBB Samples
Mean Proportion
80%
60%
40%
20%
0%
0
1
2
MAYO
MSBB
CB TC APC IFG PHG STG
Astrocyte
Neuron
c
All Knight-ADRC Samples
Cell Type
Astrocyte
Neuron
*
30.3
23.8
n =56
n =47
59.4
64.9
60%
40%
20%
0%
20%
40%
60%
80%
APOEε4-
APOEε4+
b
APOEε4 Allele Counts in Mayo & MSBB AD Only
Mean Proportion
d
Knight-ADRC LOAD Only
*
29.1
23.2
n =40
n =44
61.6
65.2
APOEε4-
APOEε4+
e
Cell Type
Astrocyte
Neuron
n.s.
*
****
TREM2 carriers
non-TREM2 carriers
Controls
Mean Proportion
35.7
26.6
23.7
23.5
24.4
17.6
17.7
18.6
29.9
23.4
30.8
26.3
57.8
63.8
66.5
66.3
68.2
73.5
73.1
71.3
62
66.6
58.6
62.2
APC
IFG
PHG
STG
APC
IFG
PHG
STG
APC
IFG
PHG
STG

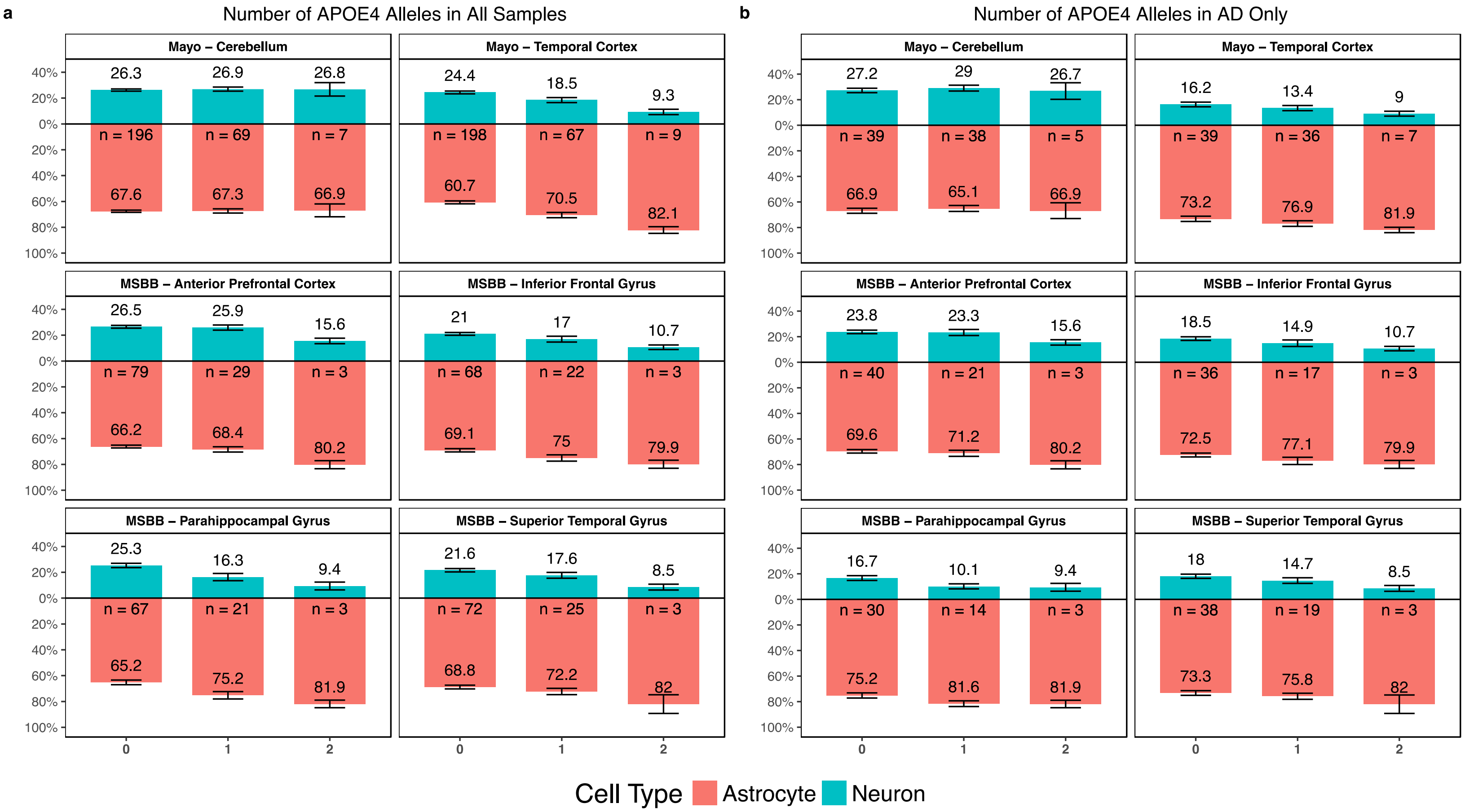


Figure 2.29 Neurons and astrocytes distributions for samples included in the Mayo Clinic and Mount Sinai Brain Bank (MSBB) stratified by APOE ε4 allele. Neuronal (blue) and astrocyte (red) proportions. a) AD Cases and controls. b) Restricted to AD cases

Table 2.16 Gene specific cellular proportion analysis for Knight-ADRC and Mount Sinai Brain Bank studies

Variant Carriers	**Sample Size**	**Neuron**		**Astrocyte**		**Oligodendrocyte**		**Microglia**	
Knight-ADRC	N	Effect	P-value	Effect	P-value	Effect	P-value	Effect	P-value
PLD3 vs Control	49	-0.1	1.6×10^{-04}	0.13	2.84×10^{-03}	-0.03	6.17×10^{-02}	7.05×10^{-04}	7.89×10^{-01}
TREM2 vs Control	36	-0.07	7.93×10^{-02}	0.11	1.05×10^{-02}	-0.03	4.9×10^{-02}	1.65×10^{-03}	5.84×10^{-01}
Sporadic AD vs Control	45	-0.11	5.45×10^{-03}	0.13	2.95×10^{-04}	-0.02	4.55×10^{-01}	-3.48×10^{-03}	1.13×10^{-01}
APOE ε4+ vs *APOE* ε4- LOAD cases & c	100	-0.06	2.11×10^{-02}	0.05	5.35×10^{-02}	0.01	3.72×10^{-01}	-8.09×10^{-04}	6.31×10^{-01}
APOE ε4+ vs *APOE* ε4- LOAD cases on	84	-0.06	2.69×10^{-02}	0.03	2×10^{-01}	0.03	1.4×10^{-02}	-8.31×10^{-04}	6.21×10^{-01}
CDR corrected	84	-0.06	2.78×10^{-02}	0.03	2.05×10^{-01}	0.03	1.16×10^{-02}	-1.05×10^{-03}	5.37×10^{-01}
Braak corrected	73	-0.06	3.66×10^{-02}	0.03	3.72×10^{-01}	0.03	4.51×10^{-03}	-1.14×10^{-03}	5.93×10^{-01}
Mount Sinai Brain Bank - Multi-region									
AD *TREM2* carriers vs Control	301	-0.03	3.57×10^{-01}	0.03	3.19×10^{-01}	-2.08×10^{-03}	7.87×10^{-01}	-2.68×10^{-03}	8.67×10^{-02}
AD non-carriers *TREM2* vs Control	882	-0.07	1.91×10^{-08}	0.08	1.25×10^{-08}	-3.36×10^{-03}	4.79×10^{-01}	-2.89×10^{-04}	7.97×10^{-01}
AD *TREM2* vs AD non-*TREM2*	673	0.05	1.98×10^{-02}	-0.05	1.58×10^{-02}	2.12×10^{-03}	7.76×10^{-01}	-2.13×10^{-03}	1.74×10^{-01}
CDR corrected	673	0.04	5.83×10^{-02}	-0.04	4.46×10^{-02}	1.68×10^{-03}	8.19×10^{-01}	-1.92×10^{-03}	2.22×10^{-01}
Braak corrected	642	0.05	1.3×10^{-02}	-0.05	2.7×10^{-02}	-1.82×10^{-03}	8.13×10^{-01}	-2.66×10^{-03}	1.28×10^{-01}
Mean plaque counts corrected	673	0.05	2×10^{-02}	-0.05	1.59×10^{-02}	1.73×10^{-03}	8.15×10^{-01}	-2.2×10^{-03}	1.5×10^{-01}
APOE ε4 counts all samples	556	-0.04	2.6×10^{-03}	0.05	1.31×10^{-03}	-0.01	4.47×10^{-02}	-3.58×10^{-04}	6.53×10^{-01}
APOE ε4 counts AD cases	225	-0.03	4.01×10^{-02}	0.03	4.23×10^{-02}	-4.52×10^{-03}	3.73×10^{-01}	-5.13×10^{-04}	6.78×10^{-01}
CDR corrected	225	-0.03	2.02×10^{-02}	0.03	2.03×10^{-02}	-4.86×10^{-03}	3.19×10^{-01}	-4.91×10^{-04}	6.93×10^{-01}
Braak corrected	198	-0.03	7.35×10^{-02}	0.04	4.89×10^{-02}	-0.01	8.54×10^{-02}	-1.08×10^{-03}	4.12×10^{-01}

Next, I analyzed the cellular composition in *PLD3* carriers (N = 33). *PLD3* carriers exhibited significantly lower relative proportion of neurons compared to controls (β = -0.10; p = 1.60×10^{-04}; **Figure 2.27d**) and a significant higher relative proportion of astrocytes (β = 0.13; p = 2.84×10^{-03}; **Table 2.16; Figure 2.27d**). Sporadic AD non-carrier cases also exhibited significantly lower relative proportion of neurons compared to controls (β = -0.11; p = 5.45×10^{-03}) and significant higher relative proportion of astrocytes (β = 0.13; p = 2.95×10^{-04}; **Table 2.16; Figure 2.27d**). The cell proportion between sporadic AD non-carriers and *PLD3* carriers did not show any significantly difference (p > 0.05).

Finally, I performed similar analyses with *TREM2* carriers. *TREM2* is involved in the immune response and its role in amyloid-β deposition or clearance remain controversial[263]. My analysis on the Knight-ADRC data showed significantly higher relative astrocytic proportion in AD affected *TREM2* carriers (N = 20) compared to controls (β = 0.11; p = 1.05×10^{-02}; **Table 2.16**; **Figure 2.27d**). Despite *TREM2* carriers presented lower neuron relative proportion compared to controls, this difference was not statistically significant (p>0.05; **Table 2.16**; **Figure 2.27d**). I analyzed whether the *TREM2* carriers provided sufficient power to detect a significant association. My empirical estimates showed that *TREM2* sample size provides 96% of power to detect an association with an effect size comparable to that observed for sporadic AD (β = -0.11). I also investigated the cellular proportion of the eleven *TREM2* carriers in the MSBB dataset. The multi-region analysis showed *TREM2* carriers do not show a significant difference in relative neuronal proportion compared to controls (p > 0.05; **Table 2.16; Figure 2.28e**), whereas in the AD *TREM2* non-carriers the relative neuronal and astrocytic proportions are significantly different from controls (β = -0.07; p = 1.91×10^{-08} and β = 0.08; p = 1.25×10^{-08} respectively; **Table 2.16**; **Figure 2.28e**).

In fact, my analyses indicate that *TREM2* carriers have a unique cellular brain composition distinct than the other AD cases. *TREM2* brains showed significantly higher relative neuronal proportion ($\beta = 0.05$; $p = 1.98\times10^{-02}$) and significantly lower relative astrocyte proportion than the AD *non*-carries ($\beta = -0.05$; $p = 1.58\times10^{-02}$; **Table 2.16**). The distribution of CDR, mean number of amyloid plaques and Braak staging do not differ between strata. Nonetheless, I verified that the cellular proportions were still significant after correcting for each of those variables (**Table 2.16**). These results suggested that the mechanism that lead to disease in *TREM2* carriers is less neuron-centric than in the general AD population.

2.5 Discussion

I have developed, optimized and validated a digital deconvolution approach to infer cell composition from bulk brain gene expression that integrates publicly available cell-type specific expression data while addressing the heterogeneity of the phenotypic differences of samples and technical characteristics of transcriptome ascertainment. I acknowledge that the accuracy of this platform might be affected by the phenotypic diversity of the reference panel or the disease-induced dysregulation of genes it includes. However, the deconvolution approach proved to be robust to the genes included in the reference panel, as I demonstrated that the proportions it inferred are not driven by the expression of any single gene. This platform produced reliable cell proportion estimates, as was shown by the evaluation of independent datasets of iPSC-derived neurons and microglia, mice cortical neurons (**Figure 2.22**) and simulated chimeric libraries.

I used this approach to deconvolve studies that include large number of neuropathologically defined AD and control brains with their transcriptome ascertained in distinct brain regions and observed consistently significant lower relative neuronal proportion and increased relative astrocyte proportions in the cerebral cortex suggesting neuronal loss and astrocytosis. Compatible with other studies, I also identified that the altered cellular proportion is also significantly associated with decline in cognition and Braak staging[239]. In contrast, I did not identify a significant difference in the cellular population structure in the cerebellum, a region not affected in AD (**Table 2.10; Figure 2.25a**).

We generated RNA-Seq data from brains carrying pathogenic mutations in *APP, PSEN1, PSEN2*, which cause alterations in Aβ processing and lead to ADAD, and also generated RNA-Seq from brains of LOAD and neuropath-free controls. I observed altered cell composition in both ADAD and LOAD compared to controls. However, I identified that ADAD brains have a different cell-type composition than disease-stage-matched LOAD, as the ADAD has a significantly lower relative neuronal proportion and more pronounced astrocytosis. Given the specific cellular population structure of the *TREM2* carriers, I compared the neuronal and astrocytic relative proportion of ADAD to that of LOAD non-carriers of variants in *TREM2* and observed significant differences ($\beta = -0.09$ and $p = 6.89\times10^{-03}$ for neurons and $\beta = 0.10$; $p = 1.49\times10^{-03}$ for astrocytes). This indicates that the difference of the relative proportion between ADAD and LOAD are not driven by *TREM2* carrier brains. Based on my results, I would hypothesize that this change in Aβ processing of ADAD would lead to more direct to neuronal death than the pathological processes of LOAD. Similarly, decreased neuronal and increased astrocyte relative proportions were significantly associated with *APOEε4* allele. It has been reported APOE ε4 allele increase the risk for AD by affecting APP metabolism or Aβ clearance[44, 151], suggesting a direct link between APP metabolism and neuronal death.

In contrast, the analysis of the Knight-ADRC brains showed that the neuronal relative proportion decrease is less pronounced in *TREM2* carriers than in other LOAD cases. I replicated this finding in a multi-area analysis from the MSBB dataset. These results may implicate that

TREM2 risk variants lead to a cascade of pathological events that differ from those occurring in sporadic AD cases, which is also consistent with the known biology of *TREM2*. Further longitudinal neuroimaging analysis are required to validate my findings. *TREM2* is involved in AD pathology through microglia mediated pathways, implicated on altered immune response and inflammation[54]. Recent studies in *TREM2* knock-out animals showed that fewer microglia cells were found surrounding Aβ plaques with impaired microgliosis[275]. Furthermore, *TREM2* deficiency was reported to attenuate tauopathy against brain atrophy[168]. I found no significant difference in the proportion of microglia between AD cases and controls. However, I found significantly decreased microglia in brains exhibiting pathological aging (**Table 2.12; Figure 2.24**), proving that these studies are sufficiently powered to identify significant differences. In any case, I cannot rule out the possibility of a change in the activation stage of microglia in these individuals. Overall, these results suggest that *TREM2* affects AD risk through a slightly different mechanism to that of ADAD or LOAD in general. Therefore, other pathogenic mechanisms should contribute to disease. I believe that a detailed modeling of immune response cells, reflecting the alternative microglia activation states, will generate more accurate profiles to elucidate the immune cell distribution in AD.

2.6 Conclusions

There is a large interest in the scientific community to use brain expression studies to try to identity novel pathogenic mechanism in AD and to identify novel therapeutic targets. These efforts are generating a large amount of bulk RNA-Seq data, as single-cell RNA (scRNA-Seq) from human brain tissue in large sample size is not feasible. Single-cell sorting needs to be performed with fresh tissue[115], which restrains the analysis of highly characterized fresh-frozen brains collected by AD research centers. My results indicate that digital deconvolution methods can accurately infer relative cell distributions from brain bulk RNA-Seq data, but I recognize the importance of obtaining traditional neuropathological measures to validate the results I observed. Having this approach validated for AD can have an important impact in the community, because digital deconvolution analyses 1) can reveal distinct cellular composition patterns underlying different disease etiologies; 2) can provide additional insights about the overall pathologic mechanisms underlying different mutations carriers for variants as in genes such as *TREM2*, *APOE*, *APP*, *PSEN1* and *PSEN2*; 3) can correct the effect that altered cell composition and genetic statuses have in addition to downstream transcriptomic analyses and lead to novel and informative results; 4) can help the analysis of highly informative frozen brains collected over the years.

In conclusion, my study provides a reliable approach to enhance our understanding of the fundamental cellular mechanisms involved in AD and enable the analysis of large bulk RNA-Seq data that may lead to novel discoveries and insights into neurodegeneration.

Chapter 3: The *TMEM106B* rs1990621 protective variant is associated with increased neuronal proportion

3.1 Abstract

Background: In previous studies, I observed decreased neuronal and increased astrocyte proportions in AD cases in parietal brain cortex by using a deconvolution method for bulk RNA-Seq. These findings suggested that genetic risk factors associated with AD etiology have a specific effect in the cellular composition of AD brains. The goal of this study is to investigate if there are genetic determinants for brain cell compositions.

Methods: Using cell type composition inferred from transcriptome as a disease status proxy, I performed cell type association analysis to identify novel loci related to cellular population changes in disease cohort. We imputed and merged genotyping data from seven studies in total of 1,669 samples and derived major CNS cell type proportions from cortical RNA-Seq data. I also inferred RNA transcript integrity number (TIN) to account for RNA quality variances. The model I performed in the analysis was: normalized neuronal proportion ~ SNP + Age + Gender + PC1 + PC2 + median TIN.

Results: A variant rs1990621 located in the *TMEM106B* gene region was significantly associated with neuronal proportion (p=6.40×10^{-07}) and replicated in an independent dataset. The association passed genome-wide multiple test correction in the multi-tissue meta-analysis (p=9.42×10^{-09}) and joint analysis (p=7.66×10^{-10}). This variant is in high LD with rs1990622 (r^2 = 0.98) which was previously identified as a protective variant for FTD with TDP-43 inclusion. Further analyses indicated that this variant is associated with increased neuronal proportion in participants with neurodegenerative disorders, not only in AD cohort but also in cognitive normal elderly cohort. However, this effect was not observed in a younger schizophrenia cohort with a mean age of death < 65. The second most significant loci for neuron proportion was

APOE, which suggested that using neuronal proportion as an informative endophenotype could help identify loci associated with neurodegeneration.

Conclusion: This result suggested a common pathway involving *TMEM106B* shared by aging groups in the present or absence of neurodegenerative pathology may contribute to cognitive preservation and neuronal protection.

3.2 Introduction

3.2.1 Alzheimer’s disease in the context of multi-cell type interactions

Although neuronal loss and synapse dysfunction are the preceding events of cognitive deficits in Alzheimer’s disease (AD), neurons do not work or survive by themselves. These delicate organelles require supports through intimate collaborations within themselves and with other cell types[125]. The microenvironment of cellular crosstalk, interaction, balance, and circuits maintained by neurons, astrocytes, microglia, oligodendrocytes, and other vascular cells are essential for the brain to carry out functions and fight against insults.

AD associated risk factors identified across the genome also point to the involvements of multi-cell types apart from neurons[125, 161]. APOE4 is related to lipid metabolism and mostly expressed in astrocyte and microglia[56]. Other lipid metabolism related risk genes are *ABCA7* identified in all cell types[130, 161], *CLU* in astrocyte and oligodendrocyte precursor cells[119, 160, 161], and *SORL1* in astrocyte[161]. Research interests in the roles of inflammatory response to toxic stimuli or microbial infection have been escalating recently, and AD risk genes associated with immune response including *TREM2*[114, 144, 246], *PLCG2*[246], *ABI3*[246], *CR1*[160, 161], *CD33*[130, 161], *HLA-DRB5–HLA-DRB1*[161], and *INPP5D*[161] are mostly expressed in microglia and macrophages. *BIN1* expressed in microglia, oligodendrocyte, and neurons[161], and *PICALM* expressed in microglia and endothelial cells[119, 161] are associated with endocytosis.

In a normal functional brain, astrocytes, microglia, and oligodendrocytes provide trophic supports to neurons and various cell type specific functions. Astrocytes confer multiple functions to fulfill neurons' metabolic needs[250] including but not limited to providing substrates for oxidative phosphorylation[210], exerting regulation of excitatory CNS neurotransmitter glutamate[76, 93], and serving as bidirectional communication nodes that talk to both neurons and blood vessels and modulate their activities in an arrangement of functional entities named neurovascular units[227, 244, 260]. Microglia surveil in the extracellular space and look for pathogens or debris to engulf through phagocytosis. Oligodendrocyte provides insulation to neurons by wrapping around the axons with myelin sheath. However, in an AD diseased brain, these supporting cells may become double-edged swords that play beneficial and/or harmful roles as disease progresses. Amyloid-β accumulation and clearance are the central events of the amyloid cascade hypothesis. Both astrocyte and microglia have been involved in response to the toxic stimuli of amyloid plaques. During the early stage, microglia[124, 126, 127] and astrocytes[126, 209, 225] accumulate around plaques to phagocytose or degrade those in a protective manner. However, as disease progresses, the chronic and prolonged activation of microglia and astrocytes will be provoked into a damaging pro-inflammatory state and a vicious circle that exacerbate pathology in a harmful manner. Evidence suggested that increased inflammatory cytokine secretion in microglia, and increased production of complement cascade components, and impaired glutamate regulation (unregulated glutamate activity can cause neuronal excitatory cell death)[76] may contribute to synaptic loss which ultimately leads to cognitive deficits. Disrupted neuronal plasticity due to myelin loss and dysfunctional neurovascular units further exacerbate the dreadful situation and destroy the harmony of the multi-cell type microenvironment.

3.2.2 Cell type composition inferred from bulk RNA-Seq deconvolution

Apart from disturbed homeostatic processes and impaired circuits integrity, cell type composition or proportion is also altered. Brains affected by AD exhibits neuronal loss, oligodendrocyte loss, astrocytosis, and microgliosis. However, the specific effects that pathological mutations and risk variants have on brain cellular composition are often ignored. To investigate the changes of cerebral cortex cell-type population structure and account for the associated confounding effects in downstream analysis, I developed an *in-silico* deconvolution method to infer cellular composition from RNA-Seq data, which has been documented in my previous publication[172], and explained in depth in **Chapter 2**. In summary, I firstly assembled a reference panel to model the transcriptomic signature of neurons, astrocytes, oligodendrocytes and microglia. The panel was created by analyzing expression data from purified cell lines. I evaluated various digital deconvolution methods and selected the best performing ones for my primary analyses. I tested the digital deconvolution accuracy on induced pluripotent stem cell (iPSC) derived neurons and microglia, and neurons derived from Translating Ribosome Affinity Purification followed by RNA-Seq. Finally, I verified its accuracy with simulated admixture with pre-defined cellular proportions.

Once the deconvolution approach was optimized, I calculated the cell proportion in AD cases and controls from different brain regions of LOAD and ADAD datasets. I found that neuronal and astrocyte relative proportions differ between healthy and diseased brains, and also differ among AD cases that carry different genetic risk variants. Brain carriers of pathogenic mutations in *APP*, *PSEN1* or *PSEN2* presented lower neuronal and higher astrocytes relative proportions compared to sporadic AD. Similarly, *APOE* ε4 carriers also showed decreased neuronal and increased astrocyte relative proportions compared to AD non-carriers. In contrast, carriers of variants in *TREM2* risk showed a lower degree of neuronal loss than matched AD cases in multiple independent studies. These findings suggest that different genetic risk factors associated with AD etiology may have gene specific effects in the cellular composition of AD brains.

3.2.3 Use cell type composition in cell type QTL (cQTL) to identify novel loci for AD risk

In a recently published study named PsychENCODE[272], a very similar deconvolution approach as reported in my previous study[172] was taken to infer cell type composition from RNA-Seq data predominantly drawn from psychiatric disorder cohorts. From the cell fractions inferred from bulk RNA-Seq data, they found that cell type composition differences can account for more than 88% of bulk tissue expression variation observed across the population with a ±4% variance on a per-subject level. Using cell type compositions as quantitative traits, the authors identified a non-coding variant closed to the *FZD9* gene that is associated with both *FZD9* gene expression and the proportion of excitatory layer 3 neurons[272]. Interestingly, deletion variants found previously upstream of *FZD9* were associated with cell composition changes in Williams syndrome[45], a developmental disorder exhibits mild to moderate intellectual disabilities with learning deficits and cardiovascular problems. This observation re-emphasized the importance of incorporating cell type composition into RNA-Seq analysis pipeline even in psychiatric disorder cohorts without dramatic changes in cellular composition, not mention the necessity of such practice in neurodegeneration disorders that have significant changes in cell type composition. It also demonstrated the great potential of using relative abundance of specific cell types in identifying novel variants and genes implicated in disease. However, it is unclear if this finding is only applicable to psychiatry-relate traits or it is a more general finding.

In this study, I utilized cell-type proportions inferred from my deconvolution method[172] to perform cell type QTL analysis in a dataset enriched for AD cases in search for potential new loci that are associated with neurodegeneration disorders. We imputed and merged genotyping or whole genome sequencing data from seven studies - five centered on neurodegeneration (N = 1,125), one schizophrenia cohort (N = 414), and GTEx multiple tissue controls (N = 130). From cortical RNA-Seq data, I derived cell fractions of four major CNS cell types, including neuron, astrocyte, microglia, and oligodendrocyte. Using normalized neuronal proportion as quantitative trait, I identified a variant rs1990621 located in the *TMEM106B* gene region significantly associated with neuronal proportion variation in all cohorts except schizophrenia subjects. This variant is in high LD with rs1990622 ($r^2 = 0.98$), which was previously identified as a protective variant in FTD cohorts[266]. Variants in this region have also been found to be associated with AD with TDP-43 pathology[229], and downregulation of *TMEM106B* is observed in AD brains[234] . In conclusion, I have identified a variant associated with neuronal proportion with potential protective effect in neurodegeneration disorders.

3.3 Methods

3.3.1 Study participants

The participants were sourced from seven studies with a total sample size of 1,669 (**Table 3.1**). Among those, five studies are mainly focused on neurodegenerative disorders including Alzheimer’s disease (N = 681), frontotemporal dementia (N = 11), progressive supranuclear palsy (N = 82), pathological aging (N = 29), Parkinson Disease (N = 1), as well as cognitive normal individuals (N = 540). These samples come from the Mayo, MSSM, Knight

ADRC, DIAN, and ROSMAP studies as described in table 3.1. To compare with the neurodegenerative disorders, I also included schizophrenia (N = 210) and bipolar disorders (N = 34) participants from the CommonMind study (**Table 3.1**). Additionally, two studies, MSSM and GTEx, contain multi-tissue data that include some participants contribute more than one tissue (**Table 3.1**).

3.3.2 Standard protocol approvals, registrations and patient consents

The protocol of DIAN and Knight-ADRC studies have been approved by the review board of Washington University in St. Louis. The protocol of Mayo dataset was approved by the Mayo Clinic Institutional Review Board (IRB). All neuropsychological, diagnostic and autopsy protocols of MSSM dataset were approved by the Mount Sinai and JJ Peters VA Medical Center Institutional Review Boards. The religious orders study and the memory and aging project of ROSMAP was approved by the IRB of Rush University Medical Center. The NIH Common Fund's GTEx program protocol was reviewed by Chesapeake Research Review Inc., Roswell Park Cancer Institute's Office of Research Subject Protection, and the institutional review board of the University of Pennsylvania. Within CommonMind consortium, the MSSM sample protocol was approved by Icahn School of Medicine at Mount Sinai IRB; the Pitt sample protocol was approved by the University of Pittsburgh's Committee for the Oversight of Research involving the Dead and IRB for Biomedical Research; the Penn sample protocol was approved by the Committee on Studies Involving Human Beings of the University of Pennsylvania. All participants were recruited with informed consent for research use.

Table 3.1 Demographic information for cohorts included in the study. AD: Alzheimer's Disease; FTD: frontal temporal dementia; PSP: progressive supranuclear palsy; PA: pathological aging; PD: Parkinson's Disease; SCZ: schizophrenia; BP: bipolar disease; OTH: other unknown dementia or no diagnosis information.

	N	Region	Age	% Male	RIN	TIN	Control	AD	FTD	PSP	PA	PD	SCZ	BP	OTH
							Discovery								
ROSMAP	523	1	86.6 ± 4.59	35.4	7.07 ± 0.99	73.2 ± 5.13	114	338	0	0	0	0	0	0	71
							Replication								
Mayo	260	1	80.4 ± 8.37	48.1	8.16 ± 0.903	77.4 ± 5.94	69	80	0	82	29	0	0	0	0
MSSM	219	4	84 ± 7.32	35.6	6.42 ± 1.77	76.4 ± 2.52	49	170	0	0	0	0	0	0	0
Knight ADRC	108	1	83.1 ± 12	42.6	6.44 ± 1.2	79.4 ± 1.91	13	77	11	0	0	1	0	0	6
DIAN	15	1	50.9 ± 7.08	73.3	5.55 ± 1.09	78.9 ± 0.99	0	15	0	0	0	0	0	0	0
GTEx	130	3	58.2 ± 9.91	67.7	6.92 ± 0.846	73.8 ± 2.97	125	1	0	0	0	0	0	0	4
CommonMind	414	1	64.6 ± 18	62.3	7.67 ± 0.899	50 ± 7.21	170	0	0	0	0	0	210	34	0
Replication Total	1,146						426	343	11	82	29	1	210	34	10
Merged Total	1,669						540	681	11	82	29	1	210	34	81

Table 3.2 General information of seven studies evolved in the analysis. TCX: temporal cortex; PAR: parietal cortex; CTX: cortex; FCX: frontal cortex; DLPFC: dorsal lateral prefrontal cortex. BM9: dorsal lateral prefrontal cortex; BM10: Anterior prefrontal cortex; BM22: superior temporal gyrus; BM24: ventral anterior cingulate cortex; BM36: parahippocampal gyrus; BM44: inferior frontal gyrus. Mean coverage unit is million.

Dataset	Brain Region	Library Type	Read Length	mRNA Enrichment	Sequencer	Mean Coverage	DNA type	Reference
				Discovery				
ROSMAP	DLPFC	Paired end	101	ploy-A selection	HiSeq 2000	99.2 ± 29.29	WGS	Bennett 2012; Bennett 2012
				Replication				
Mayo	TCX	Paired end	101	ploy-A selection	HiSeq 2000	158.31 ± 34.04	Genotype	Allen 2016
MSSM	BM10 BM22 BM36 BM44	Single end	100	rRNA depletion	HiSeq 2500	35.96 ± 10.04	WGS	Wang 2018
Knight-ADRC	PAR	Paired end	150	rRNA depletion	HiSeq 4000	137.87 ± 21.81	Genotype	Li 2018
DIAN	PAR	Paired end	150	rRNA depletion	HiSeq 4000	149.82 ± 19.68	Genotype	Li 2018
GTEx	BM24 CTX FCX	Paired end	76	ploy-A selection	HiSeq 2000	48.28 ± 13.2	WGS	GTEx 2013; Battle 2017
CommonMind	BM9	Paired end	100	rRNA depletion	HiSeq 2500	86 ± 21.12	Genotype	Fromer 2016

3.3.3 Data collection and generation

Cortical tissues from various locations of post-mortal brains were collected (**Table 3.2**). RNA was extracted from lysed tissues and prepared into libraries of template molecules ready for subsequent next-generation sequencing steps. Ribosomal RNAs constitute 80%-90% of total RNAs, which are not the targets of this study. To focus on mRNA quantification usually researchers would either remove excessive rRNAs or enrich for mRNAs during RNA-Seq library preparation. In this study, DIAN[172], Knight ADRC[172], MSSM[274], and CommonMind[95] took a rRNA depletion approach to removed ribosomal RNA from total RNAs to retain a higher mRNA content. Whereas, Mayo[9], ROSMAP[27, 28], and GTEx[6, 24] took a poly-A enrichment approach to enrich mRNAs from total RNAs. Genotype information were also collected and sequenced correspondingly. RNA-Seq paired with genotype data for each participant were either sequenced at Washington University for DIAN and Knight-ADRC studies or downloaded from public database for all the other studies. Please see **Table 3.2** and each study reference(s) for more data collection and generation specifications.

3.3.4 Data QC and preprocessing

Genetic Data

Stringent quality control (QC) steps were applied to each genotyping array or sequence data. The minimum call rate for single nucleotide polymorphisms (SNPs) and individuals was 98% and autosomal SNPs not in Hardy-Weinberg equilibrium (p-value < 1×10^{-06}) were excluded. X-chromosome SNPs were analyzed to verify gender identification. Unanticipated duplicates and cryptic relatedness (Pihat ≥ 0.25) among samples were tested by pairwise genome-wide estimates of proportion identity-by-descent. EIGENSTRAT[215] was used to calculate principal components. The 1000 Genomes Project Phase 3 data (October 2014), SHAPEIT v2.r837[67], and IMPUTE2 v2.3.2[134] were used for phasing and imputation. Individual genotypes imputed with probability < 0.90 were set to missing and imputed genotypes with probability ≥0.90 were analyzed as fully observed. Genotyped and imputed variants with MAF < 0.02 or IMPUTE2 information score < 0.30 were excluded. WGS data quality is censored by filtering out reads with sequencing depth DP < 6 and quality GQ < 20 followed by similar QC approaches as described above for genotyping data. After the QC, all studies including imputed genotype and WGS data was merged into a binary file using Plink for downstream analysis. PCA and IBD analyses were performed on the merged binary files using Plink to keep European ancestry and unrelated participants (**Figure 3.1** and **Figure 3.2**).

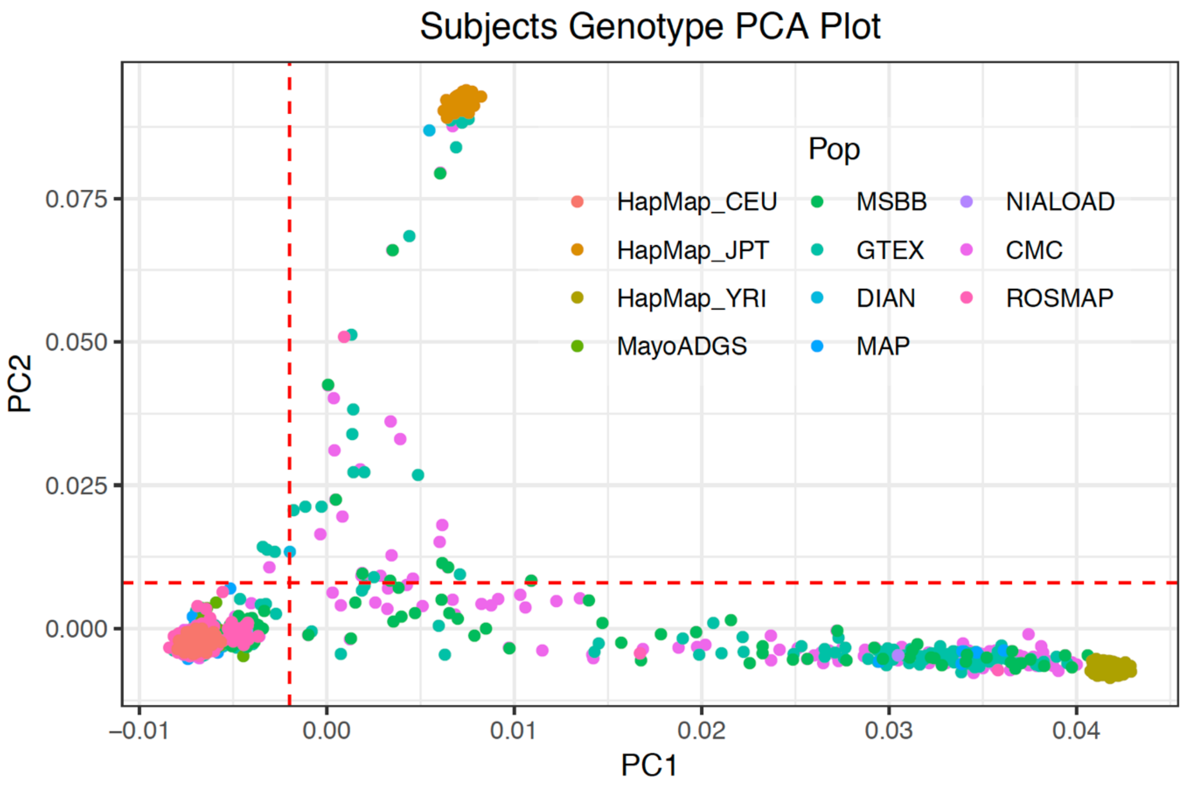


Figure 3.1 Genomic PCA analysis. Genotype data PCA analysis was performed to select European ancestry subjects with PC1 < -0.002 and PC2 < 0.008 with red dotted cut-off lines. HapMap_CEU: HapMap Utah residents with Northern and Western European ancestry; HapMap_JPT: HapMap Japanese in Tokyo, Japan; HapMap_YRI: HapMap Yoruba in Ibadan, Nigeria; MayoADGS: Mayo Clinic study participants; MSBB: MSSM study participants; GTEX: GTEx study participants; DIAN: DIAN study participants; MAP: Knight-ADRC participants; NIALOAD: Knight-ADRC participants; CMC: CommonMind participants; ROSMAP: ROSMAP participants.

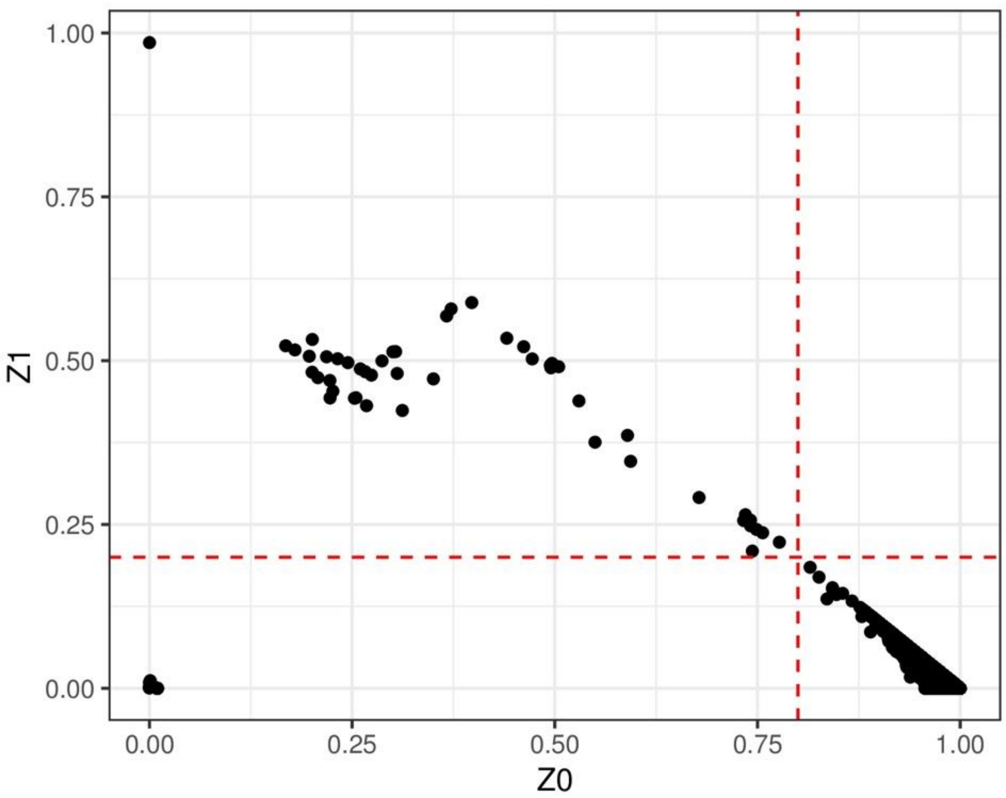


Figure 3.2 Genomic IBD analysis. IBD analysis was performed to select unrelated subjects with Z0 > 0.8 and Z1 < 0.2 with red dotted cut-off lines. When there are related individuals, one individual will be dropped from the related pair.

<u>Expression Data</u>

FastQC was applied to RNA-Seq data to examine various aspects of sequencing quality[231]. Outlier samples with high rRNA contents or low sequencing depth were removed from the pool. The remaining samples were aligned to human GRCh37 primary assembly using Star with 2-Pass Basic mode (ver 2.5.4b)[74]. Alignment metrics were ascertained by applying Picard CollectRnaSeqMetrics[4] including reads bias, coverage, ribosomal contents, coding bases, and etc. Following which, transcript integrity number (TIN) for each transcript was calculated on aligned bam files using RSeQC tin.py[273] (ver 2.6.5). RNA-Seq coding gene and transcript expression was quantified using Salmon transcript expression quantification (ver 0.7.2) with GENCODE *Homo sapiens* GRCh37.75 reference genome[208].

Four major central nerve system cell type proportions were inferred from RNA-Seq gene expression quantification output as documented in my previous deconvolution study[172]. To briefly explain the deconvolution process, I firstly assembled a reference panel to model the transcriptomic signature of neurons, astrocytes, oligodendrocytes and microglia from purified single cell tissue sources respectively. Using the reference panel and the method population-specific expression analysis[157] (PSEA, also named meanProfile in CellMix implementation[102]), I calculated four cell type proportions for each subject bulk RNA-Seq data. For each brain tissue collection site of each study, outlier values for each cell type proportion were removed. Mean values for each cell type of each tissue in each study were subtracted from the deconvolution results to center all the distributions to zero mean (**Figure 3.3**). Phenotype information from all studies were merged and unified to the same coding paradigm to enable downstream joint analysis; for example, males are all coded as 1 and females are 2.

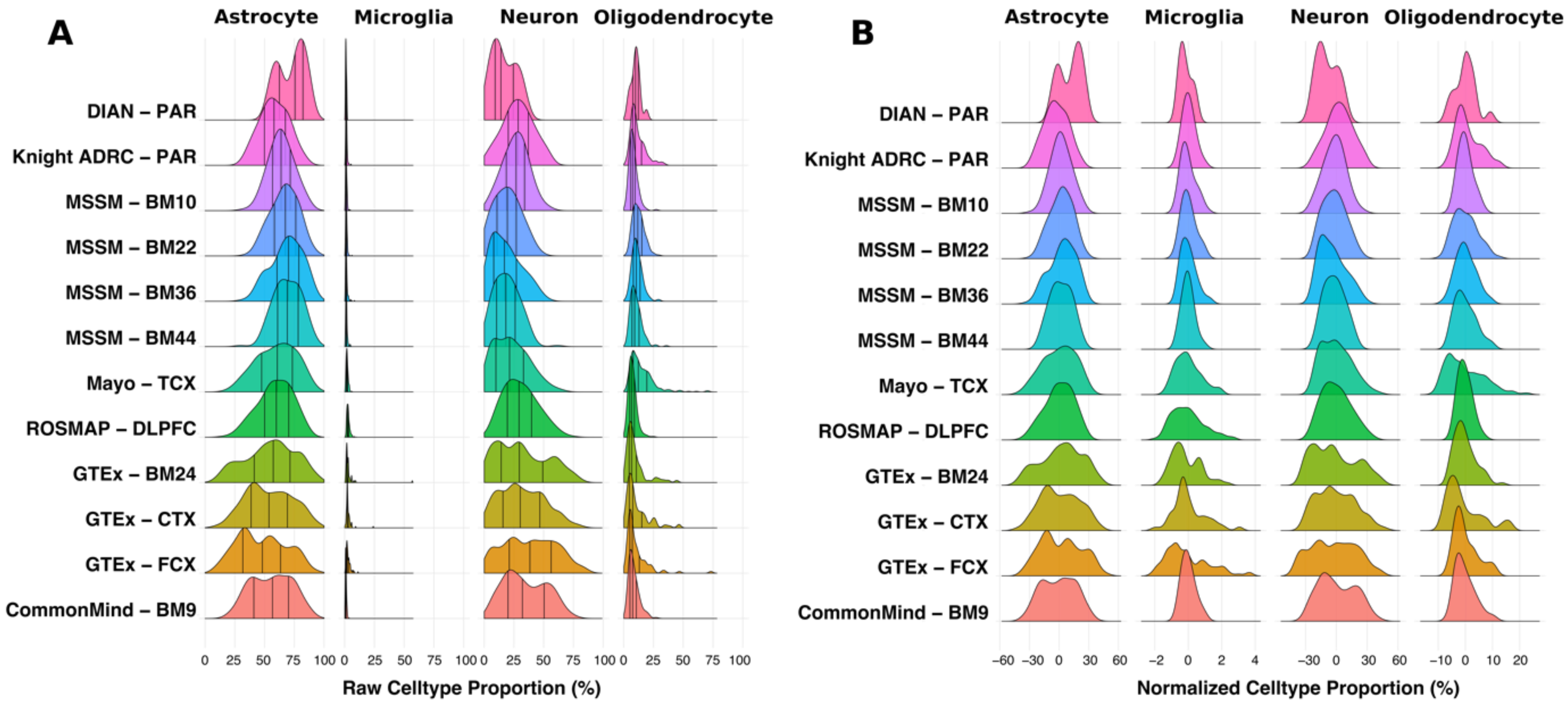


Figure 3.3 Major CNS cell type proportions derived from RNA-Seq datasets with each row representing each tissue of each study. A) raw cell type proportions inferred from the data with vertical bars indicating quantiles within each tissue and each cell type. B) cell type proportions were normalized by subtracting the mean from each tissue deconvolution result after removing outliers.

3.3.5 Data analysis

For the discovery phase, ROSMAP dataset was analyzed with linear regression model employed in Plink[217] using normalized neuronal proportion to run quantitative trait analysis. Age, sex, PC1, PC2, and median TIN were used as covariates to account for potential genetic, phenotypic or technical heterogeneity. TIN is calculated directly from post-sequencing results that captures RNA degradation by measuring mRNA integrity directly[273]. Results were depicted as Manhattan plots using R (ver 3.4.3) qqman package[261] (ver 0.1.4).

For the replication phase, all the other studies except ROSMAP were combined and prepressed to run meta-tissue QTL analysis because MSSM and GTEx contain samples with multiple cortical tissues. Meta-Tissue software installation and data preprocessing were conducted following a four-step instruction documented in the developer website: http://genetics.cs.ucla.edu/metatissue/install.html. Meta-tissue[253] processing pipeline calls two main functions, firstly MetaTissueMM[253] and then followed by Metasoft[117]. MetaTissueMM applies a mixed model to account for the heterogeneity of multiple tissue QTL effects. Metasoft performs the meta-analysis while proving a more accurate random effect p-value for multiple tissue analysis and a m-value based on Bayesian inference to indicate how likely a locus is a QTL in each tissue. Similarly, results were depicted as Manhattan plots and visually examined.

For the final merging phase, both discovery and replication studies were combined to maximize sample size. Apart from meta-tissue analysis by each tissue of each study, a split by disease status analysis was also performed in the final merging phase. Samples from each tissue of each study were also split into disease categories. Resultant subcategory with less than 20 subjects were removed from the analysis to avoid false results due to too small sample size. Similar data preparation and analysis pipeline were enforced as documented above.

QTL analysis results were uploaded to Fuma (v1.3.3d)[276] to annotation significant SNPs (p-value < 10^{-06}) with GWAScatalog (e91_r2018-02-06) and ANNOVAR (updated 2017-07-17). Gene-based analysis was also performed by Magma (v1.06)[63] implemented in Fuma.

3.3.6 Data availability

Mayo: https://www.synapse.org/#!Synapse:syn5550404

MSSM: https://www.synapse.org/#!Synapse:syn3157743

ROSMAP: https://www.synapse.org/#!Synapse:syn3219045

CommonMind: https://www.synapse.org/#!Synapse:syn2759792

GTEx: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2

Knight-ADRC: https://www.synapse.org/#!Synapse:syn12181323

According to the data request terms, DIAN data are available upon request: http://dian.wustl.edu

3.4 Results

3.4.1 Study design

The ROSMAP study containing 523 subjects will be the discovery dataset, and the other six studies are collapsed into replication dataset with 1,146 subjects. Altogether, I have assembled a set of cortical RNA-Seq data comprised of 1,669 participants predominantly focused on neurodegenerative disorders from seven sources (**Figure 3.4**, **Table 3.1**). Collectively, Mayo, MSSM, Knight ADRC, and ROSMAP studies contributed 664 sporadic AD cases. Apart from sporadic AD, 15 subjects from DIAN study and 2 from Knight-ADRC also harbor *PSEN1*, *PSEN2*, and *APP* mutations that exhibit familial AD inheritance pattern. Other neurodegenerative disorders, including progressive supranuclear palsy (PSP), pathological aging (PA), frontal temporal dementia (FTD), and Parkinson's Disease (PD), are mainly drawn from Mayo and Knight ADRC datasets. Other psychiatric disorders including schizophrenia and bipolar disorders are contributed by the CommonMind study. Besides, 540 control subjects cleared of cognitive dementia or neuropsychiatric symptoms were also included. MSSM and GTEx also included multiple tissue data, which were collected from multiple regions of the same subjects that allow us to perform region specific comparison within the same cohort.

Discovery analysis was performed in ROSMAP study. In the replication phase, all the other studies were merged to replicate signals identified from the discovery ROSMAP set. Because GTEx and MSSM contain multiple cortical regions collected from the same subjects, I also applied meta-tissue software[253] specifically designed for multi-tissue QTL analysis to perform a mixed model analysis with random effects that account for correlated measurements from multi-tissue individuals. To attain the largest available sample size for this study, the discovery and replication sets were merged to perform the merged multi-tissue QTL analysis in a search for additional signals hidden in previously separated discovery or replication analysis due to lack of power. After merged analysis, the cohorts were split into four major disease status groups (AD, control, schizophrenia, other non-AD neurodegenerative disorders) to explore how different disease strata could impact the results.

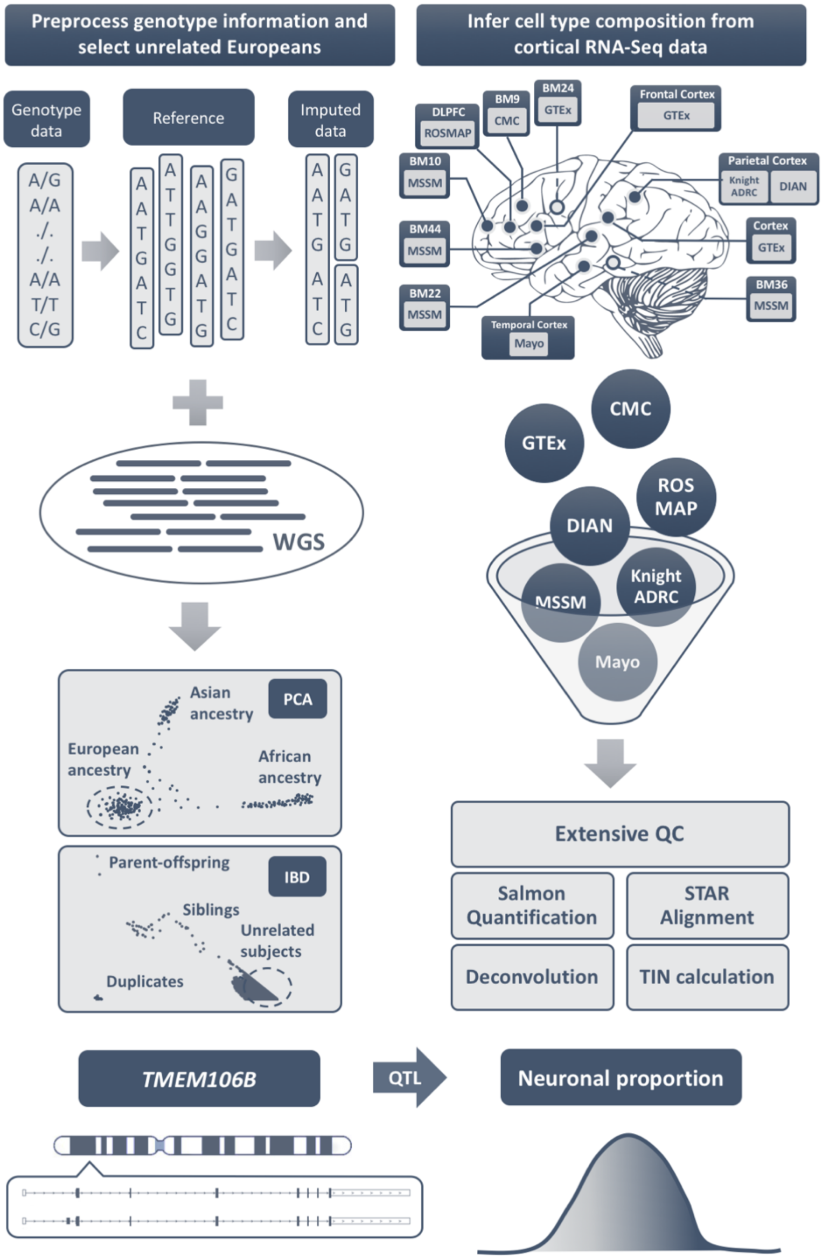


Figure 3.4 Study design. RNA-Seq and paired genotype or WGS data were accessed and preprocessed for downstream analysis. Genotype data was censored based on my quality control criteria and imputed as needed. WGS and imputed genotype were merged and followed by PCA and IBD procedures to select unrelated European ancestry subjects. RNA-Seq data was quality checked with FastQC and aligned to human GRCh37 primary assembly with Star, from which TIN was inferred with RSeQC to account for RNA integrity variances that I later incorporated into the analysis. Gene expression were quantified from unaligned RNA-Seq with psedo-aligner Salmon for deconvolution procedure. Cell type composition comprised of four major CNS cell type proportions were inferred by performing deconvolution procedure on gene expression quantification results. Using cell type proportions as quantitative traits, I identified loci in *TMEM106B* gene region associated with neuronal proportion in my assembled dataset.

3.4.2 *TMEM106B* variants associated with neuronal proportion

During discovery phase, ROSMAP dataset (N = 484 after removing outliers from total number of 523 subjects) was used to perform cell type proportion QTL analysis. Using normalized neuronal proportion as a quantitative trait, the QTL analysis identified more than 10 peaks that passed genome wide suggestive threshold ($<1.0\times10^{-05}$, **Figure 3.5AB, Table 3.3**). However, only one signal rs1990621 (chr7: 12283873) were replicated with p-value = 7.41×10^{-04} in the replication dataset (N = 1,052) combining all the other datasets except ROSMAP (**Figure 3.5CD**). When the discovery and replication datasets were merged to attained a larger sample size (N = 1,536), rs1990621 major allele C is negatively associated with neuronal proportion with p-value = 9.42×10^{-09} (**Figure 3.6AB, Figure 3.7AC**), which means the minor allele G is associated with increased neuronal proportion in my assembled datasets focusing on neurodegenerative disorders.

Table 3.3 rs1990621 (chr7:12283873) major allele C is significantly associated with decreased neuronal proportions. Therefore, G allele (MAF = 0.4658) is significantly associated with increased neuronal proportions.

Dataset	Brain Region	Ref Allele	Sample Size	Beta	SE	P-value
Discovery	DLPFC	C	484	-0.3	0.06	6.40×10^{-07}
Replication	Multiple	C	1,052	-0.13	0.04	7.41×10^{-04}
Merged meta-tissue	Multiple	C	1,536	-0.16	0.05	9.42×10^{-09}

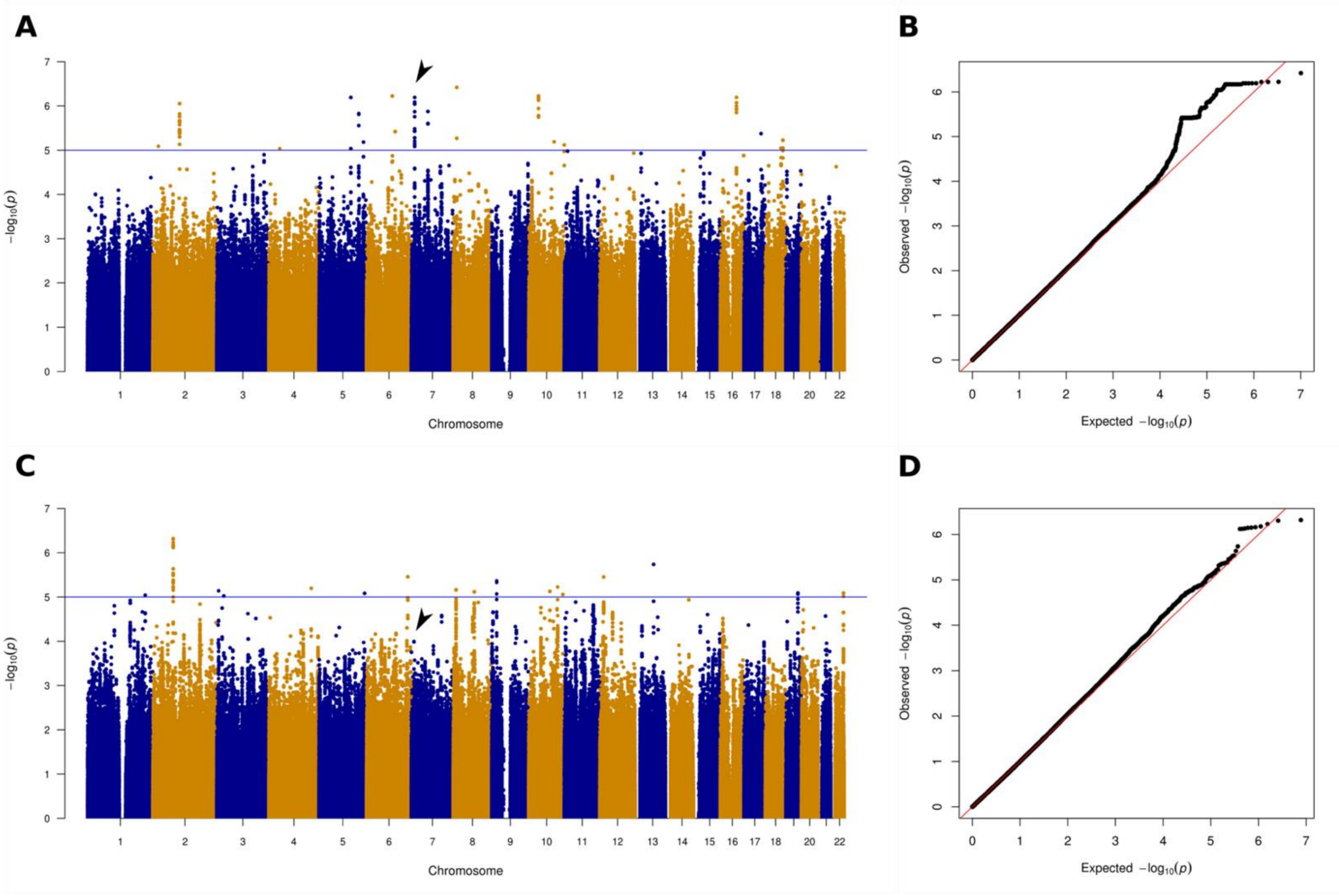


Figure 3.5 Discovery and replication phases Manhattan and QQ plots. Loci located in chromosome 7 were associated with neuronal proportion in ROSMAP discovery dataset and replicated in replication dataset. A) Discovery set Manhattan plot showed seven peaks associated with neuronal proportion at suggestive threshold. The peak located in chromosome 7 was labeled, which is for rs1990621 with p-value = 6.4×10^{-07}. B) QQ plot of the discovery phase analysis. C) Replication set Manhattan plot showed that the peak located in chromosome 7 replicated the signal identified during discovery phase with p-value = 7.41×10^{-04}. D) QQ plot of the replication phase analysis.

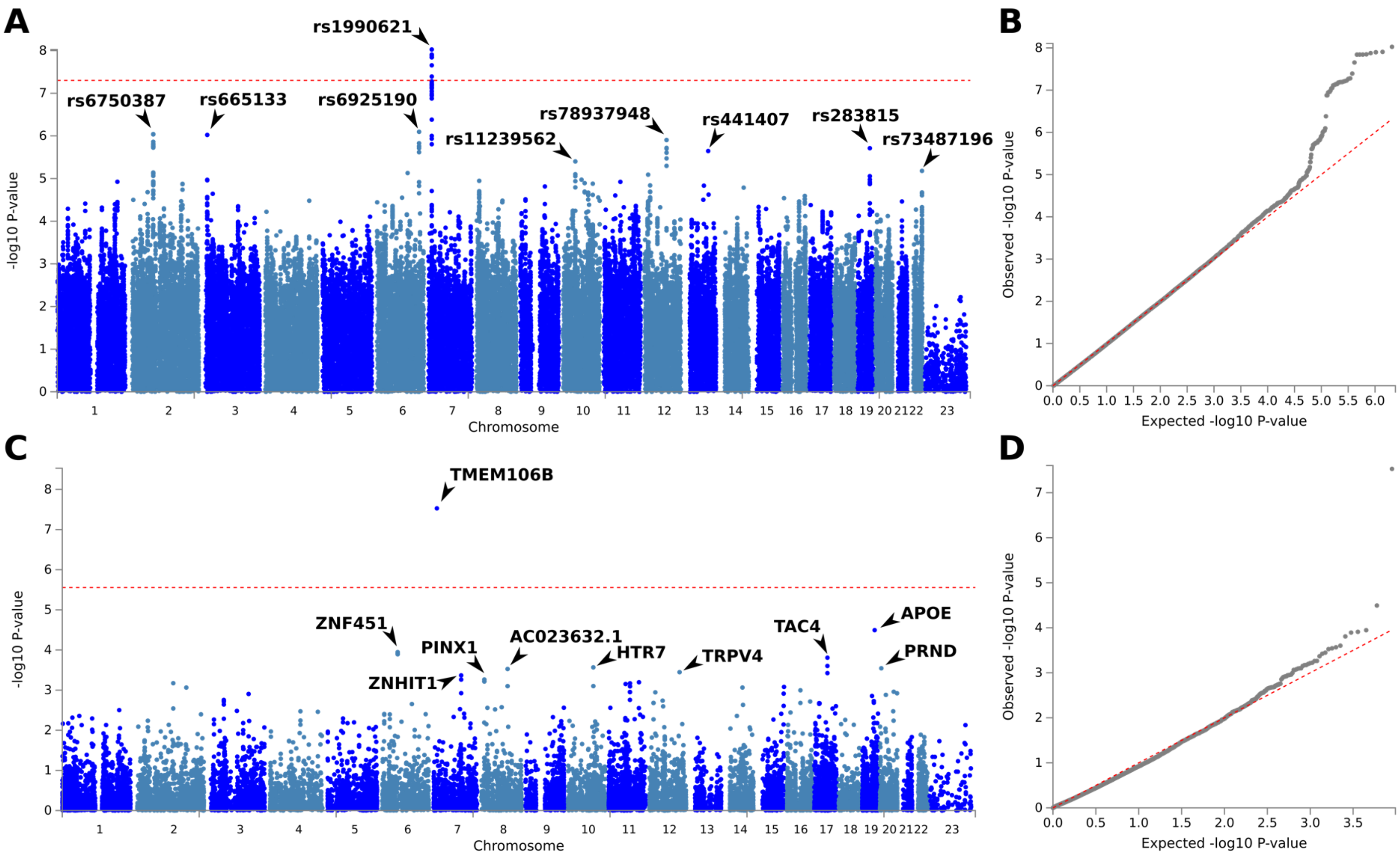


Figure 3.6 SNP-based and gene-based meta-analysis. rs1990621 located in chromosome 7 *TMEM106B* gene region was significantly associated with neuronal proportion in cortical RNA-Seq dataset. A) Manhattan plot showed SNP-based genome-wide significant hit located in chromosome 7 with other suggestive SNP hits labeled. B) QQ plot of the SNP-based analysis. C) Manhattan plot showed gene-based genome-wide significant hit located in chromosome 7 with other suggestive gene hits labeled. D) QQ plot of the gene-based analysis.

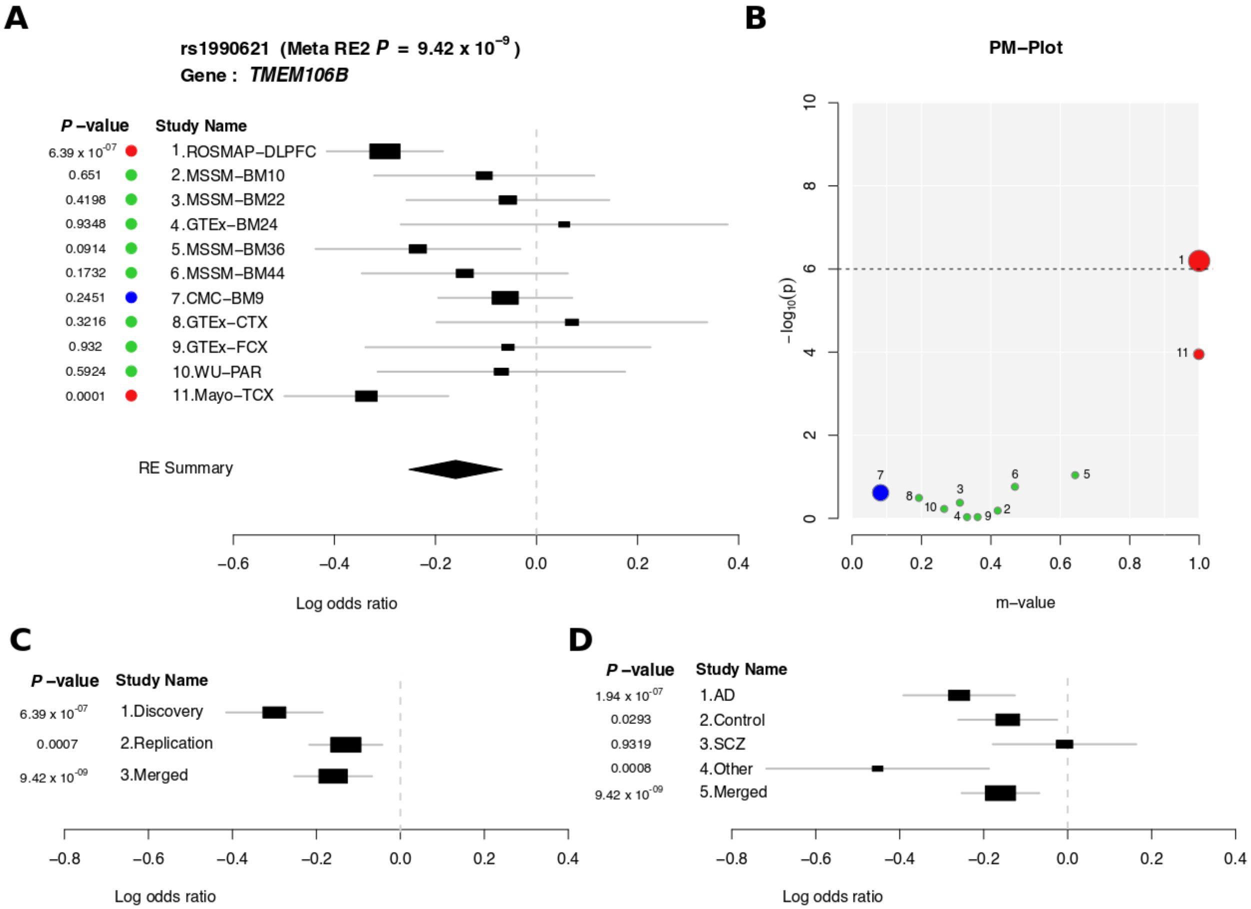


Figure 3.7 Meta-Tissue analysis results of rs1990621. A) Forest plot showed p-value and confidence interval for rs1990621 for each tissue site of each dataset that included in the Meta-Tissue analysis. Summary random effect was depicted at the bottom as RE Summary. B) PM-Plot of rs1990621 while combining both p-value (y axis) and m-value (x axis). Red dot indicates that the variant is predicted to have an effect in that particular dataset, blue dot means that the variant is predicted to not have an effect, and green dot represents ambiguous prediction. C) Forest plot p-value and confidence interval for rs1990621 for discovery, replication, and merged datasets. D) Forest plot p-value and confidence interval for rs1990621 when splitting the merged dataset into four main disease categories.

Noticeably, in both replication and merged analyses, multi-tissue data were involved that provided additional power but also posed challenges to the analysis, the same issue faced by the GTEx study[24, 55]. Compared to a tissue-by-tissue approach, multiple tissues collected from the same subject may help identify QTL by aggregating evidence from multiple tissues, which is similar to a meta-analysis of combining each study. However, one violation of such approach is that the tissues collected from the same subject are presumably highly correlated since they shared the same genetic architecture. Thus, it violates the assumption of independency for carrying out a standard meta-analysis[253]. Another challenge of the multi-tissue QTL is the heterogeneity of the effects, which means a variant may have different effects on different tissues. To resolve these issues, I applied the Meta-Tissue analytic pipeline[253] (http://genetics.cs.ucla.edu/metatissue/) specifically designed for multi-tissue QTL, the same approach that GTEx took to analyze their multi-tissue data. As shown in **Figure 3.7A**, Meta-Tissue analysis results of rs1990621 for the merged analysis were displayed as a forest plot with 95% confidence interval and p-value labeled for each tissue of each study. Among them, MSSM and GTEx are multi-tissue studies while the others are single-tissue studies. Meta-Tissue used a linear mixed model to capture the multi-tissue correlation within MSSM and GTEx respectively. Regarding the effect heterogeneity, Meta-Tissue calculated a m-value[117] to predict if a variant has an effect in a tissue. M-value is similar to the posterior probability of association based on the Bayes factor[117] but with differences specifically designed for detecting whether an effect is present in a study included in a meta-analysis. **Figure 3.7B** is a PM-Plot that integrates evidences from both frequentist (p-value) and Bayesian (m-value) sides to interpret the heterogeneity of multi-tissue QTL effects. Variant rs1990621 in ROSMAP and Mayo studies have m-values greater than 0.9, are predicted to have an effect and color coded with red. In CMC study, the m-value is less than 0.1, so it is predicted to not to have an effect and color coded with blue. All the other studies with m-value between 0.1 and 0.9 are predicted with ambiguous effect and color coded with green. Based on the forest plot and PM-Plot, the variant does have effect heterogeneity across different tissues and studies. In this case, random-effect model will be more suitable to account for effect heterogeneity. Therefore, summary random effect and p-value were reported for the analysis.

Apart from multi-tissue QTL, a single-tissue joint analysis was also performed. In this case, one tissue region was drawn from the multi-tissue data to avoid violating the independency assumption. Specifically, BM36 and frontal cortex tissue were selected to represent MSSM and GTEx study respectively. Study sites were coded as dummy variables to account for potential batch effects. In this joint analysis, the variant rs1990621 is also the top hit with p-value = 7.66×10^{-10}.

3.4.3 Neuronal protective effect of *TMEM106B* variants observed in neurodegenerative disorders and normal aging participants

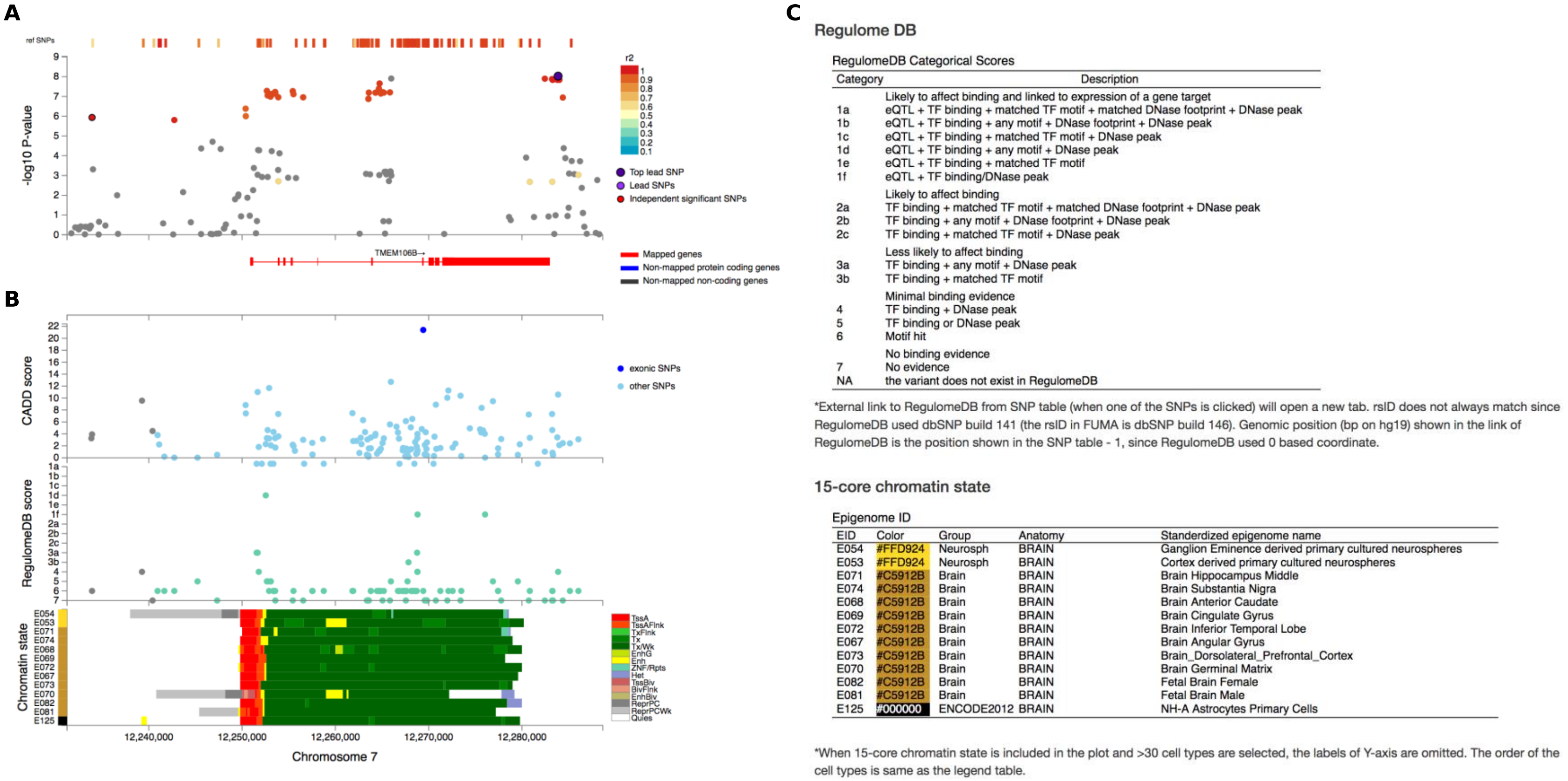
To explore the effect in different disease categories, the merged dataset was stratified based on disease: AD, other non-AD neurodegenerative disorders, schizophrenia and control. Signification associations between rs1990621 and neuronal proportion were observed in AD (p-value = 1.95×10^{-07}), other non-AD neurodegenerative (p-value = 8.19×10^{-04}), and cognitive normal control (p-value = 2.94×10^{-02}) cohorts, but not in schizophrenic cohort (p-value = 9.32×10^{-01}, **Table 3.4**, **Figure 3.7D**). The effect of the variant was more prominent in neurodegenerative cohorts and aging controls with mean age of death greater than 65 years old. However, it was absent from younger cohorts such as GTEx controls and CommonMind schizophrenia participants. Thus, this variant seems to be associated with a neuronal protection mechanism shared by any aging process in the present or absence of neuropathology.

Table 3.4 rs1990621 (chr7:12283873) major allele C is significantly associated with decreased neuronal proportions in AD, Control, and other non-AD neurodegenerative disorders. SCZ: schizophrenia; other: other non-AD neurodegenerative disorders, including progressive supranuclear palsy and pathological aging. BM9: dorsal lateral prefrontal cortex. TCX: temporal cortex.

Disease	Brain Region	Ref Allele	Sample Size	Beta	SE	P-value
AD	Multiple	C	639	-0.26	0.07	1.95×10^{-07}
Control	Multiple	C	476	-0.14	0.06	2.94×10^{-02}
SCZ	BM9	C	189	-0.01	0.09	9.32×10^{-01}
Other	TCX	C	103	-0.45	0.14	8.19×10^{-04}

3.4.4 Functional annotation of rs1990621

The variant rs1990621 is located in the *TMEM106B* gene region where other variants in high LD linkage are also located and labeled in **Figure 3.8A**. Although the CADD score and RegulomeDB score for this variant are not remarkably high to suggest any functional consequences (**Figure 3.8BC**), this variant is in high LD with rs1990622 (r^2 = 0.98), a *TMEM106B* variant previous identified to be associated with FTD risk[266], particularly in granulin precursor (*GRN*) mutation carriers[57, 92]. *TMEM106B* is expressed in neurons and microglia, with highest protein expression detected in the late endosome/lysosome compartments of neurons[36, 163, 237, 248]. A nonsynonymous variant rs3173615, which is also in high LD with rs1990621 (r^2 = 0.98), located in the exon 6 of *TMEM106B* (the dark blue dot in **Figure 3.8B**) produces two protein isoforms (p.T185S) that affect TMEM106B protein level through protein degradation mechanism[36, 49, 200].



Regulome DB

RegulomeDB Categorical Scores

Category	Description
	Likely to affect binding and linked to expression of a gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit
	No binding evidence
7	No evidence
NA	the variant does not exist in RegulomeDB

*External link to RegulomeDB from SNP table (when one of the SNPs is clicked) will open a new tab. rsID does not always match since RegulomeDB used dbSNP build 141 (the rsID in FUMA is dbSNP build 146). Genomic position (bp on hg19) shown in the link of RegulomeDB is the position shown in the SNP table - 1, since RegulomeDB used 0 based coordinate.

15-core chromatin state

Epigenome ID

EID	Color	Group	Anatomy	Standerdized epigenome name
E054	#FFD924	Neurosph	BRAIN	Ganglion Eminence derived primary cultured neurospheres
E053	#FFD924	Neurosph	BRAIN	Cortex derived primary cultured neurospheres
E071	#C5912B	Brain	BRAIN	Brain Hippocampus Middle
E074	#C5912B	Brain	BRAIN	Brain Substantia Nigra
E068	#C5912B	Brain	BRAIN	Brain Anterior Caudate
E069	#C5912B	Brain	BRAIN	Brain Cingulate Gyrus
E072	#C5912B	Brain	BRAIN	Brain Inferior Temporal Lobe
E067	#C5912B	Brain	BRAIN	Brain Angular Gyrus
E073	#C5912B	Brain	BRAIN	Brain_Dorsolateral_Prefrontal_Cortex
E070	#C5912B	Brain	BRAIN	Brain Germinal Matrix
E082	#C5912B	Brain	BRAIN	Fetal Brain Female
E081	#C5912B	Brain	BRAIN	Fetal Brain Male
E125	#000000	ENCODE2012	BRAIN	NH-A Astrocytes Primary Cells

*When 15-core chromatin state is included in the plot and >30 cell types are selected, the labels of Y-axis are omitted. The order of the cell types is same as the legend table.

Figure 3.8 Variant rs1990621 functional annotation and local plot. A) Local plot showed the zoom-in view of the hit in chromosome 7 with the top lead SNP rs1990621 labeled with dark purple. Nearby SNPs were also mainly located in the *TMEM106B* gene region and color coded with LD r2 thresholds. B) Bottom panel showed combined CADD score, RegulomeDB score, and Chromatin state of the region shown in the top panel. C) Regulome DB and chromatin state explanation.

3.4.5 The impact of other neurodegenerative risk loci on neuronal proportion

To investigate what other AD or FTD variants might have an effect in neuronal proportion QTL analysis, I extracted results for 38 SNPs examined in two large scale genome wide association studies, AD focused (Lambert et al.[161]) and FTD focused (Ferrari et al.[87]) studies. Among those, only variants located in *TMEM106B* and *APOE* gene regions passed genome wide significant or suggestive threshold. Both rs1990622 (**Figure 3.9A**) and rs2075650 (**Figure 3.9B**) were found to be associated with FTD reported in Ferrari et al., which were associated with neuronal proportion in this study (**Table 3.5**). The top signals in *APOE* region are rs283815, rs769449, and rs429358 with p-value < 1.22×10^{-05}. Note that rs429358 is one of the two SNPs that determine *APOE* isoforms. Remember that *APOE* ε4 alleles, coded by rs429358(C) and rs7412(C), confers the largest effect for AD risk. I observed that the C allele of rs429358 was associated with decreased neuronal proportion, but no association observed between rs7412 and neuronal proportion.

In a gene-based analysis of my neuronal proportion QTL, *TMEM106B* (p-value = 2.96×10^{-08}) is the only gene that passed genome-wide significant threshold followed by *APOE* (p-value = 3.2×10^{-05}), the most important gene for sporadic AD risk (**Figure 3.6CD**). Previous GWAS for AD risk performed with the International Genomics of Alzheimer's Project (IGAP) data stratified by *APOE* genotype showed that AD risk is significantly influenced by the interaction between *APOE* and *TMEM106B*[146]. Together with my observation of cellular composition QTL, these results suggest a potential interaction of *TMEM106B* and *APOE* may play a role in affecting AD risk/vulnerability and cellular composition balance between neurons and astrocytes, and the endosome and lysosome compartments might be the location that the interaction takes place.

Table 3.5 Neuronal proportion cQTL p-values were reported for variants previously identified in AD risk (by Lambert et al.), FTD risk (by Ferrari et al.), and FTD-TDP risk (by Van Deerlin et al.) studies.

								cQTL		**AD risk**		**FTD risk**		**FTD-TDP risk**	
SNP	**CHR**	**BP**	**Gene**	**Minor**	**Major**	**MAF**	**SNP proxy**	**Effect (Major)**	**p-value**	**OR (Minor)**	**p-value**	**OR (Minor)**	**p-value**	**OR (Minor)**	**p-value**
rs6656401	1	207,692,049	*CR1*	A	G	0.1573	-	0.09	5.59×10^{-02}	1.18	5.7×10^{-24}	-	-	-	-
rs730482	2	127,894,484	*BIN1*	T	A	0.3265	rs6733839	0.03	3.73×10^{-01}	1.22	6.9×10^{-44}	-	-	-	-
rs35349669	2	234,068,476	*INPP5D*	T	C	0.4256	-	2.33×10^{-03}	9.80×10^{-01}	1.08	3.2×10^{-08}	-	-	-	-
rs190982	5	88,223,420	*MEF2C*	G	A	0.3531	-	-3.08×10^{-04}	5.88×10^{-01}	0.93	3.2×10^{-08}	-	-	-	-
rs1980493	6	32,363,215	*BTNL2*	G	A	0.1443	-	0.02	7.83×10^{-01}	-	-	0.775	1.57×10^{-08}	-	-
rs3129871	6	32,406,342	*HLA-DRA*	A	C	0.3736	-	0.04	3.72×10^{-01}	-	-	0.961	3.43×10^{-01}	-	-
rs3129882	6	32,409,530	*HLA-DRA*	G	A	0.4353	-	0.06	1.04×10^{-01}	-	-	1.086	3.36×10^{-02}	-	-
rs9268856	6	32,429,719	*DRB5*	A	C	0.27	-	0.02	8.12×10^{-01}	-	-	0.809	5.51×10^{-09}	-	-
rs9268877	6	32,431,147	*DRB5*	A	G	0.4268	-	0.05	1.91×10^{-01}	-	-	1.204	1.05×10^{-08}	-	-
rs9271192	6	32,578,530	*HLA-DRB5–HLA-DRB1*	C	A	0.2677	-	0.04	4.31×10^{-01}	1.11	2.9×10^{-12}	-	-	-	-
rs10948363	6	47,487,762	*CD2AP*	G	A	0.2488	-	0.01	8.62×10^{-01}	1.1	5.2×10^{-11}	-	-	-	-
rs1020004	7	12,255,778	*TMEM106B*	G	A	0.315	-	-0.1	1.35×10^{-03}	-	-	1.03	4.59×10^{-01}	0.6	5.00×10^{-11}
rs6966915	7	12,265,988	*TMEM106B*	A	G	0.4609	-	-0.16	1.24×10^{-08}	-	-	1.07	1.21×10^{-01}	0.61	1.63×10^{-11}
rs1990622	7	12,283,787	*TMEM106B*	G	A	0.4673	-	-0.16	1.44×10^{-08}	-	-	1.08	7.88×10^{-02}	0.61	1.08×10^{-11}
rs2718058	7	37,841,534	*NME8*	G	A	0.3861	-	-3.76×10^{-04}	9.99×10^{-01}	0.93	4.8×10^{-09}	-	-	-	-
rs1476679	7	100,004,446	*ZCWPW1*	G	A	0.2655	-	0.02	7.71×10^{-01}	0.91	5.6×10^{-10}	-	-	-	-
rs11771145	7	143,110,762	*EPHA1*	A	G	0.3708	-	0	9.67×10^{-01}	0.9	1.1×10^{-13}	-	-	-	-
rs28834970	8	27,195,121	*PTK2B*	C	T	0.3318	-	0.07	6.38×10^{-02}	1.1	7.4×10^{-14}	-	-	-	-
rs1532277	8	27,466,181	*CLU*	T	C	0.3618	rs9331896	-1.42×10^{-03}	9.92×10^{-01}	0.86	2.8×10^{-25}	-	-	-	-
rs3849942	9	27,543,281	*C9orf72/MOB3B*	T	C	0.2253	-	-0.01	8.94×10^{-01}	-	-	1.166	4.38×10^{-04}	-	-
rs10838725	11	47,557,871	*CELF1*	C	T	0.2748	-	0.07	8.60×10^{-02}	1.08	1.1×10^{-08}	-	-	-	-
rs7124974	11	59,906,972	*MS4A6A*	T	G	0.3645	rs983392	0.05	2.05×10^{-01}	0.9	6.1×10^{-16}	-	-	-	-
rs10792832	11	85,867,875	*PICALM*	A	G	0.34	-	-0.01	4.50×10^{-01}	0.87	9.3×10^{-26}	-	-	-	-
rs2380093	11	87,803,455	*RAB38*	T	C	0.1323	rs1386330	-0.03	4.87×10^{-01}	-	-	1.05	3.35×10^{-01}	-	-
rs10128715	11	87,872,076	*RAB38/CTSC*	A	G	0.1217	rs74977128	-0.03	7.29×10^{-01}	-	-	1.815	3.06×10^{-08}	-	-
rs302668	11	87,876,911	*RAB38*	C	T	0.3077	-	0.08	4.57×10^{-02}	-	-	0.814	2.44×10^{-07}	-	-
rs302665	11	87,879,627	*RAB38*	G	A	0.2464	rs302652	0.07	3.65×10^{-02}	-	-	0.73	2.02×10^{-08}	-	-
rs7106306	11	87,929,167	*RAB38/CTSC*	G	C	0.1146	rs16913634	-0.07	1.85×10^{-01}	-	-	0.964	0.71	-	-
rs11218343	11	121,435,587	*SORL1*	C	T	0.04487	-	-0.12	1.35×10^{-01}	0.77	9.7×10^{-15}	-	-	-	-
rs17125944	14	53,400,629	*FERMT2*	G	A	0.08421	-	0.1	1.71×10^{-01}	1.14	7.9×10^{-09}	-	-	-	-
rs10498633	14	92,926,952	*SLC24A4-RIN3*	A	C	0.2128	-	-0.01	8.32×10^{-01}	0.91	5.5×10^{-09}	-	-	-	-
rs242557	17	44,019,712	*MAPT*	A	G	0.3558	-	-0.01	8.89×10^{-01}	-	-	0.853	4.82×10^{-03}	-	-
rs8070723	17	44,081,064	*MAPT*	G	A	0.2077	-	-0.02	7.48×10^{-01}	-	-	1.201	2.80×10^{-04}	-	-
rs1460595	18	29,045,257	*DSG2*	A	G	0.03763	rs8093731	0.05	3.81×10^{-01}	0.73	1.0×10^{-04}	-	-	-	-
rs4147929	19	1,063,443	*ABCA7*	A	G	0.159	-	-0.04	4.78×10^{-01}	1.15	1.1×10^{-15}	-	-	-	-
rs2075650	19	45,395,619	*TOMM40/APOE*	G	A	0.1407	-	0.2	2.11×10^{-04}	-	-	1.304	8.81×10^{-07}	-	-
rs3865444	19	51,727,962	*CD33*	A	C	0.2871	-	-0.05	2.02×10^{-01}	0.94	3.0×10^{-06}	-	-	-	-
rs7274581	20	55,018,260	*CASS4*	G	A	0.1045	-	-0.02	3.01×10^{-01}	0.88	2.5×10^{-08}	-	-	-	-

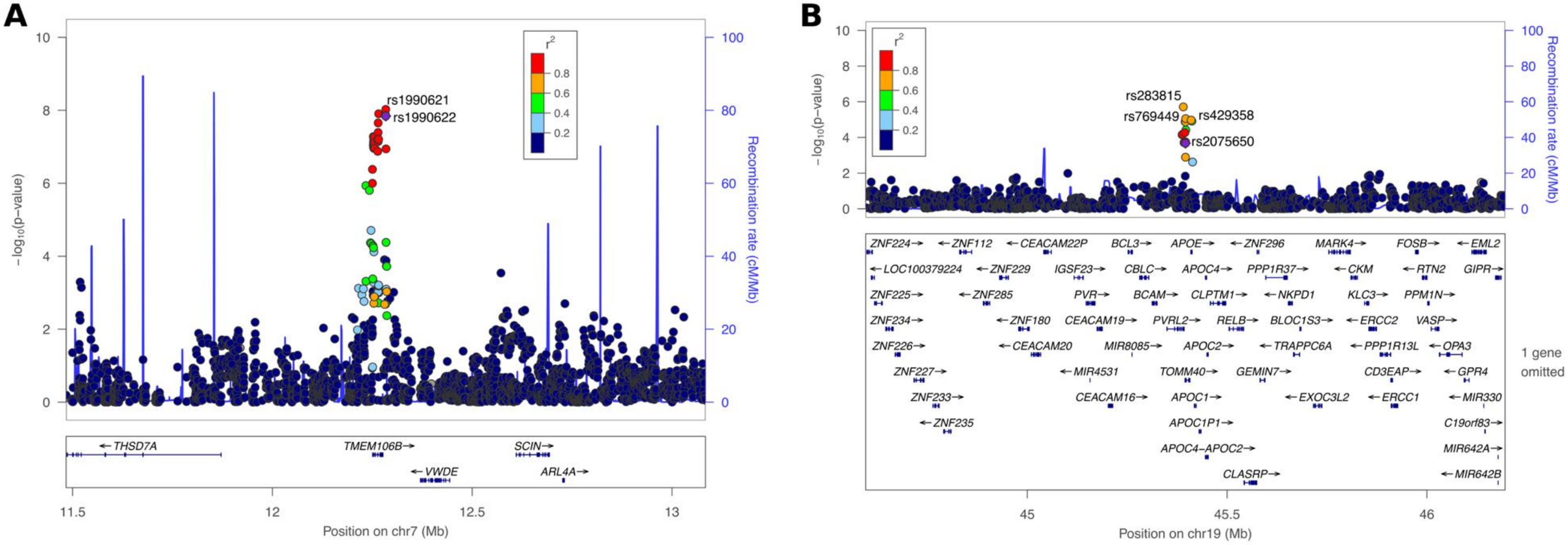


Figure 3.9 *TMEM106B* and *TOMM40/APOE* regions local plot. A) Local plot showed the zoom-in view of the hit in chromosome 7 with target SNP rs1990622 labeled with dark purple, and the top leading SNP is rs1990621. Nearby SNPs were also mainly located in the *TMEM106B* gene region and color coded with LD r2 thresholds. B) Local plot showed the zoom-in view of the hit in chromosome 19 with target SNP rs2075650 labeled with dark purple, and the top three leading SNPs are rs283815, rs769449, and rs429358. Nearby SNPs were also mainly located in the TOMM40/APOE gene region and color coded with LD r2 thresholds. One gene omitted in this region is SNRPD2.

3.5 Discussions

The common variant rs1990622 in *TMEM106B* was first identified to be associated with FTD with TDP-43 inclusions[266]. Hyper-phosphorylated and ubiquitinated TDP-43 is the major pathological protein for FTD and ALS[198], which is also present in a broader range of neurodegenerative disorders, including AD[10], Lewy body disease[196], and hippocampal sclerosis[10]. Recent study also suggested distinct TDP-43 types present in non-FTD brains, typical TDP-43 α-type and newly characterized β-type[145]. TDP-43 α-type is the typical form conventionally observed in temporal, frontal and brainstem regions. TDP-43 β-type is characterized by its close adjacency to neurofibrillary tangles, which is predominantly observed in limbic system, including amygdala, entorhinal cortex, and subiculum of the hippocampus. These findings suggested that pathologic TDP-43 protein that closely associated with *TMEM106B* variants might be the common pathologic substrate linking these neurodegenerative disorders. Multiple lines of evidence have merged and shown that protective variants in *TMEM106B* are associated with attenuated cognitive deficits or better cognitive performance in ALS[268], hippocampal sclerosis[191], presymptomatic FTD[214], and aging groups with various neuropathological burden[280] or in the absence of known brain disease[222]. My study identified a protective variant rs1990621 of *TMEM106B* is associated with increased neuronal proportion in participants with neurodegenerative disorders and normal aging in non-demented controls. However, this effect is not observed in a younger schizophrenia cohort with a mean age of death less than 65 years old. This result suggested a common pathway involving *TMEM106B* shared by aging groups in the present or absence of neurodegenerative pathology that may contribute to cognitive preservation and neuronal protection.

My study has demonstrated that a protective variant rs1990621 identified in *TMEM106B* gene region may exert neuronal protection function in aging groups. A protein coding variant rs3173615 in high LD with rs1990621 (r^2 = 0.98) produces two protein isoforms (p.T185S). The S185 allele is protective and the protein carrying this amino acid is degraded faster than the risk variant T185. Thus, the risk allele of this coding variant leads to increased TMEM106B protein level[36, 49, 200]. *TMEM106B* overexpression results in enlarged lysosomes and lysosomal dysfunction[36, 295]. It has also been shown that TMEM106B may interact with PGRN (the precursor protein for granulin) in lysosome[200]. Although rs3173615 is not included in my genomic data, it is in complete linkage disequilibrium with rs1990621 and rs1990622. It is worth pointing out that the minor allele of rs1990622, which has a protective effect in FTD, is in-phase with the minor allele of rs1990621, which is associated with increased neuronal proportion in my analysis. Despite the fact that my dataset is focused on neurodegeneration, I only have 11 verified FTD cases suggesting that TMEM106B might have a general neuronal protection role in neurodegeneration apart from FTD.

This observation suggested that a potential involvement of *TMEM106B* in the endosome/lysosome pathway may play a role in neurodegenerative disorder risk or vulnerability. Neuronal survival requires continuous lysosomal turnover of cellular contents through endocytosis and autophagy[202]. Impaired lysosomal function reduces lysosomal degradative efficiency, which leads to abnormal build-up of toxic components in the cell. Impaired lysosomal system has been found to be associated with a broad range of neurodegenerative disorders, including AD[201], Parkinson disease[12, 233, 278], Huntington disease[90, 271], FTD[167], ALS[90], Niemann-Pick disease type C[154, 204], Creutzfeldt-Jakob disease[166], Charcot-Marie Tooth disease type 2B[243], Neuronal ceroid lipofuscinoses (Batten disease)[155, 156], autosomal dominant hereditary spastic paraplegia[220], Chediak-Higashi syndrome[159], inclusion body myositis[14], and osteopetrosis[141]. Considering the extensive involvements of lysosomal/endosomal compartments in neurodegenerative disorders, it has been proposed that a long and chronic process of abnormal metabolic changes during aging has led to the accumulation of toxic materials[202]. When lifespan increases especially in the sporadic forms of neurodegenerative disorders, failures to degrade these waste products break the proteostasis and the balance maintained by the multicellular interactions, and trigger subsequent chain reactions that lead to neuronal death and outbreaks of various neurodegenerative disorders due to different genetic susceptibilities and other disease etiologies. Although each neurodegenerative disorder has its own characteristic proteopathy, the boundaries of protein pathology distribution are never clear-cut across different disorders. In fact, copathology or nonspecific pathology of proteopathy have been observed in most autopsies of neurodegenerative disorders, such as TDP-43 discussed above, Lewy body, α-synuclein[82], and etc. My observation of lysosomal gene *TMEM106B* associated with neuronal proportion in aging cohorts suggests that the lysosomal pathway might be involved in the common mechanism underlying a broad range of neurodegenerative disorders or aging process in general that contribute to neuronal cell death.

My study has demonstrated the great potential of using cell type composition as quantitative traits to identify QTLs associated with the changes in cell fractions. This approach is more powerful for disorders that involve considerably changes in cellular composition, for example, neurodegenerative disorders, and normal conditions during developmental or aging processes. The development of recent single cell studies will greatly increase the resolution in advancing our knowledge of cellular population changes. More detailed fine mapping of cellular composition from single cell studies together with machine learning algorithms, bulk RNA-Seq deconvolution will be more accurately capturing cellular fraction changes in the samples, such as different types of neurons or different states of astrocytes or microglia. Regarding scalability, this single cell powered bulk deconvolution approach is preferable for carrying out such cell type composition QTL analysis, because due to the high cost of performing single cell studies, bulk RNA-Seq is more financially feasible to scale up, and with larger sample sizes more hidden signals will be unrevealed with increased statistical power.

To conclude, I have identified a protective variant rs1990621 in *TMEM106B* associated with increased neuronal proportion through bulk RNA-Seq deconvolution and cell type proportion QTL analysis. This observation also replicated previous findings of the protective variant rs1990622 in FTD risk, which is in high LD with rs19990621[266]. Besides, I also observed the C allele of rs429358 (codetermine *APOE* ε4 isoform with rs7412 C allele) associated with decreased neuronal proportion as it was hypothesized. It suggested potential involvements of both *APOE* and *TMEM106B* in neuronal protection mechanisms underlying neurodegenerative and normal aging processes, and supported previous observation of interactions between these two genes[146] in AD cohort. I speculate that *TMEM106B* related lysosomal changes might be involved in the common pathway underlying neuronal death and astrocytosis in neurodegenerative disorders and normal aging cohorts. With larger sample size and higher deconvolution resolution, this approach will reveal more biologically relevant and novel loci associated with changes in cellular composition that are important for understanding both disease etiology and healthy aging.

Chapter 4: System biology approaches revealed transcriptomic profiles of *TREM2* and *PSEN1*

4.1 Abstract

Background: Using network analysis approaches, previous studies had revealed several hub gene or pathways that were verified or later identified as key players in disease etiology underlying AD. In sporadic AD, previous studies from the lab have identified *MS4A* gene cluster significantly associated with soluble TREM2 level in CSF. However, from GWAS result it was unclear which *MS4A* gene is the key regulator of *TREM2*. In autosomal dominant AD, mutations in the *APP*, *PSEN1* and *PSEN2* genes and lead to familial early onset AD. However, the downstream pathogenic events triggered by these risk and pathogenic variants are still not fully understood. By employing an integrative network approach, I aim to more accurately identify which gene is the key regulator of *TREM2* in sporadic AD and the downstream genes and pathways altered by *PSEN1* mutation in autosomal dominant AD.

Methods: To determine which one of the *MS4A* genes are implicated in TREM2 biology, I employed alternative approaches to explore gene regulatory networks from RNA-Seq data. To identify causal genes under the genomic locus identified by the CSF TREM2 GWAS, I combined weighted correlation network analysis (WGCNA) method to identify a module that includes *TREM2* co-expressed genes. Then I used Bayesian network inference to learn causality. To study the downstream effects of Mendelian mutations in *PSEN1* associated with Autosomal Dominant Alzheimer's disease, I applied a seed-based approach to study the genes that are significantly co-expressed with *PSEN1,* and constructed gene networks using WGCNA. This analysis includes both *PSEN1* mutation carriers, non-carriers and nenuropathological-free controls. The network was annotated with gene differential expression, cell type information, and functional pathway analysis.

Results: My analysis indicated that *MS4A4A* and *MS4A6A are* in TREM2 module, and inferred that *MS4A4A* is the key regulator of *TREM2*. For the downstream events of the Mendelian gene *PSEN1*, I identified 47 genes only present in control cohort that were potentially disrupted in *PSEN1* mutation carriers; I also found 13 genes only present in *PSEN1* mutation carriers but not in control cohort that are potentially acquired as downstream transcriptional events altered by *PSEN1* mutations. Among them, I highlighted the genes *LMNA*, *DOCK1*, and *DYNC1LI2* and discussed them in detail, that were previously associated with Alzheimer's Disease.

Conclusions: My study demonstrated the potential of using both system-based and seed-based network approaches in replicating and discovering AD related genes and their interactions. In sporadic AD cohort, I identified *MS4A4A* might be a key regulator for *TREM2*. In autosomal dominant AD cohort, I identified total of 60 genes that are lost or acquired in the *PSEN1* associated pathways.

4.2 Introduction

4.2.1 From polygenic to omnigenic - a network interpretation

To tackle disease etiology, one important theme for diseases with genetic components is to figure out how genetic variants explain phenotypic variability. A monogenic theory inspired by Gregor Mendel's work states that one disease could be explained by one mutated gene following a Mendelian inheritance pattern. Examples are familial early onset neurodegenerative disorders such as autosomal dominant AD, which can be explained by rare mutations in one single gene that results in high disease penetrance. However, diseases with complex traits, for example late onset sporadic AD, do not follow this pattern. GWAS performed in sporadic AD studies have identified dozens of variants across the genome, and many of those are common variants with low to medium effects (**Figure 1.2**)[148]. This pattern is more similar to quantitative genetics inspired by Ronald Aylmer Fisher's infinitesimal model that a quantitative trait is influences by an infinitely large number of genes. Accumulations of large number of common variants within multiple genes explain much of the heritability. These polygenic effects together with complex interactions with the environment are often observed in diseases with complex traits. Based on empirical evidences, the polygenic model has later been expanded to an "omnigenic" model for complex traits[31]. They proposed that core genes may have strong and direct effects on disease risk, but they only account for a small portion of total heritability. Any variants that have disease relevant tissue specific effects may contribute nontrivial effects on disease risk. The variants may exert their effects on core genes through highly interconnected gene regulatory network such that the collection of the small effects together explain the missing heritability[31].

4.2.2 Network analysis as a powerful tool

Network biology[21, 22, 269] has demonstrated great success in understanding biological systems and identifying disease related factors. With large scale data collection and advancement in computational powers, various types of networks have been proposed to capture the interactions among different elements (zoom-in view) and to gain a systematic view of the topological and dynamical properties of a biological system (zoom-out view). On a *gene* level, gene regulations can be depicted as gene regulatory networks as mentioned above in the omnigenic model. The modular and hierarchical organization of gene regulatory networks captures the information flow from regulators to their binding sites. On an *RNA* level, microarray and RNA-Seq technologies have enabled generations of transcriptome-wide gene co-expression networks to understand associations among transcripts and gene expression synchronicity. On a *protein* level, yeast two-hybrid screens are able to identify protein-protein interactions *in vivo*[89]. On a *cellular* level, brain neural networks generated from tracer injections have captured the topology of neural signaling highways underlying cognitive functions[85]. On a *system* level, networks generated from functional connectivity MRI facilitate understanding of how different parts of the brain segregated into functional modules and communicating intrinsically without explicit task being performed[213]. On a *disease* level, human disease interactome has been proposed to identify shared pathways among known or unknown comorbidities that shed lights on drug repurposing[50, 183]. On a *population* level, social networks could help identify key features of infectious diseases, such as risks for acquisition and effective interventions, through learning social aspects of disease transmission.

While offering the capability of modeling any type of biological interactions at different levels, network analyses provide a unique and powerful approach that can combine elements or layers from different modalities and produce integrated models to study interactions among multiple networks. One example of an integrative network analysis of combining obesity and social network showed that the social ties among friends have a larger effect on obesity risk than genetic risk factors (**Figure 4.1**)[19, 52].

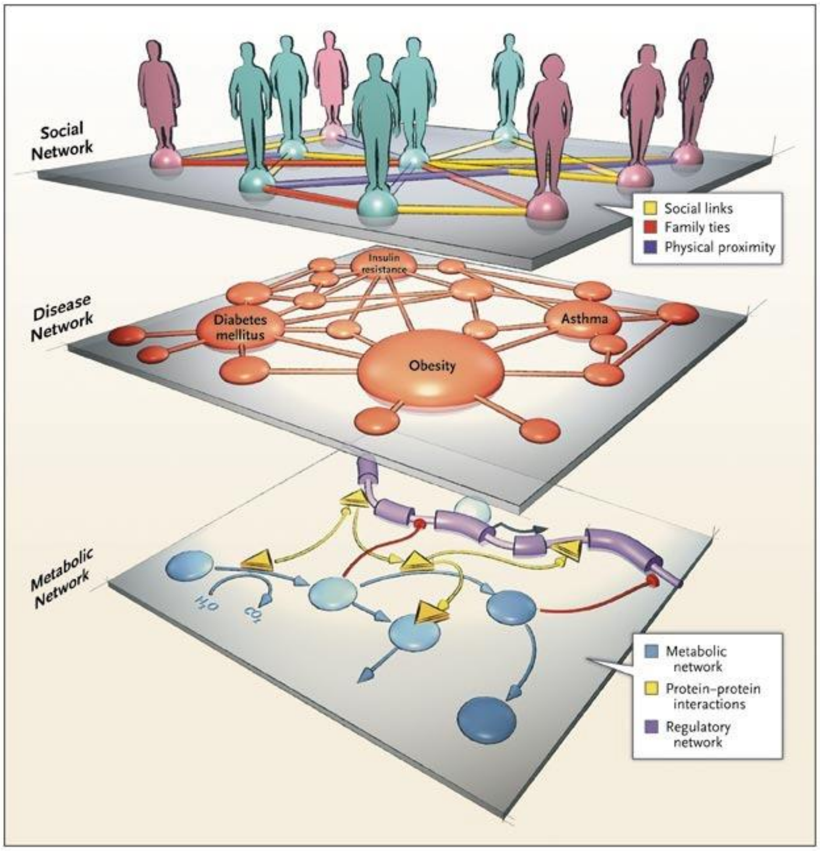


Figure 4.1 Network medicine from obesity to the "diseasome". Reproduced with permission from Barabasi[19]. Copyright Massachusetts Medical Society.

4.2.3 Basic concepts in gene co-expression networks

In biological systems, networks are defined as a collection of biological elements and interactions among them. In the context of a gene co-expression network, nodes are genes and edges are relations between the expression of a pair of genes, such as Pearson's correlation. If the network is unweighted, an edge means the correlation between a gene pair is above a hard cut-off, and there will be no edge if the correlation is below the cutoff. If the network is weighted, the edge represents weighted magnitude of correlation between a gene pair. In an undirected network, such as results from WGCNA, there is no directionality in the connection between two nodes. In a directed network, such as directed acyclic network generated from Bayesian network inference or structural equation modeling (SEM), information flow will be depicted as directed arrow pointing from parent node to child node. This chapter focuses on static and deterministic networks that capture network structural topology from a single RNA-Seq snapshot when RNAs are extracted from the biological system. Dynamic and stochastic networks can be generated with temporal data and data with biological noises. Causal network can also be generated with intervention data.

Regarding the global topology of a network, three common types of networks are showed in **Figure 4.2**. Two important metrics to describe network property are degree distribution and adjacency matrix. Degree is the number of edges a node has with other nodes (denoted as k), and degree distribution is the probability distribution of these degrees over the whole network denoted as p(k). Adjacency matrix is used to represent whether a pair of nodes are connected. It is a matrix with 0 and 1; value 1 represents the nodes are connected and adjacent to each other, and 0 represents that the nodes are unconnected. Undirected graph adjacency matrix is symmetrical and directed graph adjacency matrix is unsymmetrical. In a random network, the degree distribution p(k) follow a Poisson distribution, and connected nodes are randomly distributed in the adjacency matrix. Network path is another important metric, which measures how many steps required to connect two nodes in the network, and minimal number of steps is called the shortest path length. One important property of a random network is that the shortest path length is much smaller than a regular network. Many theoretic and functional works in network science are based on random network, however, it has two limitations – first, it does not have local clustering structure with its randomly distributed connections depicted in adjacency matrix; second, the degree distribution of nodes following a Poisson distribution do not account for the formation of hubs that mostly observed in real world networks. To resolve the first limitation of lacking local clustering, a Watts-Strogatz model has been proposed to generate graphs with small-world property by rewiring edges from a regular lattice network with random probabilities[277]. This random rewiring process created long-range connections with small path length like a random graph while retaining high local clustering properties of a regular network. Thus, the degree distribution of small world network is similar to a random network but with high local clustering property shown in adjacency matrix. Later a scale free model is proposed based on empirical evidence of real world networks including World Wide Web and citation patterns in science[20]. This model explained the hubs observed in many real-world networks that a random network model does not explain. They found that large scale complex networks exhibit a high degree of self-organizing phenomena that networks expand by adding new vertices to already well established highly connected vertices, which explained how hub nodes emerged from chaos. In a scale-free network, the degree distribution of nodes decays as a power law. This feature holds true in any scales of the network – hence the name “scale-free” network. Rigorous modeling is required to examine the power law distribution of empirical data to determine if a network fulfills the criteria[53].

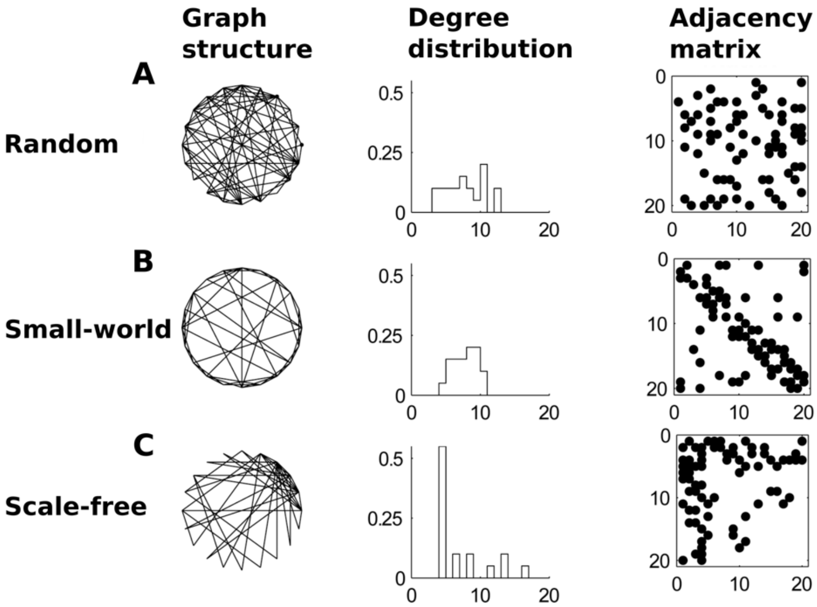


Figure 4.2 Random, small-world, and scale-free network properties. A) random network with 73 connections among 20 nodes assigned randomly; B) Small-world network with high local clustering and short average path lengths with 'hub and spoke' architecture; C) scale-free network with 'hub and spoke' architecture maintained at multiple spatial scales. Image reproduced from Stobb et al.[251] with permission.

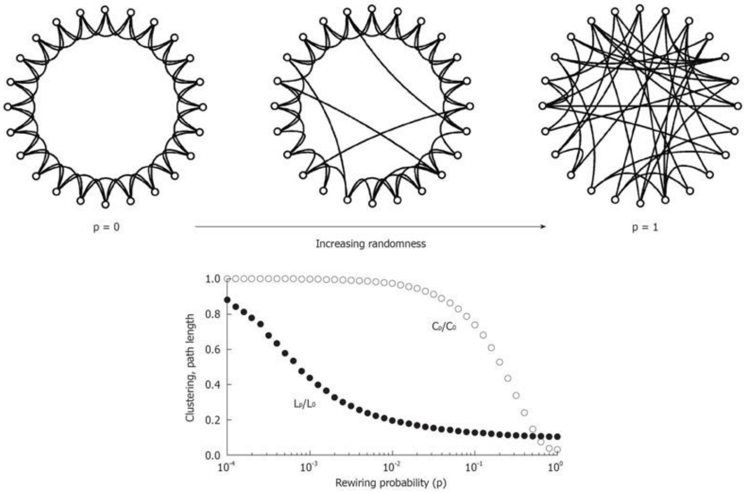


Figure 4.3 Small-world network generated from Watts-Strogatz model. Image from Watts et al.[277] with permission.

4.2.4 Network analysis in Alzheimer's disease

Gene co-expression network analysis in late-onset AD cohorts had revealed several hub gene or pathways that were verified or later identified as key players in disease etiology[184, 287]. For example, a system based approach using WGCNA had been applied to microarray data derived from prefrontal cortex tissues collected from LOAD patients and 173 non-demented healthy controls. They identified an immune response related module, which contains an important AD risk gene *TREM2*. Rare variants in *TREM2* have been found to be associated with sporadic AD risk with moderate effect[114, 144, 148]. The rare *TREM2* variant p.R47H (rs75932628) carriers exhibit increased AD risk by a range from 1.7-fold to 3.4-fold[112, 212]. In the *TREM2* module identified from network analysis, Zhang et al. focused on *TYROBP(DAP12)* gene which was identified as an adapter protein for *TREM2[30]* (**Figure 4.4**).

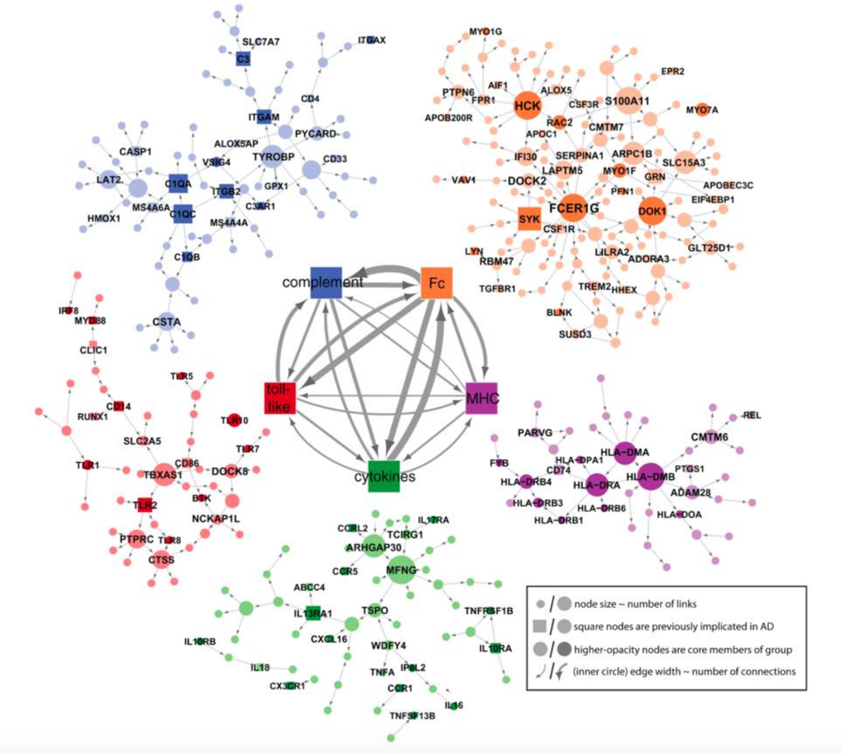


Figure 4.4 TREM2 module identified in LOAD cohorts. A module enriched for immune function and pathways contains *TREM2*, *TYROBP*, and *MS4A* gene clusters. Image from Zhang et al.[287] with permission.

Our lab has recently identified that the *MS4A* gene cluster is a key regulator of soluble TREM2 in CSF[70]. Using GWAS of a large number of subjects (N = 813) we identified a locus in chromosome 7 that shows significant association with CSF TREM2 levels (rs1582763; $p=1.15 \times 10^{-15}$), and replicated in an independent dataset (N = 580). Several members of *MS4A* gene cluster, *MS4A4A* and *MS4A6A* are tagged by this signal, which makes difficult the identification of the causal *MS4A* gene affecting CSF TREM2 levels. I envision that transcriptomic data would provide additional orthogonal evidence. To identify the relationship between TREM2 and MS4A gene cluster, I performed a system-based approach on sporadic AD cohort and focused on the module containing *TREM2* in the first part of this chapter.

In the second part of this chapter, I focused on the study of the transcriptomics downstream analyses of genetic mutations causal of autosome dominant AD (ADAD). Mutations in the *amyloid-beta precursor protein* (*APP*) and presenilin genes (*PSEN1* and *PSEN2*)[43] cause ADAD which is typically associated with Mendelian inheritance pattern and early-onset (30 ~ 50 years old) disease symptoms. Although the disease phenotype may be a consequence of an abnormality in a single effector gene product particularly in the Mendelian form of AD, given the highly-interactive functional crosstalk within biological organism and the complex disease etiology of AD, this dysregulation by a single gene may intertwine with various pathological processes and altered downstream events that interact in a complex network. To investigate the etiological heterogeneity of AD, I performed a seed-based network analysis in an ADAD cohort, carriers of *PSEN1* gene mutation, to identify altered downstream transcriptional events triggered by this ADAD gene. By annotating the network with differential gene expression, cell type information, and functional pathway analysis, I identified some target genes and related pathways that were either disrupted or emerged in PSEN1 mutation carriers through network analysis.

4.3 Methods

4.3.1 Samples

TREM2 Study

I accessed (AMP-AD portal synapse ID = 3157743) RNA-Seq data from 219 AD and non-demented control brains from Mount Sinai School of Medicine (MSSM) ascertained from four cortical regions: anterior prefrontal cortex (APC), inferior frontal gyrus (IFG), superior temporal gyrus (STG), and parahippocampal gyrus (PHG) (**Table 4.1**). Data retrieval and collection of MSSM, Knight-ADRC, and DIAN have been documented in detail in **Chapter 2**.

Table 4.1 *TREM2-MS4A* Study Demographic

	N	Age	% Male	RIN	TIN	Control	AD
MSSM	219	84 ± 7.32	35.6	6.42 ± 1.77	76.4 ± 2.52	49	170

PSEN1 Study

RNA-Seq was generated for 15 *PSEN1* carrier brains from The Dominantly Inherited Alzheimer Network (DIAN) and 14 non-demented controls from The Charles F. and Joanne Knight Alzheimer's Disease Research Center (Knight ADRC)[153]. We identified three additional participants from the Knight-ADRC study with *PSEN1* (p.A79V, p.I143T, p.S170F) mutations (**Table 4.2**).

I accessed (AMP-AD portal synapse ID = 3157743) RNA-Seq data from 67 non-demented control brains from Mount Sinai School of Medicine (MSSM) ascertained from four cortical regions: anterior prefrontal cortex (APC), inferior frontal gyrus (IFG), superior temporal gyrus (STG), and parahippocampal gyrus (PHG).

GEO replication data was accessed from GSE39420, which is collected from 14 patients (7 sporadic EOADs and 7 monogenic familial ADs with *PSEN1* mutation) and 7 neurologically healthy controls. Samples were hybridized in a Human Gene 1.1 microarray from Affymetrix[13].

Table 4.2 *PSEN1* Study Demographic

	N	Age	% Male	% ApoE4+	Control	*PSEN1*	EOAD
MSSM	67	80.1 ± 8.39	44.8	10.4	67	0	0
Knight-ADRC	17	82.3 ± 18.5	35.3	17.6	14	3	0
DIAN	15	49.1 ± 7.14	66.7	14.3	0	15	0
GEO	21	55.6± 7.65	76.2	14.3	7	7	7

4.3.2 Data processing and quality control

TREM2 Study

Data QC and preprocessing of MSSM dataset have been documented in **Chapter 2.3.2**. MSSM cases and controls gene expressions were derived from Star alignment to human GRCh37 primary assembly and quantification using --quantMode TranscriptomeSAM. Because low expressed genes tend to reflect noise and produce insignificant correlation, I removed genes with gene counts less than 4 in more than 75% subjects. To normalize gene expression respect to library size, regularized logarithm transformation was applied to raw counts of the gene expression using rlog function from DESeq2 R package. ComBat function was applied to the data to remove potential batch effect. After which, a linear regression model was applied to regress out covariate on a per-gene basis. Covariates factors included in the model are PC1, PC2, PC3 inferred from genomic PCA analysis to account for ethnic stratification; sex, age at death, post-mortem index, RIN to account for general demographics and RNA-Seq tissue pre-sequencing quality; and RNA-Seq post-sequencing metrics such as ribosomal contents, mapped reads number (uniquely mapped and multi-mappers that mapped to multiple loci) to account for alignment performances. The residuals from the linear regression were used as inputs to infer gene expression correlation. Apart from gene expression values, cognitive performance measurement CDR and Tau pathology load measurement Braak staging values were also added for computing correlation, from which I could infer what genes are closely correlated with these clinical and pathological traits. Besides, cellular composition inferred from the dataset as documented in **Chapter 2** deconvolution procedure were also added to later correlation computing and network construction. Cell type information inferred from the data using deconvolution method were also added as nodes. For each region, the top third most variable genes were selected. A joint set of the top genes from all four regions accounted for a fifth of the whole transcriptome genes. The top fifth most variable genes together with CDR, Braak staging, and cell type proportion were selected to run multi-tissue correlation for later network construction.

PSEN1 Study

Data QC and preprocessing of MSSM, Knight-ADRC, and DIAN have been documented in **Chapter 2.3.2**. For the RNA-Seq data from MSSM gene expression were performed similarly as the TREM2-MS4A study as described above, the only difference is that only non-demented controls from MSSM were used to construct the network to avoid the confounding factors from dramatic neuronal loss or astrocytosis in ADAD as I observed and documented in **Chapter 2**.

For microarray data from GSE39420, probes were mapped to corresponding genes, and each gene expression value was derived from averaging probe expression values for each gene. Each gene expression level was adjusted for subject sex, age at death, and post-mortem interval hours by fitting a linear regression model and the residuals were derived for use in downstream analyses. These covariate adjustment procedures were performed for control, FAD-PSEN1, and EOAD separately.

4.3.3 Repeated measures correlation

I accessed RNA-Seq data MSSM ascertained from four cortical regions: APC, IFG, STG, and PHG. Because there is more than one tissue collected from the same subject, the assumption of independent observations when applying standard correlation methods is violated. To aggregate data collected from multiple tissues in MSSM, I integrated the measures from these four brain regions by running repeated measures correlation tool (rmcorr R package)[18] to calculate repeated measures correlation (rmcorr) of the MSSM controls, which is a statistical technique to determine the overall within-individual relationship among paired measures assessed on two or more occasions.

4.3.4 Network construction

TREM2 study system-wide network construction

With a system-wide approach, Weighted Gene Co-Expression Network Analysis (WGCNA)[164] was applied to the transcriptome rmcorr matrix derived from MSSM AD cases and control subjects. WGCNA enforces the connectivity to exhibit a power-law distribution. This power-law distribution renders a scale-free topology to the network, which will be applied to the top fifth most variable genes across all four regions. Both dynamic tree cutting and static tree cutting with signed networks and minimum module size 200 were used, but it has been shown that the dynamic tree cutting and signed network might be more biologically relevant compared to static tree cut and unsigned network[289]. Raising the correlation to a power will help reduce the noise of the correlations in the adjacency matrix. To select the appropriate power value, pickSoftThreshold function[288] from the WGCNA package was applied to the rmcorr matrix to select the power when the network most resemble a scale-free graph while keeping the highest network connectivity.

TREM2 study Bayesian network construction

To infer probabilistic relationships between the nodes in a network module, Bayesian network analysis was applied to the module derived from the system-wide WGCNA results that contains *TREM2* gene. Normalized and covariate adjusted gene expression matrix from parahippocampal region of MSSM dataset was used, and only genes in the same module with *TREM2* were included in the Bayesian network construction. Using the discretize function from bnlearn R package[193], the continuous gene expression data were transformed into quantiles for later network structure and parameter learning. To infer or measure the degree of confidence of arc strength of Baysian network, 200 nonparametric bootstrap iterations were applied to the data to estimate the relative frequency (strength) of every possible arc[94] using the boot.strength function implemented in bnlearn package. Arcs with strength more than 0.38 and probability of direction more than 0.5 were kept, and final network was derived from averaging across 200 bootstraps. The network was plotted using graphviz.plot function with highlighted v-structure arcs. Genes related to AD risk and genes related to top pathway results (adjusted p-value less than 0.01) were color coded respectively.

PSEN1 study seed-based network construction

Using a seed-based approach, I pre-selected genes that are co-expressed with *PSEN1* to build a co-expression network. The genes that are correlated with *PSEN1* expression are potentially linked to AD pathology by assuming 'guilt-by-association'. In this seed-based approach, I used *PSEN1* gene as a bait to expand to genes that co-expressed/correlated with *PSEN1* in *PSEN1* variant carriers and/or healthy controls.

To build upon the MSSM controls co-expression network, I first selected genes that are significantly correlated with *PSEN1* in the MSSM control rmcorr results at correlation p-value < 0.05. Then the genes from parietal *PSEN1* carriers and healthy control subjects that overlapped with the significant correlated gene in MSSM controls were selected following the two criteria: (1) The gene correlation with gene *PSEN1* in parietal dataset has to fall into the 95% confident intervals of rmcorr results of the MSSM controls; (2) The correlation direction in parietal dataset has to be congruent with the direction of the MSSM controls rmcorr correlation.

4.3.5 Network robustness evaluation

By applying a bootstrapped version of WGCNA (rWGCNA)[100], it will reduce potential bias introduced by outlier samples. I performed 50 iterations of network construction with randomly selected 66.66% of the total samples. The resulting 50 networks will be merged into one large, final consensus network. The robustness of the networks was evaluated by comparing those to the final consensus network.

4.3.6 Network functional annotation

Differential gene expression analysis using DESeq2 R package, and various pathway analysis and gene enrichment analysis were applied to annotation the genes in the network, such as Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology Enrichment Analysis (GO term), and PANTHER implemented in Enrichr online interactive tool[47, 158]. Cell type composition derived from deconvolution, CDR, and Braak staging were labeled in the network to examine cell-type specific enriched modules and modules related to cognitive and pathological measurements. To annotate the genes, the list of interested genes were queried with FUMA's GENE2FUNC online tools[276] to investigate tissue specificity, reported GWAS catalog genes, TF targets, microRNA targets, and etc.

Differential gene expression and correlation analyses were repeated in the independent *PSEN1* dataset (GSE39420) collected from 14 patients (7 EOAD and 7 FAD-PSEN1) and 7 neurologically healthy controls.

4.4 Results

4.4.1 Study design

TREM2 Study

For the TREM2 study, both AD sporadic cases (N = 170) and cognitive normal controls (N = 49) were included to build the network (**Table 4.1**). Tissues from four brain regions (**Figure 4.5**) were collected and followed by RNA-Seq data generation. The same extensive QC and quantification processes were applied as documented in **Chapter 2**. Repeated measure correlation was applied to gene expression quantification results from four regions to derive a gene expression correlation matrix. With a system-based approach, top 1/5 most variable genes across the four regions were used to construct networks with scale free topology using WGCNA. The module containing *TREM2* gene was further analyzed with Bayesian network inference and functionally annotated with pathway analysis.

PSEN1 Study

As discussed in **Chapter 2**, I observed brain carriers of pathogenic mutations in *APP*, *PSEN1* or *PSEN2* presented lower neurons and higher astrocytes relative proportions compared to sporadic AD and controls. With such extensive neurodegeneration, I concerned that the network built with *PSEN1* mutation carriers would be confounded by the destructive consequence of neurodegeneration. To investigate early transcriptional events trigged by pathogenic mutations in *PSEN1*, I built the network with cognitive normal controls and genes correlated with *PSEN1* using a seed-based approach. First, I selected genes that are correlated with *PSEN1* in both cognitive normal controls and *PSEN1* mutation carriers. Then I built a network based on correlations of these selected genes in controls across four different brain regions in MSSM to capture a control network topology. Then nodes were annotated in regard to its correlation with *PSEN1* in either controls or *PSEN1* mutation carriers to infer their functional roles in early disease progression. I hypothesized that network modules enriched with genes correlated with *PSEN1* in mutation carriers could be involved in early transcriptome changes in *PSEN1* mutation carriers.

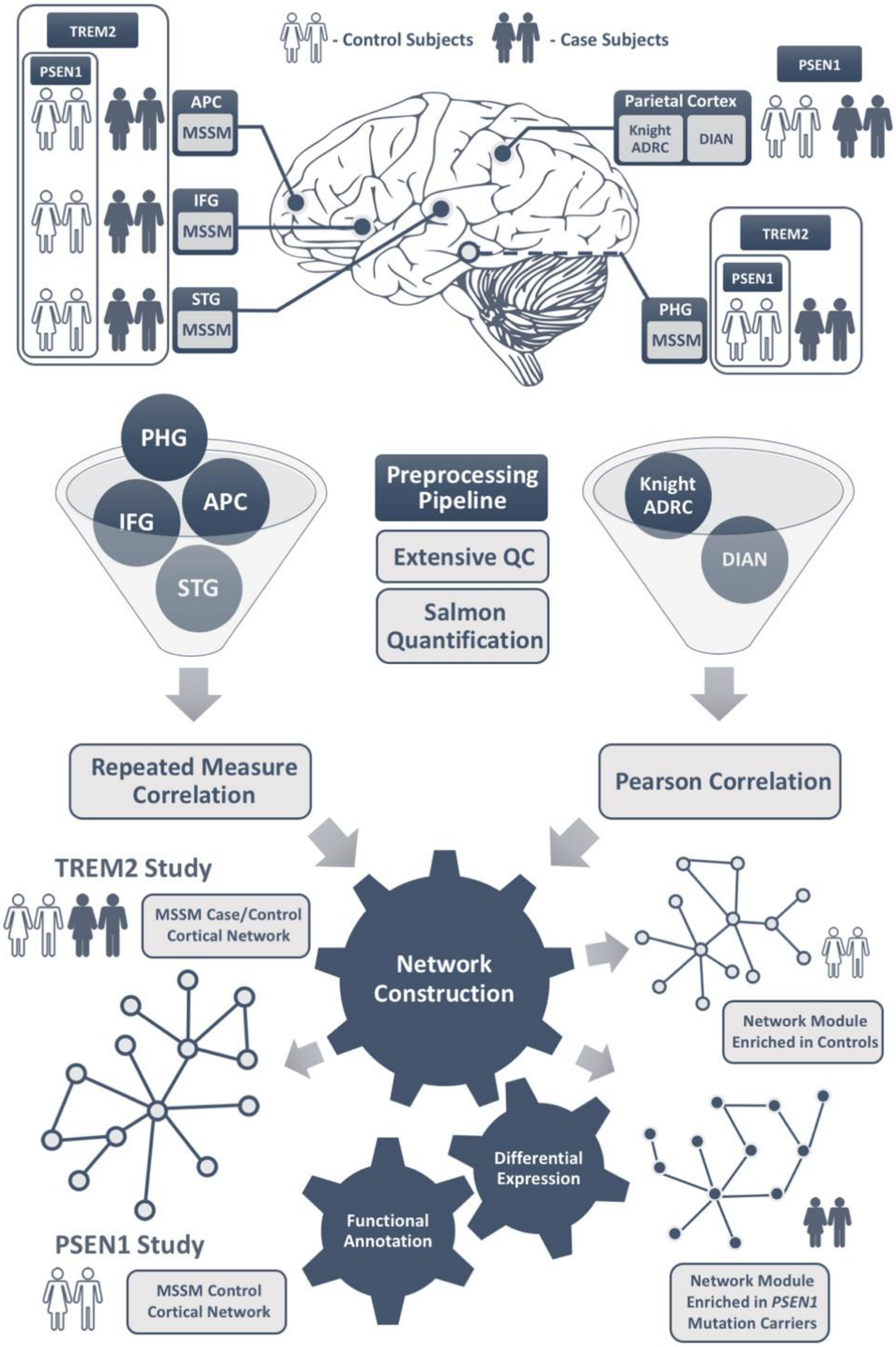


Figure 4.5 Study design. For TREM2 study, network was constructed from MSSM AD sporadic cases and controls. For PSEN1 study, network topology was built from MSSM control participants and functional modules were annotated in related to *PSEN1* variant carriers and healthy controls.

4.4.2 *MS4A* gene clusters are associated with *TREM2* in AD sporadic and control network analysis

TREM2 Study WGCNA network construction

The 20% of most variable genes from four regions of AD sporadic cases and controls were selected to construct the network using WGCNA. The network connection or topology was built upon gene co-expression patterns. Genes that are connected in a co-expression network are correlated or synchronized in their expression pattern[97, 223]. In a gene expression correlation matrix, any pair of genes would yield a non-zero value, which means any gene pair could be correlated to some extent and perfect independency does not exist. Thus, the challenge is how to determine the threshold at which two genes are considered as co-expressed. An intuitive way to solve this problem is to select a value as a hard cut-off. A binary value is assigned to each gene pair; it will be 1 if the gene pair correlation is above the threshold or 0 if it is below the threshold. This binary scenario may occur in certain biological systems such as neuron firing, but it does not fit gene expression patterns. Rather than this hard thresholding approach, the WGCNA researchers proposed a soft thresholding approach that assigns a weight to each gene pair to derive a weighted gene co-expression network. The network also bears a scale-free topology, which has been shown to be more biologically relevant by both theoretical and empirical evidences[288]. The property of scale-free network is defined as the probability of a node that is connected with k other node decays as a power law. This type of network contains a few hubs with high degree of connectivity compared to the vast majority of non-hub nodes with low connectivity. The network is highly resistant to attacks on non-hub nodes, but removing hub nodes will be deleterious or change the network topology dramatically. The hub nodes are essential for survival in biological systems proved by yeast protein network studies[118, 142].

To build a scale free network using WGCNA, taking my study as an example, the first step was to define gene co-expression similarity using repeated measure correlation derived from gene expression of four brain regions. Compared to all the other network analysis that use single region transcriptional data to build a region-specific network, this study integrated data from four brain regions to build a network representing the cerebral cortex. Repeated measure correlation was applied instead of Pearson correlation because the four cortical regions are not independently collected, thus the assumption of independent observations when applying standard correlation methods was violated. Then the similarity (correlation) matrix was transformed into an adjacency matrix with appropriate power parameters fitting the power adjacency function[288]:

$$a_{ij} = \text{power}(s_{ij}, \beta) \equiv |s_{ij}|^{\beta} \qquad \textbf{(Equation 4.1)}$$

As discussed above, the scale-free network nodes follow the power law:

$$p(k) \sim k^{-\gamma} \qquad \textbf{(Equation 4.2)}$$

To define a scale-free topology criterion, R^2 is the model fitting index of the linear model that regresses log(p(k)) on log(k) based on **Equation 4.2**, and higher R^2 means a better model fit to the scale free topology. R^2 increased as power increased (**Figure 4.6A**), thus higher power would yield a network more resembling scale free topology. However, there was also a tradeoff between scale free network resemblance and connectivity (**Figure 4.6B**), and too sparse networks forfeit too much connection information. In order to construct a network with scale-free topology and reasonable connectivity, the first power value that passed $R^2 = 0.8$ was picked to fulfill scale-free topology criterion while retaining high enough connections to investigate nodes relationships. To detect modules in the co-expression network, the topological overlap dissimilarity was measured and a topological overlap matrix (TOM) was derived to reflect relative interconnectedness between two nodes. Based on TOM, hierarchical clustering and different tree-cutting approaches were applied to determine module boundaries and gene memberships[165]. **Figure 4.7A** showed different modules derived from dynamic tree cutting on hierarchical clustering of the transcriptome wide TOM matrix containing the top 20% most variable genes. The dynamic tree cutting is a top-down approach that interactively decomposes and combines cluster branches until the assignment of module becomes stable. One problem of this dynamic tree cutting method is that it may fail to assign some tree branches, although it was not an issue in my TREM2 study (**Figure 4.7A**). To resolve this potential problem, **Figure 4.7B** showed another module assignment of the same hierarchical clustering tree with a dynamic hybrid cutting method, which is a bottom-up approach that improves the detection of any unassigned branches. Noticeably, for the module I was interested in containing *TREM2* gene (**Figure 4.7AB** labeled with star) the two tree cutting methods produced almost identical assignments with only 12 (out of 456) more genes in the hybrid tree cutting. The TREM2 modules derived from both methods showed almost identical gene memberships and network topology (**Figure 4.7CD**). *TREM2* gene is expressed in microglia and involved in immune responses in AD. Not surprisingly AD risk related genes *HLA-DRB1* and *HLA-DRB5*, which play central roles in immune system by presenting peptides derived from antigens, were also present in the same module with *TREM2*. Another sporadic AD risk gene in this module is *DSG2*, which is important for cell to cell adhesion functions, and its cytoplasmic domain anchors the cytoskeleton by interacting with plaque proteins in the desmosome-intermediate filament complex[60]. What really caught my attention is the microglial *MS4A* gene cluster that was located in the same co-expression module with *TREM2*, because we have recently observed that common variants in the *MS4A* region were significantly associated with elevated CSF soluble TREM2 level (rs1582763; p-value = 1.15×10^{-15})[70].

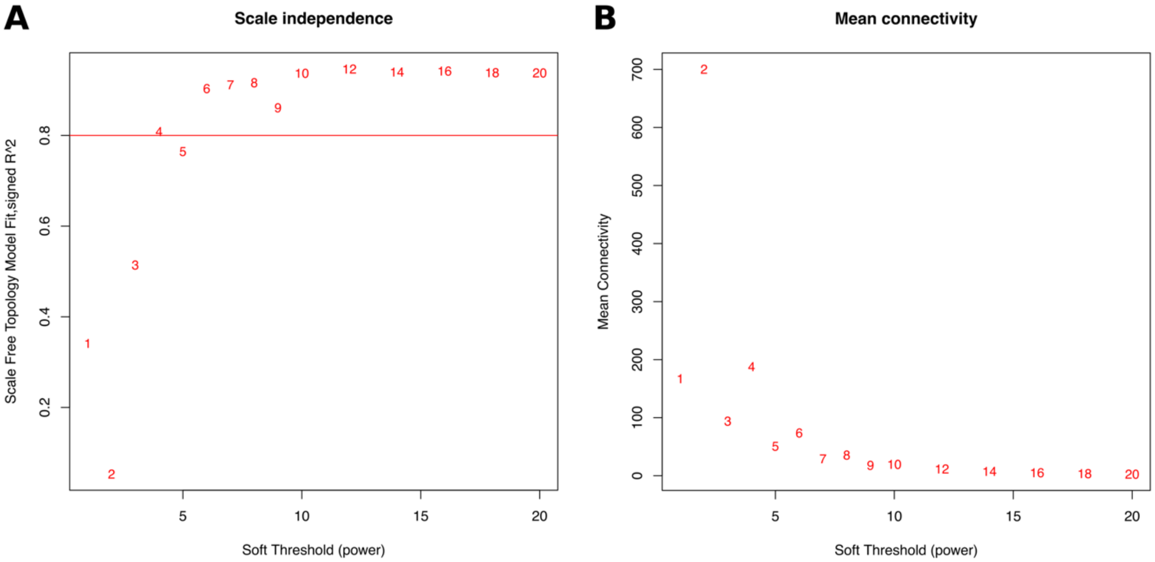


Figure 4.6 TREM2 network soft thresholding. A) R^2 is the scale-free model fitting index. Higher R^2 means a better model fit to the scale free topology. R^2 was plotted for power values ranging from 1 to 20. A R^2 = 0.8 cut-off line was plotted. B) However, there was also a tradeoff between scale free network resemblance and connectivity as the mean connectivity decreases as the power increases.

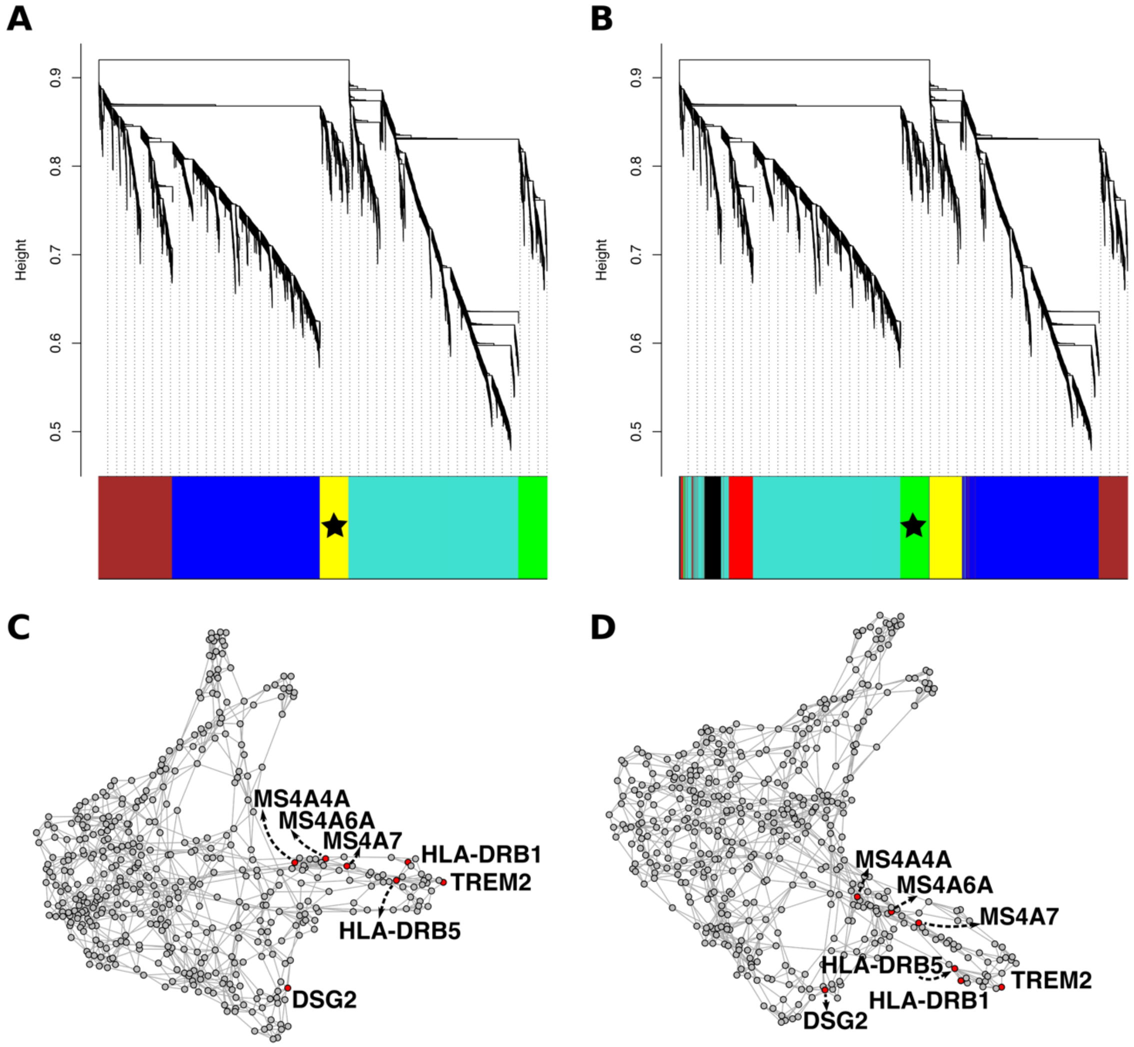


Figure 4.7 TREM2 network TOM clustering and gene module assignment. A) Dynamic tree cutting B) Dynamic hybrid cutting. C) TREM2 module derived from dynamic tree cutting. D) TREM2 module derived from dynamic hybrid cutting. The module contained TREM2 is starred.

In the previous study of network analysis in LOAD using WGCNA, Zhang et al. identified an immune/microglia module from prefrontal cortex microarray data. In this module containing *TREM2*, *TYROBP* scored the highest based on both regulatory strength and differential expression, which is an adaptor protein of TREM2. Apart from these two genes, *MS4A4A* and *MS4A6A* were also located in the same module. Although I identified *MS4A4A* and *MS4A6A* in the same module in my analysis, I seemed to have missed an important finding from

Zhang et al., which is *TYROBP*. Apart from the fact that my data is generated from RNA-Seq that is different from their microarray data, my analysis also aggregated samples from non-independent multiple tissues through repeated measures correlation. To recapitulate their finding and to understand why I missed *TYROBP* in my *TREM2* module, I performed a robust WGCNA analysis on single tissue region from anterior prefrontal cortex (BM10), which is the same region where they identified the *TREM2-TYROBP* module. The robust WGCNA is a resampling version of a regular WGCNA that I randomly resampled two thirds of the total sample size for each iteration, and repeated the process for 50 iterations to assess module assignment robustness. In this analysis, I focused on four genes that are *TREM2*, *MS4A4A*, *MS4A6A*, and *TYROBP*. *MS4A4A* and *MS4A6A* are from the same *MS4A* gene cluster located nearby on chromosome 11. Their expression patterns are highly correlated so they are included in this robustness assessment as positive controls. In the full run with all available samples from BM10 (N = 181), *MS4A4A*, *MS4A6A,* and *TYROBP* were assigned into module 12, colored as tan module in **Figure 4.8** labeled with star. *TREM2* was not assigned to any module in the full-size run. It worth mentioning that in the previous study, Zhang et al. also applied Bayesian network analysis and pathway analysis to further annotate the submodules from their *TREM2-TYROBP* network. In their pathway analysis, *MS4A4A*, *MS4A6A,* and *TYROBP* were segregated into the complement pathway, whereas *TREM2* was assigned to a separate pathway named Fc receptor system. This observation suggested that compared to *TREM2-TYROBP* co-expression, there might be a stronger relationship within *MS4A4A-MS4A6A-TYROBP* co-expression pattern. My robust WGCNA results supported this hypothesis but more in-depth simulations with 1000 runs are required to further validate this hypothesis.

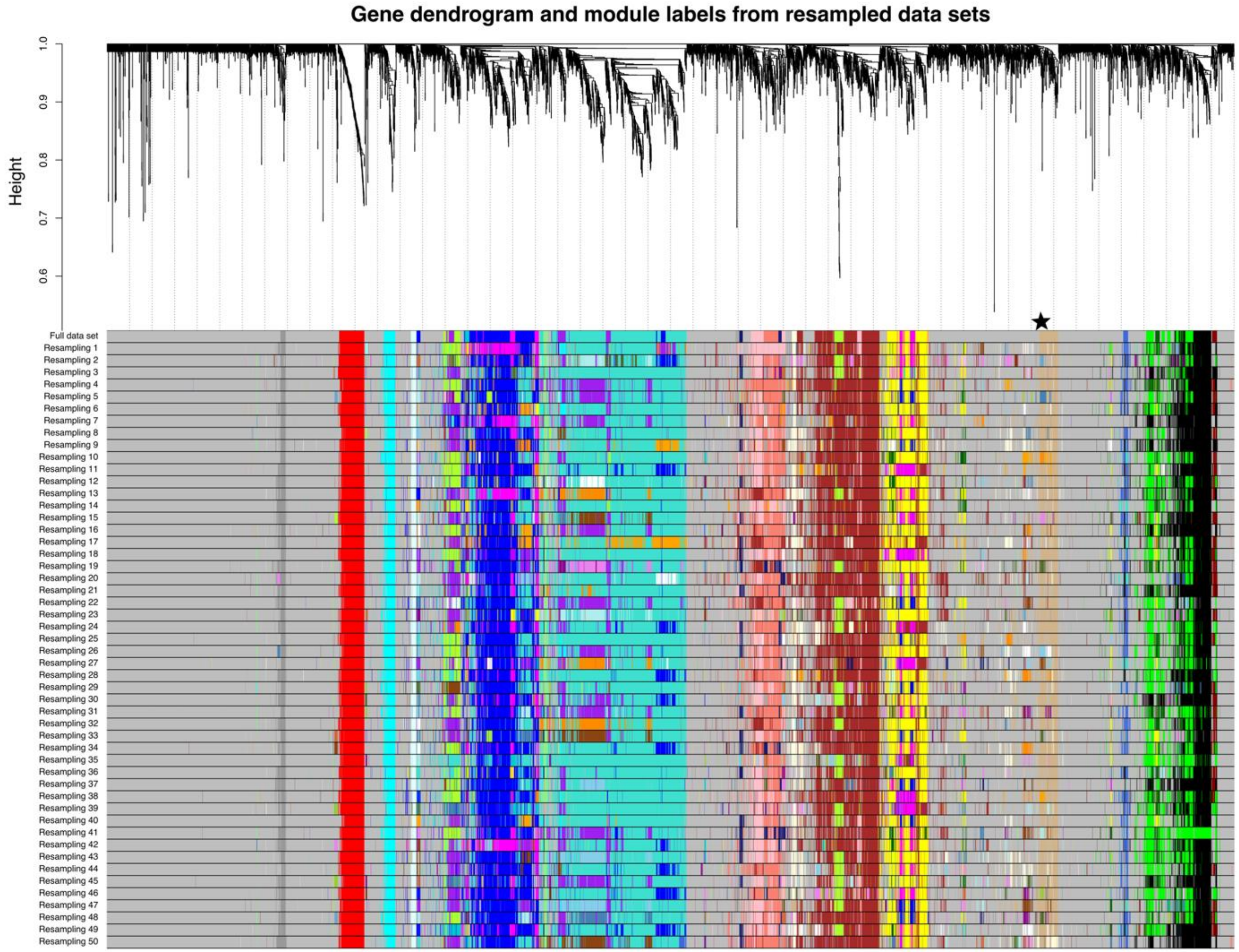


Figure 4.8 Robust WGCNA of anterior prefrontal cortex. Top hierarchical tree is derived from full dataset with all the subjects. Two thirds of the total sample size were randomly sampled for each resampling run. One full size run and total of 50 resampling runs with consensus module assignment were shown as color bars. The module contained *TREM2* is starred.

Table 4.3 Robust WGCNA simulation of *TREM2* module

RUN	Module Assignment				Dendrogram Postion			
	TREM2	*MS4A4A*	*MS4A6A*	*TYROBP*	*TREM2*	*MS4A4A*	*MS4A6A*	*TYROBP*
Full data set	0	12	12	12	5713	5770	5769	5808
Resampling 1	12	12	12	12	6741	6745	6744	6798
Resampling 2	37	0	0	0	2513	2558	2559	2486
Resampling 3	12	0	0	12	2569	2541	2540	2574
Resampling 4	0	12	12	12	1309	1354	1353	1450
Resampling 5	30	30	30	30	2538	2536	2535	2525
Resampling 6	0	12	12	27	2694	2766	2765	2645
Resampling 7	0	12	12	12	2028	1892	1950	1972
Resampling 8	0	12	12	12	1529	3569	3568	3583
Resampling 9	0	12	12	12	4876	5171	5177	5196
Resampling 10	0	26	26	26	766	6371	6370	6366
Resampling 11	0	12	12	25	2931	2829	2828	2583
Resampling 12	12	12	12	12	6498	6566	6565	6533
Resampling 13	0	0	12	12	4449	4421	4481	4490
Resampling 14	12	12	12	12	3593	3605	3604	3594
Resampling 15	32	32	32	12	1709	1695	1694	1731
Resampling 16	0	12	12	29	6272	6785	6787	6637
Resampling 17	0	12	12	12	657	4664	4663	4740
Resampling 18	12	12	12	12	2010	2108	2107	2034
Resampling 19	33	33	33	0	2559	2532	2529	2593
Resampling 20	12	12	12	12	4903	4918	4922	4924
Resampling 21	0	0	12	12	1688	1726	1849	1766
Resampling 22	0	0	0	12	3187	3452	3451	3553
Resampling 23	12	12	12	12	2681	2711	2710	2697
Resampling 24	0	12	12	12	5691	5959	5958	5892
Resampling 25	12	0	0	12	4642	4595	4589	4650
Resampling 26	28	28	28	27	2832	2803	2802	2892
Resampling 27	0	29	29	28	2055	2727	2726	2707
Resampling 28	0	0	0	12	1525	2750	2749	2709
Resampling 29	0	12	12	12	1361	1779	1778	1863
Resampling 30	0	12	12	12	1573	2146	2147	2183
Resampling 31	0	12	12	34	1982	2150	2149	2067
Resampling 32	0	12	12	12	1636	1655	1654	1701
Resampling 33	33	0	0	0	2853	2845	2844	2640
Resampling 34	12	12	12	12	6690	6647	6646	6728
Resampling 35	0	12	12	12	4430	4601	4599	4584
Resampling 36	12	0	0	12	6349	6302	6301	6440
Resampling 37	0	0	0	12	5409	5405	5404	5455
Resampling 38	0	25	25	12	2083	2216	2215	2262
Resampling 39	0	12	12	12	6470	6793	6792	6720
Resampling 40	12	12	12	12	6788	6798	6797	6844
Resampling 41	0	0	12	12	4116	4314	4448	4397
Resampling 42	0	12	12	12	6723	6739	6737	6655
Resampling 43	0	12	12	12	3958	4249	4248	4286
Resampling 44	0	12	12	12	5442	5892	5897	5914
Resampling 45	0	12	12	12	4219	4434	4438	4461
Resampling 46	0	0	0	12	3833	4361	4360	4297
Resampling 47	0	12	12	29	4060	4299	4298	4214
Resampling 48	33	0	0	12	6676	6669	6668	6742
Resampling 49	12	12	12	12	3809	3790	3803	3878
Resampling 50	12	12	12	12	4885	4861	4860	4917

TREM2 Study Bayesian network construction

To further infer the relationship between *TREM2* and *MS4A*, Bayesian network inference was applied to the WGCNA module containing *TREM2* gene. Bayesian network infers factorized probability distribution from gene expression, and incorporates it into directed acyclic graphical representation of the network. As shown in **Figure 4.9**, compared to undirected graphs from WGCNA, Bayesian network inference produces directed graphs with arrows pointing from parent nodes to child nodes. Directed acyclic graph has no directed cycles, which generates a topological ordering of nodes. The resultant network was produced from averaging across 200 bootstraps with both connection strength and direction probability hard cutoffs. In the averaged final network, there are numerous singletons, doubletons, and small branches with less than 10 nodes. The large central branch containing *TREM2* was labeled with blue, and previously identified AD risk genes were labeled with red. It worth noticing that there was a gene regulatory cascade from *MS4A6A* to *MS4A4A* to *TREM2*, suggesting a potential regulatory effect of *MS4A* genes on *TREM2*. This observation has been supported by our recent *in-vitro* MS4A knockdown study using cultured macrophage[70]. We observed that the soluble TREM2 level in cell culture was decreased after *MS4A4A* knockdown but it does not respond to *MS4A6A* knockdown. The knockdown and Bayesian network results together suggest *MS4A4A* might be the direct regulator of *TREM2*.

Pathway analysis was performed to further annotate the function of this TREM2 module, and top 8 functional pathways were labeled in the graph. The central large branch containing *TREM2*, *MS4A* genes, and *HLA-DRB* genes were involved in immune system related pathways, such as bacterial or parasitic infections, autoimmune disease systemic lupus erythematosus, tolllike receptors mediated pathogen recognition, phagocytosis and complement cascades. The second largest branch next to the blue branch was enriched with ribosomal pathway. Compared to the pathway analysis results from Zhang et al. *TREM2* module, my top 8 functional pathways with many overlapped genes replicated 4 out of 5 pathways that they highlighted for their *TREM2* network.

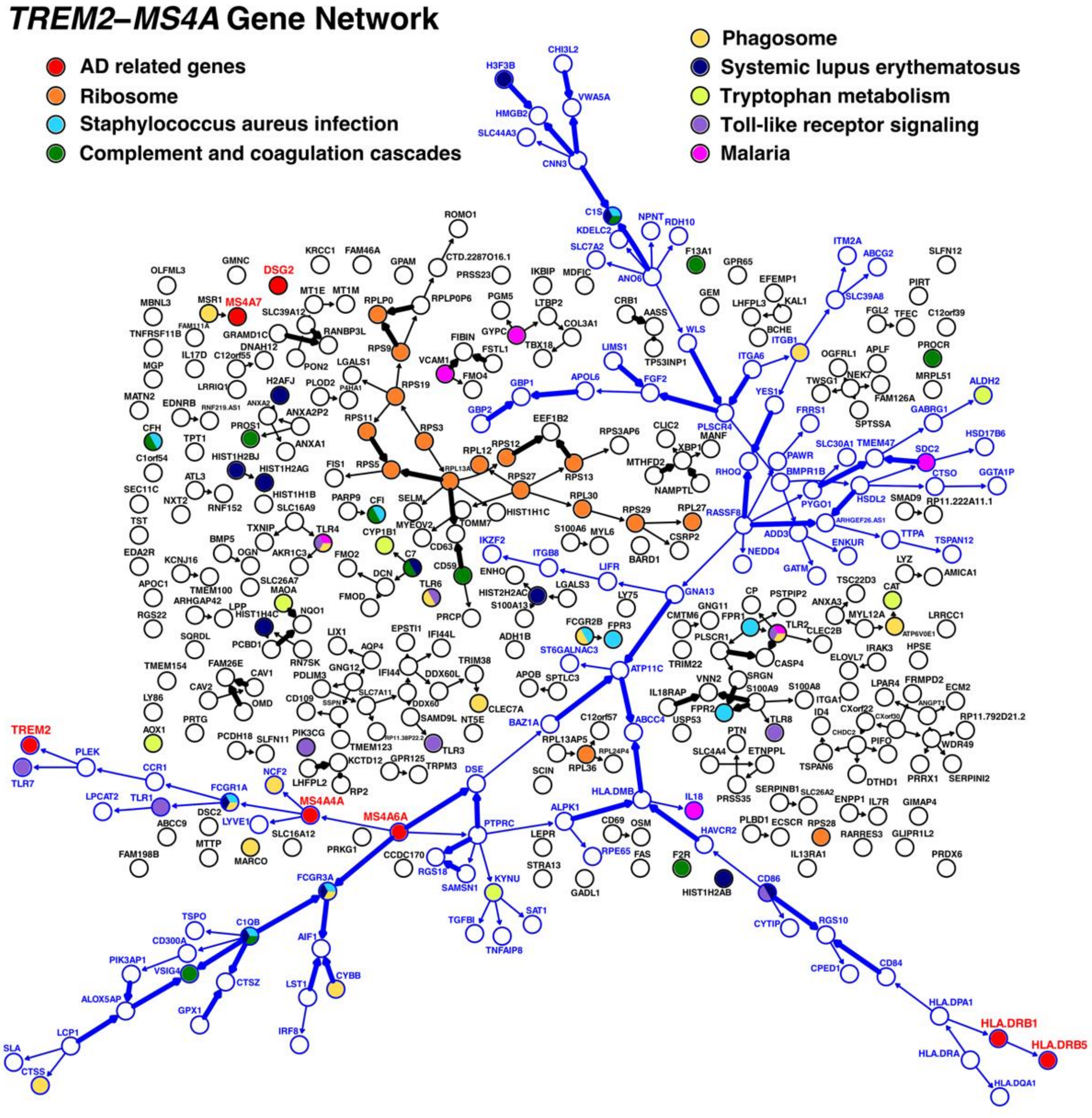


Figure 4.9 TREM2-MS4A Gene Network. AD related and top eight pathways are color coded accordingly. The largest branch is color coded in blue. V-structures are labeled in bold.

4.4.3 Disrupted and acquired genes identified in network module of *PSEN1* mutation carriers

PSEN1 study WGCNA network construction

Similar to the TREM2 study, I accessed RNA-Seq data from four cortical regions of MSSM[3], and estimated the gene expression correlation combining all four areas using repeated measure correlation. There are two key methodological differences in the approach I employed to model PSEN1 network: first, only non-demented controls were included to construct the network; second, instead of a system-based approach selecting most variable genes from the whole transcriptome, only genes correlated with *PSEN1* expression pattern were included. The genes that are correlated with *PSEN1* expression are potentially linked to AD pathology by assuming 'guilt-by-association'. With this seed-based approach using *PSEN1* as a bait, I pre-selected genes that are co-expressed with *PSEN1* gene in both *PSEN1* mutation carriers (N = 18) and non-carrier non-demented controls (N = 14) to build a co-expression network (**Figure 4.10**). Collectively, total of 5,809 genes correlated with *PSEN1* were selected from *PSEN1* non-carriers (**Figure 4.10A**) and mutation carriers (**Figure 4.10B**) that were jointly overlapped with genes correlated with *PSEN1* from MSSM non-demented control participants.

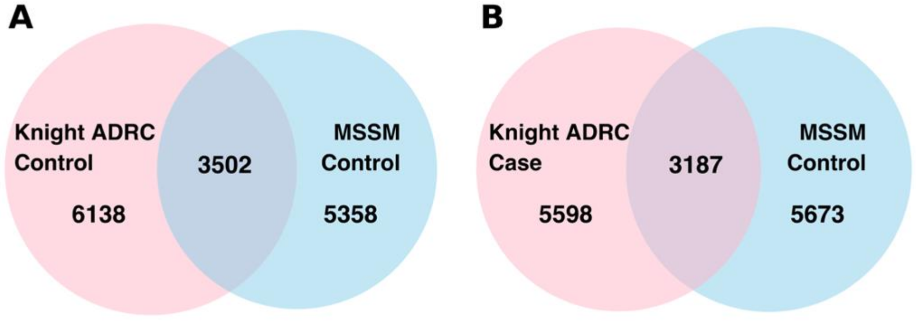


Figure 4.10 PSEN1 seeded network gene selection. Select genes from *PSEN1* mutation carriers and non-carriers correlated with MSSM control *PSEN1* gene expression.

With similar soft thresholding power selection and hierarchical clustering tree cutting approaches, modules containing *PSEN1* were produced for both dynamic tree cutting (**Figure 4.11AC**) and dynamic hybrid cutting (**Figure 4.11BD**). Compared to the dynamic tree cutting module (**Figure 4.11A**), the dynamic hybrid cutting (**Figure 4.11B**) lost branches to the left of the module and gained branches to the right of the module labeled by arrow heads. Based on a human visual inspection of the hierarchical clustering tree, the gain of the right branches might be reasonable but the loss of the left branches seems to be spurious. However, the dynamic hybrid cutting subdivided the *PSEN1* module and grouped the leftmost branches to the large module on the left of *PSEN1* module (**Figure 4.11B** turquoise module), due to which the dynamic hybrid cutting *PSEN1* module lost *BACE1*, *BIN1*, and oligodendrocyte proportion measures (**Figure 4.11D**) compared to dynamic tree cutting (**Figure 4.11C**). Based on queries of *PSEN1* in Brain RNA-Seq database of human brain purified cell type specific expression analysis[291] (http://www.brainrnaseq.org/; **Figure 4.12A**) and our single nuclei RNA-Seq dataset[66] (http://ngi.pub/snuclRNA-seq/; **Figure 4.12B**), *PSEN1* is mostly expressed in oligodendrocyte, which support dynamic tree cutting *PSEN1* module assignment.

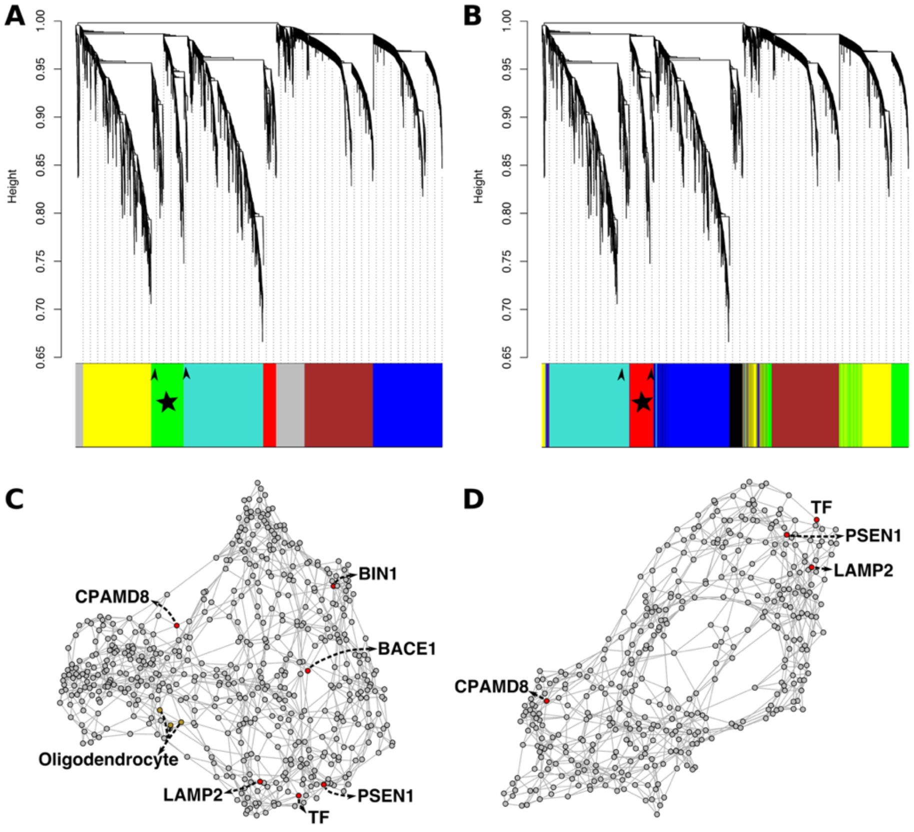


Figure 4.11 PSEN1 network TOM clustering and gene module assignment. A) Dynamic tree cutting B) Dynamic hybrid cutting. C) PSEN1 module derived from dynamic tree cutting. D) PSEN1 module derived from dynamic hybrid cutting. The module contained PSEN1 is starred.

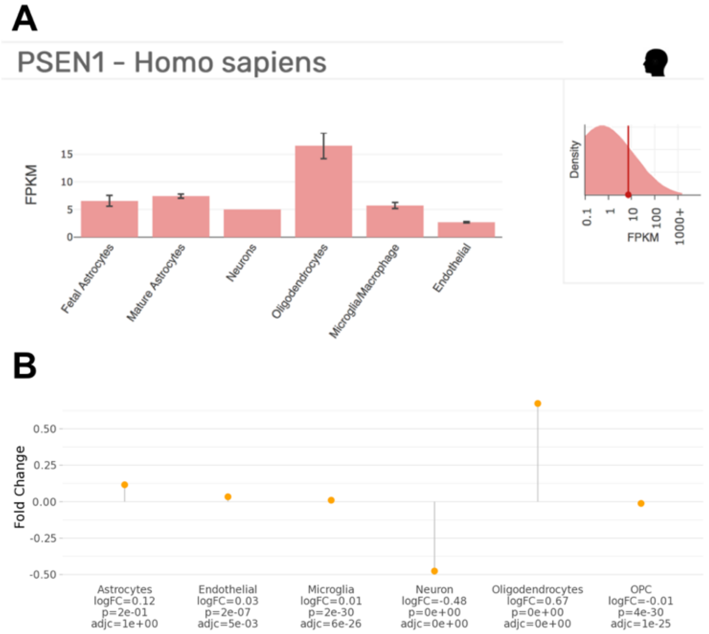


Figure 4.12 *PSEN1* is mostly expressed in oligodendrocyte. A) *PSEN1* expression in purified cell type specific expression analysis[291]. B) *PSEN1* expression in our single nuclei RNA-Seq dataset[66].

With the dynamic tree cutting *PSEN1* module, I performed differential correlation and differential expression analysis and functional annotation to identify genes that are altered in *PSEN1* mutation carriers. The analysis was replicated in another independent *PSEN1* mutation dataset. The genes were categorized into five groups and depicted as a Venn diagram[120] to show the overlaps among the groups (**Figure 4.13**). In group A, there were 268 genes that are significantly correlated with *PSEN1* expression in Knight ADRC non-demented controls. In group B, there were 264 genes that are significantly correlated with *PSEN1* expression in Knight ADRC *PSEN1* mutation carriers. In group C, there were 269 genes differentially expressed significantly comparing Knight ADRC *PSEN1* mutation carriers to non-demented controls. In group D, there were 25 genes that were differentially correlated with *PSEN1* when comparing correlation derived from group A to group B[73]. In group E, there were 64 genes specifically replicated in GEO *PSEN1* mutation carrier data (removing genes replicated in GEO PSEN1 non-carrier early onset AD cohort). Genes from combinations of A + C, A + D, and A + C + D (total of 47 genes) were potentially disrupted genes that only exist in control cohort but disappeared in *PSEN1* mutation carrier cohort. Genes from combinations of B + E, B + C + E, and B + C + D + E (total of 13 genes) were potentially emerged risk genes that were differentially correlated with *PSEN1* and/or differentially expressed when comparing *PSEN1* mutation carriers to non-carriers, and they were not correlated with *PSEN1* in control cohort.

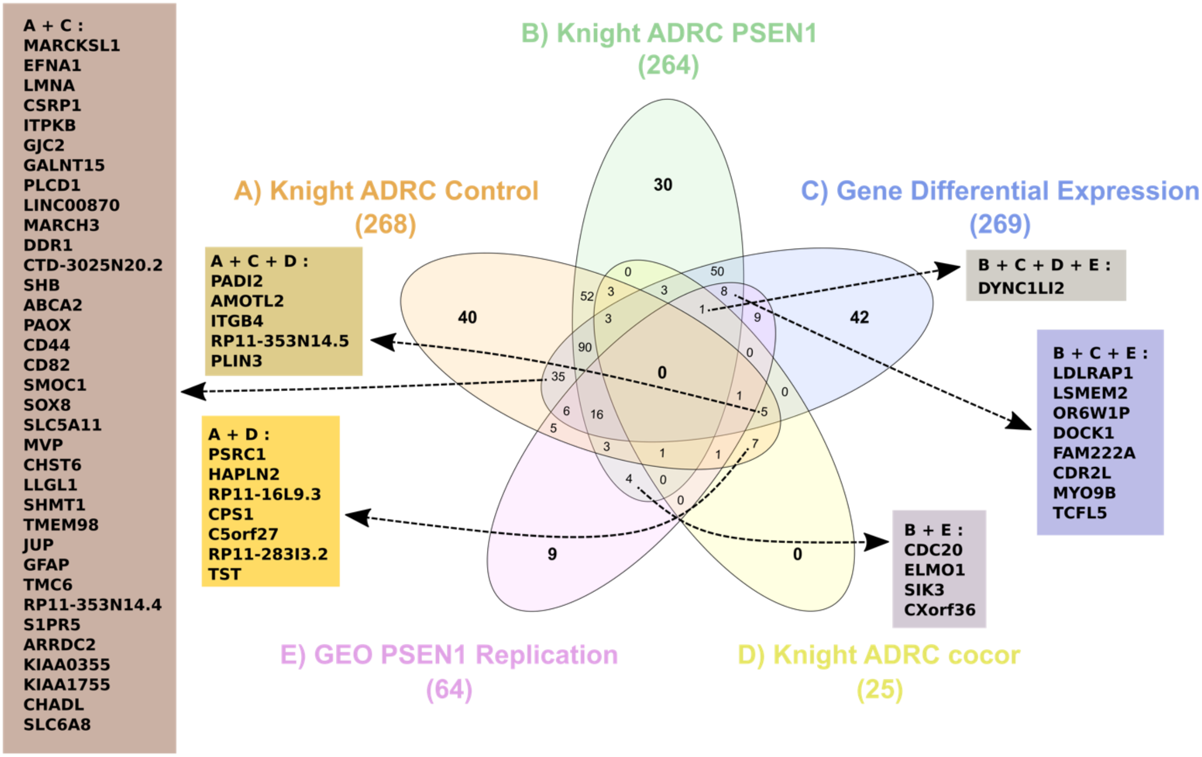


Figure 4.13 PSEN1 module genes annotation. A) Genes correlated with *PSEN1* expression in Knight ADRC non-demented controls. B) Genes significantly correlated with *PSEN1* expression in Knight ADRC *PSEN1* mutation carriers. C) Genes differentially expressed significantly comparing Knight ADRC *PSEN1* mutation carriers to non-demented controls. D) Genes differentially correlated with *PSEN1* when comparing correlation derived from group A to group B. E) Genes specifically replicated in GEO *PSEN1* mutation carrier data.

Pathway analysis were performed to the disrupted (N = 47) gene groups, emerged (N = 13) gene groups, separately and combined. In the disrupted group, arrhythmogenic right ventricular cardiomyopathy (ARVC) pathway reached significance after multiple testing correction, which contains *JUP*, *ITGB4*, and *LMNA*. Interestingly, a missense mutation (p.Asp333Gly) in *PSEN1* were previously reported in severe progressive dilated cardiomyopathy[170] suggesting a shared common pathway between heart disease and autosomal dominant AD were disrupted in *PSEN1* mutation carriers. Among the three ARVC related genes, *LMNA* is particularly interesting because mutation in this gene alone could result in an extremely rare disease called Hutchinson-Gilford progeria syndrome (1 in 4 million newborns worldwide). Patients with this disease went through a rapid “premature aging” process with growth failure, fat and hair loss, age-looking skin, and death at average of 14.6 years largely due to cardiac disease or cerebrovascular disease[262]. Although due to very limited autopsy samples, no pathological signs related to dementia or Alzheimer’s disease was identified, this accelerated aging process manifested in this disease is intriguing because aging is the most important risk factor for AD. Previously in another independent study, we also observed a significant increase of *LMNA* expression in AD brain and significantly associated with plaque load[279]. Notice that Alzheimer’s disease presenilin pathway containing *JUP* and *CD44* was also significant in the disrupted group pathway analysis. In the emerged group, *ELMO1* and *DOCK1* are related to shigellosis, bacterial invasion of epithelial cells, and integrin signaling pathway (**Table 4.4**). Interestingly, our group have identified a circular form of *DOCK1* (*circDOCK1*) was significantly up-regulated in ADAD cohort with predominantly *PSEN1* mutation carriers. *DYNC1LI2* encodes light intermediate chain 2 of dynein protein, which is a family of cytoskeletal motor proteins. It is not surprised to see this gene is related to Huntington Disease pathway. When combing both disrupted and emerged groups, apart from ARVC, shigellosis also reached significance after multiple testing correction. Shigella, a type of Gram-negative bacteria, could induce shigellosis after infection with symptoms including diarrhea, fever, and stomach cramps. The infection usually last for 5 to 7 days but there has been evidence suggesting Gram-negative bacterial molecules, such as lipopolysaccharide and *E coli* K99 pili protein, are associated with AD risk and colocalized with amyloid plaques[286]. Shigella infection related pathway might be involved as an emerged immune response to presenilin pathway disruption in *PSEN1* mutation carriers or directly related to sporadic AD susceptibility.

Table 4.4 *PSEN1* module disrupted and emerged genes pathway analysis

Type	Source	Term	Overlap	P-value	Adjusted P-value	Z-score	Combined Score	Genes
Disrupted	**KEGG**	Arrhythmogenic right ventricular cardiomyopathy (ARVC)_Homo sapiens_hsa05412	3/74	7.01×10^{-04}	3.51×10^{-02}	-1.82	13.22	*JUP;ITGB4;LMNA*
		Inositol phosphate metabolism_Homo sapiens_hsa00562	2/71	1.21×10^{-02}	1.16×10^{-01}	-1.92	8.47	*ITPKB;PLCD1*
		Biosynthesis of amino acids_Homo sapiens_hsa01230	2/74	1.31×10^{-02}	1.16×10^{-01}	-1.71	7.40	*CPS1;SHMT1*
		Hypertrophic cardiomyopathy (HCM)_Homo sapiens_hsa05410	2/83	1.63×10^{-02}	1.16×10^{-01}	-1.72	7.08	*ITGB4;LMNA*
		Dilated cardiomyopathy_Homo sapiens_hsa05414	2/90	1.90×10^{-02}	1.16×10^{-01}	-1.72	6.80	*ITGB4;LMNA*
		ECM-receptor interaction_Homo sapiens_hsa04512	2/82	1.59×10^{-02}	1.16×10^{-01}	-1.62	6.71	*ITGB4;CD44*
		Phosphatidylinositol signaling system_Homo sapiens_hsa04070	2/98	2.23×10^{-02}	1.16×10^{-01}	-1.67	6.35	*ITPKB;PLCD1*
		Carbon metabolism_Homo sapiens_hsa01200	2/113	2.90×10^{-02}	1.32×10^{-01}	-1.45	5.15	*CPS1;SHMT1*
	Pather	Alzheimer disease-presenilin pathway_Homo sapiens_P00004	2/99	2.27×10^{-02}	6.24×10^{-02}	-1.28	4.86	*JUP;CD44*
		Arginine biosynthesis_Homo sapiens_P02728	1/6	1.40×10^{-02}	6.24×10^{-02}	-0.91	3.88	*CPS1*
Acquired	**KEGG**	Shigellosis_Homo sapiens_hsa05131	2/65	7.93×10^{-04}	7.97×10^{-03}	-1.74	12.45	*ELMO1;DOCK1*
		Bacterial invasion of epithelial cells_Homo sapiens_hsa05100	2/78	1.14×10^{-03}	7.97×10^{-03}	-1.76	11.94	*ELMO1;DOCK1*
		Vasopressin-regulated water reabsorption_Homo sapiens_hsa04962	1/44	2.82×10^{-02}	1.32×10^{-01}	-1.90	6.78	*DYNC1LI2*
	Panther	Integrin signalling pathway_Homo sapiens_P00034	2/156	4.46×10^{-03}	8.91×10^{-03}	-1.78	9.64	*ELMO1;DOCK1*
		Huntington disease_Homo sapiens_P00029	1/124	7.77×10^{-02}	7.77×10^{-02}	-1.56	3.97	*DYNC1LI2*
Combined	**KEGG**	Shigellosis_Homo sapiens_hsa05131	3/65	9.82×10^{-04}	4.36×10^{-02}	-1.74	12.07	*ELMO1;DOCK1;CD44*
		Arrhythmogenic right ventricular cardiomyopathy (ARVC)_Homo sapiens_hsa05412	3/74	1.43×10^{-03}	4.36×10^{-02}	-1.79	11.70	*JUP;ITGB4;LMNA*
		Inositol phosphate metabolism_Homo sapiens_hsa00562	2/71	1.93×10^{-02}	1.66×10^{-01}	-1.89	7.45	*ITPKB;PLCD1*
		Biosynthesis of amino acids_Homo sapiens_hsa01230	2/74	2.08×10^{-02}	1.66×10^{-01}	-1.64	6.36	*CPS1;SHMT1*
		Bacterial invasion of epithelial cells_Homo sapiens_hsa05100	2/78	2.30×10^{-02}	1.66×10^{-01}	-1.64	6.17	*ELMO1;DOCK1*
		Hypertrophic cardiomyopathy (HCM)_Homo sapiens_hsa05410	2/83	2.58×10^{-02}	1.66×10^{-01}	-1.62	5.93	*ITGB4;LMNA*
		ECM-receptor interaction_Homo sapiens_hsa04512	2/82	2.52×10^{-02}	1.66×10^{-01}	-1.53	5.63	*ITGB4;CD44*
		Dilated cardiomyopathy_Homo sapiens_hsa05414	2/90	2.99×10^{-02}	1.66×10^{-01}	-1.59	5.59	*ITGB4;LMNA*
		Phosphatidylinositol signaling system_Homo sapiens_hsa04070	2/98	3.50×10^{-02}	1.78×10^{-01}	-1.55	5.21	*ITPKB;PLCD1*
		Carbon metabolism_Homo sapiens_hsa01200	2/113	4.53×10^{-02}	1.97×10^{-01}	-1.38	4.26	*CPS1;SHMT1*
	Panther	Integrin signalling pathway_Homo sapiens_P00034	3/156	1.15×10^{-02}	7.15×10^{-02}	-1.78	7.95	*ITGB4;ELMO1;DOCK1*
		Alzheimer disease-presenilin pathway_Homo sapiens_P00004	2/99	3.57×10^{-02}	8.56×10^{-02}	-1.18	3.93	*JUP;CD44*

4.5 Discussion

To interrogate highly interconnected complex system, network approach not only provides a holistic view of the overall topology but also retains the resolution to the level of each pairwise relationship. With this powerful approach, I analyzed sporadic AD cohort using a system-based approach focusing on the TREM2 module. Using WGCNA, I reconfirmed the observation of *MS4A4A* and *MS4A6A* in the TREM2 module, and inferred that *MS4A4A* might be the kay regulator of *TREM2* through Bayesian network analysis. This finding replicated previous network analysis that observed the same *TREM2-MS4A* pattern, and also provided a network explanation of our recent finding about the strong association between *MS4A* gene cluster and soluble TREM2 level in CSF. I then adapted the approach and applied to ADAD cohorts specially with *PSEN1* mutation carriers. With tissues collected from multiple cortical regions in the context of non-demented controls, I constructed a healthy cortical network topology. The network nodes are comprised of genes that are significantly co-expressed with *PSEN1* in both and *PSEN1* mutation carriers and non-demented controls with no *PSEN1* mutation. With functional annotation of the network nodes, I identified 47 genes only present in control cohort that were potentially disrupted in *PSEN1* mutation carriers; I also identified 13 genes only present in *PSEN1* mutation carriers but not in control cohort that were potentially emerged as downstream transcriptional events acquired into the *PSEN1* network. In this list of genes, I highlighted three genes, *LMNA*, *DOCK1*, and *DYNC1LI2*, that are associated to AD in additional studies.

LMNA encodes lamin protein[281], which is important for nucleoskeleton structure in providing mechanical support to nuclear envelope[61, 245]. Apart from the premature aging syndrome mentioned above in **Chapter 4.4.3**, other disease phenotypes include muscular dystrophy, lipodystrophy, and cardiomyopathy. In relation to AD, it has been shown that lamin dysfunction has led to neuronal death in tauopathy in *Drosophila*, which is also conserved in human tauopathy[96]. In an interactome study that produced a comprehensive map of molecular interactions derived from yeast two-hybrid and literature curated interactions[42, 182], Alzheimer's disease and heart disease are close neighbors in the network by sharing several proteins associated with both diseases. Apart from the missense mutation in *PSEN1* that leads to heart disease as I discussed above in **Chapter 4.4.3**, co-occurrence of cardiovascular disease and AD in elderly suggested additive or synergistic effects on both sides[16].

DOCK1 encodes a member of the dedicator of cytokinesis protein (Dock) family. *DOCK3* has been found to be associated with AD[149]. This gene's protein product was originally discovered in a yeast two-hybrid protein-protein interaction screening, which binds to presenilin so it was named as PBP (presenilin-binding protein) in the original paper[149]. We have recently inferred circular RNAs (circRNAs) from RNA-Seq data, which is a type of noncoding RNA that may be involved in AD through a circRNA-mediated "miRNA sponging systems"[175]. Although Dock1 does not bind to presenilin directly[149], circular form of DOCK1 transcript (circDOCK1) was found to be upregulated in both sporadic and autosomal dominant AD. These observations suggested a potential involvement of Dock family in AD pathology through the presenilin pathway.

DYNC1LI2 encodes light intermediate chain of dynein, which is a motor protein that is required for retrograde axonal transport along microtubules[236, 264]. Dysfunction of this protein leads to disruption of endosomal and lysosomal pathway[257]. Mutations in dynein has been found in several neurodegenerative diseases suggesting its essential role in neuronal survival, especially for motor neurons[48, 81, 188, 228]. In relation to AD, accumulation of amyloid precursor protein was observed in aged monkey (*Macaca fascicularis*) brains with dynein knockdown[152], and dynein-mediated endocytic dysfunction[256] with increased Rab GTPase level might be involved in this process[152].

With more high-throughput omic data being generated, integrating data from different sources and dimensions will be a promising future direction for network analysis. Apart from multi-dimensional networks, generating hierarchical network (network of networks) could also be an interesting direction. By learning more from the science of complexity, we could gain more insights into the complex systems from gene regulation[123] to human brain[173]. To conclude, my study demonstrated the potential of using both system-based and seed-based network approaches in replicating and discovering AD related genes and their interactions. In sporadic AD cohort, I identified *MS4A4A* might be a key regulator for *TREM2*. In autosomal dominant AD cohort, I identified total of 60 genes that are disrupted or emerged as a consequence of *PSEN1* mutation, in which I highlighted three genes with intriguing links to AD. Many of the remaining genes that are not discussed in detail in this chapter are also found to be associated with AD, such as *ABCA2*[176], *GFAP*[147], *CXorf36*[207], suggesting a great potential of using network analysis to generate working hypothesis.

Chapter 5: Conclusions and future directions

5.1 Dissertation work contributed to AD research

In **Chapter 2**, a deconvolution pipeline for bulk RNA-Seq was developed to account for cell type specific effects in brain tissues. Due to disease pathology, cell type balance is disrupted in Alzheimer's Disease (AD) brain, which is a key feature in neurodegeneration that has often been overlooked in transcriptome research. With deconvolution methods to better delineate cell population changes in disease condition, it would help interpret results and reveal transcriptional changes in a cell type specific manner.

In **Chapter 3,** using cell type proportion as quantitative trait, a common pathway underlying aging brains has been identified in the presence or absence of neurodegenerative disease symptoms. A protective variant of *TMEM106B*, which was previously identified with a protective effect in FTD, was identified to be associated with neuronal proportion in aging brains, suggesting a common pathway underlying neuronal protection and cognitive reservation in elderly. This extended analysis yield from deconvolution results from **Chapter 2** demonstrated one promising application of deconvolution followed by cell type QTL analysis in identifying new genes or pathways underlying neurodegeneration or aging in general.

In **Chapter 4**, using network analysis I replicated and reconfirmed the co-expression pattern between *MS4A* gene cluster and *TREM2* in sporadic AD, from which further evidence was inferred from Bayesian network analysis to show that *MS4A4A* might be a potential regulator of *TREM2* that is validated by *in-vitro* experiments. In Autosomal Dominant AD (ADAD) cohort, disrupted and acquired genes were identified from *PSEN1* mutation carriers. Among the genes, previous identified AD related gene and pathways were revealed together with novel findings. These results demonstrated the great potential of applying network approach in identifying disease associated genes and the interactions among them.

5.2 Aging, proteinopathy, and neurodegeneration

Apart from clinical criteria, diagnosis of neurodegenerative disorders heavily hinges on proteinopathies observed either from longitudinal brain imaging and Cerebrospinal fluid data and postmortems neuropathology. Although each disease has its own characteristic protein(s), pathological proteins from a different neurodegenerative disorder are also observed in patient autopsies. For example, AD is characterized by amyloid β and tauopathy; Parkinson’s Disease is characterized by alpha-synuclein; FTD is characterized by TDP-43 pathology. However, in AD post-mortem tissues alpha-synuclein and TDP-43 are often detected in. Similarly, amyloid β and tau are also present in PD. TDP-43 is not only identified in FTD brains, but it also presents in various neurodegenerative disorders, such as AD, ALS, hippocampal sclerosis as I have discussed in **Chapter 3.5**. Based on what I have learnt from the field and observations obtained from this dissertation work, a model is summarized and depicted in **Figure 5.1**.

Initially, soluble protein substrates in their normal folding states are performing normal functions in the brains. However, due to triggering-events that could be genetic, lifestyle or environmental factors, such as microbial infection, traumatic brain injury, toxic metal exposure, and even lack of sleep, the normal protein substrates become insoluble through a mis-folding process and start to aggregate. Regarding the initial formation of protein aggregates, a seeding hypothesis has been proposed[139]. Taking amyloid formation as an example, an initial nucleation event generates a seed that later initiates the following pathogenic accumulation of amyloid β proteins in an exponential way of aggregation. Then this pathogenic aggregation starts to propagate based on well-established prion observations that a diseased protein could convert a normal protein into a diseased state so that this process is able to propagate to other unaffected proteins. One study also demonstrated that the proteins with prion-like characteristics are also transmissible from patient to patient through injection of cadaveric pituitary-derived growth hormone[140] that induced AD pathology in patients without genetic risk factors.

What type of mis-folded proteins are generated depends on the genetic background of the individual. For example, if one person carries mutations or risk variants in AD risk genes, such as *APP*, *PSEN1*, *PSEN2*, *APOE, TREM2*, this individual will be susceptible to amyloid β and tau aggregation. Carriers of PD risk variants, in genes such as *PARK2*, *LRR2*, *PINK1*, *SNCA*, will be susceptible to alpha-synuclein aggregation. And similarly, carriers of variants in in FTD risk genes, such as *MAPT*, *GRN* or *C9orf72*, will be susceptible to tau and TDP-43 aggregation. As it was discussed above, different categories of proteinopathies are not mutually exclusive in their distribution, which means alpha-synuclein protein aggregation could be observed in AD patients with amyloid β and tau pathologies. Thus, a patient as depicted in **Figure 5.1** is clinically diagnosed with AD, but this individual also harbors multiple other proteinopathies that have not reached the threshold for clinical manifestation. This observation could be explained by an omnigenic model that apart from core pathways of AD risk genes, a collection of low-effect risk genes for other neurodegenerative disorders contribute non-trivial effects to the pathologic protein aggregation process. Apart from risk factors, there are also protective factors that could help ameliorate the damaging effect brought by toxic proteins that protect individuals with protein aggregation pathology but no cognitive deficits.

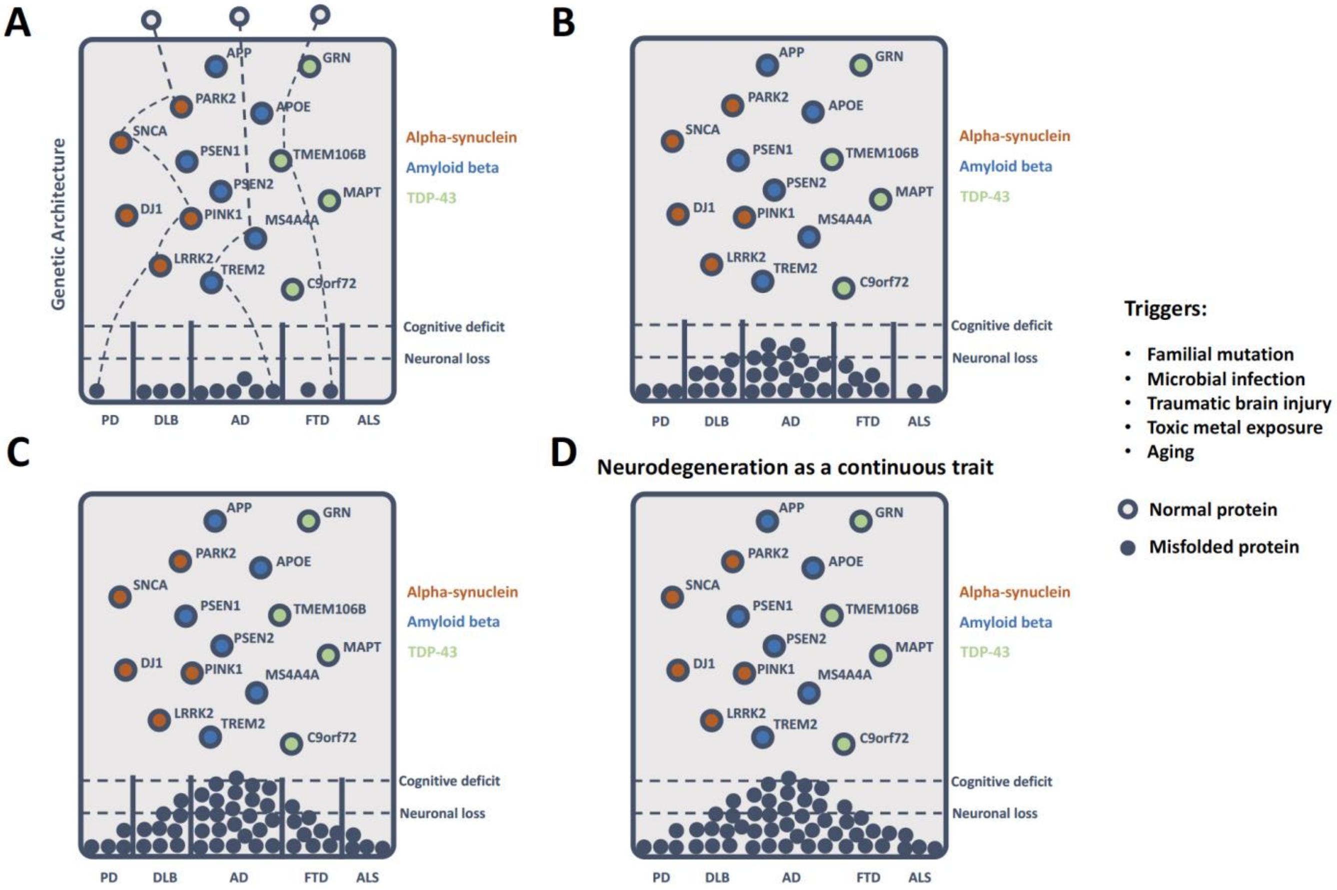


Figure 5.1 Proteinopathy model in neurodegenerative disorders. Triggered by various factors such as familial mutation, microbial infection, traumatic brain injury, toxic metal exposure, or general aging process, normal proteins may become different types of misfolded proteins based on a subject's specific genetic or epigenetic architecture as shown in A) that form alpha-synuclein, amyloid beta, or TDP-43 proteinopathies that belong to different neurodegenerative disease categories. B) Accumulated proteinopathies may lead to neuronal loss. C) As disease progresses, the subject may experience cognitive deficit with more accumulated proteinopathies. D) A patient diagnosed with AD may have alpha-synuclein and TDP-43 deposition apart from amyloid plaques that suggests neurodegeneration shall be considered as a continuous trait rather than distinctive disease categories.

5.3 Future directions in developing therapies for AD

A prevalence study focusing on preclinical and clinical AD prevalence showed that there are eight times more people in pre-symptomatic phase than people in clinical phase[39]. People with preclinical AD have either amyloidosis or neurodegeneration or both, but have no clinical manifestation of symptoms. This data showed great promise for preventive therapies that by either slowing down disease progression or delaying onset, it will be possible to protect a large number of people from developing AD during their lifetime. Currently, more than half of the disease modifying therapies are focusing on anti-amyloid approach, including immunotherapy, BASE inhibitor, and anti-aggregation. However, there have no signs of any success so far in this path. One problem with some clinical trial designs are the patients recruited for the trials are usually in their middle or late stage of AD. As it has been shown above, during the long preclinical phase, especially in the sporadic form of neurodegenerative disorders, neuropathological and neurodegenerative changes have occurred long before any clinical symptoms. However, by the time of clinical manifestation, it is usually too late for any treatment due to the extensive brain damage. Clearance of aggregated proteins may not be effective to compensate cognitive deficit due to massive neuronal loss. If neuronal loss is not reversible in the current situation, early intervention, amyloid clearance or trigger-targeting therapies should be performed earlier during disease progression to target people still in their pre-symptomatic phases. Apart from the amyloid pathway, other potential pathways to target are the lysosomal and autophagy pathway to facilitate toxic protein degradation. In addition, targeting immune pathway could prevent vicious cycles of proinflammatory responses and boost immune resilience to infections. All these pre-clinical therapies heavily rely on achieving early diagnosis in the general population. To advance early diagnosis, the development of non-invasive and accurate diagnosis tools is highly demanded to predict disease at early stages before symptom onset. We also need to advance our understanding of genetic and environmental triggers of AD to identify and better target susceptible cohorts.

Given the complexity of neurodegenerative disorders, we also need to tailor therapies differently for people with different genetic backgrounds. This idea of precision medicine has gained success in other medical fields, for example cancer treatment, from which we can learn. With better understanding of AD etiology, test results shall be interpreted in the context of person specific genetic architecture. Studies with more diverse cohorts shall be supported to understand ethnic and sex differences, which should also be considered in evaluating drug responses, dosages and side effects during clinical trials. Last but not least, cognitive or motor function improvement therapies and health care facilities specially designed for people with dementia or motor deficits are also needed for patients with neurodegenerative disorders. Regarding recent interests in finding protective factors, neuronal protection therapies could prevent the major detrimental consequence. As documented in this dissertation work, using cell type compositions inferred from deconvolution as disease endophenotypes, I identified protective variants in *TMEM106B* gene that may have neuronal protection effect in general aging groups independent of disease status, which could help understand the relationship between aging and neuronal survival and be a potential target for neuronal protection therapies.

References

1 AMPAD Knowledge Portal BroadiPSC RNAseq https://www.synapse.org/ - !Synapse:syn3607401

2 AMPAD Knowledge Portal Mayo Clinic RNAseq https://www.synapse.org/ - !Synapse:syn5550404

3 AMPAD Knowledge Portal Mount Sinai Brain Bank RNAseq https://www.synapse.org/ - !Synapse:syn3157743

4 Broad Institute The Picard Pipeline http://broadinstitute.github.io/picard/

5 (1997) Consensus recommendations for the postmortem diagnosis of Alzheimer's disease. The National Institute on Aging, and Reagan Institute Working Group on Diagnostic Criteria for the Neuropathological Assessment of Alzheimer's Disease. Neurobiol Aging 18: S1-2

6 (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580-585 Doi 10.1038/ng.2653

7 UConn StemCell Core Broad iPSC deposited in the AMP-AD https://www.synapse.org/ - !Synapse:syn36074012016

8 Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4: e6098 Doi 10.1371/journal.pone.0006098

9 Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS et al (2016) Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data 3: 160089 Doi 10.1038/sdata.2016.89

10 Amador-Ortiz C, Lin WL, Ahmed Z, Personett D, Davies P, Duara R et al (2007) TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer's disease. Ann Neurol 61: 435-445 Doi 10.1002/ana.21154

11 Andrews S (2010) Fastqc User Manual http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3 Analysis Modules/

12 Anglade P, Vyas S, Javoy-Agid F, Herrero MT, Michel PP, Marquez J et al (1997) Apoptosis and autophagy in nigral neurons of patients with Parkinson's disease. Histology and histopathology 12: 25-31

13 Antonell A, Llado A, Altirriba J, Botta-Orfila T, Balasa M, Fernandez M et al (2013) A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. Neurobiol Aging 34: 1772-1778 Doi 10.1016/j.neurobiolaging.2012.12.026

14 Askanas V, Engel WK (2001) Inclusion-body myositis: newest concepts of pathogenesis and relation to aging and Alzheimer disease. J Neuropathol Exp Neurol 60: 1-14

15 Association As (2018) 2018 Alzheimer's disease facts and figures. Alzheimer's & Dementia 14: 367-429

16 Attems J, Jellinger KA (2014) The overlap between vascular disease and Alzheimer's disease--lessons from pathology. BMC Med 12: 206 Doi 10.1186/s12916-014-0206-2

17 Bai F, Zhang Z, Yu H, Shi Y, Yuan Y, Zhu W et al (2008) Default-mode network activity distinguishes amnestic type mild cognitive impairment from healthy aging: a combined structural and resting-state functional MRI study. Neuroscience letters 438: 111-115

18 Bakdash JZ, Marusich LR (2017) Repeated Measures Correlation. Frontiers in Psychology 8: Doi 10.3389/fpsyg.2017.00456

19 Barabasi AL (2007) Network medicine--from obesity to the "diseasome". N Engl J Med 357: 404-407 Doi 10.1056/NEJMe078114

20 Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512

21 Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12: 56-68 Doi 10.1038/nrg2918

22 Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113 Doi 10.1038/nrg1272

23 Barker WW, Luis CA, Kashuba A, Luis M, Harwood DG, Loewenstein D et al (2002) Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the State of Florida Brain Bank. Alzheimer Disease & Associated Disorders 16: 203-212

24 Battle A, Brown CD, Engelhardt BE, Montgomery SB (2017) Genetic effects on gene expression across human tissues. Nature 550: 204-213 Doi 10.1038/nature24277

25 Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ et al (2014) Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. PLoS Genet 10: e1004606 Doi 10.1371/journal.pgen.1004606

26 Benitez BA, Cruchaga C (2013) TREM2 and neurodegenerative disease. N Engl J Med 369: 1567-1568 Doi 10.1056/NEJMc1306509#SA4

27 Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS (2012) Overview and findings from the religious orders study. Curr Alzheimer Res 9: 628-645

28 Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS (2012) Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res 9: 646-663

29 Blennow K, Hampel H, Weiner M, Zetterberg H (2010) Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. Nat Rev Neurol 6: 131-144 Doi 10.1038/nrneurol.2010.4

30 Bouchon A, Hernandez-Munain C, Cella M, Colonna M (2001) A DAP12-mediated pathway regulates expression of CC chemokine receptor 7 and maturation of human dendritic cells. J Exp Med 194: 1111-1122

31 Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169: 1177-1186 Doi 10.1016/j.cell.2017.05.038

32 Braak H, Braak E (1997) Frequency of stages of Alzheimer-related lesions in different age categories. Neurobiol Aging 18: 351-357

33 Braak H, Braak E (1990) Neurofibrillary changes confined to the entorhinal region and an abundance of cortical amyloid in cases of presenile and senile dementia. Acta Neuropathol 80: 479-486

34 Braak H, Braak E (1991) Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol 82: 239-259

35 Braak H, Braak E (1995) Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol Aging 16: 271-278; discussion 278-284

36 Brady OA, Zheng Y, Murphy K, Huang M, Hu F (2013) The frontotemporal lobar degeneration risk factor, TMEM106B, regulates lysosomal morphology and function. Hum Mol Genet 22: 685-695 Doi 10.1093/hmg/dds475

37 Brennand KJ The hiPSC Neurons and NPCs study (MSSMiPSC) deposited in the AMP-AD.

38 Brennand KJ, Simone A, Jou J, Gelboin-Burkhart C, Tran N, Sangar S et al (2011) Modelling schizophrenia using human induced pluripotent stem cells. Nature 473: 221-225 Doi 10.1038/nature09915

39 Brookmeyer R, Abdalla N, Kawas CH, Corrada MM (2018) Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States. Alzheimers Dement 14: 121-129 Doi 10.1016/j.jalz.2017.10.009

40 Buchman AS, Boyle PA, Wilson RS, Beck TL, Kelly JF, Bennett DA (2009) Apolipoprotein E e4 allele is associated with more rapid motor decline in older persons. Alzheimer disease and associated disorders 23: 63-69

41 Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS et al (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci 28: 264-278 Doi 10.1523/jneurosci.4178-07.2008

42 Caldera M, Buphamalai P, Müller F, Menche J (2017) Interactome-based approaches to human disease. Current Opinion in Systems Biology 3: 88-94

43 Campion D, Dumanchin C, Hannequin D, Dubois B, Belliard S, Puel M et al (1999) Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. Am J Hum Genet 65: 664-670 Doi 10.1086/302553

44 Castellano JM, Kim J, Stewart FR, Jiang H, DeMattos RB, Patterson BW et al (2011) Human apoE isoforms differentially regulate brain amyloid-beta peptide clearance. Sci Transl Med 3: 89ra57 Doi 10.1126/scitranslmed.3002156

45 Chailangkarn T, Trujillo CA, Freitas BC, Hrvoj-Mihic B, Herai RH, Yu DX et al (2016) A human neurodevelopmental model for Williams syndrome. Nature 536: 338-343 Doi 10.1038/nature19067

46 Chan G, White CC, Winn PA, Cimpean M, Replogle JM, Glick LR et al (2015) CD33 modulates TREM2: convergence of Alzheimer loci. Nat Neurosci 18: 1556-1558 Doi 10.1038/nn.4126

47 Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV et al (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14: 128 Doi 10.1186/1471-2105-14-128

48 Chen X-J, Xu H, Cooper HM, Liu Y (2014) Cytoplasmic dynein: a key player in neurodegenerative and neurodevelopmental diseases. Science China Life Sciences 57: 372-377

49 Chen-Plotkin AS, Unger TL, Gallagher MD, Bill E, Kwong LK, Volpicelli-Daley L et al (2012) TMEM106B, the risk gene for frontotemporal dementia, is regulated by the microRNA-132/212 cluster and affects progranulin pathways. J Neurosci 32: 11213-11227 Doi 10.1523/jneurosci.0521-12.2012

50 Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabasi AL et al (2018) Network-based approach to prediction and population-based validation of in silico drug repurposing. Nature communications 9: 2691 Doi 10.1038/s41467-018-05116-5

51 Chikina M, Zaslavsky E, Sealfon SC (2015) CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. Bioinformatics 31: 1584-1591 Doi 10.1093/bioinformatics/btv015

52 Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. N Engl J Med 357: 370-379 Doi 10.1056/NEJMsa066082

53 Clauset A, Rohilla Shalizi C, Newman MEJ (2007) Power-law distributions in empirical data. arXiv e-prints, City

54 Colonna M (2003) TREMs in the immune system and beyond. Nat Rev Immunol 3: 445-453 Doi 10.1038/nri1106

55 Consortium GT (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660 Doi 10.1126/science.1262110

56 Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261: 921-923

57 Cruchaga C, Graff C, Chiang HH, Wang J, Hinrichs AL, Spiegel N et al (2011) Association of TMEM106B gene polymorphism with age at onset in granulin mutation carriers and plasma granulin protein levels. Arch Neurol 68: 581-586 Doi 10.1001/archneurol.2010.350

58 Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R et al (2014) Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature 505: 550-554 Doi 10.1038/nature12825

59 Cruchaga C, Kauwe JSK, Nowotny P, Bales K, Pickering EH, Mayo K et al (2012) Cerebrospinal fluid APOE levels: an endophenotype for genetic studies for Alzheimer's disease. Hum Mol Genet 21: 4558-4571 Doi 10.1093/hmg/dds296

60 D.J. Abrams JES (2017) Diseases of the Intercalated Disc. In: John Lynn Jefferies BCB, Jeffrey Robbins, Jeffrey A. Towbin (ed) Cardioskeletal Myopathies in Children and Young Adults, City

61 Dahl KN, Kalinowski A (2011) Nucleoskeleton mechanics at a glance. J Cell Sci 124: 675-678 Doi 10.1242/jcs.069096

62 De Jager PL, Yang HS, Bennett DA (2018) Deconstructing and targeting the genomic architecture of human neurodegeneration. Nat Neurosci 21: 1310-1317 Doi 10.1038/s41593-018-0240-z

63 de Leeuw CA, Mooij JM, Heskes T, Posthuma D (2015) MAGMA: generalized gene-set analysis of GWAS data. Plos Comput Biol 11: e1004219 Doi 10.1371/journal.pcbi.1004219

64 De Strooper B, Annaert W (2010) Novel research horizons for presenilins and gamma-secretases in cell biology and disease. Annu Rev Cell Dev Biol 26: 235-260 Doi 10.1146/annurev-cellbio-100109-104117

65 Del-Aguila JL, Fernandez MV, Jimenez J, Black K, Ma SM, Deming Y et al (2015) Role of ABCA7 loss-of-function variant in Alzheimer's disease: a replication study in

European-Americans. Alzheimers Research & Therapy 7: Doi ARTN 7310.1186/s13195-015-0154-x

66 Del-Aguila JL, Li Z, Dube U, Mihindukulasuriya KA, Budde JP, Fernandez MV et al (2019) Single- nuclei RNA sequencing from human brain to study for Mendelian and sporadic AD. bioRxiv: 593756 Doi 10.1101/593756

67 Delaneau O, Marchini J (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature communications 5: 3934 Doi 10.1038/ncomms4934

68 Deleidi M, Isacson O (2012) Viral and inflammatory triggers of neurodegenerative diseases. Sci Transl Med 4: 121ps123 Doi 10.1126/scitranslmed.3003492

69 Dementia As (2019) 2019 Alzheimer's disease facts and figures. Alzheimer's & Dementia 15: 321-387

70 Deming Y, Filipello F, Cignarella F, Hsu S, Mikesell R, Li Z et al (2018) The MS4A gene cluster is a key regulator of soluble TREM2 and Alzheimer disease risk. bioRxiv: 352179 Doi 10.1101/352179

71 Deming Y, Li Z, Kapoor M, Harari O, Del-Aguila JL, Black K et al (2017) Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. Acta Neuropathol 133: 839-856 Doi 10.1007/s00401-017-1685-y

72 DIAN Dominantly Inherited Alzheimer Network http://www.dian-info.org/. Accessed 2017-05-10

73 Diedenhofen B, Musch J (2015) cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10: e0121945 Doi 10.1371/journal.pone.0121945

74 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21 Doi 10.1093/bioinformatics/bts635

75 Dominy SS, Lynch C, Ermini F, Benedyk M, Marczyk A, Konradi A et al (2019) Porphyromonas gingivalis in Alzheimer’s disease brains: Evidence for disease causation and treatment with small-molecule inhibitors. Science Advances 5: eaau3333

76 Dong XX, Wang Y, Qin ZH (2009) Molecular mechanisms of excitotoxicity and their relevance to pathogenesis of neurodegenerative diseases. Acta pharmacologica Sinica 30: 379-387 Doi 10.1038/aps.2009.24

77 Douvaras P, Sun B, Wang M, Kruglikov I, Lallos G, Zimmer M et al (2017) Directed Differentiation of Human Pluripotent Stem Cells to Microglia. Stem Cell Reports 8: 1516-1524 Doi 10.1016/j.stemcr.2017.04.023

78 Echavarri C, Aalten P, Uylings HB, Jacobs HI, Visser PJ, Gronenschild EH et al (2011) Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. Brain Struct Funct 215: 265-271 Doi 10.1007/s00429-010-0283-8

79 Efthymiou AG, Goate AM (2017) Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. Mol Neurodegener 12: 43 Doi 10.1186/s13024-017-0184-x

80 Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11: 446

81 Eschbach J, Dupuis L (2011) Cytoplasmic dynein in neurodegeneration. Pharmacology & therapeutics 130: 348-363

82 Espay AJ, Vizcarra JA, Marsili L, Lang AE, Simon DK, Merola A et al (2019) Revisiting protein aggregation as pathogenic in sporadic Parkinson and Alzheimer diseases. Neurology 92: 329-337 Doi 10.1212/wnl.0000000000006926

83 Farfel JM, Yu L, Buchman AS, Schneider JA, De Jager PL, Bennett DA (2016) Relation of genomic variants for Alzheimer disease dementia to common neuropathologies. Neurology 87: 489-496 Doi 10.1212/wnl.0000000000002909

84 Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R et al (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. Jama 278: 1349-1356

85 Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1: 1-47

86 Fernandez MV, Black K, Carrell D, Saef B, Budde J, Deming Y et al (2016) SORL1 variants across Alzheimer's disease European American cohorts. Eur J Hum Genet 24: 1828-1830 Doi 10.1038/ejhg.2016.122

87 Ferrari R, Hernandez DG, Nalls MA, Rohrer JD, Ramasamy A, Kwok JB et al (2014) Frontotemporal dementia and its subtypes: a genome-wide association study. Lancet Neurol 13: 686-699 Doi 10.1016/s1474-4422(14)70065-1

88 Fields JA, Ferman TJ, Boeve BF, Smith GE (2011) Neuropsychological assessment of patients with dementing illness. Nature Reviews Neurology 7: 677

89 Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340: 245-246 Doi 10.1038/340245a0

90 Filimonenko M, Stuffers S, Raiborg C, Yamamoto A, Malerod L, Fisher EM et al (2007) Functional multivesicular bodies are required for autophagic clearance of protein aggregates associated with neurodegenerative disease. J Cell Biol 179: 485-500 Doi 10.1083/jcb.200702115

91 Fillenbaum GG, van Belle G, Morris JC, Mohs RC, Mirra SS, Davis PC et al (2008) Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. Alzheimers Dement 4: 96-109 Doi 10.1016/j.jalz.2007.08.005

92 Finch N, Carrasquillo MM, Baker M, Rutherford NJ, Coppola G, Dejesus-Hernandez M et al (2011) TMEM106B regulates progranulin levels and the penetrance of FTLD in GRN mutation carriers. Neurology 76: 467-474 Doi 10.1212/WNL.0b013e31820a0e3b

93 Fonnum F (1984) Glutamate: a neurotransmitter in mammalian brain. J Neurochem 42: 1-11

94 Friedman N, Goldszmidt M, Wyner A (2013) Data Analysis with Bayesian Networks: A Bootstrap Approach. arXiv e-prints, City

95 Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM et al (2016) Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci 19: 1442-1453 Doi 10.1038/nn.4399

96 Frost B, Bardai FH, Feany MB (2016) Lamin dysfunction mediates neurodegeneration in tauopathies. Current Biology 26: 129-136

97 Gaiteri C, Ding Y, French B, Tseng GC, Sibille E (2014) Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes, brain and behavior 13: 13-24

98 Gaiteri C, Mostafavi S, Honey CJ, De Jager PL, Bennett DA (2016) Genetic variants in Alzheimer disease - molecular and brain network approaches. Nat Rev Neurol 12: 413-427 Doi 10.1038/nrneurol.2016.84

99 Gan L, Cookson MR, Petrucelli L, La Spada AR (2018) Converging pathways in neurodegeneration, from genetics to mechanisms. Nat Neurosci 21: 1300-1309 Doi 10.1038/s41593-018-0237-7

100 Gandal MJ, Haney JR, Parikshak NN, Leppa V, Ramaswami G, Hartl C et al (2018) Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science 359: 693-697 Doi 10.1126/science.aad6469

101 Gatz M, Pedersen NL, Berg S, Johansson B, Johansson K, Mortimer JA et al (1997) Heritability for Alzheimer's disease: the study of dementia in Swedish twins. J Gerontol A Biol Sci Med Sci 52: M117-125

102 Gaujoux R, Seoighe C (2013) CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics 29: 2211-2212 Doi 10.1093/bioinformatics/btt351

103 Gaujoux R, Seoighe C (2012) Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect Genet Evol 12: 913-921 Doi 10.1016/j.meegid.2011.08.014

104 Glenner GG, Wong CW (1984) Alzheimer's disease and Down's syndrome: sharing of a unique cerebrovascular amyloid fibril protein. Biochem Biophys Res Commun 122: 1131-1135

105 Glenner GG, Wong CW (1984) Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. Biochem Biophys Res Commun 120: 885-890

106 Goate A, Hardy J (2012) Twenty years of Alzheimer's disease-causing mutations. J Neurochem 120 Suppl 1: 3-8 Doi 10.1111/j.1471-4159.2011.07575.x

107 Golub VM, Brewer J, Wu X, Kuruba R, Short J, Manchi M et al (2015) Neurostereology protocol for unbiased quantification of neuronal injury and neurodegeneration. Front Aging Neurosci 7: 196 Doi 10.3389/fnagi.2015.00196

108 Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M et al (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS One 6: e27156 Doi 10.1371/journal.pone.0027156

109 Graff-Radford NR, Crook JE, Lucas J, Boeve BF, Knopman DS, Ivnik RJ et al (2007) Association of low plasma Abeta42/Abeta40 ratios with increased imminent risk for mild cognitive impairment and Alzheimer disease. Arch Neurol 64: 354-362 Doi 10.1001/archneur.64.3.354

110 Greenberg SM, Briggs ME, Hyman BT, Kokoris GJ, Takis C, Kanter DS et al (1996) Apolipoprotein E epsilon 4 is associated with the presence and earlier onset of hemorrhage in cerebral amyloid angiopathy. Stroke 27: 1333-1337

111 Greicius MD, Srivastava G, Reiss AL, Menon V (2004) Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proceedings of the National Academy of Sciences 101: 4637-4642

112 Guerreiro R, Hardy J (2013) TREM2 and neurodegenerative disease. N Engl J Med 369: 1569-1570 Doi 10.1056/NEJMc1306509

113 Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E et al (2013) TREM2 variants in Alzheimer's disease. N Engl J Med 368: 117-127 Doi 10.1056/NEJMoa1211851

114 Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E et al (2013) TREM2 variants in Alzheimer's disease. N Engl J Med 368: 117-127 Doi 10.1056/NEJMoa1211851

115 Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M et al (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 14: 955-958 Doi 10.1038/nmeth.4407

116 Hafkemeijer A, van der Grond J, Rombouts SA (2012) Imaging the default mode network in aging and dementia. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1822: 431-441

117 Han B, Eskin E (2012) Interpreting meta-analyses of genome-wide association studies. PLoS Genet 8: e1002555 Doi 10.1371/journal.pgen.1002555

118 Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV et al (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430: 88-93 Doi 10.1038/nature02555

119 Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML et al (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 41: 1088-1093 Doi 10.1038/ng.440

120 Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R (2015) InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 16: 169 Doi 10.1186/s12859-015-0611-3

121 Hebert LE, Weuve J, Scherr PA, Evans DA (2013) Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology 80: 1778-1783 Doi 10.1212/WNL.0b013e31828726f5

122 Heiman M, Kulicke R, Fenster RJ, Greengard P, Heintz N (2014) Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). Nature protocols 9: 1282-1291 Doi 10.1038/nprot.2014.085

123 Heltberg ML, Krishna S, Jensen MH (2019) On chaotic dynamics in transcription factors and the associated effects in differential gene regulation. Nature communications 10: 71

124 Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL et al (2015) Neuroinflammation in Alzheimer's disease. Lancet Neurol 14: 388-405 Doi 10.1016/s1474-4422(15)70016-5

125 Henstridge CM, Hyman BT, Spires-Jones TL (2019) Beyond the neuron-cellular interactions early in Alzheimer disease pathogenesis. Nature reviews Neuroscience 20: 94-108 Doi 10.1038/s41583-018-0113-1

126 Heppner FL, Ransohoff RM, Becher B (2015) Immune attack: the role of inflammation in Alzheimer disease. Nature reviews Neuroscience 16: 358-372 Doi 10.1038/nrn3880

127 Hickman S, Izzy S, Sen P, Morsett L, El Khoury J (2018) Microglia in neurodegeneration. Nat Neurosci 21: 1359-1369 Doi 10.1038/s41593-018-0242-x

128 Hippenmeyer S, Vrieseling E, Sigrist M, Portmann T, Laengle C, Ladle DR et al (2005) A developmental switch in the response of DRG neurons to ETS transcription factor signaling. PLoS biology 3: e159 Doi 10.1371/journal.pbio.0030159

129 Hippius H, Neundörfer G (2003) The discovery of Alzheimer's disease. Dialogues in clinical neuroscience 5: 101

130 Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM et al (2011) Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 43: 429-435 Doi 10.1038/ng.803

131 Holtman IR, Raj DD, Miller JA, Schaafsma W, Yin Z, Brouwer N et al (2015) Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. Acta Neuropathol Commun 3: 31 Doi 10.1186/s40478-015-0203-5

132 Holtzman DM, Morris JC, Goate AM (2011) Alzheimer's disease: the challenge of the second century. Sci Transl Med 3: 77sr71 Doi 10.1126/scitranslmed.3002369

133 Holtzman DM, Morris JC, Goate AM (2011) Alzheimer’s disease: the challenge of the second century. Sci Transl Med 3: 77sr71-77sr71

134 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955-959 Doi 10.1038/ng.2354

135 Hye A, Lynham S, Thambisetty M, Causevic M, Campbell J, Byers HL et al (2006) Proteome-based plasma biomarkers for Alzheimer's disease. Brain 129: 3042-3050 Doi 10.1093/brain/awl279

136 Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB et al (2018) NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia 14: 535-562

137 Jack CR, Bennett DA, Blennow K, Carrillo MC, Feldman HH, Frisoni GB et al (2016) A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. Neurology 87: 539-547

138 Jack Jr CR, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW et al (2010) Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9: 119-128

139 Jarrett JT, Lansbury PT, Jr. (1993) Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? Cell 73: 1055-1058

140 Jaunmuktane Z, Mead S, Ellis M, Wadsworth JD, Nicoll AJ, Kenny J et al (2015) Evidence for human transmission of amyloid-β pathology and cerebral amyloid angiopathy. Nature 525: 247

141 Jentsch TJ (2007) Chloride and the endosomal-lysosomal pathway: emerging roles of CLC chloride transporters. The Journal of physiology 578: 633-640 Doi 10.1113/jphysiol.2006.124719

142 Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41-42 Doi 10.1038/35075138

143 Jiang S, Wen N, Li Z, Dube U, Del Aguila J, Budde J et al (2018) Integrative system biology analyses of CRISPR-edited iPSC-derived neurons and human brains reveal deficiencies of presynaptic signaling in FTLD and PSP. Translational psychiatry 8: 265 Doi 10.1038/s41398-018-0319-z

144 Jonsson T, Stefansson H, Ph DS, Jonsdottir I, Jonsson PV, Snaedal J et al (2012) Variant of TREM2 Associated with the Risk of Alzheimer's Disease. N Engl J Med: Doi 10.1056/NEJMoa1211103

145 Josephs KA, Murray ME, Tosakulwong N, Weigand SD, Serie AM, Perkerson RB et al (2019) Pathological, imaging and genetic characteristics support the existence of distinct TDP-43 types in non-FTLD brains. Acta Neuropathol 137: 227-238 Doi 10.1007/s00401-018-1951-7

146 Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, Naj AC et al (2016) A novel Alzheimer disease locus located near the gene encoding tau protein. Mol Psychiatry 21: 108-117 Doi 10.1038/mp.2015.23

147 Kamphuis W, Middeldorp J, Kooijman L, Sluijs JA, Kooi EJ, Moeton M et al (2014) Glial fibrillary acidic protein isoform expression in plaque related astrogliosis in Alzheimer's disease. Neurobiol Aging 35: 492-510 Doi 10.1016/j.neurobiolaging.2013.09.035

148 Karch CM, Goate AM (2015) Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol Psychiatry 77: 43-51 Doi 10.1016/j.biopsych.2014.05.006

149 Kashiwa A, Yoshida H, Lee S, Paladino T, Liu Y, Chen Q et al (2000) Isolation and characterization of novel presenilin binding protein. Journal of neurochemistry 75: 109-116

150 Khachaturian ZS (1985) Diagnosis of Alzheimer's disease. Arch Neurol 42: 1097-1105

151 Kim J, Basak JM, Holtzman DM (2009) The role of apolipoprotein E in Alzheimer's disease. Neuron 63: 287-303 Doi 10.1016/j.neuron.2009.06.026

152 Kimura N, Inoue M, Okabayashi S, Ono F, Negishi T (2009) Dynein dysfunction induces endocytic pathology accompanied by an increase in Rab GTPases: a potential mechanism underlying age-dependent endocytic dysfunction. J Biol Chem 284: 31291-31302 Doi 10.1074/jbc.M109.012625

153 KnightADRC Knight-Alzheimer’s Disease Research Center http://alzheimer.wustl.edu/

154 Ko DC, Milenkovic L, Beier SM, Manuel H, Buchanan J, Scott MP (2005) Cell-autonomous death of cerebellar purkinje neurons with autophagy in Niemann-Pick type C disease. PLoS Genet 1: 81-95 Doi 10.1371/journal.pgen.0010007

155 Koike M, Shibata M, Ohsawa Y, Nakanishi H, Koga T, Kametaka S et al (2003) Involvement of two different cell death pathways in retinal atrophy of cathepsin D-deficient mice. Molecular and cellular neurosciences 22: 146-161

156 Koike M, Shibata M, Waguri S, Yoshimura K, Tanida I, Kominami E et al (2005) Participation of autophagy in storage of lysosomes in neurons from mouse models of neuronal ceroid-lipofuscinoses (Batten disease). Am J Pathol 167: 1713-1728 Doi 10.1016/s0002-9440(10)61253-9

157 Kuhn A, Thu D, Waldvogel HJ, Faull RL, Luthi-Carter R (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat Methods 8: 945-947 Doi 10.1038/nmeth.1710

158 Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44: W90-97 Doi 10.1093/nar/gkw377

159 Kypri E, Schmauch C, Maniak M, De Lozanne A (2007) The BEACH protein LvsB is localized on lysosomes and postlysosomes and limits their fusion with early endosomes. Traffic (Copenhagen, Denmark) 8: 774-783 Doi 10.1111/j.1600-0854.2007.00567.x

160 Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M et al (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 41: 1094-1099 Doi ng.439 [pii] 10.1038/ng.439

161 Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C et al (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45: 1452-1458 Doi 10.1038/ng.2802

162 Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C et al (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45: 1452-1458 Doi 10.1038/ng.2802

163 Lang CM, Fellerer K, Schwenk BM, Kuhn PH, Kremmer E, Edbauer D et al (2012) Membrane orientation and subcellular localization of transmembrane protein 106B (TMEM106B), a major risk factor for frontotemporal lobar degeneration. J Biol Chem 287: 19355-19365 Doi 10.1074/jbc.M112.365098

164 Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559 Doi 10.1186/1471-2105-9-559

165 Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24: 719-720 Doi 10.1093/bioinformatics/btm563

166 Laszlo L, Lowe J, Self T, Kenward N, Landon M, McBride T et al (1992) Lysosomes as key organelles in the pathogenesis of prion encephalopathies. J Pathol 166: 333-341 Doi 10.1002/path.1711660404

167 Lee JA, Beigneux A, Ahmad ST, Young SG, Gao FB (2007) ESCRT-III dysfunction causes autophagosome accumulation and neurodegeneration. Curr Biol 17: 1561-1567 Doi 10.1016/j.cub.2007.07.029

168 Leyns CEG, Ulrich JD, Finn MB, Stewart FR, Koscal LJ, Remolina Serrano J et al (2017) TREM2 deficiency attenuates neuroinflammation and protects against neurodegeneration in a mouse model of tauopathy. Proceedings of the National Academy of Sciences: 201710311 Doi 10.1073/pnas.1710311114

169 Leyns CEG, Ulrich JD, Finn MB, Stewart FR, Koscal LJ, Remolina Serrano J et al (2017) TREM2 deficiency attenuates neuroinflammation and protects against neurodegeneration in a mouse model of tauopathy. Proc Natl Acad Sci U S A 114: 11524-11529 Doi 10.1073/pnas.1710311114

170 Li D, Parks SB, Kushner JD, Nauman D, Burgess D, Ludwigsen S et al (2006) Mutations of presenilin genes in dilated cardiomyopathy and heart failure. Am J Hum Genet 79: 1030-1039 Doi 10.1086/509900

171 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760 Doi 10.1093/bioinformatics/btp324

172 Li Z, Del-Aguila JL, Dube U, Budde J, Martinez R, Black K et al (2018) Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure. Genome Med 10: 43 Doi 10.1186/s13073-018-0551-4

173 Lipsitz LA, Goldberger AL (1992) Loss of'complexity'and aging: potential applications of fractals and chaos theory to senescence. Jama 267: 1806-1809

174 Liu JZ, Erlich Y, Pickrell JK (2017) Case-control association mapping by proxy using family history of disease. Nat Genet 49: 325-331 Doi 10.1038/ng.3766

175 Lukiw W (2013) Circular RNA (circRNA) in Alzheimer's disease (AD). Frontiers in genetics 4: 307

176 Mace S, Cousin E, Ricard S, Genin E, Spanakis E, Lafargue-Soubigou C et al (2005) ABCA2 is a strong genetic risk factor for early-onset Alzheimer's disease. Neurobiol Dis 18: 119-125 Doi 10.1016/j.nbd.2004.09.011

177 Maher B (2008) Personal genomes: The case of the missing heritability. Nature News 456: 18-21

178 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ et al (2009) Finding the missing heritability of complex diseases. Nature 461: 747

179 Mattsson N, Andreasson U, Zetterberg H, Blennow K (2017) Association of plasma neurofilament light with neurodegeneration in patients with Alzheimer disease. JAMA neurology 74: 557-566

180 Mayeux R, Tang MX, Jacobs DM, Manly J, Bell K, Merchant C et al (1999) Plasma amyloid β‐peptide 1–42 and incipient Alzheimer's disease. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 46: 412-416

181 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303 Doi 10.1101/gr.107524.110

182 Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J et al (2015) Uncovering disease-disease relationships through the incomplete interactome. Science 347: 1257601

183 Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J et al (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 347: 1257601 Doi 10.1126/science.1257601

184 Miller JA, Woltjer RL, Goodenbour JM, Horvath S, Geschwind DH (2013) Genes and pathways underlying regional and cell type changes in Alzheimer's disease. Genome Med 5: 48 Doi 10.1186/gm452

185 Mintun MA, Larossa GN, Sheline YI, Dence CS, Lee SY, Mach RH et al (2006) [11C]PIB in a nondemented population: potential antecedent marker of Alzheimer disease. Neurology 67: 446-452 Doi 10.1212/01.wnl.0000228230.26044.a4

186 Mirra SS, Hart MN, Terry RD (1993) Making the diagnosis of Alzheimer's disease. A primer for practicing pathologists. Arch Pathol Lab Med 117: 132-144

187 Mirra SS, Heyman A, McKeel D, Sumi SM, Crain BJ, Brownlee LM et al (1991) The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology 41: 479-486

188 Moore JK, Sept D, Cooper JA (2009) Neurodegeneration mutations in dynactin impair dynein-dependent nuclear migration. Proceedings of the National Academy of Sciences 106: 5147-5152

189 Morris JC (1993) The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology:

190 Morris JC (1997) Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. Int Psychogeriatr 9 Suppl 1: 173-176; discussion 177-178

191 Murray ME, Cannon A, Graff-Radford NR, Liesinger AM, Rutherford NJ, Ross OA et al (2014) Differential clinicopathologic and genetic features of late-onset amnestic dementias. Acta Neuropathol 128: 411-421 Doi 10.1007/s00401-014-1302-2

192 Murray ME, Dickson DW (2014) Is pathological aging a successful resistance against amyloid-beta or preclinical Alzheimer's disease? Alzheimers Res Ther 6: 24 Doi 10.1186/alzrt254

193 Nagarajan R, Scutari M, Lbre S (2013) Bayesian Networks in R: with Applications in Systems Biology. Springer Publishing Company, Incorporated, City

194 Nagy Z, Esiri MM, Joachim C, Jobst KA, Morris JH, King EM et al (1998) Comparison of pathological diagnostic criteria for Alzheimer disease. Alzheimer disease and associated disorders 12: 182-189

195 Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Dore V et al (2018) High performance plasma amyloid-beta biomarkers for Alzheimer's disease. Nature 554: 249-254 Doi 10.1038/nature25456

196 Nakashima-Yasuda H, Uryu K, Robinson J, Xie SX, Hurtig H, Duda JE et al (2007) Co-morbidity of TDP-43 proteinopathy in Lewy body related diseases. Acta Neuropathol 114: 221-229 Doi 10.1007/s00401-007-0261-2

197 Narayanan M, Huynh JL, Wang K, Yang X, Yoo S, McElwee J et al (2014) Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. Molecular systems biology 10: 743 Doi 10.15252/msb.20145304

198 Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT et al (2006) Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science 314: 130-133 Doi 10.1126/science.1134108

199 Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y et al (2015) Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12: 453-457 Doi 10.1038/nmeth.3337

200 Nicholson AM, Finch NA, Wojtas A, Baker MC, Perkerson RB, 3rd, Castanedes-Casey M et al (2013) TMEM106B p.T185S regulates TMEM106B protein levels: implications for frontotemporal dementia. J Neurochem 126: 781-791 Doi 10.1111/jnc.12329

201 Nixon RA (2005) Endosome function and dysfunction in Alzheimer's disease and other neurodegenerative diseases. Neurobiol Aging 26: 373-382 Doi 10.1016/j.neurobiolaging.2004.09.018

202 Nixon RA, Yang DS, Lee JH (2008) Neurodegenerative lysosomal disorders: a continuum from development to late age. Autophagy 4: 590-599

203 Olsen I, Singhrao SK (2015) Can oral infection be a risk factor for Alzheimer's disease? Journal of oral microbiology 7: 29143 Doi 10.3402/jom.v7.29143

204 Pacheco CD, Kunkel R, Lieberman AP (2007) Autophagy in Niemann-Pick C disease is dependent upon Beclin-1 and responsive to lipid trafficking defects. Hum Mol Genet 16: 1495-1503 Doi 10.1093/hmg/ddm100

205 Paduraru M, Ciobica A, Mavroudis I, Fotiou D, Baloyannis S (2012) Hippocampal neuronal loss in the CA1 and CA3 areas of Alzheimer's disease patients. Psychiatria Danubina 24: 152-158

206 Parikshak NN, Gandal MJ, Geschwind DH (2015) Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat Rev Genet 16: 441-458 Doi 10.1038/nrg3934

207 Patrick E, Rajagopal S, Wong H-KA, McCabe C, Xu J, Tang A et al (2017) Dissecting the role of non-coding RNAs in the accumulation of amyloid and tau neuropathologies in Alzheimer’s disease. Molecular neurodegeneration 12: 51

208 Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14: 417-419 Doi 10.1038/nmeth.4197

209 Pekny M, Pekna M, Messing A, Steinhauser C, Lee JM, Parpura V et al (2016) Astrocytes: a central element in neurological diseases. Acta Neuropathol 131: 323-345 Doi 10.1007/s00401-015-1513-1

210 Pellerin L, Pellegri G, Bittar PG, Charnay Y, Bouras C, Martin JL et al (1998) Evidence supporting the existence of an activity-dependent astrocyte-neuron lactate shuttle. Developmental neuroscience 20: 291-299 Doi 10.1159/000017324

211 Perrin RJ, Fagan AM, Holtzman DM (2009) Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature 461: 916

212 Pottier C, Wallon D, Rousseau S, Rovelet-Lecrux A, Richard AC, Rollin-Sillaire A et al (2013) TREM2 R47H variant as a risk factor for early-onset Alzheimer's disease. J Alzheimers Dis 35: 45-49 Doi 10.3233/jad-122311

213 Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA et al (2011) Functional network organization of the human brain. Neuron 72: 665-678 Doi 10.1016/j.neuron.2011.09.006

214 Premi E, Grassi M, van Swieten J, Galimberti D, Graff C, Masellis M et al (2017) Cognitive reserve and TMEM106B genotype modulate brain damage in presymptomatic frontotemporal dementia: a GENFI study. Brain 140: 1784-1791 Doi 10.1093/brain/awx103

215 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909

216 Price JL, Morris JC (1999) Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. Ann Neurol 45: 358-368

217 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575

218 Rajarajan P, Gil SE, Brennand KJ, Akbarian S (2016) Spatial genome organization and cognition. Nature reviews Neuroscience 17: 681-691 Doi 10.1038/nrn.2016.124

219 Readhead B, Haure-Mirande JV, Funk CC, Richards MA, Shannon P, Haroutunian V et al (2018) Multiscale Analysis of Independent Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus. Neuron 99: 64-82.e67 Doi 10.1016/j.neuron.2018.05.023

220 Reid E, Connell J, Edwards TL, Duley S, Brown SE, Sanderson CM (2005) The hereditary spastic paraplegia protein spastin interacts with the ESCRT-III complex-associated endosomal protein CHMP1B. Hum Mol Genet 14: 19-38 Doi 10.1093/hmg/ddi003

221 Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J et al (2010) Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC Bioinformatics 11: 27 Doi 10.1186/1471-2105-11-27

222 Rhinn H, Abeliovich A (2017) Differential Aging Analysis in Human Cerebral Cortex Identifies Variants in TMEM106B and GRN that Regulate Aging Phenotypes. Cell Syst 4: 404-415.e405 Doi 10.1016/j.cels.2017.02.009

223 Richiardi J, Altmann A, Milazzo A-C, Chang C, Chakravarty MM, Banaschewski T et al (2015) Correlated gene expression supports synchronous activity in brain networks. Science 348: 1241-1244

224 Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G et al (2011) Integrative genomics viewer. Nature biotechnology 29: 24-26 Doi 10.1038/nbt.1754

225 Rodriguez-Arellano JJ, Parpura V, Zorec R, Verkhratsky A (2016) Astrocytes in physiological aging and Alzheimer's disease. Neuroscience 323: 170-182 Doi 10.1016/j.neuroscience.2015.01.007

226 Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F et al (2007) The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat Genet 39: 168-177

227 Rouach N, Koulakoff A, Abudara V, Willecke K, Giaume C (2008) Astroglial metabolic networks sustain hippocampal synaptic transmission. Science 322: 1551-1555 Doi 10.1126/science.1164022

228 Rubinsztein DC, Ravikumar B, Acevedo-Arozena A, Imarisio S, O’Kane CJ, Brown SD (2005) Dyneins, autophagy, aggregation and neurodegeneration. Autophagy 1: 177-178

229 Rutherford NJ, Carrasquillo MM, Li M, Bisceglio G, Menke J, Josephs KA et al (2012) TMEM106B risk variant is implicated in the pathologic presentation of Alzheimer disease. Neurology 79: 717-718 Doi 10.1212/WNL.0b013e318264e3ac

230 Ryan NS, Nicholas JM, Weston PS, Liang Y, Lashley T, Guerreiro R et al (2016) Clinical phenotype and genetic associations in autosomal dominant familial Alzheimer's disease: a case series. Lancet Neurol 15: 1326-1335 Doi 10.1016/S1474-4422(16)30193-4

231 S. A (2010) FastQC: a quality control tool for high throughput sequence data. City

232 Salta E, De Strooper B (2017) Noncoding RNAs in neurodegeneration. Nature reviews Neuroscience 18: 627-640 Doi 10.1038/nrn.2017.90

233 Sarkar S, Davies JE, Huang Z, Tunnacliffe A, Rubinsztein DC (2007) Trehalose, a novel mTOR-independent autophagy enhancer, accelerates the clearance of mutant huntingtin and alpha-synuclein. J Biol Chem 282: 5641-5652 Doi 10.1074/jbc.M609532200

234 Satoh J, Kino Y, Kawana N, Yamamoto Y, Ishida T, Saito Y et al (2014) TMEM106B expression is reduced in Alzheimer's disease brains. Alzheimers Res Ther 6: 17 Doi 10.1186/alzrt247

235 Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH et al (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology 43: 1467-1472

236 Schnapp BJ, Reese TS (1989) Dynein is the motor for retrograde axonal transport of organelles. Proceedings of the National Academy of Sciences 86: 1548-1552

237 Schwenk BM, Lang CM, Hogl S, Tahirovic S, Orozco D, Rentzsch K et al (2014) The FTLD risk factor TMEM106B and MAP6 control dendritic trafficking of lysosomes. Embo j 33: 450-467 Doi 10.1002/embj.201385857

238 Selkoe DJ (2001) Alzheimer's disease: genes, proteins, and therapy. Physiol Rev 81: 741-766

239 Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT (2011) Neuropathological alterations in Alzheimer disease. Cold Spring Harbor perspectives in medicine 1: a006189 Doi 10.1101/cshperspect.a006189

240 Sheline YI, Morris JC, Snyder AZ, Price JL, Yan Z, D'Angelo G et al (2010) APOE4 allele disrupts resting state fMRI connectivity in the absence of amyloid plaques or decreased CSF Abeta42. J Neurosci 30: 17035-17040 Doi 10.1523/jneurosci.3987-10.2010

241 Sheline YI, Raichle ME, Snyder AZ, Morris JC, Head D, Wang S et al (2010) Amyloid plaques disrupt resting state default mode network connectivity in cognitively normal elderly. Biological psychiatry 67: 584-587

242 Shen-Orr SS, Gaujoux R (2013) Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Current opinion in immunology 25: 571-578 Doi 10.1016/j.coi.2013.09.015

243 Shirk AJ, Anderson SK, Hashemi SH, Chance PF, Bennett CL (2005) SIMPLE interacts with NEDD4 and TSG101: evidence for a role in lysosomal sorting and implications for Charcot-Marie-Tooth disease. J Neurosci Res 82: 43-50 Doi 10.1002/jnr.20628

244 Simard M, Arcuino G, Takano T, Liu QS, Nedergaard M (2003) Signaling at the gliovascular interface. J Neurosci 23: 9254-9262

245 Simon DN, Wilson KL (2011) The nucleoskeleton as a genome-associated dynamic 'network of networks'. Nature reviews Molecular cell biology 12: 695-708 Doi 10.1038/nrm3207

246 Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J et al (2017) Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet 49: 1373-1384 Doi 10.1038/ng.3916

247 Srinivasan K, Friedman BA, Larson JL, Lauffer BE, Goldstein LD, Appling LL et al (2016) Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. Nature communications 7: 11295 Doi 10.1038/ncomms11295

248 Stagi M, Klein ZA, Gould TJ, Bewersdorf J, Strittmatter SM (2014) Lysosome size, motility and stress response regulated by fronto-temporal dementia modifier

TMEM106B. Molecular and cellular neurosciences 61: 226-240 Doi 10.1016/j.mcn.2014.07.006

249 Steinberg S, Stefansson H, Jonsson T, Johannsdottir H, Ingason A, Helgason H et al (2015) Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat Genet 47: 445-447 Doi 10.1038/ng.3246

250 Stobart JL, Anderson CM (2013) Multifunctional role of astrocytes as gatekeepers of neuronal energy supply. Frontiers in cellular neuroscience 7: 38 Doi 10.3389/fncel.2013.00038

251 Stobb M, Peterson JM, Mazzag B, Gahtan E (2012) Graph theoretical model of a sensorimotor connectome in zebrafish. PLoS One 7: e37292 Doi 10.1371/journal.pone.0037292

252 Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS et al (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci U S A 90: 1977-1981

253 Sul JH, Han B, Ye C, Choi T, Eskin E (2013) Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet 9: e1003491 Doi 10.1371/journal.pgen.1003491

254 Sweeney MD, Sagare AP, Zlokovic BV (2018) Blood-brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. Nat Rev Neurol 14: 133-150 Doi 10.1038/nrneurol.2017.188

255 Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663-676 Doi 10.1016/j.cell.2006.07.024

256 Tan SC, Scherer J, Vallee RB (2011) Recruitment of dynein to late endosomes and lysosomes through light intermediate chains. Mol Biol Cell 22: 467-477 Doi 10.1091/mbc.E10-02-0129

257 Tan SC, Scherer J, Vallee RB (2011) Recruitment of dynein to late endosomes and lysosomes through light intermediate chains. Molecular biology of the cell 22: 467-477

258 Tang M, Ryman DC, McDade E, Jasielec MS, Buckles VD, Cairns NJ et al (2016) Neurological manifestations of autosomal dominant familial Alzheimer's disease: a comparison of the published literature with the Dominantly Inherited Alzheimer Network observational study (DIAN-OBS). Lancet Neurol 15: 1317-1325 Doi 10.1016/S1474-4422(16)30229-0

259 Tierney MC, Fisher RH, Lewis AJ, Zorzitto ML, Snow WG, Reid DW et al (1988) The NINCDS-ADRDA Work Group criteria for the clinical diagnosis of probable Alzheimer's disease: a clinicopathologic study of 57 cases. Neurology 38: 359-364

260 Tsacopoulos M, Magistretti PJ (1996) Metabolic coupling between glia and neurons. J Neurosci 16: 877-885

261 Turner SD (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv: 005165 Doi 10.1101/005165

262 Ullrich NJ, Gordon LB (2015) Hutchinson-Gilford progeria syndrome. Handbook of clinical neurology 132: 249-264 Doi 10.1016/b978-0-444-62702-5.00018-4

263 Ulrich JD, Ulland TK, Colonna M, Holtzman DM (2017) Elucidating the Role of TREM2 in Alzheimer's Disease. Neuron 94: 237-248 Doi 10.1016/j.neuron.2017.02.042

264 Vallee RB, Williams JC, Varma D, Barnhart LE (2004) Dynein: An ancient motor protein involved in multiple modes of transport. Journal of neurobiology 58: 189-200 Doi 10.1002/neu.10314

265 van de Leemput J, Boles NC, Kiehl TR, Corneo B, Lederman P, Menon V et al (2014) CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. Neuron 83: 51-68 Doi 10.1016/j.neuron.2014.05.013

266 Van Deerlin VM, Sleiman PM, Martinez-Lage M, Chen-Plotkin A, Wang LS, Graff-Radford NR et al (2010) Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. Nat Genet 42: 234-239 Doi 10.1038/ng.536

267 Van Hoesen GW, Augustinack JC, Dierking J, Redman SJ, Thangavel R (2000) The parahippocampal gyrus in Alzheimer's disease. Clinical and preclinical neuroanatomical correlates. Ann N Y Acad Sci 911: 254-274

268 Vass R, Ashbridge E, Geser F, Hu WT, Grossman M, Clay-Falcone D et al (2011) Risk genotypes at TMEM106B are associated with cognitive impairment in amyotrophic lateral sclerosis. Acta Neuropathol 121: 373-380 Doi 10.1007/s00401-010-0782-y

269 Vidal M, Cusick ME, Barabasi AL (2011) Interactome networks and human disease. Cell 144: 986-998 Doi 10.1016/j.cell.2011.02.016

270 Villemagne VL, Dore V, Burnham SC, Masters CL, Rowe CC (2018) Imaging tau and amyloid-beta proteinopathies in Alzheimer disease and other conditions. Nat Rev Neurol 14: 225-236 Doi 10.1038/nrneurol.2018.9

271 Vonsattel JP, DiFiglia M (1998) Huntington disease. J Neuropathol Exp Neurol 57: 369-384

272 Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP et al (2018) Comprehensive functional genomic resource and integrative model for the human brain. Science 362: Doi 10.1126/science.aat8464

273 Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S et al (2016) Measure transcript integrity using RNA-seq data. BMC Bioinformatics 17: 58 Doi 10.1186/s12859-016-0922-z

274 Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q et al (2018) The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data 5: 180185 Doi 10.1038/sdata.2018.185

275 Wang Y, Ulland TK, Ulrich JD, Song W, Tzaferis JA, Hole JT et al (2016) TREM2-mediated early microglial response limits diffusion and toxicity of amyloid plaques. J Exp Med 213: 667-675 Doi 10.1084/jem.20151948

276 Watanabe K, Taskesen E, van Bochoven A, Posthuma D (2017) Functional mapping and annotation of genetic associations with FUMA. Nature communications 8: 1826 Doi 10.1038/s41467-017-01261-5

277 Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440-442 Doi 10.1038/30918

278 Webb JL, Ravikumar B, Atkins J, Skepper JN, Rubinsztein DC (2003) Alpha-Synuclein is degraded by both autophagy and the proteasome. J Biol Chem 278: 25009-25013 Doi 10.1074/jbc.M300227200

279 Wen N, Li Z, Harari O, Cruchaga C, Karch C (2018) NUCLEOSKELETON DYSREGULATION IS MEDIATED BY LMNA IN ALZHEIMER’S DISEASE. Alzheimer's & Dementia: The Journal of the Alzheimer's Association, City, pp 1115

280 White CC, Yang HS, Yu L, Chibnik LB, Dawe RJ, Yang J et al (2017) Identification of genes associated with dissociation of cognitive performance and neuropathological burden: Multistep analysis of genetic, epigenetic, and transcriptional data. PLoS medicine 14: e1002287 Doi 10.1371/journal.pmed.1002287

281 Worman HJ (2012) Nuclear lamins and laminopathies. J Pathol 226: 316-325

282 Wright AL, Zinn R, Hohensinn B, Konen LM, Beynon SB, Tan RP et al (2013) Neuroinflammation and neuronal loss precede Abeta plaque deposition in the hAPP-J20 mouse model of Alzheimer's disease. PLoS One 8: e59586 Doi 10.1371/journal.pone.0059586

283 Yang HS, Yu L, White CC, Chibnik LB, Chhatwal JP, Sperling RA et al (2018) Evaluation of TDP-43 proteinopathy and hippocampal sclerosis in relation to APOE epsilon4 haplotype status: a community-based cohort study. Lancet Neurol 17: 773-781 Doi 10.1016/s1474-4422(18)30251-5

284 Yu L, Boyle PA, Leurgans S, Schneider JA, Bennett DA (2014) Disentangling the effects of age and APOE on neuropathology and late life cognitive decline. Neurobiol Aging 35: 819-826 Doi 10.1016/j.neurobiolaging.2013.10.074

285 Yu L, Boyle PA, Nag S, Leurgans S, Buchman AS, Wilson RS et al (2015) APOE and cerebral amyloid angiopathy in community-dwelling older persons. Neurobiol Aging 36: 2946-2953 Doi 10.1016/j.neurobiolaging.2015.08.008

286 Zhan X, Stamova B, Jin LW, DeCarli C, Phinney B, Sharp FR (2016) Gram-negative bacterial molecules associate with Alzheimer disease pathology. Neurology 87: 2324-2332 Doi 10.1212/wnl.0000000000003391

287 Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA et al (2013) Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell 153: 707-720 Doi 10.1016/j.cell.2013.03.030

288 Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 4: Article17 Doi 10.2202/1544-6115.1128

289 Zhang B, Langfelder P, Horvath S (2007) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24: 719-720 Doi 10.1093/bioinformatics/btm563

290 Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S et al (2014) An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J Neurosci 34: 11929-11947 Doi 10.1523/JNEUROSCI.1860-14.2014

291 Zhang Y, Sloan SA, Clarke LE, Caneda C, Plaza CA, Blumenthal PD et al (2016) Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals

Transcriptional and Functional Differences with Mouse. Neuron 89: 37-53 Doi 10.1016/j.neuron.2015.11.013

292 Zhao Y, Raichle ME, Wen J, Benzinger TL, Fagan AM, Hassenstab J et al (2017) In vivo detection of microstructural correlates of brain pathology in preclinical and early Alzheimer Disease with magnetic resonance imaging. Neuroimage 148: 296-304 Doi 10.1016/j.neuroimage.2016.12.026

293 Zhong Y, Wan YW, Pang K, Chow LM, Liu Z (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics 14: 89 Doi 10.1186/1471-2105-14-89

294 Zhou P, Zhang Y, Ma Q, Gu F, Day DS, He A et al (2013) Interrogating translational efficiency and lineage-specific transcriptomes using ribosome affinity purification. Proc Natl Acad Sci U S A 110: 15395-15400 Doi 10.1073/pnas.1304124110

295 Zhou X, Sun L, Brady OA, Murphy KA, Hu F (2017) Elevated TMEM106B levels exaggerate lipofuscin accumulation and lysosomal dysfunction in aged mice with progranulin deficiency. Acta Neuropathologica Communications 5: 9 Doi 10.1186/s40478-017-0412-1

Zeran Li

PhD in Neuroscience

Washington University in St. Louis

(765)269-6689

zeranli@wustl.edu

Education

Washington University – St. Louis, MO, USA Aug 2011 – May 2019

Doctor of Philosophy, May 2019

Major in Neuroscience

Cognitive, Computational & Systems Neuroscience Curriculum Pathway

Purdue University – West Lafayette, IN, USA Aug 2008 – May 2010

Bachelor of Science, May 2010

Major in Biology

China Agricultural University – Beijing, China Sep 2005 – July 2008

Bachelor of Science, May 2010

Major in Biological Science

Research Experience

Carlos Cruchaga Lab, Washington University, St. Louis, MO Sept 2015 – Present

Project: Integrative Analysis to Investigate Complex Interaction in Alzheimer's Disease

- Developed a semi-supervised NMF deconvolution pipeline to resolve tissue heterogeneity in bulk RNA-seq.
- Analyzed genetic variants and transcriptome data in Alzheimer's disease to better understand disease etiology and to identify potential therapeutic targets.

- Applied machine learning algorithms and graph-based network analysis to next generation sequencing and proteome data to identify risk factors associated with Alzheimer's Disease.
- Collaborated on multi-omic projects including CSF and plasma protein, cell free RNA, circular RNA, single nuclei RNA sequencing data analysis.

John Pruett Lab, Washington University, St. Louis, MO Nov 2012 – Aug 2014

Project: Exploring Cortical Biomarkers for ASD with fcMRI Functional Parcellation

- Applied community detection and image gradient detection algorithms to analyze infant brain imaging data to identify functional areal differences in infants and adults and discover pre-clinical biomarkers for Autism disorders in infants.

Yuk Fai Leung Lab, Purdue University, West Lafayette, IN Jan 2009 – May 2011

Project: The Role of Phenylthiourea in Zebrafish Eye Growth Regulation

- Measured the smaller eyes caused by phenylthiourea (PTU) treatment.
- Proposed and tested the thyroid hormone hypothesis that the smaller eye in the PTU-treated larvae may be caused by a lower thyroid activity.
- Published one research article and one review article as first authors during undergraduate research.
- Received two internal research awards and one external research grant during undergraduate research.

De Ye Lab, China Agricultural University, Beijing, China Mar 2007 – Jan 2008

Project: KD616 Promoter of Arabidopsis Thaliana Clone, Localization, and Expression Analyses

- Constructed a bacterial plasmid of predicted *KD616* promoter fused with *GUS* reporter gene, infected *Arabidopsis Thaliana*, selected and stained the seeds of transgenic plants to observe GUS signals.

Employment

Research Assistant Aug 2010 – May 2011

Yuk Fai Leung Lab, Purdue University, West Lafayette, IN

- Conduct research to elucidate the role of phenylthiourea in zebrafish eye growth regulation.
- Conduct routine fish room and lab maintenance.
- Assisted in performing immunostaining, quantitative PCR and statistical analyses using R for several ongoing projects.

Biology Tutor Aug 2009 – May 2010

Department of Biological Sciences, Purdue University, West Lafayette, IN

- Tutored freshmen and sophomores in introductory level biology courses.

Selected Publications

- **Li Zeran**, Del-Aguila JL, Dube U, et al. Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure. Genome Med. 2018 Jun 8;10(1):43.
- **Li Zeran**, Farias FG, Dube U, et al. The *TMEM106B* rs1990621 protective variant is also associated with increased neuronal proportion. bioRxiv 2019.
- **Li Zeran**, Ptak D, Zhang L, Walls EK, Zhong W, Leung YF (2012) Phenylthiourea specifically reduces zebrafish eye size. *PLoS One* 7(6): e40132.
- **Li Zeran**, Zhang L, Leung YF (2013) Use of the zebrafish model to study refractive error. *Expert Review of Ophthalmology* 8 (1): 1 – 3.
- Deming Y, **Li Zeran**, Kapoor M, et al. Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. *Acta Neuropathol.* 2017;133(5):839-856.
- Del-Aguila JL, **Li Zeran**, Dube Umber, et al. Single-nuclei RNA sequencing from human brain to study for Mendelian and sporadic AD. bioRxiv 2019.
- Deming Y, **Li Zeran**, Benitez BA, Cruchaga C. Triggering receptor expressed on myeloid cells 2 (TREM2): a potential therapeutic target for Alzheimer disease? *Expert Opin Ther Targets*. 2018 Jun 11.
- Jiang S, Wen N, **Li Zeran**, et al. Integrative system biology analyses of CRISPR-edited iPSC-derived neurons and human brains reveal deficiencies of presynaptic signaling in FTLD and PSP. *Transl Psychiatry*. 2018;8(1):265.
- Deming Y, Filipello F, Cignarella F, Hsu S, Mikesell R, **Li Zeran**, et al. The MS4A gene cluster is a key regulator of soluble TREM2 and Alzheimer disease risk. bioRxiv 2018.

Academic Achievements and Awards

Boeing Patent Challenge Second Prize	Dec 2015
Graduated with Distinction/Honor Curriculum	May 2010
Dr. William H. Phillips Undergraduate Research Internship	Mar 2010
Sigma Xi Grants-in-Aid for Vision-related Research	Jan 2010
Sandy and Zippy Ostroy Research Experience for Undergraduates Award	Dec 2009

References

Carlos Cruchaga, Professor

Washington University in St. Louis, Missouri, USA

cruchagac@wustl.edu

(314) 286-0546

Oscar Harari, Assistant Professor

Washington University in St. Louis, Missouri, USA

harario@wustl.edu

(314) 273-1862

Sharlee Climer, Assistant Professor

University of Missouri, St. Louis, Missouri, USA

climers@umsl.edu

(314) 516-4985

Yuk Fai Leung, Associate Professor

Purdue University, West Lafayette, Indiana, USA

yfleung@purdue.edu

(765) 496-3153