Weill Cornell Medicine

Analysis of Next Generation Sequencing Data Final Project Presentation

Differential Expression Analysis of RNA-seq Data Derived from Alzheimer's Patient Microglia

Jake Sauter
M.S. Computational Biology

04.19.21

Presentation Structure

1. Introduction
2. Results
3. Methods
4. Discussion

Weill Cornell Medicine

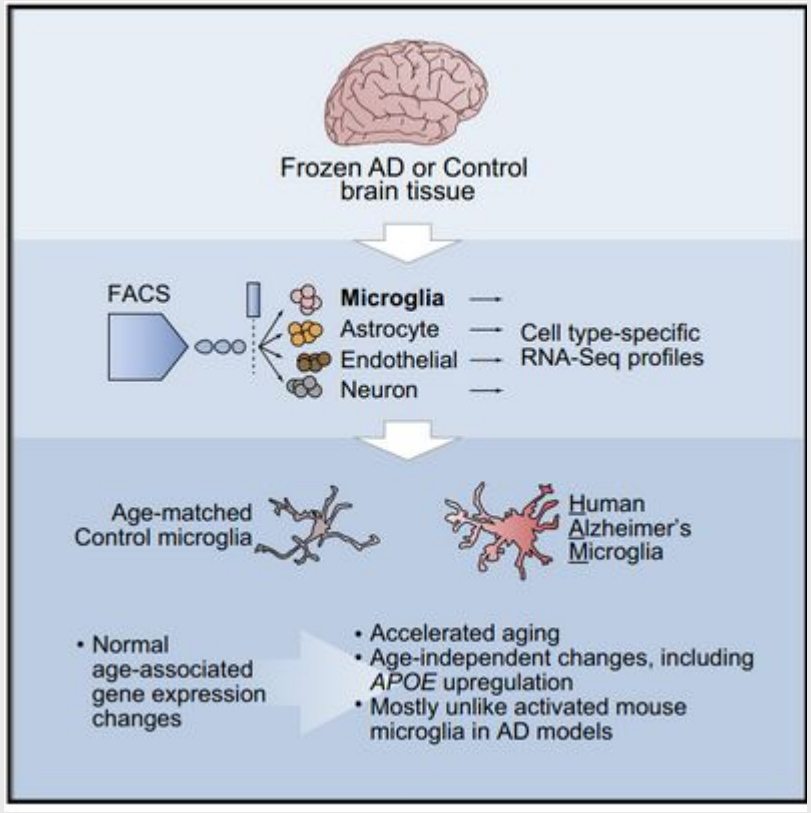
Introduction

Alzheimer's Disease

- **Alzheimer’s disease (AD) effects approximately 5.8 million people in the the United States ages 65 and older**
  - **Early sign and symptoms of the disease centering around forgetfulness**
  - **Impairment of memory follows, until loss of ability to carry out everyday tasks**
- **There is currently no treatment that cures AD or alters the disease process in the brain**
  - **Important to continue research in the area**
  - **Possibly NAD therapies:** **Nicotinamide riboside restores cognition through an upregulation of proliferator-activated receptor-γ coactivator 1α regulated β-secretase 1 degradation and mitochondrial gene expression in Alzheimer's mouse models**

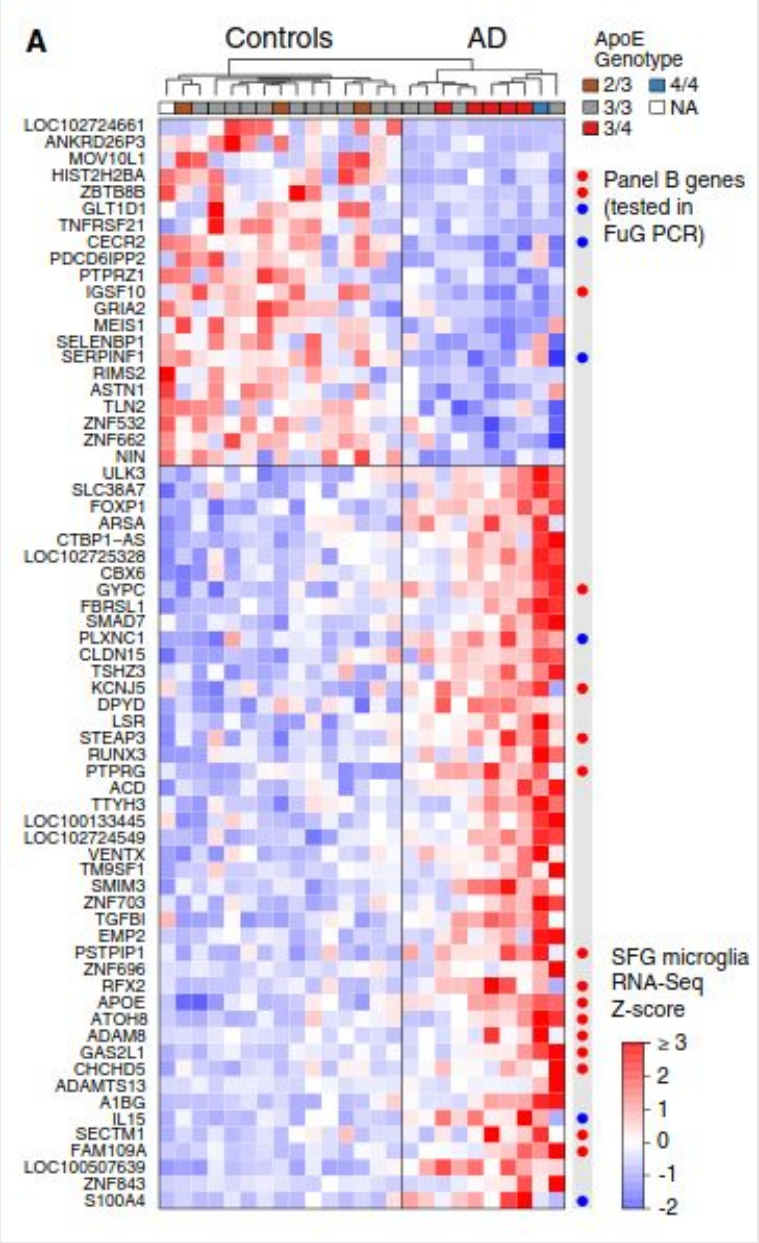
Alzheimer's Patient Microglia Exhibit Enhanced Aging and Unique Transcriptional Activation

- **Recent genetic studies of humans have identified brain specific myeloid cells (microglia) as a potential key cell type controlling an individual's risk of acquiring Alzheimer's Disease**
- **This study identified 45 differentially expressed genes between Control and AD Microglia (Myeloid) cells**



Study Hypothesis

- **Given scrutinous quality control and industry standard tools such as STAR and DESeq2, I will achieve a different set of differentially expressed genes than shown final in the reference publication**
  - **"Sorted cell and whole tissue RNA-Seq data were analyzed using the GSNAP aligner and HTSeqGenie"**



Weill Cornell Medicine

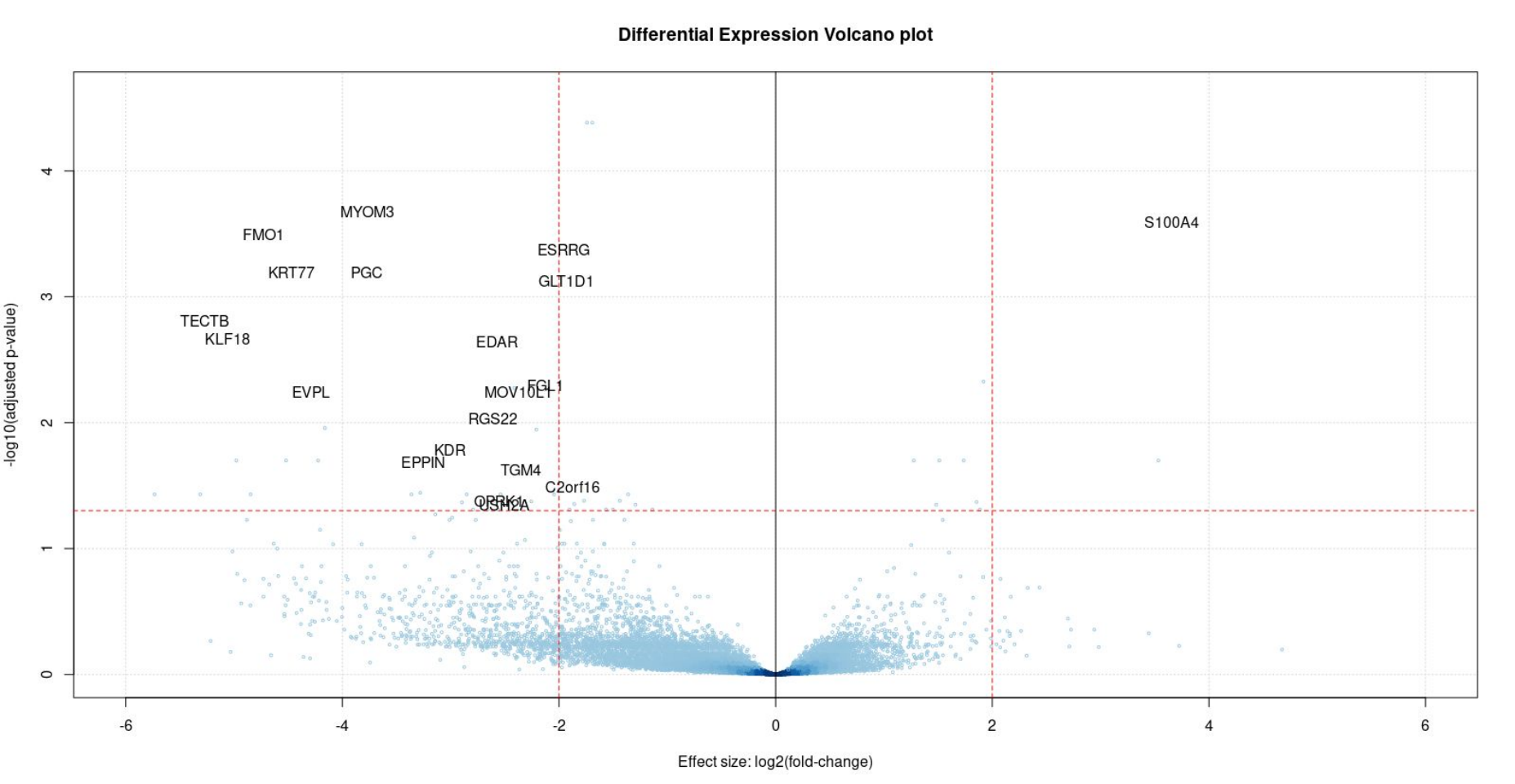
Results

Differentially Expressed Genes

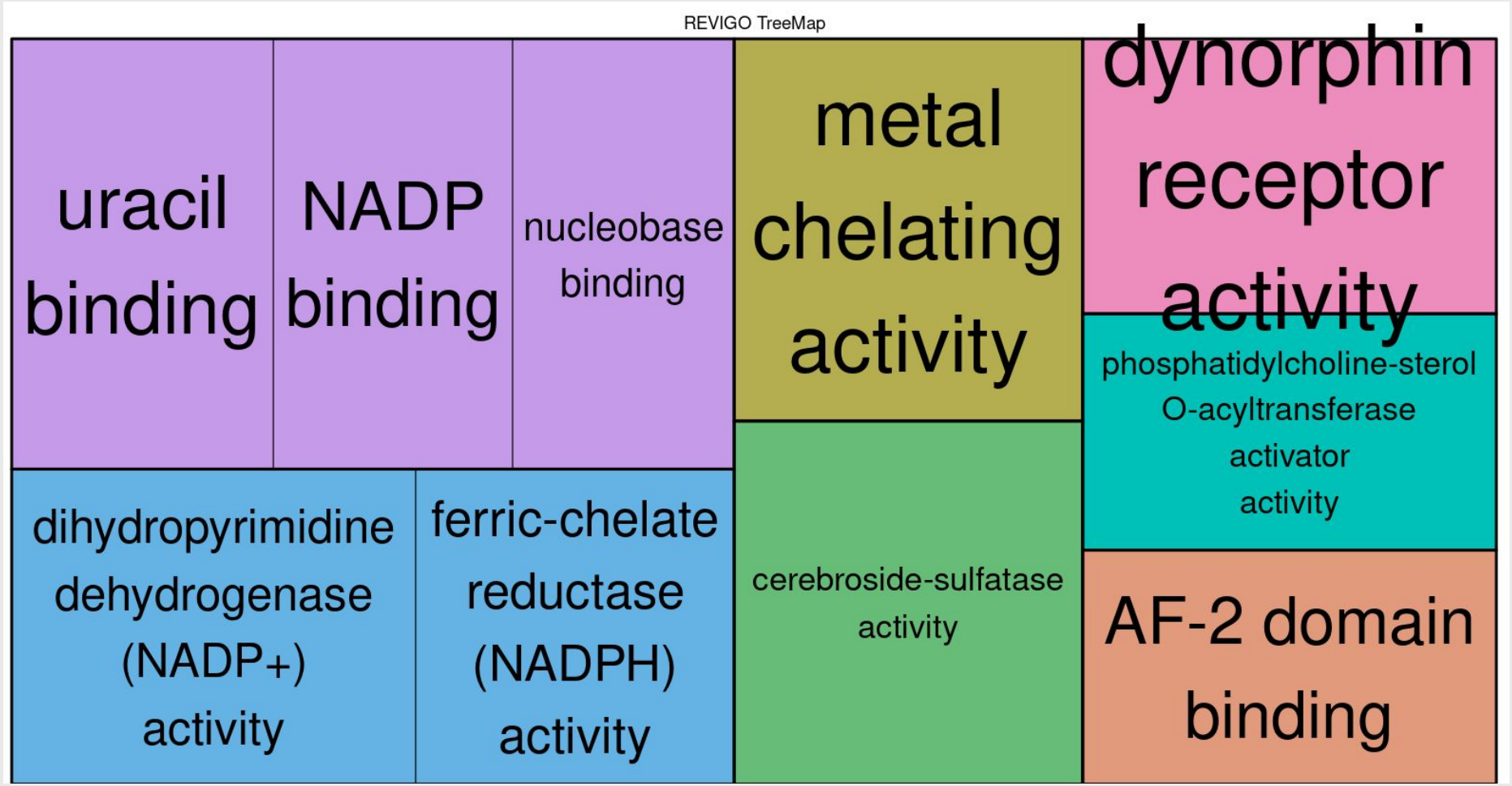
- **Confirmed 14 genes to be differentially expressed between control and AD clinical groups**
- **Same 14 genes appear in paper as gathered literature-supported AD risk genes**

	padj	log2FoldChange
SERPINF1	0.000041	1.7407
CECR2	0.000041	1.6923
TGFBI	0.004718	-1.9188
GLT1D1	0.005319	2.4133
ARSA	0.019995	-1.2752
PTPRG	0.019995	-1.7361
S100A4	0.019995	-3.5340
APOE	0.019995	-1.5115
MOV10L1	0.037131	2.5397
ASTN1	0.041758	1.4381
DPYD	0.042672	-1.8551
STEAP3	0.044855	-1.4818
IGSF10	0.048952	1.5660
PSTPIP1	0.048952	-1.8852

Differential Expression Volcano Plot



Gene Ontology Treemap -- Revigo



Weill Cornell Medicine

Methods

Downloading Data

- **Data Available through NCBI GEO Accession Number**
  - **SRA Runs Selector**
  - **113 isolated-cell samples**
  - **Age, APOE genotype, Sex, PMI metadata available**

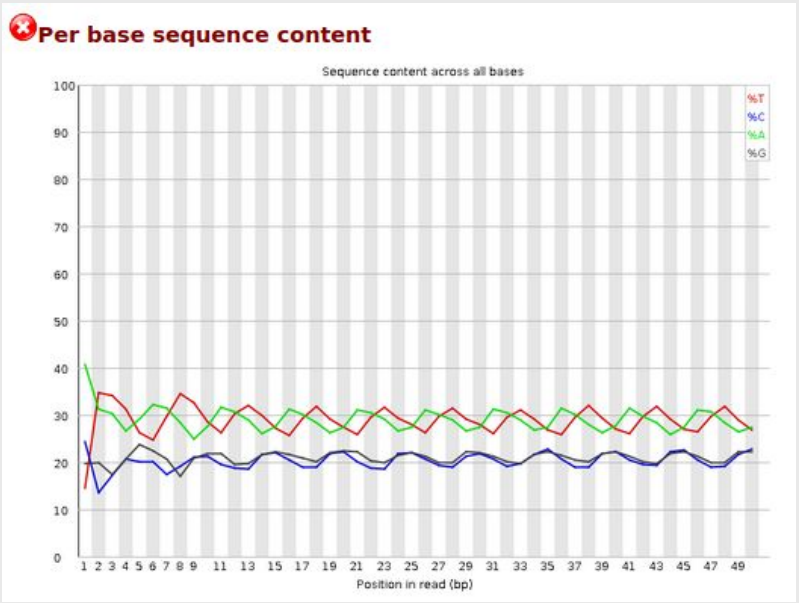
Run	BioSample	APOE	Bases	Bytes	Cell_type	Diagnosis
SRR8440443	SAMN10741236	3/3	2.27 G	694.72 Mb	myeloid	Control
SRR8440444	SAMN10741241	3/3	2.19 G	683.21 Mb	endothelial	Control
SRR8440445	SAMN10741240	3/3	2.17 G	700.85 Mb	neuron	Control
SRR8440446	SAMN10741239	3/4	2.44 G	791.87 Mb	neuron	AD
SRR8440447	SAMN10741238	NA	2.14 G	648.79 Mb	myeloid	Control
SRR8440448	SAMN10741237	3/4	2.16 G	664.66 Mb	myeloid	AD
SRR8440449	SAMN10741235	2/3	2.16 G	653.94 Mb	myeloid	Control
SRR8440450	SAMN10741234	2/3	2.99 G	940.45 Mb	endothelial	Control
SRR8440451	SAMN10741233	3/3	2.91 G	902.51 Mb	endothelial	AD
SRR8440452	SAMN10741232	3/3	2.46 G	793.50 Mb	neuron	Control
SRR8440453	SAMN10741231	3/3	2.34 G	766.49 Mb	neuron	AD
SRR8440454	SAMN10741230	3/3	2.47 G	789.37 Mb	neuron	AD
SRR8440455	SAMN10741229	3/3	2.24 G	829.08 Mb	endothelial	Control
SRR8440456	SAMN10741228	3/3	4.10 G	1.24 Gb	astrocyte	Control
SRR8440457	SAMN10741286	2/3	10.49 G	3.24 Gb	astrocyte	Control
SRR8440458	SAMN10741285	3/3	2.01 G	650.22 Mb	neuron	Control

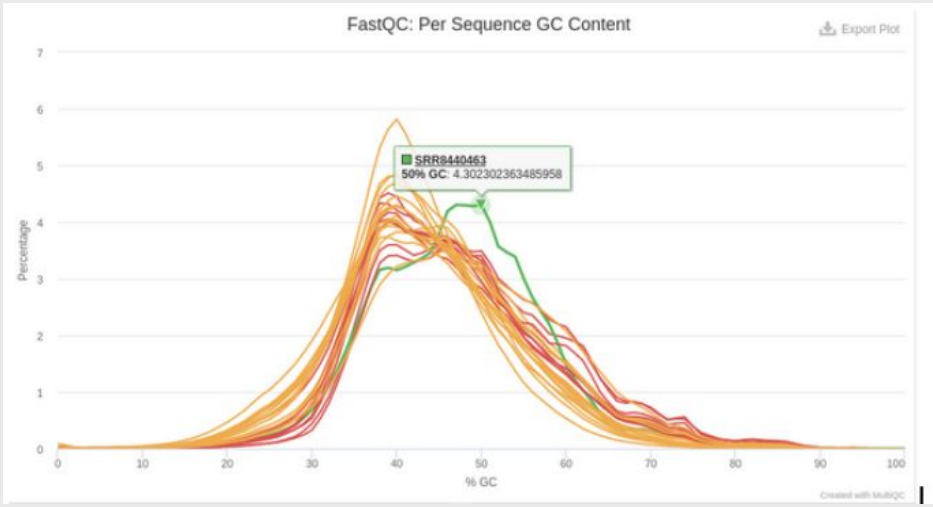
Raw Data FastQC

- At first detected Illumina adapters
  - TrimGalore
- Some samples had cyclic GC content
  - No per-tile information available
- Skewed GC content found in a sample
  - 72 year old male with AD and 4/4 APOE genotype

Overrepresented sequences

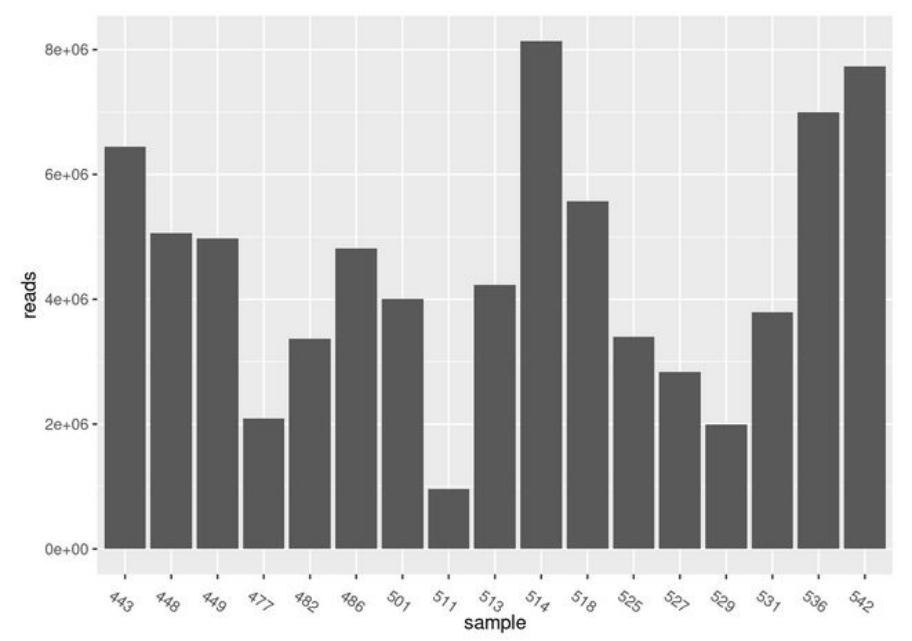
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC	121904	0.34849652269629916	TruSeq Adapter, Index 5 (100% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATG	43845	0.12534313917196513	TruSeq Adapter, Index 5 (100% over 49bp)





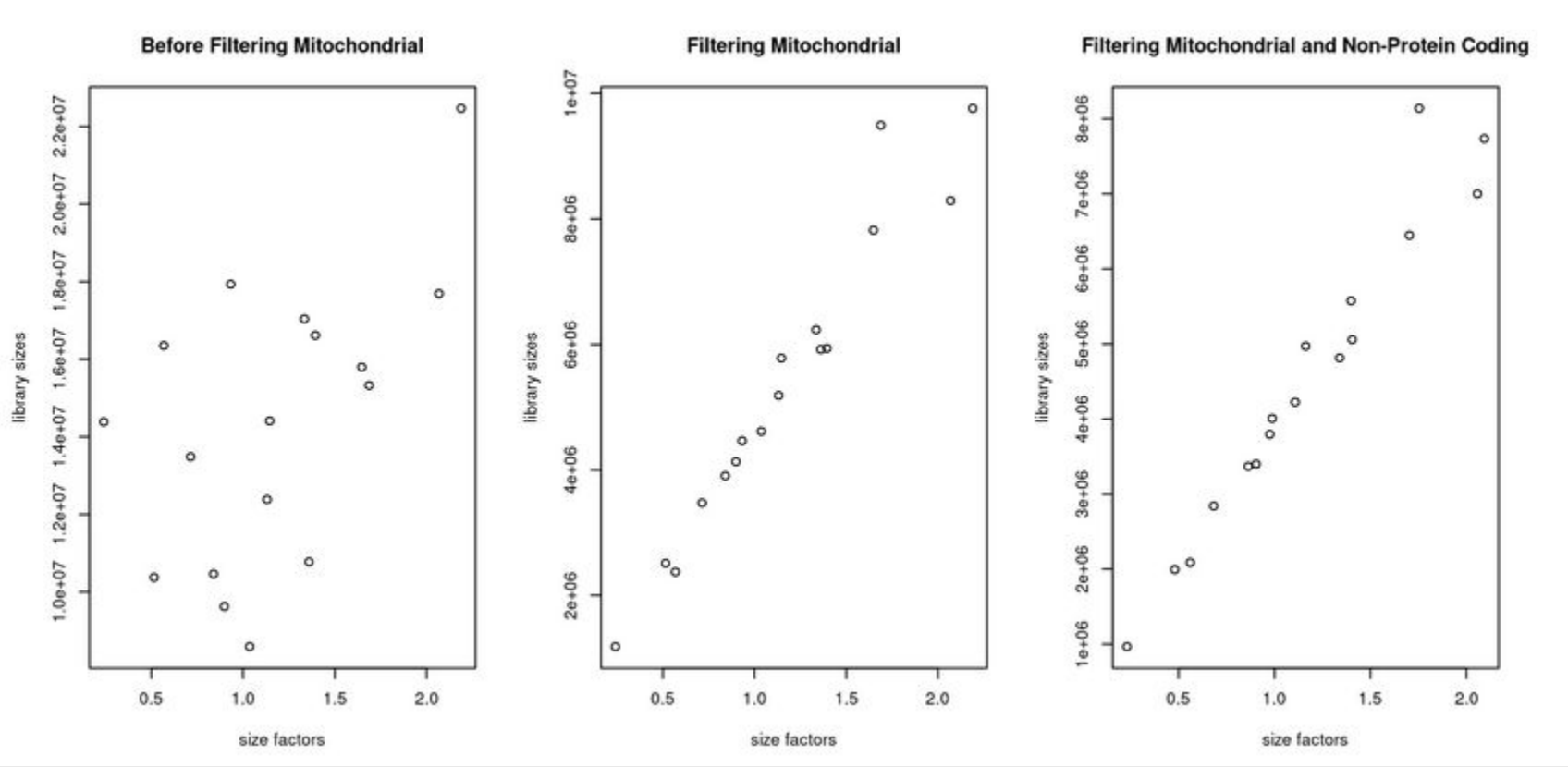
RNA-Seq Alignment with STAR

- **Same 72 YO male sample seen with skewed GC content also has 33.79% of reads failing to map to human HG38 Genome**
- **Other samples look fairly uniform**



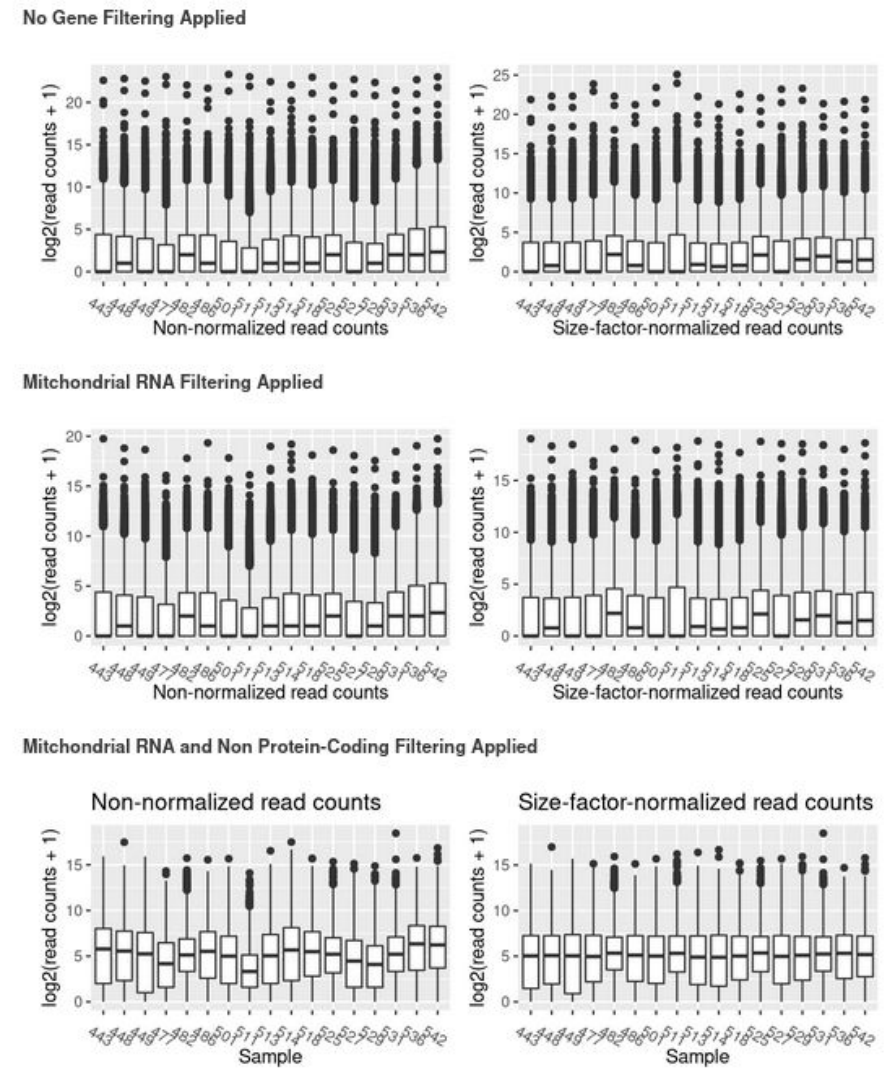

```
Percentage of unmapped reads for SRR8440443: 5.90%
Percentage of unmapped reads for SRR8440447: 4.17%
Percentage of unmapped reads for SRR8440448: 4.04%
Percentage of unmapped reads for SRR8440449: 8.38%
Percentage of unmapped reads for SRR8440463: 33.79%
Percentage of unmapped reads for SRR8440477: 5.14%
Percentage of unmapped reads for SRR8440482: 6.54%
Percentage of unmapped reads for SRR8440484: 3.65%
Percentage of unmapped reads for SRR8440486: 8.13%
Percentage of unmapped reads for SRR8440488: 5.19%
Percentage of unmapped reads for SRR8440501: 6.98%
Percentage of unmapped reads for SRR8440511: 5.47%
Percentage of unmapped reads for SRR8440513: 6.79%
Percentage of unmapped reads for SRR8440514: 4.97%
Percentage of unmapped reads for SRR8440517: 5.76%
Percentage of unmapped reads for SRR8440518: 5.02%
Percentage of unmapped reads for SRR8440524: 11.28%
Percentage of unmapped reads for SRR8440525: 9.51%
Percentage of unmapped reads for SRR8440527: 7.05%
Percentage of unmapped reads for SRR8440529: 6.67%
Percentage of unmapped reads for SRR8440531: 10.64%
Percentage of unmapped reads for SRR8440536: 7.61%
Percentage of unmapped reads for SRR8440538: 7.43%
Percentage of unmapped reads for SRR8440539: 12.34%
Percentage of unmapped reads for SRR8440542: 8.16%
```

Read Count Normalization -- Size Factors



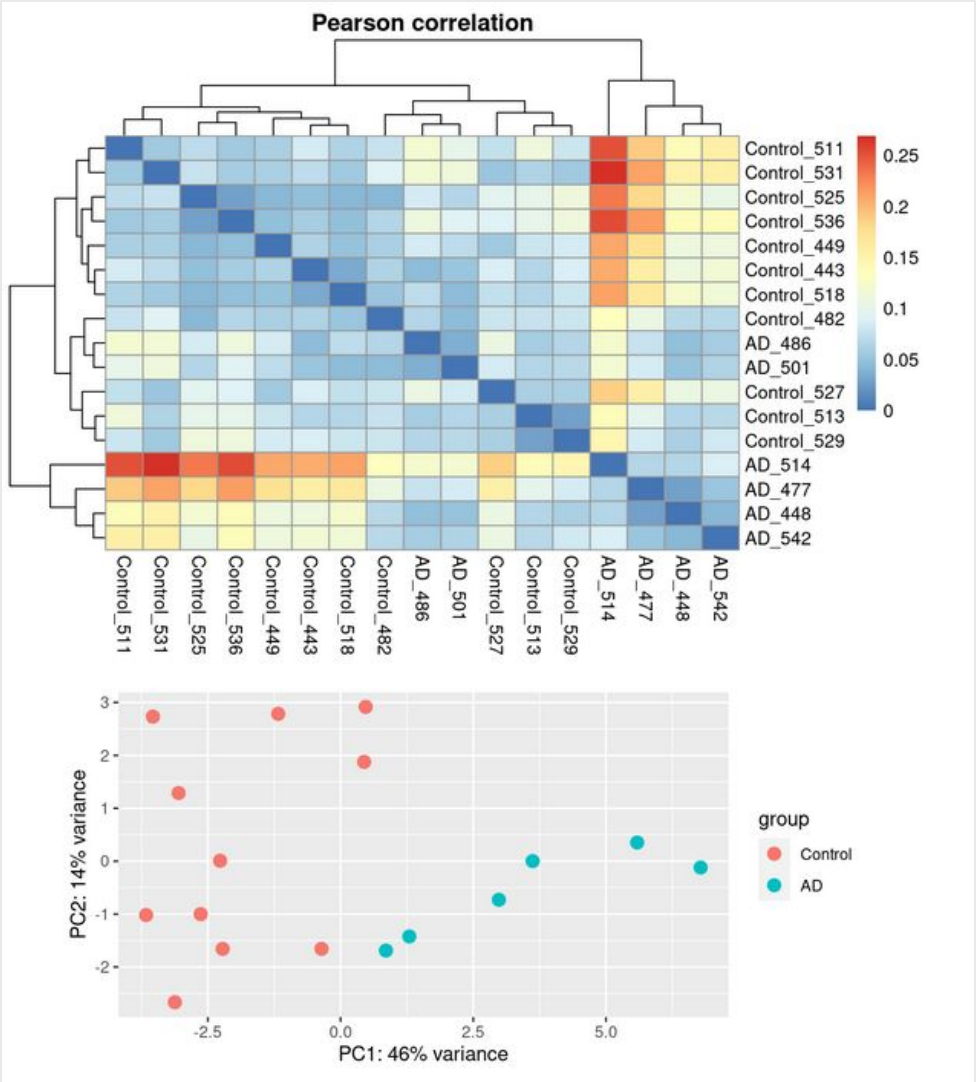
Gene Filtering

- Tissue from frozen samples
  - High lncRNA and Mitochondrial RNA content
- Filtered for only non-Mitochondrial protein-coding genes
  - Study does not, and identifies many mitochondrial associated differentially expressed genes



Literature-Supported Gene Set

- Did not find clinical groups to be rlog normalized gene expression, or top 300-1000 variable genes
- Study identifies 25 literature-supported AD-risk genes
  - Use in exploratory data analysis phase

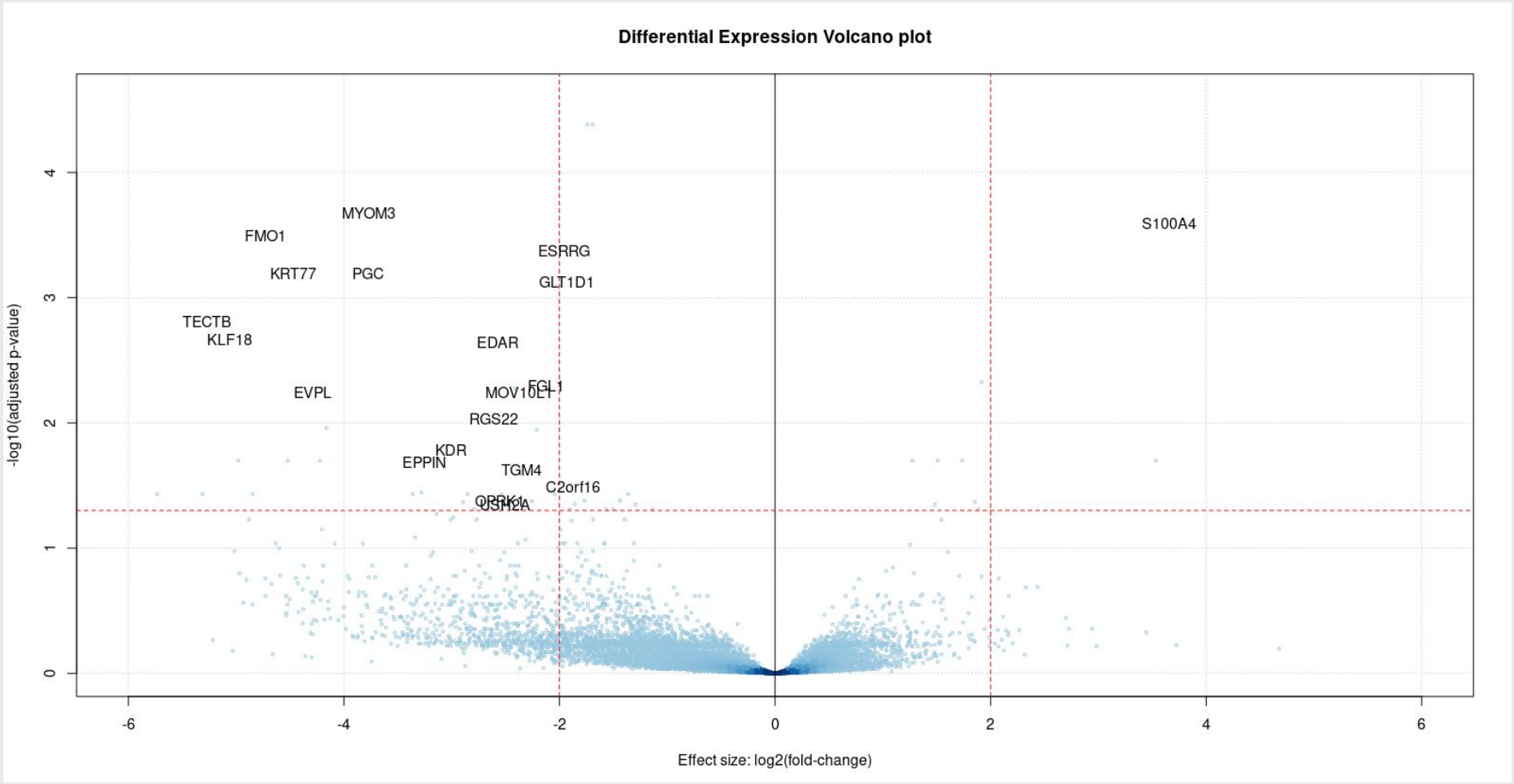


Differentially Expressed Genes

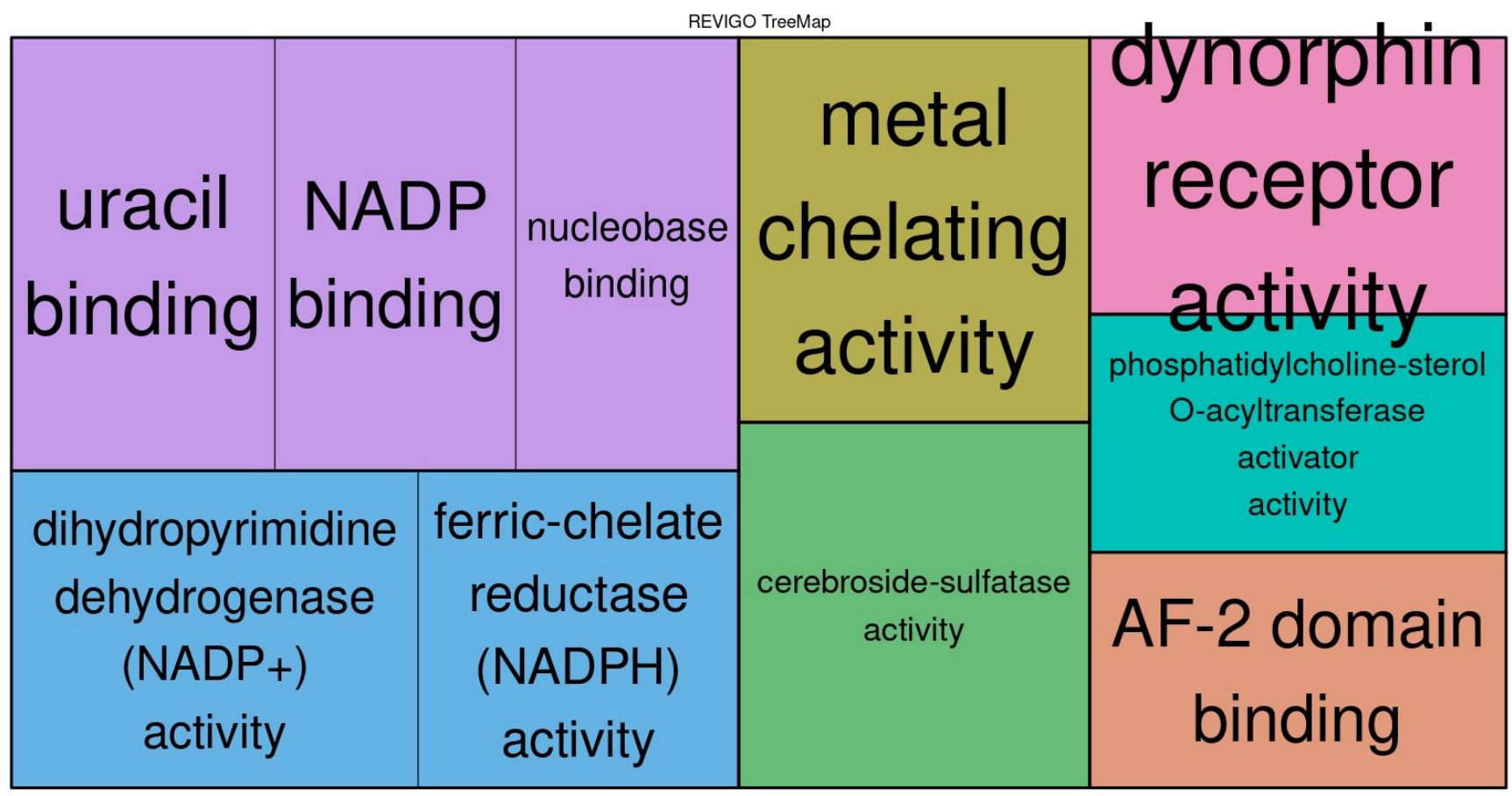
- **Confirmed 14 genes to be differentially expressed between control and AD clinical groups**
- **Same 14 genes appear in paper as gathered literature-supported AD risk genes**

	padj	log2FoldChange
SERPINF1	0.000041	1.7407
CECR2	0.000041	1.6923
TGFBI	0.004718	-1.9188
GLT1D1	0.005319	2.4133
ARSA	0.019995	-1.2752
PTPRG	0.019995	-1.7361
S100A4	0.019995	-3.5340
APOE	0.019995	-1.5115
MOV10L1	0.037131	2.5397
ASTN1	0.041758	1.4381
DPYD	0.042672	-1.8551
STEAP3	0.044855	-1.4818
IGSF10	0.048952	1.5660
PSTPIP1	0.048952	-1.8852

Differential Expression Volcano Plot



Gene Ontology Treemap -- Revigo



Weill Cornell Medicine

Discussion

Interesting Identified Genes

SERPINF1 (serpin family F member 1) – The encoded protein is secreted and strongly **inhibits angiogenesis**. In addition, this protein is a **neurotrophic factor** involved in neuronal differentiation in retinoblastoma cells. Mutations in this gene were found in individuals with osteogenesis imperfecta.

CECR2 (CECR2 histone acetyl-lysine reader) – Involved in chromatin remodeling, and may additionally play a role in **DNA damage response**. The encoded protein functions as part of an ATP-dependent complex that is involved in neurulation.

ARSA (arylsulfatase A) – Defects in this gene lead to metachromatic leucodystrophy (MLD), a progressive demyelination disease which results in a **variety of neurological symptoms** and ultimately death.

S100A4 (S100 calcium binding protein A4) – S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the **regulation of a number of cellular processes such as cell cycle progression and differentiation**.

	padj	log2FoldChange
SERPINF1	0.000041	1.7407
CECR2	0.000041	1.6923
TGFBI	0.004718	-1.9188
GLT1D1	0.005319	2.4133
ARSA	0.019995	-1.2752
PTPRG	0.019995	-1.7361
S100A4	0.019995	-3.5340
APOE	0.019995	-1.5115
MOV10L1	0.037131	2.5397
ASTN1	0.041758	1.4381
DPYD	0.042672	-1.8551
STEAP3	0.044855	-1.4818
IGSF10	0.048952	1.5660
PSTPIP1	0.048952	-1.8852

Weill Cornell Medicine