UNIVERSITY OF CALIFORNIA
Los Angeles

Peripheral inflammation in neurodegenerative diseases

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Neuroscience

by

Daniel Christopher Nachun

2018

ABSTRACT OF THE DISSERTATION

Peripheral inflammation in neurodegenerative diseases

by

Daniel Christopher Nachun
Doctor of Philosophy in Neuroscience
University of California, Los Angeles, 2018
Professor Giovanni Coppola, Chair

This thesis constitutes an exhaustive analysis of peripheral blood gene expression across a diverse set of neurodegenerative disease. The first manuscript included in the thesis focuses on the analysis of peripheral blood gene expression in Friedreich's ataxia, a rare pediatric onset neurodegenerative disease caused by an autosomal recessive repeat expansion in the *FXN* gene, where the genetic basis of the disease is fully understood. The second manuscript takes a similar approach but instead focuses on neurodegenerative disorders with complex genetics and later onset, in particular Alzheimer's disease (AD), mild cognitive impairment (MCI), and five disorders in the frontotemporal dementia (FTD) spectrum: behavioral variant FTD (bvFTD), semantic variant primary progressive aphasia (svPPA), and non-fluent variant primary progressive aphasia (nfvPPA), progressive supranuclear palsy (PSP) and corticobasal syndrome (CBS). The initial focus of the project was to find specific gene expression biomarker candidates to build a biomarker panel, and to develop predictive models for disease status or severity from gene expression in blood. It became clear as the thesis progressed that both of these goals were not feasible, because expression changes in individual genes were too subtle and noisy to make viable biomarkers, and machine learning models had no predictive power for disease status or severity. However, systems

level of analysis of the peripheral blood transcriptome with weighted gene co-expression network analysis (WGCNA) revealed evidence of an increased innate immune inflammatory response in monocytes and neutrophils. This inflammatory response was found to overlap strongly with microglia-expressed genes, particularly those genes found to be affected in post-mortem AD brains. Because of this overlap with microglial genes, the genes in the inflammatory response in blood are also enriched for genetic risk for AD as determined by genome wide association studies (GWAS). The remarkable similarity of this inflammatory response across a wide array of neurodegenerative diseases warrants further investigation, particularly to determine how and why inflammatory signals enter peripheral blood from the central and peripheral nervous system in the diseases and whether this inflammation is pathological or protective and should be a target for future therapeutic interventions.

The dissertation of Daniel Christopher Nachun is approved.

Daniel H Geschwind

Stefan Horvath

Roel A Ophoff

Giovanni Coppola, Committee Chair

University of California, Los Angeles

2018

Contents

ABSTRACT OF THE DISSERTATION **ii**

Contents **v**

List of Figures **ix**

List of Tables **xv**

Acknowledgements **xvii**

Vita **xviii**

Education xviii

Publications xviii

1 Introduction **1**

1.1 Neurodegeneration 1

1.2 Peripheral blood 3

1.3 Motivation for project 4

Bibliography **5**

2 Friedreich's Ataxia **6**

2.1 Abstract 6

2.2 Introduction 7

2.3 Subjects and samples 8

2.4 Results 8

Differential Expression 8
Regression with clinical phenotypes 10
Enrichment analysis of DE and FDS-associated genes 11
Overlap with other datasets 12
Weighted Gene Coexpression Network Analysis (WGCNA) 13
Diagnosis 14
Functional disability stage 15
Enrichment analysis of significant modules 16
Cell type deconvolution 17
qPCR and array validation 18
2.5 Discussion 20
2.6 Methods 23
RNA collection and microarray hybridization 23
Array preprocessing 23
Removal of confounding covariates 24
Differential expression 24
Regression with functional disability stage and other phenotypic measures . . 25
Gene coexpression network analysis 25
Overlap with other datasets 26
Cell type deconvolution 26
Gene set annotation 27
qPCR validation 27
Data Availability 29
Ethics Statement 29
2.7 Description of analysis of other datasets 29
GSE11204 29
GSE30933 30
RNAi mouse 30

Bibliography **37**

3 Dementia **41**

3.1 Introduction . . . 41

3.2 Results . . . 42

Differential Expression . . . 42

AD and MCI . . . 42

Sex differences in AD and MCI . . . 44

FTD disorders . . . 44

ApoE genotype . . . 46

Network Analysis . . . 47

AD and MCI . . . 47

FTD disorders . . . 50

Enrichment of network modules for cell type specific genes in blood . . . 51

Relationship between innate immune modules . . . 53

Cell type composition . . . 54

Enrichment for AD genetic risk . . . 56

Enrichment for microglia genes . . . 58

DNA methylation . . . 62

Differential Methylation . . . 62

Methylation and aging . . . 63

3.3 Discussion . . . 63

3.4 Methods . . . 65

Gene expression array preprocessing . . . 65

Methylation array preprocessing . . . 66

Linear Modeling and Residualization . . . 66

Differential Expression and Differential Methylation . . . 67

Methylation aging . . . 67

Weighted Gene Co-expression Network Analysis (WGCNA) . . . 67

Cell type composition . . . 68

Pathway enrichment analysis . . . 69

Enrichment for genetic risk with MAGMA . . . 69

Enrichment for genetic risk using sLDSR 69
Enrichment for cell type specific expression 70

Bibliography **99**

List of Figures

2.1 **Figure 2-1. Differential expression identifies 829 genes differentially expressed between patients and controls and 1078 genes differentially expressed between patients and carriers.** Volcano plot of all genes in patient vs. control and patient vs. carrier comparisons. The fold change is on the x-axis, and the logBF is on the y-axis. Blue indicates a gene that is significantly downregulated (logBF > 0.5, pp > 0.95), while red indicates a gene that is significantly upregulated. 9

2.2 **Figure 2-2. Regression of gene expression with functional disability stage (FDS) identifies 1508 genes significantly associated with FDS.** Volcano plot of all genes in FDS regression. The regression coefficient is in the x-axis, and the logBF is in the y-axis. Blue indicates a gene with a significant negative regression coefficient (logBF > 0.5), while red indicates a gene with a significant positive coefficient. 10

2.3 **Figure 2-3. Enrichment analysis identifies biological pathways that are significantly overrepresented in differentially expressed and FDS-associated genes.** Bar plot of most representative enrichment term for each gene set in the y-axis. The label on the right is the pathway, and the number in parentheses is the size of the overlap between the gene set and pathway. The logBF on the x-axis, and is statistically significant at logBF > 0.5 (marked by red line). 12

2.4 **Figure 2-4. Overlap of differentially expressed genes with other datasets.** The number in the top of each cell in the heatmap is the number of transcripts in the overlap and the number in parentheses is the logBF of a hypergeometric overlap test. LogBF > 0.5 is considered significant. T3 = 12 weeks old, T4 = 16 weeks old, T5 = 20 weeks old. See supplemental text for additional descriptions of the datasets and analytic procedures. 13

2.5 **Figure 2-5. WGCNA identifies the pink, green, and black modules as significantly different across clinical status.** **a** Cluster dendrogram and color assignment for all transcripts in the full dataset. **b** Cluster dendrogram and heatmap of eigengene correlations. **c** Violin plots showing eigengene posterior estimates for the pink, green, and black modules. The 95% credible intervals are between the smaller top and bottom lines and median estimate is the larger middle line. . . . 15

2.6 **Figure 2-6. WGCNA identifies the magenta, yellow, and red modules as significantly associated with functional disability stage (FDS).** **a** Cluster dendrogram and color assignment for all transcripts in the patients with FDS available. **b** Cluster dendrogram and heatmap of eigengene correlations. **c** Scatterplots showing relationship of FDS in the x-axis with eigengene expression in the y-axis for the magenta, yellow, and red modules. 16

2.7 **Figure 2-7. Enrichment analysis identifies biological pathways that are significantly overrepresented in WGCNA modules.** Bar plot of most representative enrichment term for each gene set in the y-axis. The label on the right is the pathway, and the number in parentheses is the size of the overlap between the gene set and pathway. The logBF on the x-axis, and is statistically significant at logBF > 0.5. 17

2.8 **Figure 2-8. Cell type deconvolution analysis.** Boxplots showing cell type proportion of 7 cell types in patients, carriers and controls. 18

2.9 **Figure 2-9. qPCR and array validation of top 3 DE genes in 32 patients and 32 age- and sex-matched controls.** Boxplots showing the relative expression of the top 3 DE genes to the median value of the control samples. 21 patient and 16 control samples (marked with closed circles) were new and not previously included in the analysis. Top: microarray data, bottom: qPCR. 19

2.10 **Figure 2-S1. Clinical status is significantly associated with age, sex, batch and RNA integrity number (RIN).** Diagnostic plots showing the relationship between clinical status and **a** age, **b** sex, **c** batch, **d** RIN, and **e** site. 32

2.11 **Figure 2-S2. Functional disability stage is significantly associated with age.** Diagnostic plots showing the relationship between functional disability stage and **a** age, **b** sex, **c** batch, **d** RIN, and **e** site. 33

2.12 **Figure 2-S3. No overlap in differentially expressed genes is observed with DRG and cerebellum in RNAi mouse**. Heatmaps showing the overlap of up- and down-regulated transcripts in our datasets with the tissues in the RNAi mouse. The number in the top of each cell is the number of transcripts in the overlap and the number in parentheses is the logBF of a hypergeometric overlap test. LogBF > 0.5 is considered significant. 34

2.13 **Figure 2-S4. Functional disability stage (FDS) is not associated with cell type proportions.** Scatterplots showing functional disability stage vs. cell type proportion. 35

3.1 **Figure 3-1. a** Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles. **b** Bar plots of enrichment of significantly upregulated genes for neutrophil degranulation (GO:0043312) with the logBF on the x-axis. 43

3.2 **Figure 3-2. a,b** Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles. **c** Correlation plot of the pairwise correlation of logFC values of FTD disorders vs. control, with the dendrogram on top showing hierarchical clustering of disorders. 46

3.3 **Figure 3-3. a-c,** Violin plots of posterior estimate of mean eigengene values for each diagnosis, with the median and 5% and 95% quantiles indicated by lines. **c** Bar plot of enrichment of genes in each module for neutrophil degranulation (GO:0043312) with the logBF on the x-axis. 48

3.4 **Figure 3-4. a** Violin plot of posterior estimate of mean eigengene values for each diagnosis in the brown module, with the median and 5% and 95% quantiles indicated by lines. **b** Correlation plot of the pairwise correlation of the mean differences in eigengene values of FTD disorders vs. control, with the dendrogram on top showing hierarchical clustering of disorders. 50

3.5 **Figure 3-5. a,b** Bar plots of enrichment of the top 300 genes in each WGCNA module for monocyte- and neutrophil-specific genes in the AD and FTD networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend. 52

3.6 **Figure 3-6. a** UpSet plot of overlap between WGCNA modules. Total module size is in the bar plot on the left and overlap sizes are in the bar plots on the top with first bar showing genes in all 4 modules, and the other bars showing genes unique to each set. **b** Correlation plot of connectivity correlation between the brown modules in the AD, female AD and FTD networks and yellow module in the male AD network. 53

3.7 **Figure 3-7. a,b** Box plots of blood cell type composition estimated from gene expression in AD, MCI and control (**a**) and FTD disorders (**b**). 55

3.8 **Figure 3-8. a,b** Bar plots of enrichment of the top 300 genes in each WGCNA module for genetic risk for AD as estimated by MAGMA in the AD and FTD networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. 57

3.9 **Figure 3-9. a,b** Bar plots of enrichment of the top 300 genes in each WGCNA microglial genes in the AD and FTD networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend. 60

3.10 **Figure 3-10. a,b** Bar plots of enrichment of the top 300 genes in the brown module of the AD network for the module in each post-mortem network most enriched for microglial genes. The logBF is on the x-axis and the number of genes in each overlap is in parentheses next to the y-axis label. . . . 61

3.11 **Figure 3-S1. a,b** Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles. 71

3.12 **Figure 3-S2.** Bar plots of enrichment of significantly upregulated genes for neutrophil degranulation (GO:0043312) with the logBF on the x-axis. 72

3.13 **Figure 3-S3.** Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles. 73

3.14 **Figure 3-S4. a-d,** Violin plots of posterior estimate of mean eigengene values for each diagnosis, with the median and 5% and 95% quantiles indicated by lines. **d** Bar plot of enrichment of genes in each module for neutrophil degranulation (GO:0043312) with the logBF on the x-axis. 74

3.15 **Figure 3-S5. a,b** Bar plots of enrichment of the top 300 genes in each WGCNA module for monocyte- and neutrophil-specific genes in the AD male and female networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend. 75

3.16 **Figure 3-S6.** Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses. 76

3.17 **Figure 3-S6. (cont)** Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses. 77

3.18 **Figure 3-S6. (cont)** Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses. 78

3.19 **Figure 3-S6. (cont)** Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses. 79

3.20 **Figure 3-S7.** Box plots of blood cell type composition estimated from methylation in AD, MCI and control (**a**) and FTD disorders (**b**). 80

3.21 **Figure 3-S8. a,b** Bar plots of enrichment of the top 300 genes in each WGCNA module for genetic risk for AD as estimated by MAGMA in the AD male and female networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. 81

3.22 **Figure 3-S9. a-d** Labeled heatmaps of enrichment of all network modules for other GWAS studies. Each cell shows the -$\log_{10}$ p-value of the enrichment. 82

3.23 **Figure 3-S10. a,b** Bar plots of association of all module memberships with AD risk for all modules in (**a**) AD and FTD networks and (**b**) AD male and female networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. 83

3.24 **Figure 3-S11. a,b** Line plots of sensitivity analysis showing all cutoffs (**a**) or all cutoffs until the first cutoff that is not significant (**b**). The x-axis is the size of the cutoff of genes ranked by membership in the module in each plot title, and the y-axis show the -$\log_{10}$ p-value. 84

3.25 **Figure 3-S12.** Bar plots of enrichment of the top 300 genes in each WGCNA module for genetic risk for AD as estimated by sLDSR in all four networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. . 85

3.26 **Figure 3-S13.** Bar plots of enrichment of the top 300 genes in each WGCNA module for microglial genes in the AD male and female networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend. 86

3.27 **Figure 3-S14.** Labeled heatmaps of enrichment of all network modules for all CNS cell types using the Zhang et. al. dataset. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses. 87

3.28 **Figure 3-S15.** Labeled heatmaps of enrichment of all network modules for all CNS cell types using the Wang et. al. dataset. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses. 88

3.29 **Figure 3-S16.** Bar plots of enrichment of the top 300 genes in the brown module of the FTD and AD female networks and yellow module in the AD male network for the module in each post-mortem network most enriched for microglial genes. The logBF is on the x-axis and the number of genes in each overlap is in parentheses next to the y-axis label. 89

3.30 **Figure 3-S16. (cont)** Bar plots of enrichment of the top 300 genes in the brown module of the FTD and AD female networks and yellow module in the AD male network for the module in each post-mortem network most enriched for microglial genes. The logBF is on the x-axis and the number of genes in each overlap is in parentheses next to the y-axis label. 90

3.31 **Figure 3-S17.** Volcano plots of the log fold change (logFC) in CpG islands, promoters and gene bodies on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The methylation annotation and number of DM features are shown in the plot titles. 91

3.32 **Figure 3-S17. (cont)** Volcano plots of the log fold change (logFC) in CpG islands, promoters and gene bodies on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The methylation annotation and number of DM features are shown in the plot titles. 92

3.33 **Figure 3-S17. (cont)** Volcano plots of the log fold change (logFC) in CpG islands, promoters and gene bodies on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The methylation annotation and number of DM features are shown in the plot titles. 93

3.34 **Figure 3-S18.** Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge. 94

3.35 **Figure 3-S18. (cont)** Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge. 95

3.36 **Figure 3-S18. (cont)** Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge. 96

3.37 **Figure 3-S18. (cont)** Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge. 97

3.38 **Figure 3-S18. (cont)** Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge. 98

List of Tables

Table 1-1. Summary of typical symptoms, affected CNS/PNS regions and proteinopathies found in neurodegenerative diseases. 1

Table 1-2. Summary of cell types in peripheral blood. 3

Table 2-1. Summary of subject demographics. 8

Acknowledgements

Chapter 1 includes material from the following manuscript now in press:

Nachun, Daniel; Gao, Fuying; Isaacs, Charles; Strawser, Cassandra; Yang, Zhongan; Dokuru, Deepika; Van Berlo, Victoria; Sears, Renee; Farmer, Jennifer; Perlman, Susan; Lynch, David R; Coppola, Giovanni. Peripheral blood gene expression reveals an inflammatory transcriptomic signature in Friedreich's ataxia patients. Hum Mol Genet. 2018; doi:10.1093/hmg/ddy198

Giovanni Coppola was the principal investigator, David Lynch and Susan Perlman provided patient samples, Deepika Dokuru, Zhongan Yang, Victoria Van Berlo, and Renee Sears prepared samples, Charles Isaacs, Cassandra Strawser, Jennifer Farmer, Zhongan Yang, and Deepika Dokuru handled clinical data, and Fuying Gao helped with data analysis. Funding was provided by the Friedreich's Ataxia Research Alliance, the Muscular Dystrophy Association and the National Institutes of Health.

Chapter 2 includes material for a manuscript in preparation. Authorship is as follows:

Nachun, Daniel; Gao, Fuying; Yang, Zhongan; Dokuru, Deepika; Van Berlo, Victoria; Sears, Renee; Ramos, Eliana Marisa; Karydas, Anna; Boxer, Adam; Rosen, Howard; Miller, Bruce; Coppola, Giovanni.

Giovanni Coppola was the principal investigator, Bruce Miller, Howard Rose, and Adam Boxer provided patient samples, Anna Karydas, Eliana Marisa Ramos, Deepika Dokuru and Zhongan Yang provided or processed clinical data, Deepika Dokuru, Zhongan Yang, Renee Sears, and Victoria Van Berlo processed samples, and Fuying Gao helped with data analysis.

Vita

Education

B.S., Biology, June 2012

Williams College

Williamstown, MA

Thesis: *DFOXO, an Insulin Pathway Component, Acts in Cholinergic Motor Neurons to Moduleate Bang-Sensitivity in Slamdance Mutants*

Adviser: Dr. Derek Dean

GPA: 3.33/4.0

High School Diploma, June 2008

Westside High School

Omaha, NE

GPA: 3.96/4.0

Publications

Peripheral blood gene expression reveals an inflammatory transcriptomic signature in Friedreich's ataxia patients. **Nachun D**, Gao F, Isaacs C, Strawser C, Yang Z, Dokuru D, Van Berlo V, Sears R, Farmer J, Perlman S, Lynch D, Coppola G. Human Molecular Genetics, 2018. *In press.*

Inflammatory peripheral biomarkers in Alzheimer's disease, mild cognitive impairment and frontotemporal dementia. **Nachun D**, Gao F, Yang Z, Dokuru D, Van Berlo V, Sears R, Karydas A, Boxer AL, Rosen H, Miller BL, Coppola G. *In preparation*

Small intestine crypts exhibit methylation age deceleration in healthy tissue. Lewis S, **Nachun D**, Horvath SH, Coppola G, Jones DL. *In preparation*

Methylation aging in an in vitro glial model of Huntington's disease. Foti R, **Nachun D**, Drigabzu A, Coppola G, Goldman S. *In preparation*

Transcriptional profiling of isogenic iPS-derived Friedreich's ataxia sensory neurons. Lai J, **Nachun D**, Petrosyan L, Throesch B, Baldwin K, Coppola G, Gottesfeld JM, Soragni E. *In preparation*

Transcription profiling of sleep-deprived mice reveals decreased synaptic plasticity and increased stress response. Aaling N, **Nachun D**, Coppola G, Goldman S, Nedergaard M. *In preparation*

Peripheral blood gene expression biomarkers in Friedreich's Ataxia. **Nachun D**, Gao F, Isaacs C, Yang Z, Dokuru D, Van Berlo V, Sears R, Farmer J, Perlman S, Lynch D, Coppola G. *Friedreich's Ataxia Research Meeting*, King of Prussia, PA, September, 2016.

Peripheral blood gene expression biomarkers in Friedreich's Ataxia. **Nachun D**, Gao F, Isaacs C, Yang Z, Dokuru D, Van Berlo V, Sears R, Farmer J, Perlman S, Lynch D, Coppola G. *Society for Neuroscience Annual Meeting*, San Diego, CA, November, 2016.

Peripheral blood gene expression biomarkers in Friedreich's Ataxia. **Nachun D**, Gao F, Isaacs C, Strawser ,C Yang Z, Dokuru D, Van Berlo V, Sears R, Farmer J, Perlman S, Lynch D, Coppola G. *International Ataxia Research Conference*, Pisa, Italy, September, 2017.

Chapter 1

Introduction

1.1 Neurodegeneration

Neurodegenerative diseases pose an enormous and growing public health problem, having been identified as the fourth most common cause of death in the United States in 2010 [1]. More concerningly, while most of the top 10 causes of death in the United States and worldwide such as cancer, cardiovascular diseases, infections, or accidents are to some extent preventable or treatable, there are no known interventions to prevent or cure neurodegenerative disease. Consequently, there is an enormous desire to better understand neurodegeneration so that prevention or treatment becomes possible. The goal of this thesis is not to identify any possible treatments – this work is the domain of experimental biology, whereas the work presented here is entirely observational. Instead my goal is to enhance the understanding of neurodegeneration so that future experimental work may ultimately lead to the desired interventions.

Before describing my findings, it is important to summarize what is already known about neurodegeneration so that my results can be placed in their proper context. An exhaustive review of neurodegenerative diseases is beyond the scope of both this chapter and this thesis – the reviews referenced should be consulted for this purpose. Table 1-1 shows a simplified summary of clinical features, affected brain regions, and pathologies of the complex genetic disorders discussed in this thesis [2].

Disease	Typical Symptoms	CNS/PNS Region	Proteinopathy
Alzheimer's Disease (AD)	Memory loss; confusion and disorientation	Hippocampus	Amyloid beta, MAPT
Mild Cognitive Impairment (MCI)	Memory loss (less severe than AD)	Hippocampus	Amyloid beta, MAPT
Parkinson's Disease (PD)	Tremors; shuffling	Midbrain structures	SNCA, MAPT
Amyotrophic Lateral Sclerosis (ALS)	Muscle weakness; loss of motor control	Neuromuscular junction	TDP-43, FUS
Behavioral variant frontotemporal dementia (bvFTD)	Emotional and personality changes; impaired decision making	Frontal and temporal lobes	MAPT, TDP-43, FUS
Semantic variant primary progressive aphasia (svPPA)	Loss of semantic meaning of words	Temporal lobe	TDP-43
Non-fluent variant primary progressive aphasia (nfvPPA)	Loss of grammar; incorrect ordering of words	Frontal lobe	MAPT
Progressive supranuclear palsy (PSP)	Atypical parkinsonism	Midbrain structures	MAPT
Corticobasal syndrome (CBS)	Rigidity and akinesia	Motor cortex	MAPT

Table 1-1. Summary of typical symptoms, affected CNS/PNS regions and proteinopathies found in neurodegenerative diseases.

While there are pathogenic variants which can cause inherited familial forms of these disorders, they only account for a small percentage of the total number of cases of these diseases. However, there are rare neurodegenerative disorders which are truly monogenic, being caused by a single pathogenic variant that is either recessive (two copies needed for disease) or dominant (only one copy needed for disease). One of the most well known examples of this is the dominant disorder Huntington's Disease, caused by a CAG expansion in the

HTT gene. An example of a recessive monogenic neurodegenerative disorder is described in Chapter 2, Friedreich's ataxia (FRDA), which is caused by a GAA repeat expansion in the FXN gene. It is characterized by a gait dysfunction (ataxia), loss of sensation and motor control, and other symptoms outside the nervous system including cardiomyopathy and diabetes. It is also noteworthy that unlike both the more common neurodegenerative disorders in Table 2-1 and Huntington's Disease, FRDA does not lead to protein aggregation in any affected cell types. The age of onset of FRDA is also much younger than most neurodegenerative diseases, with most patients exhibiting symptoms before the age of 25.

1.2 Peripheral blood

All of the biological data presented in this thesis is collected from peripheral blood, which – like the nervous system – is composed of a complex set of cell types with highly specialized functions. Table 1-2 groups cell types in blood into the adaptive and innate immune system and other cell types, and summarizes the role of each of the three categories.

Adaptive Immune System	**Innate Immune System**	**Other cells**
T-cells	Monocytes	Megakaryocytes
B-cells	Dendritic cells	Platelets
Natural killer cells	Neutrophils	Red blood cells
	Eosinophils	
	Basophils	
Targets pathogens for removal with specific antibodies	Clears pathogens in a non-specific manner; clears cellular debris from injury and normal cell death	Megaryocytes produce platelets for clotting; Red blood cells transport oxygen and carbon dioxide

Table 1-2. Summary of cell types in peripheral blood.

Peripheral blood has been studied in neurodegenerative disorders for many years, and several important findings have been made. Toxic oligomers of amyloid beta [3], SNCA [4],

MAPT [5], and TDP-43 [6] have been found in the blood serum of patients with neurodegenerative disease and can be predictive of disease pathology [3]. Many studies have found evidence of increased concentrations of inflammatory cytokines in blood such as IL-6, IL-1B, and CRP [7], but it remains unknown if this inflammation is pathological or protective. Gene expression and methylation have also been studied in neurodegenerative diseases, but these studies have had low sample sizes or statistical confounds and typically do not consider the biology of blood when interpreting their results. Despite these limitations, the largest recent study of the peripheral transcriptome in AD and MCI patients, for example, found evidence of an increased immune response and a decrease in mitochondrial translation [8].

1.3 Motivation for project

My goals for this thesis were to improve upon previous studies of peripheral blood gene expression and methylation in neurodegenerative diseases in several ways. I planned to use much larger samples sizes and very rigorous statistical analysis to increase my statistical confidence in my findings. I also planned to use systems biology approaches to incorporate existing biological data relevant to blood, such as cell type-specific gene expression, into my interpretation of my results. I believe that a rigorous and thorough analysis of peripheral blood gene expression and methylation in neurodegeneration is useful even though blood cells do not exhibit obvious pathology. While this does place some limits on the understanding that can be obtained about the biology of neurodegeneration from blood, the low cost and ease of accessibility of blood compared to imaging and post-mortem tissue makes it worthwhile to study.

Bibliography

1. Murray, C. J. L. & Lopez, A. D. Measuring the global burden of disease. en. *N. Engl. J. Med.* **369,** 448–457 (Aug. 2013).
2. Elahi, F. M. & Miller, B. L. A clinicopathological approach to the diagnosis of dementia. en. *Nat. Rev. Neurol.* **13,** 457–476 (Aug. 2017).
3. Nakamura, A. *et al.* High performance plasma amyloid-β biomarkers for Alzheimer's disease. en. *Nature* **554,** 249–254 (Feb. 2018).
4. Bengoa-Vergniory, N., Roberts, R. F., Wade-Martins, R. & Alegre-Abarrategui, J. Alpha-synuclein oligomers: a new hope. en. *Acta Neuropathol.* **134,** 819–838 (Dec. 2017).
5. Majerova, P. *et al.* Microglia display modest phagocytic capacity for extracellular tau oligomers. en. *J. Neuroinflammation* **11,** 161 (Sept. 2014).
6. Zondler, L. *et al.* Impaired activation of ALS monocytes by exosomes. en. *Immunol. Cell Biol.* **95,** 207–214 (Feb. 2017).
7. Lai, K. S. P. *et al.* Peripheral inflammatory markers in Alzheimer's disease: a systematic review and meta-analysis of 175 studies. en. *J. Neurol. Neurosurg. Psychiatry* **88,** 876–882 (Oct. 2017).
8. Lunnon, K. *et al.* Mitochondrial dysfunction and immune activation are detectable in early Alzheimer's disease blood. en. *J. Alzheimers. Dis.* **30,** 685–710 (2012).

Chapter 2

Friedreich's Ataxia

2.1 Abstract

Transcriptional changes in Friedreich's ataxia (FRDA), a rare and debilitating recessive Mendelian neurodegenerative disorder, have been studied in affected but inaccessible tissues – such as dorsal root ganglia, sensory neurons, and cerebellum – in animal models or small patient series. However, transcriptional changes induced by FRDA in peripheral blood, a readily accessible tissue, have not been characterized in a large sample. We used differential expression, association with disability stage, network analysis, and enrichment analysis to characterize the peripheral blood transcriptome and identify genes that were differentially expressed in FRDA patients (n=418) compared to both heterozygous expansion carriers (n=228) and controls (n=93, 739 individuals in total), or were associated with disease progression, resulting in a disease signature for FRDA. We identified a transcriptional signature strongly enriched for an inflammatory innate immune response. Future studies should seek to further characterize the role of peripheral inflammation in FRDA pathology and determine its relevance to overall disease progression.

2.2 Introduction

Friedreich's ataxia (FRDA, OMIM 229300) is a rare autosomal recessive disorder characterized by progressive ataxia, significant loss of motor control, cardiomyopathy, and diabetes. The disorder is usually caused by an intronic trinucleotide (GAA) repeat expansion in the highly conserved gene frataxin (*FXN*, ENSG00000165060), whose protein product is essential to the formation of iron-sulfur cluster complexes. These complexes are necessary for the proper functioning of a large number of proteins, particularly those involved in mitochondrial metabolism. FRDA is a result of *FXN* haploinsufficiency, and complete loss of *FXN* is embryonic lethal [1]. FRDA patients exhibit a 70-80% reduction of *FXN* expression levels compared to unaffected individuals [2]. Heterozygous expansion carriers exhibit a modest reduction in *FXN* expression and do not develop clinical symptoms.

FXN deficiency causes a number of pathologies at the cellular level (reviewed in [3]). A large build up in mitochondrial iron and reduced function of antioxidant proteins lead to an increase in reactive oxygen species, which lead to severe oxidative stress characterized by damage to proteins, DNA, and lipid membranes. These effects can ultimately lead to degeneration and cell death, particularly in post-mitotic cells with very high metabolic activity, such as large neurons, cardiomyocytes, and pancreatic islet cells [4], but many of the affected pathways are universal to the function of all eukaryotic cells, and a more subtle transcriptional response may be present in peripheral tissues not clinically involved, but readily available for study in large cohorts. In addition, because FRDA results in severe metabolic stress and eventual loss of cells of the peripheral and central nervous system, this may lead to a peripheral immune response that can be detected at the level of gene expression in blood immune cells. To explore these hypotheses, we collected the largest series to date of RNA from peripheral blood from FRDA patients, carriers, and controls, and performed microarray-based gene expression analysis. We identified an inflammatory disease-associated signature which in part overlaps with previous datasets from patients and animal models. The entire dataset is available to the FRDA community in a web-based application (REPAIR) for data mining and additional analyses.

Status	Male	Female	Total	Age	GAA1 length
Patient	221 (53%)	197 (47%)	418	25 ± 11.9	900 ± 185
Carrier	89 (39%)	139 (61%)	228	50 ± 17.8	N/A
Control	53 (57%)	40 (43%)	93	37 ± 10.4	N/A
Total	363	376	739		

Table 2-1. Summary of subject demographics.

2.3 Subjects and samples

739 subjects were enrolled at two sites, UCLA and the Children's Hospital of Philadelphia (CHOP). Table 2-1 provides a basic summary of the demographics of the subjects. Subjects were divided into three groups based on clinical diagnosis. Patients were those subjects clinically diagnosed with FRDA (n=418) and in most (90.6%) the approximate number of GAA repeats in the *FXN* gene was also determined via PCR [5] to serve as molecular confirmation, as well as an indirect measure of disease severity. Eighteen patients were compound heterozygotes with one repeat expansion and one loss-of-function point mutation in *FXN* (Table 2-S1). Carriers were those subjects carrying one expanded *FXN* allele and one normal allele (n=228). Most carriers were parents of patients, who are obligate carriers. Control subjects consisted of individuals known not to have any relatives with FRDA. Because carriers and controls are phenotypically indistinguishable, we checked blood frataxin levels in 95 enrolled controls [6] and excluded 2 subjects with frataxin levels lower than the range observed in homozygote expansion carriers, leaving 93 controls for further analyses.

2.4 Results

Differential Expression

In order to identify a peripheral signature related to FRDA pathology, we fit linear models for each transcript. At a cutoff of $\log_{10}$ Bayes Factor > 0.5 (logBF, see Methods) comparing the

alternate model containing disease status to the null model without it, after accounting for a number of potential confounders (see Methods and Figure 2-S1-2), 1115 transcripts were significant for the effect of disease status across all three groups. To identify transcripts that were significantly differentially expressed (DE) across pairwise comparisons, we computed posterior probabilities and identified transcripts for each pairwise comparison where the posterior probability (pp) of differential expression was greater than 0.95 (see Methods). The global false discovery rate (FDR) for each set of DE transcripts in each comparison was also computed as described in Methods. Of the 1115 transcripts identified as being significantly affected by disease status, 829 transcripts were DE between patients and controls (global FDR = **0.012**), 1078 between patients and carriers (global FDR = **0.0017**) and 182 between carriers and controls (global FDR = **0.018**) (Figure 2-1a-b, Table 2-S2). The observation that more genes were DE in patients vs. carriers compared to patients vs. controls is likely due the much larger number of carriers (228) compared to controls (93), which provides stronger statistical support to small expression changes.

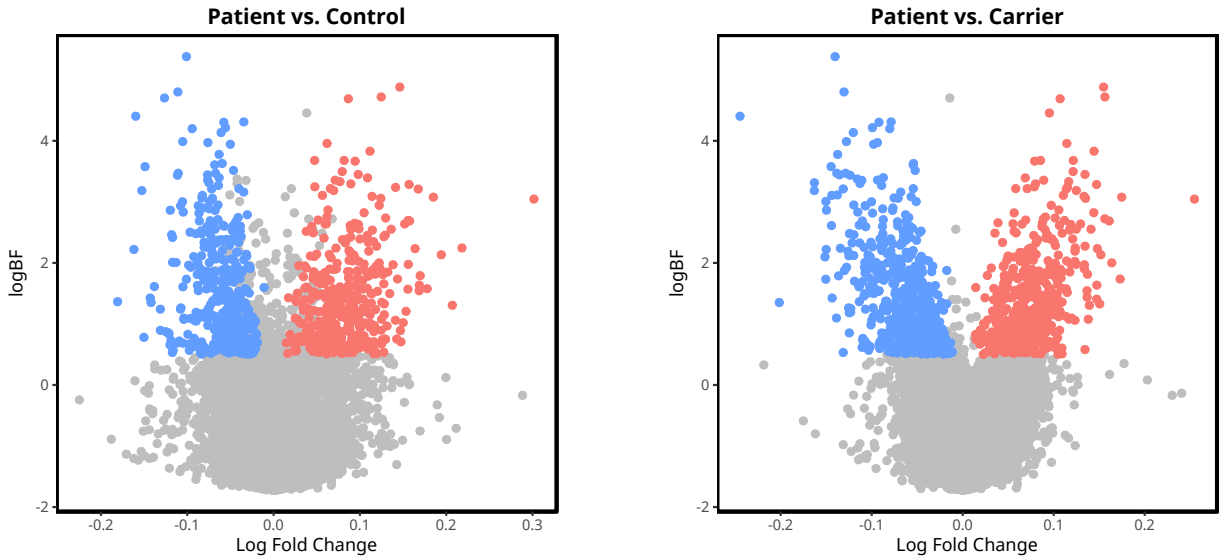

Figure 2-1. Differential expression identifies 829 genes differentially expressed between patients and controls and 1078 genes differentially expressed between patients and carriers. Volcano plot of all genes in patient vs. control and patient vs. carrier comparisons. The fold change is on the x-axis, and the logBF is on the y-axis. Blue indicates a gene that is significantly downregulated (logBF > 0.5, pp > 0.95), while red indicates a gene that is significantly upregulated.

Regression with clinical phenotypes

Several phenotypic measures can be used to quantify disease severity in FRDA patients. A direct clinical measure is the Functional Disability Stage (FDS) score developed for the Friedreich's Ataxia Rating Scale [7], which rates patients on a scale from 0-6 based upon their mobility, with 0 indicating no impairment and 6 complete disability. Two less direct measures of disease severity are the disease duration in years and the size of the shorter GAA repeat expansion in patients, GAA1. We used linear modeling to identify transcripts with significant positive or negative linear relationships with each phenotypic measure. At a cutoff of logBF > 0.5, comparing the alternate model with the phenotypic measure to the null model without it, we identified 1508 transcripts significantly associated with FDS (global FDR = **0.0028**, Figure 2-2, Table 2-S3), 280 transcripts significantly associated with GAA1 (global FDR = **0.0043**), and 13 transcripts significantly associated with disease duration (global FDR = **0.006**). In all 3 analyses, all genes with logBF > 0.5 also had a posterior probability > 0.95.

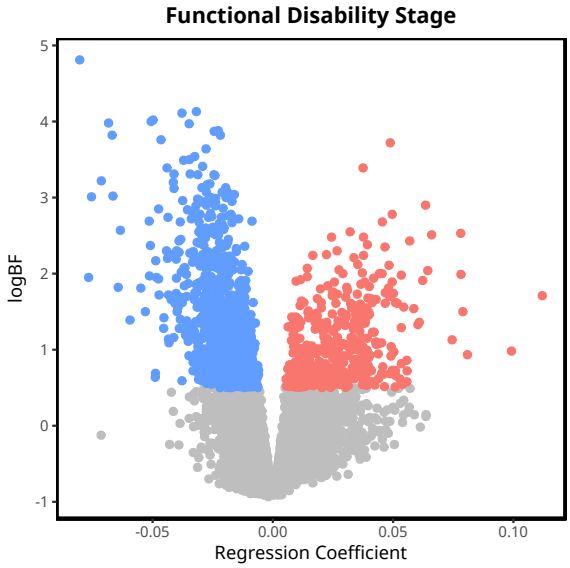

Figure 2-2. Regression of gene expression with functional disability stage (FDS) identifies 1508 genes significantly associated with FDS. Volcano plot of all genes in FDS regression. The regression coefficient is in the x-axis, and the logBF is in the y-axis. Blue indicates a gene with a significant negative regression coefficient (logBF > 0.5), while red indicates a gene with a significant positive coefficient.

Enrichment analysis of DE and FDS-associated genes

We used enrichment analysis to identify biological pathways that were significantly overrepresented in DE or FDS-associated genes (Figure 2-3). In genes that were significantly upregulated in patients compared to carriers and controls, we identified a very strong enrichment for one specific process: neutrophil degranulation (patient vs. control: 58 genes, logBF = **22.2**, patient vs. carrier: 70 genes, logBF = **26.6**). There was weaker enrichment for downregulated genes in general, with the strongest term relating to T-cell differentiation (patient vs. control: 12 genes, logBF = **6.26**, patient vs. carrier: 14 genes, logBF = **6.67**). This enrichment is supported by the presence of numerous T-cell marker genes (*CCR7, CD8A, GZMK, CD3D, CD27*) in the most downregulated genes in patients. These results indicate the presence of subtle but robust changes in peripheral blood gene expression associated with the presence of a pathogenic mutation in FRDA.

Remarkably, enrichment analysis identified the same top term for genes positively associated with FDS: neutrophil degranulation (38 genes, logBF = **5.78**). Negatively associated genes had weaker overall enrichment, which was primarily centered around RNA processing (mRNA splicing: 39 genes, logBF = **4.28**; tRNA modification: 12 genes, logBF = **3.8**; rRNA modification: 32 genes, logBF = **3.25**). Although not significantly enriched, several of the most negatively associated genes (*CD79A*, *GZMB*) are also lymphocyte marker genes, recapitulating the decrease in similar genes seen in differential expression.

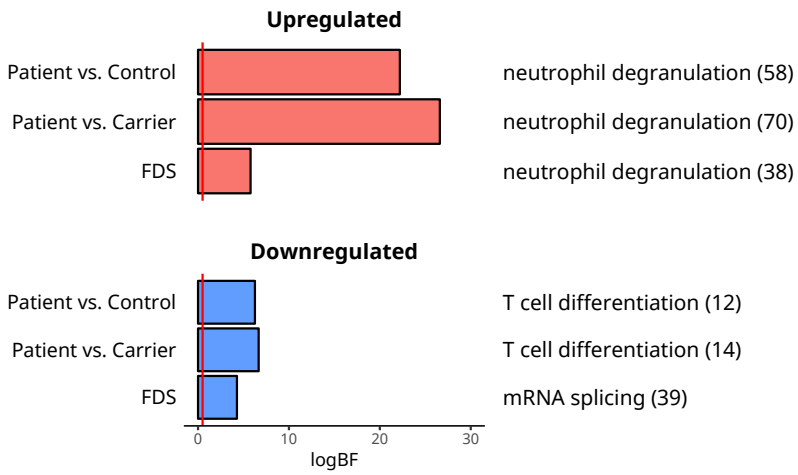

Figure 2-3. Enrichment analysis identifies biological pathways that are significantly overrepresented in differentially expressed and FDS-associated genes. Bar plot of most representative enrichment term for each gene set in the y-axis. The label on the right is the pathway, and the number in parentheses is the size of the overlap between the gene set and pathway. The logBF on the x-axis, and is statistically significant at logBF > 0.5 (marked by red line).

Overlap with other datasets

Gene expression changes associated with frataxin deficiency have previously been studied in a number of models, including transgenic mice, as well as peripheral blood. Two human datasets from peripheral blood (GSE11204, GSE30933, see supplemental text for descriptions of each dataset), and one mouse dataset (RNAi mouse, [8]) were analyzed using the same differential expression workflow used with our data. We considered upregulated (logFC > 0) and downregulated (logFC < 0) transcripts separately (or positive and negative regression coefficients for FDS-associated transcripts) and the overlaps were computed for patients vs. controls, patients vs. carriers, and FDS regression in each direction of change. Thirteen comparisons had a logBF greater than 0.5 (Figure 2-4), indicating that our DE and FDS associated genes were significantly enriched for genes enriched in differential expression in other datasets.

In all cases, the overlap was only observed in the upregulated genes. Six of the enriched comparisons originate from the patient vs. control and carrier vs. control contrasts from a previously published peripheral blood dataset (GSE30933), while the other seven enrichments are seen in DE genes seen in heart tissue collected at several developmental time-

points in a novel mouse model of FRDA [8]. No enrichment was seen for the other previously published peripheral blood dataset (GSE11204). There was also no enrichment observed in DE genes in cerebellum and dorsal root ganglion (DRG) tissue collected from the same mouse model (Figure 2-S3).

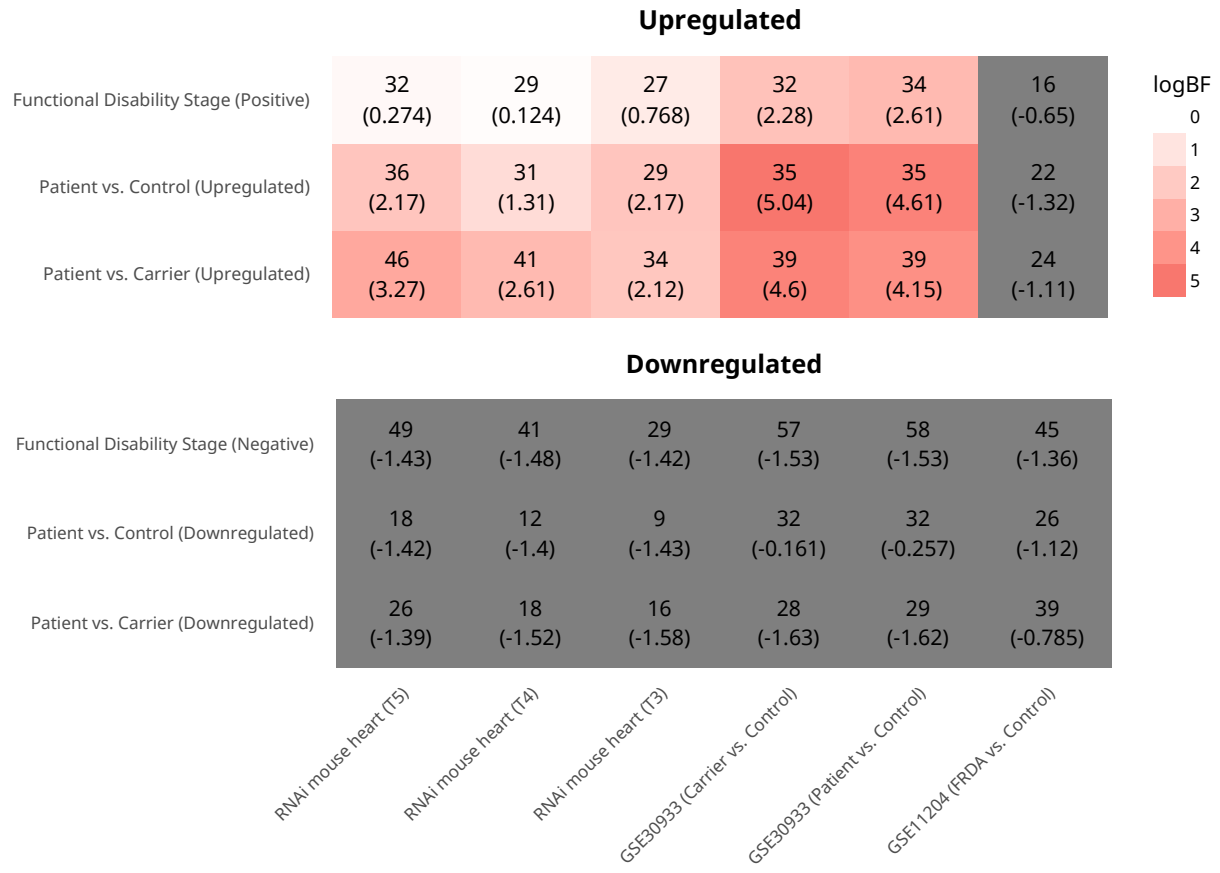

Figure 2-4. Overlap of differentially expressed genes with other datasets. The number in the top of each cell in the heatmap is the number of transcripts in the overlap and the number in parentheses is the logBF of a hypergeometric overlap test. LogBF > 0.5 is considered significant. T3 = 12 weeks old, T4 = 16 weeks old, T5 = 20 weeks old. See supplemental text for additional descriptions of the datasets and analytic procedures.

Weighted Gene Coexpression Network Analysis (WGCNA)

WGCNA is a powerful method for the identification of groups of coexpressed transcripts [9–12]. We first identified modules in our dataset, then used module eigengenes (see Methods) as summary measures for each module to determine if any modules were significantly dif-

ferent across our diagnostic groups, or related to disease progression, using the same linear model designs as in the previous analyses.

Diagnosis

First, we assessed the relationship with diagnostic groups. We identified 7 distinct coexpression modules in the complete dataset (Figure 2-5a-b, Table 2-S4). Three of the seven modules had a logBF > 0.5 for the alternate model compared to the null (Figure 2-5c): pink (logBF = **2.17**), green (logBF = **1.38**), and black (logBF = **1.67**). To identify the specific pairwise differences in the eigengene values, we also computed posterior probabilities and false discovery rates for the contrasts previously described for differential expression (patient vs. control FDR = **0.0015**, patient vs. carrier FDR = **0.0017**). The pink module eigengene was significantly higher in patients than in controls (logFC = **0.011**, pp = **0.994**) and carriers (logFC = **0.012**, pp = **1.0**), while no difference was observed between carriers and controls. The green module eigengene was also higher in patients compared with controls (logFC = **0.012**, pp = **0.998**) and carriers (logFC = **0.009**, pp = **1.0**). Conversely, the black module eigengene was significantly decreased in patients compared to controls (logFC = **-0.008**, pp = **0.970**) and carriers (logFC = **-0.012**, pp = **1.0**).

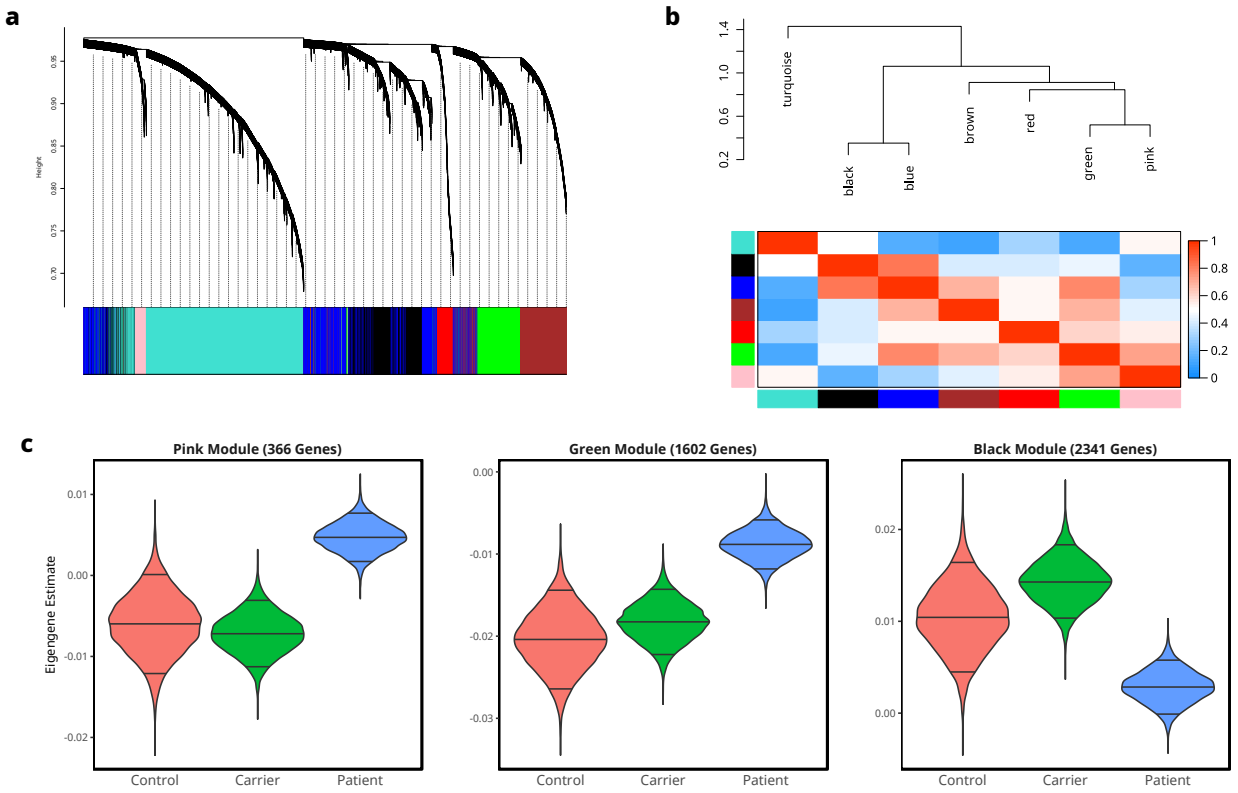

Figure 2-5. WGCNA identifies the pink, green, and black modules as significantly different across clinical status. **a** Cluster dendrogram and color assignment for all transcripts in the full dataset. **b** Cluster dendrogram and heatmap of eigengene correlations. **c** Violin plots showing eigengene posterior estimates for the pink, green, and black modules. The 95% credible intervals are between the smaller top and bottom lines and median estimate is the larger middle line.

Functional disability stage

We also used WGCNA to identify groups of coexpressed genes correlated with FDS. Using the same subset of patients as in the regression with FDS, we identified 8 modules (Fig. 6a-b, Table 2-S5), and used the same linear model designs described for regression with FDS to determine if any eigengenes were significantly associated with FDS. Three modules had a logBF > 0.5 for the alternate model compared to the null (global FDR = **1.0×10^{-4}**, Figure 2-6c): the magenta module (coef. = **0.0071**, logBF = **1.45**, pp = **0.999**), the yellow module (coef. = **-0.0093**, logBF = **3.01**, pp = **1.0**), and the red module (coef. = -**0.0077**, logBF = **1.91**, pp = **1.0**).

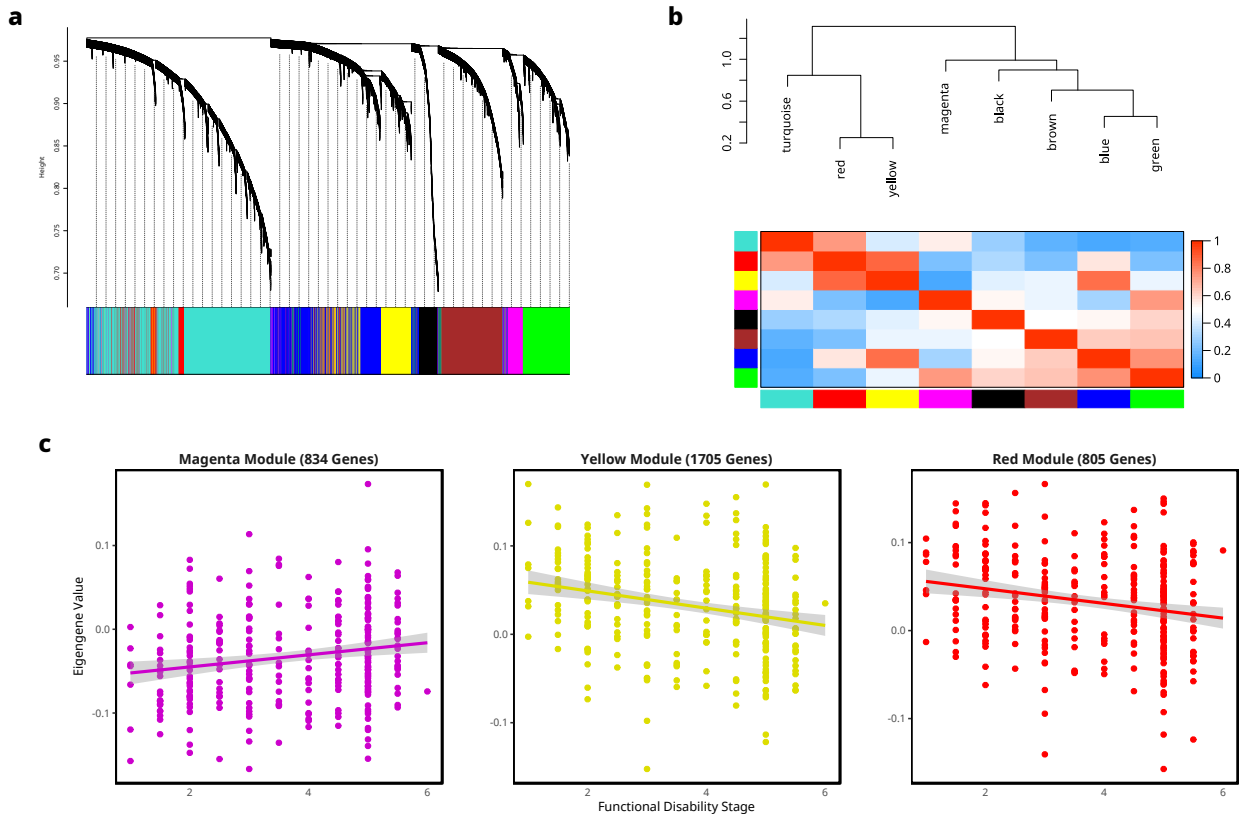

Figure 2-6. WGCNA identifies the magenta, yellow, and red modules as significantly associated with functional disability stage (FDS). a Cluster dendrogram and color assignment for all transcripts in the patients with FDS available. **b** Cluster dendrogram and heatmap of eigengene correlations. **c** Scatterplots showing relationship of FDS in the x-axis with eigengene expression in the y-axis for the magenta, yellow, and red modules.

Enrichment analysis of significant modules

Similar to the approach taken with DE and FDS-associated genes, we used enrichment analysis with Enrichr to identify biological pathways which were overrepresented in our significant WGCNA modules (Figure 2-7). In the status network, the pink module was highly enriched for neutrophil degranulation (43 genes, logBF = **12.0**), the same process seen in upregulated genes in differential expression and genes positively associated with FDS. The green module exhibited even stronger enrichment for neutrophil degranulation (156 genes, logBF = **38.8**). The likely reason the green module is separate from the pink module is that the green eigengene is slightly increased in carriers, while the pink module shows no difference between carriers and controls. Finally, the black module, while showing weaker enrichment overall,

did contain a large number of genes involved in rRNA modification (49 genes, logBF = **1.47**).

In the FDS network, we found that the magenta module was strongly enriched for the same inflammatory response, neutrophil degranulation (75 genes, logBF = **14.2**), as seen in the pink module in the status network. Enrichment analysis of yellow module indicated enrichment for rRNA processing (44 genes, logBF = **2.98**) and the mitochondrial respiratory chain complex (20 genes, logBF = **1.89**). Finally, the red module was strongly enriched for translation, especially mitochondrial translation (27 genes, logBF = **4.44**).

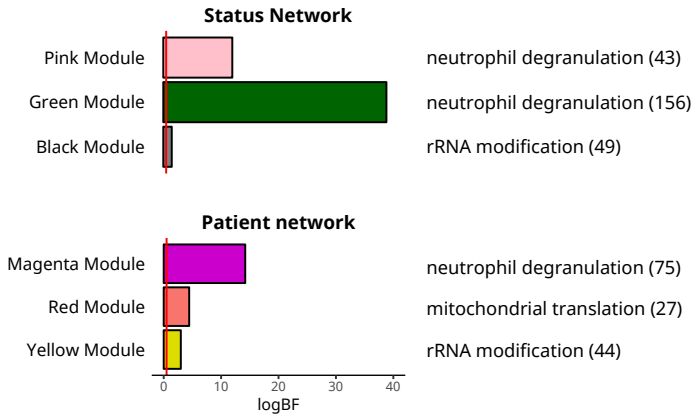

Figure 2-7. Enrichment analysis identifies biological pathways that are significantly overrepresented in WGCNA modules. Bar plot of most representative enrichment term for each gene set in the y-axis. The label on the right is the pathway, and the number in parentheses is the size of the overlap between the gene set and pathway. The logBF on the x-axis, and is statistically significant at logBF > 0.5.

Cell type deconvolution

Changes in cell type composition could in theory explain some of the changes in gene expression we observed between patients, carriers, and controls and with FDS. The availability of cell-type specific transcriptomes in well-studied tissues such as peripheral blood has led to the development of tools to estimate the proportion of cell types in a sample known to contain a mixed population of cells. We used the *CellMix* tool [13] with an existing cell-type specific peripheral blood dataset [14] to estimate cell type proportions in our full dataset (patients, carriers, and controls) and the subset of FRDA patients we used for regression of gene expression with FDS, after regressing out the effects of collinear variables.

After comparing cell type proportions in our 3 disease status groups, only the proportion of natural killer cells was significantly different (logBF = **2.45**, Figure 2-8) and pairwise testing found the proportion underwent small but significant decrease in patients compared to both carriers and controls (patient vs. control: diff. = **-0.0159**, pp = **1.0**; patient vs. carrier: diff. = **-0.0094**, pp = **0.999**). We also regressed cell type proportion with FDS but found no significant associations (Figure 2-S4).

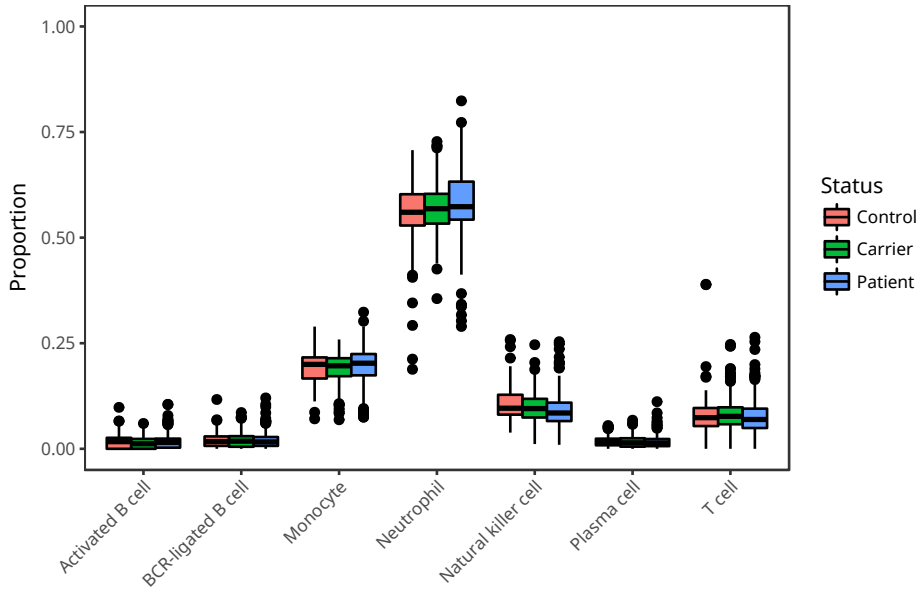

Figure 2-8. Cell type deconvolution analysis. Boxplots showing cell type proportion of 7 cell types in patients, carriers and controls.

qPCR and array validation

We validated our differential expression changes of the top three DE genes between patients and controls, *MMP9*, *DYSF*, and *ANPEP*, using quantitative polymerase chain reaction (qPCR) in 32 patients (including 21 additional samples not previously included in the analysis) and 32 age and sex-matched controls (including 16 new samples, Figure 2-9). We also analyzed the corresponding array data for samples for which this was available (14/32 controls and 22/32 patients). In the qPCR data, there were no significant differences between patients and controls for *MMP9* (p < **0.08,** logFC = **-0.0006**, Mann-Whitney U test), *DYSF* (p < **0.76**, logFC

= **0.022**) or *ANPEP* (p < **0.26**, logFC = **-0.007**). In the corresponding array data, *MMP9* was significantly increased in patients (p < **0.013**, logFC = **0.92**), while *ANPEP* (p < **0.13**, logFC = **0.41**) and *DYSF* (p < **0.34**, logFC = **0.15**) were upregulated but did not reach statistical significance. These results show that we have biologically validated our results with a small number of independent microarrays, but that qPCR is less powered to detect small expression differences between patients and controls, likely because of small sample size and noisier quantification.

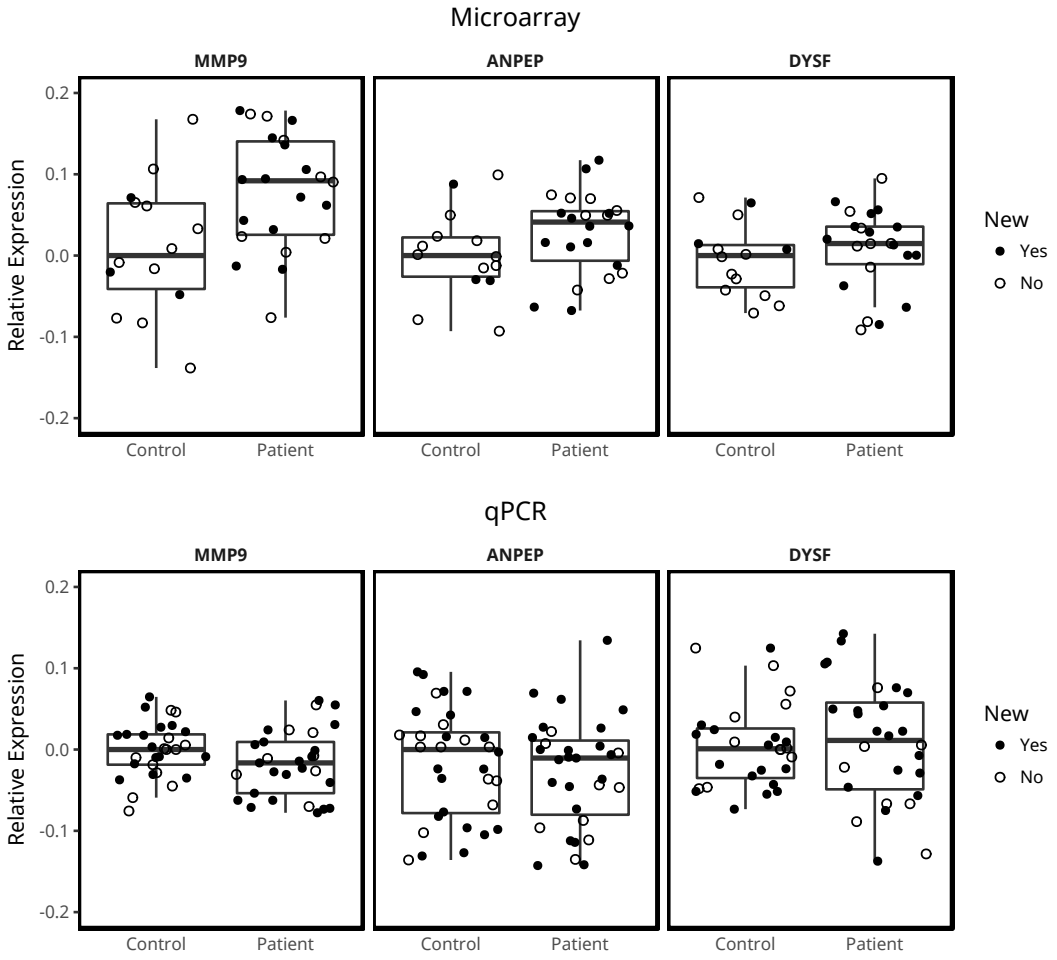

Figure 2-9. qPCR and array validation of top 3 DE genes in 32 patients and 32 age- and sex-matched controls. Boxplots showing the relative expression of the top 3 DE genes to the median value of the control samples. 21 patient and 16 control samples (marked with closed circles) were new and not previously included in the analysis. Top: microarray data, bottom: qPCR.

2.5 Discussion

We report the first large-scale analysis of peripheral gene expression in patients with FRDA, heterozygous mutation carriers, and controls. After conservative data processing and strict statistical thresholds, we identified the transcripts with either robust differences between patients and controls, or correlated with FDS. In addition, network methods identified coordinated groups of genes with biological significance.

The most striking finding across our analyses was the robust enrichment for increased expression in patients of inflammatory genes, particularly those involved in neutrophil degranulation, an important innate response to tissue injury and infection which has also been implicated in chronic inflammation [15]. It is not possible to determine from this data whether the inflammatory response observed peripherally is part of the disease pathogenesis, or merely a response to stress induced by *FXN* deficiency. In other chronic inflammatory disorders, activation of neutrophils and other components of the innate immune response is a key component of the disease [16]. A growing body of literature also supports the involvement of both innate and adaptive immune responses in neurodegeneration, including Parkinson's disease, Alzheimer's disease, and Frontotemporal dementia [17]. Many of our top DE genes and network hub genes are clearly linked to the innate immune response. These include several peptidases (*MMP9*, *ANPEP*, *MME*), a regulator of peptidase activity (*PI3*), two carbonic anhydrases (*CA1*, *CA4*), and genes regulating neutrophil degranulation (*NCF4*, *DYSF*, *STX3*).

We also identified a strong enrichment for a decrease in transcription and translation associated with FRDA, both when comparing patients to controls and carriers, and when examining the relationship with disease severity. Our DE genes and hub genes include *RPL14*, a ribosome component, as well as a chaperone protein (*TTC4*), and an rRNA processing gene (*DDX47*). It has long been known that oxidative stress, like that induced by FXN deficiency, leads to a decrease in translation [18], which may explain these changes.

Several of our DE or hub genes have been identified as being relevant to other neurodegenerative disorders. Both *PROK2* and *AQP9* were identified as being DE in peripheral blood in Huntington's disease [19]. Mutations in dysferlin (*DYSF*) have been identified as a cause of limb-girdle muscular dystrophy [20], and mutations in alpha-synuclein (*SNCA*) have been

identified in familial cases of Parkinson's disease [21]. However, the relevance of these genes to the pathology of FRDA cannot be ascertained from this study.

We observed a fairly consistent overlap with our previous independent peripheral blood study including 41 subjects. The GSE11204 dataset, while also partly collected in peripheral blood (the part of the dataset collected from cell lines was not analyzed because phenotypic data were not available), was severely confounded by batch effect which might explain the poor overlap. The intriguing overlap with genes that are DE in the heart of a novel mouse model for FRDA may indicate there are some similar inflammatory processes occurring in the heart. We speculate that the complete lack of overlap with corresponding CNS tissues (DRG and cerebellum) in the same FRDA mouse model is caused by large differences in structure and function between cells of the CNS and peripheral blood and the smaller number of genes identified as DE in CNS tissues of the mouse model compared to the heart.

The detection of large numbers of genes significantly associated with FDS is intriguing given that this is a high-level clinical measurement and was collinear with age (whose effects were removed from the data before regressing with FDS). Although it is less sensitive than FARS, FDS is easier to collect in large series, and is a fairly direct measurement of disease severity, so it is biologically plausible that genes would be positively or negatively associated with it. We were also intrigued to note that the same inflammatory response which appears to differentiate patients from controls and carriers is also positively associated with disease severity. By contrast, the enrichments seen for downregulated genes in patients and genes inversely associated with severity were generally weaker, though still consistently including transcription.

The relatively poor detection of genes associated with GAA1 or disease duration is likely to due to several issues. Both measures were collinear with age and are not direct measures of disease severity. Furthermore, somatic mosaicism may introduce differences in GAA1 length in blood compared to affected tissues such as the spinal cord, heart, or pancreas.

Although we found enrichment for a number of cell type-specific signatures in our data, cell type deconvolution revealed no change in proportion of neutrophils as estimated from gene expression data, leading us to hypothesize that the large increase in neutrophil degranulation persistently seen across different analyses is not due to an absolute change in

neutrophil count. We only a small decrease in natural killer cells in patients, which may explain the decrease in lymphocyte activation observed in differential expression, although alterations in adaptive immune responses have been observed in neurodegenerative disease [17]. Complete blood cell counts should be used to properly characterize what changes, if any, occur in cell type composition, and cell type-specific transcriptomes, especially of neutrophils, should be generated to identify which genes are undergoing changes in expression in individual blood cell types.

It is also important to recognize the limitations of studying a neurodegenerative disease like FRDA by quantifying gene expression in peripheral blood. The fold changes and regression coefficients with FDS we observed are quite small in magnitude when compared with what is typically observed in model systems and post-mortem studies. Due to the scale of the study, we were not able to control some factors associated with the sample collection that could increase the variability in our gene expression signal, such as fasting, exercise, and the time of day the sample was collected. We cannot determine conclusively whether these factors may have confounded our study but we anticipate they would likely reduce our power to detect effects, making our results more conservative.

The inflammation occurring in FRDA is not an acute response to an infection or a traumatic injury; instead it is likely to be similar to the chronic low-grade inflammation observed in other neurodegenerative and inflammatory disorders [17]. Our large sample size and rigorous correction for potential confounders has provided the statistical power to identify a broad inflammatory signature. No individual gene can fully quantify the inflammatory response and other cellular pathology, but in aggregate these genes provide insights into the effects of *FXN* deficiency. A further strength of our large sample size is that we can capture more of the genetic variation across FRDA patients than is logistically feasible in model systems and post-mortem studies, which makes our results relevant for a broader range of patients.

Future studies of FRDA in humans should characterize the peripheral inflammatory state of FRDA patients, and seek to identify whether this inflammation contributes to the pathology of the disease, or is merely a response to stresses induced by it. In particular, proteomic cytokine profiling and immune cell activity assays, could provide valuable biomarkers beyond gene expression.

2.6 Methods

The full pipeline and code used for all of the analyses is available on Github (https://github.com/coppolalab/FRDA_pipeline) and a summary is provided in this section.

RNA collection and microarray hybridization

Peripheral blood was collected in Paxgene tubes and frozen before RNA extraction, which was performed using a semi-automated system (Qiacube). Subjects were not specifically instructed to fast or refrain from exercise, and the time of collection was not uniform. RNA quantity was assessed with Nanodrop (Nanodrop Technologies) and quality with the Agilent Bioanalyzer (Agilent Technologies), which generated an RNA Integrity number (RIN) for each sample. Total RNA (200 ng) was amplified, biotinylated, and hybridized on Illumina HT12 v4 microarrays, as per manufacturer's protocol, at the UCLA Neuroscience Genomics Core. Slides were scanned using an Illumina BeadStation and signal extracted using the Illumina BeadStudio software (Illumina, San Diego CA).

Array preprocessing

Array preprocessing was performed using the standard pipeline from the *lumi* package [22] which is designed specifically for Illumina microarrays. Raw intensities were normalized using variance-stabilized transformation [23] and interarray normalization was performed with robust spline normalization. 17 outliers were removed from the full dataset and 6 outliers were removed the patient-only dataset using sample-wise connectivity *z*-scores. Batch effect correction was performed using ComBat from the *sva* package [24]. Probes were filtered by detection score and unannotated probes were dropped. Duplicate probes for the same gene were dropped using the maxMean method with the collapseRows function [25] from the *WGCNA* package, which only keeps probes with the highest mean expression across all of the samples. After all probe filtering steps, 16099 probes were used for analysis of the full dataset, and 15198 probes for the patient-only dataset.

Removal of confounding covariates

Age and sex were found to be collinear with disease status (Figure 2-S1). To account for this, the effects of both covariates were fitted and removed using the median posterior estimates from linear models for each gene made with the *BayesFactor* package [26, 27].

Differential expression

Differential expression between patients, carriers, and controls was assessed using Bayesian model comparison on linear models for each gene generated with the *BayesFactor* package [26, 27]. Bayesian model comparison produces Bayes factors instead of p-values for assessing significance. A Bayes factor (BF) is the ratio of the probabilities of two models, and reflects the amount of information gained in terms of variance explained when adding one or more variables to a model. Because age and sex were already removed due to collinearity, only disease status and RIN were available to use as variables. The full model containing the intercept, disease status and RIN was compared to the null model containing only the intercept and RIN. Bayes factors were log-transformed to $\log_{10}$ Bayes factors to place them on a more practical scale [28], and a $\log_{10}$ Bayes factor (logBF) of 0.5 was used as a cutoff for significance of the alternative model to the null model [29]. Although we are fitting a separate model for each gene and thus running thousands of tests, Bayes factors do not require adjustment for multiple comparisons because they are model comparisons [26].

Posterior estimates of the regression coefficients were generated using 10,000 iterations of Monte Carlo Markov chain sampling with a random seed set to 12345 to guarantee reproducibility. We then specified 3 contrasts: patient-control, patient-carrier and carrier-control. For contrast, the posterior samples were subtracted from each other in the order specified to produce an estimate of the difference in expression between the two groups. The median of this estimate was treated as the log fold change (logFC). The posterior probability of the pairwise comparison being in the same direction as the logFC was defined as the number of posterior samples that were non-zero and had the same sign as the logFC.

The Bayesian false discovery rate (FDR) for each pairwise comparison is 1 - posterior probability of the comparison, so we used a posterior probability of 0.95 as our threshold for pair-

wise significance, so that the FDR for individual genes would be less than 5%. The global FDR for a pairwise comparison was computed by taking the mean of the FDR values for all of the genes that were found to be significantly DE for that comparison (adapted from [30, 31]).

Regression with functional disability stage and other phenotypic measures

Several phenotypic measures were available in a large subset of the FRDA patients (n = 308), including functional disability stage (FDS) from the Friedreich's Ataxia Rating Scale (FARS), the shorter of the two GAA repeat expansions (GAA1), and the disease duration (the difference between age of onset and age at draw). Patients that were compound heterozygotes with one loss-of-function *FXN* variant on one allele and a repeat expansion on the other were excluded from this analysis. Age was found to be collinear with all 3 measures (Figure 2-S2) and was removed using the same linear modeling with *BayesFactor* previously described.

Similar to the approach used for differential expression, linear models for each gene were fitted using *BayesFactor*. The full model containing the intercept, the continuous phenotype (FDS, GAA1, or disease duration), sex and RIN, was compared to the null model without the continuous phenotypes and $\log_{10}$ Bayes factors were computed. Posterior estimates of the coefficients were generated using the same parameters described above, and posterior probabilities were defined as the number of samples in an estimate that were non-zero and whose sign was opposite that of the median estimate. The same thresholds of logBF > 0.5 and posterior probability > 0.95 were used to assess significance of the linear relationship between gene expression and the continuous phenotypes, and the global FDR was computed as described for differential expression.

Gene coexpression network analysis

Weighted gene co-expression network analysis (WGCNA) was run on 1) the full set of samples; and 2) the subset of patients with complete phenotypic information described above. Only batch effect was removed using ComBat, as the network construction step must be performed on data that has not any source of biological variation removed. The pipeline from the *WGCNA* package was used as previously reported [9]. A signed network with a soft power

of 6 was generated, and a module dissimilarity threshold of 0.2 was used to merge correlated modules. Hub genes were identified in network modules using scaled connectivity, the ratio of a specific gene's within-module connectivity to the maximum within-module connectivity in that module.

Eigengene values, summarizing gene expression within each module, were compared across disease status using the same linear model approach described for differential expression, with age and sex being regressed out before fitting the final models. Posterior estimates of the model parameters were generated using the same parameters previously described. Similar to the approach used for genes, module eigengenes with a logBF > 0.5 when comparing the alternative model to the null were considered different across conditions, and pairwise comparisons were also considered significant if their 95% credible intervals did not overlap. For regression with continuous phenotypes, the same linear modeling, removal of age effect, and posterior estimation as that described for regression of genes was used with the module eigengenes. An eigengene with logBF > 0.5 was considered to have a significant linear relationship with the continuous phenotype.

Overlap with other datasets

We compared our results to 2 other human datasets from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/): GSE11204 [32] and GSE30933 [33], as well as a dataset generated on a novel mouse model of frataxin deficiency [8]. The same workflow used to identify DE genes in our data was applied to these datasets, with adjustments made to account for platform differences. Complete descriptions of the datasets and analytic procedures are available in the supplemental text. Enrichment was tested using the $\log_{10}$ Bayes Factor computed from a hypergeometric overlap test [34] implemented in *BayesFactor*.

Cell type deconvolution

Cell type deconvolution was performed using the quadratic programming method [35] implemented by the *CellMix* package [13], which provides a peripheral blood dataset [14] that can be used to estimate proportions of cell types in transcriptomic data. Deconvolution was

run on the raw, unprocessed array data as recommended, although outliers were removed so that only the samples used in the final analysis were used to compute cell type proportions. The proportions were separately estimated in the full group of patients, carriers and controls, as well as the subset of patients used for phenotype regression.

The significance of differences in proportions of cell types across patients, carriers, and controls was separately assessed for each cell type using the same Bayesian model comparison and posterior probability estimation described for differential expression. The effects of age and sex were removed by linear regression from the raw expression data before running *CellMix* as described for differential expression, as both variables were confounded with disease status. The significance of regression of FDS with cell type proportion was also determined using the same Bayesian model comparison and posterior probability estimation described for differential expression. The effect of age was removed by linear regression from the raw expression data before running *CellMix* as described for phenotype regression because it was confounded with FDS.

Gene set annotation

Enrichment of genes for specific ontologies and pathways was analyzed using the following datasets downloaded from Enrichr ([36, 37], RRID:SCR_001575): GO Biological Process 2015 (RRID:SCR_002811), GO Molecular Process 2015 (RRID:SCR_002811), KEGG 2016 (RRID:SCR_012773), Reactome 2016 (RRID:SCR_003485). Enrichment scores were computed using a $\log_{10}$ Bayes Factor computed using the same hypergeometric contingency table implemented in *BayesFactor* [34] used for overlap testing.

qPCR validation

Taqman qPCR was used to validate expression changes observed for the top 3 genes, in 32 patients and 32 age- and sex-matched controls. 8/32 (25%) patients and 11/32 (34%) controls were new samples that had not been studied previously, therefore in addition to being a technical validation, this is also partly a biological confirmation of our findings. RNA was converted to cDNA using the Invitrogen Superscript III First-Strand Synthesis System. The

TaqMan TM Gene Expression Assay was then used to detect gene expression in the following three target genes: *MMP9* (Taqman, Hs00957562_m1), *ANPEP* (Taqman, Hs00174265_m1), and *DYSF* (Taqman, Hs01002513_m1). *RPLP0* (Taqman, Hs99999902_m1), *GAPDH* (Taqman, Hs02758991_g1), and β-*Actin* (Applied Biosystems, 4326315E) were used as reference genes. 3 technical replicates for each reaction, resulting in 9 replicates for each biological sample for a total of 576 PCR amplifications. The real-time PCR was carried out on a LightCycler 480 (Roche) instrument and the C_t values were retrieved using the instrument software.

C_t values for the 3 targets genes and 3 reference genes were normalized to a dilution curve as previously described [38] and outliers were identified and removed in two steps. First, data were standardized by subtracting the mean and dividing by the median absolute deviation for each pair target and reference genes separately (i.e. only *MMP9* with *RPLP0* as reference). Any reaction with a standardized score with absolute value great than 2 was excluded, resulting in a total of 44/576 *MMP9* reactions, 43/576 *ANPEP* reactions and 50/475 *DYSF* reactions being excluded. After removing these outliers, the median value across all remaining technical replicates for each gene in each subject was computed. Median expression values per subject were again standardized by median and MAD and any subject whose standardized score had an absolute value greater than 2 was excluded. This resulted in 6 subjects being excluded for *MMP9*, 1 subjects for *ANPEP*, and 8 subjects being excluded for *DYSF*. The significance of the difference in expression between patients and controls for each gene was assessed using the Mann-Whitney U test because the expression values were not normally distributed. Data from corresponding arrays was processed using the same array preprocessing pipeline previously described, except that age and sex were not regressed out because they were no longer confounded with disease status.

To maintain consistency with the qPCR analysis, the Mann-Whitney U test was also used to assess the significance of the differences between patients and controls for each gene in the array data. For 9 patients and 11 controls, the array used was from a different time point than the one analyzed in the original DE analysis, providing both technical and biological validation for those subjects.

Data Availability

All raw gene expression data is available for download in NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/gds) under accession number GSE102008. An interactive differential expression analysis interface for the data is available in the REPAIR database (https://coppolalab.ucla.edu/account/login). Finally, interactive visualizations of our network analysis are available on our website (https://coppolalab.ucla.edu/gclabapps/nb/browser?id=FRDA_Gene%20Expression%20Network%20-%20Diagnosis;ver=, https://coppolalab.ucla.edu/gclabapps/nb/browser?id=FRDA_Gene%20Expression%20Network%20-%20FDS;ver=).

Ethics Statement

Protocols for acquisition of data from subjects were approved by the Institutional Review Boards of UCLA and Children's Hospital of Philadelphia, and consent for data to be used was obtained from subjects or the appropriate legal guardian.

2.7 Description of analysis of other datasets

GSE11204

This dataset [32] consists of microarray data run on the Agilent-012097 Human 1A Microarray (V2) G4110B from cell lines or peripheral blood from FRDA patients and controls. We only chose to analyze the peripheral blood because the cell line data were missing important phenotypic information such as sex. Since the controls and patients were run in separate batches, the dataset is fully confounded by batch effect. Preprocessing of raw array data downloaded from GEO was completed using workflow provided by *limma*. Background correction was performed using the normexp method, within array normalization used the loess method, and between array normalization used quantile normalization. Duplicate probes were collapsed using *collapseRows* from the *WGCNA* package. Differentially expressed transcripts were identified using the same Bayesian model comparison described for the main

dataset, with the full model containing disease status, age, and sex, and the null model containing only age and sex.

GSE30933

This dataset [33] consists of microarray data run on the Illumina HumanRef-v8.0 platform from peripheral blood in 10 FRDA patients, 10 carriers and 9 controls. Preprocessing was completed using a modified version of the workflow used to preprocess the main dataset. Raw data was log2-transformed because some QC columns were unavailable. Inter-array normalization was performed using robust spline normalization. Duplicate probes were collapsed using *collapseRows* from the *WGCNA* package. Differentially expressed transcripts were identified using the same Bayesian model comparison described for the main dataset, with the full model containing disease status and the null model containing only the intercept, as no other covariates were available.

RNAi mouse

This dataset [8] consists of microarray data run on the Illumina MouseRef-8 v2.0 platform from a novel RNAi-based mouse model of FRDA. Two controls were provided: the RNAi transgenic mouse given no doxycycline, and a wildtype mouse with no RNAi construct given doxycycline (to test the effects of doxycycline exposure on their own). For all groups except the rescue, animals were sacrificed at 5 time points (0, 3, 12, 16 and 20 weeks of age), with 12 replicates for each disease/control group at each time point. The preprocessing pipeline was identical to that used in the main dataset except that mouse annotation was used instead of human. Mouse gene symbols were converted to HomoloGene IDs that could be directly compared with HomoloGene IDs for human gene symbols.

We focused on the disease vs. control comparisons. Differentially expressed transcripts were identified using the same Bayesian model comparison described for the main dataset, with the full model containing genotype, drug treatment, timepoint, the three 2-way interactions (genotype x drug treatment, genotype x timepoint, treatment x timepoint), and the 3-way interaction(genotype x drug treatment x timepoint), as well as age, sex, weight

and RIN as covariates. The null model contained only timepoint and four covariates previously described. Model-level significance was determined using logBF > 0.5 as for the other datasets. Pairwise significance at each timepoint for the joint comparison of the transgenic DOX-treated mouse vs. the transgenic mouse given no drug and the transgenic DOX-treated mouse vs. the DOX-treated wild type mouse was determined by computing the joint posterior probability of both comparisons have the same sign as each other and their respective median parameter estimates.

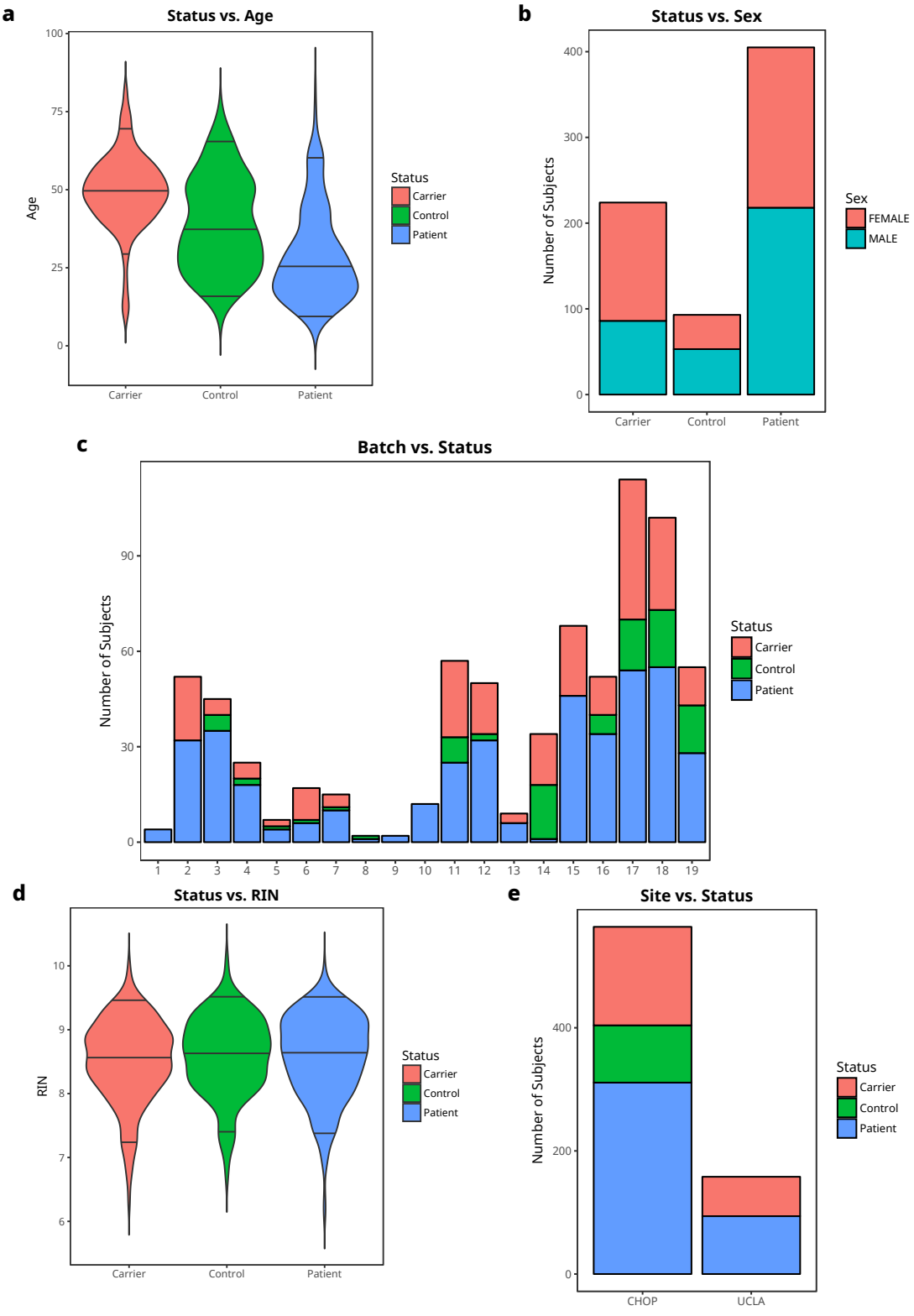

Figure 2-S1. Clinical status is significantly associated with age, sex, batch and RNA integrity number (RIN). Diagnostic plots showing the relationship between clinical status and **a** age, **b** sex, **c** batch, **d** RIN, and **e** site.

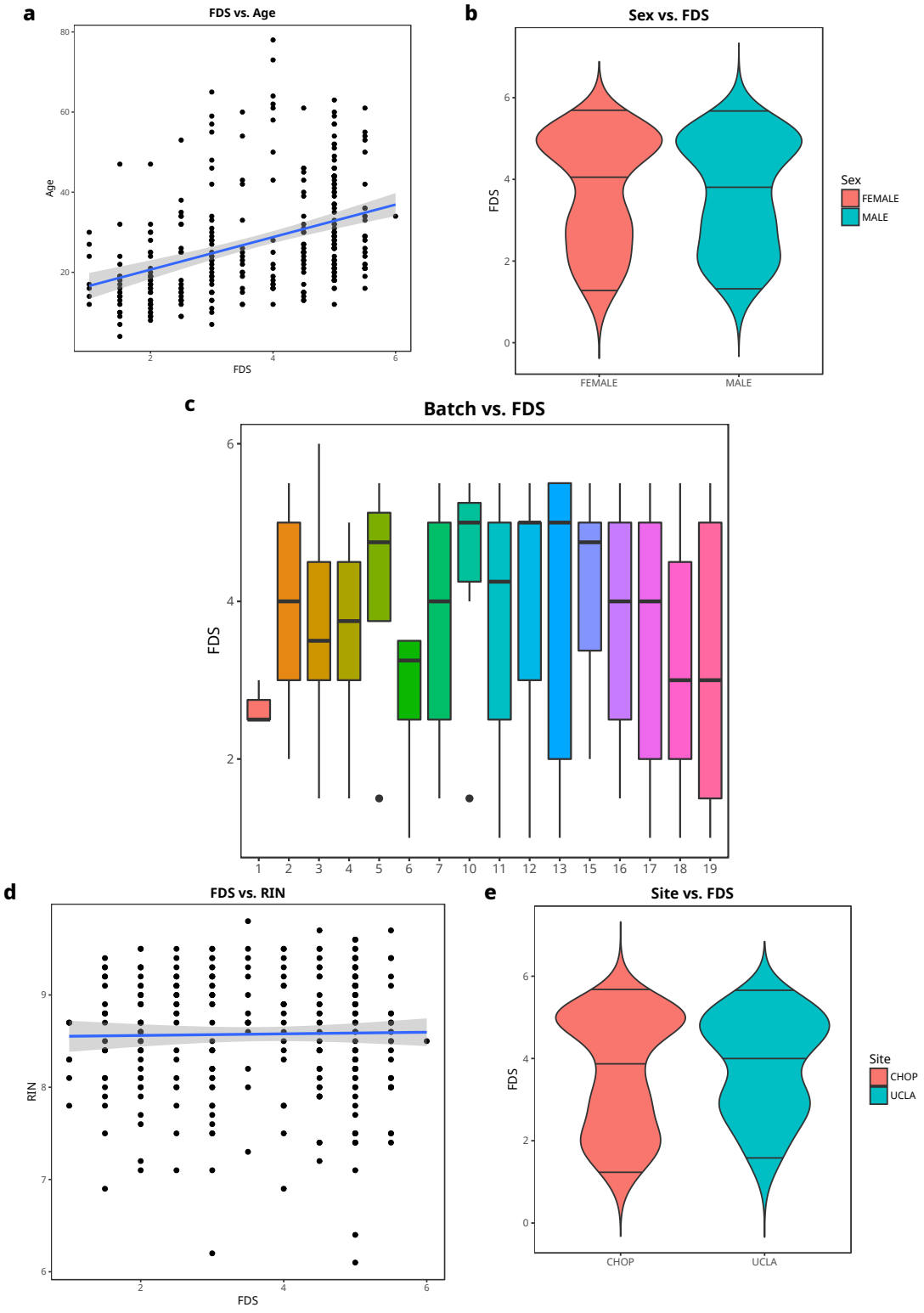

Figure 2-S2. Functional disability stage is significantly associated with age. Diagnostic plots showing the relationship between functional disability stage and **a** age, **b** sex, **c** batch, **d** RIN, and **e** site.

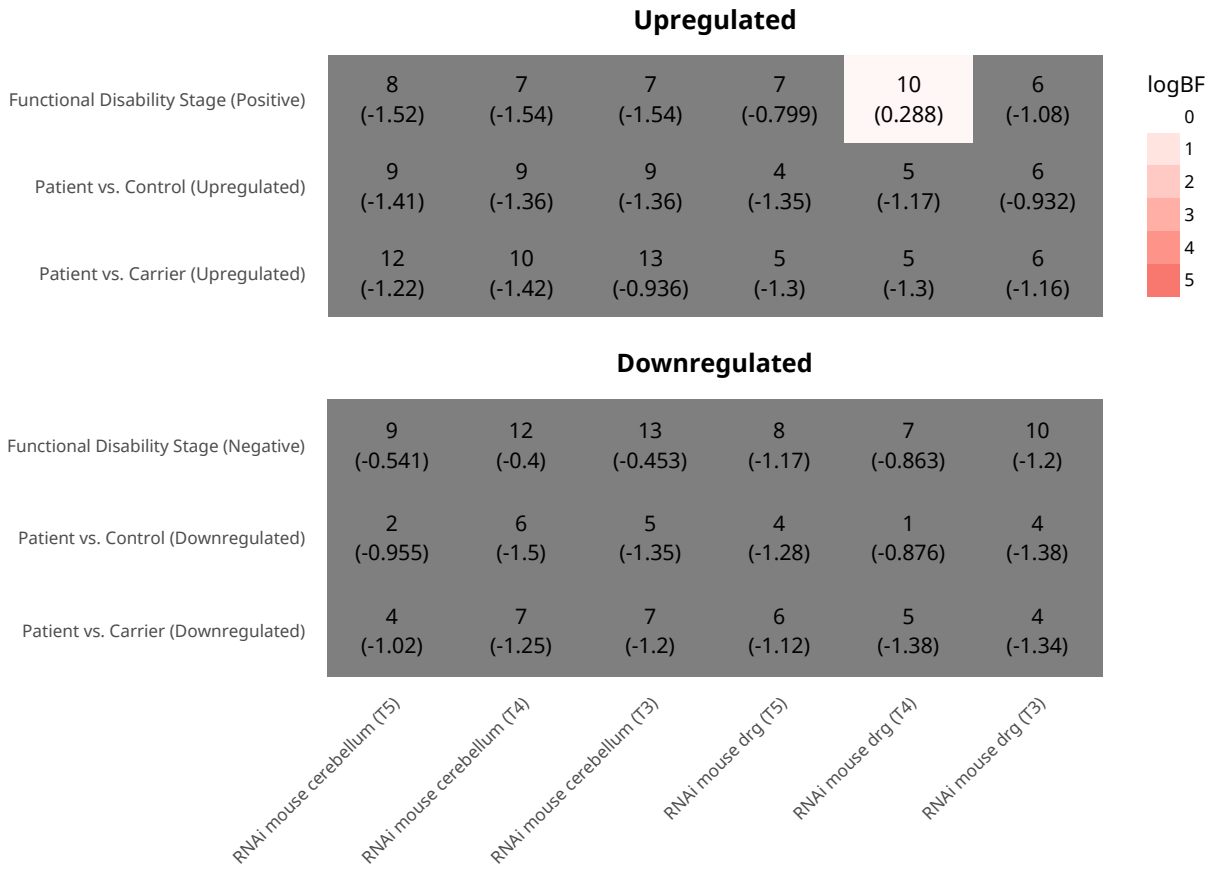

Figure 2-S3. No overlap in differentially expressed genes is observed with DRG and cerebellum in RNAi mouse. Heatmaps showing the overlap of up- and down-regulated transcripts in our datasets with the tissues in the RNAi mouse. The number in the top of each cell is the number of transcripts in the overlap and the number in parentheses is the logBF of a hypergeometric overlap test. LogBF > 0.5 is considered significant.

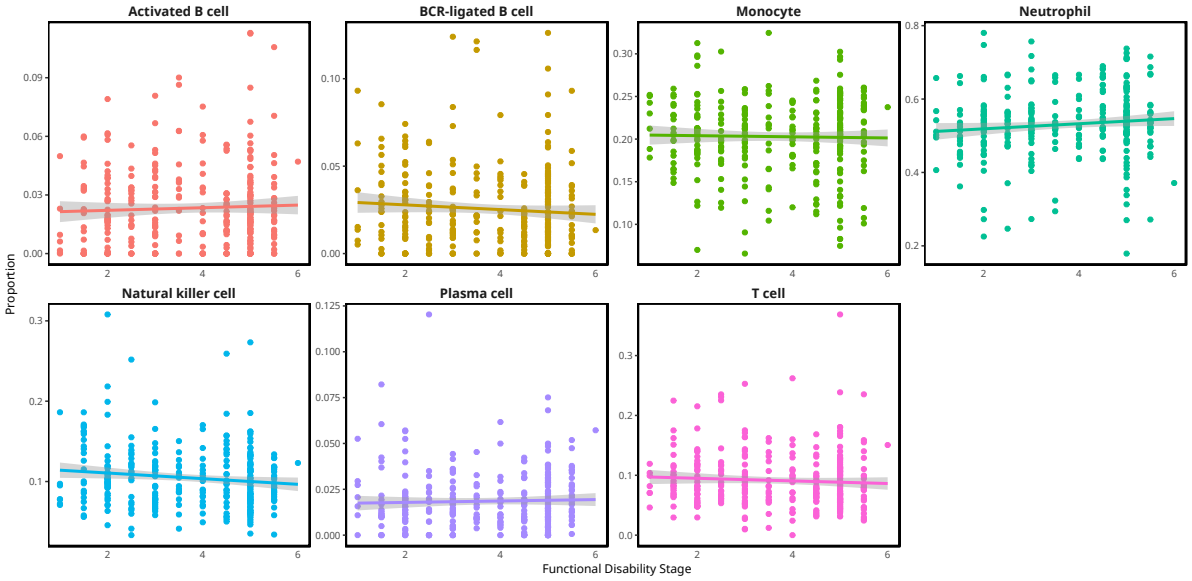

Figure 2-S4. Functional disability stage (FDS) is not associated with cell type proportions. Scatterplots showing functional disability stage vs. cell type proportion.

Table 2-S1. Table of point mutations in FRDA patients.

Table 2-S2. Annotated tables of differential expression for all genes in the transcriptome, and lists of significantly DE genes when comparing patients and controls (n=829), patients and carriers (n=1078) and carriers and controls (n=182).

Table 2-S3. Annotated table of regression with functional disability stage (FDS), GAA1, and disease duration for all genes in the transcriptome, and lists of significantly associated with FDS (n=1508), GAA1 (n=280), and disease duration (n=13).

Table 2-S4. Annotated table of network statistics for all genes in the transcriptome in the diagnosis co-expression network. kTotal = total connectivity, kWithin = within module connectivity, kOut = connectivity outside module, kDiff = kWithin - kOut, kscaled = scaled within module connectivity, MM = module membership (correlation with module eigengene).

Table 2-S5. Annotated table of network statistics for all genes in the transcriptome in the FDS co-expression network. kTotal = total connectivity, kWithin = within module connectivity, kOut = connectivity outside module, kDiff = kWithin - kOut, kscaled = scaled within module connectivity, MM = module membership (correlation with module eigengene).

Bibliography

1. Cossée, M. *et al.* Inactivation of the Friedreich ataxia mouse gene leads to early embryonic lethality without iron accumulation. *Hum. Mol. Genet.* **9,** 1219–1226 (2000).
2. Gottesfeld, J. M., Rusche, J. R. & Pandolfo, M. Increasing frataxin gene expression with histone deacetylase inhibitors as a therapeutic approach for Friedreich's ataxia. *J. Neurochem.* **126,** 147–154 (2013).
3. Pastore, A. & Puccio, H. Frataxin: A protein in search for a function. *J. Neurochem.* **126,** 43–52 (2013).
4. Cnop, M., Mulder, H. & Igoillo-Esteve, M. Diabetes in Friedreich ataxia. *J. Neurochem.* **126,** 94–102 (2013).
5. Campuzano, V. *et al.* Triplet Repeat Expansion Friedreich ' s Ataxia : Autosomal Recessive Disease Caused by an lntronic GAA Triplet Repeat Expansion. *Science* **271,** 1423–1427 (1996).
6. Lazaropoulos, M. *et al.* Frataxin levels in peripheral tissue in Friedreich ataxia. en. *Ann Clin Transl Neurol* **2,** 831–842 (Aug. 2015).
7. Bürk, K., Schulz, S. R. & Schulz, J. B. Monitoring progression in Friedreich ataxia (FRDA): The use of clinical scales. *J. Neurochem.* **126,** 118–124 (2013).
8. Chandran, V. *et al.* Inducible and reversible phenotypes in a novel mouse model of Friedreich's Ataxia. en. *eLife Sciences* **6,** e30054 (Dec. 2017).
9. Langfelder, P. *et al.* Integrated genomics and proteomics define huntingtin CAG length–dependent networks in mice. *Nat. Neurosci.* **19,** 623–633 (2016).

10. Seyfried, N. T. *et al.* A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. en. *Cell Syst* **4,** 60–72.e4 (Jan. 2017).

11. Wu, Y. E., Parikshak, N. N., Belgard, T. G. & Geschwind, D. H. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. en. *Nat. Neurosci.* **19,** 1463–1476 (Nov. 2016).

12. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. en. *Nature* **540,** 423–427 (Dec. 2016).

13. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. en. *Bioinformatics* **29,** 2211–2212 (Sept. 2013).

14. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. en. *PLoS One* **4,** e6098 (July 2009).

15. Caielli, S., Banchereau, J. & Pascual, V. Neutrophils come of age in chronic inflammation. en. *Curr. Opin. Immunol.* **24,** 671–677 (Dec. 2012).

16. Gernez, Y., Tirouvanziam, R. & Chanez, P. Neutrophils in chronic inflammatory airway diseases: can we target them and how? en. *Eur. Respir. J.* **35,** 467–469 (Mar. 2010).

17. Amor, S. *et al.* Inflammation in neurodegenerative diseases–an update. en. *Immunology* **142,** 151–166 (June 2014).

18. Shenton, D. *et al.* Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. en. *J. Biol. Chem.* **281,** 29011–29021 (Sept. 2006).

19. Mastrokolias, A. *et al.* Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. en. *Eur. J. Hum. Genet.* **23,** 1349–1356 (Jan. 2015).

20. Liu, J. *et al.* Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy. en. *Nat. Genet.* **20,** 31–36 (Sept. 1998).

21. Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. en. *Science* **276,** 2045–2047 (June 1997).

22. Du, P., Kibbe, W. A. & Lin, S. M. lumi: A pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–1548 (2008).

23. Lin, S. M., Du, P., Huber, W. & Kibbe, W. A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* **36,** 1–9 (2008).

24. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).

25. Miller, J. a. *et al.* Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics* **12,** 322 (2011).

26. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56,** 356–374 (2012).

27. Rouder, J. N. & Morey, R. D. Default Bayes Factors for Model Selection in Regression. *Multivariate Behav. Res.* **47,** 877–903 (2012).

28. Jeffreys, H. *Theory of probability* (Oxford University Press, 1961).

29. Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90,** 773–795 (1995).

30. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc.* **96,** 1151–1160 (2001).

31. Efron, B. Microarrays, Empirical Bayes and the Two-Groups Model. *Stat. Sci.* **23,** 1–22 (2008).

32. Haugen, A. C. *et al.* Altered gene expression and DNA damage in peripheral blood cells from Friedreich's ataxia patients: Cellular model of pathology. *PLoS Genet.* **6** (2010).

33. Coppola, G. *et al.* A gene expression phenotype in lymphocytes from friedreich ataxia patients. *Ann. Neurol.* **70,** 790–804 (2011).

34. Jamil, T. *et al.* Default "Gunel and Dickey" Bayes factors for contingency tables. en. *Behav Res* **49,** 638–652 (Apr. 2017).

35. Gong, T. *et al.* Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. en. *PLoS One* **6,** e27156 (Nov. 2011).

36. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (2013).

37. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–W97 (July 2016).

38. Bustin, S. A. *et al.* The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clin. Chem.* **55,** 611–622 (Apr. 2009).

Chapter 3

Dementia

3.1 Introduction

Neurodegenerative forms of dementia constitute a heterogeneous group of diseases characterized by progressive loss of cognitive function driven by dysfunction and loss of neurons and glia in the central nervous system (CNS). Clinicians categorize dementias by the type of cognitive dysfunction observed in patients, such as memory loss, speech difficulties, or decline in executive function [1]. Pathological studies have identified specific proteinopathies such as accumulation of amyloid beta, MAPT, TDP-43, FUS, and SNCA in certain brain regions which have some association with clinical symptoms [2, 3]. Genetic studies have also implicated both rare and common variants in risk for dementia, some of which may be linked to associated proteinopathies. Notably, several genes harboring disease-associated variants, such as *TREM2* [4] and *GRN* [2], and others [5] are not directly associated with proteinopathy, but rather are involved in inflammatory response and innate immunity and are primarily expressed in microglia, implicating inflammation in the CNS as relevant to dementia pathophysiology [6].

Comparatively little is known about whether peripheral inflammation, in particular inflammation mediated by white blood cells (WBCs), is altered in dementia patients. Previous studies, often including small numbers of samples, have reported some evidence of increased peripheral inflammation in the most common form of dementia, Alzheimer's disease (AD) [7], as well as Parkinson's disease, the most common neurodegenerative motor disorder [8], while

it is unknown whether peripheral inflammation is altered in the spectrum of disorders associated with frontotemporal dementia (FTD), including behavioral variant FTD (bvFTD), semantic variant and nonfluent variant primary progressive aphasia (svPPA and nfvPPA), progressive supranuclear palsy (PSP), and corticobasal syndrome (CBS).

We collected peripheral blood transcriptomics and methylation data in a large sample of patients with AD, mild cognitive impairment (MCI), FTD spectrum disorders, and healthy controls, to determine whether a signal associated with disease is detectable in peripheral blood in dementia patients. We found evidence of a consistent increase in a transcriptomic inflammatory innate immune response in neutrophils and monocytes in AD, PSP, and nfvPPA, and a sex-specific response in MCI. This inflammatory response gene set was significantly enriched for genetic risk for AD and genes expressed in microglia. We also show that this innate immune response could not be identified in methylation data and that it was not a result of changes in cell type composition.

3.2 Results

Peripheral blood was collected from 1387 individuals for gene expression analysis and 664 individuals for DNA methylation analysis. Gene expression was quantified using Illumina HT12 v4 microarrays, and we quantified DNA methylation genome-wide using Illumina HumanMethylation450k microarrays. We only considered subjects with unambiguous diagnoses of control, AD, MCI, bvFTD, svPPA, nfvPPA, PSP or CBS, and also removed RNA samples with RIN < 6.0. After filtering by these criteria, we had 1044 RNA samples (283 control, 299 AD, 193 MCI, 85 bvFTD, 54 PSP, 47 nfvPPA, 45 svPPA, 38 CBS) and 605 DNA methylation samples (289 control, 144 AD, 22 MCI, 45 bvFTD, 38 PSP, 20 nfvPPA, 47 svPPA).

Differential Expression

AD and MCI

We used Bayesian linear modeling (see Methods) to identify transcripts which were differentially expressed (DE) between controls and patients with AD and MCI. We identified a signif-

icant confound between age and diagnosis in the full dataset which we resolved by stratifying our samples so that only subjects between the ages of 60 and 90 were included. After this stratification process as well as the removal of expression outliers, we had 229 control samples, 198 AD samples, and 124 MCI samples. At A significance cutoff of logBF > 0.5 and posterior probability > 0.95, we identified 444 DE genes between AD and controls (global FDR = **0.0073**, Figure 3-1a), 451 between MCI and controls (global FDR = **0.0057**), and 280 between AD and MCI (global FDR = **0.012**). Enrichment analysis of differentially expressed genes (Figure 3-1b) revealed a significant enrichment for neutrophil degranulation (GO:0043312), a key component of the innate immune response, in upregulated genes in AD vs. Control (39 genes, logBF = **11.20**) and MCI vs. Control (28 genes, logBF = **4.31**), and a weaker enrichment in AD vs. MCI (13 genes, logBF = **1.91**).

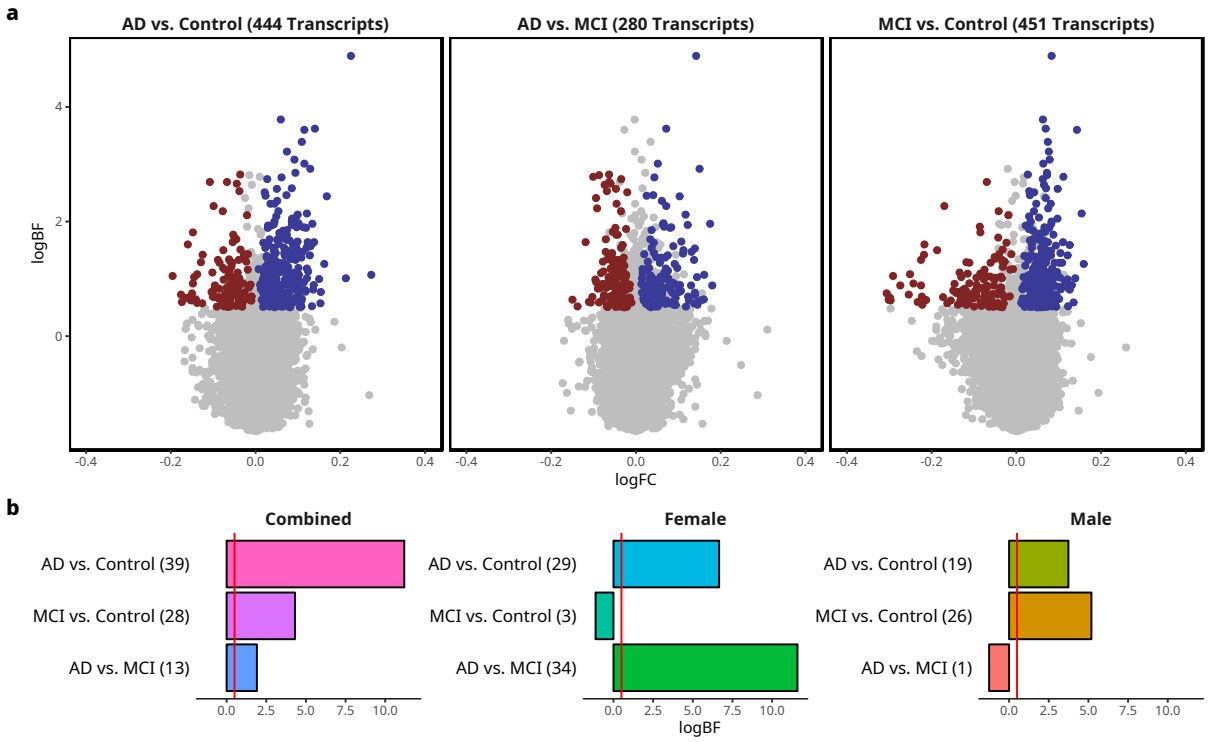

Figure 3-1. **a** Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles. **b** Bar plots of enrichment of significantly upregulated genes for neutrophil degranulation (GO:0043312) with the logBF on the x-axis.

Sex differences in AD and MCI

Studies of the epidemiology of dementia, particularly AD, have identified sex differences in disease risk including consistent evidence that females have a higher risk of developing AD [9]. We partitioned our samples by sex to determine if there were differences in the inflammatory response we observed in the full dataset in males and females. Each dataset was preprocessed separately, resulting in 251 male samples (94 controls, 93 AD, 64 MCI) and 269 female samples (123 controls, 94 AD, 52 MCI). The number of DE genes in both females (AD vs. Control: 355 genes, global FDR = **0.0076**; MCI vs. Control: 293 genes, global FDR = **0.011**; AD vs. MCI: 468 genes, global FDR = **0.0055**, Figure 3-S1a) and males (AD vs. Control: 280 genes, global FDR = **0.0076**; MCI vs. Control: 393 genes, global FDR = **0.011**; AD vs. MCI: 157 genes, global FDR = **0.0055**, Figure 3-S1b) was comparable to the analysis on the full dataset. However, the enrichment in neutrophil degranulation (Figure 3-1b) in upregulated genes in MCI vs. Control was observed in males (26 genes, logBF = **5.19**) but not in females MCI vs. Control (3 genes, logBF = **-1.13**). This indicates that, while males and females MCI vs. controls show similar numbers of DE genes, male MCI subjects exhibit an inflammatory response in peripheral blood similar to that in AD subjects, whereas an inflammatory response is not detectable in female MCI subjects.

FTD disorders

We applied the same linear modeling approach to identify transcripts which were differentially expressed between FTD disorders and control (268 controls, 75 bvFTD, 53 PSP, 45 nfvPPA, 43 svPPA, 35 CBS). A significant age confound was also observed with diagnosis in this subset of the data, and it could not be resolved using stratification and was instead removed using residualization (see Methods). Using the significance cutoffs previously described, we identified 175 DE genes in bvFTD vs. control (global FDR = **0.013**, Figure 3-2a), 189 genes in nfvPPA vs. control (global FDR = **0.016**), 81 genes in svPPA vs. control (global FDR = **0.02**), 257 genes in PSP vs. control (global FDR = **0.0073**, Figure 3-2b) and 48 genes in CBS vs. control (global FDR = **0.023**). Enrichment analysis of upregulated genes in each disease vs. control (Figure 3-S2) revealed significant enrichment for neutrophil degranulation in PSP vs. control

(18 genes, logBF = **2.08**) and nominally significant enrichment in bvFTD vs. control (6 genes, logBF = **1.04**), but no enrichment for the other disorders. We did not attempt to partition our FTD disorder data by sex because our samples sizes for these disorders were much lower than for AD and MCI.

We next asked how similar the transcriptomic effects of each FTD-spectrum disorder were to the other disorders. To visualize this, we correlated the logFC values of disease vs. control for all genes in each FTD disorder with the corresponding logFC values in each of the other disorders and clustered the diseases based on these correlation values (Figure 3-2c). We found that nfvPPA and PSP, both known to be tauopathies [10, 11] clustered with each other and away from the other categories, suggesting the presence of a tau-related signal.

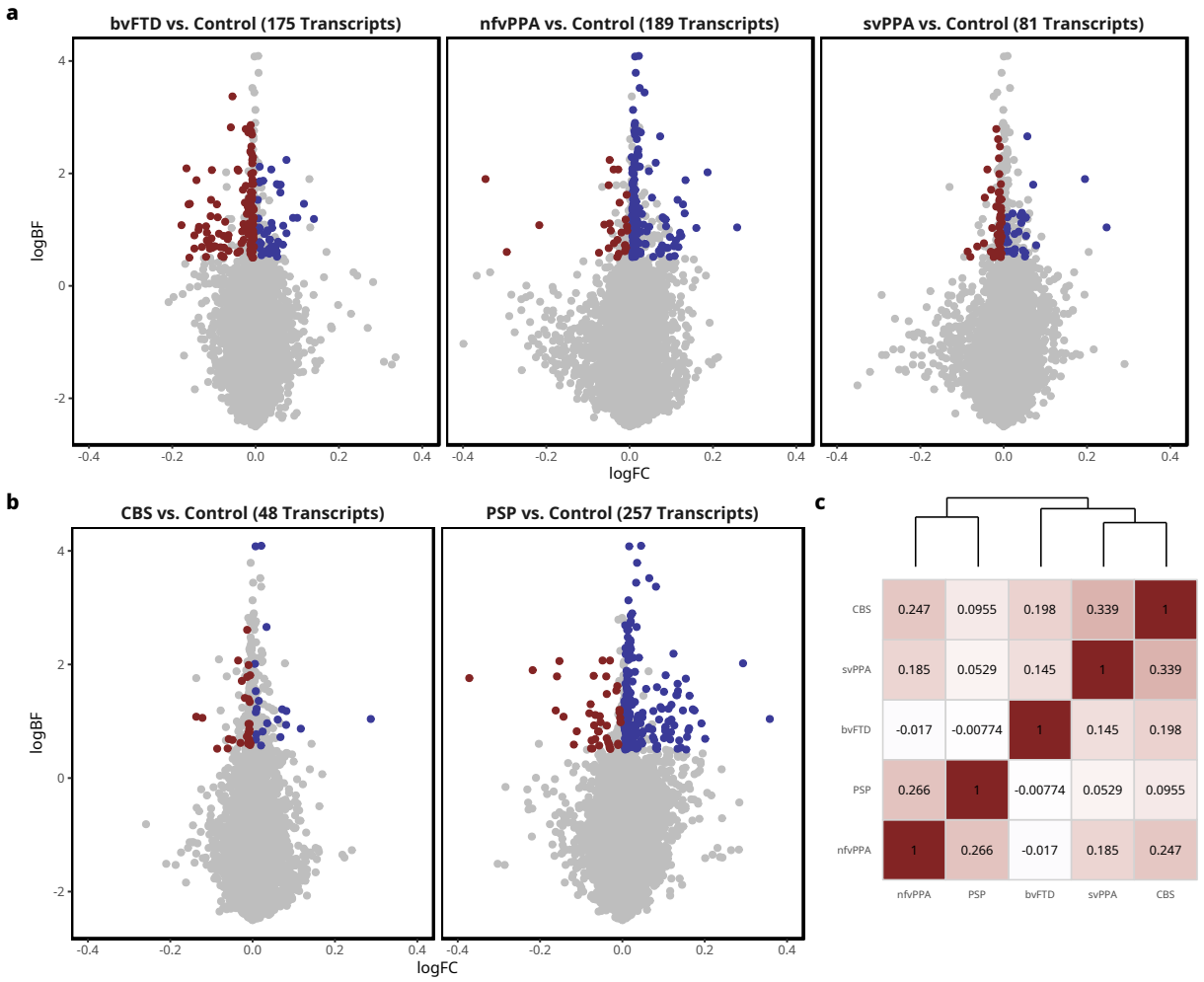

Figure 3-2. **a,b** Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles. **c** Correlation plot of the pairwise correlation of logFC values of FTD disorders vs. control, with the dendrogram on top showing hierarchical clustering of disorders.

ApoE genotype

The E4 allele in the apolipoprotein E gene (*APOE*) is the strongest known common variant risk for AD [12] and homozygous E4 carriers have an odds ratio of 11-13 for developing AD compared to homozygous E3 carriers, while the rare E2 allele is protective for AD [13]. We genotyped 525 of the samples in our cohort and identified 196 E4 carriers and 329 E4 non-carriers, and found no confound between *APOE* genotype and age or sex. We identified 737 DE genes (global FDR = **0.0020**) between E4 carriers and non-carriers (Figure 3-S3), with marginally sig-

nificant enrichment for neutrophil degranulation in upregulated genes (27 genes, logBF = **0.412**). Because there are many controls and MCI subjects who carry the E4 allele, we also restricted our comparison of E4 carriers and non-carriers to AD patients (78 E4 carriers, 104 E4 non-carriers). We identified 371 DE genes (global FDR = **0.0040**, Figure 3-S3) between E4 carriers and non-carriers, also with marginally significant enrichment for neutrophil degranulation (logBF = **0.431**).

Network Analysis

Our differential expression results showed that there was an enrichment for an innate immune transcriptional response, and also pointed to a possible effect of sex on MCI vs. control gene expression signatures. We used weighted gene co-expression network analysis (WGCNA) to identify clusters of co-expressed transcripts (also known as modules) which are often highly enriched for specific biological pathways [14–20]. The gene expression pattern in a module across samples can be summarized by the first principal component of the expression values of all the genes in given module, or eigengene. These eigengenes were analyzed using the same linear modeling approach used for differential expression, with the same significance cutoffs. We used this method on the same data subsets analyzed with differential expression to better understand the innate immune response signature we identified throughout our analyses.

AD and MCI

WGCNA identified 18 modules of co-expressed genes in the full dataset of AD, MCI, and control subjects. Disease status was a significant predictor for the magenta (logBF = **1.28**, Figure 3-S4a) and brown (logBF = **0.723**, Figure 3-3a) modules. Pairwise comparisons for the magenta module showed a significant increase in AD vs. control (diff. = **0.0085**, pp. = **0.982**) and MCI vs. control (diff. = **0.017**, pp. = **1.0**) and a marginally significant decrease in AD vs. MCI (diff. = -**0.0081**, pp. = **0.958**). In the brown module (Figure 3-3a), AD (diff. = **0.012,** pp = **0.998**) and MCI (diff. = **0.011**, pp = **0.991**) were significantly increased vs. Control, while no significant difference was observed between AD and MCI (diff. = **0.0012**, pp = **0.595**). Both

modules showed significant enrichment for neutrophil degranulation (magenta: 45 genes, logBF = **0.67**, Figure 3-S4e; brown: 123 genes, logBF = **25.2**, Figure 3-3d) but clearly showed different expression patterns with regards to MCI.

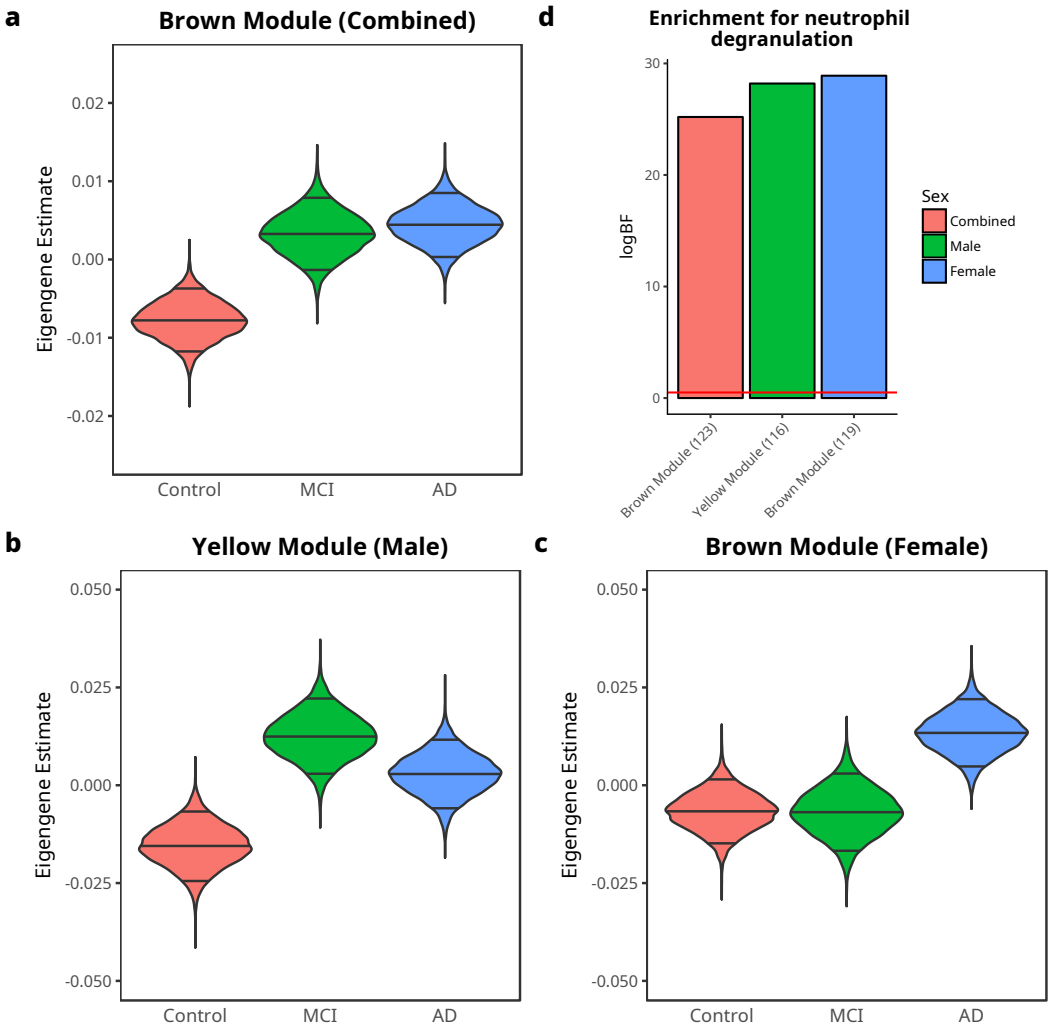

Figure 3-3. a-c, Violin plots of posterior estimate of mean eigengene values for each diagnosis, with the median and 5% and 95% quantiles indicated by lines. **c** Bar plot of enrichment of genes in each module for neutrophil degranulation (GO:0043312) with the logBF on the x-axis.

Since our previous analyses pointed to a possible effect of sex in expression patterns, we regenerated our co-expression network in males and females separately. In the male samples, we identified two out of 19 modules where diagnosis had a significant effect: the green module (logBF = **0.945**, Figure 3-S4b) and the yellow module (logBF = **0.619**, Figure 3-3b). The pairwise comparisons for the green module found that it was significantly increased

in both AD (diff. = **0.0159**, pp. = **0.962**) and MCI vs. control (diff. = **0.0312**, pp = **1.0**), but not significantly different for AD vs. MCI (diff. = **-0.0155**, pp. = **0.947**). Similarly, the yellow module was also significantly increase in both AD (diff. = **0.0184**, pp. = **0.980**) and MCI vs. control (diff. = **0.0279**, pp. = **0.998**) but was not different between AD and MCI (diff. = **-0.00965**, pp. = **0.844**). Enrichment analysis of both modules revealed that the yellow module was strongly enriched for neutrophil degranulation (116 genes, logBF = **28.2**, Figure 3-3d) while the green module was only weakly enriched for the same pathway (51 genes, logBF = **0.901**, Figure 3-S4e).

In female samples we also identified two modules (out of 23) where diagnosis had a significant effect: the black (logBF = **1.69**, Figure 3-S4d) and the darkgrey module (logBF = **1.57**, Figure 3-S4c). The pairwise comparisons for the black module found MCI vs. control was significantly increased (diff. = **0.0326**, pp = **1.0**) and AD vs. MCI was significantly decreased (diff. = **0.0345**, pp = **1.0**), but there was no significant difference between AD and control (diff. = **-0.00198**, pp = **0.60**). In contrast, the darkgrey module was significantly increased in AD vs. control (diff. = **0.0185**, pp = **0.988**) and AD vs. MCI (diff = **0.0361**, pp = **1.0**) and was significantly decreased in MCI vs. control (diff. = **-0.0178**, pp = **0.967**). We also found that the brown module, while not significant for model comparison (logBF = **0.212**, Figure 3-3c), still showed a significant increase in AD vs. control (diff. = **0.020**, pp. = **0.993**) and AD vs. MCI (diff. = **0.020**, pp. = **0.977**) but no difference in MCI vs. control (diff. = **-0.0001**, pp. = **0.505**). Enrichment analysis of these modules revealed significant enrichment in the darkgrey module (26 genes, logBF = **7.11**, Figure 3-S4e) and brown module (119 genes, logBF = **28.9**, Figure 3-3d) for neutrophil degranulation but no enrichment for the same pathway in the black module (7 genes, logBF = **-0.735**, Figure 3-S4e). These results confirmed our observation in the differential expression analysis that while both male and female AD subjects show evidence of increased inflammatory response, an increased inflammatory response in MCI is only observed in males.

FTD disorders

WGCNA in the cohort of FTD disorders and controls identified 14 modules, none of which identified disease status as a significant overall predictor. However, pairwise comparisons to controls identified the brown module as associated with PSP (upregulated, diff. = **0.0127**, pp. = **0.982**, Figure 3-4a) and nfvPPA (upregulated diff. = **0.0107**, pp. = **0.953**). Enrichment analysis of this module found that it was strongly enriched for neutrophil degranulation (brown: 153 genes, logBF = **40.5**). These results support the notion that a transcriptional signature associated with increased innate immune response is present in PSP vs. control but not in bvFTD vs. control, and provide some support for an increased innate immune response in nfvPPA vs. control, which could not be detected with differential expression.

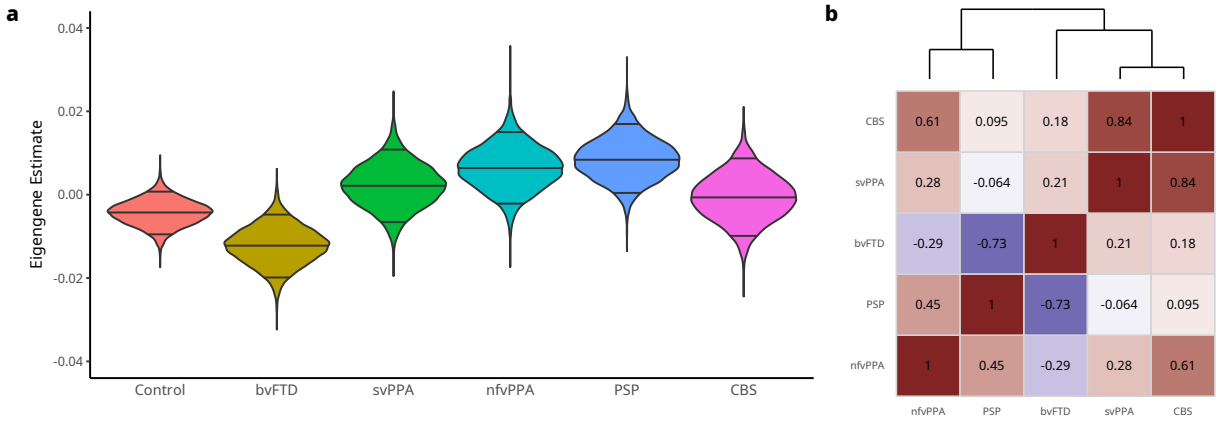

Figure 3-4. a Violin plot of posterior estimate of mean eigengene values for each diagnosis in the brown module, with the median and 5% and 95% quantiles indicated by lines. **b** Correlation plot of the pairwise correlation of the mean differences in eigengene values of FTD disorders vs. control, with the dendrogram on top showing hierarchical clustering of disorders.

Like we had previously done with the differential expression data in the FTD disorders, we correlated the mean differences between disease and control for each eigengene in one disease with the mean differences in the other diseases to understand how similar the disease effects were at the level of co-expression modules. We observed a disease clustering similar to what observed in the DE analysis: nfvPPA and PSP forming one cluster, svPPA and CBS forming another cluster, and bvFTD remaining unique (Figure 3-4b).

Enrichment of network modules for cell type specific genes in blood

Because we observed clear evidence of neutrophil degranulation in all of our WGCNA networks, we wanted to determine whether our modules were enriched for genes specific to cell types in blood. We used the pSI [21, 22] tool to identify cell type-specific genes from an existing dataset [23] and then tested for significant overlap with the top 300 genes in each of our modules. As shown in Figure 3-5, the brown module in both the AD (neutrophils: 24 genes, logBF = **8.11**; monocytes: 44 genes, logBF = **21.2**) and FTD (neutrophils: 23 genes, logBF = **7.02**; monocytes: 41 genes, logBF = **17.9**) networks was clearly the most significantly enriched for cell type specific genes in neutrophils and monocytes. Similarly, in the male and female AD networks, the yellow (neutrophils: 25 genes, logBF = **9.02**; monocytes: 51 genes, logBF = **28**) and brown (neutrophils: 26 genes, logBF = **9.46**; monocytes: 43 genes; logBF = **19.9**) modules, respectively, were also strongly enriched for cell type specific genes in neutrophils and monocytes (Figure 3-S5). Importantly, in all analyses we identified other modules not affected by disease which are also enriched for neutrophils and monocytes, indicating that not all of the transcriptome associated with these cell types is affected by disease. We also show that the modules we found to be enriched for neutrophil and monocyte genes are not enriched for any other blood cell types (Figure 3-S6).

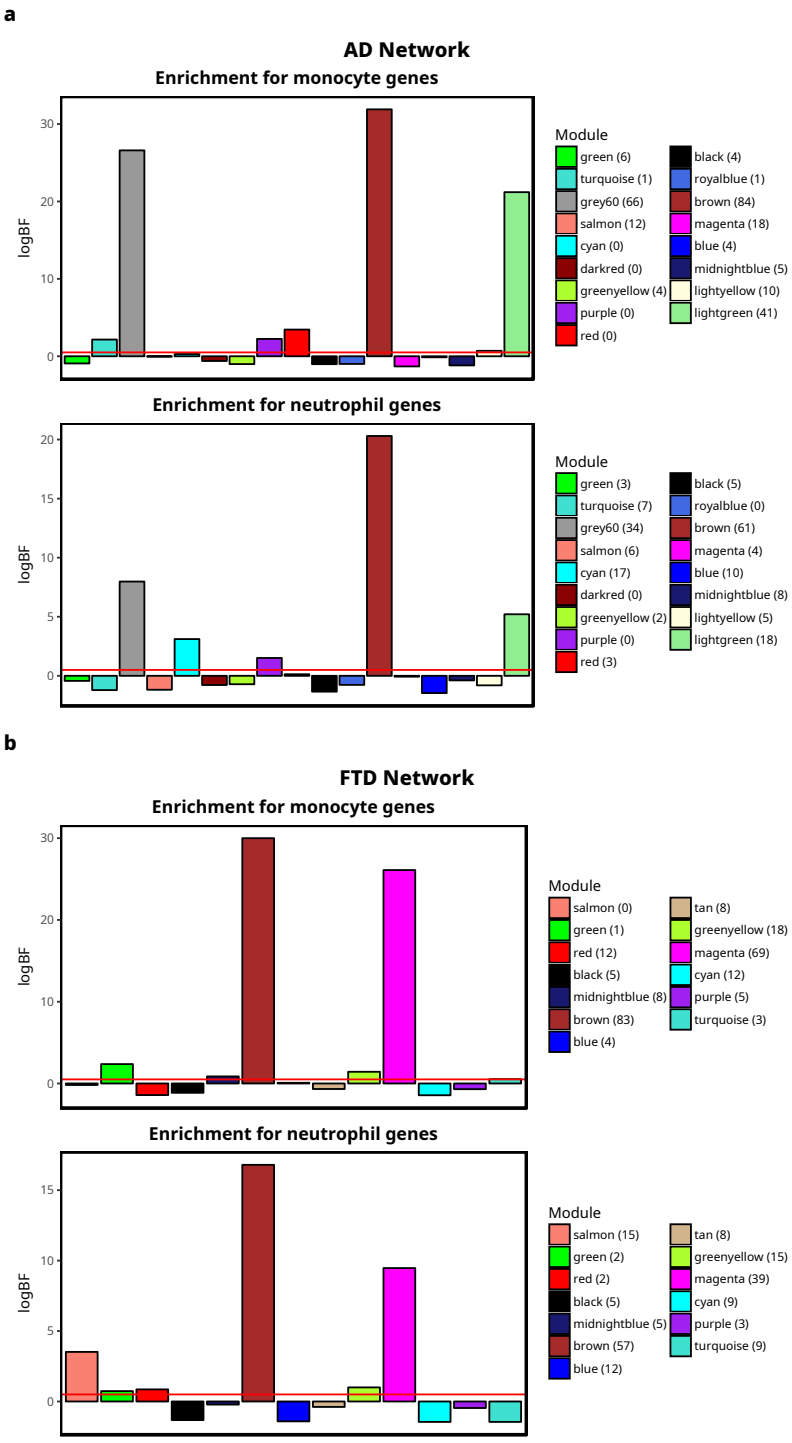

Figure 3-5. a,b Bar plots of enrichment of the top 300 genes in each WGCNA module for monocyte- and neutrophil-specific genes in the AD and FTD networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend.

Relationship between innate immune modules

After identifying modules enriched in innate immune response in neutrophils and monocytes and significantly affected by disease in the AD/MCI, sex-specific AD/MCI, and FTD networks, we wanted to know how similar the brown module in the full AD/MCI network, female AD/MCI network and FTD network and the yellow module in the male AD/MCI network were to each other, as these were the modules with the strongest enrichment for neutrophil- and monocyte-specific genes that were significantly increased in disease. We found that 692 genes were shared between all four modules, corresponding to more than 50% of the genes in the individual modules (Figure 3-6a). We also wanted to determine how similar these modules were in terms of their overall connectivity within the co-expression network. To quantify this we computed the pairwise connectivity correlation (see Methods) We find that the correlation coefficient is greater than 0.9 for all pairwise combinations of these modules (Figure 3-6b), indicating they are representing the same innate immune pathway in all four networks.

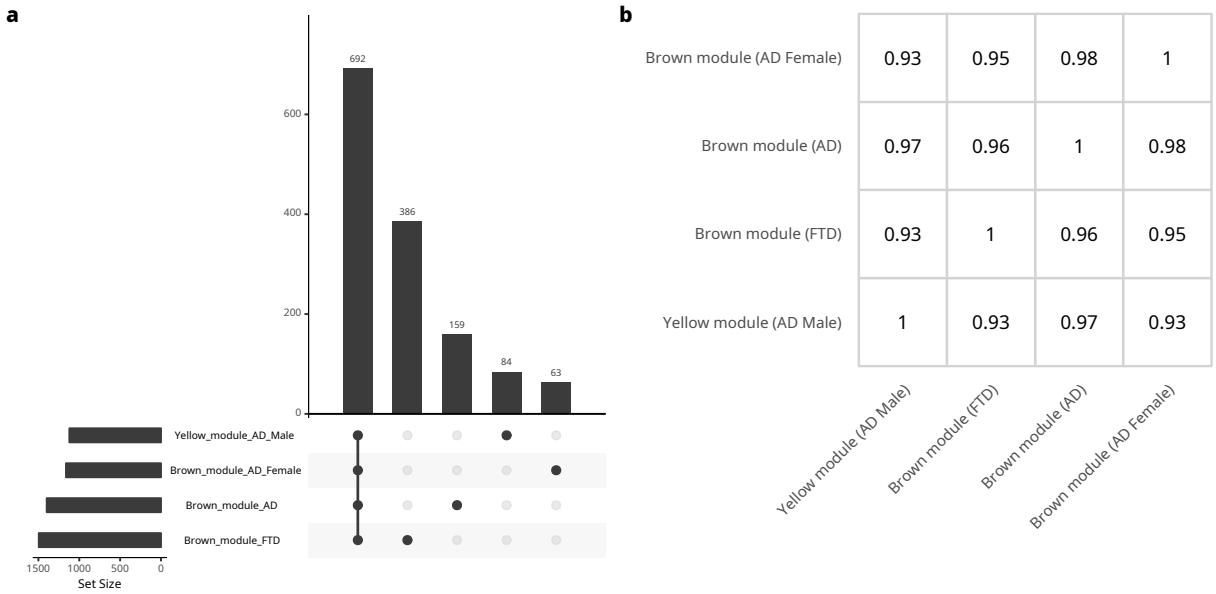

Figure 3-6. **a** UpSet plot of overlap between WGCNA modules. Total module size is in the bar plot on the left and overlap sizes are in the bar plots on the top with first bar showing genes in all 4 modules, and the other bars showing genes unique to each set. **b** Correlation plot of connectivity correlation between the brown modules in the AD, female AD and FTD networks and yellow module in the male AD network.

Cell type composition

One potential confounding factor in our identification of an enrichment in an innate immune response in upregulated genes in AD and FTD disorders is a change in cell type composition, in particular an increase in neutrophils that could explain the enrichment for neutrophil degranulation. We used both our gene expression data and methylation data to determine cell type composition of our blood samples to determine if there were any changes in cell type composition.

We used the *CellMix* package [24] to determine cell type composition using expression data. This tool uses gene expression profiles from FACS-sorted cell types in peripheral blood to estimate the proportion of most common types of cells seen in blood. Figure 3-7a shows the percent composition of cell types in the cohort of AD, MCI and control and Figure 3-7b shows the percent composition of cell types in the FTD disorders and controls. In the FTD disorders, the effect of age on cell type composition was removed with residualization before analysis. In both cohorts, no significant effect of diagnosis was seen on the proportions of any cell type, indicating that a change cell type composition as determined by gene expression was not responsible for the innate immune response signal detected in the differential expression and network analysis.

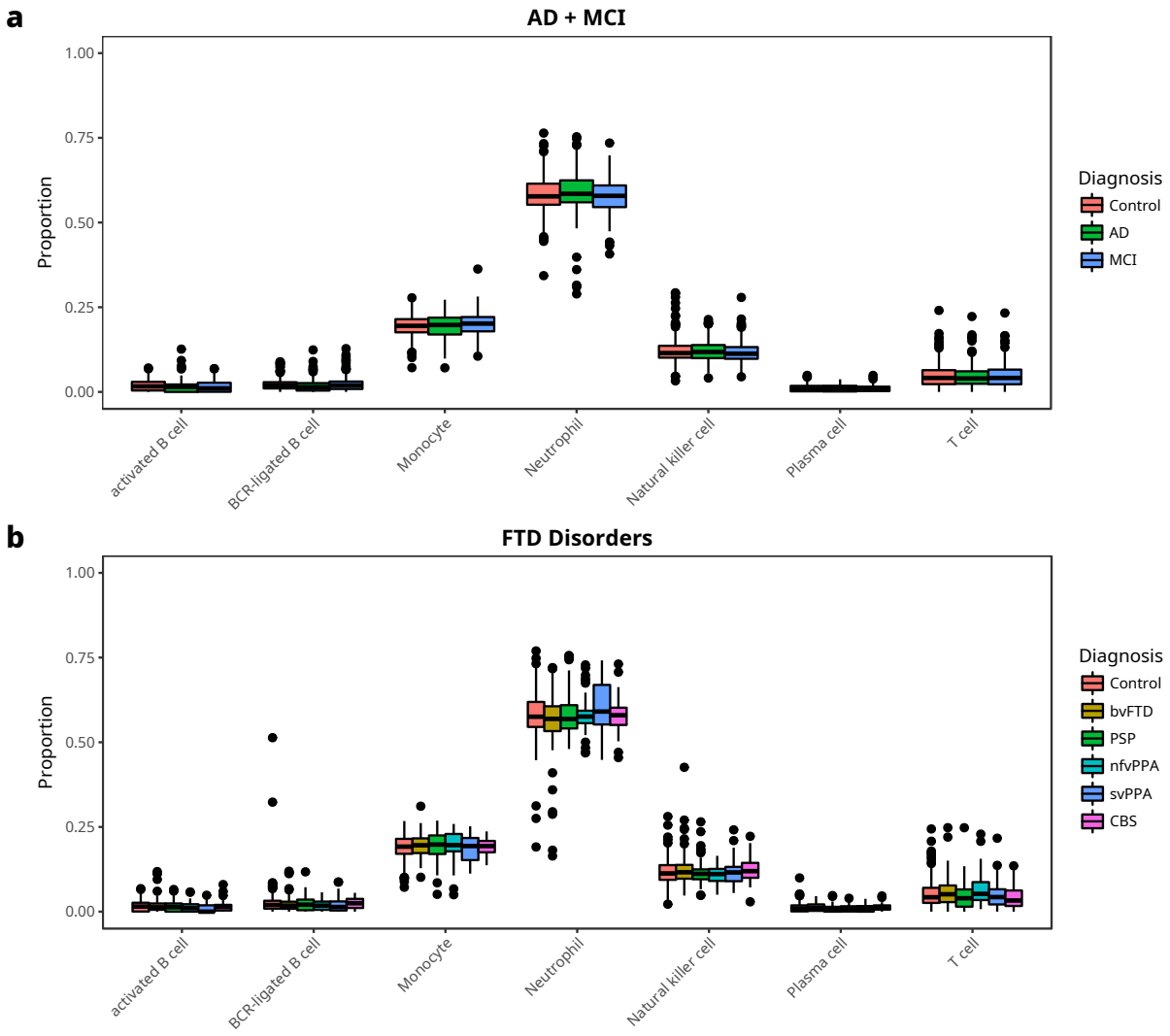

Figure 3-7. a,b Box plots of blood cell type composition estimated from gene expression in AD, MCI and control (**a**) and FTD disorders (**b**).

To confirm this, we also used our methylation data (described under DNA methylation) to estimate cell type composition, as DNA methylation is a more stable and robust marker of cell type composition than gene expression [25]. Figure 3-S7a shows the percent composition of cell types in AD, MCI and control subjects and Figure 3-S7b shows the percent composition of cell types in the FTD disorders and controls. Similar to gene-expression based estimates of cell type composition, there was no significant effect of diagnosis on cell type composition estimated using methylation, further supporting the notion that the increased innate immune response seen in the gene expression data in dementia subjects is not confounded

by changes in cell composition.

Enrichment for AD genetic risk

After identifying a consistent innate immune response signature in AD, MCI and FTD disorders in our co-expression networks, we wanted to investigate the possibility that the immune response modules we identified were enriched for genes associated with the genetic risk for AD resulting from common genetic variation.

We used MAGMA [26], a tool designed to test the gene sets for enrichment for genetic association with a trait using genome wide association study (GWAS) data. The GWAS we used was generated by the International Genomics of Alzheimer's Project (IGAP) using 17,008 AD cases and 37,154 controls and genotyped or imputed 7,055,881 single nucleotide polymorphisms (SNPs) [27].

We used the top 300 genes in each module from each of our networks as the gene sets to be tested for enrichment for genetic association. In the full AD/MCI network, we found the brown ($p < \mathbf{0.00181}$) and magenta ($p < \mathbf{0.0227}$) modules (the only modules enriched for neutrophil degranulation that increased in AD or MCI) to be significantly and exclusively enriched for genetic risk for AD (Figure 3-8a). Similarly in the sex-specific male network (Figure 3-S8), we only see robust enrichment for the yellow ($p < \mathbf{0.00231}$) and green ($p < \mathbf{0.0123}$) modules while in the sex-specific female network (Figure 3-S8) we only see robust enrichment for the brown module ($p < \mathbf{2.8 \times 10^{-6}}$). Finally, this trend holds in the FTD network (Figure 3-8b), where robust enrichment is exclusive of the brown module as well. We note this is consistent with our finding that the brown module in the AD/MCI, female AD/MCI, and FTD networks and the yellow module in the male AD/MCI network were all highly correlated with each other, therefore their enrichment for genetic risk should also be correlated.

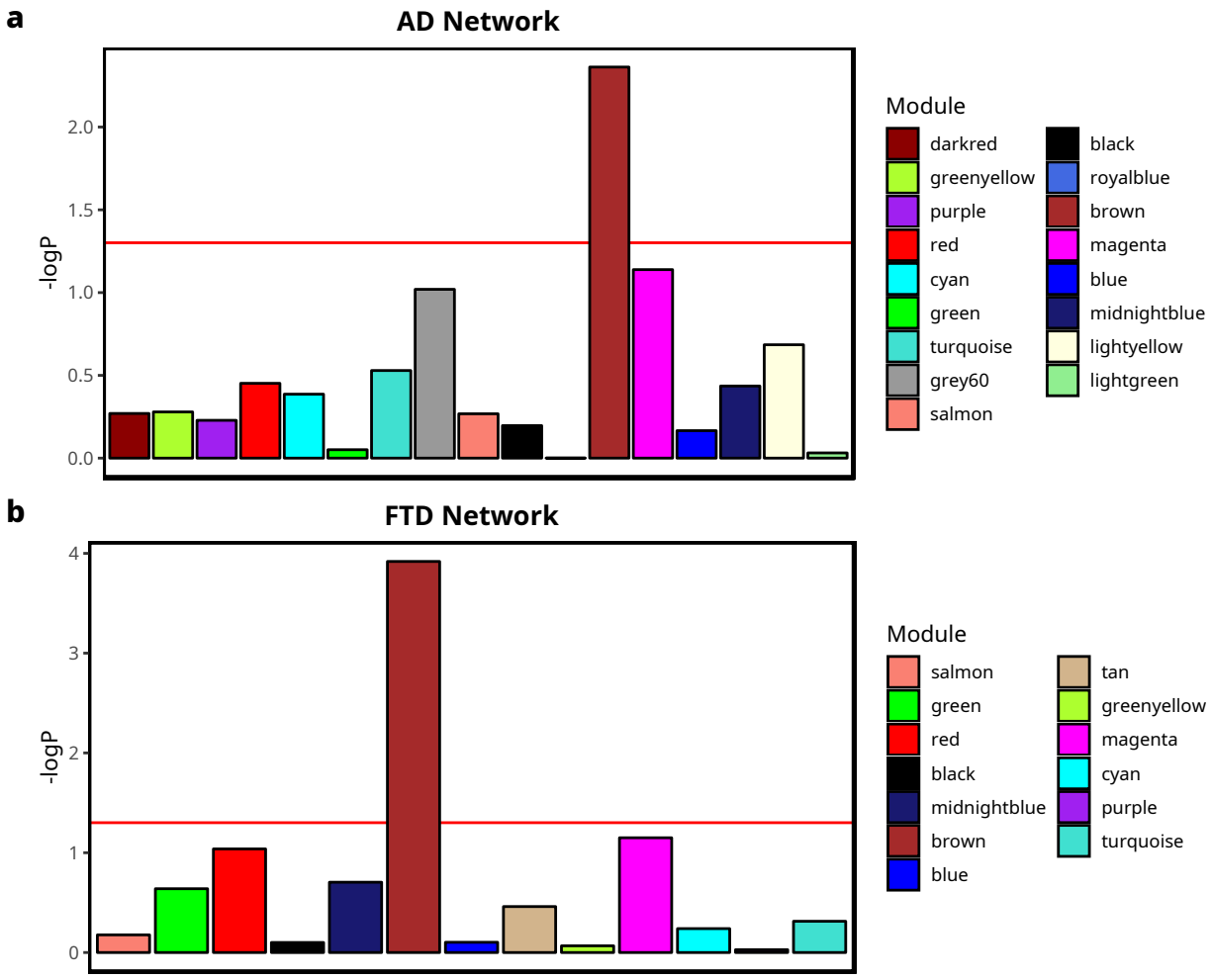

Figure 3-8. a,b Bar plots of enrichment of the top 300 genes in each WGCNA module for genetic risk for AD as estimated by MAGMA in the AD and FTD networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes.

We also tested whether this enrichment was specific to AD or was seen with other GWAS studies. We observed no enrichment, in any of the networks, for the innate immune modules described above for genetic risk for amyotrophic lateral sclerosis [28], age of onset in Huntington's Disease [29], type 2 diabetes [30], schizophrenia [31], or major depression [32], indicating that this enrichment for genetic risk is specific to AD (Figure 3-S9).

We chose not to use all of the genes in our co-expression modules because we wanted to focus on the genes most specific to the modules. To quantify the importance of using more specific genes from our gene sets, we performed two alternative analyses with MAGMA. First, we treated module membership - the correlation coefficient between each gene in the

network and each eigengene in the network – as continuous covariates. We found that in contrast to the gene set analysis, this analysis eliminated significant enrichment for any of our modules (Figure 3-S10), indicating that were no modules in which a significant correlation existed between module membership and genetic risk for AD. Second, we also wanted to show that significant enrichment in the brown module in AD, FTD, and AD female networks and the yellow module in the male AD network was not restricted to only the top 300 genes. We ranked all genes a given network by their membership in the module of interest and then generated cutoffs in step sizes of 100. As seen in Figure 3-S11, significant enrichment is seen at many cutoff sizes, showing that the enrichment observed in the top 300 genes is observed at other cutoff sizes, and therefore is robust to the arbitrary choice of this filter..

An alternative approach to identifying gene sets enriched for genetic association with traits is provided by stratified LD score regression (sLDSR) which partitions heritability of a trait between a gene set of interest and a control gene set, while accounting for background effects of heritability on specific SNPs like epigenetic modifications and enhancers [33]. This model is more conservative than MAGMA and is affected by the overall heritability of trait as estimated by the GWAS. The IGAP GWAS exhibited relatively low heritability (0.069), limiting our ability to partition that heritability among gene sets and making the model more conservative. We find that for top 300 genes in all of the co-expression modules in the AD, male AD and FTD networks, no modules meet significance after Bonferroni correction for multiple comparisons (Figure 3-S12), though the brown module in FTD was nominally significant. However the brown module in the female AD network was marginally significant after Bonferroni correction for multiple comparisons ($p < 0.0511$), providing a partial validation of our findings with MAGMA using an independent technique.

Enrichment for microglia genes

After observing a significant enrichment for genetic risk in the modules associated with an innate immune response in our co-expression networks, we hypothesized that this could be because the innate immune response in the peripheral transcriptome overlaps with the transcriptome of microglia, which mediate innate immune responses in the CNS. To test this, we used a cell type specific gene expression dataset collected from human brains [34] and

the pSI tool [21, 22] to identify cell type-specific genes in microglia as well as neurons, astrocytes, oligodendrocytes, and endothelial cells and tested the same genes sets from network modules analyzed in MAGMA for enrichment for these cell type specific genes. In the both the full AD network (26 genes, logBF = **5.92**) and the FTD network (23 genes, logBF = **4.16**), we found that the brown module was strongly enriched for microglia-expressed genes (Figure 3-9), and showed little or no enrichment for the other cell types in the CNS (Figure 3-S14). Similarly for the male and female AD networks, we found significant enrichment of microglia-expressed genes in the yellow (26 genes, logBF = **6.02**) and brown modules (24 genes, logBF = **4.76**), respectively (Figure 3-S13), and little or no enrichment for other cell types (Figure 3-S14). We also confirmed these patterns of enrichment with a second independently derived microglia-specific dataset [35] (Figure 3-S15). We note that in all cases other modules in the network also showed enrichment for microglia-specific genes, but these modules were not significantly increased in disease, suggesting that our analyses identify a subset of microglia-expressed genes which are associated with disease status, and detectable in peripheral blood.

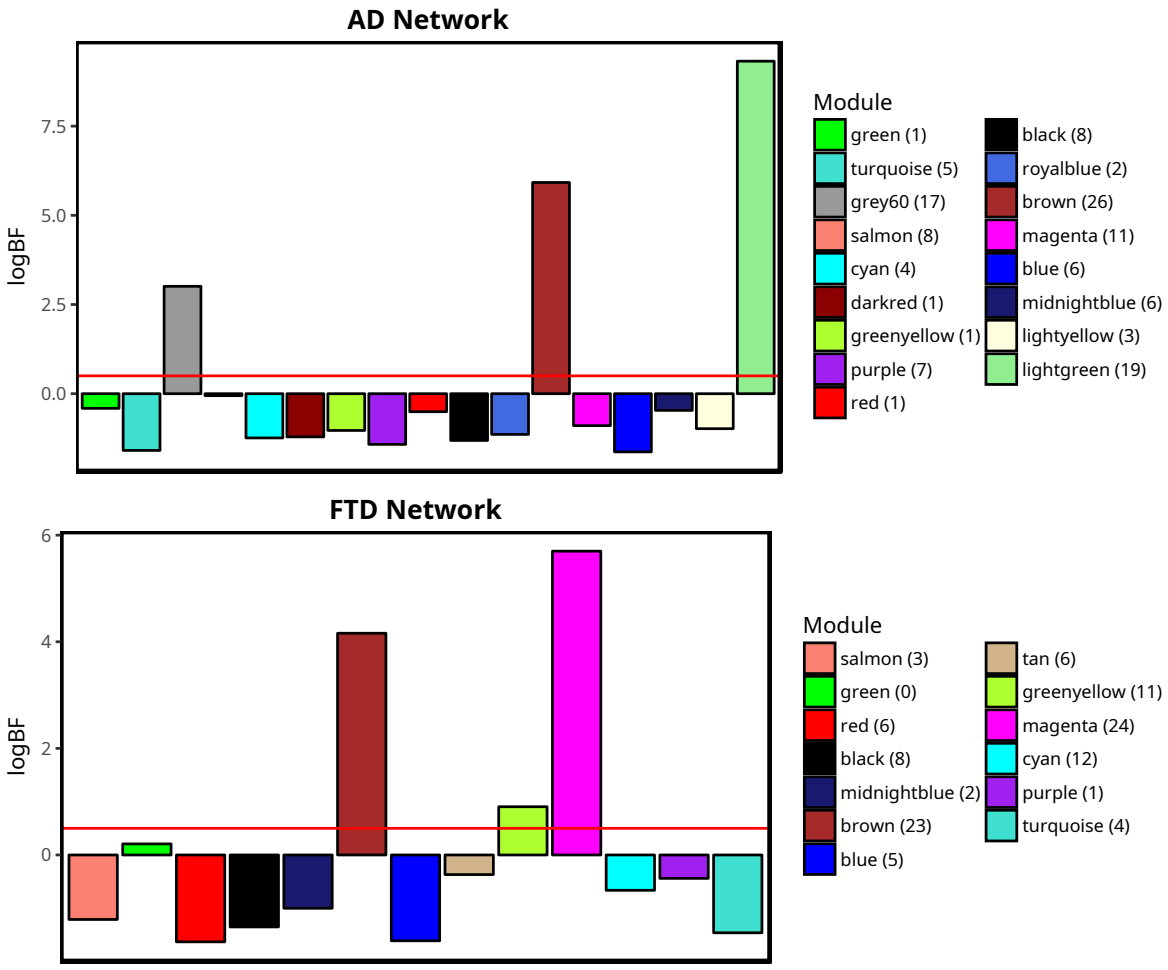

Figure 3-9. a,b Bar plots of enrichment of the top 300 genes in each WGCNA microglial genes in the AD and FTD networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend.

We also used co-expression networks constructed from post-mortem brain samples to provide additional confirmation of our finding that co-expression modules in blood were enriched for microglial genes. Co-expression modules from 27 networks generated from 6 datasets were tested for enrichment in microglial genes as described above, and the most significantly enriched module in each network was tested for enrichment in the same blood modules tested for enrichment of microglial genes directly. Figure 3-10a shows the enrichment in the brown module from our AD network for the microglial module identified in 17 brain regions taken from the same cohort of AD and control subjects [35] and run on the same microarray platform. We were intrigued to find that enrichment was strongest in a

number of brain regions most affected by Alzheimer's pathology, including the hippocampus and parahippocampal gyrus, as well as the posterior cingulate cortex [36] and caudate nucleus [37]. Figure 3-10b shows enrichment of the brown module from our AD network for a number of other microglial modules. These include modules from the amygdala and nucleus accumbens from the same study but run on a different platform, one identified in a microarray dataset taken from the DLPFC [38, 39], one identified in a network built from samples taken from patients with five psychiatric disorders: autism, schizophrenia, bipolar disorder, depression and alcoholism [40–49], and microglial modules from two RNA-Seq datasets, one collected from the temporal cortex and cerebellum which also includes PSP subjects [50], and the other from Brodmann areas 10, 22, 36 and 44 (doi:10.7303/syn3159438). Networks were already available for all datasets except the RNA-Seq data set from the temporal cortex and cerebellum, for which we built networks using the same parameters described for the blood network (see Methods). Figure 3-S16 show the same information as Figure 3-9 for the AD male, AD female and FTD networks, although we note that because the brain networks were generated using data only from AD patients and included both males and females, they are not as comparable to the other three blood networks.

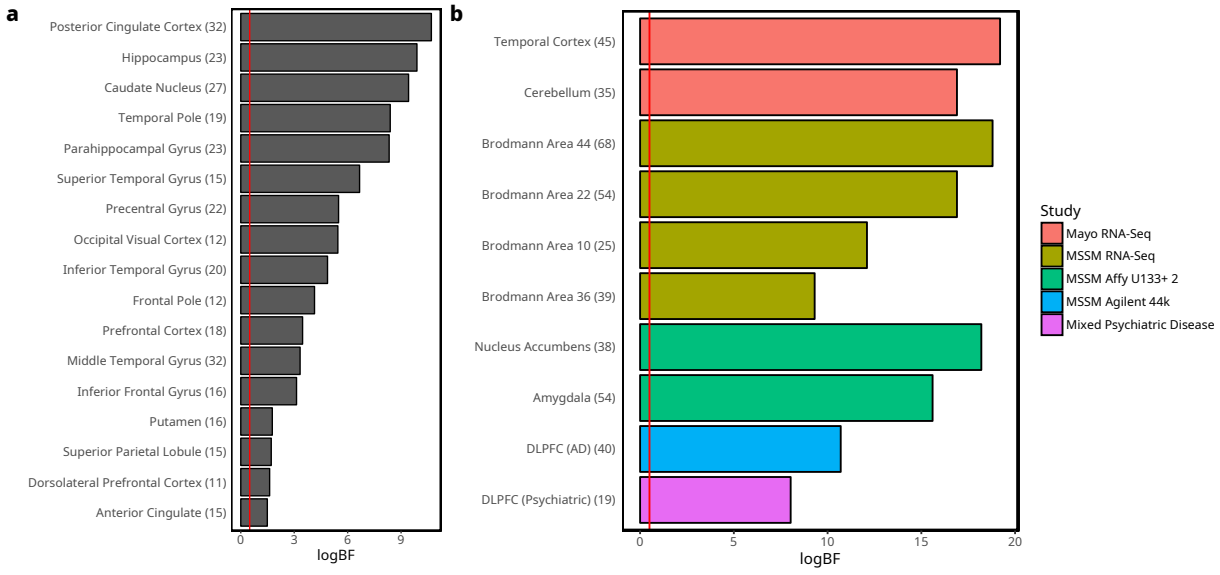

Figure 3-10. a,b Bar plots of enrichment of the top 300 genes in the brown module of the AD network for the module in each post-mortem network most enriched for microglial genes. The logBF is on the x-axis and the number of genes in each overlap is in parentheses next to the y-axis label.

DNA methylation

Once we identified a transcriptomic signature associated with an increased innate immune response in AD and FTD disorders, we wanted to understand whether there were any epigenetic changes that could be mediating some of these expression differences. Our DNA methylation samples contained AD subjects as well as FTD disorders and controls and mostly overlapped with the gene expression subjects (530 shared).

Differential Methylation

We used the *RnBeads* workflow to preprocess our array data and aggregate methylation levels across promoters, gene bodies, and CpG islands (see Methods for details). We then transformed the beta values to M-values [51], which are suitable for analysis using linear models, allowing the use of the same linear modelling approach described for differential expression to analyze differential methylation (DM) across our diagnostic groups. In the AD vs. control comparison (125 AD, 113 controls, Figure 3-S17), we identified 87 DM promoters (global FDR = **0.0042**), 106 DM gene bodies (global FDR = **0.0039**), and 104 DM CpG islands (global FDR = **0.0036**).

In the FTD disorders (44 bvFTD, 46 svPPA, 38 PSP, 20 nfvPPA, 156 controls, Figure 3-S17), we identified 194 DM promoters (global FDR = **0.015**), 147 DM gene bodies (global FDR = **0.012**), and 334 DM CpG islands (global FDR = **0.014**) for bvFTD vs. control, 118 DM promoters (global FDR = **0.021**), 64 DM gene bodies (global FDR = **0.018**), and 226 DM CpG islands (global FDR = **0.020**) for nfvPPA vs control, 166 DM promoters (global FDR = **0.011**), 163 DM gene bodies (global FDR = **0.0076**), and 267 DM CpG islands (global FDR = **0.012**) for svPPA vs. control, and 194 DM promoters (global FDR = **0.0059**), 109 DM gene bodies (global FDR = **0.011**) and 372 DM CpG islands (global FDR = **0.0060**) for PSP vs. control. There were fewer DM genes in general for AD vs. control and each other than in the expression data, while in the FTD disorders, the number of DM genes was usually comparable the number of DE genes.

We annotated promoters, gene bodies, and CpG islands with gene symbols as described in Methods and used Enrichr to test for pathway enrichment. We found that both hypo- and hypermethylated gene sets exhibited poor enrichment in general for both cohorts, indicating

that the DM genes did not converge on clear pathways like the expression data did.

Methylation and aging

While did not observe compelling differences between disease and control in differential methylation, we also wanted to know if our dementia samples showed any evidence of age acceleration when comparing their chronological age to that estimated using a methylation-based predictor, or methylation biological clock [52]. We computed two methylation ages - PhenoAge and GrimAge. PhenoAge is directly estimated from 513 CpGs identified with a penalized linear regression model which are predictive of age, while GrimAge estimates 8 blood protein measures from different sets of CpGs, and then estimates methylation age using these predicted protein measures, as well as chronological age and sex. For both measures, while we confirmed that the methylation-based age predictor is accurate in predicting the chronological age of each subject we found no significant effect of diagnosis on methylation age when fitting chronological age, sex, and batch in the null model, and these same variables as well as diagnosis in the alternate model (Figure 3-S18).

3.3 Discussion

Our comprehensive analysis of peripheral blood gene expression and methylation in AD, MCI and FTD disorders detected a transcriptional signature indicative of an increased inflammatory response in neutrophils and monocytes in AD, MCI in males, and PSP, and modest evidence for this increase in nfvPPA. Whether this response is also increased in svPPA and CBS is unclear, although their transcriptional profiles are correlated, whereas bvFTD may show evidence of decreased inflammation. We further showed that the inflammatory response we identified using network analysis is significantly and specifically enriched for genetic risk for AD, even when considering the inflammatory response seen in nfvPPA and PSP. Finally, we showed that this genetic enrichment is likely driven by the strong overlap between the transcriptional inflammatory response we see in neutrophils and monocytes with genes which are expressed in microglia. This enrichment for microglial genes is also seen even more prominently when looking at co-expression modules enriched for microglial genes in post-

mortem brains from AD patients. We also demonstrated that this transcriptional response is not driven by changes in methylation or methylation aging or by differences in cell type composition of blood between diseases and control.

The most likely reason for the correlation between transcriptional changes in neutrophils and monocytes in blood and microglia in the CNS is that the inflammatory signals responsible for activating the microglial response are also present in blood. Inflammatory peptides are increased in the serum of AD patients [7], and there is evidence that monocytes are responsible for clearing circulating amyloid beta [53], tau [54], alpha-synuclein [55], and TDP-43 [56]. We saw no clear evidence of enrichment of neuron-, astrocyte- or oligodendrocyte-specific genes in any of the co-expression modules we identified in blood, which is unsurprising given that these other CNS cell types do not have analogous cell types in blood. However, this is consistent with post-mortem gene and protein co-expression networks in AD brains [39], which showed no evidence of enrichment for genetic AD risk in neurons and astrocytes, and only weak enrichment for oligodendrocytes. By contrast, many psychiatric disorders show strong enrichment for genetic risk in neurons [49] and no enrichment in microglia, which may limit the value of studying peripheral blood gene expression in those disorders.

The observation that a transcriptional inflammatory response in PSP and nfvPPA would be enriched for AD genetic risk is intriguing because it implies that there is some overlap in inflammatory signaling across diseases. GWAS studies for FTD disorders are currently very underpowered, but these results predict that, at least for PSP and nfvPPA, some genetic risk may be mediated through microglial genes and may be correlated with microglia-associated genetic risk seen in AD. Similarly, the finding that bvFTD does not show an increased inflammatory response in blood is consistent with the lack of any enrichment for genetic risk in any of our co-expression modules for ALS, a disorder which is comorbid with bvFTD and shares many genetic risk factors [2], and would predict that microglial inflammatory responses are not as relevant to disease pathology as they are in AD.

The sex difference we identified in the inflammatory response in MCI patients may be related the sex differences in the clinical presentation of MCI. Previous studies have found that females show less memory impairment than males with an equivalent level of neurodegeneration [57, 58], and this may lead to underdiagnosis of MCI in females. In our data, we saw

no evidence of an inflammatory response in female MCI subjects, but saw that in male MCI subjects, the inflammatory response was as strong as that seen in AD subjects. This could arise because of different prodromal AD trajectories in females whereby impaired cognition would be detectable only when it has progressed to AD itself, while in males the cognitive impairment would be detectable during earlier stages as MCI, before progression to AD.

The lack of any relationship between the transcriptional inflammatory response we observed and changes in methylation in blood supports the idea that the we are observing a non-cell autonomous response to inflammatory signaling rather than a response that is driven by epigenetic changes. Previous studies have similarly reported a lack of DNA methylation changes in peripheral blood associated with AD [59], and our analysis in a larger sample size confirms this finding. Analysis of methylation post-mortem AD brains identified only a small number of significantly altered CpGs [60], although many of those CpGs were located in genes of interest to AD such as *ABCA7* and *BIN1*.

3.4 Methods

Gene expression array preprocessing

Illumina HT-12 v4 microarrays were preprocessed using the *lumi* pipeline [61] as previously described (FRDA). Expression values were normalized using the variance-stabilized transformation [62] and robust spline normalization was used for inter-array normalization. Probes with a detection score p-value greater than 0.01 were dropped, as were probes which were unannotated. Duplicated probes for the same transcript were dropped using the *collapseRows* function [63] from the *WGCNA* package. Outliers were removed based on connectivity z-scores [64]. Batch correction was performed using ComBaT [65] and any batch with less than 8 samples was dropped to allow for more robust estimation of batch effects. Subsets of data such as those partitioned by sex were preprocessed separately, meaning that different sets of samples were identified as outliers, and different batches were dropped based on the number of samples per batch in a given subset of data.

Methylation array preprocessing

Illumina HumanMethylation450K microarrays were preprocessed using the *RnBeads* pipeline [66]. Probes outside of a CpG context or located on SNPs or sex chromosomes were removed, as were probes with missing values or a standard deviation below 0.005. The *greedyCut* algorithm implemented by *RnBeads* was used to remove probes with low detection score p-values and samples identified as outliers. The *noob* method was used for background correction [67] and beta mixture quantile normalization [68] was used for normalizing methylation levels. Beta values were converted to M-values [51] so they would have Gaussian distributions suitable for use in linear models. Batch correction was done using ComBaT [65].

Linear Modeling and Residualization

All linear models used in the study were built with the *BayesFactor* package [69, 70] which implements full Bayesian linear regression. The default Cauchy prior for mean effect size and inverse Wishart prior for effect variance were used. Bayes factors comparing the alternate model containing the variable of interest (usually diagnosis) to the null model without the variable of interest were computed using the default approximation method and log10-transformed. A log10-transformed Bayes factor (logBF) greater than 0.5 was considered significant, a threshold consistent with previous analyses (FRDA PAPER). Posterior distributions for all model parameters were estimated using Gibbs sampling with 10,000 iterations, the median of the distributions was used as the maximum a posteriori (MAP) estimate of each model parameter. When needed, residuals were computed for residualization by subtracting the MAP estimated effects of variables from the original data.

Posterior probabilities for individual pairwise contrasts being non-zero were computed by finding the proportion of posterior samples with a sign opposite of the MAP estimate for that parameter. A posterior probability of 0.95 or greater was considered significant. The global FDR for differential expression and differential methylation analyses was computed by taking the average of the posterior probabilities for all genes declared significant for an individual analysis.

Differential Expression and Differential Methylation

Linear models as described above were used to identify differentially expressed (DE) and differentially methylated (DM) genes in normalized, batch corrected data. The alternate model for each gene included the variable of interest (either diagnosis or ApoE allele) along with age and sex as covariates, unless age was confounded, in which case it was removed by residualization before fitting the final model. The null model contained only the covariates. DE and DM genes were defined as genes with a logBF > 0.5 for diagnosis or ApoE allele. Furthermore, in models in which diagnosis had more than 2 values, the posterior probability of the pairwise difference between two diagnoses had to exceed 0.95 to be considered DE or DM. Correlation between differential expression across FTD disorders was computed using biweight midcorrelation of logFC values of each diagnosis vs. control and disease were clustered using average-linked hierarchical clustering.

Methylation aging

Methylation data was reprocessed with no removal of samples or probes. Background correction and normalization were performed as previously described for the DM analysis. Methylation ages were estimated using the published methylation clock model [52]. Significance of diagnosis on methylation aging was assessed using a linear model with chronological age, sex, and diagnosis in the alternate model and chronological age and sex in the null model for the AD and MCI data. For the FTD disorders, the effect of chronological age on methylation was removed with residualization because chronological age was confounded with diagnosis, and the alternate and null models were the same as those used with AD and MCI data except that chronological age was excluded from both models.

Weighted Gene Co-expression Network Analysis (WGCNA)

We constructed gene co-expression networks using the WGCNA package [15]. The expression data used for each network was normalized and corrected for batch effect but no other covariates were removed. Signed adjacency matrices with a soft power of 12 were computed using biweight midcorrelation [71], converted to topological overlap matrices [64] and

clustered using the default average-linked hierarchical clustering. Modules were identified using dynamic tree cutting of the hierarchical clustering tree [72] with cut height of 0.995 and a deepSplit parameter of 2. Module eigengenes were computed from the first principal component of the expression values of the genes in each module, and correlated modules were merged using a dissimilarity threshold of 0.2. We determined eigengene significance using linear models with the same design and significance cutoffs as described above, and removed confounding covariates using residualization on the eigengenes before fitting the final model.

We computed connectivity correlation between modules from different networks by first computing the correlation between the module eigengene of interest and all of the genes in its own network to create an eigengene connectivity vector. We identified the intersection of all genes shared between all of the networks, and subsetted the eigengene connectivity vectors so that they only contained these shared genes. Finally we computed the pairwise correlation between these connectivity vectors to obtain the connectivity correlation.

Cell type composition

Cell type composition was estimated from both gene expression and methylation data. For gene expression data, we used the *CellMix* package [24] to estimate cell type composition from the raw unnormalized data with the quadratic programming model using the default blood cell type dataset provided by the package [73]. For methylation data, we used the *estimateCellCounts* function from the *minfi* package [74] to estimate cell type composition from raw unnormalized data with a previously described regression model [75]. Cell type compositions were converted from proportions to M-values to give them gaussian distributions and confounding covariates were removed using residualization. The same linear models used differential expression and methylation were used to identify if significant differences in cell type composition could be found between diagnostic groups.

Pathway enrichment analysis

Pathway enrichment was analyzed using the GO Biological Process 2017b dataset downloaded from Enrichr [76]. A Bayesian hypergeometric overlap test [77] was used to determine if overlap between a given gene set and all gene sets in the GO Biological Process was significant, with a logBF > 0.5 being defined as significant.

Enrichment for genetic risk with MAGMA

We used the MAGMA tool [26] to compute enrichment for genetic risk in the top 300 genes our co-expression network modules. MAGMA aggregates the genetic association Z-scores for individual SNPs in a given gene into a single gene-level score, and then fits a linear mixed effects model which tests whether membership in a gene set significantly increases the association Z-score while accounting for linkage disequilibrium between genes. The resulting p-value represents the probability of observing the difference in Z-score between the gene set members and non-members under the null hypothesis that the difference is 0. All GWAS studies and annotations used hg19/GRCh37 coordinates. Annotation of SNPs to HGNC symbols was done using Ensembl 87, the last release available for hg19, and no window was added up or downstream of genes. Linkage disequilibrium between genes was estimated using 1000 Genomes Phase 3 [78] European samples and synonymous SNPs were dropped. SNP-level association statistics were aggregated using the default mean method. The sensitivity analysis performed in modules of interest used the same parameters as the main gene set enrichment except that gene sets were instead created by ranking all genes by correlation with the appropriate eigengene and taking increasing cutoffs in increments of 100. Regression with module membership was computed by using module membership as a continuous covariate.

Enrichment for genetic risk using sLDSR

We also used stratified LD score regression (sLDSR) [33] as a validation method for genetic enrichment. SNPs from 1000 Genome Phase 3 [78] European samples were annotated using Ensembl 87 with no window around the gene as with MAGMA and LD scores for SNPs in each

gene set and its corresponding control (all other genes present in the WGCNA network) were computed and filtered to only include SNPs found in HapMap3 [79]. LD scores in each gene set were compared to LD scores in the corresponding control set using the baseline model and regression weights provided by the original publication.

Enrichment for cell type specific expression

We used pSI ([21, 22]) to estimate enrichment of co-expression network modules for expression of cell type specific genes from published data sets in peripheral blood [23] and the CNS [34]. Cell type specific genes were identified computing the specific index using the 100 permutations considering genes with an expression cutoff of the 5% quantile of all expression values. Genes with a specific index p-value < 0.01 were included in each cell type specific list. In blood, we identified 353 basophil-specific genes, 372 naive B-cell-specific genes, 168 mature B-cell-specific genes, 104 myeloid dendritic cell-specific genes, 44 eosinophil-specific genes, 226 neutrophil-specific genes, 272 megakaryocyte-specific genes, 296 monocyte-specific genes, 66 mature NK-cell-specific genes, 49 memory CD8+ T-cell-specific genes, 187 naive CD8+ T-cell-specific genes, 115 naive CD4+ T-cell-specific genes and 83 memory CD4+ T-cell-specific genes. In the CNS, we identified 433 microglia-specific genes, 343 endothelial cell-specific genes, 400 neuron-specific genes, 295 astrocyte-specific genes, and 196 oligodendrocyte-specific genes. Enrichment was tested using hypergeometric overlap testing as described for pathway enrichment. Enrichment for cell type specific modules identified in post-mortem brain co-expression networks in AD [39] and psychiatric disorders [49] was also tested using hypergeometric overlap testing.

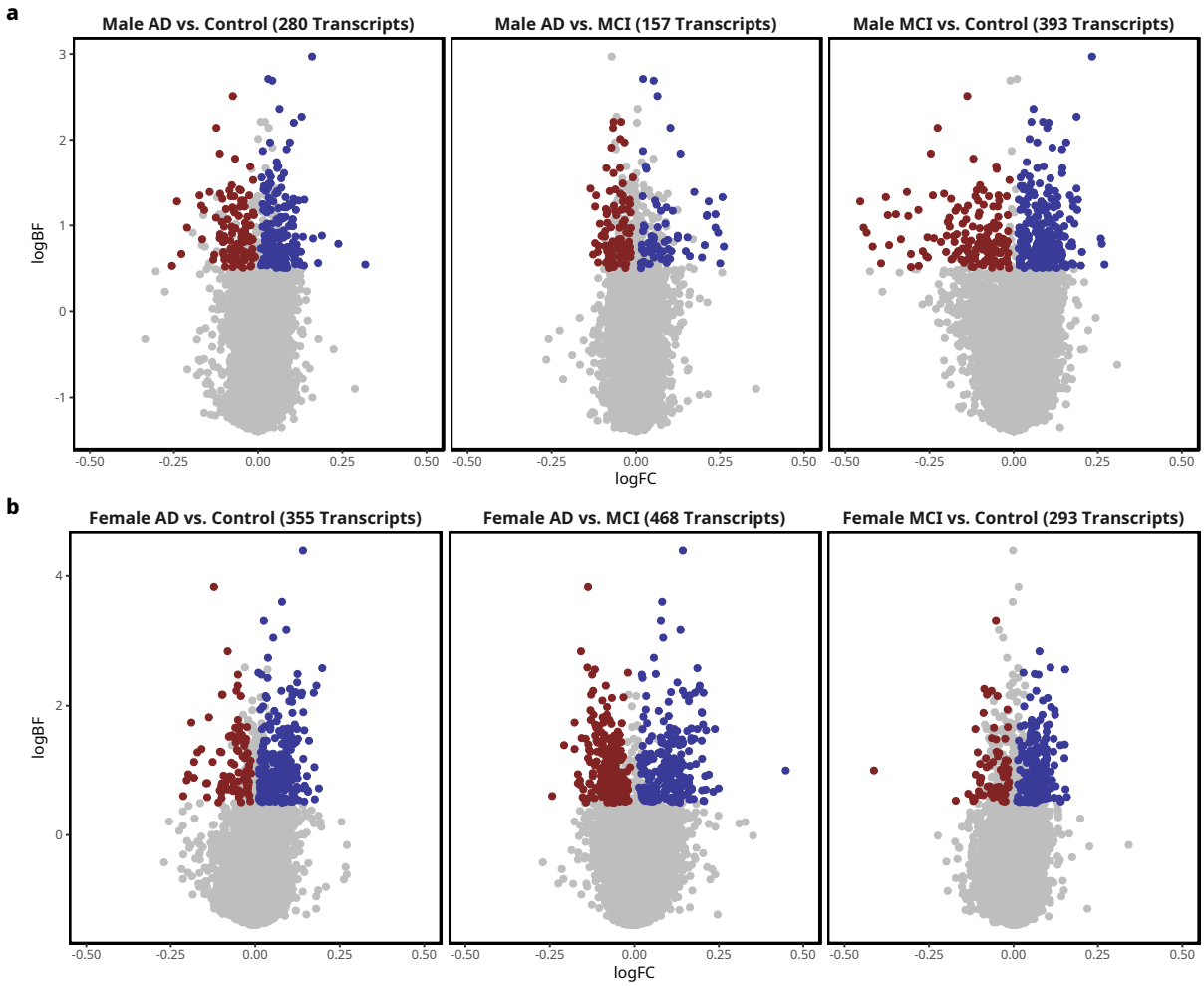

Figure 3-S1. a,b Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles.

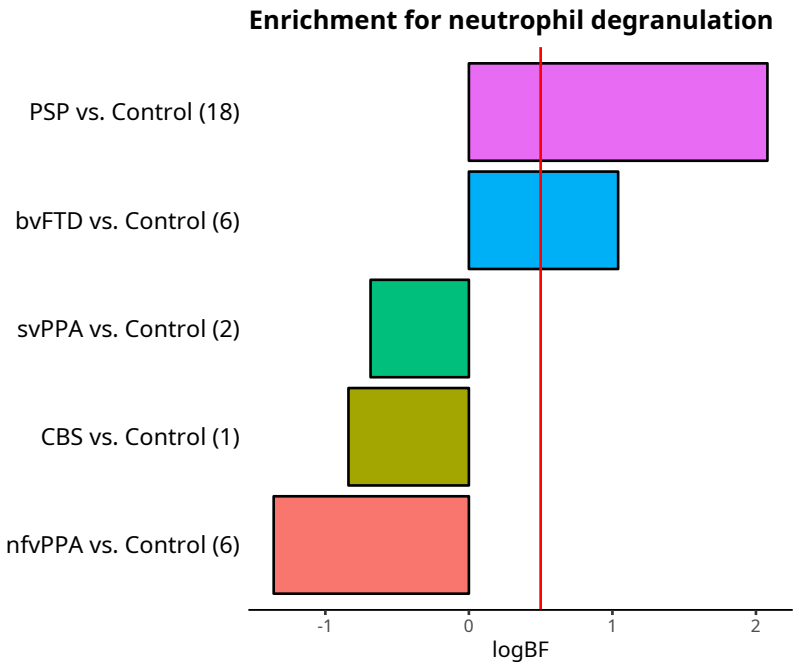

Figure 3-S2. Bar plots of enrichment of significantly upregulated genes for neutrophil degranulation (GO:0043312) with the logBF on the x-axis.

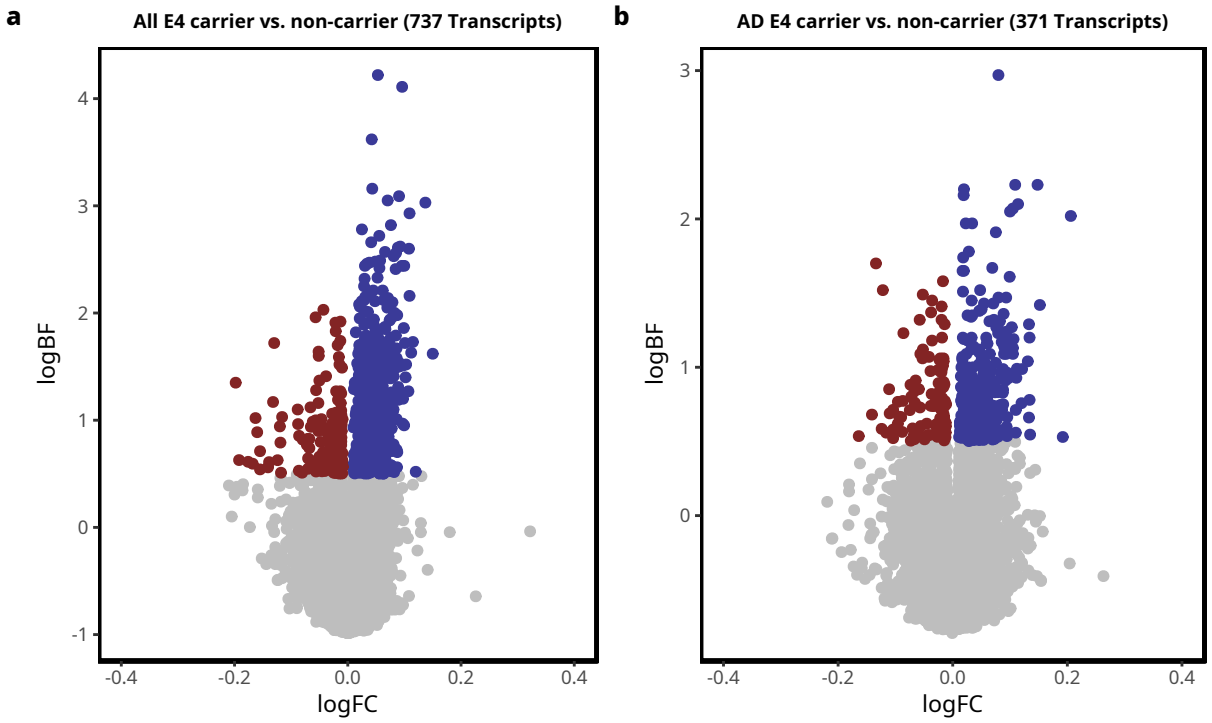

Figure 3-S3. Volcano plots of the log fold change (logFC) in gene expression on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The contrast and number of DE genes are shown in the plot titles.

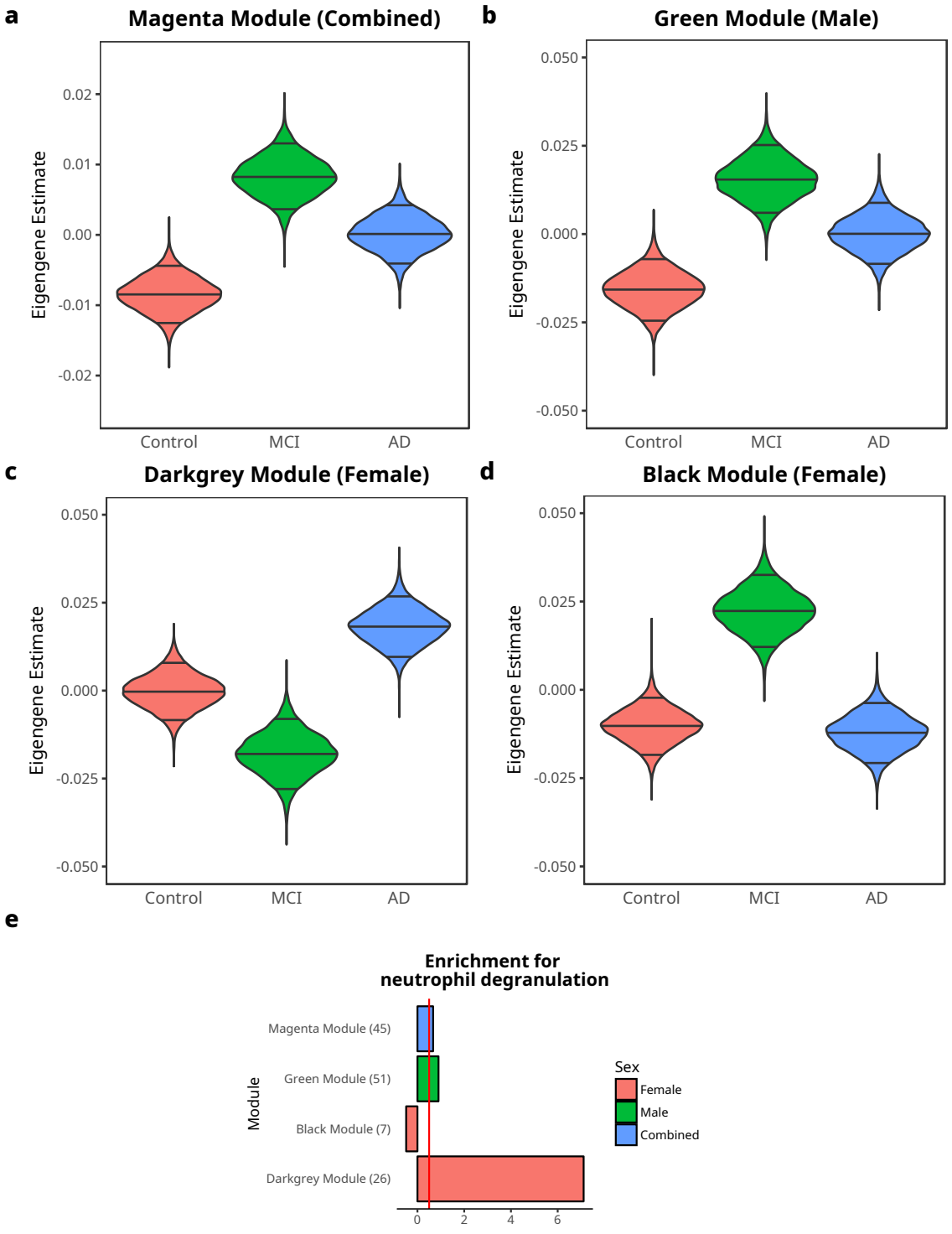

Figure 3-S4. a-d, Violin plots of posterior estimate of mean eigengene values for each diagnosis, with the median and 5% and 95% quantiles indicated by lines. **d** Bar plot of enrichment of genes in each module for neutrophil degranulation (GO:0043312) with the logBF on the x-axis.

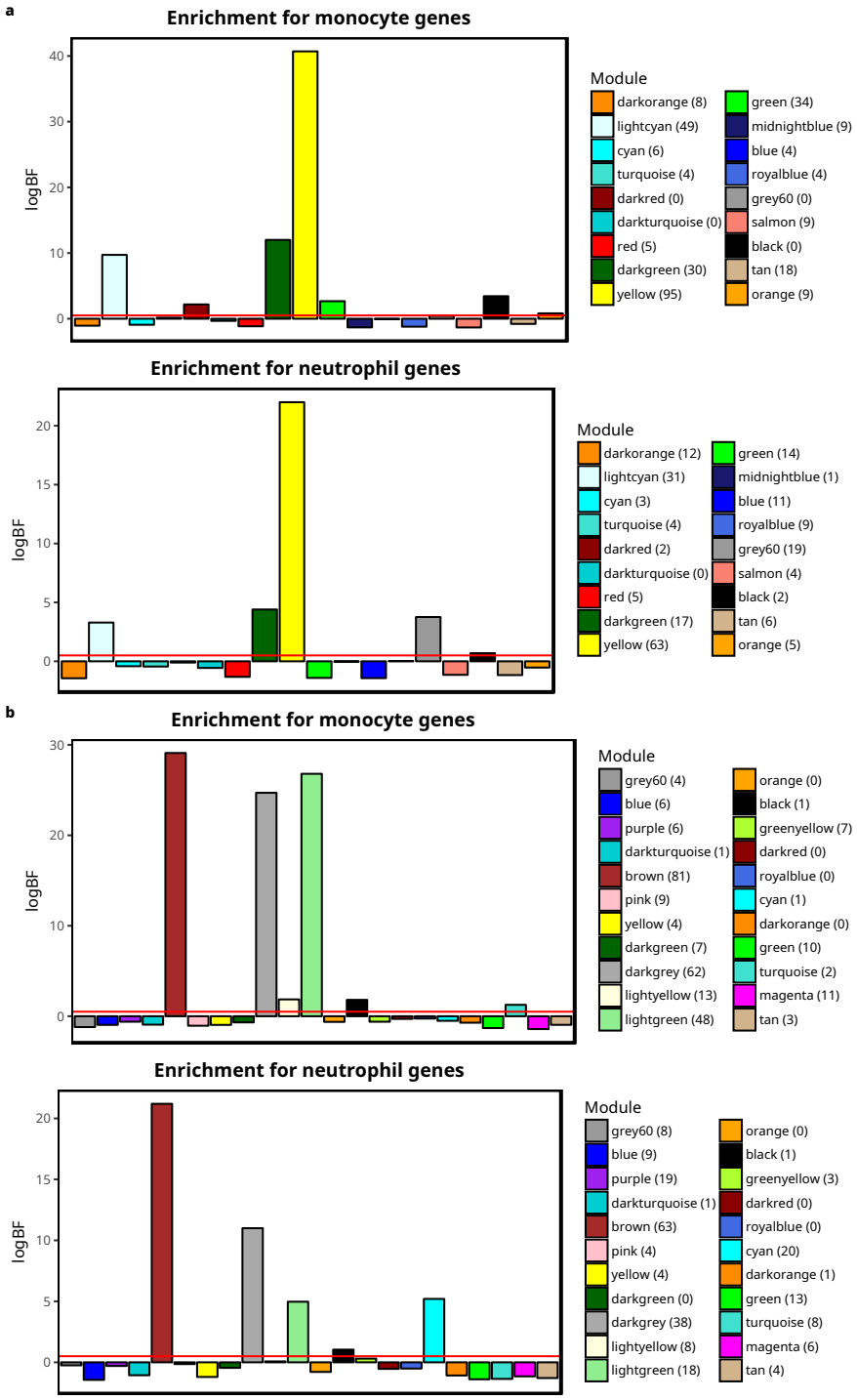

Figure 3-S5. a,b Bar plots of enrichment of the top 300 genes in each WGCNA module for monocyte- and neutrophil-specific genes in the AD male and female networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend.

Figure 3-S6. Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses.

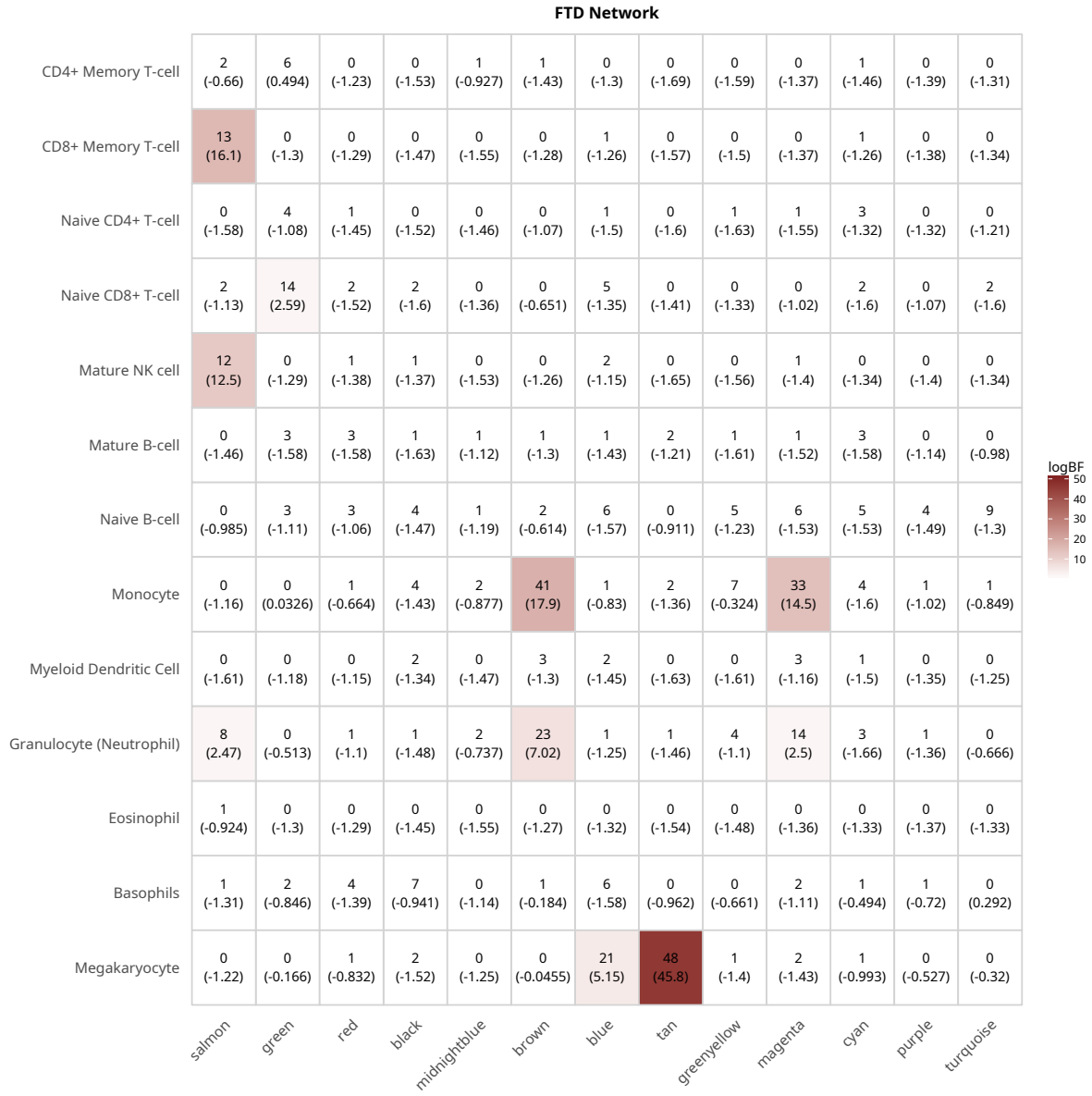

Figure 3-S6. (cont) Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses.

Figure 3-S6. (cont) Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses.

Figure 3-S6. (cont) Labeled heatmaps of enrichment of all network modules for all blood cell types. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses.

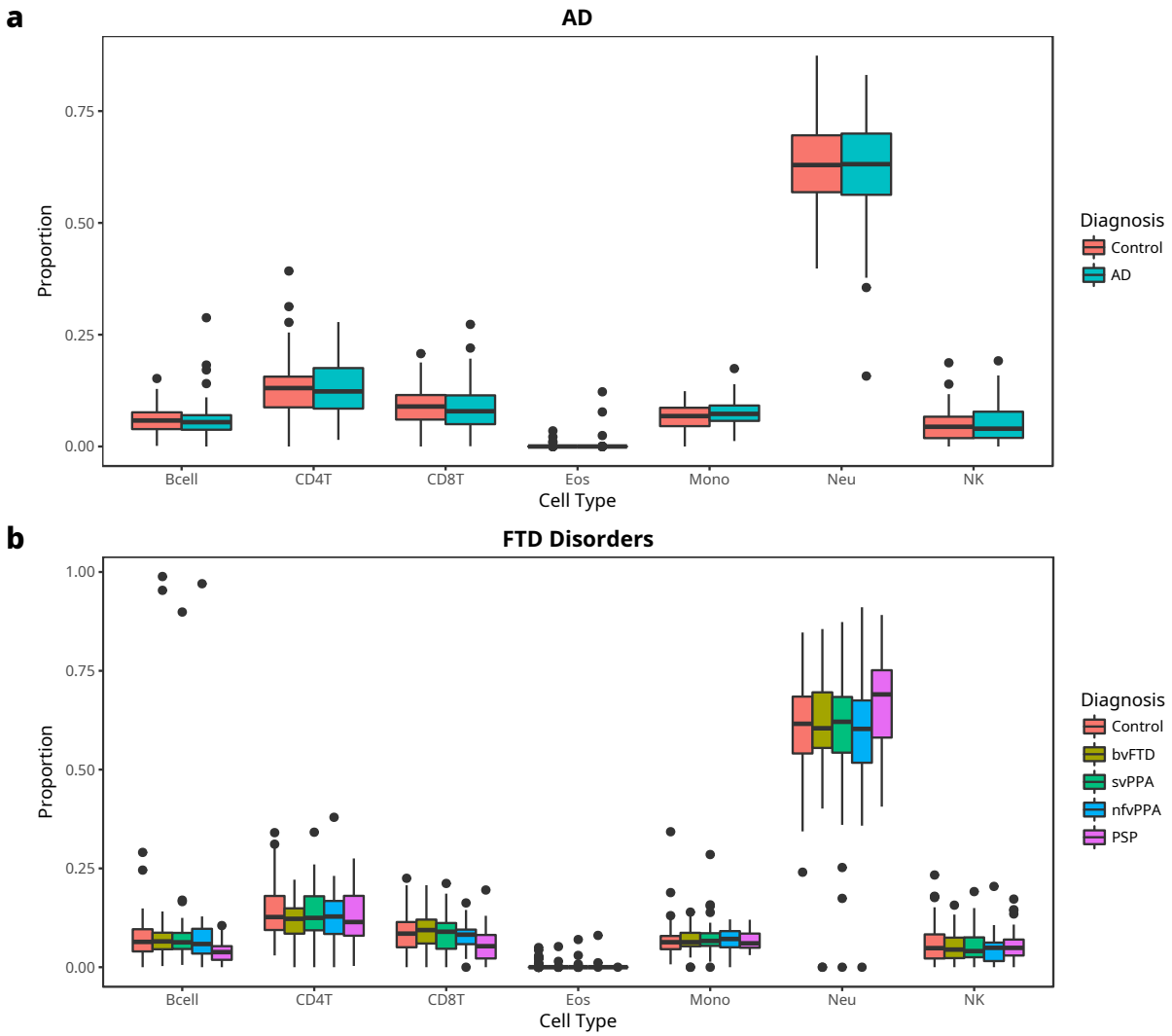

Figure 3-S7. Box plots of blood cell type composition estimated from methylation in AD, MCI and control (**a**) and FTD disorders (**b**).

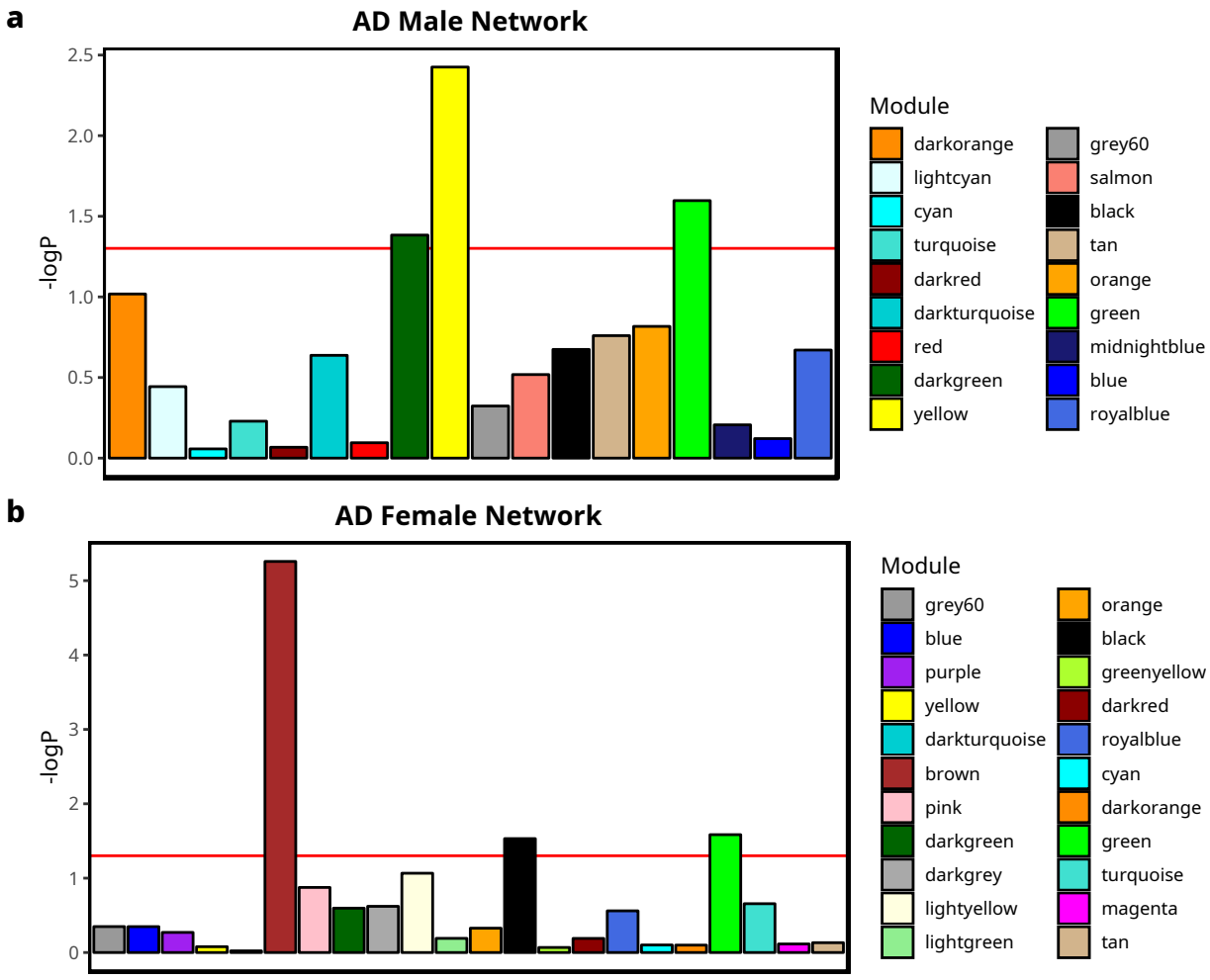

Figure 3-S8. a,b Bar plots of enrichment of the top 300 genes in each WGCNA module for genetic risk for AD as estimated by MAGMA in the AD male and female networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes.

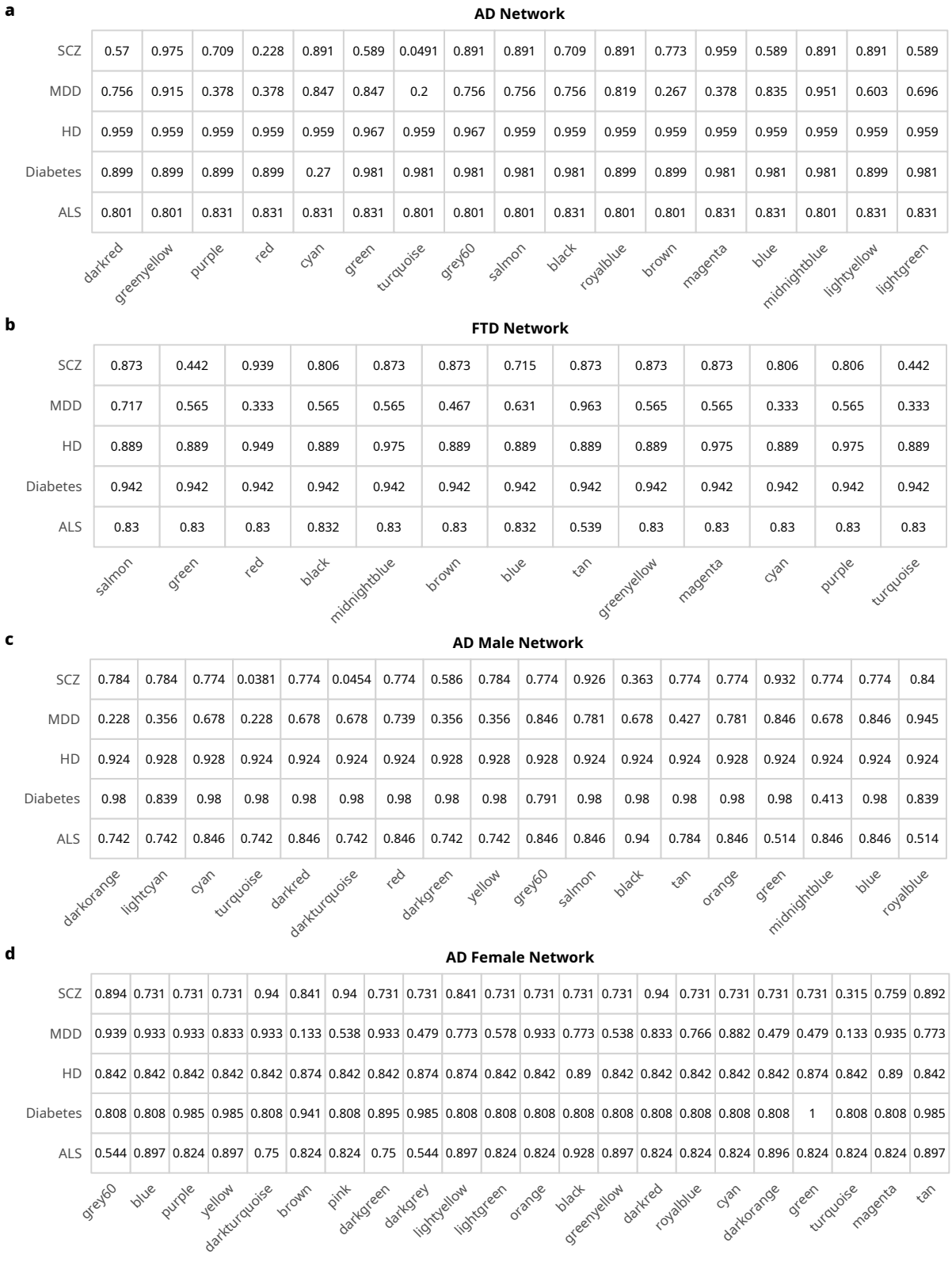

Figure 3-S9. a-d Labeled heatmaps of enrichment of all network modules for other GWAS studies. Each cell shows the -$\log_{10}$ p-value of the enrichment.

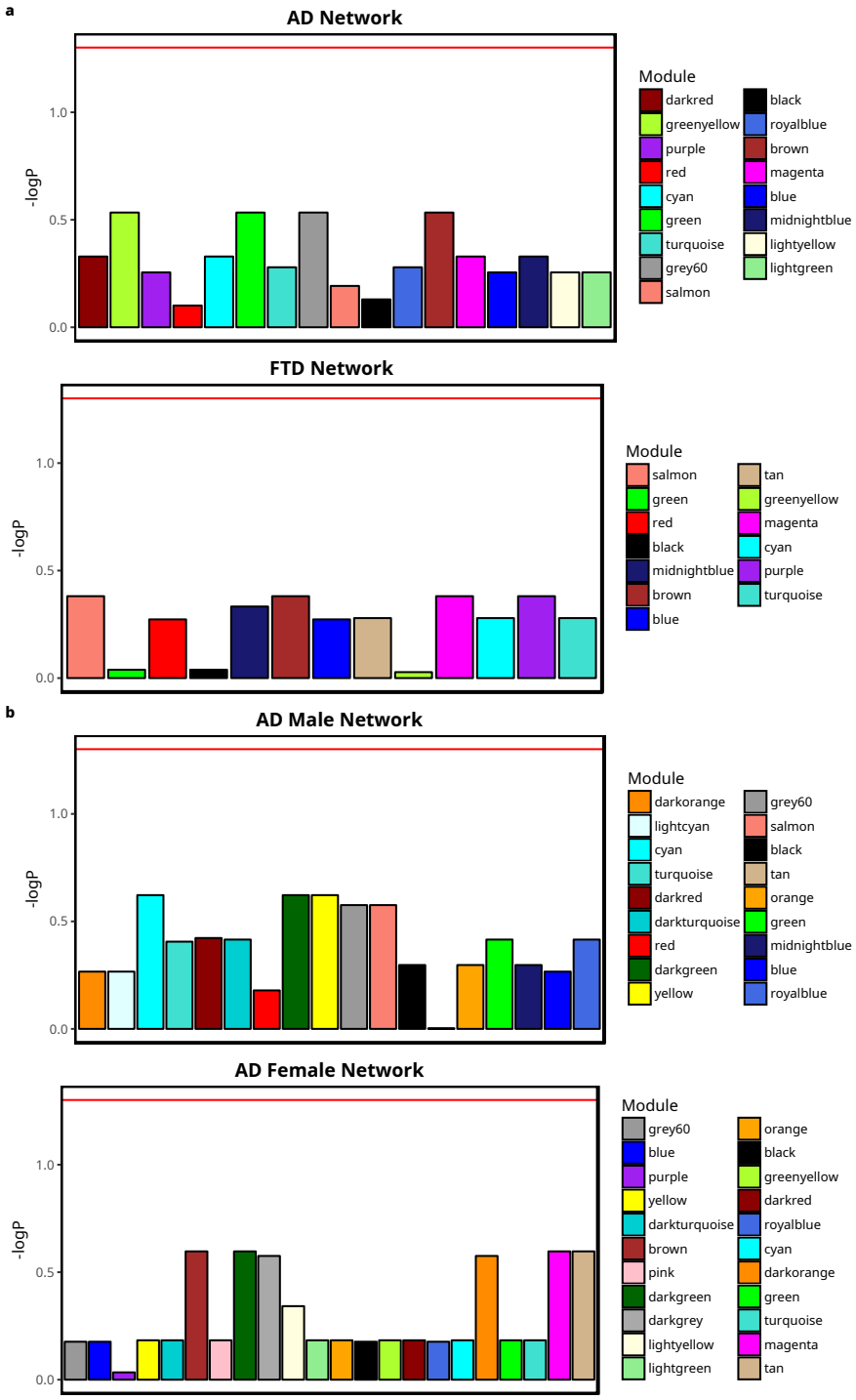

Figure 3-S10. a,b Bar plots of association of all module memberships with AD risk for all modules in (**a**) AD and FTD networks and (**b**) AD male and female networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes.

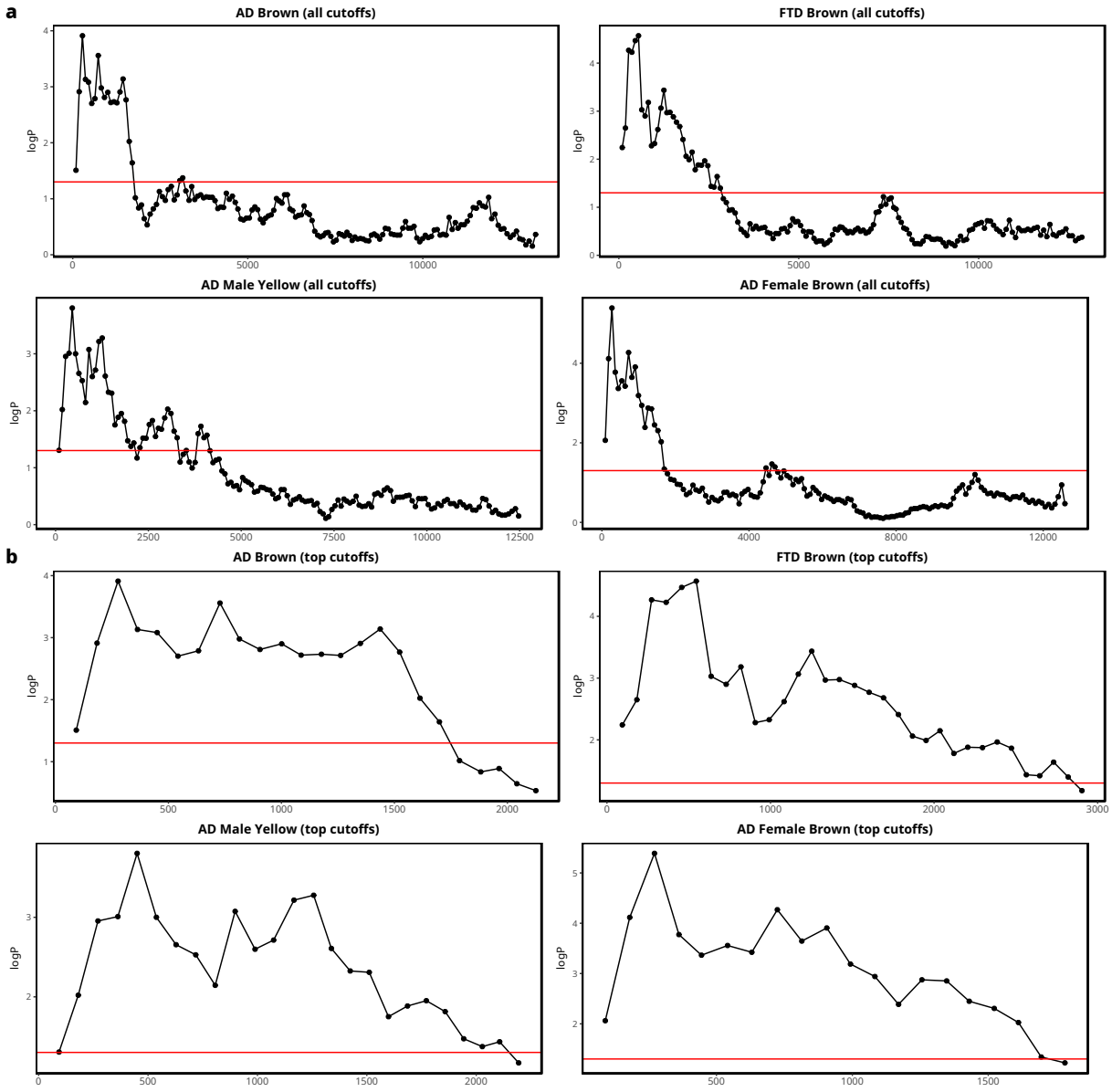

Figure 3-S11. a,b Line plots of sensitivity analysis showing all cutoffs (**a**) or all cutoffs until the first cutoff that is not significant (**b**). The x-axis is the size of the cutoff of genes ranked by membership in the module in each plot title, and the y-axis show the -$\log_{10}$ p-value.

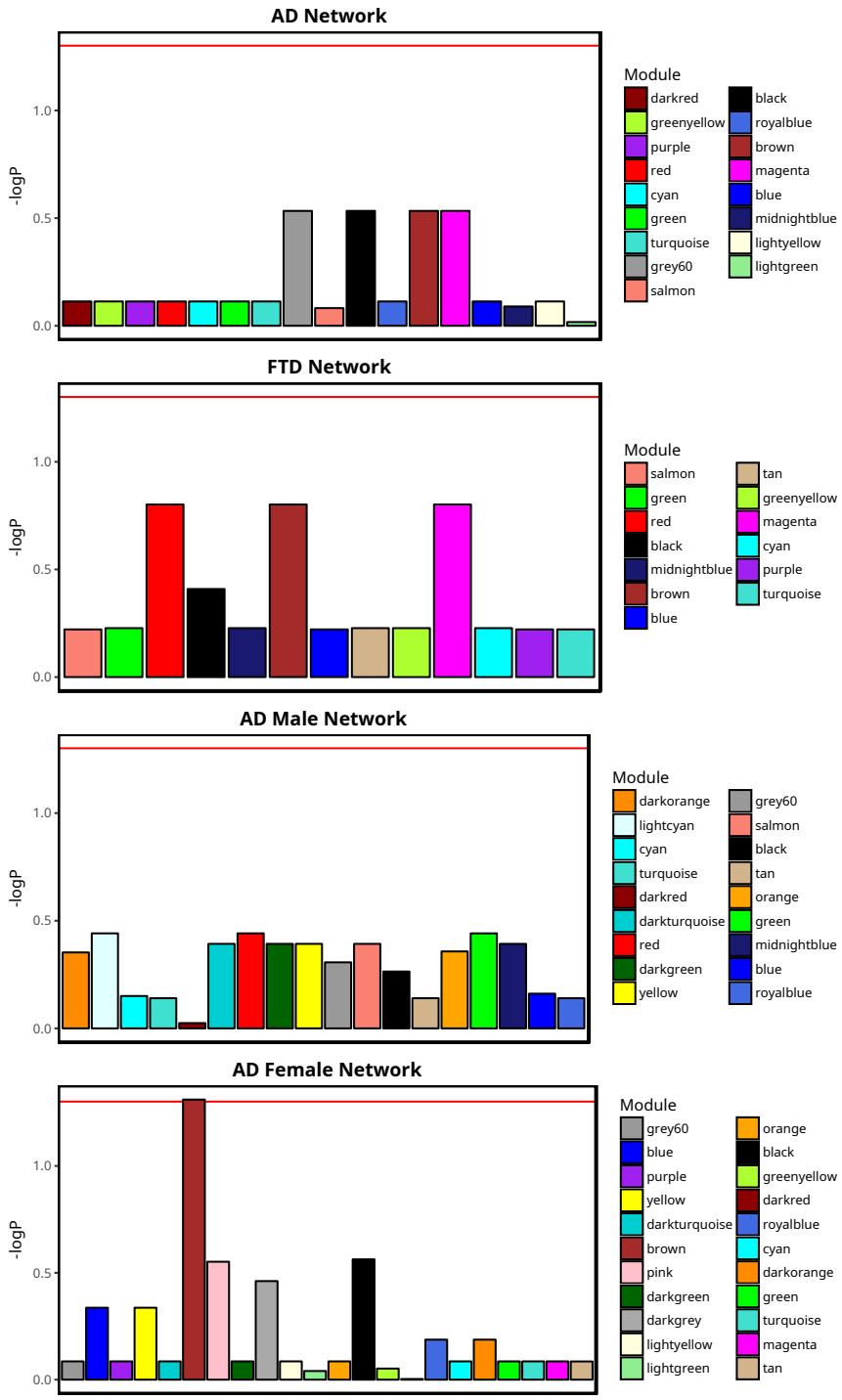

Figure 3-S12. Bar plots of enrichment of the top 300 genes in each WGCNA module for genetic risk for AD as estimated by sLDSR in all four networks. The -$\log_{10}$ p-value (adjusted for multiple comparisons) is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes.

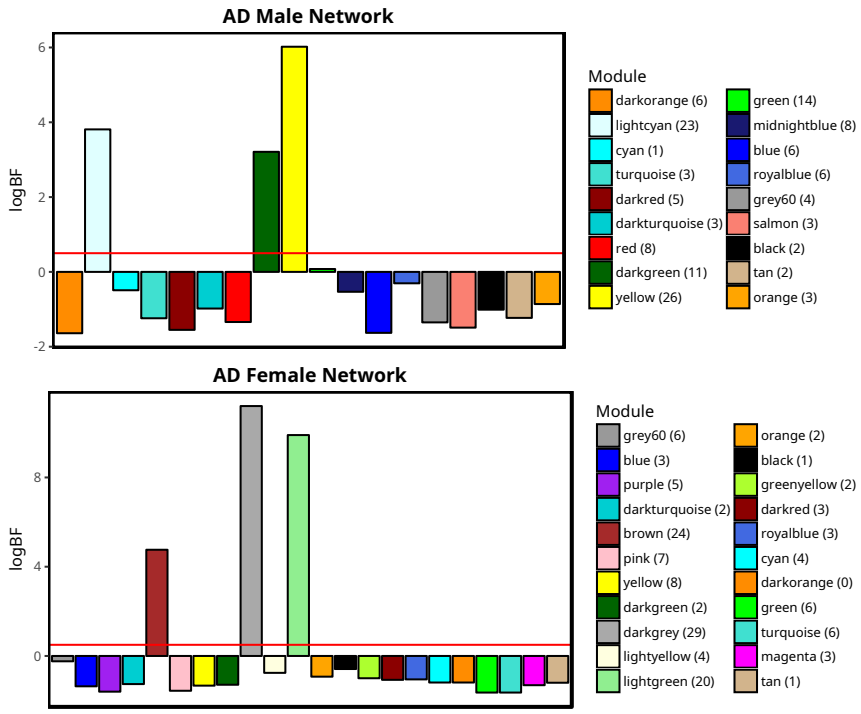

Figure 3-S13. Bar plots of enrichment of the top 300 genes in each WGCNA module for microglial genes in the AD male and female networks. The logBF is on the y-axis and modules are ordered by the hierarchical clustering of their eigengenes. The number of genes in each overlap is in parentheses next to the module name in the legend.

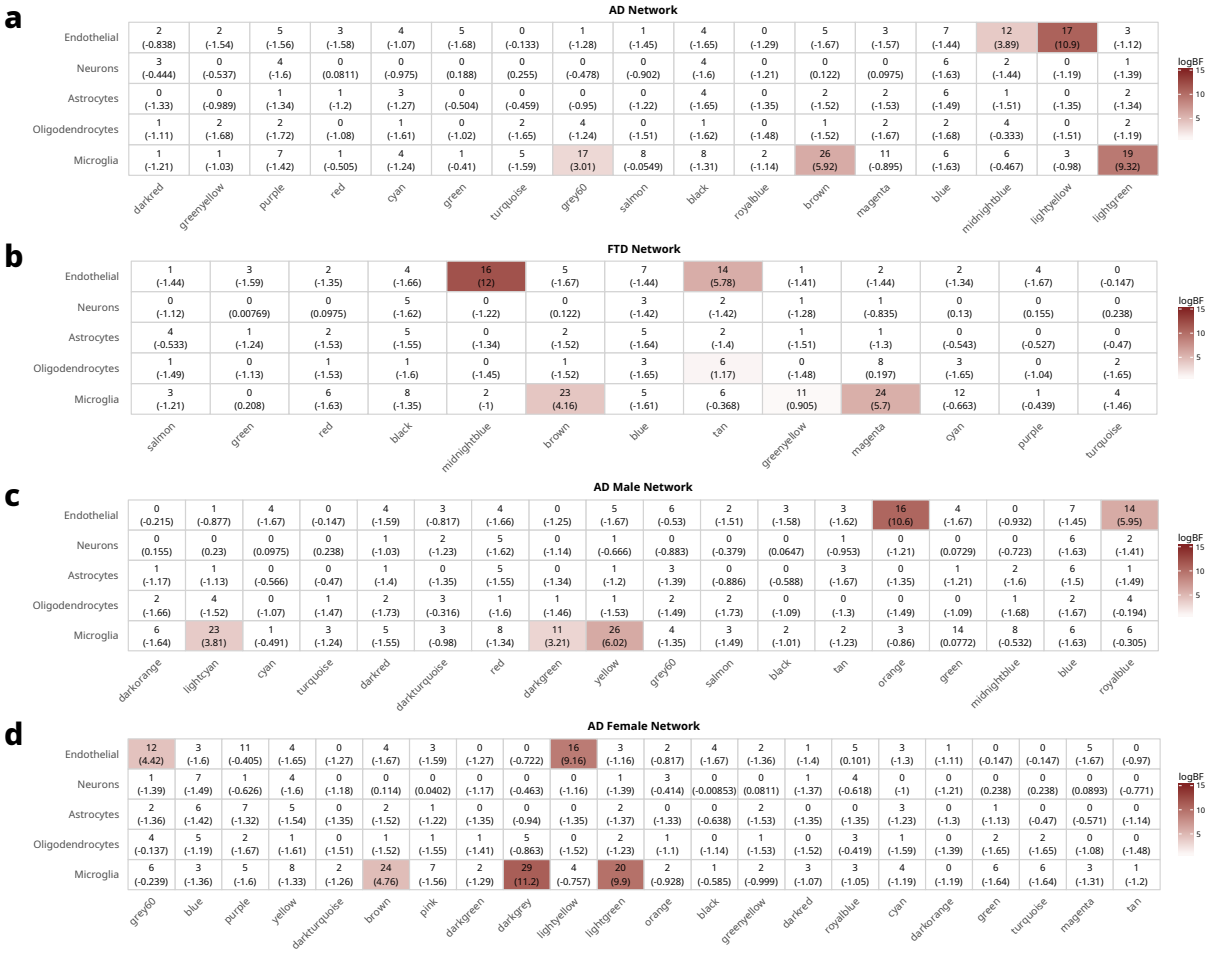

Figure 3-S14. Labeled heatmaps of enrichment of all network modules for all CNS cell types using the Zhang et. al. dataset. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses.

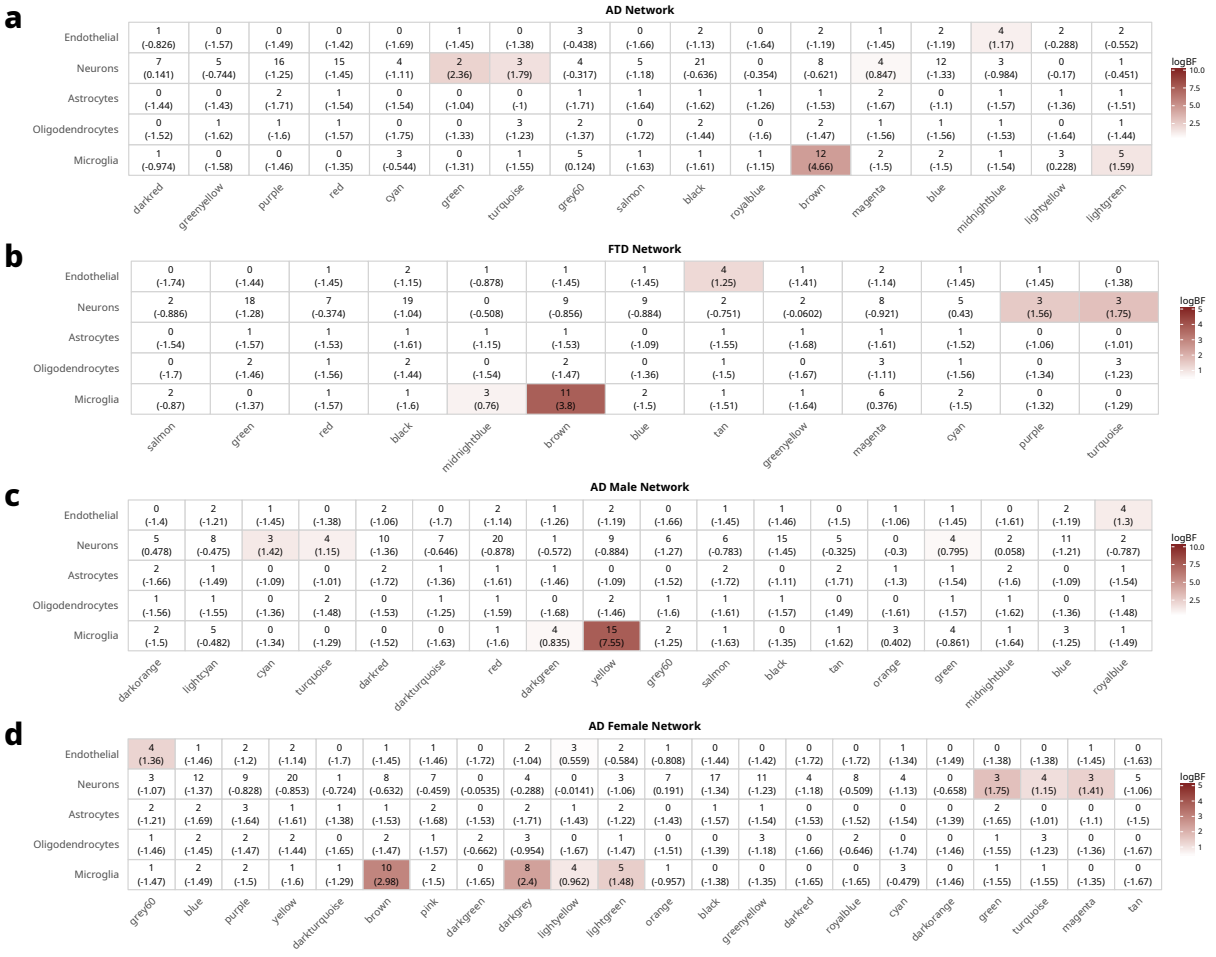

Figure 3-S15. Labeled heatmaps of enrichment of all network modules for all CNS cell types using the Wang et. al. dataset. Each cell shows the number of genes in the overlap and the logBF of the overlap significance in parentheses.

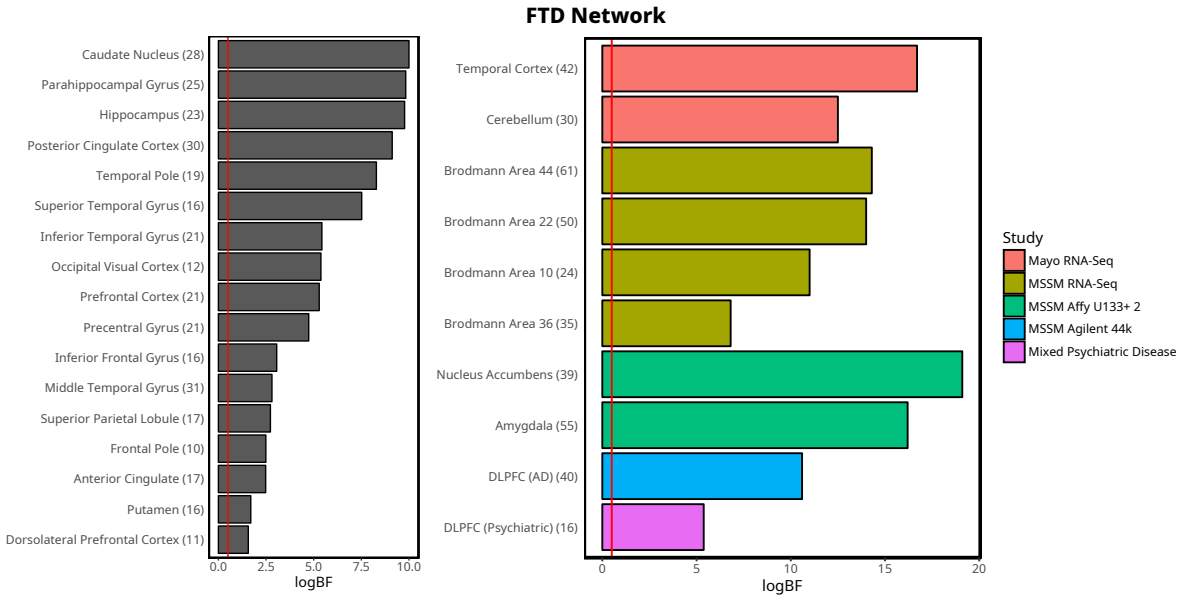

Figure 3-S16. Bar plots of enrichment of the top 300 genes in the brown module of the FTD and AD female networks and yellow module in the AD male network for the module in each post-mortem network most enriched for microglial genes. The logBF is on the x-axis and the number of genes in each overlap is in parentheses next to the y-axis label.

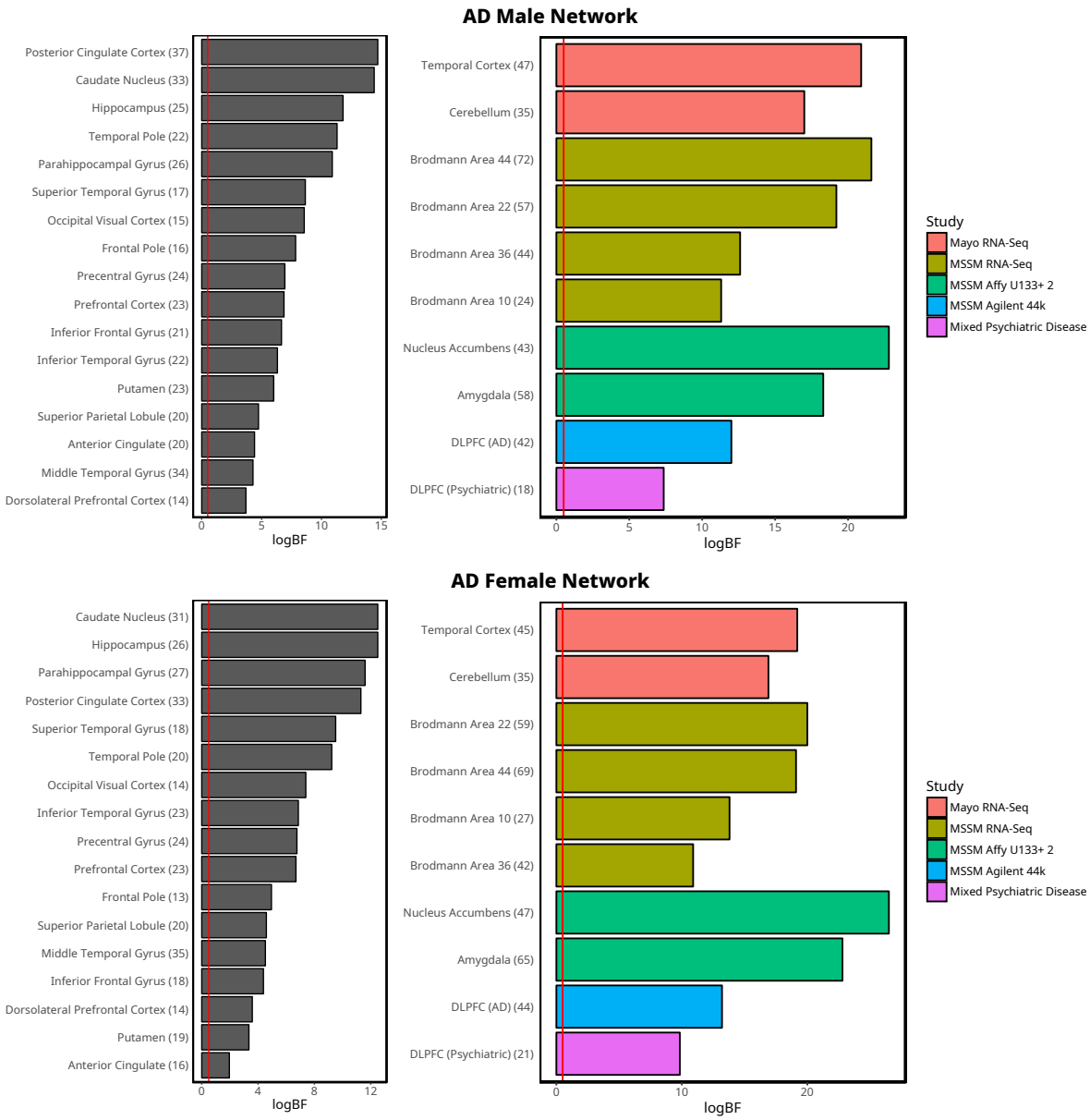

Figure 3-S16. (cont) Bar plots of enrichment of the top 300 genes in the brown module of the FTD and AD female networks and yellow module in the AD male network for the module in each post-mortem network most enriched for microglial genes. The logBF is on the x-axis and the number of genes in each overlap is in parentheses next to the y-axis label.

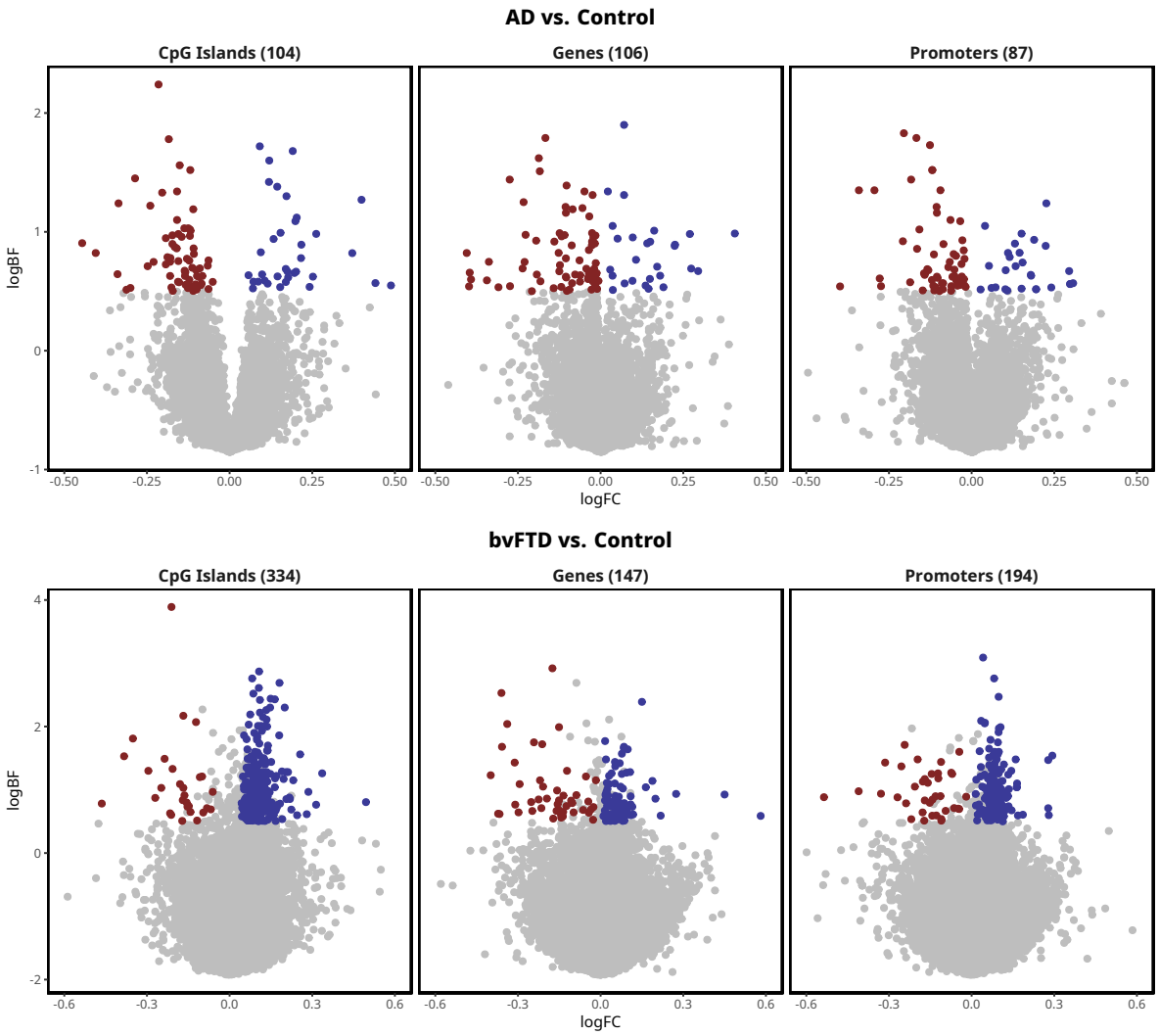

Figure 3-S17. Volcano plots of the log fold change (logFC) in CpG islands, promoters and gene bodies on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The methylation annotation and number of DM features are shown in the plot titles.

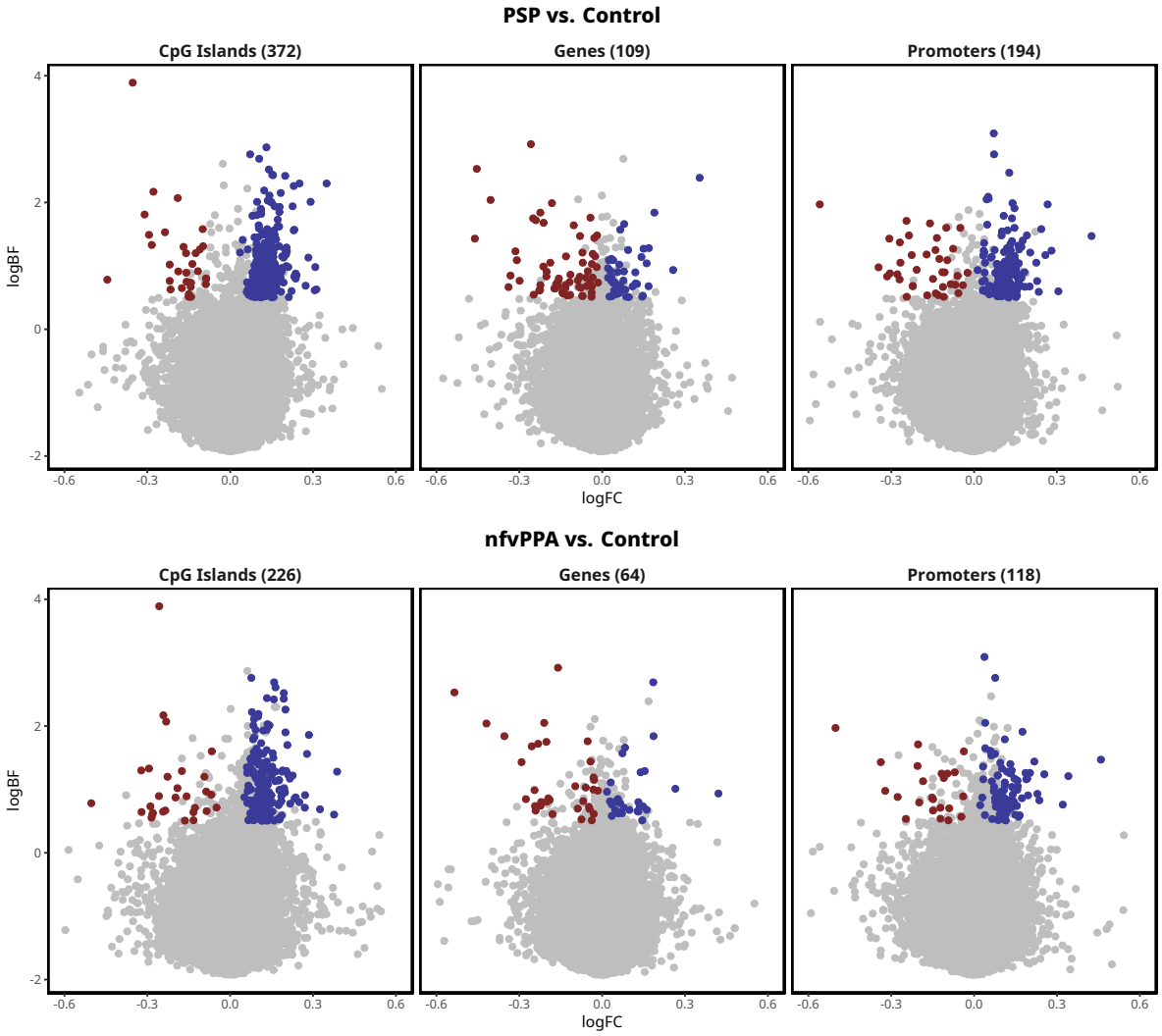

Figure 3-S17. (cont) Volcano plots of the log fold change (logFC) in CpG islands, promoters and gene bodies on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The methylation annotation and number of DM features are shown in the plot titles.

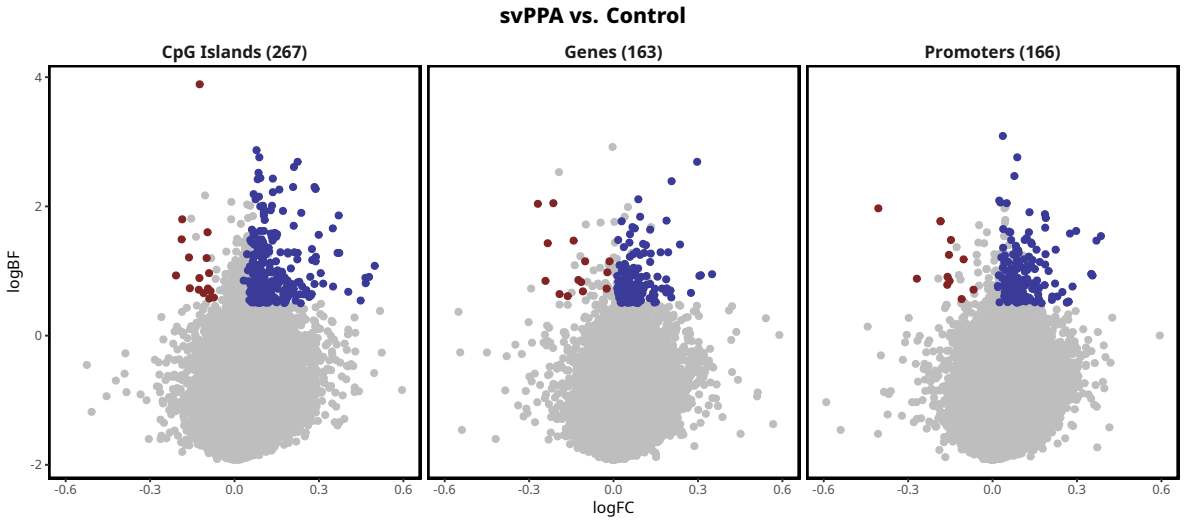

Figure 3-S17. (cont) Volcano plots of the log fold change (logFC) in CpG islands, promoters and gene bodies on the x-axis versus the $\log_{10}$ Bayes Factor (logBF) on the y-axis, with downregulated genes in blue and upregulated genes in red. Only genes with a significant pairwise probability greater than 0.95 were colored. The methylation annotation and number of DM features are shown in the plot titles.

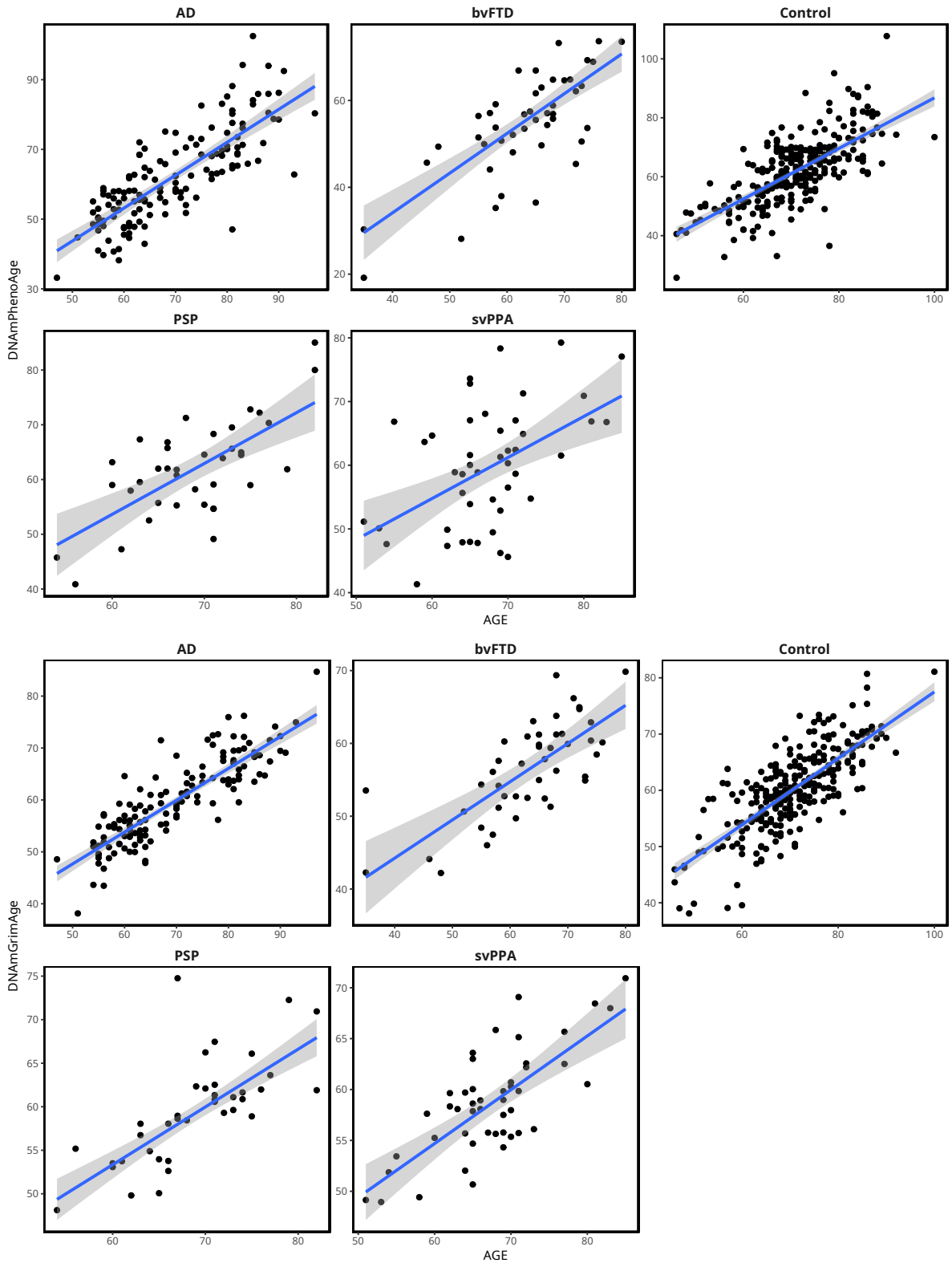

Figure 3-S18. Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge.

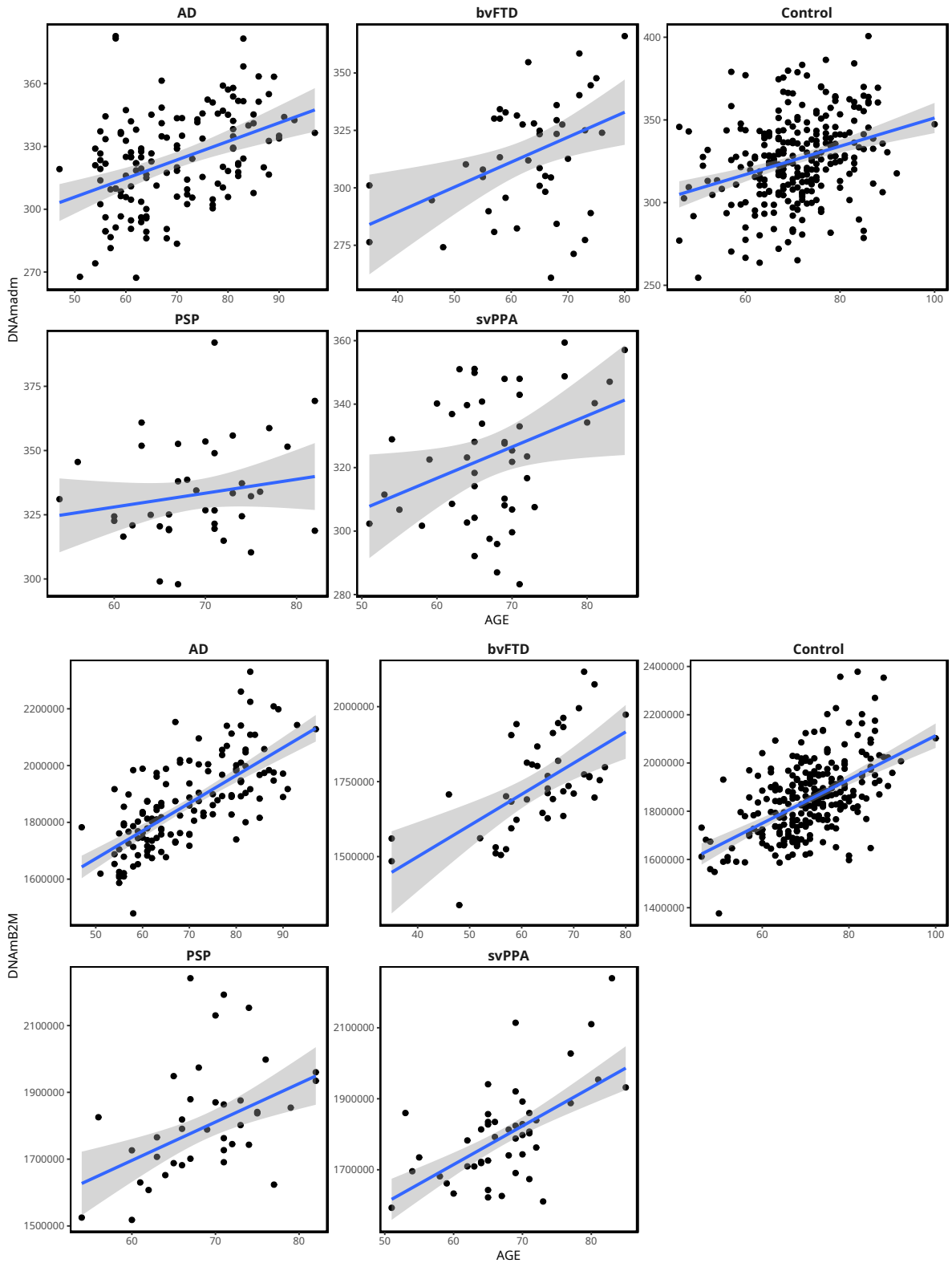

Figure 3-S18. (cont) Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge.

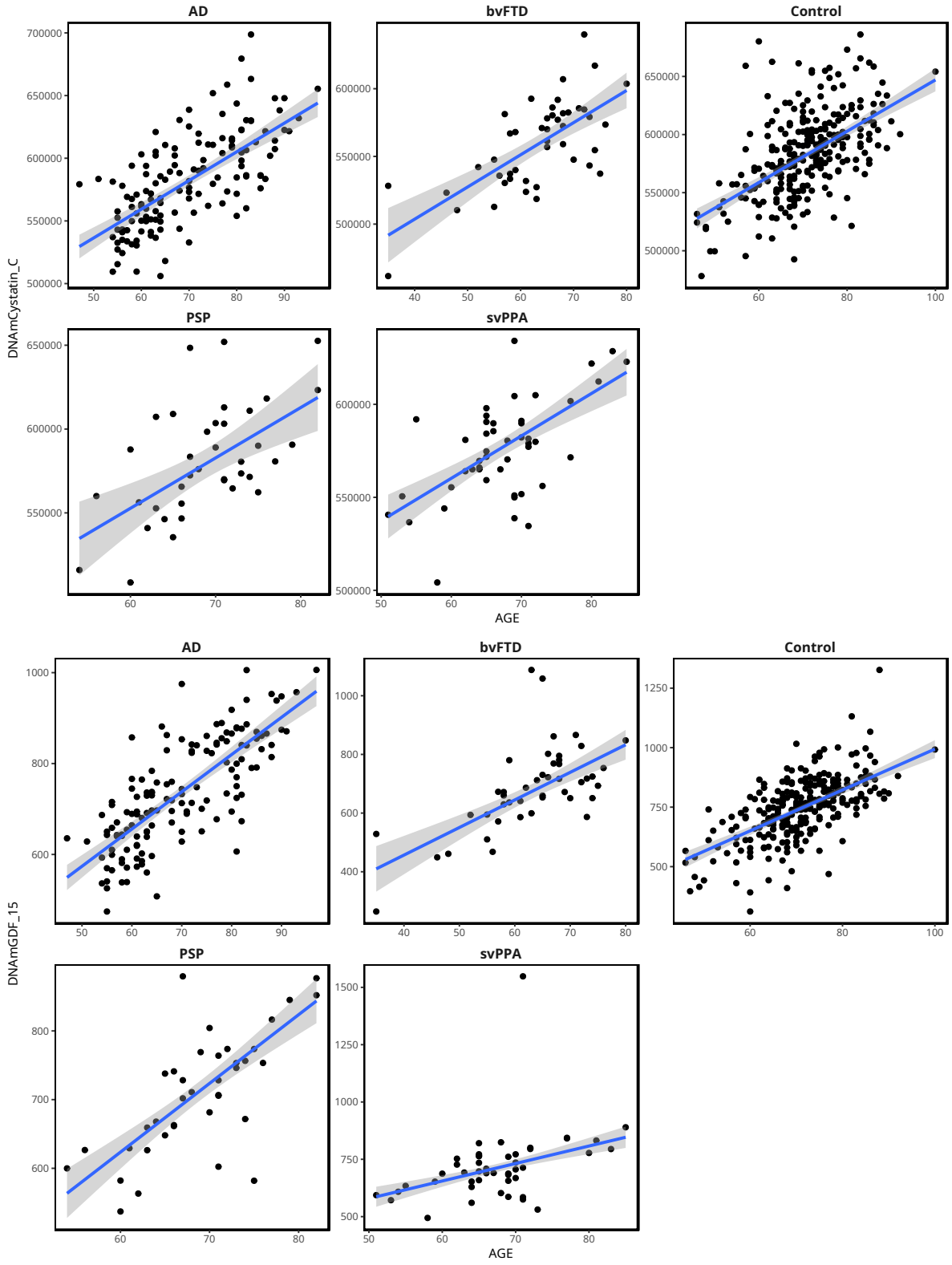

Figure 3-S18. (cont) Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge.

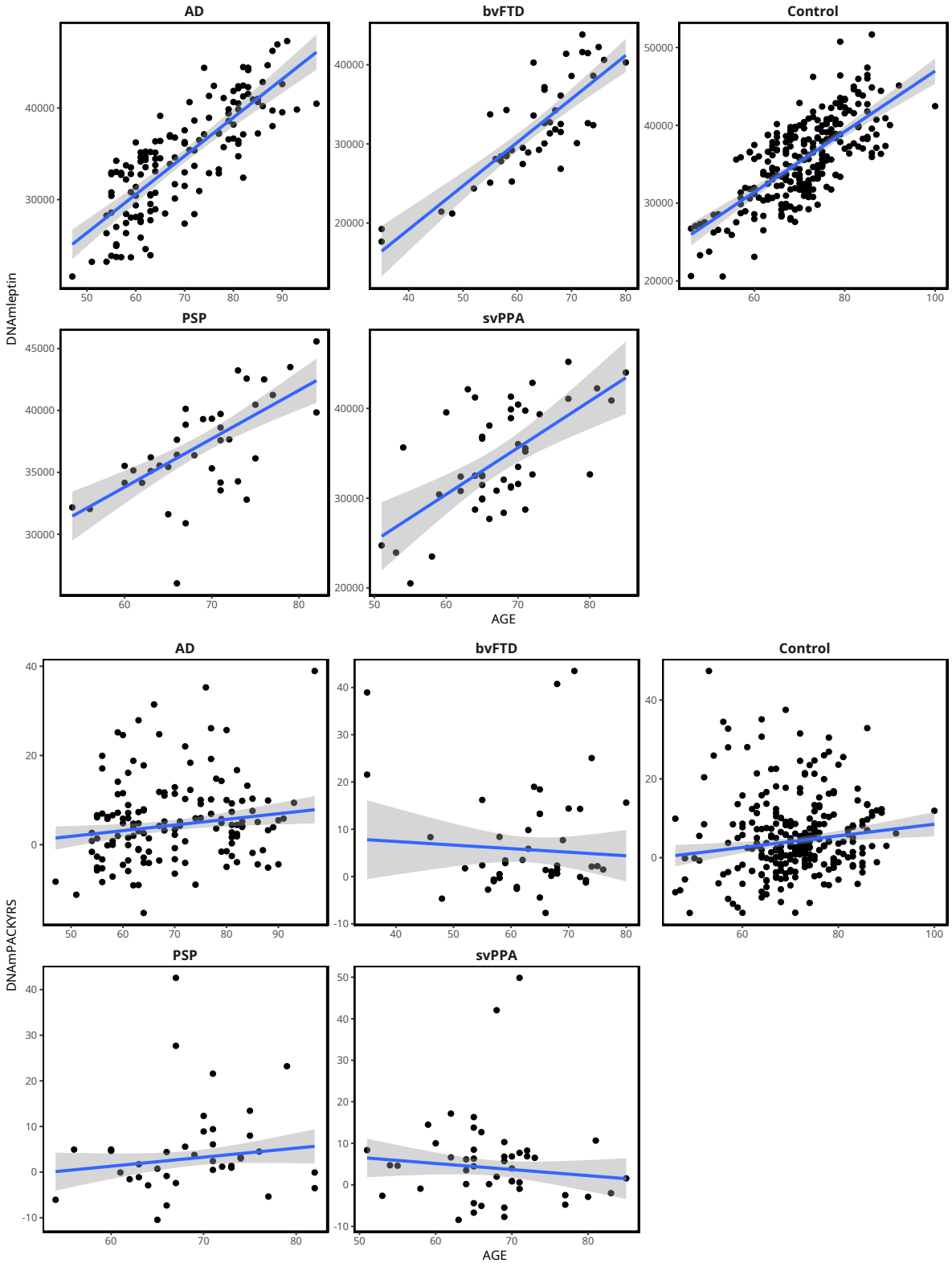

Figure 3-S18. (cont) Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge.

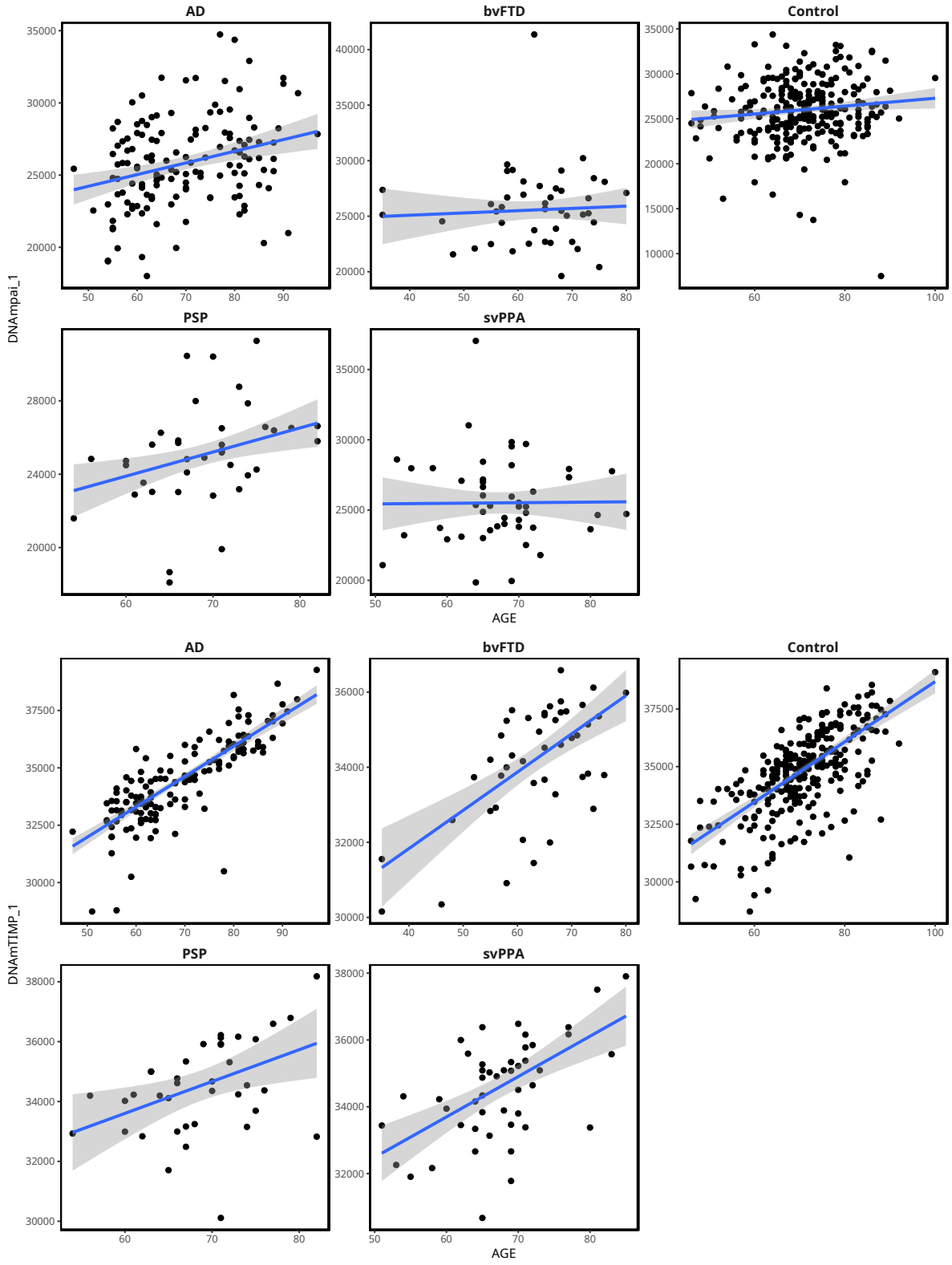

Figure 3-S18. (cont) Scatterplots of chronological age versus DNAmPhenoAge, DNAmGrimAge, and the 8 variables which are used to estimate DNAmGrimAge.

Bibliography

1. Elahi, F. M. & Miller, B. L. A clinicopathological approach to the diagnosis of dementia. en. *Nat. Rev. Neurol.* **13,** 457–476 (Aug. 2017).
2. Bang, J., Spina, S. & Miller, B. L. Frontotemporal dementia. en. *Lancet* **386,** 1672–1682 (Oct. 2015).
3. Ugalde, C. L., Finkelstein, D. I., Lawson, V. A. & Hill, A. F. Pathogenic mechanisms of prion protein, amyloid-β and α-synuclein misfolding: the prion concept and neurotoxicity of protein oligomers. en. *J. Neurochem.* **139,** 162–180 (Oct. 2016).
4. Clayton, K. A., Van Enoo, A. A. & Ikezu, T. Alzheimer's Disease: The Role of Microglia in Brain Homeostasis and Proteopathy. en. *Front. Neurosci.* **11,** 680 (Dec. 2017).
5. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. en. *Nat. Genet.* (July 2017).
6. Amor, S. *et al.* Inflammation in neurodegenerative diseases–an update. en. *Immunology* **142,** 151–166 (June 2014).
7. Lai, K. S. P. *et al.* Peripheral inflammatory markers in Alzheimer's disease: a systematic review and meta-analysis of 175 studies. en. *J. Neurol. Neurosurg. Psychiatry* **88,** 876–882 (Oct. 2017).
8. Dufek, M. *et al.* Serum inflammatory biomarkers in Parkinson's disease. en. *Parkinsonism Relat. Disord.* **15,** 318–320 (May 2009).
9. Fisher, D. W., Bennett, D. A. & Dong, H. Sexual dimorphism in predisposition to Alzheimer's disease. en. *Neurobiol. Aging* (Apr. 2018).

10. Spinelli, E. G. *et al.* Typical and atypical pathology in primary progressive aphasia variants. en. *Ann. Neurol.* **81,** 430–443 (Mar. 2017).

11. Eser, R. A. *et al.* Selective Vulnerability of Brainstem Nuclei in Distinct Tauopathies: A Postmortem Study. en. *J. Neuropathol. Exp. Neurol.* **77,** 149–161 (Feb. 2018).

12. Mahley, R. W. & Rall Jr, S. C. Apolipoprotein E: far more than a lipid transport protein. en. *Annu. Rev. Genomics Hum. Genet.* **1,** 507–537 (2000).

13. Freudenberg-Hua, Y., Li, W. & Davies, P. The Role of Genetics in Advancing Precision Medicine for Alzheimer's Disease-A Narrative Review. en. *Front. Med.* **5,** 108 (Apr. 2018).

14. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** Article17 (2005).

15. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

16. Plaisier, C. L. *et al.* A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLoS Genet.* **5** (2009).

17. Chandran, V. *et al.* A Systems-Level Analysis of the Peripheral Nerve Intrinsic Axonal Growth Program. en. *Neuron* **89,** 956–970 (Mar. 2016).

18. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 17402–17407 (2006).

19. Shirasaki, D. I. *et al.* Network organization of the huntingtin proteomic interactome in mammalian brain. *Neuron* **75,** 41–57 (2012).

20. Langfelder, P. *et al.* Integrated genomics and proteomics define huntingtin CAG length–dependent networks in mice. *Nat. Neurosci.* **19,** 623–633 (2016).

21. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. en. *Nucleic Acids Res.* **38,** 4218–4230 (July 2010).

22. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. en. *J. Neurosci.* **34,** 1420–1431 (Jan. 2014).

23. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. en. *Cell* **144,** 296–309 (Jan. 2011).

24. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. en. *Bioinformatics* **29,** 2211–2212 (Sept. 2013).

25. Baron, U. *et al.* DNA methylation analysis as a tool for cell typing. en. *Epigenetics* **1,** 55–60 (Jan. 2006).

26. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. en. *PLoS Comput. Biol.* **11,** e1004219 (Apr. 2015).

27. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).

28. Van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **10** (2016).

29. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. en. *Cell* **162,** 516–526 (July 2015).

30. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. en. *Nat. Genet.* **44,** 981–990 (Sept. 2012).

31. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. en. *Nature* **511,** 421–427 (July 2014).

32. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. en. *Nat. Genet.* **50,** 668–681 (May 2018).

33. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. en. *Nat. Genet.* **50,** 621–629 (Apr. 2018).

34. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. en. *Neuron* **89,** 37–53 (Jan. 2016).

35. Wang, M. *et al.* Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. en. *Genome Med.* **8,** 104 (Nov. 2016).

36. Leech, R. & Sharp, D. J. The role of the posterior cingulate cortex in cognition and disease. en. *Brain* **137,** 12–32 (Jan. 2014).

37. Jiji, S., Smitha, K. A., Gupta, A. K., Pillai, V. P. M. & Jayasree, R. S. Segmentation and volumetric analysis of the caudate nucleus in Alzheimer's disease. en. *Eur. J. Radiol.* **82,** 1525–1530 (Sept. 2013).

38. Narayanan, M. *et al.* Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. en. *Mol. Syst. Biol.* **10,** 743 (July 2014).

39. Seyfried, N. T. *et al.* A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. en. *Cell Syst* **4,** 60–72.e4 (Jan. 2017).

40. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).

41. Garbett, K. *et al.* Immune transcriptome alterations in the temporal cortex of subjects with autism. en. *Neurobiol. Dis.* **30,** 303–311 (June 2008).

42. Chow, M. L. *et al.* Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. en. *PLoS Genet.* **8,** e1002592 (Mar. 2012).

43. Chen, C. *et al.* Two gene co-expression modules differentiate psychotics and controls. en. *Mol. Psychiatry* **18,** 1308–1314 (Dec. 2013).

44. Maycox, P. R. *et al.* Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. en. *Mol. Psychiatry* **14,** 1083–1094 (Dec. 2009).

45. Iwamoto, K., Bundo, M. & Kato, T. Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. en. *Hum. Mol. Genet.* **14,** 241–253 (Jan. 2005).

46. Narayan, S. *et al.* Molecular profiles of schizophrenia in the CNS at different stages of illness. en. *Brain Res.* **1239,** 235–248 (Nov. 2008).

47. Chang, L.-C. *et al.* A conserved BDNF, glutamate- and GABA-enriched gene module related to human depression identified by coexpression meta-analysis and DNA variant genome-wide association studies. en. *PLoS One* **9,** e90980 (Mar. 2014).

48. Ponomarev, I., Wang, S., Zhang, L., Harris, R. A. & Mayfield, R. D. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. en. *J. Neurosci.* **32,** 1884–1897 (Feb. 2012).

49. Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. en. *Science* **359,** 693–697 (Feb. 2018).

50. Allen, M. *et al.* Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. en. *Sci Data* **3,** 160089 (Oct. 2016).

51. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. en. *BMC Bioinformatics* **11,** 587 (Nov. 2010).

52. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14,** R115 (2013).

53. Thériault, P., ElAli, A. & Rivest, S. The dynamics of monocytes and microglia in Alzheimer's disease. en. *Alzheimers. Res. Ther.* **7,** 41 (Apr. 2015).

54. Majerova, P. *et al.* Microglia display modest phagocytic capacity for extracellular tau oligomers. en. *J. Neuroinflammation* **11,** 161 (Sept. 2014).

55. Bengoa-Vergniory, N., Roberts, R. F., Wade-Martins, R. & Alegre-Abarrategui, J. Alpha-synuclein oligomers: a new hope. en. *Acta Neuropathol.* **134,** 819–838 (Dec. 2017).

56. Zondler, L. *et al.* Impaired activation of ALS monocytes by exosomes. en. *Immunol. Cell Biol.* **95,** 207–214 (Feb. 2017).

57. Sundermann, E. E. *et al.* Better verbal memory in women than men in MCI despite similar levels of hippocampal atrophy. en. *Neurology* **86,** 1368–1376 (Apr. 2016).

58. for the Alzheimer's Disease Neuroimaging Initiative *et al.* Does the Female Advantage in Verbal Memory Contribute to Underestimating Alzheimer's Disease Pathology in Women versus Men? *J. Alzheimers. Dis.* **56** (ed Pike, K.) 947–957 (Feb. 2017).

59. Yokoyama, A. S., Rutledge, J. C. & Medici, V. DNA methylation alterations in Alzheimer's disease. en. *Environ Epigenet* **3,** dvx008 (May 2017).

60. De Jager, P. L. *et al.* Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. en. *Nat. Neurosci.* **17,** 1156–1163 (Sept. 2014).

61. Du, P., Kibbe, W. A. & Lin, S. M. lumi: A pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–1548 (2008).

62. Lin, S. M., Du, P., Huber, W. & Kibbe, W. A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* **36,** 1–9 (2008).

63. Miller, J. a. *et al.* Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics* **12,** 322 (2011).

64. Dong, J. & Horvath, S. Understanding network concepts in modules. *BMC Syst. Biol.* **1,** 24 (2007).

65. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** 1724–1735 (2007).

66. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11,** 1138–1140 (2014).

67. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41,** e90–e90 (2013).

68. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29,** 189–196 (2013).

69. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56,** 356–374 (2012).

70. Rouder, J. N. & Morey, R. D. Default Bayes Factors for Model Selection in Regression. *Multivariate Behav. Res.* **47,** 877–903 (2012).

71. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13,** 328 (2012).

72. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24,** 719–720 (2008).

73. Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z. & Clark, H. F. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. en. *PLoS One* **4,** e6098 (July 2009).

74. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14,** 293 (2013).

75. Houseman, E. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).

76. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–W97 (July 2016).

77. Jamil, T. *et al.* Default "Gunel and Dickey" Bayes factors for contingency tables. en. *Behav Res* **49,** 638–652 (Apr. 2017).

78. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. en. *Nature* **526,** 68–74 (Oct. 2015).

79. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. en. *Nature* **467,** 52–58 (Sept. 2010).